Chi-squared Tests of Interval and Density Forecasts, and the Bank of England's Fan Charts

Kenneth F. Wallis

Department of Economics
University of Warwick
Coventry CV4 7AL, UK
[K.F.Wallis@warwick.ac.uk]

Revised November 2001

Acknowledgement The first version of this paper was written during a period of study leave granted by the University of Warwick and spent at the Economic Research Department, Reserve Bank of Australia and the Economics Program RSSS, Australian National University; the support of these institutions is gratefully acknowledged.

Abstract This paper reviews recently proposed likelihood ratio tests of goodness-of-fit and independence of interval forecasts. It recasts them in the framework of Pearson chi-squared statistics, and considers their extension to density forecasts. The use of the familiar framework of contingency tables increases the accessibility of these methods to users, and allows the incorporation of two recent developments, namely a more informative decomposition of the chi-squared goodness-of-fit statistic, and the calculation of exact small-sample distributions. The tests are applied to two series of density forecasts of inflation, namely the US Survey of Professional Forecasters and the Bank of England fan charts. This first evaluation of the fan chart forecasts finds that whereas the current-quarter forecasts are well-calibrated, this is less true of the one-year-ahead forecasts. The fan charts fan out too quickly, and the excessive concern with the upside risks was not justified over the period considered.

Keywords Forecast evaluation; interval forecasts; density forecasts; likelihood ratio tests; chi-squared tests; exact inference; Bank of England inflation forecasts

JEL classification C53, E37

1. Introduction

Interval forecasts and density forecasts are being increasingly used in practical real-time forecasting. An interval forecast of a variable specifies the probability that the future outcome will fall within a stated interval; this usually uses round numbers such as 50% or 90% and states the interval boundaries as the corresponding percentiles. A density forecast is an estimate of the complete probability distribution of the possible future values of the variable. As supplements to point forecasts they each provide a description of forecast uncertainty, whereas no information about this is available if only a point forecast is presented, a practice which is being increasingly criticized in macroeconomic forecasting. Density forecasts are more directly used in decision-making in the fields of finance and risk management. Tay and Wallis (2000) provide a survey of applications of density forecasting in macroeconomics and finance, and the examples in the present paper come from the former field.

Evaluating the accuracy of interval and density forecasts is similarly receiving increasing attention. For interval forecasts the first question is whether the coverage is correct *ex post*, that is, whether the relative frequency with which outcomes are observed to fall in their respective forecast intervals is equal to the announced probability. Christoffersen (1998) argues that this *unconditional* hypothesis is inadequate in a time-series context, and defines an efficient sequence of interval forecasts with respect to a given information set as one which has correct *conditional* coverage. He presents a likelihood ratio framework for conditional coverage testing, which combines a test of unconditional coverage with a test of independence. This supplementary hypothesis is directly analogous to the requirement of lack of autocorrelation of orders greater than or equal to the forecast lead time in the errors of a sequence of efficient point forecasts. It is implemented in a two-state (the outcome lies in the forecast interval or not) Markov chain, as a likelihood ratio test of the null hypothesis that successive observations are statistically independent, against the alternative hypothesis that the observations are from a first-order Markov chain.

For density forecasts the first question again concerns goodness-of-fit. Two classical methods of testing goodness-of-fit - the likelihood ratio and Pearson chi-squared tests - proceed 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 forecast densities with the observed relative frequencies. It is usually recommended to use classes with equal probabilities, so the class boundaries are quantiles; similarly, a standard way of reporting densities is in terms of their quantiles. For a sequence of density forecasts these change over time, of course. This approach reduces the density forecast to a k-interval forecast and sacrifices information, but the distinction between unconditional and conditional coverage extends to this case, and a likelihood ratio test of independence can be developed in a k-state Markov chain, generalising Christoffersen's proposal above.

It is well known that the likelihood ratio tests and Pearson chi-squared tests for these problems are asymptotically equivalent; for general discussion and references to earlier literature see Stuart, Ord and Arnold (1999, Ch. 25). In discussing this equivalence for the Markov chain tests they develop, Anderson and Goodman (1957) note that the chi-squared tests, which are of the form used in contingency tables, have the advantage that, "for many users of these methods, their motivation and their application seem to be simpler", and this point of view prompts the present line of enquiry. In this paper we accordingly explore the equivalent chi-squared tests for the hypotheses discussed above. The term chi-squared tests is used here and in the title of the paper to refer to Pearson's chi-squared statistic, memorable to multitudes of students as the formula $\Sigma(O-E)^2/E$. Asymptotically it has the c^2 distribution under the null hypothesis.

Two recent extensions to this framework are also considered. The first is the "rearrangement" by Anderson (1994) of the chi-squared goodness-of-fit test to provide more information on the nature of departures from the null hypothesis, in respect of specific features of the empirical distribution such as its location, scale and skewness. Second, since our attention is focussed on applications in macroeconomics, where sample sizes are as yet small, we consider the exact finite-sample distribution of the chi-squared statistic. While the contingency table literature contains an extensive discussion of finite-sample behaviour, going back to Fisher's exact test - see Yates (1984) for a review of methods for 2×2 tables - it is only relatively recently that convenient computational routines have become easily available. The calculation of exact *P*-values based on the permutational distribution of the test statistic uses methods surveyed by Mehta and Patel (1998), implemented in StatXact-4 (Cytel Software Corp.).

