Evaluating Interval and Probability Forecasts

Unlike point forecasts discussed in Forecasting Note #2, interval and probability forecasts convey the uncertainty of a future prediction, and thus are more complex to evaluate.

Interval and probability forecasts are statements about the likelihood that future outcomes of a given variable will either fall within a certain range (interval forecast), or will occur a certain fraction of the time (probability forecast). In practice, we do not observe the entire probability distribution of the outcome at each point in time, but rather observe only a single realization of the predicted variable – the "actual" value. This means that we cannot evaluate probability or density forecasts by simply comparing them to past distributions. This note explains how to evaluate and validate probability and density forecasts.

Evaluating a Single Interval Forecast

Consider an interval forecast that takes the form of a prediction that:

- with probability two-thirds, the predicted outcome of the variable in question will lie between zero and six percent;
- with probability one-third, the outcome will be either negative, or above six percent.

Suppose the actual outcome turns out to be seven percent and, thus, falls outside the interval forecast. This observed outcome would be somewhat surprising given the interval forecast, whereby the outcome is predicted to be twice as likely to fall inside the [0, 6] interval as it is to fall outside. Still, outcomes outside the [0,6] interval are expected to occur one-third of the time, and so are not unlikely.

After observing a single realization of the variable, as in this example, it is generally not possible to tell whether the interval forecast was appropriate or "correct". This is especially true the closer the probability associated with the interval forecast is to 50%. For example, suppose that the interval forecast states that the variable will fall in the interval [1,5] with 50% probability. Since it is equally likely that the outcome falls inside or outside this 50% interval forecast, we do not learn about the quality of the forecast from observing a single realization. In contrast, if an interval forecast has probability 99% associated with it, outcomes that fall outside of the predicted interval would be informative, as they are unlikely to occur and thus might be indicative of a poorly specified interval forecast.

Testing that the Interval Forecast is Correct on Average

To evaluate the quality of an interval forecast model, it is often necessary to observe interval forecasts and corresponding outcomes at several points in time, to obtain a track record of the forecasts. This track record makes it possible to tell whether the coverage of the interval forecast is correct on average, and the longer the track record, the better the ability to evaluate. For an α% interval forecast, we could expect that the outcome falls inside the interval forecast α% of the time, and outside the interval forecast (100- α)% of the time. For example, for α = 50%, we would expect the outcome to fall inside the interval forecast as often as it falls outside it.

Suppose we observe, based on the track record of the variable and interval forecasts, that the variable's values fall outside the 50% interval forecast [1,5] 70% of the time. In this example, the 50% interval forecast is too narrow on average and could be widened to, say, [-1, 7]. If, instead, we observe that the outcome falls outside of the 50% interval forecast 30% of the time, the original interval forecast may be too wide and could be narrowed to, say, [2, 4].

Figure 1 plots the time-series of the growth rate of Apple's quarterly revenue (blue line) against point forecasts generated by a simple time-series model for the period 2004Q2-2015Q3, a total of 46 quarters (dashed black line). We see that the point forecasts are much smoother than actual revenue data – a common feature of predictions.

