The Supply Chain Approach to Planning and Procurement Management

The supply chain approach models stochastic events influencing a manufacturing organization's shipment and inventory performance in the same way that a mechanical engineer models tolerance buildup in a new product design. The objectives are to minimize on-hand inventory and optimize supplier response times.

by Gregory A. Kruger

This paper describes the processes and equations behind a reengineering effort begun in 1995 in the planning and procurement organizations of the Hewlett-Packard Colorado Springs Division. The project was known as the *supply chain* project. Its objectives were to provide the planning and procurement organizations with a methodology for setting the best possible plans, procuring the appropriate amount of material to support those plans, and making up-front business decisions on the costs of inventory versus supplier response time (SRT),* service level to SRT objectives, future demand uncertainty, part lead times, and part delivery uncertainty. The statistical modeling assumptions, equations, and equation derivations are documented here.

Basic Situation

Consider a factory building some arbitrary product to meet anticipated customer demand. Since future demand is always an uncertainty, planning and procurement must wrestle with the task of setting plans at the right level and procuring the appropriate material. The organization strives to run the factory between two equally unattractive scenarios: not enough inventory and long SRTs, or excessive inventory but meeting SRT goals. In fact, more than one organization has found itself with the worst of both worlds—huge inventories and poor SRTs.

The supply chain project focused on characterizing the various stochastic events influencing a manufacturing organization's shipment and inventory performance, modeling them analogously to the way a mechanical engineer would model a tolerance buildup in a new product design.

Problem Formulation

For a particular product, a factory will incur some actual demand each week, that is, it will incur demand D_i in week i, for i = 1, 2, 3, ... From a planning and procurement perspective, the problem is that looking into the future the D_i are unknown.

Let P_i be the plan (or forecast) for week i in the future. Now for each week, the actual demand can be expressed as the planned demand plus some error: $D_i = P_i + e_i$.

The MRP (material requirements planning) system, running at intervals of R weeks, evaluates whether to order more material to cover anticipated demand, and if the decision is to order, how much to order. Given a lead time of L weeks to take delivery of an order placed to a supplier now for some part, the material in the supply pipeline must cover real demand for the next L+R weeks. By supply pipeline we mean the quantity of the part already on hand at the factory plus the quantity in orders placed to the supplier and due for delivery over the next L weeks.

For simplicity, assume for the remainder of this discussion that we are dealing with a part unique to one product and used only once in building the product. We will remove these constraints later but for now it will help to focus on the key concepts.

Define X to be the unknown but actual demand the factory will experience for this part over the next L+R weeks:

$$X = \sum_{i=1}^{L+R} D_i = \sum_{i=1}^{L+R} \left(P_i + e_i\right).$$

In statistical terminology, X is a random variable, that is, we cannot say with certainty the value it will take next, but with some assumptions about the nature of the planning errors (e_i), the distribution of X can be characterized. Specifically, we will make the assumption that the e_i are distributed according to the Gaussian (normal) distribution with mean zero and

* In standard terminology, SRT stands for "supplier response time." In this case, a better term would be "shipment response time," because the supplier being referred to is HP and not one of HP's suppliers. In this paper, we use the standard terminology for SRT, but the word "supplier" in all other contexts means one of HP's suppliers.

variance σ^2 (see Fig. 1). The assumption that the mean of the e_i is zero says that our plans are unbiased, that is, the factory is not consistently overestimating or underestimating future demand. Thus, the average of the differences between the plan and the actual demand over a reasonable period of time would be about zero. The normal distribution is symmetric, so we are saying there is equal probability in any week of actual demand being above or below plan. The variance measures how large the planning errors can get in either direction from zero.

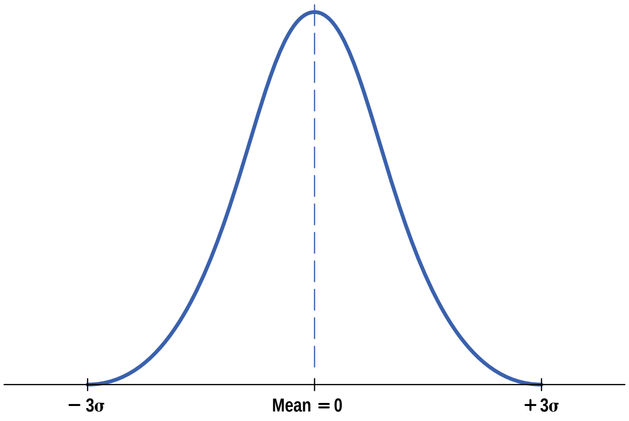

Fig. 1. *Assumed normal distribution of planning errors.*

We would like to know both the expected value of X and its variance. Knowing these two values will form the basis for the ultimate decision rules for replenishment order sizes placed to the supplier for our part.