Two datasets are used below. The first is the U.S. Survey of Professional Forecasters' (SPF) density forecasts of inflation, previously analyzed by Diebold, Tay and Wallis (1999). The forecast densities are reported numerically, as histograms, and have no particular functional form. The fact that their skewness and kurtosis vary over time is well-established (Lahiri and Teigland, 1987), but there is no underlying model of this variation. The series of forecasts and outcomes are shown in Figure 1, where plotted percentiles have been obtained by linear interpolation of the published histograms. The second example is the Bank of England Monetary Policy Committee's density forecasts of inflation, which date from the establishment of the Committee in mid-1997. A specific functional form is assumed, with three parameters that determine location, scale and skewness. The underlying point forecast is based on a macroeconometric model, to ensure internal consistency across the range of variables considered, but like most macroeconomic forecasts, it is also subject to judgemental adjustment. In particular the skewness is based on the collective judgement of the ninemember Committee about the balance of risks to the forecast; this varies over time, again in an unmodelled way.

In both cases the absence of strict stationarity or of relatively simple constant-parameter models of time-varying moments, such as conditional heteroscedasticity, precludes the application of some recent developments in the forecast evaluation literature. One is the use of kernel estimation of the true conditional density function - see Li and Tkacz (2001) for a recent proposal and references to earlier literature - while another focuses on the effect of parameter estimation error on the evaluation procedures, following West (1996). Simple parametric models, such as members of the ARCH extended family, are more prevalent in applications in finance, but as noted by Tay and Wallis (2000), there are as yet few attempts to model and forecast skewness and excess kurtosis, which are well-known features of the unconditional distributions of many financial series.

An alternative group of goodness-of-fit tests, well-established in the literature, is based on the probability integral transformation. For a density forecast whose distribution function is $F(.)$, this is simply defined as $z = F(y)$, where y is the observed outcome: z is the forecast probability of observing an outcome no greater than that actually realised. If $F(.)$ is correct, then z has a uniform $U[0,1]$ distribution. If a sequence of density forecasts is correctly conditionally calibrated then, analogously to the no-autocorrelation requirement discussed above, Diebold, Gunther and Tay (1998) show that the corresponding z-sequence is iid $U[0,1]$; they present histograms of z for visual assessment of unconditional uniformity, and various autocorrelation tests. Diebold, Tay and Wallis (1999) use the Kolmogorov-Smirnov test on the sample distribution function of z in their evaluation of the SPF inflation forecasts. Comparison of the two approaches to testing is a subject of future research.

The remainder of this paper is organized as follows. Unconditional coverage and goodness-of-fit tests are discussed in Section 2, and tests of independence in Section 3; these are combined into joint tests of conditional coverage in Section 4. These sections use the SPF forecasts as illustrations, then Section 5 evaluates the Bank of England Monetary Policy Committee's density forecasts of inflation using these techniques.

2. Unconditional coverage and goodness-of-fit tests

For a sequence of interval forecasts with *ex ante* coverage probability p, the *ex post* coverage is p, the proportion of occasions on which the observed outcome lies in the forecast interval, and we wish to test the hypothesis of correct coverage. If in n observations there are n_1 outcomes falling in their respective forecast intervals and n_0 outcomes falling outside, then $p = n_1 / n$. From the binomial distribution the likelihood under the null hypothesis is

$$L(p) \propto (1-p)^{n_0} p^{n_1}$$

and the likelihood under the alternative hypothesis, evaluated at the maximum likelihood estimate p, is

$$L(p) \propto (1-p)^{n_0} p^{n_1} .$$

The likelihood ratio test statistic $-2\log[L(p)/L(p)]$, which is denoted $\mathrm{LR}_{\mathrm{uc}}$ by Christoffersen, is then

$$\mathrm{LR}_{\mathrm{uc}} = 2[n_0 \log(1-p)/(1-p) + n_1 \log(p/p)] .$$

The asymptotically equivalent chi-squared statistic is the square of the usual standard normal test statistic of a sample proportion, namely

$$X^2 = n(p-\pi)^2/\pi(1-\pi)\,.$$

Each can be compared to asymptotic critical values of the χ^2 distribution with one degree of freedom.

For interval forecasts the calibration of each tail individually may be of interest, as noted by Christoffersen. If the forecast is presented as a central interval, with equal tail probabilities, then the expected frequencies under the null hypothesis of correct unconditional coverage are $n(1-\pi)/2,\, n\pi,\, n(1-\pi)2$ respectively, and the chi-squared statistic comparing these with the observed frequencies has two degrees of freedom. This is a step towards goodness-of-fit tests for complete density forecasts, where the choice of the number of classes into which to divide the observed outcomes is typically related to sample size. Dividing the range of the variable into equiprobable classes is usually recommended with power considerations in mind; we denote the frequencies with which the outcomes fall into k such classes as $n_i,\, i=1,\ldots,k,\quad \Sigma n_i = n$. Equivalently, the range of the z-transform can be similarly divided, with class boundaries $j/k,\ j=0,1,\ldots,k,$ and the respective observed frequencies are the same. The chi-squared statistic for testing goodness-of-fit is

$$X^2 = \Sigma(n_i - n/k)^2/(n/k) = (k/n)\Sigma n_i^2 - n$$

and the likelihood ratio test statistic is

$$\mathrm{LR} = 2\Sigma n_i \log(kn_i/n)\,.$$

Each has a limiting χ^2 distribution with $k-1$ degrees of freedom under H_0.