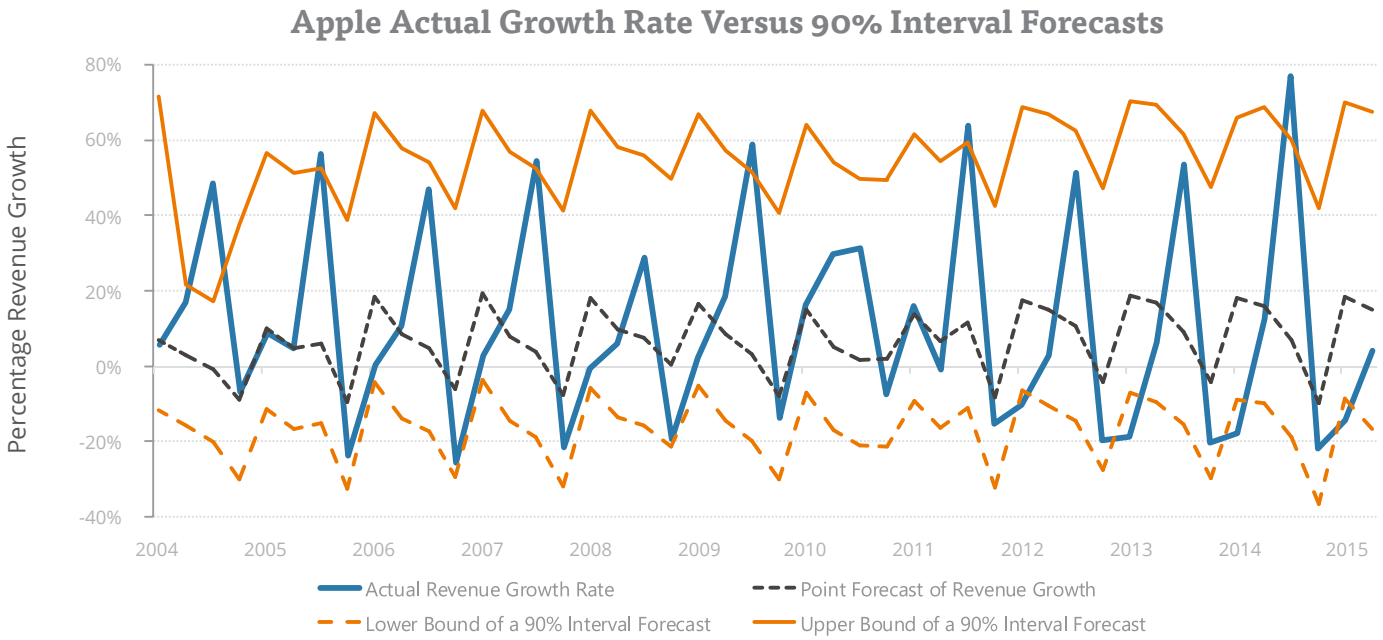

Figure 1 shows actual revenues, predicted revenues, and upper and lower bounds for the 90% interval forecast of Apple's one-quarter-ahead revenues.

As noted above, point forecasts do not provide the degree of uncertainty surrounding the prediction, so the graph also shows a 90% interval forecast, whose upper and lower bounds are marked in orange in the graph. This series of interval forecasts predicts that the actual revenue growth for Apple will fall inside these bounds with 90% probability, or 90% of the time. They further predict that the revenue growth may fall below the lower bound about 5% of the time, and above the upper bound about 5% of the time. Comparing the frequency with which outcomes fall inside a sequence of interval forecasts, that is, measuring how often the blue line falls inside the orange lines, is a test of whether the width of the interval forecast is correct on average – a test of the so-called "unconditional coverage". Such tests can be quantified using an inclusion indicator, which is a variable that takes a value of one whenever the outcome falls inside the interval forecast, and is zero otherwise. Testing that an interval forecast has the right unconditional coverage amounts to testing that the time-series of the inclusion indicator is equal to one about 90% on average over time. Said another way, the test measures whether the inclusion indicator has the right mean, which should be equal to the coverage rate, α – in this example, 90%. This test is similar to the test of unbiasedness for point forecasts, discussed in our Forecast Note #2.

Returning to Figure 1, whenever the actual value falls outside the orange bands, the indicator value takes a value of zero. In our example for Apple in Figure 1, we have 46 quarterly observations, and thus would expect that the actual value would fall outside the interval forecast around 10% of 46 cases, or around five times.

Figure 2 shows the time series for the inclusion indicator. Out of the 46 quarterly revenue forecasts, 10 are outside the 90% interval forecast, and the inclusion indicator obtains a value of zero for these instances. Moreover, we see a clear pattern in the indicator variable towards the end of the sample: the indicator variable takes a value of zero in the second quarter of each of the last four years, i.e., from 2012 onwards. This interval forecast may thus be too narrow, as well as possibly too predictable in the later period, as we will discuss below.

