Standard deviation can seriously damage your wealth.

Synopsis
Standard deviation (SD) is used to set safety stocks, while forecasting is used to estimate lead time usage. The two combine to determine the stock and capacity commitment an organisation must make to achieve its service level goals.
While forecast error is widely understood, standard deviation error is not. Nor, crucially, is the interaction between the two errors. Sometimes a low side forecast will be compensated by a high side estimate of SD, at other times a low side forecast will be made worse by a low side SD estimate (and so on).
In this paper, researcher and consultant Bill Brockbank proposes that

1. The two errors compound more often than they cancel or partly cancel
2. Other methods, though not field tested, might give better results

One such method is proposed while the research continues.

A note on computerised stock algorithms, both explicit and implicit.
A human forecasting system has an inbuilt gatekeeper. The forecaster decides whether to accept or reject changes suggested by their investigations.
By this and other such methods the forecast is made **explicit**.
By contrast, many system forecasts are implicit, and never exposed to critical scrutiny. For example, where the systemised outcome is a suggested Purchase Order the forecast and standard deviation calculations are implicit.
Where the results mislead it's common to blame the forecast.
Maybe the forecast is as good as it can be, the problem is with the SD estimate?

The origins of this research.

When one company's figures show a $7m a year gap between actual and ideal stockholding with no discernable pattern in the product by product differences, the problem *must have multiple causes*. We needed to investigate and challenge every component (human or system) of the calculation. In this case, standard deviation was just one (the second largest) of more than 5 contributory causes. We estimate SD was causing $750,000 of extra costs a year in the UK alone.

The size of this prize was a spur to delve into an area we felt had been under-researched. Indeed the study was in part a legacy of a deep seated unease with both the theoretical SD calculation and its computer implementation.

We therefore set out to answer 2 questions

1. Is past SD a good predictor of future SD?
2. Even if it is a bad predictor, do the forecast and SD errors cancel? In other words, does it matter? Are we often right for the wrong reason – the errors have cancelled or partly cancelled.

Along the way we found just how much SD has been overlooked; it had become the 'forgotten partner' in the whole forecasting for stock arena.

Some light hearted comparisons are helpful for the way they illuminate the difference in mindshare.

Demand forecasting has 2.18 million web references, vs 7,500 for standard deviation. Forecasting has 57 books in Amazon, SD has no books, one paper. Forecasting has an institute (the [American] Institute of Business

(Tel 01279-726806, FAX 01279-726603, Mobile 07802-768078)

Forecasting) which – the last conference I attended – had absolutely nothing on standard deviation. Several professional societies have forecasting SIG's (Special Interest Groups) – ORS (Operation Research Society) is just one. None have SD SIGs

Method.

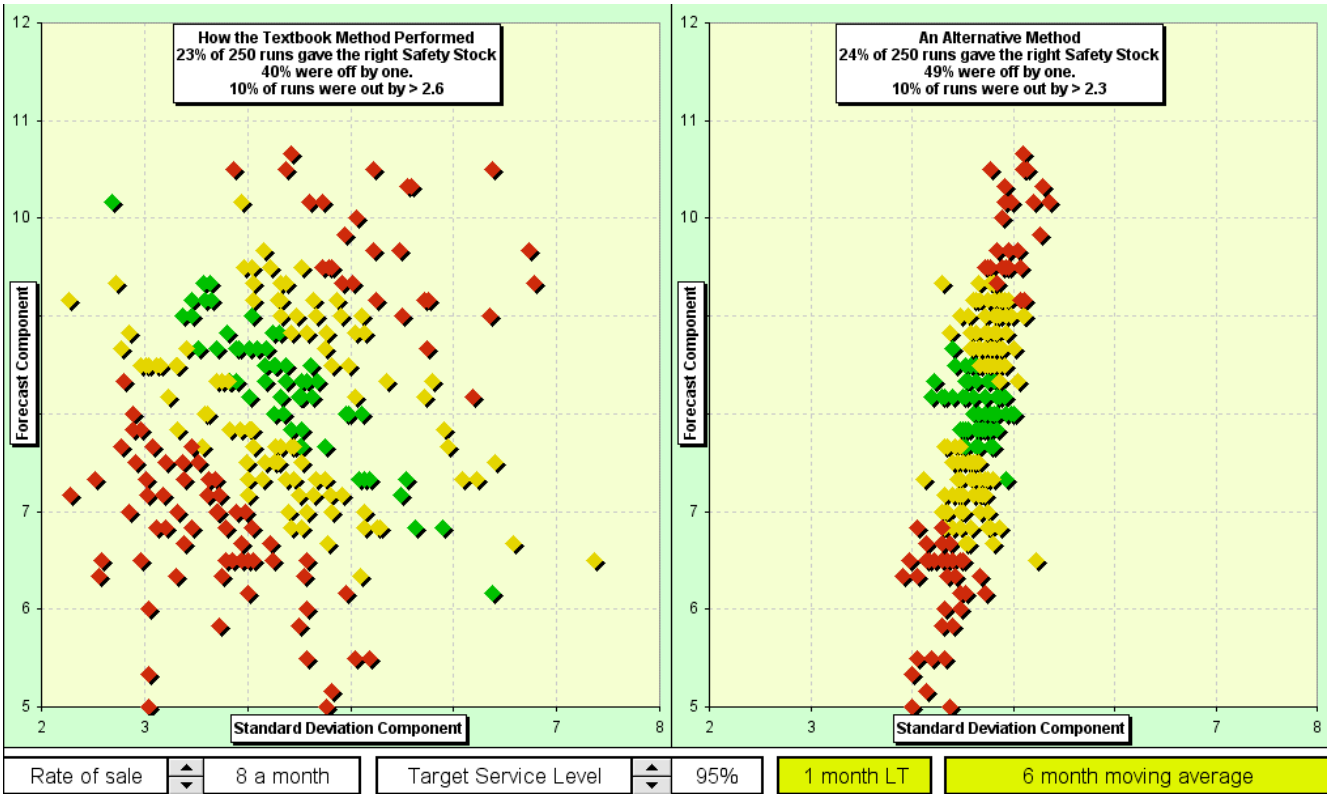
The method simulates a common (though not recommended) computerised safety stock algorithm.