The asymptotic distribution of the test statistic rests on the asymptotic k-variate normality of the multinomial distribution of the observed frequencies. Placing these in the $k\times 1$ vector $\boldsymbol{x}$, under the null hypothesis this has mean $\boldsymbol{\mu} = (n/k,\ldots,n/k)'$ and covariance matrix

$$\boldsymbol{V} = (n/k)[\boldsymbol{I} - \boldsymbol{ee'}/k]$$

where $\boldsymbol{e}$ is a $k \times 1$ vector of ones. This matrix is singular, with rank $k-1$, thanks to the restriction $\sum n_i = n$. The usual derivation of the limiting c^2 result proceeds by dropping one class and inverting the non-singular covariance matrix of the remaining $k-1$ n_i's: see, for example, Stuart and Ord (1994, Example 15.3). Alternatively, defining the generalized inverse $\boldsymbol{V}^-$, the result that $(\boldsymbol{x}-\boldsymbol{m})'\boldsymbol{V}^-(\boldsymbol{x}-\boldsymbol{m})$ has a c^2 distribution with $k-1$ degrees of freedom can be used directly: see, for example, Pringle and Rayner (1971, p.78). Since the matrix in square brackets above is symmetric and idempotent it coincides with its generalized inverse, and the chi-squared statistic given in the preceding paragraph is equivalently written as

$$X^2 = (\boldsymbol{x}-\boldsymbol{m})'[\boldsymbol{I} - \boldsymbol{ee'}/k](\boldsymbol{x}-\boldsymbol{m})/(n/k).$$

There exists a $(k-1) \times k$ transformation matrix $\boldsymbol{A}$ such that

$$\boldsymbol{AA'} = \boldsymbol{I}, \qquad \boldsymbol{A'A} = [\boldsymbol{I} - \boldsymbol{ee'}/k],$$

hence defining $\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{m})$ the statistic can also be written as

$$X^2 = \boldsymbol{y'y}/(n/k),$$

where the $k-1$ components $y_i^2/(n/k)$ are independently distributed as c^2 with one degree of freedom under H_0. Anderson (1994) introduces this device in order to focus upon particular moments or characteristics of the distribution of interest. For example, with $k = 4$ and

$$\boldsymbol{A} = \tfrac{1}{2}\begin{bmatrix} 1 & 1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix}$$

the three components focus in turn on departures from the null distribution with respect to location, scale and skewness. Such decompositions are potentially more informative about the nature of departures from the null hypothesis than the single "portmanteau" goodness-of-fit statistic.

For the SPF density forecasts of inflation shown in Figure 1, the frequencies of outcomes falling in 4 equiprobable classes can be easily read from the graph. The class boundaries are the quartiles, and ordering the classes from the lowest to the highest values gives the observed frequencies 2, 16, 3, 7. The chi-squared statistic has the value 17.43, suggesting that the null hypothesis of correct distribution should be rejected. The decomposition suggested in the preceding paragraph gives the contributions of departures with respect to location, scale and skewness as 2.29, 3.57 and 11.57 respectively. The major departure is with respect to skewness, resulting from the relative absence of large negative inflation surprises, as noted by Diebold, Tay and Wallis (1999), and the preponderance in the last twelve years of the sample of outcomes that are close to but just below the median.

3. Tests of independence

A test of independence against a first-order Markov chain alternative is based on a matrix of transition counts $[n_{ij}]$, where n_{ij} is the number of observations in state i at time $t-1$ and j at t. The maximum likelihood estimates of the transition probabilities are the cell frequencies divided by the corresponding row totals.

For an interval forecast there are two states - the outcome lies inside or outside the interval - and these are denoted 1 and 0 respectively. The estimated transition probability matrix is

$$\boldsymbol{P} = \begin{bmatrix} 1-p_{01} & p_{01} \\ 1-p_{11} & p_{11} \end{bmatrix} = \begin{bmatrix} n_{00}/n_{0\bullet} & n_{01}/n_{0\bullet} \\ n_{10}/n_{1\bullet} & n_{11}/n_{1\bullet} \end{bmatrix}$$

where replacing a subscript with a dot denotes that summation has been taken over that index. The likelihood evaluated at $\boldsymbol{P}$ is

$$L\left(\boldsymbol{P}\right) \propto (1-p_{01})^{n_{00}}\, p_{01}{}^{n_{01}} (1-p_{11})^{n_{10}}\, p_{11}{}^{n_{11}} .$$

The null hypothesis of independence is that the state at t is independent of the state at $t-1$, that is, $p_{01} = p_{11}$, and the maximum likelihood estimate of the common probability is $p = n_{\bullet 1}/n$. The likelihood under the null, evaluated at p, is

$$L(p) \propto (1-p)^{n_{\bullet 0}}\, p^{n_{\bullet 1}} .$$

This is identical to $L(p)$ defined in the previous section if the first observation is ignored. The likelihood ratio test statistic is then

$$\mathrm{LR}_{\mathrm{ind}} = -2\log[L(p)/L(\boldsymbol{P})]$$

which is asymptotically distributed as c^2 with one degree of freedom under the independence hypothesis.

This likelihood ratio test is asymptotically equivalent to the chi-squared test of independence in a 2×2 contingency table. Alternatively denoting the matrix of observed frequencies as

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

this has the familiar expression

$$X^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} .$$

Equivalently, it is the square of the standard normal test statistic for the equality of two binomial proportions. It is sometimes proposed to use the Yates continuity correction, so that the c^2 distribution better approximates the discrete distribution of X^2. The adjusted statistic is

$$X_c^2 = \frac{n(|ad - bc| - \frac{1}{2}n)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

Computer packages are now available to compute the exact distribution by enumerating all possible tables that give rise to a value of the test statistic greater than or equal to that observed and cumulating their null probabilities. Thus there is no reason for the Yates correction, nevertheless the following example provides an interesting illustration of its use.

