FINANCIAL OPTIONS RESEARCH CENTRE

University of Warwick

An Evaluation of Tests of Distributional Forecasts

Pablo Noceti
Jeremy Smith
and
Stewart Hodges

August 1999
(this version, January 2000)

Financial Options Research Centre
Warwick Business School
University of Warwick
Coventry
CV4 7AL
Phone: (01203) 524118

FORC Preprint: 2000/102

AN EVALUATION OF TESTS OF DISTRIBUTIONAL FORECASTS

Pablo Noceti*
F.O.R.C.
Warwick Business School
University of Warwick

Jeremy Smith
Department of Economics
University of Warwick

And

Stewart Hodges*
F.O.R.C.
Warwick Business School
University of Warwick

August 1999 (this version, January 2000)

Abstract

Traditionally, forecasters have concentrated on the point forecasts from their models. This has been increasingly seen as deficient, as individuals are not indifferent to the uncertainty associated with these forecasts. Consequently, more recently attention has been focused on the distribution associated with forecasts. This paper investigates the size and power of a number of (distribution free) tests for distributional forecasts.

*Funding from the ESRC ROPA award number R022250134 and FORC corporate members is gratefully acknowledged.

1. Introduction

Traditionally, economic forecasters have simply reported point forecasts of their models (see for example Wallis (1995) and Diebold and Lopez, (1996)). However, individuals are not indifferent to the varying degrees of uncertainty associated with these forecasts. Christofersen (1998) analysed how to evaluate interval forecasts, such that the interval is wider in more volatile periods and narrower in relatively stable periods. Subsequently, attention has focussed on the evaluation of the whole density function associated with a particular forecast, see Dawid (1984), and more recently, Diebold *et al* (1998a).

Interest in the formulation and evaluation of density forecasts is closely related to advances in the area of risk management and the density forecasts of financial asset returns, return volatilities and portfolio losses. In particular, the Value at Risk measure used for bank capital requires the estimation of a specific percentage point for a given time horizon. However, given that the models are capable of providing forecasts of the whole distribution of returns, this seems rather inefficient. It therefore becomes necessary to have a test to assess the accuracy of the density forecast model.

Diebold *et al* (1998a) show that using a loss function to evaluate density forecasts it is impossible to rank two incorrect forecasts so that all users agree with the ranking, but that, when the forecast density coincides with the true data generating process, it will be preferred by all users. Their approach is based on the fact that when the forecast and true generating processes are the same, their probability integral transform should be independently and identically distributed U(0,1). They prefer to use graphical tools, such as plots and correlograms to assess uniformity, based on the idea that statistical tests are non-constructive, in the sense that they do not give an insight into the reasons for rejection. This approach is then applied in Diebold, *et al* (1999) to evaluate inflation density forecasts and, extended to a multivariate framework, in Diebold *et al* (1998b) and Clements and Smith (1999).

In the next section, we briefly review the idea of the probability integral transform used to transform the distribution of forecasts into U(0,1) variates and the literature which has predominantly used the Kolmogorov-Smirnov (KS) test to assess the validity of their forecast densities. In section 3 we present Monte Carlo experiments to demonstrate the low power of the KS test to detect a bias in the mean and variance of the forecast distribution. Section 4 introduces a series of alternative (distribution free) tests based on the empirical density function and compares their power with that of the KS statistic for biases in the mean and variance of normal data, while section 5 focuses on non-normal data generating processes (DGP's) and specifically on their skewness and kurtosis. Section 6 offers some concluding remarks.

2. The Probability Integral Transform

Given a conditional density $f_t(y_t / \psi_{t-1})$ and a density of one step ahead forecasts $p_{t-1}(y_t)$, made at time *t-1* for the variable of interest (y_t) using the set of all available information at time *t-1* (ψ_{t-1}), they can be related through the Probability Integral Transform (z_t) defined as:

$$z_t = \int_{-\infty}^{y_t} p_{t-1}(u)du = P_{t-1}(y_t)$$

If the series of one-step ahead forecasts $p_{t-1}(y_t)$ coincides with the series of conditional densities $f_t(y_t / \psi_{t-1})$ (and assuming $p_{t-1}^{-1}(y_t)$ is strictly positive and bounded over the support y_t), then z_t should be independent and identically distributed (i.i.d) Uniform (0,1). Therefore, a way to assess whether the forecasts are a good representation of the true data generating process is to test the independence and uniformity of the series of probability integral transforms $\{z_t\}$. Unfortunately, this is a joint test and failure to accept the independence-uniformity hypothesis can be due to the z series being either non-uniform and/or non-i.i.d.

In Diebold, *et al* (1999) and Clements and Smith (1999) independence is investigated by testing for non-zero autocorrelations in the first three moments of the z_t series and then, conditional on independence, uniformity is tested using the Kolmogorov-Smirnov (KS)

statistic. Unfortunately, the power of the KS test statistic to detect non-uniformity is rather low (see, for example, Stephens 1974).

3. Size and Power of the Distributional Test Using Kolmogorov-Smirnov (KS) Statistic:

By far, the most commonly employed test for distributional forecasts is the well-known KS test for goodness of fit. In this section we describe the KS statistic (D) and examine its power for samples of various sizes when the null is N(0,1) but the data generating process (DGP) has a different mean or a different variance. Let z_j be the theoretical cumulative distribution function under the null, and A_j the empirical cumulative distribution, for $j = 1, \ldots, n$, where n is number of observations.
Then, the Kolmogorov difference is defined as $D = Max_j \{abs(z_j - A_j)\}$ (see Neave and Worthington (1988) for a complete description of the mechanics of the test). Seven different sample sizes: n = 50, 100, 225, 450, 900 and 3600 were considered. Simulations to assess the power of the tests were done with 1000 replications and 10000 for size calculations[1].

Power to Detect a Bias in the Mean:

We analysed the power of this test to detect a bias in the mean in the forecast distribution. We test a null hypothesis of N(0,1) data, when the DGP is distributed N(μ,1), with $\mu = \delta \times (1/\sqrt{n})$, so that the mean of the true process is equal to the mean of the hypothesised distribution increased by δ standard errors of the mean, with $\delta \in (2,8)$. The

[1] Given a set of parameters θ, that define a distribution, the probability of rejecting the null hypothesis when it is false for a particular set $\theta = \theta_1$ (and a given confidence level), is called the power of the test. On the other hand, the size of the test (or significance) is the maximum probability of rejection (given a critical value) when the null is true.

lower point ($\delta = 2$) gives the approximate upper limit of a 95% confidence interval for the mean[2].