Monthly 'Sales' for a single product are randomly generated about a mean from a Poisson or similar skew distribution. In other words, the sales have the demand variability one would expect of a rational market with no acquisition cost or history of shortage.

The sales populate 12 months in each of 1,000 notionally different years. In fact the underlying demand is the same, each month or year's sales are just samples from the population and therefore vary purely through sampling error.

Using 8 different forecast methods, each year calculates a month ahead forecast, and a historic SD. These are used to calculate a cycle + safety stock target over an array of lead times and service levels, always using the correct transforms. The cycle and safety components are kept separate so we can later determine which component is causing what part of the total error.

Since the 'correct' cycle and safety stock are already known from the base data, we can now compare and classify the computerised predictions with this base.



Results.

Across a wide range of rates of sale, forecast methods, lead times and service levels, standard deviation compounded the forecast error (in the total of cycle + safety stock) more often than it cancelled or partly cancelled it.

(Tel 01279-726806, FAX 01279-726603, Mobile 07802-768078)

Communicating the Results

The intended audience were busy line managers with no grounding in statistics. We needed something punchy and very visual to capture such a wide range of input variables and outcomes.

We set up a competition between the conventional and proposed alternative method (above). To win, one method had to have more 'hits' ***and*** fewer bad misses ***and*** 900 of the 1000 points closer than the other method. In each run, the computer randomly chose and recalculated 10 cases from the ~4,500 possible combinations in the model.

The proposed method won 85% of trials and tied the rest. The textbook method never won.

Other lines of enquiry

We looked at stepped demand, considered building in trend or seasonality, or extending the range of forecast methods.

Stepped demand showed nothing new. Past SD is simply a poor predictor of future SD (colloquially, "SD is a poor predictor of itself") and that error reinforces the forecast error more than it cancels. In the words of one, "if SD were a forecast method we would never use it. Once you isolate SD from all the noise around it, it's simply too bad a forecast of itself to be usable"

In the circumstances, making demand more complex by inducing a step would prove nothing new.

This is ironic. Step changes in demand are a special case, a minority subset of the whole. Yet the only reason we reforecast is because we think there might have been a step change. If there is no change, there's no reason to re-forecast. *"To find a needle in a haystack, don't start by adding hay"*

It seems to us that if the simple case SD (no trend, no seasonality) is unsound then complication will simply increase the decibels on something already too noisy to use.

We think these nuances are red herrings – the SD problem is structural, and the occasional stoke of luck at particular settings should not divert us from the search for a better method.

The search for alternative methods.

In forecasting it's widely accepted that we can predict a family with more accuracy than an individual within the family. This is the principle underlying the insurance and assurance markets, that swings cancel roundabouts. It may be widely understood; it's less widely practiced. Individuals and (especially) computer systems forecast at the SKU (Stock Keeping Unit, i.e. product), and sometimes SKU + shop level *because they can,* not because they should.

"Dynamic response to real time sales data" (a direct quote from a software sales brochure) starts with an assumption that demand changed, and that we should therefore do something about it – in this case change the store stock. In a couple of case studies we've shown how this doubles the shop out-of-stock. A 'change' signalled by a sale or lack of sale (and subsequent reforecast) isn't necessarily a *change in demand*, it may be pure luck.

Grouping products into 'volatility families' seemed a fruitful place to start a search for an alternative to SKU by SKU SD calculation. Can products be grouped by volatility, and the SD for the group be used for each member?