An interval forecast with probability 0.5 is implicit in the SPF density forecasts of inflation, as the inter-quartile range represented by the boxes in Figure 1. The matrix of transition counts for these data is

$$\begin{bmatrix} 5 & 4 \\ 3 & 15 \end{bmatrix}$$

which gives a chi-squared statistic of 4.35, reduced to 2.69 by the Yates continuity correction. At the 5% significance level the adjusted statistic indicates that the null hypothesis should not be rejected, whereas the original Pearson statistic indicates the opposite, using the asymptotic $c^2(1)$ critical value.

The exact *P*-value for the observed table is given by StatXact-4 as 0.072, which leads to the same decision as the adjusted statistic. This is a “conditional” test in the sense that it treats the marginal totals as fixed: this issue has been long debated, and the consensus of Yates (1984) and his discussants is that this is “the only rational test”. Conditioning on the margins of the observed contingency table for the purpose of inference eliminates nuisance parameters and is justified by sufficiency and ancillarity principles; it does not require that the margins are actually fixed in the data generating process. However it implies that the resulting (hypergeometric) distribution is highly discrete in small samples. Given the marginal totals of a 2×2 table, the entry in any one cell determines the other three. The top left cell in our example can take integer values between 0 and 8, thus the chi-squared statistic has only 9 possible values, whose probabilities under H_0 are given by the hypergeometric distribution. To test H_0 the *P*-value is the sum of these probabilities for values of the statistic greater than or equal to that observed. This in turn has rather few possible values, hence a formal test using a conventional significance level such as 0.05 is in general *conservative*, since its actual size is smaller than this. In the present example, the next possible value of the chi-squared statistic is 5.68, which has *P*-value 0.026. A less conservative test can be based on the *mid P-value*, which is half the probability of the observed statistic plus the probability of values greater than that observed, here equal to 0.049. Alternatively, rather than deciding whether to "accept" or "reject" the null hypothesis, one can simply regard the *P*-value as a measure of the degree to which the data support H_0.

Various generalizations of the 2×2 test are immediate. For interval forecasts each tail may be considered separately, as noted above, so that the Markov chain has three states, as the outcome lies in the lower tail, the forecast interval, or the upper tail. The resulting chi-squared statistic (with 4 degrees of freedom) is used as an "autocorrelation" measure by Granger, White and Kamstra (1989) in evaluating their interval forecasts based on ARCH-quantile estimators, although they do not make explicit the Markov chain framework. For the SPF density forecasts in Figure 1 the matrix of transition frequencies becomes

$$\begin{bmatrix} 0 & 2 & 0 \\ 2 & 15 & 1 \\ 0 & 2 & 5 \end{bmatrix},$$

which disaggregates the a, b and c cells of the previous 2×2 array. In the event, the a entry of 5 is simply relocated to the bottom right corner, as the result of the first six outcomes all lying in the upper quartile of the forecast densities. The chi-squared statistic is 13.38, the major contribution coming from this bottom right cell. The asymptotic P-value based on $\chi^2(4)$ is 0.008 and the exact P-value 0.007. The distinction between the two tails of the forecast densities and the initial sequence of observations in the upper tail results in stronger evidence against independence than that provided by the 2×2 table.

The generalization to density forecasts grouped into k classes is also immediate, although with sample sizes that are typical in practical macroeconomic forecasting the resulting table is likely to be sparse once k gets much beyond 2 or 3, as in the above example. Conventional time-series tests based on the z-series, as used by Diebold *et al.* (1998, 1999), are then likely to be more informative. They also facilitate the investigation of possible higher-order dependence, given the disadvantage of the Markov chain approach that the dimension of the transition matrix increases with the square of the order of the chain. However, such expansion can be avoided if a particular periodicity is of interest, by defining transitions from time $t-l$ to t, where l is the length of the period, ignoring intermediate movements, as noted by Clements and Taylor (2001). This would be appropriate in testing the efficiency of a quarterly series of one-year-ahead forecasts, for example, by checking independence at lag four.

4. Joint tests of coverage and independence

Christoffersen proposes a likelihood ratio test of conditional coverage as a joint test of unconditional coverage and independence. This is a test of the null hypothesis of Section 2 against the alternative of Section 3, and the likelihood ratio test statistic is

$$\mathrm{LR}_{\mathrm{cc}} = -2\log[L(p)/L(\boldsymbol{P})].$$

Again ignoring the first observation the test statistics obey the relation

$$\mathrm{LR}_{\mathrm{cc}} = \mathrm{LR}_{\mathrm{uc}} + \mathrm{LR}_{\mathrm{ind}}.$$

Asymptotically $\mathrm{LR}_{\mathrm{cc}}$ has a c^2 distribution with two degrees of freedom under the null hypothesis. The alternative hypothesis for $\mathrm{LR}_{\mathrm{ind}}$ and $\mathrm{LR}_{\mathrm{cc}}$ is the same, and these tests form an ordered nested sequence.

The asymptotically equivalent chi-squared test compares the observed contingency table with the expected frequencies under the joint hypothesis of row independence and correct coverage probability p. The statistic has the usual form $\Sigma(O-E)^2/E$ where the observed and expected frequencies are respectively

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} (1-p)(a+b) & p(a+b) \\ (1-p)(c+d) & p(c+d) \end{bmatrix}.$$

It has two degrees of freedom since the column proportions are specified by the hypothesis under test and not estimated. The statistic is equal to the sum of the squares of two standard normal test statistics of sample proportions, one for each row of the table.