Table 1 shows the results for sample sizes up to 3600 and gives the % of rejections out of 1000 replications

Table 1: N(0,1) v N(μ,1), μ = δ/√n

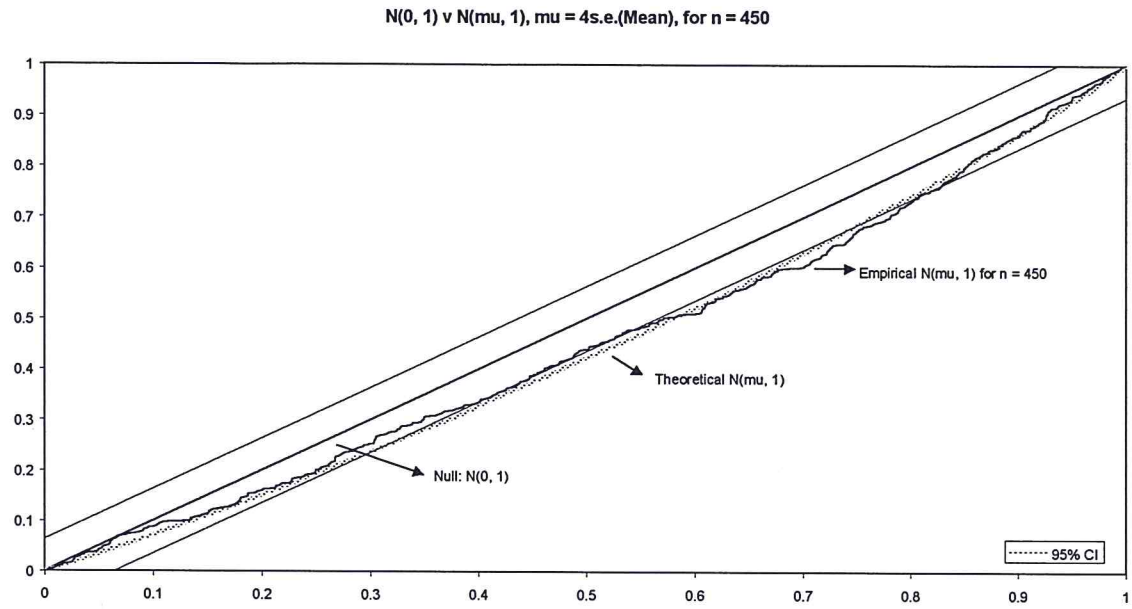
	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)
δ	n = 50	100	225	450	900	3600
2	38.6 (18.1)	37.2 (17.7)	39.9 (20.2)	40.3 (19.5)	39.5 (20.6)	36.7 (16.4)
4	92.0 (77.5)	92.3 (79.8)	92.9 (78.6)	94.6 (80.9)	93.9 (80.6)	93.1 (80.0)
6	100 (99.2)	100 (99.5)	100 (99.2)	100 (99.7)	99.7 (99.5)	99.9 (99.4)
8	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)

For $\delta = 2$, it shows that the KS statistic cannot detect deviations from the mean, when the null hypothesis is N(0,1), with an average of approximately 39% rejections for a 95% confidence level, and 19% for 99%. To get an acceptable level of rejection when applying the KS test to the cumulative distributions of normal data, the bias should be of at least 4 standard errors above the mean and this is independent of the sample size considered.

Figure 1 shows the theoretical and empirical cumulative probabilities when $\delta = 4$ for $n = 450$. Given that the cumulative probability of a variable distributed U(0,1) is its own value, the theoretical probabilities will plot like a 45 degree line between the origin and (1,1).

[2] The upper limit of a 95% confidence interval for the mean is given by: $average + t^{0.975}_{n-1} \times (stdev/\sqrt{n})$. For the samples considered, the t-value is approximately equal to 2.

Figure 1:



We can see from the picture that a N(μ,1) with μ corresponding to a factor of 4 s.e.(mean) is only just outside the 95% limits of the KS statistic.

Power to Detect a Bias in the Variance:

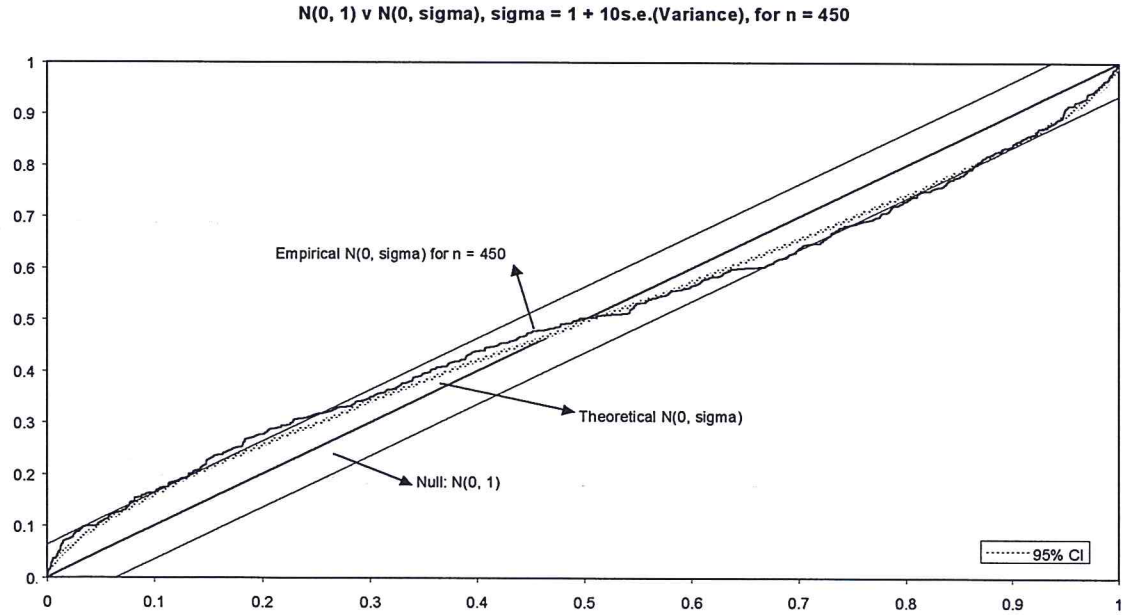
The same methodology is used to test the power of the KS statistic in detecting deviations from uniformity from a $N(0,\sigma^2)$, σ^2 being the variance of the DGP and equal to the hypothesised variance increased by two standard errors (of the variance): $\sigma^2 = 1 + [\delta \times (\sqrt{2}/\sqrt{n})]$, when the forecasted distribution was N(0,1).

Table 2: N(0,1) v $N(0,\sigma^2)$, $\sigma^2 = 1 + [\delta \times (\sqrt{2}/\sqrt{n})]$

	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)
δ	n = 50	100	225	450	900	3600
2	9.2 (2.3)	8.5 (1.9)	8.3 (2.4)	10.2 (2.0)	9.1 (1.9)	8.7 (2.0)
4	18.5 (6.3)	18.7 (6.2)	22.8 (6.2)	25.6 (5.6)	24.9 (7.1)	25.9 (7.0)
8	43.8 (20.4)	59.9 (24.8)	72.3 (35.5)	79.3 (46.1)	86.9 (51.1)	92.8 (64.3)
10	60.7 (29.1)	75.7 (40.9)	87.4 (55.3)	95.1 (70.1)	97.1 (80.8)	99.6 (92.1)
14	78.2 (44.7)	94.9 (71.4)	99.4 (91.3)	99.9 (97.9)	100 (99.6)	100 (100)
20	93.1 (71.1)	99.8 (93.8)	100 (99.4)	100 (100)	100 (100)	100 (100)

As shown in Table 2, the power of the KS test for uniformity to detect biases in the variance is even lower than for biases in the mean, with an average number of rejections out of 1000 replications of approximately 9% for a 95% confidence and approximately 2% for a 99% confidence. As before, a factor of 2 approximately gives to the upper limit of a 95% confidence interval for the variance (this approximation is more accurate for larger n[3]). As shown in Table 2, for the KS test to give a reasonably powerful test for distributional forecasts with a biased variance, the bias should be of at least 10 standard errors (of the variance) for a 95% confidence level, and of 14 standard errors for a 99% confidence level. There is a small increase in the power of the test as the sample size increases, with the % of rejections going from 87.4% for a factor $\delta = 10$ and a sample size of 225, to 95.1% for 450, 97.1% for 900 and 99.6% for 3600 observations at the 95% confidence level. Figure 2 shows the empirical and theoretical cumulative probabilities for a normal distribution with the variance increased by 10 standard errors (of the variance).

