Sensitivity of the Chi-Squared Goodness-of-Fit Test
to the Partitioning of Data

Gianna Boero

Jeremy Smith

And

Kenneth F. Wallis

No 694

WARWICK ECONOMIC RESEARCH PAPERS

DEPARTMENT OF ECONOMICS

THE UNIVERSITY OF
WARWICK

Sensitivity of the chi-squared goodness-of-fit test to the partitioning of data

Gianna Boero, Jeremy Smith and Kenneth F. Wallis

Department of Economics
University of Warwick
Coventry CV4 7AL, UK

January 2004

Abstract In this paper we conduct a Monte Carlo study to determine the power of Pearson's overall goodness-of-fit test as well as the "Pearson analog" tests (see Anderson (1994)) to detect rejections due to shifts in variance, skewness and kurtosis, as we vary the number and location of the partition points. Simulations are conducted for small and moderate sample sizes. While it is generally recommended that to improve the power of the goodness-of-fit test the partition points are equiprobable, we find that power can be improved by the use of non-equiprobable partitions.

Keywords: Pearson's Goodness-of-fit test; Distributional assumptions; Monte Carlo; Normality; partitions.

JEL classification: C12, C14

Acknowledgements: The authors would like to thank Gordon Anderson and Peter Burridge as well as seminar participants at the ESEM conference 2003 for helpful comments and suggestions.

1. Introduction

Goodness of fit or the degree of correspondence between observed outcomes and expected outcomes based upon a postulated distribution is a cornerstone of classical statistics. The two classical nonparametric approaches to testing goodness of fit (as surveyed by Stuart, Ord and Arnold (1999, Ch. 25)) are (i) Pearson's goodness-of-fit (X^2) test, which involves grouping data into classes and comparing observed outcomes to those hypothesised under some null distribution; and (ii) Kolmogorov-Smirnov (K-S) test, which involves comparing the empirical cumulative distribution function (cdf) with a cdf obtained under some null hypothesis.

Anderson (1994) has devised a method to decompose the X^2 test into a series of individual component tests (see Boero, Smith and Wallis, 2004), in which each component test focuses on a different moment of the distribution, in an attempt to provide more information on the nature of any rejection of the null hypothesis by the X^2 test.

In spite of the wide use of the X^2 test it is still not clear how many partition points (class intervals) should be used in the construction of this test, and how these class intervals should be formed. However, it is generally recommended that researchers use equiprobable partitions, see Stuart, Ord and Arnold (1999).

In this paper we conduct a systematic analysis of the power of the X^2 test in relation to the number of partitions points. We also investigate the sensitivity of the power of the X^2 test, as well as its component tests, to the location of the partition points, that is, to the choice between equiprobable or non-equiprobable splits. The power of the X^2 test is examined with respect to departures in variance, skewness and kurtosis from a null distribution of a N(0,1). The power of the X^2 and its component tests is compared to that of the K-S statistic, additionally, we compare these tests to

the chi-squared test (for the population variation), when looking at variance departures and to the Jarque and Bera (1980) (J-B) test when looking at departures in skewness or kurtosis.

The paper proceeds as follows. Section 2 gives a brief description of the X^2 test and the literature on the choice of the number and location of the partition points. We also discuss Anderson's (1994) method of decomposing the X^2 test into its component tests, as well as the K-S test. Section 3 outlines the alternative distributions used to generate artificial data under the alternative hypotheses against which we test the null hypothesis of N(0,1). Section 4 reports the results of the X^2 and its component tests to detect departures from the null hypothesis using both equiprobable and non-equiprobable partitions. Finally, in Section 5 we summarise the main results and offer some concluding remarks.

2. Goodness-of-fit tests

2.1. The chi-squared goodness-of-fit (X^2) test

Pearson's classical goodness-of-fit (X^2) test proceeds by dividing the range of the variable into k mutually exclusive classes and comparing the probabilities of outcomes falling in these classes given by the hypothesised distribution with the observed relative frequencies. With class probabilities $p_i > 0,\ i = 1,\ldots,k$, $\sum_{i=1}^{k} p_i = 1$ and observed class frequencies $n_i > 0,\ i = 1,\ldots,k$, $\sum_{i=1}^{k} n_i = N$, the test statistic is

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

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 X^2 test with different numbers of partitions (k) 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, have focused on asymptotic results and have assumed equiprobable classes (such that, $p_i = 1/k, i = 1,\ldots,k$). Mann and Wald (1942) have suggested equiprobable class intervals and develop a formula for the optimal choice of the number of classes, which depends on the sample size, N, and the level of significance. For equiprobable splits the formula for the choice of cells is $k = 3.765(N\text{-}1)^{0.4}$, at the 5% significance level. Table 1 reports values of k for selected values of N based upon this formula. Although Mann and Wald's recommendation (p.307) is based on asymptotic theory, they suggest that the results hold approximately for sample sizes as low as 200 and may be true for considerably smaller samples.

Table 1

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

The advantages of the Mann and Wald technique are that the application of the formula removes the subjective element from the choice of the number and width of the classes and equiprobable classes are easy to use and lead to unbiased tests (see also Gumbel, 1943, and Cohen and Sackrowitz, 1975). However, various numerical studies have presented empirical evidence to show that the value of k proposed by Mann and Wald is too large, resulting in loss of power in many situations. Williams (1950) indicates that the value of k as given by the Mann and Wald formula 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, N. In a comparison of the power of the X^2 test and the likelihood ratio (LR) 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 quite similar to 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 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 (p. 959).

Most of the results summarised above are based on asymptotic theory. Only a limited number of cases have been examined to validate the asymptotic theory, Kallenberg et al. (1985) present some results for N=50 and N=100, while Koehler and Gan (1990) report results for N=20, 50 and 100. In contrast little is known for cases with non-equiprobable partitions. In general, therefore the evidence that has been produced is often contradictory, leaving the problems of how many partitions to use and how to choose them largely unresolved.

2.2. X^2 component tests or 'Pearson analog' tests

Anderson (1994) presented a method for decomposing the X^2 test into (k-1) independent $\chi^2(1)$ component tests as:

$$X_j^2 = \upsilon_j^2 / n\sigma_j^2 \quad j = 1,\ldots,k-1$$

where $v_j = i_j^{'}(x-\mu)$, x is a vector (kx1) of observed frequencies with mean μ, and i_j is a set of k dimensioned vectors (the decompositions are only really possible for k=4, 8, 16 and 32). For example, for $k = 8$, we can write the first four vectors of i_j as:

$$i_m = \{1 \quad 1 \quad 1 \quad 1 \quad -1 \quad -1 \quad -1 \quad -1\}$$

$$i_{sc} = \{1 \quad 1 \quad -1 \quad -1 \quad -1 \quad -1 \quad 1 \quad 1\}$$

$$i_{sk} = \{1 \quad 1 \quad -1 \quad -1 \quad 1 \quad 1 \quad -1 \quad -1\}$$

$$i_k = \{1 \quad -1 \quad -1 \quad 1 \quad 1 \quad -1 \quad -1 \quad 1\}$$

Defining $p = (p_1,\ldots,p_8)$, where p_j are the class probabilities, the variance is written:

$$\sigma_j^2 = 1 - (i_j p')^2 \quad j = m, sc, sk, k$$

From the form of the i_j vectors, the first 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, with 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. There are three remaining component tests, but these have no obvious interpretation. Boero, Smith and Wallis (2004) present a more theoretical derivation

of the component tests and show that the independence of the component tests does not hold when using non-equiprobable splits, although each component test still has a $\chi^2(1)$ distribution. An application of this "Pearson analog" test to the comparison of income distributions is given by Anderson (1996). It has also been used in density forecast evaluation by Wallis (2003) and Boero and Marrocu (2003).

While 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 for k=8, assuming that the eight partitions are obtained by dividing each of the original four partitions 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 j^{th} partition as F_j. The first and last partitions are open-ended, thus with F_0 =0 and F_k =1 the partition probabilities satisfy, $p_j = F_j - F_{j-1}$, j=1,…,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 the four sign

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

2.3. Kolmogorov-Smirnov test

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

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

where Z is the theoretical cdf under the null hypothesis, and A is the empirical cdf.

3. Experiments used in the Monte Carlo study

The Monte Carlo experiments in this paper 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 variance, skewness or kurtosis. In this section we outline the nature of these alternative distributions.

Experiment A: Non-unit variance

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

Experiment B: Skewed distributions

B_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, …, 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. This distribution has been used in a recent study by Noceti, Hodges and Smith (2003) who summarise the results from a Monte Carlo study of the relative power of some distributional tests.