We will use the following notation: E(x) represents the expected value of the random variable x, and V(x) represents the variance of the random variable x.

Before launching into the derivation of the expected value of the real demand over the next L + R weeks, note that L itself is a random variable. When an order is placed with the supplier, delivery does not always come exactly on the acknowledgment date. There is some uncertainty associated with when the replenishment order will arrive. Like the planning errors, we will assume that the delivery errors are normally distributed about zero. Thus:

$$E(X) = E\left(\sum_{i=1}^{L+R}\left(P_i + e_i\right)\right) \cong \sum_{i=1}^{E(L)+R} E\left(P_i + e_i\right)$$

$$= \sum_{i=1}^{\mu_L+R}\left(E\left(P_i\right) + E\left(e_i\right)\right) = \sum_{i=1}^{\mu_L+R} P_i.$$

The result will be precisely correct when the P_i are stationary (that is, the plan is a constant run rate) and will serve as an approximation when the P_i are nonstationary.

Determining the variance of X is more involved because the limit of the summation, L + R, is a random variable. The derivation can be be found in Appendix I. The result is:

$$V(X) \cong \left(\mu_L + R\right)\sigma_e^2 + \overline{P}_{L+R}^2\sigma_L^2.$$

where σ_e is the standard deviation of the errors e_i, σ_L is the standard deviation of L, and $\overline{P}_{L+R}$ is the average of the plan over L + R weeks.

The standard deviation of demand is the square root of this result. In practice, we estimate the standard deviation of demand by:

$$\hat{\sigma}_X \cong \sqrt{\left(\overline{L} + R\right)s_{DE}^2 + \overline{P}_{L+R}^2 s_{LE}^2},$$

where $\overline{L}$ is the average lead time from the supplier of this part, R is the review period, s_{DE}^2 is the variance of the difference between the weekly plan and the actual weekly demand, and s_{LE}^2 is the variance of the difference between the date requested and the date received. Lee, Billington, and Carter[1] give the same result when modeling the demand at a distribution center within a supply chain.

Knowing the variance of the demand uncertainty over L + R weeks, we can develop a decision rule for determining the amount of inventory to carry to meet the actual demand the desired percent of the time.

We define the *order-up-to level* as:

$$\text{Order-up-to Level} = \sum_{i=1}^{L+R} P_i + Z_{1-\alpha}\hat{\sigma}_X,$$

where $Z_{1-\alpha}$is the standard normal value corresponding to a probability α of stocking out. $Z_{1-\alpha}\hat{\sigma}_X$ is called the safety stock.

We define the *inventory position* as follows:

$$\begin{aligned}\text{Inventory Position} = {} & \text{On-Hand Quantity} \\ & + \text{On-Order Quantity} \\ & - \text{Back-Ordered Quantity.}\end{aligned}$$

The purchase order size decision rule each R weeks for replenishment of this part becomes:

$$\begin{aligned}\text{New Order Quantity} = {} & \text{Order-up-to Level} \\ & - \text{Inventory Position.}\end{aligned}$$

We are simply trying to keep the order-up-to level of material in the supply pipeline over the next $L+R$ weeks, knowing we have a probability α of stocking out.

As you can see, the basic idea behind the statistical calculation of safety stock is straightforward. In practice, a number of complicating factors must be accounted for before we can make this technology operational. The list of issues includes:

- The chosen frame of reference for defining and measuring future demand uncertainty
- The impact of SRT objectives on inventory requirements
- The translation from part service level to finished product service level
- Appropriate estimates for demand and supply uncertainty upon which to base the safety stock calculations
- Purchasing constraints when buying from suppliers
- The hidden effect of review period on service level performance
- The definition of service level.