Although the chi-squared statistics for the separate and joint hypotheses are asymptotically equivalent to the corresponding likelihood ratio test statistics, in finite samples they obey the additive relation among those statistics given above only approximately, and not exactly.

For the matrix of transition counts of the SPF inter-quartile range forecasts given on page 10 the chi-squared test statistic for the joint hypothesis is 8.11. Its exact P-value in the two binomial proportions model is 0.018, indicating rejection of the joint hypothesis. This confirms the conclusion reached by Diebold, Tay and Wallis (1999) by less formal methods. The chi-squared statistic testing unconditional coverage on the column totals is 4.48, and with the test statistic for independence from Section 3 of 4.35 we note the lack of additivity.

5. Bank of England fan chart forecasts

The Bank of England has published a density forecast of inflation in its quarterly *Inflation Report* since February 1996. The forecast is represented graphically as a set of prediction intervals covering 10%, 20%,…,90% of the probability distribution, of lighter shades for the outer bands. This is done for inflation forecasts one to nine quarters ahead, and since the dispersion increases and the intervals "fan out" as the forecast horizon increases, the result has become known as the "fan chart". Contrary to the initial suggestion of Thompson and Miller (1986) to use the selective shading of *quantiles* "to draw attention away from point forecasts and toward the *uncertainty* in forecasting," the Bank's preferred presentation is based on the shortest intervals for the assigned probabilities, hence the tail probabilities are typically unequal, moreover they are not reported. An example is shown in Figure 2, whereas Figure 3 shows an alternative presentation of the same forecast based on percentiles, as recommended by Wallis (1999). These differ because the distribution is asymmetric, usually positively skewed, and the accompanying discussion often emphasises the distinction between the "upside risks" and the "downside risks" to the forecast. The forecast is represented analytically by the two-piece normal distribution (John, 1982; Wallis, 1999), for which probabilities can be readily calculated from standard normal tables once values have been assigned to the underlying parameters that determine its location, dispersion and skewness.

The introduction of new arrangements for the operation of monetary policy in 1997 saw the establishment of the Monetary Policy Committee (MPC), which in particular adopted the Bank's existing forecasting practice. Evaluations of forecasts published to date, in the *Inflation Reports* of August 1999 and August 2000, have analysed "The MPC's forecasting record" beginning with its first inflation projection published in August 1997, focussing on the one-year-ahead forecasts. These forecasts, extended to the first three years, are shown in the upper panel of Table 1, together with inflation outcomes and associated *z*-values calculated via formulae given by Wallis (1999). The definition of inflation is the annual percentage growth in the quarterly Retail Prices Index excluding mortgage interest payments (RPIX, Office for National Statistics code CHMK). Strictly speaking, the forecasts are conditional projections, which assume that interest rates remain at the level just set by the MPC. Nevertheless it is argued that they can be evaluated as unconditional forecasts, comparing the mean projections with actual outcomes, for example, since inflation does not react rapidly to interest rate changes and, in any event, the actual profile of interest rates has turned out relatively close to the constant interest rate assumption.

Point forecast evaluations usually focus on the conditional expectation, the mean of the forecast density, and the *Inflation Report* evaluations of the one-year-ahead forecasts do likewise, despite the focus on the mode, the most likely outcome, in the MPC's forecast commentary and press releases. With the usual definition of forecast error as outcome minus forecast, the mean forecasts in the upper panel of Table 1 have an average error of -0.20: on average inflation has been overestimated by 0.2 percentage points. The standard error of the mean is 0.11, hence the null hypothesis of unbiasedness would be rejected against the onesided alternative of an upward bias at the 5% significance level. The standard deviation of the forecast errors is 0.38, indicating that the standard deviation used in preparing the fan chart is an overestimate. This supports the comment (*Inflation Report*, August 2000, p.63) that, having based the fan chart variances on forecast errors over the past ten years, outturns that tend to lie close to the centre of the distribution suggest that recent forecast errors have been smaller than in the past.

Reducing the density forecasts to interval forecasts based on the central 50% interval or inter-quartile range, as in the SPF example above, gives the sequence of states defined in Section 3 as

1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 0, 1.

The matrix of transition counts for this sequence is

$$\begin{bmatrix} 2 & 2 \\ 2 & 5 \end{bmatrix}$$

with chi-squared statistic 0.505. This is one of only five possible values of the conditional test statistic, with mid *P*-value 0.386. The test takes no account of the fact that all the zero states relate to outcomes in the lower tail, however, but in this case the three-way classification of Granger *et al.* (1989) illustrated in Section 3 is not helpful, since the above 2×2 array is simply bordered below and on the right by zeros. The class frequencies for the four-way goodness-of-fit test used in the closing paragraph of Section 2 are 4, 4, 4, 0, giving a test statistic of 4.0. This provides little evidence against the null hypothesis, nevertheless the absence of outcomes in the top quartile suggests an exaggerated concern with the upside risks on the part of the MPC. The decomposition of the test statistic into components associated with location, scale and skewness coincidentally gives exactly the same weight to departures in each of the three moments, reflecting the preceding discussion. A further picture of the general departure of the fan chart forecasts from the correct density is given in Figure 4, which compares the sample distribution function of the observed *z*-values with the uniform distribution function, the 45° line representing the null hypothesis of a correct density. It is seen that the density forecasts place too much probability in the upper ranges of the inflation forecast.