B_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

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

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, …, 0.90}.

B_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≈2×*d* we set *d*={0.00, 0.025, …, 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.

C: Distributions with heavy tails.

C_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 Mandelbrot, 1963, Fama, 1965, McCulloch, 1994, and Rachev and Mittnik, 2000).

C_2*: *Anderson's kurtotic distribution (see Anderson, 1994)

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

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 to 7.0.

The results of the Monte Carlo experiments reported in this paper are based on 5000 replications for sample sizes N=25, 50, 75, 100, 150, 250, 350. All tests are undertaken at the 5% significance level. For equiprobable partitions we take k=2,4,…,40. For non-equiprobable partitions we take k = 4,8,16,32 and consider partitions which are symmetric around 0.5, such that for k=8, {F_1, F_2, F_3, 0.5, 1-F_3, 1-F_2, 1-F_1}.

4. The power of the tests

4.1. Departure from unit variance

We now analyse the power of the various tests for departures from the null hypothesis of N(0,1) due to a non-unit variance (experiment ***A***).

4.1.1. 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 Figure 1a for a wide range of alternatives with excess variance (thicker tails), and in Figure 1b for alternatives with variance smaller than one (thinner tails), 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).

From Figures 1a and 1b it is evident that, for most alternatives, power is maximised for a value k in the interval four to ten. 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 for thick-tailed alternatives, while there seems to be considerable loss in power for thinner-tailed alternatives for values of k greater than 10.

Table 2 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 δ<1, and at

k around 8 and 10 for alternatives with δ>1. Moreover, the power of the X^2 test to detect departures from a unit variance is slightly 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 2 report the power of the K-S test and the χ^2 test for the variance of a population, $(N\text{-}1)s^2/\sigma^2$, with N-1 degrees of freedom. 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 χ^2 test.

4.1.2. Non-equiprobable classes

As discussed earlier the power of the scale component test (PCSc) (and hence the X^2 test) depends upon F_2 and F_6. In experiment ***A*** 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 to 0.3 through steps of 0.025. For k=8, Figures 2a and 2b 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 2b shows that the power of the PCSc test unambiguously falls as F_2 increases for all δ. By comparison, Figure 2a shows that the power of the X^2 test falls as F_2 increases only for δ>1, and is largely insensitive to F_2 for δ<1. For example, for the X^2 test when δ=1.2, power increases to 55% for F_2=0.15, compared with 39% for F_2=0.25, whereas when δ=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 δ<1 arises from looking at the performance of the other component tests for $\delta < 1$. We have found 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, thereby offsetting the falling power of the PCSc test in the X^2 test. 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 2b 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 intervals. For example, for $\delta = 1.2$, power increases to 67% for $F_2 = 0.15$, compared with 46% for $F_2 = 0.25$, and to 86% from 73% when $\delta = 0.8$.

Using $F_2 = 0.15$, Table 2 (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 $\delta < 1$. Also the results indicate that with non-equiprobable partitions the power of the X^2 test is not very sensitive to increasing k from 4 to 8 or 16, providing there are sufficient observations to avoid sparsity in some partitions.

Finally, in Figure 3 we summarise the power results of the X^2 test obtained for $\delta \neq 1$ and $N = 150$, using $k = 10$ equiprobable intervals, and $k = 8$ non-equiprobable intervals (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 ($\chi^2(N-1)$). As we can see, the best performance is given by the latter (chi-sq in the figure). 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.

The results discussed above, 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 that unbalanced partitions with some small partitions in the tails lead to important gains in power when the alternative has heavy tails (δ>1). These findings are consistent with the suggestion in Kallenberg et al. (1985).

4.2. Departures due to Skewness

4.2.1. 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 Figures 4 to 6. Figure 4 reports the results for the Ramberg distribution ($\boldsymbol{B_1}$), with skewness ranging between 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 remarkably 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 Figure 4 we observe that, for N=150, power increases from approximately 14% (k=4) to 58% (k=10) when skewness is 0.4, and from 28% (k=4) to 91% (k=6) and to 100% (k=8) when skewness is 0.8.

Results for different sample sizes, reported in Table 3, 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 between 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 ($\boldsymbol{B_2}$), see Figure 5, which reports the power of the X^2 test for skewness=0.5 and different sample sizes. Figure 6 reports the power of the X^2 test for Anderson's skewed distribution ($\boldsymbol{B_3}$), for N=150 and different degrees of skewness, we observe a less prominent impact on power by increasing k from 4 to 8 or 10.

4.2.2. Non-equiprobable classes

As discussed earlier 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 partitions, we have $\{F_1, F_2,\ldots,F_7\}= \{F_2/2, F_2, (0.5+F_2)/2, 0.5, 1-(0.5+F_2)/2, 1-F_2, 1-F_2/2\}$, and conduct experiments for values of F_2 in the range 0.15 to 0.3 in steps of 0.025. In Figures 7a and 7b we plot the power of the X^2 and PCSk tests, respectively, against values of F_2 for N=150 and k=8, for $\boldsymbol{B_1}$ (Ramberg) with skewness ranging from 0.2 to 0.8.

From Figures 7a and 7b 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 (X^2) is 56% (67%) compared to 25% (51%) when using 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 $\boldsymbol{B_2}$ (two-piece normal) and $\boldsymbol{B_3}$ (Anderson's skewed distribution) are qualitatively similar to those reported in Figures 7a and 7b, although for $\boldsymbol{B_3}$ the gains to using non-equiprobable partitions are smaller than observed for $\boldsymbol{B_1}$ and $\boldsymbol{B_2}$. 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.

Figures 8a and 8b plots the power of the X^2 test for $\boldsymbol{B_1}$ and $\boldsymbol{B_3}$ with N=150, for k=8 equiprobable partitions, k=8 non-equiprobable partitions (F_2=0.15), 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 3 for $\boldsymbol{B_1}$ and $\boldsymbol{B_2}$.

First of all, from Figures 8a and 8b and Table 3, we observe for all tests a significant increase in power for increasing degrees of skewness. Figure 8a and Table 3 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 3, and both alternatives $\boldsymbol{B_1}$ and $\boldsymbol{B_2}$. Figure 8a and Table 3 also report the performance of the K-S and J-B tests. The results indicate that the best performance for $\boldsymbol{B_1}$ and $\boldsymbol{B_2}$ is achieved by the J-B test and the worse performance by the K-S test. For example, for alternative $\boldsymbol{B_1}$, and skewness 0.6 (N=150), the power of the J-B test is about 80%, while it is only 25% for the K-S test. It is also clear that, with k=8 non-equiprobable partitions, 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% for skewness of 0.6).

A different ranking of the tests is obtained from the experiments conducted under $\boldsymbol{B_3}$. Figure 8b shows that, in this case, the X^2 test clearly dominates the J-B test. Finally, we notice that, differently from the results discussed above, with non-equiprobable partitions doubling k from 4 to 8 has no effect on power.

4.3. Departures due to kurtosis

4.3.1. 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 Figures 9 and 10 for $\boldsymbol{C_1}$ (Stable distribution) and $\boldsymbol{C_2}$ (Anderson's kurtotic distribution), respectively, for k varying from 2 to 40 and N=150. As we can see from Figure 9, 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 the Anderson's distribution, reported in Figure 10 where we can see that the power of the X^2 test approaches 100% for kurtosis of about 6. Moreover, for both alternatives, we notice an improvement in the power of the test moving from 4 to 8 (or 10) partitions, after which the test does not seem to be sensitive to further increases in k.

4.3.2. Non-equiprobable classes

As discussed earlier, 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. In Figures 11a and 11b we plot the power of the X^2 and PCK tests, respectively, against values of F_1 and F_3 for N=150 and k=8 classes, for $\boldsymbol{C_1}$ with $\alpha = 1.95$.

Figure 11a 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 (Figure 11b) is maximised 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 $\boldsymbol{C_2}$ (Anderson's kurtotic distribution) we have found that the power of both the X^2 test and the PCK component test was maximised at F_1=0.05 and F_3=0.45. In this case while the PCSc test has shown some power (the sample variance for this experiment was approximately unity), it was markedly smaller than the power of the PCK component test.