Figure 2:



Again, as can be seen in the plot, a variance increased by a factor 10 barely exceeds the 95% confidence interval for the KS difference.

---

[3] The upper limit of a 95% confidence interval for the variance is given by $(n-1)s^2/\chi^2_{n-1,\,0.025}$. We assumed $s^2 = 1$. For n = 50, this gives a factor = 1.55 while 2 +s.e.(variace) gives a factor 1.4, while for n = 450 the factors are 1.15 and 1.33 respectively.

4. Alternative (Distribution Free) Tests for Distributional Forecasts

In this section we introduce alternative (distribution free) ways to test distributional forecasts and analyse their power to detect biases in the mean and variance with respect to the 'real' data generating process.

4.1 Testing Distributional Forecasts by Testing Uniformity

The Kuiper Statistic:

The first alternative to the KS considered is the Kuiper test (Kuiper, 1962), this uses the same measures of deviation between the hypothesised and empirical distribution functions. Given $D^+ = max[(i/n) - z_i]$ and $D^- = [z_i - (i-1)/n]$, the at-the-step and before-the-step Kolmogorov differences respectively, the Kuiper statistic (V) is defined as: $V = D^+ + D^-$.

The Cramér-von Mises Statistic:

The Cramér-von Mises (W^2) statistic (see Cramér, 1946) uses a different measure of the deviation between the hypothesised and the empirical distribution function, and is defined as:

$$W^2 = \sum_{i=1}^{n} [z_i - (2i-1)/2n]^2 + (1/12n)$$

The Watson Statistic:

The Watson (U^2) statistic (Watson, 1961) is based on a transformation of the Cramér-von Mises statistic and is defines as $U^2 = W^2 - n(z_{mean} - ½)^2$ where $z_{mean} = \sum_i z_i/n$ (the average cumulative probability under the null).

The Anderson-Darling statistic

Anderson and Darling (1954) proposed a test of goodness of fit, focusing on differences in the tails of the cumulative distribution. The statistic (A^2) is defined as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{N}(2i-1)[\log(z_i) + \log(1-z_{-i})]$$

They compute the asymptotic percentage points for the distribution of this test statistic and find that these values are reached very rapidly for sample sizes as large as 40.

For all the other alternative tests we used the transformation and critical values proposed by Stephens (1974), who analyses their power for some simple deviations from uniformity.

4.2 Testing Distributional Forecasts by Testing the Inverse of the Cumulative Normal: Jarque-Bera and Doornik-Hansen Tests

The power of tests for normality is well documented (Doornik and Hansen, 1994). Therefore, it seems logical to transform the cumulative probabilities so that under the null hypothesis they should be distributed N(0,1) and apply a test for normality to these transformations. We define $x_i = \Phi^{-1}(z_i)$ where Φ^{-1} is the inverse of the cumulative standard normal distribution and z_i is, as before, the cumulative probabilities of the realisations under the forecasted distribution. We know that, given the probability integral transform, if the forecasted distribution and the DGP coincide, z_i is distributed U(0,1), so that, under the null, x_i should be distributed N(0,1).

The most commonly used test for normality is the Jarque-Bera test, based on the third and fourth moments of the distribution.

The Jarque–Bera statistic (*JB*) is defined as $JB = n\times[\beta_1/6 + (\beta_2 - 3)^2/24] \sim_a \chi^2(2)$ where $\sqrt{\beta_1}$ is the sample skewness, β_2 the sample kurtosis, n the sample size, and where $\sim_a$ means asymptotically distributed as.

Doornik and Hansen (1994) argue, following Bowman and Shenton (1975), that the Jarque-Bera test for normality is unsuitable except for very large sample sizes, given that the statistics $\sqrt{\beta_1}$ and β_2 are not independent, and the sample kurtosis approaches normality very slowly. Therefore, they propose a test based on transformed skewness and kurtosis, with the transformation bringing the statistic much closer to normality.

The proposed statistic (*DH*) has the form:

$DH = z_1^2 + z_2^2 \sim_{app} \chi^2(2)$

Where $\sim_{app}$ denotes approximately distributed as, and z_1 and z_2 the transformed skewness and kurtosis respectively.

Using a N(0,1), our results show that the empirical size values are within the 95% confidence interval of the theoretical values, for all tests and across all n, with exception of the Kolmogorov-Smirnov statistic, for $n < 100$ (a result also mentioned by Neave & Worthington (1988), and due to the fact that the critical values employed in the test are asymptotic) and for the Jarque-Bera test for $n < 225$.

4.3 Power of the Alternative Tests to Detect a Bias in the Mean and Variance

We performed the five alternative tests on the same data as the KS test when the true DGP is $N(\mu,1)$ and $N(0,\sigma^2)$, and the mean (variance) has been increased by a factor δ-times the standard error of the mean (variance), with δ defined as before.

These results show that both, the Anderson-Darling and Cramér-von Mises tests are considerably more powerful than the Kolmogorov-Smirnov test to detect a bias in the mean, especially when the bias is relatively small ($\delta = 2$). Of these two tests, the Anderson-Darling seems to be the most powerful, with 49.3% and 27.9% rejections for $\delta = 2$ and $n = 225$ at 95% and 99% respectively (against 47.6% and 26.2% for the Cramér-von Mises W^2 statistic and 39.9% and 20.2% for the KS statistic). These results are consistent for all sample sizes up to 3600.

Table 3: Alternative Tests for a Bias in the Mean

δ	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)
A^2	n = 50	100	225	450	900	3600
2	48.1 (26.5)	47.2 (26.5)	49.3 (27.9)	49.1 (26.7)	49.8 (27.8)	45.2 (23.4)
4	96.8 (89.6)	97.6 (90.1)	97.1 (89.8)	97.8 (91.8)	97.5 (92.0)	97.4 (91.0)
6	100 (100)	100 (100)	100 (99.9)	100 (100)	100 (99.9)	100 (100)
8	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)
W^2						
2	47.1 (23.5)	45.4 (25.1)	47.6 (26.2)	47.1 (25.3)	47.9 (26.1)	42.9 (21.3)
4	96.2 (87.7)	96.6 (87.8)	96.6 (88.2)	97.1 (90.0)	96.8 (90.4)	96.7 (88.7)
6	100 (100)	100 (99.9)	100 (99.9)	100 (100)	100 (99.8)	100 (99.9)
8	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)
U^2						
2	22.9 (7.5)	21.3 (8.8)	23.0 (9.2)	21.1 (8.2)	23.2 (7.2)	18.9 (6.5)
4	72.1 (49.9)	72.5 (48.0)	70.3 (47.8)	70.2 (46.1)	70.1 (47.0)	69.5 (48.9)
6	98.1 (92.5)	97.6 (92.3)	97.2 (90.8)	98.3 (92.0)	98.4 (90.9)	98.3 (91.5)
8	100 (99.8)	100 (99.7)	99.9 (99.9)	99.9 (99.8)	100 (99.6)	100 (99.8)
V						
2	22.7 (7.4)	22.4 (8.6)	24.2 (10.7)	21.9 (8.3)	24.5 (7.6)	20.7 (7.0)
4	75.6 (55.3)	77.5 (57.3)	76.7 (55.6)	79.2 (54.8)	79.5 (55.6)	77.5 (57.3)
6	99.1 (96.8)	99.2 (96.2)	98.8 (95.9)	99.3 (96.8)	99.4 (971)	99.2 (97.1)
8	100 (100)	100 (100)	100 (99.9)	100 (100)	100 (99.9)	100 (100)