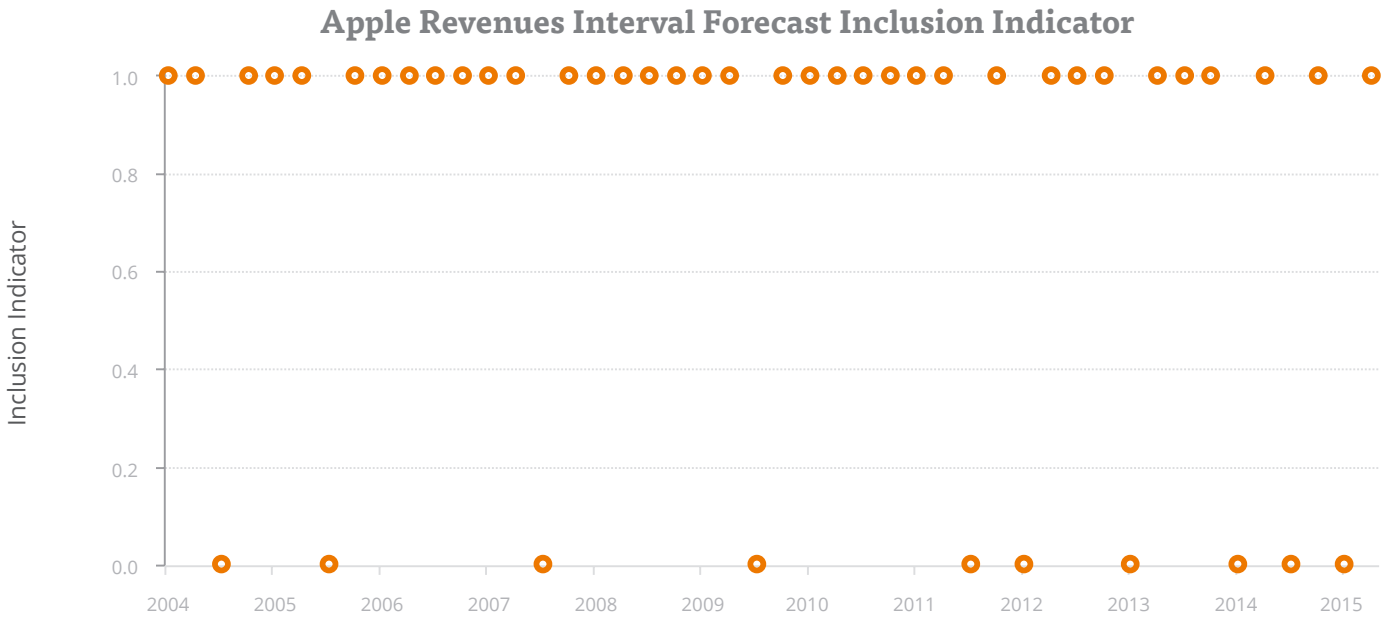

Figure 2 shows periods where Apple's actual revenues fall inside the 90% interval forecast from Figure 1 (inclusion indicator = 1) or fall outside the 90% interval forecast (inclusion indicator = 0).

Evaluating Fan Charts

As discussed in Forecast Note $1, fan charts incorporate interval forecasts reported at several levels of confidence – e.g., α = 25%, 50%, 75%, 90% - and for different forecast horizons ranging, e.g., from one quarter through a few years. We can use such merged interval forecast information to test both the quality of individual interval forecasts, as well as the quality of the joint forecasts. Said another way, we can test not only whether, say, the 50% one-quarter-ahead interval forecast for revenues is correctly specified, but also whether the interval forecasts are jointly correctly specified.

At any given forecast horizon, say, one quarter ahead, we expect that the 75% interval forecast should be wider than the 50% interval forecast. In fact, 25% of the time the outcome should fall inside the 75% interval forecast but outside the 50% interval forecast, since the former generally contains the latter. We can test this feature in the same way that we test if individual interval forecasts are correctly specified, using inclusion indicators for each interval forecast.

By comparing interval forecasts, for any significance level, and across different horizons, we can test whether the forecasting model correctly tracks the expected growth of uncertainty about a particular variable as the forecast horizon is extended.

Note that uncertainty need not always grow over time. For example, in the fall of 2008, short-term uncertainty was extremely high and interval forecasts for many economic variables would have been very wide. However, looking further into the future, e.g. at a five-year forecast horizon, the economy would have been expected to stabilize, and so the uncertainty surrounding outcomes at this longer horizon could well have been lower. In general, we only expect long-run uncertainty, and thus the width of the interval forecasts, to be greater than short-run uncertainty on average.

Testing that the Interval Forecast is Correct at Each Point in Time

In addition to having the right coverage on average, correctly specified interval forecasts have other features that can be tested. Most notably, the event that the predicted variable falls outside of the interval forecast at a given point in time should be independent of the outcome falling outside the interval forecast in the previous period or in the subsequent period.

The property of independence between consecutive outcomes falling inside or outside of their respective interval forecasts suggests other, stronger, tests. Specifically, in the same way that the forecast error associated with a point forecast should not be predictable ahead of time, the time series of the inclusion indicator for an interval forecast should not be predictable.