Figures 12 and 13 explore further the performance of the X^2 test under alternatives $\boldsymbol{C_1}$ and $\boldsymbol{C_2}$, respectively, when the test is computed using non-equiprobable partitions. The figures also report the performance of the J-B and K-S statistics. Figure 12 shows that the X^2 and PCK tests dominate the other tests in terms of power. In particular, we can see that the maximum power exhibited by the K-S test is 60%, while the power of the J-B test ranges between 20% and 50% for the stability parameter α between 2 and 1.85, and becomes comparable to that of the X^2 test for smaller values of α. Also Figure 12 illustrates some gains in power for the X^2 and PCK tests by using non-equiprobable partitions. The power of the X^2 test is, in fact, between 80% and 90% with 8 and 16 equiprobable partitions, and between 90% and 100% with non-equiprobable partitions. The partitions were formed by setting F_1=0.05 and F_3=0.45, with k=8.

In Figure 13 we report the results for $\boldsymbol{C_2}$, for X^2 computed with k=10 equiprobable partitions and k=8 non-equiprobable partitions (F_1=0.05 and F_3=0.45), and for the PCK test (F_1=0.05 and F_3=0.45). It is clear that greater gains are achieved by the X^2 test for this distribution (compared to $\boldsymbol{C_1}$) with the use of non-equiprobable partitions relative to equiprobable partitions. For example, for kurtosis of 4.4, k=10 equiprobable partitions deliver power of just above 50%, while power can reach

values above 90% with the use of k=8 non-equiprobable partitions. Table 4 presents results for different sample sizes. Comparing the performance of the X^2 test with that of the K-S and J-B tests, we can see that the X^2 test is superior to the K-S statistic for all values of k, and it is also superior to the J-B test for kurtosis higher than 3.6.

4. Conclusions

In this paper we have summarised 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. We have made suggestions for values of k to use in testing for normality against special alternatives of interest, in the case of equiprobable intervals, and show that there can be significant gains in power from the use of non-equiprobable intervals.

The simulations have been designed to detect departures from a standard normal distribution, in the presence of changes in variance, skewness and kurtosis. The relative performance of the X^2 test has been compared against that of the K-S statistic, a simple test about the variance of a population, and the J-B test.

In synthesis, two main results seem to apply in general to small and moderate sample sizes, and stand against common practical recommendations. First, our simulations indicate that the 'optimal' number of cells is smaller than that recommended in previous studies, most of which are based on asymptotic results.

Second, the use of non-equiprobable partitions 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. Specific choices of non-equiprobable partitions have shown to improve the performance of the X^2 test over the K-S test, and to reduce substantially its disadvantage with respect to the moment-based tests considered in this paper.

References

Anderson, G. (1994). "Simple tests of distributional form", *Journal of Econometrics*, 62, 265-276.

Anderson, G. (1996). "Nonparametric tests of stochastic dominance in income distributions". *Econometrica,* 64, 1183-1193.

Boero, G., Smith, J. P. and Wallis, K. F. (2004). "Decompositions of Pearson's Chi-Squared Test". *Journal of Econometrics*, forthcoming.

Boero, G. and Marrocu E. (2003). "The performance of SETAR models by regime: a conditional evaluation of interval and density forecasts", *International Journal of Forecasting*, forthcoming.

Chambers, J. M., Mallows, C. L. and Stuck, B. W. (1976). "A method for simulating stable random variables". *Journal of the American Statistical Association*, 71, 340-344.

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

Dahiya, R. C. and Gurland J. (1973). "How many classes in the Pearson chi-square test?", *Journal of the American Statistical Association*, 68, 678-89.

Fama, E. F. (1965). "The behaviour of stock market prices". *Journal of Business*, 38, 34-105.

Gumbel, E. J. (1943). "On the reliability of the classical chi-square test". *Annals of Mathematical Statistics*, 14, 253-63.

Hamdan, M.A. (1963). "The number and width of classes in chi-square test". *Journal of the American Statistical Association*, 58, 678-89.

Kallenberg, W. C. M., Oosterhoff, J. and Schriever, B. F. (1985). "The number of classes in chi-squared goodness-of-fit tests". *Journal of the American Statistical Association*, 80, 959-968.

Koelher, K. J. and Gan, F. F. (1990). "Chi-squared goodness-of-fit tests: cell selection and power". *Communications in Statistics*, B, 19, 1265.

Mandelbrot, B. (1963). "New methods in statistical economics". *Journal of Political Economy*, 71, 421-440.

Mann, H. B. and Wald, A. (1942). "On the choice of the number of class intervals in the application of the chi-square test". *Annals of Mathematical Statistics*, 13, 306-317.

McCulloch, J. H. (1994). Financial applications of stable distributions. In *Statistical methods in finance* (Maddala, G.S., and Rao, C.R., Eds.) Elsevier, pp.393-425.

Noceti, P, Smith, J. and Hodges, S. (2003). "An evaluation of tests of distributional forecasts", *Journal of Forecasting*, 22, 447-455.

Rachev, Svetlozar T. and Mittnik, S. (2000). *Stable Paretian models in finance*. Wiley.

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

Stuart, A., Ord, J.K. and Arnold, S. (1999). *Kendall's advanced theory of statistics*, 6th ed., vol. 2A. London: Edward Arnold.

Uchaikin, V. V. and Zolotarev, V. M. (1999). *Chance and stability, stable distributions and their applications*. VSP Utrecht, The Netherlands.

Wallis, K.F. (1999). "Asymmetric density forecasts of inflation and the Bank of England's fan chart". *National Institute Economic Review*, No. 167, 106-112.

Wallis, K.F. (2003). "Chi-squared tests of interval and density forecasts, and the Bank of England's fan charts". *International Journal of Forecasting*, 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", *Journal of the American Statistical Association*, 45, 77-86.

Table 2: Power of the X^2 test of N(0,1) vs N(0,d)
Alternative A

		Equiprobable splits			Non-equiprobable splits				
	δ	k=4	k=8	k=16	k=4	k=8	k=16	ch-sq.	K-S
N=25									
	0.6	47.8	40.8	27.7	44.6	26.7	13.2	96.9	12.0
	0.7	23.2	22.4	15.5	20.3	13.0	8.1	75.1	9.3
	0.8	9.1	9.6	8.9	9.0	6.7	4.8	39.9	5.5
	0.9	5.3	6.7	7.0	4.8	6.1	4.3	15.5	6.3
	1.1	4.6	5.9	6.0	7.8	9.8	9.2	18.2	6.0
	1.2	8.2	10.7	8.5	13.8	14.6	14.8	39.0	8.1
	1.3	9.5	13.1	13.1	20.7	24.2	24.2	60.6	10.8
	1.4	15.4	21.0	21.9	32.1	36.9	36.2	77.4	15.4
	1.5	18.8	31.4	33.5	43.4	47.6	50.0	88.1	20.7
N=50									
	0.6	85.6	85.3	67.1	91.0	79.6	53.2	100.0	42.9
	0.7	50.6	48.6	31.4	55.9	39.1	21.4	97.0	16.4
	0.8	19.5	19.3	13.8	21.1	15.4	8.4	67.8	8.2
	0.9	7.3	8.0	6.3	8.5	6.8	5.0	24.6	4.4
	1.1	7.5	6.7	6.0	11.2	10.6	10.8	26.3	5.5
	1.2	13.6	14.9	13.5	19.8	21.9	21.0	59.3	11.8
	1.3	19.2	24.1	24.0	34.5	38.0	37.2	83.9	16.0
	1.4	31.0	41.7	45.2	51.6	59.6	59.1	95.1	24.6
	1.5	39.2	59.4	64.1	69.5	78.6	79.7	98.8	36.0
N=100									
	0.6	99.6	99.7	98.9	100.0	99.8	98.6	100.0	90.2
	0.7	82.0	87.1	76.2	91.3	87.4	66.9	100.0	46.5
	0.8	38.0	40.4	30.5	47.4	37.5	23.6	92.9	14.9
	0.9	10.6	10.0	10.3	11.3	10.1	7.2	40.8	6.7
	1.1	9.3	11.8	9.1	15.4	13.6	15.5	40.2	9.3
	1.2	21.6	29.0	27.6	37.3	38.9	36.4	83.0	15.5
	1.3	39.2	51.7	53.2	61.8	67.6	66.2	97.7	29.1
	1.4	58.2	74.8	80.8	81.6	88.6	88.6	99.8	48.2
	1.5	74.8	90.4	93.8	95.3	97.9	98.2	100.0	70.3
N=150									
	0.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.5
	0.7	94.5	97.8	96.3	99.0	98.6	94.3	100.0	74.3
	0.8	54.5	58.9	48.2	66.9	62.3	41.5	98.7	22.4
	0.9	12.2	12.0	9.0	16.5	11.3	9.1	54.6	7.3
	1.1	11.8	15.0	11.1	20.3	18.8	16.9	51.9	9.2
	1.2	27.7	38.7	40.5	49.2	54.5	52.5	93.4	19.9
	1.3	56.2	72.3	75.7	80.4	85.7	85.4	99.7	42.0
	1.4	75.5	91.9	94.1	94.4	97.0	97.6	100.0	70.1
	1.5	91.4	98.2	99.6	99.0	100.0	100.0	100.0	89.4
N=250									
	0.6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	0.7	99.3	100.0	100.0	100.0	100.0	100.0	100.0	97.2
	0.8	77.8	85.5	79.0	90.8	89.7	76.8	100.0	47.8
	0.9	20.0	21.5	17.5	27.1	22.5	15.5	74.8	9.6
	1.1	18.6	18.9	17.7	24.9	27.0	24.5	69.7	11.8
	1.2	46.3	61.6	63.1	71.1	75.1	73.9	99.1	33.9
	1.3	79.6	94.1	95.6	95.6	97.7	97.7	100.0	67.8
	1.4	94.1	99.1	99.5	99.5	99.7	99.9	100.0	94.1
	1.5	99.8	99.9	100.0	99.9	100.0	100.0	100.0	99.6