There are significant business outcomes from managing inventory with the statistical calculation of safety stock. These include the ability to:

- Predict average on-hand inventory and the range over which physical inventory can be expected to vary
- Trade off service level and inventory
- Trade off SRT and inventory
- Plot order aging curves so that you can see how long customers may have to wait when stock-outs do occur
- Measure the impact of reducing lead times, forecasting error, and delivery uncertainty
- Measure the impact of changing review periods and minimum order quantities to the supplier
- Stabilize the orders placed to suppliers so that they are not being subjected to undue uncertainties
- Reduce procurement overhead required for manipulating orders.

Turning off the Production Plan Overdrive

Many manufacturing planning organizations have traditionally handled the uncertainties of future demand by intentionally putting a near-term overdrive into the production plan (see Fig. 2). By driving the material requirements plan (MRP) higher than expected orders, a buffer of additional material is brought into the factory to guard against the inevitable differences between forecast and actual demand. In effect, this overdrive, or front loading, functions as safety stock, although it is never called that by the materials system.

While this practice has helped many factories meet shipment demands, it has also caused frustrations with nonoptimal inventory levels. Biasing the build plan high across all products does not consider that it is unlikely that all of the products will be simultaneously above their respective forecasts. Therefore, inventories on parts common to several products tend to be excessive. Also, this approach treats all parts the same regardless of part lead times, rather than allocating safety stock inventory based upon each part's procurement lead time. The factory can easily end up with inventories too high on short lead time parts and too low on longer lead time parts. Finally, the practice of building a front-end overdrive into the plan can lead to conflict between the procurement and production planning departments. Wanting to ensure sufficient material to meet customer demand, the planning department's natural desire is to add a comfortable pad to the production plan. Procurement, aware of the built-in overdrive in the plan and under pressure to reduce inventories, may elect to second-guess the MRP system and order fewer parts than suggested. Should planning become aware that the intended safety pad is not really there, it can lead to an escalating battle between the two organizations.

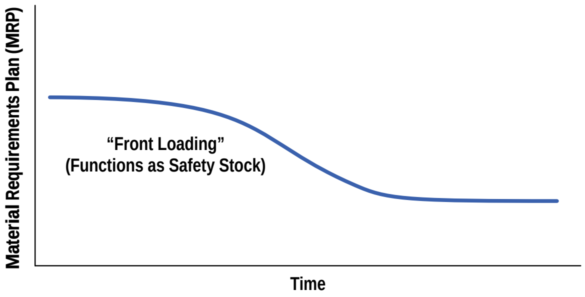

Fig. 2. *Many manufacturing planning organizations handle the uncertainties of future demand by intentionally driving the material requirements plan (MRP) higher than expected orders.*

Frame of Reference

Fundamental to the use of the statistical safety stock methods outlined in this paper is how one chooses to measure demand uncertainty, or in other words, what is the point of reference. The two alternative views are (see Fig. 3):

- Demand uncertainty is the difference between part consumption in the factory and planned consumption.
- Demand uncertainty is the difference between real-time customer demand and the forecast.

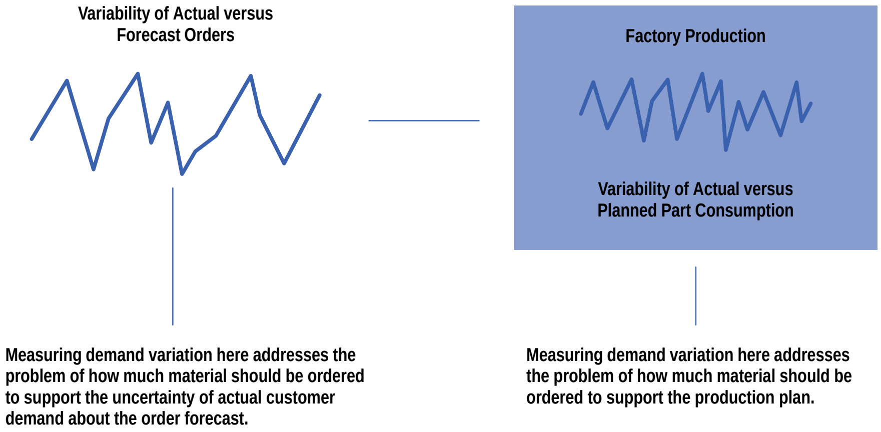
Variability of Actual versus
Forecast Orders
Factory Production
Variability of Actual versus
Planned Part Consumption
Measuring demand variation here addresses the problem of how much material should be ordered to support the uncertainty of actual customer demand about the order forecast.
Measuring demand variation here addresses the problem of how much material should be ordered to support the production plan.