As an illustration, consider again the 50% interval forecast. At a given point in time, there is a 50% chance that the actual outcome falls outside the interval forecast. Suppose that we observe that ten consecutive observations fall outside of the 50% interval forecast. If these were independent events, as they should be, then the chance of observing this stream of consecutive outcomes should only be 0.5 raised to the power ten – less than one in a thousand.

Just like a point forecast, i.e. the forecast of the most likely outcome, moves around over time, interval forecasts will also shift around as a result of changes in the underlying economic conditions. The two main drivers of such time variation in interval forecasts are shifts in the mean of the variable being forecasted, and shifts in the volatility, i.e. uncertainty, surrounding the outcome. If the mean forecast changes, but the uncertainty surrounding the outcome remains the same, the entire interval forecast shifts up or down. Conversely, if the mean forecast remains unchanged, but the uncertainty increases, the center of the interval forecast may remain the same, but the interval forecast widens. In practice, we often see movements in both the location and width of interval forecasts occurring over time. If an interval forecast properly tracks shifts in the expected outcome and the uncertainty surrounding it, we would not expect to see long "runs" of observations outside interval forecasts with a reasonably high coverage such as $\alpha = 90\%$.

If it does happen that a sequence of outcomes outside the interval forecast is observed, its nature can be suggestive of ways in which the interval forecasts are misspecified, and could be improved. To see this, suppose that a forecasting model ignores periods of high volatility and assumes constant volatility for the outcome variable, equal to its average volatility. In that situation, during spells of high volatility in the data, there is a higher chance of observing "extreme" values, i.e. very large negative or very large positive outcomes, which are likely to fall outside interval forecasts built under the assumption of constant volatility. To the extent that high-volatility periods are concentrated in time, we would expect to find similar clusters of zeroes in the inclusion indicator. Conditional on an initial observation falling outside the interval forecast, this type of pattern would allow the forecaster to predict, ahead of time, that the outcome will fall outside the interval forecast with a higher probability than is stipulated by the forecasting model. Such information would indicate a misspecified forecasting model and, thus, could allow an improvement to the performance of the model by means of incorporating time-varying volatility.

Returning to Figure 2, the tendency of Apple's actual revenues to fall outside the 90% interval forecast in the second quarter of each of the last four fiscal periods suggests that the forecasts may not sufficiently account for the seasonal component in Apple's recent revenues.

Testing Probability Forecasts

As discussed in Forecasting Note #1, probability forecasts take the form of a probability distribution for the unknown (future) outcome variable, and are statements about how likely different outcomes are in the future. The higher the probability is for a particular value of the future outcome, the more likely it is that the outcome falls at or near that value.

Just as with other types of forecasts, it can be helpful to examine the validity and "fit" of such probability forecasts, and test whether outcomes in different parts of the distribution occur with a frequency that is consistent with the probability forecast. For example, a forecasting model may indicate that the tails of the outcome distribution are relatively thick, which implies that outcomes in the tails of the distribution, far from the average, are relatively likely to occur. In other words, a thick-tailed probability forecast implies that very large changes to the outcome are more likely to occur than if the tails of the distribution were thin. A model may also indicate a left-skewed distribution, in which case large negative outcomes are more likely than large positive outcomes.

Similarly to interval forecasts, because we never observe the "true" outcome distribution, but only see a single draw from this distribution at each point in time, again we can study a sequence of probability forecasts and outcomes in order to be able to evaluate if the probability forecasts are properly specified. Poorly specified probability forecasts can appear as frequent observations of outcomes with a low predicted probability or, in the opposite case, as observations predicted to be likely to occur, but occurring rarely in actuality.

The Probability Integral Transform

One can test whether the probability forecasts are correctly specified by means of the so-called probability integral transform, or PIT. The PIT measures the probability, expressed as a fraction of one, of observing an outcome smaller than or equal to the actual realization of the variable.

For example, outcomes that are so small that they correspond to the 1%, 5% and 0.5% left tails of their respective probability forecasts would be assigned PIT values of 0.01, 0.05, and 0.005. Outcomes in the middle of the probability distribution would be assigned PIT values close to 0.5, and outcomes in the far right tail of the probability distribution would be assigned PIT values close to one. For example, a large outcome which is only expected to be exceeded with a 1% probability would be assigned a PIT value of 0.99.