Table 3: Power of the X^2 test against skewness

Alternative B_1: Ramberg distribution

	Equiprobable splits			Non-equiprobable splits				
skew	k=4	k=8	k=16	k=4	k=8	k=16	K-S	J-B
N=50								
0.5	6.1	11.1	11.9	12.5	14.6	13.4	9.8	10.9
0.6	7.1	14.5	15.6	16.6	20.4	16.6	11.9	15.9
0.7	11.1	26.0	23.2	27.5	29.8	26.8	15.8	28.0
0.8	11.9	64.1	50.1	52.4	55.1	49.6	20.8	33.9
N=100								
0.5	8.9	21.0	21.6	23.0	27.6	25.0	14.7	30.8
0.6	10.1	33.2	33.4	32.9	45.1	38.1	18.8	48.2
0.7	16.9	59.8	58.4	57.9	72.7	62.4	25.7	75.1
0.8	19.8	97.9	90.6	89.7	94.4	91.3	35.1	91.8
N=150								
0.5	11.5	30.4	33.8	34.2	45.4	37.5	19.6	55.3
0.6	13.8	51.1	53.0	49.5	66.7	59.3	24.7	80.6
0.7	22.1	82.6	83.9	79.5	94.6	86.8	35.0	96.8
0.8	28.2	100.0	99.6	97.4	99.9	99.9	48.3	99.8
N=250								
0.5	18.0	50.2	58.6	54.9	72.9	63.1	27.5	86.3
0.6	24.3	80.2	85.3	76.2	92.9	88.4	41.0	98.9
0.7	35.2	98.0	99.2	96.3	100.0	99.8	54.1	100.0
0.8	45.6	100.0	100.0	100.0	100.0	100.0	78.0	100.0

Alternative B_2: Two-piece normal

	Equiprobable splits			Non-equiprobable splits				
skew	k=4	k=8	k=16	k=4	k=8	k=16	K-S	J-B
N=50								
0.5	7.2	8.0	9.0	9.0	8.3	9.8	7.6	14.1
0.6	9.3	14.3	14.0	15.1	17.7	15.4	12.3	20.5
0.7	7.1	18.2	16.0	18.6	20.4	17.1	11.2	25.1
0.8	9.5	25.6	22.4	26.4	31.1	26.4	13.8	34.8
N=100								
0.5	8.7	14.7	16.5	18.2	19.4	17.9	13.3	32.8
0.6	9.1	24.5	25.2	24.9	29.6	25.6	14.8	44.4
0.7	10.7	37.1	34.7	39.1	45.6	38.5	19.4	65.8
0.8	13.5	57.5	54.6	54.4	66.3	56.5	22.5	77.9
N=150								
0.5	10.6	22.1	24.6	25.6	29.2	26.0	18.1	53.1
0.6	12.5	40.3	38.9	38.9	48.1	39.6	19.6	68.3
0.7	13.2	54.8	55.1	54.8	70.5	60.5	26.1	89.8
0.8	18.7	79.5	78.4	74.4	89.2	81.4	34.6	95.8
N=250								
0.5	14.8	38.0	42.3	42.0	54.6	46.6	24.2	80.6
0.6	19.5	62.3	67.0	61.4	79.7	70.2	32.1	94.0
0.7	22.1	82.6	85.8	80.7	94.7	90.6	38.9	99.3
0.8	27.4	97.7	98.9	95.2	99.8	99.0	51.6	100.0

Table 4: Power of the X^2 test against kurtosis

Alternative C_2: Anderson's kurtotic distribution

	Equiprobable splits			Non-equiprobable splits				
kurtosis	k=8	k=16	k=32	k=8	k=16	k=32	K-S	J-B
N=25								
2.0	12.4	11.2	12.5	3.3	3.7	1.6	13.8	0.1
2.8	4.8	5.5	5.9	4.7	4.6	6.8	6.0	1.1
3.2	5.4	4.5	6.8	8.6	12.3	14.8	4.3	4.1
4.0	8.2	7.4	8.3	25.4	29.0	37.4	8.0	11.3
4.8	15.5	14.1	13.2	47.8	52.1	60.7	10.6	18.7
5.6	20.1	20.8	19.6	64.3	71.9	78.5	11.5	26.3
6.4	29.0	30.4	29.0	72.5	77.6	84.9	14.2	33.8
7.2	30.1	32.9	33.2	80.3	85.4	91.2	16.6	38.7
N=50								
2.0	23.8	21.2	14.7	9.4	9.7	3.1	19.9	0.0
2.8	6.7	5.8	4.3	4.1	4.0	4.4	6.8	1.9
3.2	5.5	4.8	3.6	8.3	10.2	13.8	4.8	5.8
4.0	11.4	11.0	10.0	35.3	40.9	45.0	9.1	21.8
4.8	24.1	23.8	20.0	66.7	72.7	78.7	14.5	37.0
5.6	38.4	39.5	34.5	84.8	90.0	92.4	21.2	49.2
6.4	52.6	55.5	49.7	92.6	95.0	96.5	30.8	62.5
7.2	60.1	63.5	59.4	96.1	98.1	98.8	35.1	67.1
N=100								
2.0	47.2	46.7	37.9	39.3	37.7	15.9	28.9	14.6
2.8	6.2	5.4	5.8	4.1	4.2	3.2	6.3	1.8
3.2	5.4	5.2	5.0	7.8	9.5	13.1	4.7	8.4
4.0	20.9	20.5	18.7	55.1	60.9	64.6	13.7	33.8
4.8	48.9	47.0	44.3	90.0	93.5	93.9	26.2	61.8
5.6	70.9	72.2	69.1	98.9	98.8	99.2	44.4	78.8
6.4	86.3	86.4	83.9	99.7	100.0	99.9	61.7	88.9
7.2	92.6	95.0	93.2	100.0	100.0	100.0	71.8	93.1
N=150								
2.0	69.7	71.9	60.4	74.9	70.9	42.3	40.7	59.3
2.8	6.2	7.0	6.3	3.7	4.6	3.7	5.0	2.2
3.2	5.6	4.9	5.4	9.7	11.0	13.3	5.5	8.8
4.0	30.9	29.9	26.9	68.4	73.9	75.5	16.7	45.0
4.8	67.8	70.5	62.9	97.9	98.9	98.7	40.5	79.0
5.6	88.3	91.4	88.3	99.7	99.9	100.0	67.3	89.5
6.4	97.1	98.1	96.8	100.0	100.0	100.0	82.6	97.3
7.2	98.7	98.8	99.3	100.0	100.0	100.0	91.5	98.8

Alternative C_1: Stable distribution

	Equiprobable splits			Non-equiprobable splits				
N	k=8	k=16	k=32	k=8	k=16	k=32	K-S	J-B
α=1.975								
25	21.6	25.7	25.7	41.3	45.2	47.0	14.9	5.4
50	43.9	48.4	46.0	70.2	69.9	68.3	25.8	9.2
75	63.5	70.4	68.9	86.4	86.9	85.0	37.0	12.9
100	79.0	84.5	82.1	93.9	93.5	93.2	50.9	16.5
150	93.1	95.7	94.8	99.4	99.3	99.2	74.8	19.7
250	99.9	99.7	100.0	100.0	100.0	100.0	96.5	29.6
350	100.0	100.0	100.0	100.0	100.0	100.0	99.1	34.3