Fig. 3. *Frames of reference for measuring demand uncertainty. These two measures can be very different in a factory dedicated to steady build rates according to a build plan. In a factory fluctuating its production is response to actual orders, these two measures are more alike.*

Consider using part consumption within the factory versus build plan as the frame of reference. The function of statistical safety stocks here is to provide confidence that material is available to support the production plan. A factory with a steady-rate build plan would carry relatively little safety stock because there are only small fluctuations in actual part consumption. Of course, actual order fulfillment performance would depend upon finished goods inventory and the appropriateness of the plan. In this environment, the organization's SRT objective has no direct bearing on the safety stock calculations. The factors influencing the estimate of demand uncertainty and hence safety stock are fluctuations in actual builds from the planned build, part yield loss, and part use for reasons other than production.

If the point of reference calls for measuring demand uncertainty as the deviation between the forecast and real-time incoming customer orders, safety stock becomes a tool to provide sufficient material to meet customer demand. This factory is not running steady-state production but rather building what is required. Now the SRT objective should be included in the safety stock calculations since production does not have to build exactly to real-time demand if the SRT objective is not zero. From this perspective, statistical safety stocks, projected on-hand inventory, SRT, and service levels are all tied together, giving a picture of the investments necessary to handle marketplace uncertainty and still achieve order fulfillment goals.

In choosing between these two frames of reference for the definition of demand uncertainty it comes down to an analysis of factory complexity and timing. If factory cycle times are relatively short so that production is not far removed from customer orders, then demand uncertainty can be measured as real-time orders versus forecast. However, if factory cycle times are long so that production timing is well-removed from incoming orders, then demand uncertainty would best be measured as part consumption versus build plan.

SRT in Safety Stock Calculations

Appendix IV documents the mathematics for incorporating SRT objectives into the safety stock calculations. As has been discussed, using the SRT mathematics would be appropriate when measuring demand uncertainty as deviations of real-time customer orders from forecast. It is critical, however, that we understand how production cycle times affect the factory's actual SRT performance.

As stated in Appendix IV, if factory cycle time is considered to be zero, the SRT mathematics ensures that material sufficient to match customer orders will arrive no later than the desired number of weeks after the customer's order. Clearly, time must be allocated to allow the factory to build and test the completed product. In this paper, this production time is not the cycle time for building one unit but for building a week's worth of demand.

Care must be taken when using the SRT mathematics. Consider that the practice of booking customer orders inside the SRT window will place demands on material earlier than expected from the mathematical model given in Appendix IV. In practice, one should be conservative and use perhaps no more than half of the stated SRT as input to the safety stock model.

Part versus Product Service Level

The statistical mathematics behind the safety stock calculations are actually ensuring a service level for parts availability and not for completed product availability. This is true regardless of whether the chosen frame of reference for measuring demand uncertainty is part-level consumption or product-level orders. Since production needs a complete set of parts to build the product the question arises as to what the appropriate part service level should be to support the organization's product service level goals. Unfortunately, there is not a simple algebraic solution to this problem.

The exact answer is subject to the interdependencies among the probabilities of stocking out of any of the individual parts in the bill of materials. If we assume that the probabilities of stocking out of different parts are statistically independent, then the situation looks bleak indeed. For example, if we have a 99% chance of having each of 100 parts needed to build a finished product, independence would suggest only a $0.99^{100} = 36.6\%$ chance of having all the parts. Clearly the chance of stocking out of one part is not totally independent of stocking out of another. For example, if customer demand is below plan there is less chance of stocking out of any of the parts required. Just as clearly, there is not total dependence among parts. One supplier may be late on delivery, causing a stock-out on one part number while there are adequate supplies of other parts on the bill of materials. In the example mentioned, the truth about product service level lies between the two extremes, that is, somewhere between 0.99^{100} and 0.99.

As an operational rule of thumb, individual part service levels should be kept at 99% or greater. Of course, the procurement organization may choose to run inexpensive parts at a 99.9% or even higher service level so as never to run out. Then the service level on expensive parts can be lowered such that the factory gets the highest return on its inventory dollar. For example, a factory may run a critical, expensive part at a 95% service level while maintaining a 99.9% service level on cheaper components to achieve a product level goal of a 95% service level to the SRT objective.

