Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Fill Rates of Single-Stage and Multistage Supply Systems

Matthew J. Sobel,

MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 6, No. 1, Winter 2004, pp. 41–52
ISSN 1523-4614 | EISSN 1526-5498 | 04 | 0601 | 0041

DOI 10.1287/msom.1030.0027

Fill Rates of Single-Stage and Multistage Supply Systems

Matthew J. Sobel
Department of Operations, Weatherhead School of Management, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106-7235, matthew.sobel@case.edu

A supply system's fill rate is the fraction of demand that is met from on-hand inventory. This paper presents formulas for the fill rate of periodic review supply systems that use base-stock-level policies. The first part of the paper contains fill-rate formulas for a single-stage system and general distributions of demand. When demand is normally distributed, an exact expression uses only the standard normal distribution and density functions, and a good approximation uses only the standard normal distribution function. The second part of the paper derives the probability distribution of the finished goods inventory level for serial systems with buffer inventories between stages. This distribution leads to fill-rate formulas and the conclusion that shorter supply chains have higher fill rates.

Key words: fill rate; service level; supply system; base-stock-level policy; multistage; supply chain; logistics
History: Received: June 12, 2002; days with author: 156; revisions: 3; average review cycle time: 84 days; Senior Editor: Morris Cohen.

1. Introduction

The *fill rate* is the long-run average fraction of demand that a supply system can satisfy immediately. The research literature concerning fill rates concentrates on single-stage supply systems, although many demand processes are met from inventories at the end of multistage systems. The ways in which multiple stages can arise range from contractually distinct entities in a supply chain to multiple stages of production in a single factory.

The practical importance of the fill rate in multiple-stage systems is reflected in many articles in the trade press. For example, "Since MasterLink went live in 1995, Master Lock's *fill rate*...has zoomed from 78% to 93%" (Hoffman 1996); "Some retailers have been disappointed with the home video category because the industry's *fill rate* is about 75%, compared with about 90% for packaged goods such as toothpaste and detergent,..." (Orenstein 1999); "Three years ago, Hans Wede was brought in as president of General Marble Inc. in a last-ditch effort to save the company. Today, General Marble has a 100% *fill rate* with a 2-day lead time" (Plantz 1997); "[In order to choose the number of warehouses] He claims you need to know the...order *fill rate* required. You need to know what is more important to the customer—*fill rate*, speed, or both" (Richardson 1998); "Computer maker AST Research Inc. has launched an expensive initiative to improve customer service that is cutting help-desk response times and increasing the *fill rate* on spare-part shipments..." (Sweeney 1996).

Why study the fill rate in a periodic review model? It is more difficult to characterize and compute than in a continuous review model with a Poisson demand process (Zipkin 2000). Also, the widespread implementation of transaction processing information systems makes continuous review practical. However, in many firms, replenishment goods are shipped only periodically. Therefore, the shipment dates determine the points in time at which orders are placed.

The supply system in this paper has multiple stages arranged in series with buffer inventories between stages. The model is discrete in time; each period, quantities of items are released from the buffer inventories for further processing downstream. The system is *capacitated* if there are bounds on the quantities that can be processed during a period. Otherwise, it is *uncapacitated*. The released quantities are selected with a base-stock-level policy and end-item demands are independent and identically distributed random variables (i.i.d.r.v.s).

Section 2 reviews related research, and the model is detailed in §3. Sections 4 and 5 concern uncapacitated single-stage systems. Section 4 has new fill-rate formulas for general demand distributions. When demand is normally distributed, the new expressions yield an exact formula in §5 that can be calculated using only the standard normal distribution and density functions. The derivation of the exact formula suggests an approximation that uses only the standard normal distribution function and leads to a simple expression for the base-stock level as a function of the fill rate.

The remainder of the paper concerns multistage systems. Section 6 presents properties of echelon variables that are induced by base-stock-level policies. Section 7 obtains the probability distribution of the finished goods inventory level in multistage systems and culminates in a simple expression for the fill rate (Theorem 2). Also, a formal result confirms the intuition that shorter supply chains have higher fill rates (Corollary 1). Section 8 presents bounds on the fill rate and §9 has several numerical examples.