Although the errors of a quarterly series of optimal one-year-ahead point forecasts would be expected to exhibit low-order autocorrelation, the simple device used above indicates no substantial departure from independence in the implied interval forecasts. However, little is known about how the corresponding lack of independence in density forecasts might manifest itself and hence be efficiently detected. For this question the one-step-ahead forecasts are then of interest, although they have been neglected in published discussion to date. In practical forecasting the first thing one has to do is forecast the present, and the fan chart is no exception, the first forecast shown being for the current quarter. The MPC normally meets on the Wednesday and Thursday following the first Monday of each month, and the quarterly *Inflation Report*, containing the fan chart forecasts, is published in February, May, August and November a week after the meeting. At approximately the same time, in mid-month, the Office for National Statistics releases the previous month's Retail Prices Index. Thus the quarterly forecast can take account of no current-quarter information on the variable in question, and so can be regarded as a one-step-ahead forecast, although other economic intelligence on the first month of each quarter is clearly available to the MPC at its mid-quarter meeting. Correspondingly, the year-ahead forecasts discussed above are, in effect, five-step-ahead forecasts.

Data on the one-step-ahead forecasts and outcomes are shown in the lower panel of Table 1. The mean forecasts have a mean error of 0.001 and an RMSE of 0.17. Unlike the year-ahead forecasts, the one-quarter-ahead forecasts are unbiased, and now the forecast standard deviation in general appears to be correct. The transition frequencies for an inter-quartile range interval forecast are

$$\begin{bmatrix} 4 & 4 \\ 4 & 3 \end{bmatrix},$$

which are within rounding of the expected values under the independence hypothesis. The sample distribution function of the observed *z*-values for these forecasts and outcomes shown in Figure 5 lies much closer to the 45° degree line than in the previous case, the complete range of the densities being better represented in the data. Overall, the evidence suggests that the current-quarter forecasts are conditionally well-calibrated.

The initial conclusion from these rather short series of forecasts is that calibration problems arise as the MPC moves beyond the current-quarter forecasts: the fan charts fan out too quickly, positive inflation shocks have occurred much less frequently than the MPC expected, and on average inflation one year ahead has been overestimated.

6. Conclusion

The increasing use of interval and density forecasts is a welcome development. They help to assess and communicate future uncertainty, whereas a point forecast alone gives no guidance as to its likely accuracy. If the forecast is an input to a decision problem, then once one moves beyond the circumstances in which certainty equivalence holds - to an asymmetric cost function, for example - a point forecast is inadequate and a density forecast is required. The accuracy of interval and density forecasts in turn requires assessment, and this paper recasts some recent proposals into a familiar framework of contingency tables, which increases their accessibility for many users of these methods and allows other recent developments to be incorporated. In many important applications only small samples are available for evaluation, and the calculation of exact *P*-values for the statistics considered is advocated.

The density forecasts of inflation made by the Bank of England's Monetary Policy Committee have received much attention in the United Kingdom and in other central banks. Our first evaluation of the series of "fan chart" forecasts published since the MPC's inauguration in 1997 shows that, whereas the current-quarter forecasts perform well, the one-year-ahead forecasts, on which the Bank's evaluations have hitherto exclusively focussed, show negative bias. Uncertainty in these forecasts has been overestimated, so that the fan charts fan out too quickly, and the excessive concern with the upside risks has not been justified over this period.

References

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

Anderson, T.W. and Goodman, L.A. (1957). Statistical inference about Markov chains. *Annals of Mathematical Statistics*, 28, 89-110.

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

Clements, M.P. and Taylor, N. (2001). Evaluating interval forecasts of high-frequency financial data. Unpublished paper, University of Warwick.

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

Diebold, F.X., Tay, A.S. and Wallis, K.F. (1999). Evaluating density forecasts of inflation: the Survey of Professional Forecasters. In *Cointegration, Causality, and Forecasting: A Festschrift in Honour of Clive W.J. Granger* (R.F. Engle and H. White, eds), pp.76-90. Oxford: Oxford University Press.

Granger, C.W.J., White, H. and Kamstra, M. (1989). Interval forecasting: an analysis based upon ARCH-quantile estimators. *Journal of Econometrics*, 40, 87-96.

John, S. (1982). The three-parameter two-piece normal family of distributions and its fitting. *Communications in Statistics - Theory and Methods*, 11, 879-885.

Lahiri, K. and Teigland, C. (1987). On the normality of probability distributions of inflation and GNP forecasts. *International Journal of Forecasting*, 3, 269-279.

Li, F. and Tkacz, G. (2001). A consistent bootstrap test for conditional density functions with time-dependent data. Working Paper, Bank of Canada.

Mehta, C.R. and Patel, N.R. (1998). Exact inference for categorical data. In *Encyclopedia of Biostatistics* (P. Armitage and T. Colton, eds), pp.1411-1422. Chichester: John Wiley.

Pringle, R.M. and Rayner, A.A. (1971). *Generalized Inverse Matrices with Applications to Statistics*. London: Charles Griffin.

Stuart, A. and Ord, J.K. (1994). *Kendall's Advanced Theory of Statistics*, 6th ed., vol.1. London: Edward Arnold.

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

Tay, A.S. and Wallis, K.F. (2000). Density forecasting: a survey. *Journal of Forecasting*, 19, 235-254. Reprinted in *Companion to Economic Forecasting* (M.P. Clements and D.F. Hendry, eds). Oxford: Blackwell, 2002.