Parts Common to Multiple Products

In the problem formulation section it was assumed that we were dealing with a part unique to a single product and used only once to build that product. First, recognize that the situation in which a part is unique to a single product, but happens to be used more than once to build the product, is trivial. If the product uses a part k times then the forecasted part demand is simply k times the forecast for the product. Similarly, the standard deviation of the forecast error for the part is simply k times the standard deviation of the forecast error for the product.

The more interesting situation arises when a part is common to multiple products. We will look at two alternative approaches to handling common parts, the second method being superior to the first. In the first approach, we will assume that the forecasting errors for the products using the common part are independent of one another. Since the total forecasting error for the part can be written as the sum of the forecasting errors for each of the products using the part, the standard deviation of the part forecasting uncertainty can be easily determined.

Consider a part used in j products and used k_i times in product i, where $i = 1, 2, ..., j$. Let DE represent the forecasting or demand error. Then:

$$DE_{part} = k_1 DE_{product1} + k_2 DE_{product2} + k_3 DE_{product3}$$

$$+ \dots + k_j DE_{productj}$$

$$\sigma^2_{DEpart} = k_1^2 \sigma^2_{DEproduct1} + k_2^2 \sigma^2_{DEproduct2}$$

$$+ k_3^2 \sigma^2_{DEproduct3} + \dots + k_j^2 \sigma^2_{DEproductj}.$$

The big problem with this approach is the assumption of independence of forecasting errors among all the products using the part. If, for example, when one product is over its forecast there is a tendency for one or more of the others to be over their forecasts, the variance calculated as given here will underestimate the true variability in part demand uncertainty.

The second approach to estimating forecasting uncertainty for common parts is to explode product-level forecasts into part-level forecasts and product-level customer demand into part-level demand and measure the demand uncertainty directly at the part level. For a part common to j products we simply measure the forecast error once as the difference between the part forecast and actual part demand instead of measuring the forecast errors for the individual products and algebraically combining them as before. Any covariances between product forecasting uncertainties will be picked up in the direct measurement of the part-level forecasting errors. Clearly, this is the preferred approach to estimating part demand uncertainty, since it avoids making the assumption of forecast error independence among products using the part.

Estimation of Demand and Part Delivery Uncertainty

The whole approach to safety stocks and inventory management outlined here is dependent upon the basic premise behind any statistical sampling theory—namely, that future events can be modeled by a sample of past events. Future demand uncertainty is assumed to behave like past demand uncertainty. Future delivery uncertainty is assumed to behave like the supplier's historical track record. This raises two issues when estimating the critical inputs to the safety stock equations: robust estimation and business judgment. Both of these issues are extremely dependent upon the chosen frame of reference, that is, whether we are measuring real-time customer demand or part-level consumption on the factory floor.

From a sample size perspective we would like to have as much data as possible to estimate both demand and delivery uncertainty. However, in a rapidly changing business climate we may distrust data older than, say, six months or so. If I am measuring demand uncertainty as the deviations between real-time customer orders and the forecast, do I want to filter certain events so they do not influence the standard deviation of demand uncertainty and hence safety stocks? It may be good business practice not to allow big deals to inflate the standard deviation of demand uncertainty if those customers are willing to negotiate SRT. In statistical jargon, we want our estimates going into the safety stock equation to be robust to outliers. Naturally, if the demand uncertainty is measured as part consumption on the factory floor versus planned consumption, data filtering is not an issue. It is possible that an unusual event affecting parts delivery from a supplier may be best filtered from the data so that the factory is not holding inventory to guard against supply variability that is artificially inflated.

A common situation is the introduction of a new product. Suppose the chosen point of reference is measuring demand uncertainty as real-time customer orders versus forecast. How do we manage a new product introduction? A viable option is to use collective business judgment to set the demand uncertainty even though there is technically a sample size of zero before introduction. Prior product introductions or a stated business objective of being able to handle demand falling within $\pm\Delta$ of the plan during the early sales months can be used to establish safety stocks. In fact, the organization can compare the inventory costs associated with different assumptions about the nature of the demand volatility. Estimates of average inventory investment versus assumed demand uncertainty obtained from the statistical models can help the business team select an introduction strategy.