Table 4: Alternative tests for a Bias in the Variance

δ	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)
A^2	n = 50	100	225	450	900	3600
2	20.4 (5.9)	16.8 (4.7)	15.5 (4.1)	16.7 (3.9)	13.6 (3.2)	13.9 (2.8)
4	50.3 (23.5)	50.8 (24.4)	57.2 (22.7)	59.9 (25.9)	58.5 (24.4)	58.3 (24.2)
6	78.9 (54.3)	85.8 (59.9)	91.1 (65.9)	93.2 (71.3)	94.6 (74.8)	96.9 (79.9)
8	92.8 (73.9)	97.0 (84.6)	99.3 (92.1)	99.5 (94.8)	99.8 (97.9)	100 (98.9)
10	97.6 (89.5)	99.5 (96.1)	100 (98.8)	100 (99.7)	100 (100)	99.9 (99.9)
14	100 (99.1)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)
W^2						
2	10 (3.0)	9.2 (1.8)	8.1 (2.3)	10.1 (2.2)	7.6 (1.9)	9.6 (1.6)
4	19.3 (6.3)	23.2 (5.1)	24.7 (5.1)	28.9 (5.4)	28.5 (6.8)	29.2 (6.3)
8	50 (19.6)	67.5 (28.1)	83.4 (42.9)	88.5 (53.6)	93.7 (61.2)	98.3 (77.0)
10	68 (32.4)	83.5 (46.2)	93.5 (66.5)	98.2 (80.2)	99.6 (91.4)	99.9 (97.5)
14	87.2 (54.9)	98.7 (82.1)	99.9 (95.7)	100 (99.5)	100 (100)	100 (100)
U^2						
2	20.0 (9.0)	19.8 (8.4)	22.7 (8.4)	24.8 (9.2)	23.4 (7.4)	24.7 (9.0)
4	49.7 (28.2)	55.4 (32.6)	64.7 (39.3)	67.9 (46.9)	72.8 (47.6)	75.4 (50.3)
8	88.0 (71.3)	96.1 (86.5)	98.7 (95.3)	99.5 (97.5)	99.9 (99.3)	100 (99.9)
10	94.9 (86.7)	98.8 (95.5)	100 (99.6)	100 (99.9)	100 (100)	100 (99.9)
14	99.4 (97.1)	100 (100)	100 (100)	100 (100)	100 (100)	100 (100)
V						
2	18.2 (7.5)	17.0 (7.0)	19.8 (6.9)	20.6 (7.8)	20.6 (6.1)	22.2 (8.1)
4	46.4 (25.7)	50.8 (27.8)	58.3 (34.6)	62.8 (39.0)	66.4 (41.1)	69.0 (43.4)
8	86.1 (67.3)	94.9 (81.5)	98.4 (93.4)	98.9 (95.9)	99.8 (98.3)	100 (99.8)
10	94.2 (83.4)	98.7 (94.4)	99.9 (99.0)	100 (100)	100 (100)	100 (99.9)
14	99.2 (96.0)	100 (99.8)	100 (100)	100 (100)	100 (100)	100 (100)

Note: JB and DH tests not reported as these only test 3rd and 4th moments.

The results for a bias in the variance are not as clear as in the case of the mean. Both the Anderson-Darling A^2 statistic and Watson's U^2 statistic present very similar results, with the A^2 statistic being slightly more powerful for smaller samples (50, 100, 225) and the Watson statistic for larger samples (450, 900, 3600). Nevertheless, both tests are more powerful than the KS statistic for all sample sizes and all δ.

5 Non-Normal Data Generating Processes

We are concerned with cases when the underlying distribution is misspecified, although the first and second moments are correct. That financial asset returns exhibit characteristics that are different from those expected of normally distributed variates is now generally accepted. In particular, financial data are fat-tailed (excess kurtosis) and may exhibit skewness (asymmetry in the distribution). This is investigated using a standardised t-distribution for fat-tailedness, and a Ramberg distribution in the case of an asymmetric DGP.

5.1 Power to Detect a Bias in Excess Kurtosis when the Null is Fat-Tailed

We generated random data from a Student's t distribution with degrees of freedom = k and assumed a null Student's t distribution with $k = 10$. Some empirical studies using the t-distribution have suggested k should be around 10 for financial data (Bera and Higgins, 1993, Bookstaber and McDonald, 1987). However, more recent studies have pointed to the fact that returns may have higher kurtosis than determined by this value, or even infinite kurtosis (see for example, Hansen 1994, Noceti and Hodges, 1998). This, together with the fact that the degrees of freedom parameter is very difficult to estimate accurately (Blattberg and Gonedes, 1972) suggests we should investigate a possible downwards bias misspecification in k. We considered DGP's with $k = 3$, 5 and 10 when the null is $k = 10$.