Figure 1a: Power of X2 test against variance departure, N=150, δ>1 (equiprobable)

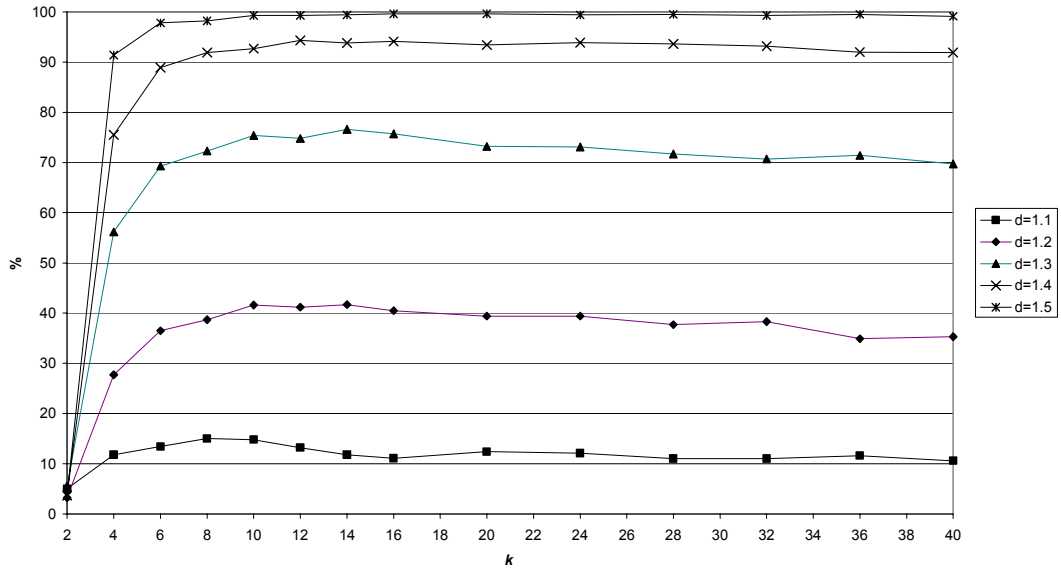

Figure 1b: Power of X^2 test against variance departure, N=150, δ<1 (equiprobable)

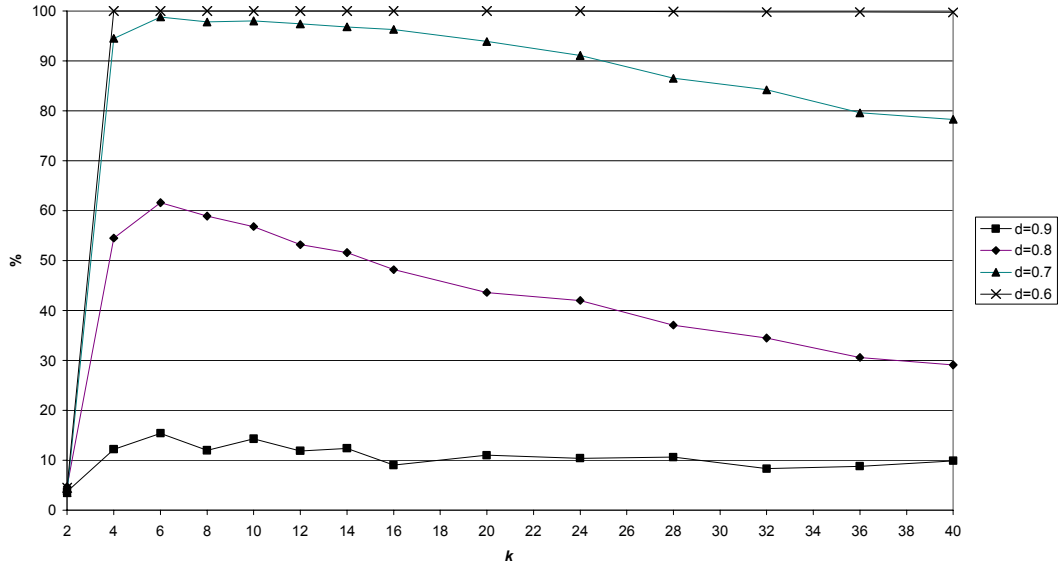

Figure 2a: Power of X^2 test against variance departure, N=150 (non-equiprobable)

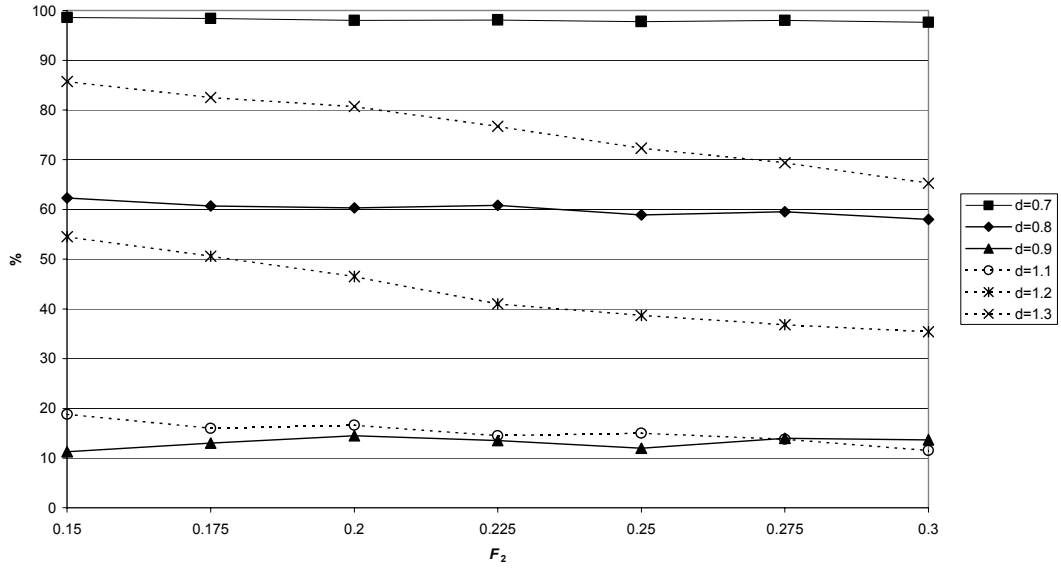

Figure 2b: Power of the PCSc test against variance departure, N=150 (non-equiprobable)