Effect of Minimum Buy Quantities and Desired Delivery Intervals

In most cases, there are constraints on the order sizes we place to our suppliers, such that replenishment orders are not exactly the difference between the theoretical order-up-to level and the inventory position. These constraints may be driven by the supplier in the form of minimum buy quantities or ourselves in the form of economic order quantities or desired delivery frequencies. The net effect of all such constraints on order sizes is to reduce the periods of exposure to stock-outs.

For example, suppose the factory's plans predict needing 100 units of some part per week. Further suppose that the ordering constraint is that we order 1000 units at a time determined by either the supplier's minimum or our economic order quantity. This order quantity represents ten weeks of anticipated demand. Once the shipment of parts arrives from the supplier, there is virtually no chance of stocking out for several weeks until just before the arrival of the next shipment. Given this observation we see that safety stock requirements actually decrease as purchase quantity constraints increase (see **Appendix V**).

Although safety stocks decrease, average on-hand inventory and the standard deviation of on-hand inventory both increase. See **Appendix III** for formula derivations of the average and the standard deviation of on-hand inventory.

Effect of Review Period

Analysis of the equation for the standard deviation of demand uncertainty given above shows that as the review period R increases, σ_X increases, thereby driving up safety stock. This makes sense because the safety stock is there to provide the desired confidence of making it through R weeks without a stock-out. However, note that the service level metric itself is changing. For $R = 1$, the service level gives the probability of making it through each week without a stock-out. For $R = 2$, the service level gives the probability of making it through two weeks, for $R = 3$, three weeks, and so on. Increasing review period therefore has an effect similar to that of minimum buy quantities. When operating at longer review periods, purchase quantities to the supplier are larger, since we are procuring to cover R weeks of future demand and not just one week of future demand. To keep the average weekly service level at the desired goal, safety stock would actually have to be throttled back as the review period increases because of less frequent periods of exposure.

Service Level Metric

Throughout this paper, service level has been defined as the probability of not stocking out over a period of time, usually on a weekly basis. There is another commonly used service level metric called the *line item fill rate* (LIFR). With the LIFR the issue is not whether stock-outs occur but rather whether there is at least the desired percentage of the required items available. For example, suppose in a week of factory production, demand for a part is 100 units but there are only 95 available. Measured in terms of LIFR, the service level is 95%.

Proponents of LIFR argue that the metric gives appropriate credit for having at least some of what is required, whereas the probability of stock-out metric counts a week in which there was 95% of the required quantity of a particular part as a stock-out.

When calculating safety stocks to a LIFR metric rather than multiplying the standard deviation of demand over the lead time plus the review period by a standard normal value, solve for k in the following approximation formula:[2]

$$\mathrm{LIFR_{goal}} = 1 - \frac{\sigma_X}{\mu_D} e^{\left(-0.92 - 1.19k - 0.37k^2\right)}.$$

where μ_D is the average weekly demand. Then the safety stock is $k\sigma_X$.

Inventory versus Service Level Exchange Curves

A useful graphical output from the statistical inventory mathematics is the inventory versus service level exchange curve as shown in Fig. 4.

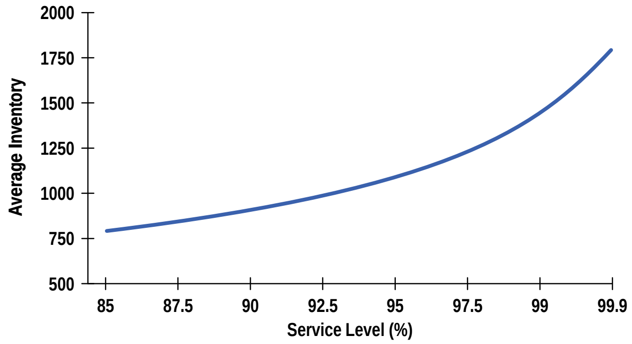

Fig. 4. *Average inventory as a function of service level.*

Such graphs demonstrate the nonlinear relationship between increasing inventory and service level given the constraints on the factory. The curve represents the operating objective. (Johnson and Davis[3] refer to this curve as the "efficient frontier.") By comparing historical inventory and service levels to the performance levels possible as indicated in Fig. 4, a factory can gauge how much room it has for improvement. In addition, procurement can determine where on the curve they should be operating based upon their cost for expediting orders. As can be seen in Fig. 4, a factory operating in the 90% service level range would get a lot of leverage from inventory money invested to move them to 95% service. However, moving from 95% to 99% service level requires more money and moving from 99% to 99.9% requires more yet. By comparing the cost (and success rate) of expediting parts to avoid stock-outs with the cost of holding inventory, the organization can determine the most cost-effective operating point.