Table 5: Power of Distributional Forecast Test, t_{10} v t_k

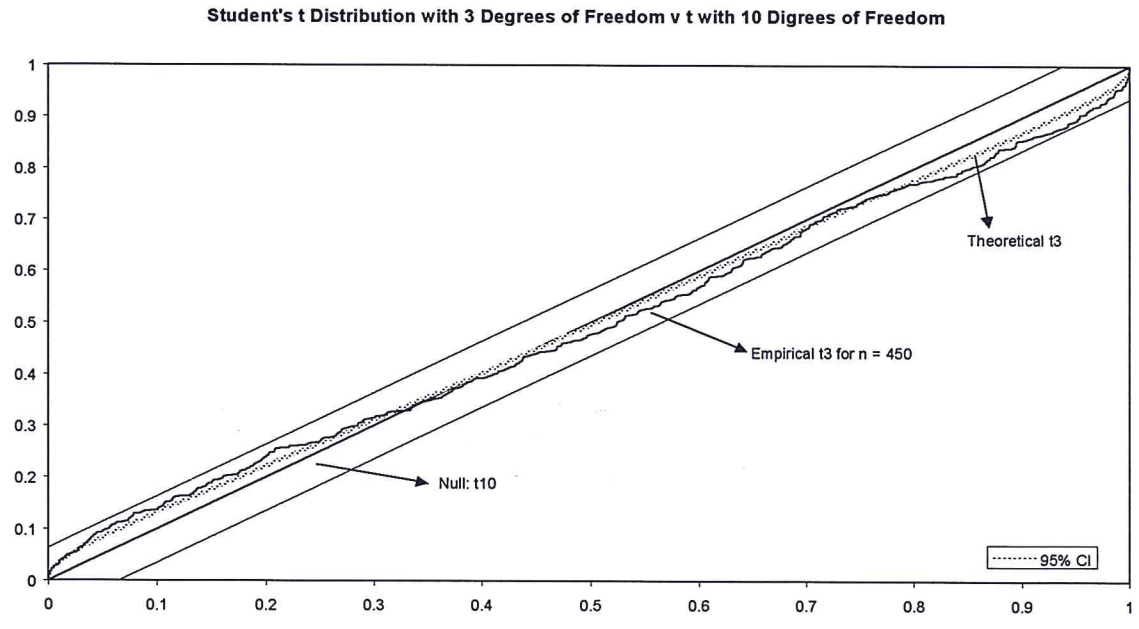
k	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)
D	n = 50	100	225	450	900	3600
3	5.1 (1.4)	8.0 (1.8)	12.2 (2.5)	21.1 (4.9)	54.8 (18.8)	100 (99.9)
5	3.0 (0.7)	5.7 (0.9)	4.8 (1.2)	6.7 (1.2)	8.6 (1.6)	39.0 (9.7)
10	4.3 (0.8)	4.7 (1.0)	5.1 (0.8)	4.9 (0.5)	5.3 (1.0)	4.9 (1.5)
100	5.1 (1.1)	5.2 (0.8)	5.9 (1.4)	5.4 (0.7)	6.8 (1.0)	25.7 (5.6)
A^2						
3	17.3 (5.7)	32.8 (11.7)	62.2 (28.7)	92.0 (68.2)	100 (98.6)	100 (100)
5	5.1 (0.9)	8.8 (1.6)	10.1 (2.3)	20.9 (4.9)	38.8 (11.5)	99.3 (91)
10	4.8 (1.6)	5.4 (1.1)	5.2 (0.9)	4.2 (1.1)	4.8 (1.2)	4.9 (1.4)
100	5.4 (0.7)	5.0 (0.9)	6.6 (1.1)	9.2 (2.0)	19.3 (3.3)	95.6 (64.5)
JB						
3	23.8 (16.2)	41.9 (31.4)	72.2 (60.2)	93.0 (87.4)	99.8 (99.0)	100 (100)
5	9.2 (5.9)	16.8 (9.0)	29.3 (20.2)	41.5 (28.3)	70.1 (53.4)	99.9 (99.1)
10	3.2 (1.6)	4.2 (1.8)	4.8 (1.6)	4.2 (2.0)	4.8 (1.2)	4.4 (1.2)
100	0.4 (0.0)	0.3 (0.0)	2.8 (0.0)	11.0 (0.5)	43.8 (11.9)	99.8 (97.4)
DH						
3	27.4 (14.5)	43.9 (26.4)	73.1 (57.8)	93.2 (86.3)	99.9 (99.0)	100 (100)
5	9.8 (3.7)	16.7 (7.0)	29.5 (16.6)	40.6 (25.4)	68.6 (50.6)	99.9 (99.0)
10	3.4 (1.2)	5.5 (1.6)	4.7 (0.9)	4.5 (1.7)	4.3 (1.0)	4.3 (1.0)
100	2.3 (0.3)	3.7 (0.2)	8.8 (1.8)	16.8 (2.7)	48.9 (19.9)	99.8 (98)
W2						
3	5.9 (1.6)	8.7 (2.3)	14.8 (3.2)	25.8 (5.5)	60.9 (19.4)	100 (99.9)
5	4.1 (0.7)	6.1 (0.9)	6.1 (0.9)	7.6 (1.6)	10.7 (2.2)	45.9 (9.1)
10	4.4 (1.4)	5.0 (0.8)	5.3 (0.8)	4.8 (0.8)	5.0 (1.3)	5.1 (1.4)
100	6.0 (1.1)	5.3 (0.9)	6.8 (1.1)	6.3 (1.2)	7.7 (1.3)	31.0 (5.3)
U2						
3	10.3 (3.9)	17.7 (5.7)	38.0 (17.7)	69.6 (43.8)	97.4 (86.0)	100 (100)
5	5.2 (0.8)	6.0 (0.9)	8.5 (1.9)	15.7 (5.0)	30.0 (12.4)	91.3 (75.6)
10	3.8 (0.9)	5.2 (0.8)	5.2 (1.0)	5.4 (1.0)	4.5 (1.5)	5.7 (0.9)
100	6.7 (1.6)	6.3 (1.6)	9.7 (2.6)	14.6 (4.0)	25.0 (9.3)	88.0 (61.9)
V						
3	10.4 (3.9)	19.2 (5.9)	40.4 (17.6)	73.2 (46.1)	98.9 (90.6)	100 (100)
5	4.4 (1.2)	5.4 (1.5)	8.8 (2.2)	16.5 (5.0)	31.2 (12.0)	94.9 (79.4)
10	4.3 (1.0)	5.4 (1.1)	4.1 (1.2)	4.3 (1.1)	4.7 (0.9)	6.7 (0.8)
100	6.4 (1.1)	6.5 (2.1)	10.0 (2.6)	14.3 (4.1)	27.1 (9.6)	91.6 (69.3)

Table 5 shows that for $k = 10$ all tests are correctly sized, though there is a suggestion of over-rejection in the JB test at the 99% level for sample sizes up to 450, and in the DH (again at 99% level) for $n = 450$ (and under-rejection at 95% level for $n = 50$). Across all other k and for all n, the DH test on the inverse of the cumulative normal tends to be more powerful at detecting differences in the kurtosis between the real and forecast

distributions than the alternative tests considered. The Kolmogorov-Smirnov, Cramér-von Mises, Kuiper and Watson statistics have very low power to detect the bias in the distribution's tails even for k as low as 3 and sample sizes up to 450 observations, with the Kuiper and Watson tests becoming reasonably powerful for larger samples. For $k = 3$, all tests seem to give 100% rejections asymptotically ($n = 3600$). When $k = 5$, however, the power of most tests falls dramatically. Both the DH and JB tests yield similar results, and seem to be the most powerful for all sample sizes. The Kolmogorov-Smirnov and Cramér-von Mises tests have very low power for all samples, while the remaining tests only seem to work for large n. As an experiment, we performed all tests on a process with an upward bias in the degrees of freedom (thinner tails than the null). In this case no test has any power for $n < 450$.

Figure 3 shows the theoretical and empirical cumulative probabilities when the data generating process is distributed Student's t with $k = 3$ but we forecasted with a Student's t distribution with $k = 10$.

Figure 3:



6.2 Asymmetry

Another common feature often found in distributions of financial asset returns is asymmetry (or significant skewness), meaning that the probabilities of getting a positive and a negative observation of the same value are not equal. To study the power of the KS and AD tests when the underlying process is skewed but we forecast with a symmetric distribution, we generated random samples from the Ramberg distribution (see Ramberg *et al*, 1979) with varying sample sizes and degrees of skewness (q).