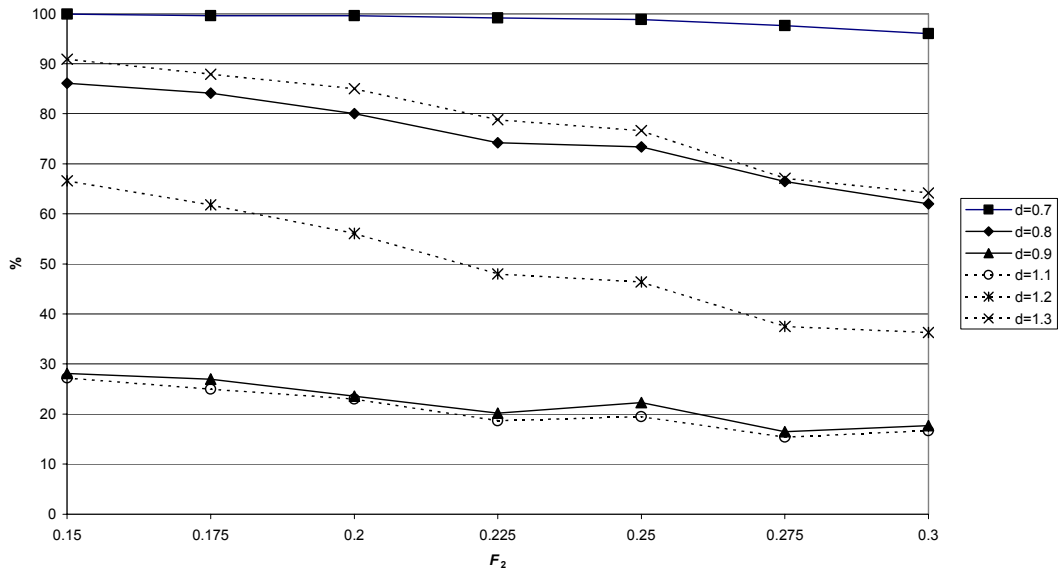

Figure 3: Power of the X^2 test against variance departure, N=150

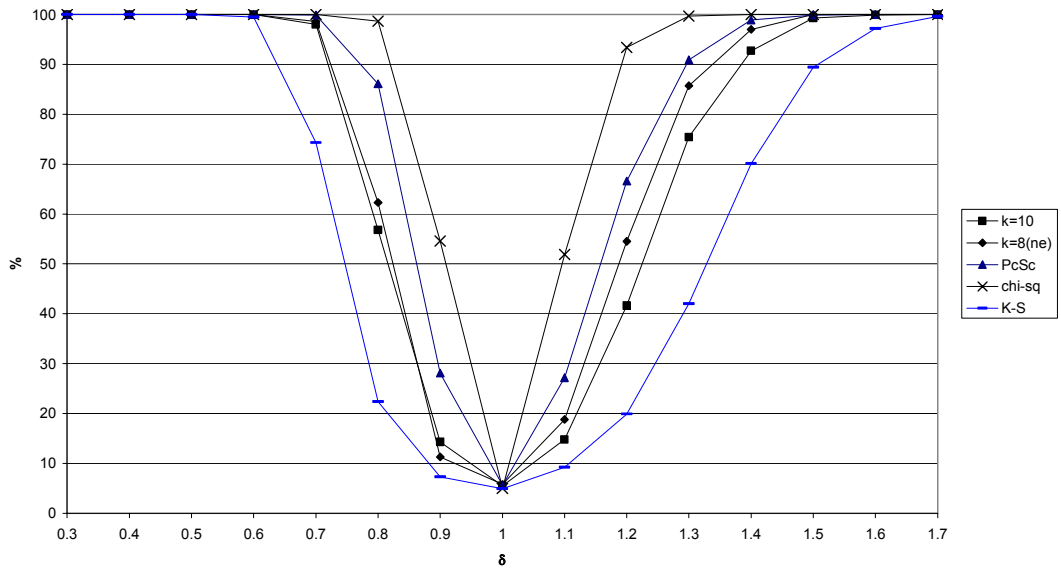

Figure 4: Power of the X^2 test against skewness (B_1: Ramberg distribution) N=150 (equiprobable)

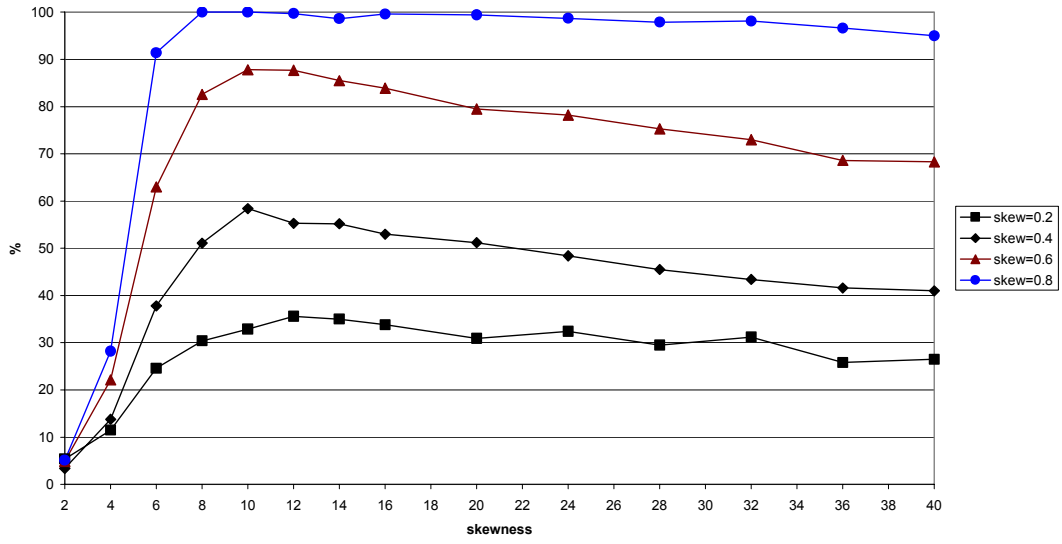

Figure 5: Power of the X^2 test against skewness (B_2: two-piece). Skew=0.5, (equiprobable)

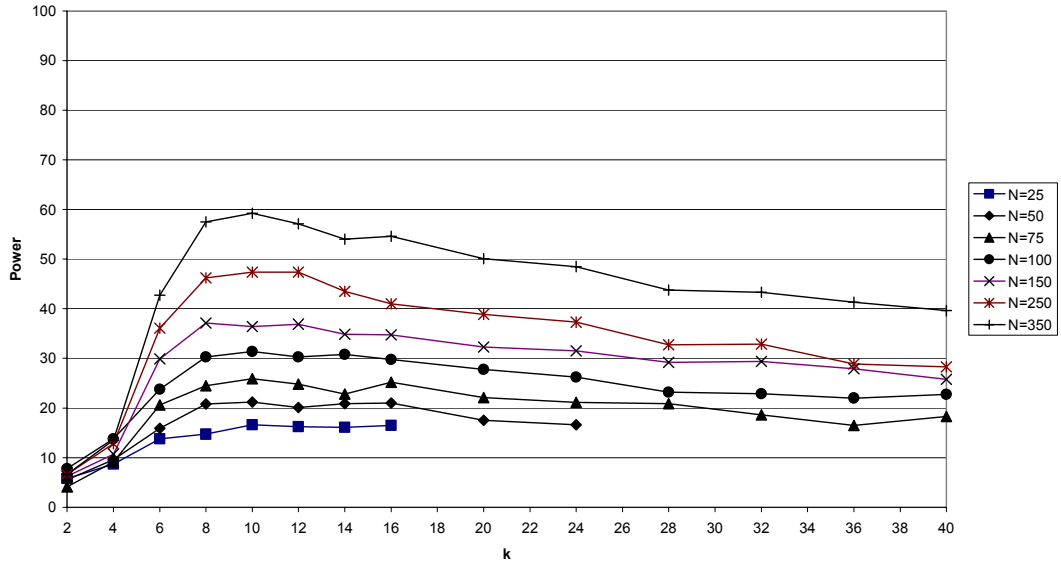

Figure 6: Power of the X^2 test against skewness (B_3: Anderson) N=150 (equiprobable)

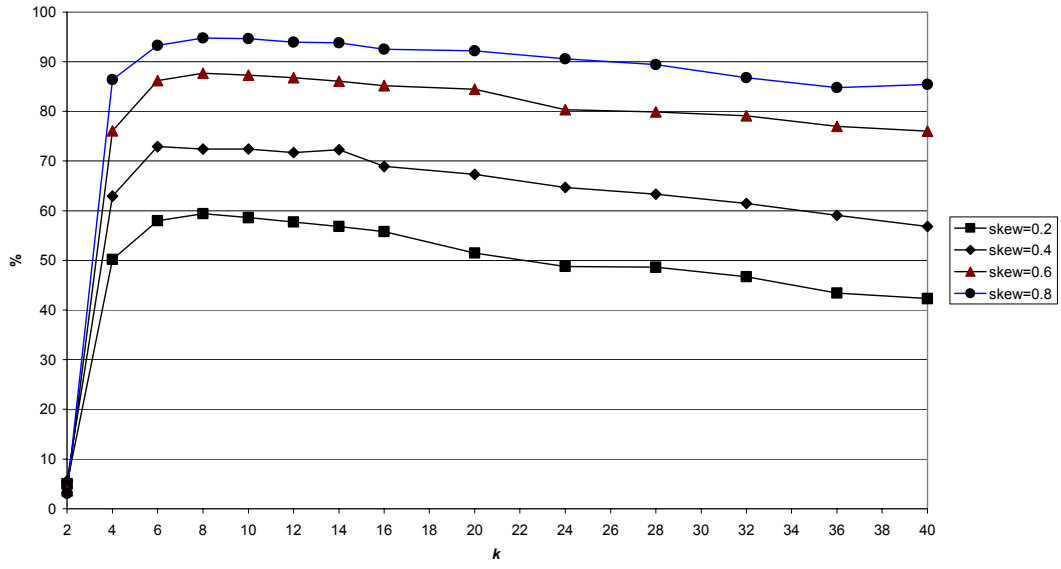

Figure 7a: Power of the X^2 test against skewness (B_1: Ramberg), N=150 (non-equiprobable)