Thompson, P.A. and Miller, R.B. (1986). Sampling the future: a Bayesian approach to forecasting from univariate time series models. *Journal of Business and Economic Statistics*, 4, 427-436.

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.

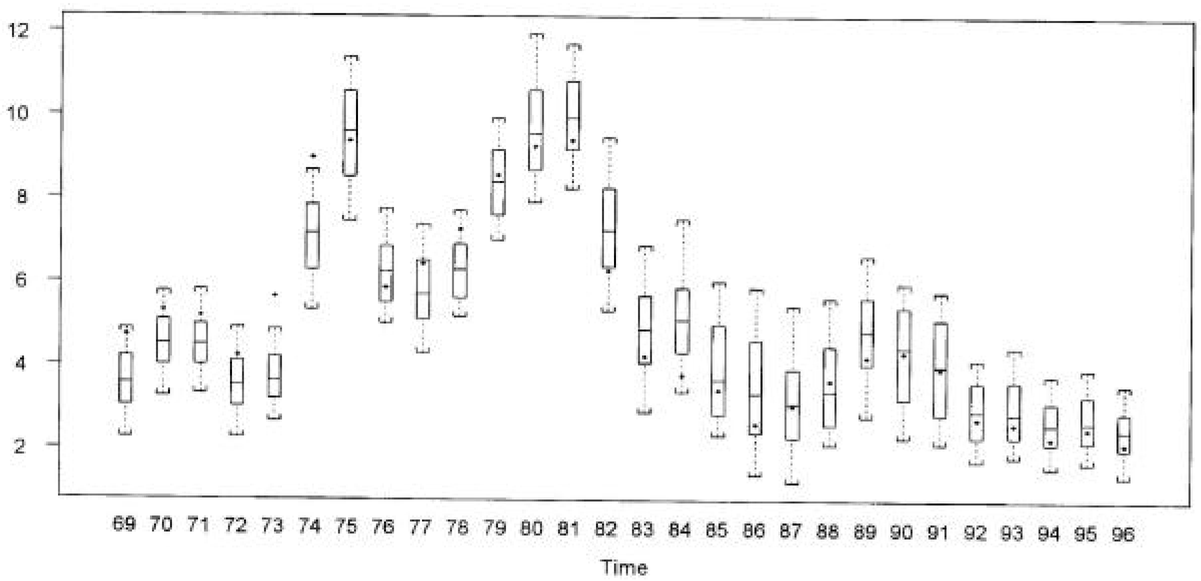
West, K.D. (1996). Asymptotic inference about predictive ability. *Econometrica*, 64, 1067-1084.

Yates, F. (1984). Tests of significance for 2x2 contingency tables (with discussion). *Journal of the Royal Statistical Society,* A, 147, 426-463.

Table 1. Bank of England Monetary Policy Committee Inflation Forecasts

Inflation Report	Mode	Mean	Std. Dev.	Outcome	*z*
		One-year-ahead forecasts			
Aug 97	1.99	2.20	0.75	2.55	0.69
Nov 97	2.19	2.84	0.61	2.53	0.37
Feb 98	2.44	2.57	0.60	2.53	0.49
May 98	2.37	2.15	0.61	2.30	0.57
Aug 98	2.86	3.00	0.60	2.17	0.08
Nov 98	2.59	2.72	0.62	2.16	0.18
Feb 99	2.52	2.58	0.62	2.09	0.22
May 99	2.23	2.34	0.59	2.07	0.33
Aug 99	1.88	2.03	0.56	2.13	0.59
Nov 99	1.84	1.79	0.55	2.11	0.72
Feb 00	2.32	2.42	0.56	1.87	0.16
May 00	2.47	2.52	0.55	2.26	0.32
		Current quarter (one-step-ahead) forecasts			
Aug 97	2.65	2.69	0.15	2.81	0.79
Nov 97	2.60	2.73	0.12	2.80	0.75
Feb 98	2.60	2.64	0.24	2.59	0.43
May 98	2.83	2.74	0.24	2.94	0.79
Aug 98	2.51	2.56	0.24	2.55	0.49
Nov 98	2.54	2.58	0.19	2.53	0.41
Feb 99	2.49	2.51	0.19	2.53	0.54
May 99	2.48	2.51	0.18	2.30	0.12
Aug 99	2.31	2.35	0.17	2.17	0.13
Nov 99	2.20	2.19	0.17	2.16	0.44
Feb 00	1.93	1.96	0.17	2.09	0.78
May 00	1.88	1.89	0.17	2.07	0.84
Aug 00	2.38	2.38	0.16	2.13	0.06
Nov 00	2.36	2.37	0.17	2.11	0.05
Feb 01	1.94	1.92	0.17	1.87	0.42
May 01	1.90	1.88	0.17	2.26	0.99

Figure 1
SPF Inflation Forecasts and Realizations, 1969-1996



Notes: The density forecasts are represented by box-and-whisker plots. The box gives the inter-quartile range of the forecast, and the inner line is the median; the tails are the 10^{th} and 90^{th} percentiles. **Inflation realizations are denoted ◆.**
Source: Diebold, Tay and Wallis (1999).