1. Would coefficient of variance (the square of the SD) be a better place to start?
2. Or SD as a percent of the square root of mean? (SD%)

(Tel 01279-726806, FAX 01279-726603, Mobile 07802-768078)

In theory they can (especially the last, for reasons beyond this paper), and there's some empirical evidence that this would work in practice.

For example, we know that spares demand for expensive items follows a predictable skew curve, while cheap spares are more erratic. This reflects a 'one to fit, one for the van' buying pattern for cheap items[1].

For the same rate of consumption, cheap spares might have a SD of 1.8 times the square root of the forecast (SD180%). Elsewhere, we know that companies who have given unreliable ex-stock service experience demand amplification, and there's some clustering of the SKU by SKU SD% for that firm.

Further, different firms experience different amounts of amplification - if a firm has higher than normal amplification on one product it will tend to have similar amplification on all products.

Colloquially, "If demand is erratic, it's universally erratic".

While this risks confusing cause and effect (poor customer service leads to demand amplification) there's every reason to believe that, behind it all, "numbers are blind". Once we manage out the causes of amplification, replacement windscreen wipers might belong to the low volatility family, while Queen Mother commemorative postage stamps would not.

There's more work to be done to unpick real data, to persuade firms to try the new methods and to persuade software companies to change their algorithms.

'Family SD%' gave consistently better results than individual SD; those are the results reported above.

One of our conclusions is provisional until we have more live trials. We need to prove that we can classify products into 'volatility index families', and that the families (and associated expulsion/adoption processes to reclassify products from one family to another) can work in practice.

Conclusion

Across a wide range of settings, SKU by SKU SD compounds the forecast error more often than it cancels or partly cancels. Using SKU by SKU SD we get unlucky more often than we get lucky, and the amount by which we are unlucky is greater.

Hypothesis

Until there's conclusive proof that a product has changed family, an individual SKU's SD is proportional to the square root of its forecast.

The proportion – or multiplier - is

1. A property of the SKU's 'volatility family', not the individual SKU
2. Almost always greater than one. If the proportion is less than one then demand is being 'gated' by the user (e.g. a car plant with fixed speed production line; hospital operating theatre supplies) In this case collaboration & visibility are better than forecasting.

Where this might lead.

Even though the research is unproven in the field, it's helpful to look at what else might be affected. This list is offered as a start.

1. By basing SD on the forecast we replace two SKU level variables with one. This should be easier to manage. Given the minefield we found during this journey, and the way SD had been overlooked as a contributory cause, anything which simplifies the calculation and its management and control would be a win *even if the methods had tied*.

[1] And maybe one for luck …

2. The rate of product 'churn' is increasing. Many consumer products are obsolete in a year. There simply isn't time to build enough history to calculate a meaningful SD. Initialising the SD% at the product family's average means we can do that part of the calculation without any history.
3. It would be an interesting side study to find how those forecasting systems which do use SD cope with new and superceding products. In this context, the forecasting system is the whole – human, machine and management control. We are not optimistic!
4. The notion of a 'gatekeeper' in any forecasting system is a useful one; in this context the SD calculation is just another forecasting system. The gatekeeper looks at exceptions and recommendations for change suggested by the computer, and applies a human judgement on whether to accept a change or not. A good gatekeeper will accept some, reject some and 'wait and see' on the remainder. This is an excellent balance of human and machine capability. The categorisation of products into volatility families suggests a gatekeeper should manage the subsequent promotion / relegation processes. The computer can still do all the donkeywork, the human applies a veto. I'm not aware of any scientific study with and without gatekeepers, but this approach and the balance of human vs. machine intelligence has always felt right to me. I've used a gatekeeper ("riding shotgun" on a focus forecasting system) with great success.
5. Making it hard to change a SD (by making the SD forecast explicit, then 'gatekeeping' changes) will induce much needed stability where now there is too much noise. Study after study shows the re-forecast to be less accurate than the original; "we made a few right at the expense of making a lot wrong". Common sense suggests that a history based forecast (be that sales or SD) should be done on sufficient data, therefore a re-forecast should only be done when there's sufficient ***new*** data. "It stays the same until we're sure it's different" should be the watchword.
6. The categorisation into volatility families might be easier then we think. There is a precedent. Focus forecasting systems require (for example) the gatekeeper to judge if an alpha (exponential smoothing factor) of 0.3 – medium responsiveness - is appropriate for a product previously set at 0.1 - low responsiveness. Pretty soon the gatekeeper develops a 'feel'. In Paul Freeman's words (when running Kellogg's focus forecasting system) "All Bran doesn't suddenly behave like Frosties, no matter how good the line of best fit might appear". Asked to rank different product types (disposable operating theatre scissors, car oil filters, flu remedies, ice cream, socks, petrol, high street shop ties and airport shop ties) into volatility sequence we saw a high degree of consensus.

(Tel 01279-726806, FAX 01279-726603, Mobile 07802-768078)