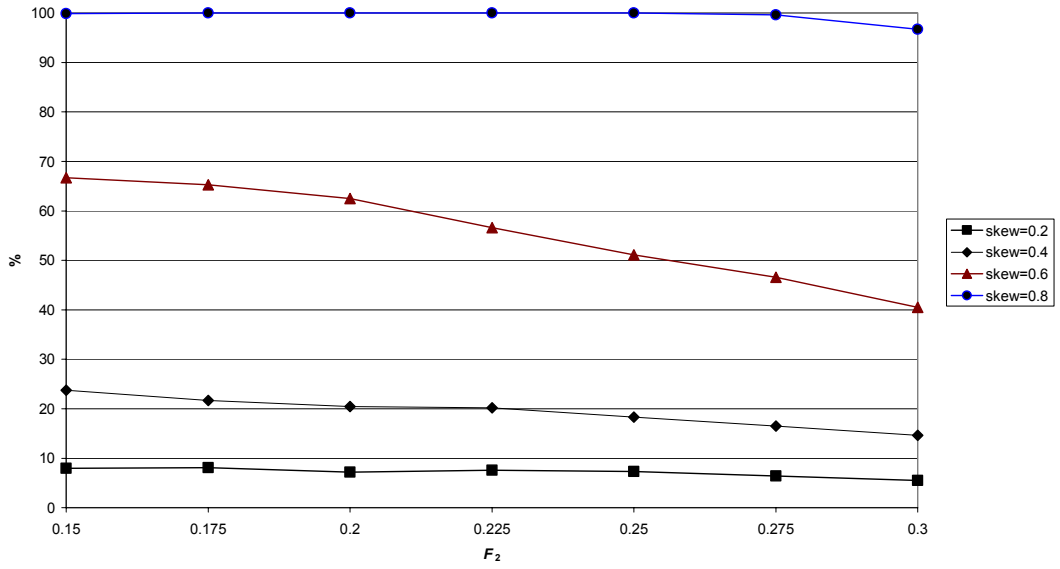

Figure 7b: Power of the PCSk test against skewness (B_1: Ramberg), N=150 (non-equiprobable)

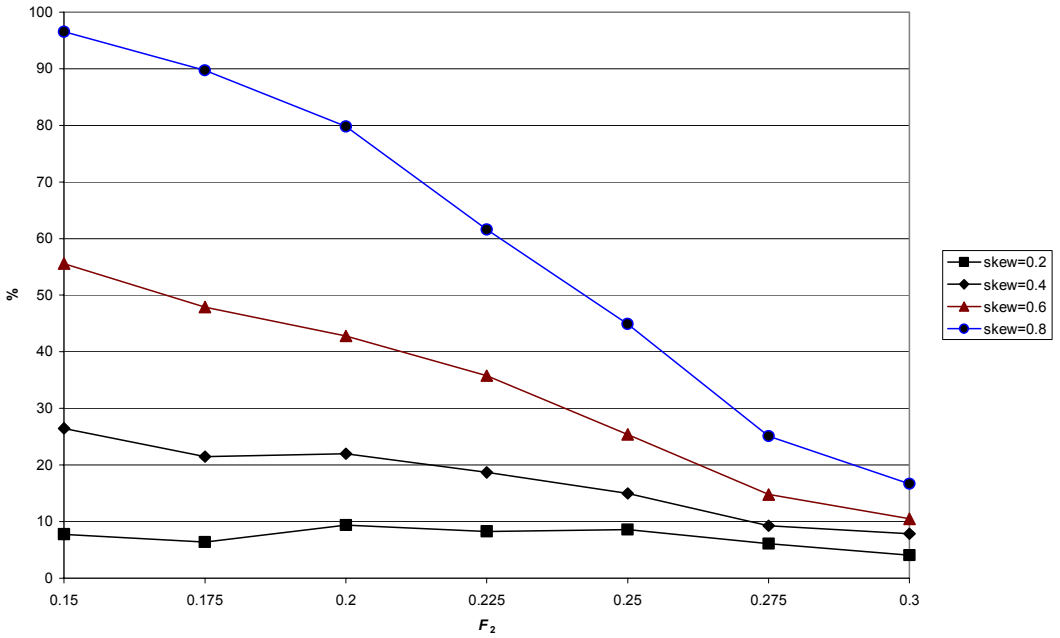

Figure 8a: Power against skewness (B_1: Ramberg), N=150

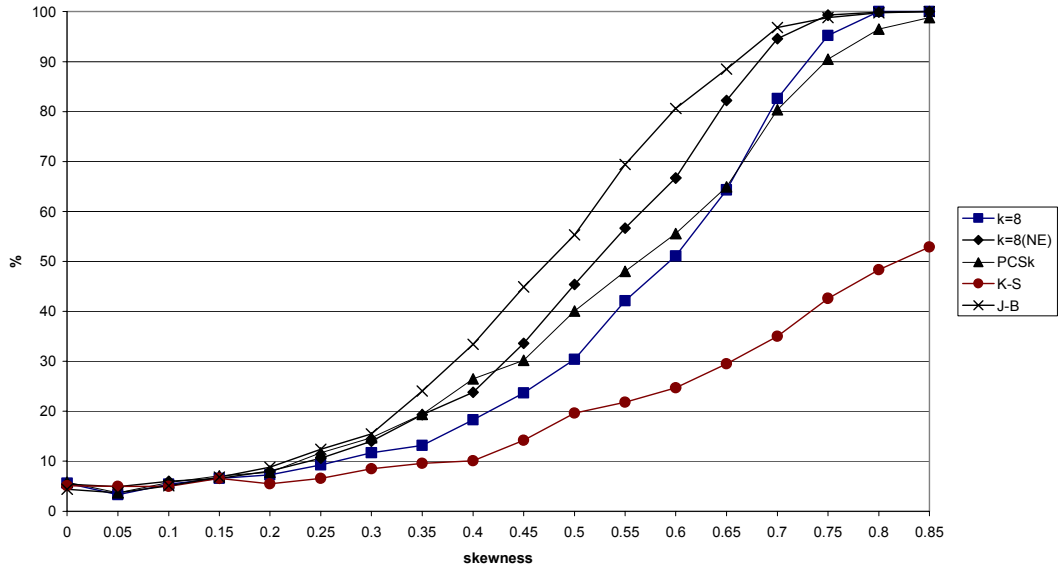

Figure 8b: Power against skewness (B_3: Anderson), N=150

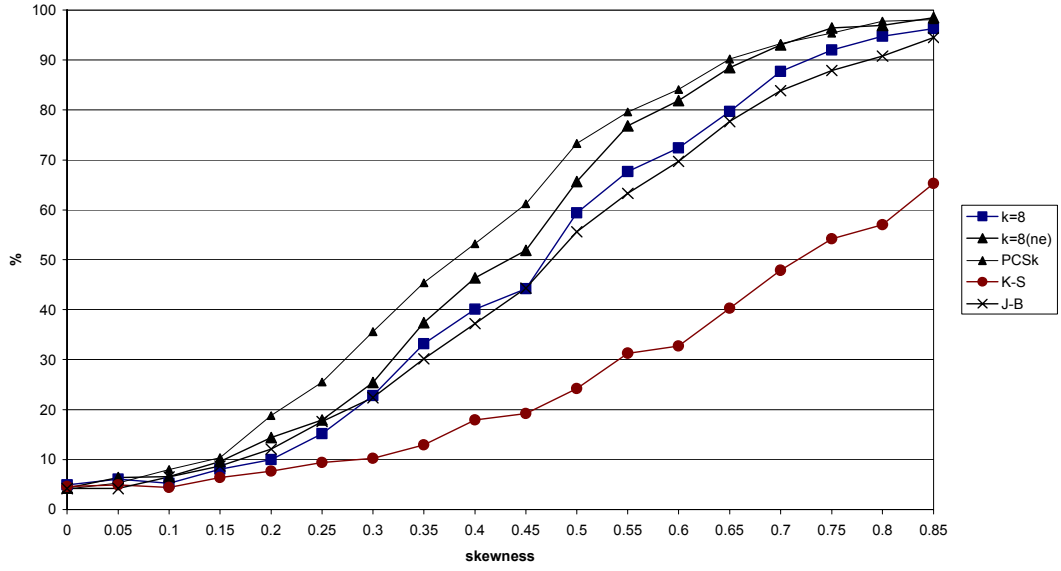

Figure 9: Power of the X^2 test agasinst kurtosis (C_1: Stable distribution), N=150