The primary results in the paper are succinct fill-rate formulas (Theorems 1 and 2), easily computable exact and approximate formulas for uncapacitated single-stage models with normal demand (§5), and clarification of the trade-off between the fill rate and the finished goods target inventory in multistage models (Proposition 4).

2. Related Research

The research on fill rate equations for periodic review models concentrates on a single item in a single-stage system with a constant lead time and a demand process consisting of i.i.d. normal random variables. See §5 for references.

Fill rate expressions are sometimes used to optimize the parameters of an inventory policy, subject to a lower bound on the fill rate induced by the policy. Tijms and Groenevelt (1984) derive an approximate fill-rate formula to optimize the selection of s in an (s, S) policy. Platt et al. (1997) use their approximation to specify approximately optimal (Q, R) policies.

Glasserman and Liu (1997) and Glasserman (1997) consider the same model as in §3, except that batch size decisions are made each period *after* observing the demand (in this paper, the batch size decisions are made *before* observing the demand). Continuing the asymptotic analysis of shortfall stochastic processes in Glasserman and Tayur (1996), these papers develop asymptotic bounds and approximations, including diffusion approximations with higher order correction terms, for fill rates and optimal base-stock levels of capacitated single-stage and multistage systems. In comparison, the fill-rate formulas here are exact and the bounds are valid without asymptotics, but they are not simple analytical expressions as are the asymptotes and diffusion limits.

It is natural to employ a continuous time model to study continuous review systems. The fill rate is easy to compute in a continuous review model with a Poisson demand process. Liu and Cheung (1997) find an optimal base-stock-level policy subject to fill rate and waiting time constraints when the demand process is Poisson. Janssen et al. (1999) determine the reorder point s in a (R, s, Q) model subject to a fill rate constraint when demands comprise a compound renewal process. Zipkin (2000) considers a single-stage model with a compound Poisson demand process. He obtains exact and approximate fill rate expressions and presents methods to minimize the system cost subject to a fill rate constraint. Tijms and Groenevelt (1984) approximately minimize cost subject to a fill rate constraint in a single-item, setup cost model. Their assumptions preclude the base-stock policies in this paper.

Several papers analyze the problem of minimizing costs subject to a fill rate constraint in a Clark and Scarf-type serial model with a Poisson demand process. Boyaci and Gallego (2001) present exact and approximate algorithms; Shang and Song (2001) present an approximate algorithm, and Axsäter (2003) gives sufficient conditions for an optimal mixed base-stock policy (with a compound Poisson demand process).

Ahire and Schmidt (1996) show that single-location fill rate expressions can be useful in a multiechelon system. They consider a single warehouse that supplies a set of retailers facing mutually independent Poisson demand processes. They compare an adaptation of the single-location approximate fill rate in Hadley and Whitin (1963) (also exploited by Johnson et al. 1995; see §5) with estimates of the exact fill rate obtained by simulation.

Song (1998) develops fill rate expressions for a single-echelon *multi-item* inventory system with demands that are a multivariate compound Poisson process. Her work is related to research on other kinds of models (see her references and de Kok 2003), but it appears to have the only expressions for fill rates in a multi-item system. The results in this paper are not special cases of those in Song (1998). Federgruen et al. (1984) minimize cost subject to a fill rate constraint in a multi-item system with a compound Poisson demand process and with setup costs for families of items.

3. Model

The following model describes a periodic review system with N stages in series. Demand arises for the end item, excess demand is backlogged, and the stages are numbered 1 through N so that stage j is downstream from stage $j+1$. Material can be processed at each stage in a single period, and items are stored at the downstream end of stage j after they are processed at stage j. As explained at the end of §6, this description encompasses some models in which the processing time at each stage is any positive integer.