Order Aging Curves

Another useful graphical output is the order aging curve. This curve in a sense tells the rest of the story about material availability to meet the SRT and service level objectives. More specifically, the curve demonstrates what type of service can be expected for SRTs shorter than the objective and how long customers can be expected to wait when you are unable to meet your SRT objective. Fig. 5 shows a family of order aging curves, each corresponding to a certain safety stock value determined by the stated SRT goal. We see, for example, that a factory holding safety stocks to support a 99% service level on a two-week SRT goal could, in fact, support a one-week SRT with a service level better than 90%. That same factory will almost surely have all orders filled no later than four weeks from receipt of customer order.

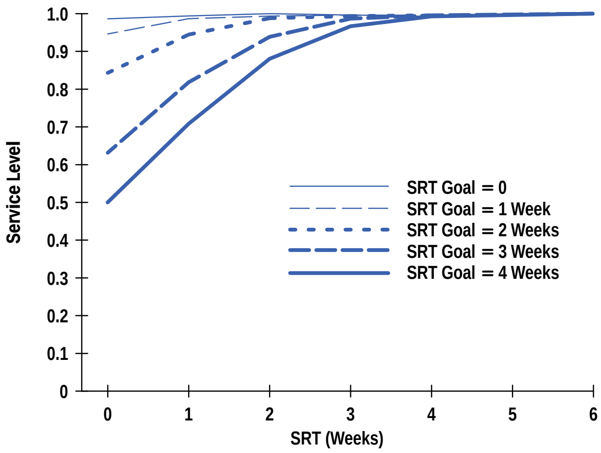

Fig. 5. *Order aging curves for differing SRT (supplier response time) goals.*

Theory versus Practice

Ultimately, the actual performance the factory experiences in the key metrics of service level to the SRT objectives and average on-hand inventory will depend upon whether the supply chain performs according to the inputs provided to the statistical model. All of the estimates are predicated upon the future supply chain parameters fluctuating within the estimated boundaries. As depicted in Fig. 6, we have built up a set of assumptions about the nature of the various uncertainties within our supply chain. If one or more of these building blocks proves to be inaccurate, the factory will realize neither the service level nor the inventory projected.

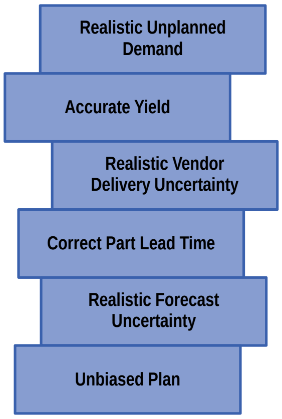
Realistic Unplanned Demand
Accurate Yield
Realistic Vendor Delivery Uncertainty
Correct Part Lead Time
Realistic Forecast Uncertainty
Unbiased Plan

Fig. 6. *Supply chain inputs. The accuracy of the estimates of service level and on-hand inventory are dependent on the validity of the inputs.*

Acknowledgments

Special thanks to Rob Hall of the HP strategic planning and modeling group and Greg Larsen of the Loveland Manufcturing Center for assistance in the development of this theory and helpful suggestions for this paper. Not only thanks but also congratulations to the process engineering, planning, and procurement organizations of the Colorado Springs Division for reengineering division processes to put supply chain theory into practice.

References

1. H.L. Lee, C. Billington, and B. Carter, "Hewlett-Packard Gains Control of Inventory and Service through Design for Localization," *Interfaces*, Vol. 23, no. 4, July-August 1993, p. 10.
2. S. Nahmias, *Production and Operations Analysis*, Richard Irwin, 1989, p. 653.
3. M.E. Johnson and T. Davis, *Improving Supply Chain Performance Using Order Fulfillment Metrics*, Hewlett-Packard Strategic Planning and Modeling Group Technical Document (Internal Use Only), 1995, p. 14.

- ▶ Go to Appendix I
- ▶ Go to Appendix II
- ▶ Go to Appendix III
- ▶ Go to Appendix IV
- ▶ Go to Appendix V
- ▶ Go to Appendix VI
- ▶ Go to Appendix VII
- ▶ Go to Next Article
- ▶ Go to Journal Home Page