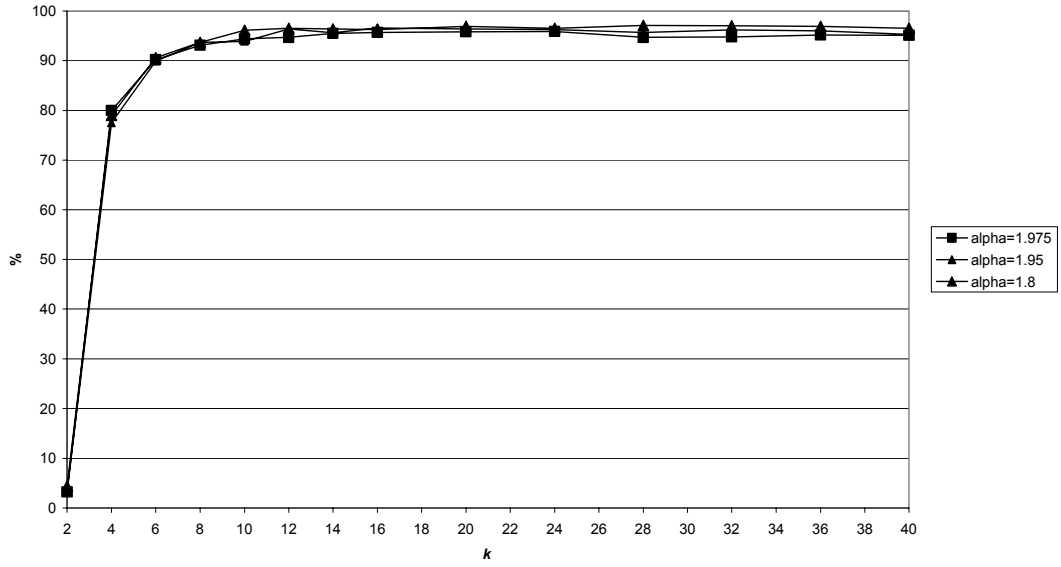

Figure 10: Power of the X^2 test against kurtosis (C_2: Anderson), N=150 (equiprobable)

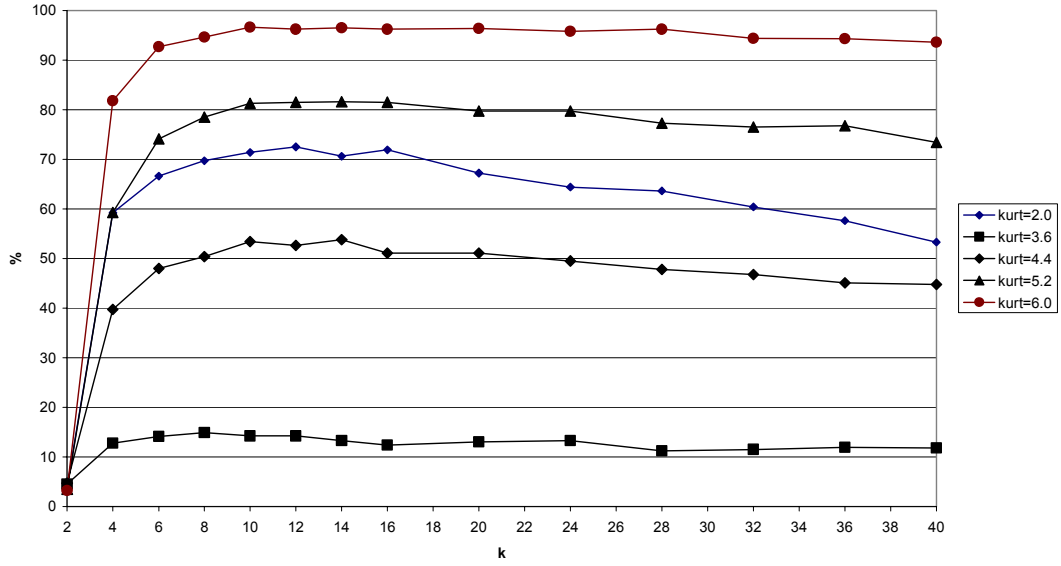

Figure 11a: Power of the X^2 test against kurtosis (C_1: Stable, α=1.975), N=100

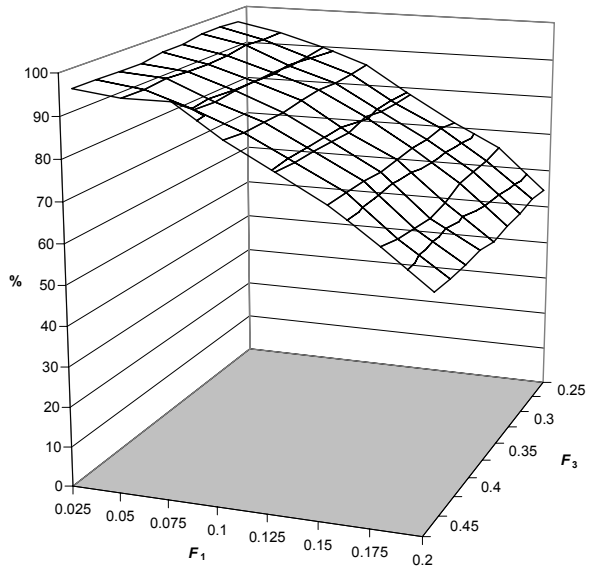

Figure 11b: Power of the PCK component test against kurtosis (C_1: Stable, α=1.975), N=100

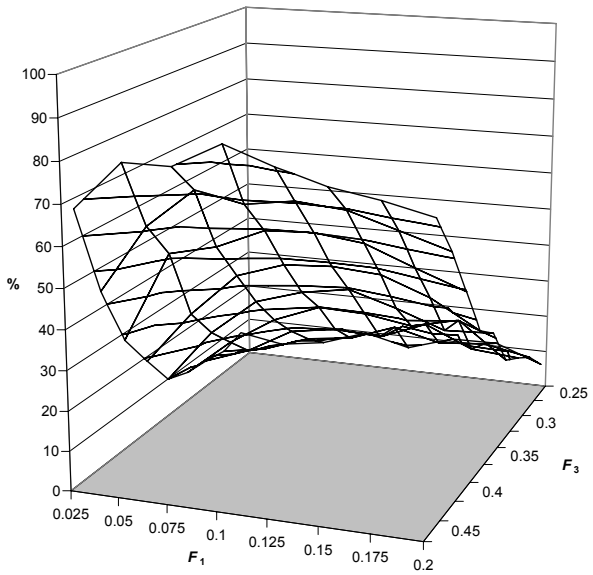

Figure 12: Power against kurtosis (C_1: Stable), N=100

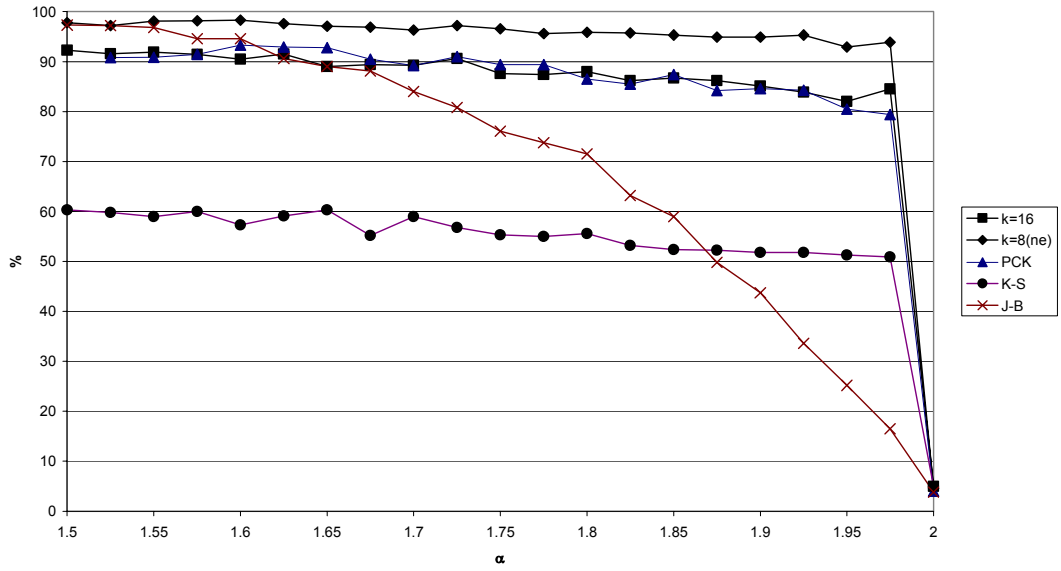

Figure 13: Power against kurtosis (C_2: Anderson) N=150