This distribution, unlike most known distributions, is not expressed as a function of the underlying variable's values, but in terms of their cumulative probabilities. It has four parameters that determine the mean, variance, skewness and kurtosis of the distribution, making it flexible in terms of attainable shapes. 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)*. λ_1, is the location parameter, λ_2 the scale parameter, and λ_3 and λ_4 shape parameters. Ramberg *et al* (1979) give tables of λ_1, λ_2, λ_3 and λ_4 for distributions with mean 0 and variance 1, for varying degrees of skewness and kurtosis.

Figure 4 shows data generated from Ramberg distributions with different degrees of skewness and excess kurtosis.

Figure 4:

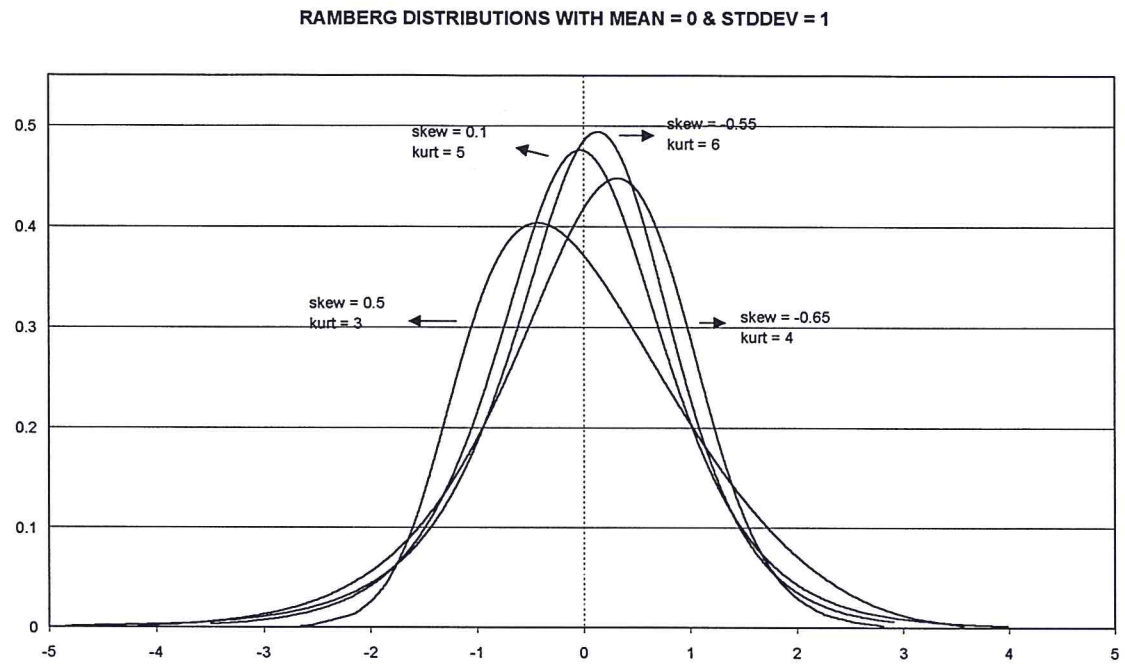
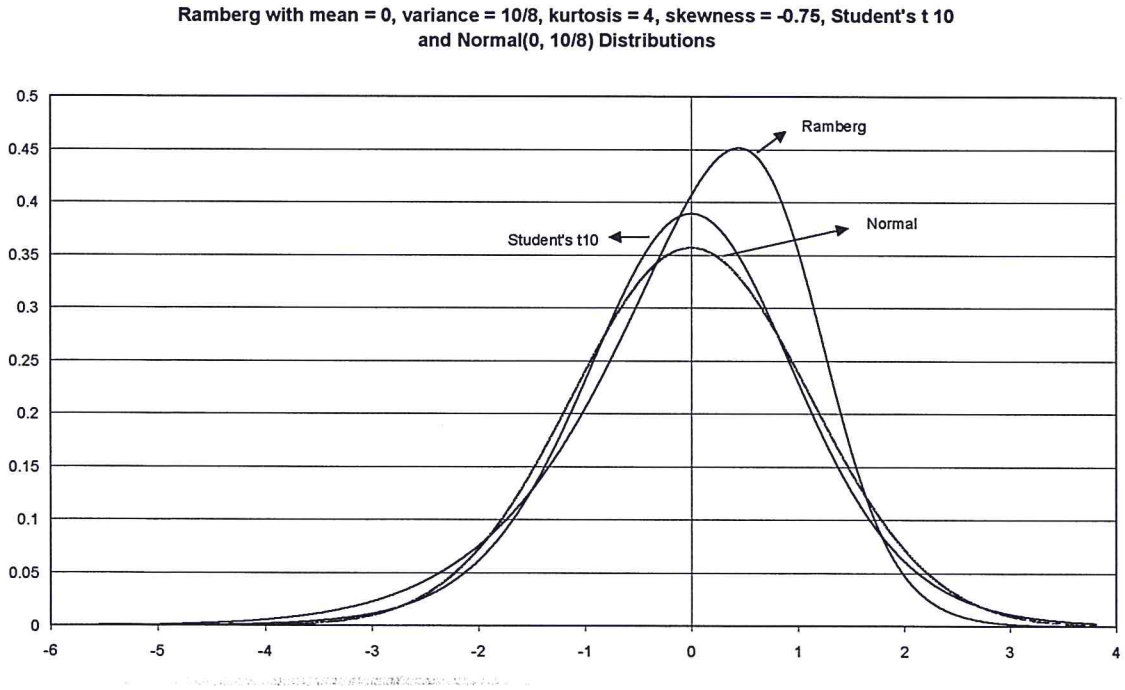


Figure 5:



Given that null hypotheses are bound to account for the significant kurtosis often found in financial return distributions, we analysed the case of a null with a kurtosis equal to 4

(Student's t with $k = 10$) when the real process has the same degree of kurtosis but varying degrees of negative skewness.

For $q = 0$ and samples up to 900 observations, almost all tests approximately return size (which is correct if we consider only the first four moments of the distribution). There seems to be over-rejection for all tests when the sample is really large (3600) except for the Cramér-von Mises test which is correctly sized for all sample sizes. This over-rejection for very large sample sizes could be due to the fact that the two distributions considered here are only equal in the first four moments.