Figure 2 The August 1997 *Inflation Report* fan chart

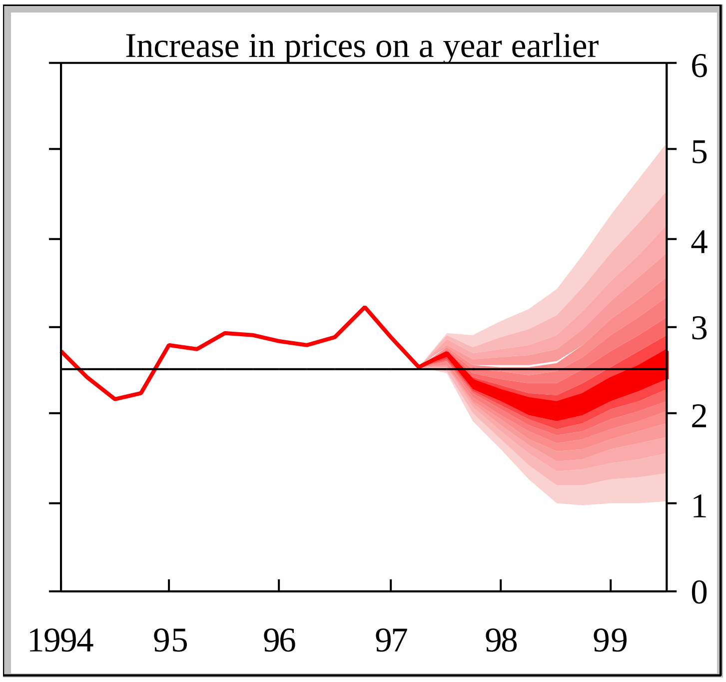


Figure 3 Alternative fan chart based on central prediction intervals

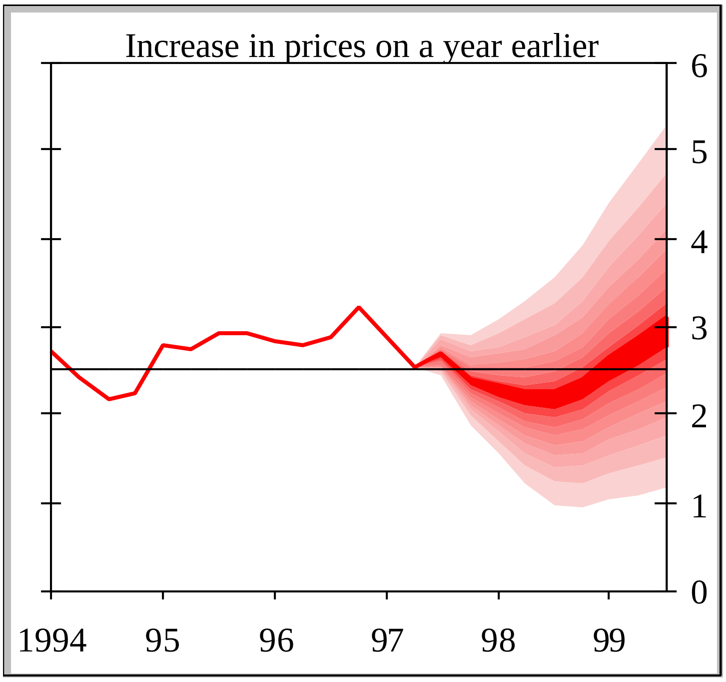


Figure 4

MPC year-ahead forecasts: distribution functions
of sample z-values (n=12) and uniform distribution

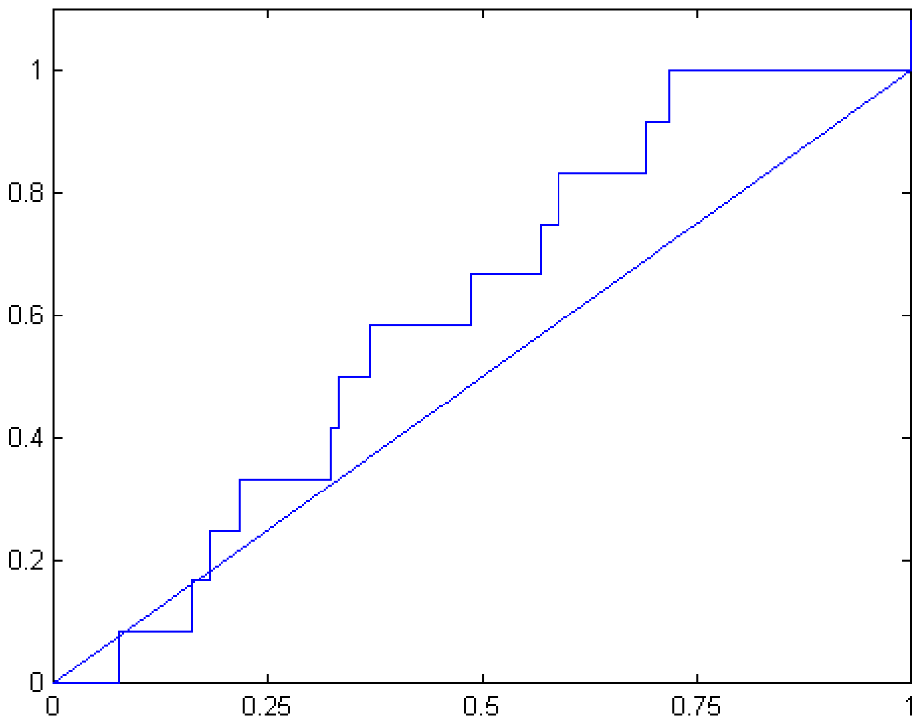


Figure 5

MPC current-quarter forecasts: distribution functions of sample z-values (n=16) and uniform distribution