For correctly specified probability forecasts, the PIT values should be uniformly distributed, so that all PIT values between zero and one are equally likely to occur. To understand this property, suppose we observe a sample of 10 probability forecasts, and 10 corresponding observations of the variable, and that three of the 10 associated outcomes are in the far left tail of the probability forecasts with PIT values of 0.01, 0.03 and 0.005. Such outcomes would be highly unlikely if the probability forecasts were correctly specified; in a sample with ten observations, we would only expect to find one observation with a PIT value below 0.10, and PIT values at or below 0.01 should occur only one percent of the time. This suggests that the probability distribution in the example underestimates the likelihood of large negative outcomes, and that its left tail is too thin.

Returning to the time-series model for Apple's quarterly revenues, **Figure 3** below shows the time-series of PIT values over the period 2004Q2 – 2015Q3. The very high Q4 revenues in each of the last four years from 2012-2015 give rise to a strong seasonality with PIT scores that are very high (above 0.88), while conversely, the very low revenues in Q2 produce PIT values below 0.05 for each of these four years. In other words, this PIT test produces values that should be unlikely to occur as often as they do in the last four years of the sample. This pattern clearly shows that the simple time-series model used to produce this probability forecast is incapable of handling the seasonality in Apple's revenue stream, in this case associated with the introduction of the iPhone.

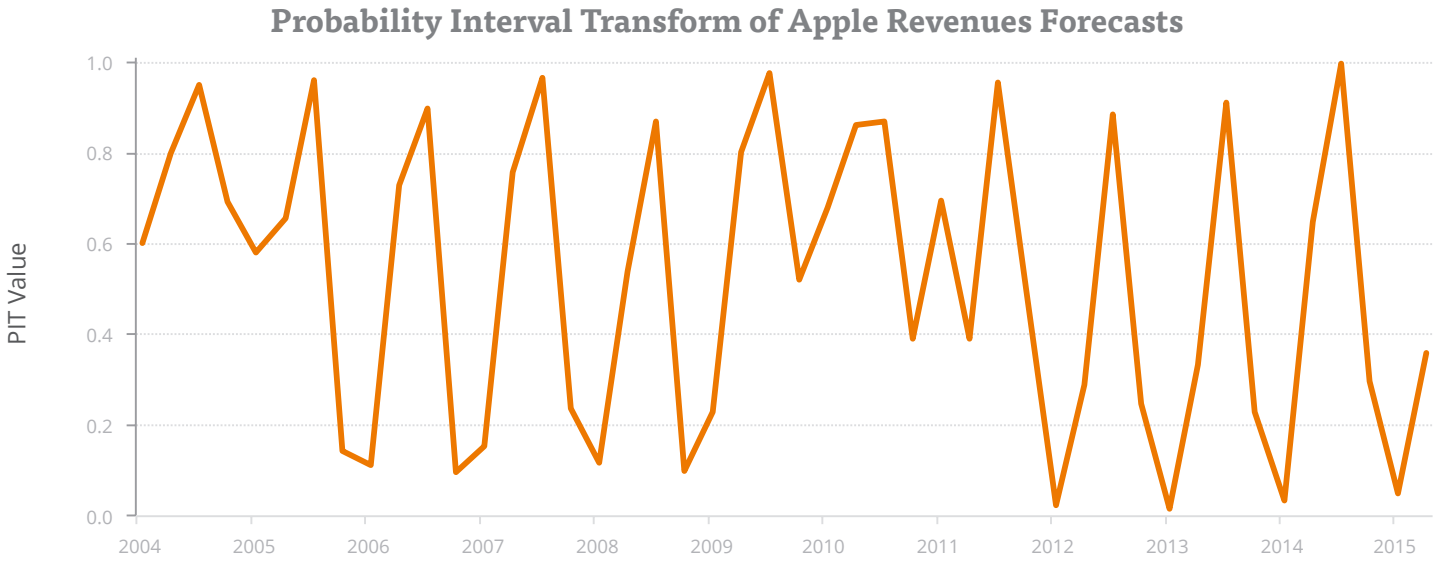

Figure 3 displays values of the probability integral transform ("PIT") for quarterly forecasts of Apple's revenues.

Figure 4 demonstrates the PIT test another way, by plotting a histogram of the PIT values. Recall that the PIT values should be uniformly distributed between zero and one. Thus, using 10 bins, as we do in the graph, we can expect around 4-5 observations in each bin if the model is correctly specified, and consequently the bins to be of similar height. We see, instead, that the number of outcomes in the lowest bin [0 – 0.1], corresponding to low outcomes relative to the forecast, equals 7, which is greater than the expected value of 4 or 5. In addition, outcomes in the two right-most bins, i.e., surprisingly high outcomes, are slightly over-represented, totaling 12 outcomes compared to 9 expected values in these two bins. Conversely, outcomes in the middle of the distribution are underrepresented. This suggests again that the model at the core of this probability forecast is not appropriate for capturing the largest negative and positive values of Apple's quarterly revenue series.