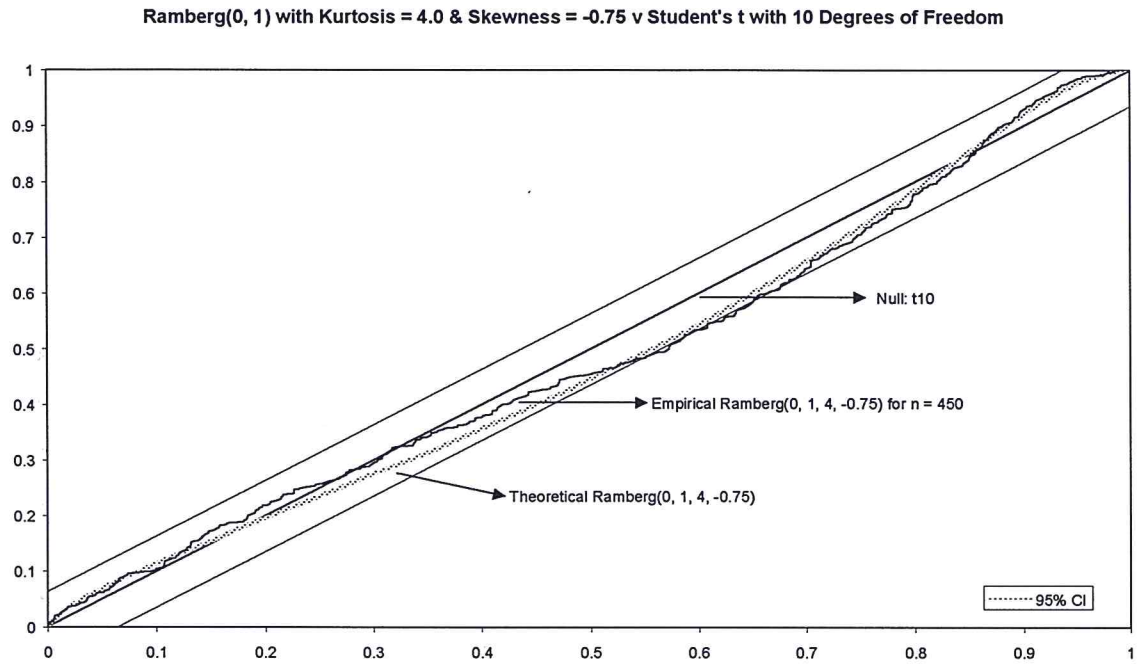
With the smaller samples considered (up to 450 observations), the DH test is the most powerful to detect skewness in the DGP when the assumed forecast distribution is symmetric, for all q. The power, however, decreases quite markedly as the skewness is reduced by a small fraction, going from 71.1% to 39.7% for $q = -1.0$ to $q = -0.85$ when $n = 50$, 59.1% to 41.5% for $q = -0.75$ to -0.65 and $n = 100$, and from 59.6% to 39.2% for $q = -0.55$ to -0.45 when $n = 225$. For $n = 450$, both the JB and the DH tests are powerful and present similar results for large q's, with the DH test still being more powerful for q below -0.75. As the sample size increases to 900, most other tests begin to detect the skewness in the underlying process, but the DH test remains more powerful for the smaller skewness parameters. For very large sample sizes, all tests seem to detect the skewness in the DGP.

Table 6: Power for Student's t with k = 10 when Underlying is Ramberg with skewness = q

q	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)	95% (99%)
D	n = 50	100	225	450	900	3600
0.00	3.7 (0.6)	3.9 (0.7)	5.1 (0.8)	4.3 (1.1)	6.1 (1.2)	7 (1.4)
-1.00	18.8 (7.5)	38.1 (18.7)	72.4 (49.4)	96 (84.9)	100 (100)	100 (100)
-0.85	11.9 (3.3)	26.8 (12.9)	51.2 (28.9)	78.7 (59.7)	98.1 (92.3)	100 (100)
-0.75	11.4 (2.4)	17.3 (5.9)	36.3 (18.1)	63.2 (39.4)	91.5 (78.6)	100 (100)
-0.65	9.3 (2.8)	14.1 (4.5)	27.4 (11.3)	50.9 (28.9)	77.9 (56.3)	100 (100)
-0.55	6.3 (1.8)	10.8 (2.5)	16.5 (6.1)	34.1 (16.6)	61.7 (38.1)	99.9 (98.2)
-0.45	5.4 (1.1)	7.8 (2)	13.7 (3.9)	22.9 (9.2)	37.7 (20.2)	95.7 (86.1)
A^2						
0.00	5.2 (0.9)	5 (0.9)	5.5 (1.1)	5.1 (1.1)	5.6 (1.3)	5.9 (0.9)
-1.00	18.1 (5.2)	42.5 (15.5)	91 (60.5)	100 (98.5)	100 (100)	100 (100)
-0.85	10 (3.2)	25.7 (9.3)	60.1 (27.3)	95.5 (71.2)	100 (100)	100 (100)
-0.75	11.6 (2.5)	17.1 (4.4)	39.4 (14)	77.7 (41.9)	99.9 (91.3)	100 (100)
-0.65	8.8 (2.2)	13.2 (3.5)	26.7 (8.2)	57.1 (24.9)	94.2 (67.7)	100 (100)
-0.55	7.2 (1.6)	10 (1.8)	16.9 (3.8)	36 (12.9)	72.3 (36.9)	100 (100)
-0.45	6 (1.5)	7 (1.4)	11.2 (2.5)	20.7 (6.8)	43.3 (16.2)	100 (98.3)
JB						
0.00	4.1 (1.4)	3.7 (1.3)	4.1 (1.1)	3.6 (1.4)	5.3 (2)	9.4 (2.4)
-1.00	29 (11.7)	87.6 (50.2)	100 (99.8)	100 (100)	100 (100)	100 (100)
-0.85	15.1 (6.4)	56.4 (25.7)	99 (88.3)	100 (100)	100 (100)	100 (100)
-0.75	12.4 (5.2)	37.2 (15.5)	90.5 (66)	99.9 (99.1)	100 (100)	100 (100)
-0.65	10.6 (4.7)	25.4 (10.5)	73.4 (44.1)	97.9 (90.6)	100 (100)	100 (100)
-0.55	8.1 (3.3)	19.4 (7.4)	49.9 (24)	87.9 (66.1)	99.6 (97.3)	100 (100)
-0.45	5.7 (2.5)	11.1 (4.8)	32.1 (13.2)	64.6 (39.8)	93.9 (79.4)	100 (100)
DH						
0.00	5.1 (1.5)	5.1 (1.1)	3.7 (1.1)	4.1 (0.8)	5.2 (1.6)	8.8 (2.1)
-1.00	71.1 (45)	98.7 (92.3)	100 (100)	100 (100)	100 (100)	100 (100)
-0.85	39.7 (17.3)	82.1 (58.8)	99.7 (98.3)	100 (100)	100 (100)	100 (100)
-0.75	28.9 (12.5)	59.1 (34.6)	95.5 (86.5)	99.9 (99.5)	100 (100)	100 (100)
-0.65	18.7 (6.1)	41.5 (20.1)	83.5 (63.1)	98.5 (94.7)	100 (100)	100 (100)
-0.55	13.9 (4.7)	28.1 (10.6)	59.6 (35.9)	91.1 (77)	99.9 (98.8)	100 (100)
-0.45	9.4 (2.1)	15.8 (5.1)	39.2 (17.5)	69 (46.7)	94.7 (83.9)	100 (100)
W^2						
0.00	5 (1)	4.3 (1)	5 (1.2)	4.3 (1)	5.5 (1.3)	5 (1.1)
-1.00	18.4 (6.5)	36.6 (16.9)	75.2 (46.6)	97.5 (86.6)	100 (99.8)	100 (100)
-0.85	11.2 (3.7)	25.4 (10.4)	51 (25.6)	82.2 (57.4)	99.4 (94.5)	100 (100)
-0.75	11 (3.1)	17.1 (4.7)	35.3 (15.9)	63.8 (35.6)	93.6 (77)	100 (100)
-0.65	9.4 (2.6)	14.2 (4.3)	25.6 (9)	48.3 (24.4)	79.4 (53.3)	100 (100)
-0.55	7.5 (1.8)	10.8 (2.5)	17.6 (5.4)	33.4 (14.7)	61.9 (33)	100 (99)
-0.45	5.9 (1.4)	7.7 (1.7)	12.9 (3.4)	22 (8.5)	36.8 (17.2)	97.6 (86.9)
U^2						
0.00	5 (1)	5 (1.6)	4.9 (1)	5.6 (1.3)	8.1 (2.4)	16.1 (4.9)
-1.00	39 (18.6)	70.2 (48.5)	97.4 (91.2)	100 (99.8)	100 (100)	100 (100)
-0.85	23.4 (8.4)	48.7 (26.4)	82.4 (63.5)	98.8 (95.6)	100 (100)	100 (100)
-0.75	18.2 (6.7)	33 (14.2)	66.2 (42.7)	92.1 (79.1)	100 (99.5)	100 (100)
-0.65	14.2 (4.9)	23.9 (9.8)	50.4 (27.7)	80.2 (61.7)	98.6 (94)	100 (100)
-0.55	10.2 (2.8)	17.2 (5.6)	32 (13.4)	61.6 (39.2)	88.5 (75.7)	100 (100)
-0.45	6.9 (1.6)	10.3 (2.9)	23.5 (8.9)	39.3 (19.4)	68.5 (45.6)	100 (99.8)
V						
0.00	4.7 (1.1)	4.8 (1.1)	4.4 (1)	5.9 (1.1)	7.8 (2.2)	15.6 (4.8)
-1.00	37.3 (17.9)	70.3 (47.3)	96.9 (91.5)	100 (100)	100 (100)	100 (100)
-0.85	22 (7.8)	46.2 (24.2)	79.9 (60.6)	98.3 (93.7)	100 (100)	100 (100)
-0.75	17.2 (5.4)	28.9 (12.6)	62.3 (39.9)	88.8 (75.1)	99.9 (99)	100 (100)
-0.65	12.7 (4)	22.4 (7.7)	47 (24.5)	77.1 (56.9)	97.8 (91)	100 (100)
-0.55	8.9 (2.7)	15.7 (5)	28.8 (12.1)	56.5 (32.9)	86.3 (71)	100 (100)
-0.45	6.1 (1.7)	9.3 (2.3)	21.6 (7.6)	36.3 (16.5)	63.8 (39.8)	100 (99.7)

Figure 6 shows the empirical and theoretical cumulative probabilities for a standard Ramberg distribution with kurtosis = 4 and skewness = -0.75, when the forecasting distribution is a Student's t with 10 degrees of freedom.

Figure 6:



Conclusions:

Increasingly, forecasters are required to supply more information than a single point forecast from their models. In the limit, they can supply information about the entire distribution from which the forecast was generated, i.e. a distributional forecast. To evaluate the validity of this distributional forecast, Dawid (1984) and later Diebold *et al* (1998) suggest using the probability integral transform, where, under the null of a correct model, the resulting cumulative probabilities are distributed i.i.d. U(0,1). Diebold *et al* (1998b) and Smith and Clemens (1999) use this probability integral transform and find that, using a Komogorov-Smirnov difference they are unable to reject the null hypothesis.

This paper shows that the Kolmogorov-Smirnov statistic is less powerful than alternative tests also based on the empirical distribution function to detect misspecifications in the forecast distributions. Of all the tests analysed, the Anderson-Darling statistic seems to be the most powerful to detect biases in the first two moments (mean and variance), with Watson's statistic being slightly more powerful in the variance case for large samples. For biases in the higher moments, however, the Doornik-Hansen test on the inverse of the cumulative normal seems to be the most powerful test. This case is particularly important in the area of finance, where it is recognised that returns are leptokurtic and, sometimes, negatively skewed. However, for small sample sizes, the test is less powerful except for cases of substantial negative skewness (q = -0.75).

References

Bera , A. K., and Higgins, M. L. (1993), "ARCH models: properties, estimation and testing", Journal of Economic Surveys, 7, 4, 305-362.

Berkowitz, J. (1999), "Evaluating the forecasts of risk models", mimeo, Federal Reserve Board, Washington, D.C.

Blattberg, R. C., and Gonedes, N. J. (1974), "A comparison of the stable and student distributions as statistical models for stock prices", Journal of Business, 47, 244-280.

Bookstaber, R. M., and McDonald, J. B. (1987), "A general distribution for describing security price returns", Journal of Business, 60, 3, 401-424.

Chatfield, C. (1993), "Calculating interval forecasts", Journal of Business and Economic Statistics, 11, 121-135.

Christoffersen, P. F. (1998) "Evaluating interval forecasts", International Economic Review, 39, 841-862.

Clements, M. P. and Smith, J. (1999), "Evaluating the forecast densities of linear and non-linear models: Applications to output growth and unemployment", mimeo University of Warwick.

Cramér, H. (1946), Mathematical Methods of Statistics, Princeton: Princeton University Press.

Dawid, A. P. (1984), "Statistical theory: The prequential approach", Journal of The Royal Statistical Society, ser. A, 147, 278-292.

Diebold, F. X. Gunther, T. A., Tay, A. S. (1998a), "Evaluating density forecasts: with applications to financial risk management", International Economic Review, 39, 863-883.

Diebold, F. X., Hahn, J., Tay, A. S. (1998b) "Real-time density forecast evaluation and calibration: Monitoring the risk of high frequency returns on foreign exchange", NBER working paper 6845.

Diebold, F. X., Tay, A. S., Wallis, K. F. (1999) "Evaluating density forecasts of inflation: The survey of professional forecasters", NBER working paper 6228. Forthcoming in R. Engle and H. White (eds.), Festschrift in Honor of C. W. J. Granger, Oxford: Oxford University Press.

Doornik, J. A., Hansen, H. (1994), "A practical test for univariate and multivariate normality", Discussion paper, Nuffield College.

Kendall, M. G., Stuart, A., Ord, J. K. (1987), Advanced Theory of Statistics, 5th Ed., vols 1 and 2. London: Charles Griffin and Co.

Kuiper, N. H.(1962), "Tests concerning random points on a circle", Proc. Koninkl. Nederl. Akad. van Wetenschappen, Series A, 63, 38-47.

Neave, H. R., Worthington, P. L. (1988), Distribution Free Tests, London: Unwin Hyman.

Ramberg, J., Dudewicz, E., Tadikamalla, P., Mykytka, A. (1979), "A probability distribution and its uses in fitting data", Technometrics, 21,201-209.

Stephens, M. A. (1974), "EDF statistics for goodness of fit and some comparisons", Journal of the American statistical Association, 69, 730-737.

Tay, A. S., Wallis, K. F. (1999), "Density forecasting: A survey", mimeo, Department of Economics, University of Warwick.

Wallis, K. F. (1995), "Large-scale macroeconometric modelling", in Pesaran, M. H. and Wickens, M. R. Handbook of Applied Econometrics: Macroeconomics: Basil Blackwell.

Watson, G. S. (1961), "Goodness-of-fit tests on a circle. I.", Biometrika, 48, 109-114.