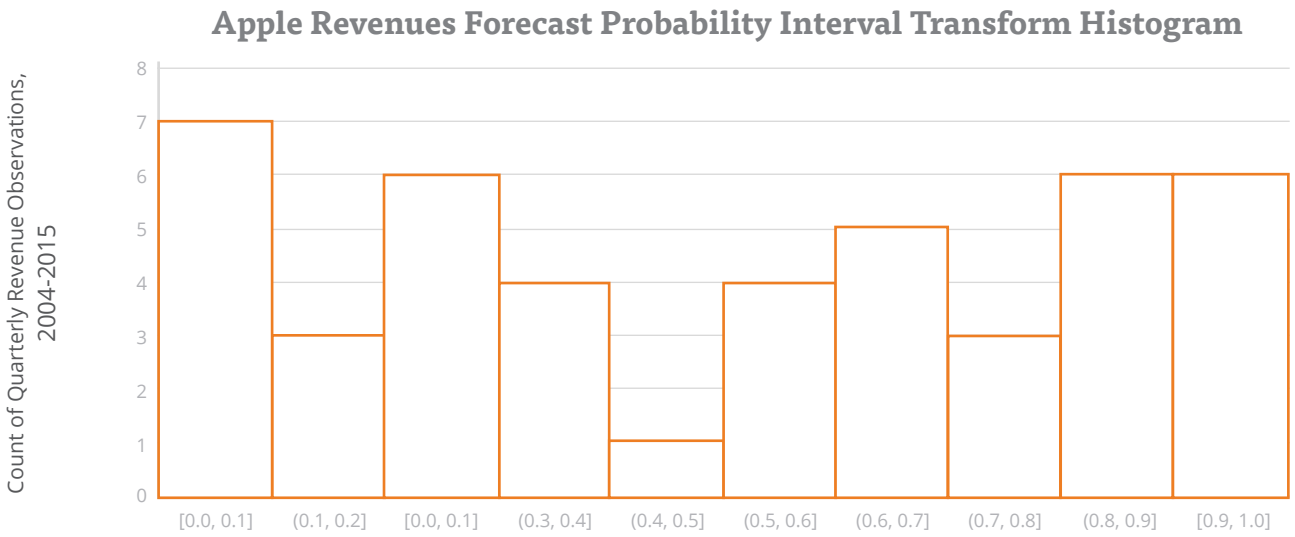

Figure 4 shows a histogram of PIT values for quarterly forecasts of Apple's revenues. The observations should be uniformly distributed with the same number of observations in each bin.

Conditional Tests of Probability Forecasts

The PIT, like the inclusion indicator variable for the interval forecasts, should not be predictable if the probability forecasts are correctly specified. We can test this property by regressing the current value of the PIT on its own past values, or by regressing the PIT on lags of other variables in the forecaster's information set. In all cases, we should expect to find no explanatory power of these variables over next period's PIT value.

Evidence of time-series dependence in the PIT values suggest that the same errors tend to be repeated by the probability forecasts. For example, if we find that the PIT values are serially correlated, so that small PIT values follow other small PIT values, this suggests that surprisingly small (large) outcomes tend to follow previous surprisingly small (large) outcomes. In that case, the forecast may be adjusted downward (upward) or the volatility estimate increased after observing an outcome that is an outlier in the probability distribution, thus reducing the chance of surprises in the same direction.

Summary

Interval and probability forecasts convey information on how the expected outcome changes over time and how much uncertainty surrounds the outcome. If specified correctly, they can also be used to measure the likelihood of unusually high or low outcomes, and how it varies over time. Because they provide additional information that is not contained in point forecasts, it is helpful to ensure that these types of forecasts are correctly specified. This can be tested using the methods described in this note.

All figures described herein are for illustrative purposes only

INTENSITY

About Intensity

Intensity generates excellence through powerful research, analysis, and expertise to solve the most complex challenges in the marketplace and courtroom. We consistently deliver reliable results that are built upon meticulous research, intense scrutiny, and scientific analysis.

Please contact us any time!

12730 High Bluff Drive Suite 300
San Diego, CA 92130
858.876.9101

intensity.com