At the beginning of period t, let x_{jt} denote the number of items that are in storage at the downstream end of stage j $(j = 2, \ldots, N;\ t = 1, 2, \ldots)$. Let x_{1t} be the analogous quantity at stage 1 minus the number of items backlogged, if any, at the beginning of period t. Let z_{Nt} be the number of items purchased from vendors in period t, and for $j < N$, let z_{jt} be the number of items that are removed from the inventory at stage $j+1$ and processed at stage j in period t. For each stage $j < N$, it is assumed that there is an upper bound $c_j \le \infty$ on the number of items that can be processed at stage j during any period. The model is *uncapacitated* if $c_j = \infty$ for every $j < N$.

Because a batch cannot exceed the number of inventoried items upstream,

$$0 \le z_{jt} \le \min\{x_{j+1,\,t}, c_j\} \qquad \text{if } 1 \le j < N;\ 0 \le z_{Nt}. \quad (1)$$

Let D_t be the demand in period t, and let D, D_1, $D_2, \ldots$ be independent, identically distributed, and nonnegative random variables with d.f. (distribution function) G and finite expectation μ. To avoid trivialities, it is assumed that $G(0) < 1$ and $\mu < c_j$ $(j = 1, \ldots, N-1)$. Let $G^{(k)}(\cdot)$ denote the k-fold convolution of $G(\cdot)$, and let $G^0(a) = 1$ (0) if $a > (<)\, 0$.

The following chronology of events occurs during each time period t: The inventory vector $(x_{1t}, \ldots, x_{Nt})$ is observed; the batch size vector $(z_{1t}, \ldots, z_{Nt})$ is chosen; items are withdrawn from inventories and processed, and finally demand occurs. Let y_{jt} be the inventory at stage j in period t after items have been processed but before demand occurs:

$$y_{1t} = x_{1t} + z_{1t}; \qquad y_{jt} = x_{jt} + z_{jt} - z_{j-1,\,t} \qquad \text{if } j > 1.$$

Let $(u)^+$ denote $\max\{u, 0\}$. The on-hand inventory that is available to satisfy demand in period t is $(y_{1t})^+$, because $y_{1t} < 0$ if the backlog is positive in period t and exceeds the number of items processed at stage 1. Because excess demand (if any) is backlogged, the inventory dynamics are as follows:

$$x_{1,\,t+1} = y_{1t} - D_t; \qquad x_{j,\,t+1} = y_{jt} \qquad (1 < j \le N). \quad (2)$$

The *fill rate*, β, is the long-run average fraction of demand that can be satisfied immediately. So,

$$\beta = \lim_{T\to\infty} E\left[\sum_{t=1}^{T} \min\{(y_{1t})^+, D_t\} \Big/ \sum_{t=1}^{T} D_t\right]. \quad (3)$$

The expectation and limit exist for the base-stock-level policies that are analyzed in subsequent sections.

4. Uncapacitated Single-Stage Systems

This section considers an uncapacitated single stage in which goods ordered in period t are available to satisfy demand in period $t+L$ (where L is a nonnegative integer), and order quantities are selected with a base-stock-level policy. Let τ be the base-stock level, so $z_{1t} = (\tau - x_{1t})^+$, $t = 1, 2, \ldots$. There is no loss of generality in assuming $x_{11} < \tau$, i.e., the initial inventory is no higher than τ (cf. Lemma 1 in §7). Consequently, $x_{1t} = \tau - \sum_{k=1}^{L} D_{t-k}$, $t \ge L+1$.

Because a single-stage system is a special case of the multistage model described in §3, the following result is a special case of Theorem 2 in §8. However, the single-stage result is easier to derive and to comprehend. Its usefulness is illustrated following the proof and in §5. If $L=0$, formula (4) corresponds to the well-known result (sometimes used to define the fill rate in operations management texts) that $1-\beta$ is the average fraction of demand that is backordered.

Theorem 1.

$$\beta = 1-E\left(\left[D_{L+1}-\left(\tau-\sum_{j=1}^{L}D_j\right)^+\right]^+\right)\Big/\mu \tag{4}$$

$$= \int_0^\tau \left[G^{(L)}(b)-G^{(L+1)}(b)\right]db\Big/\mu. \tag{5}$$

Proof.

$$\begin{aligned}
\beta &= \lim_{T\to\infty} E\left[\sum_{t=1}^{T}\min\{(x_{1t})^+, D_t\}\sum_{t=1}^{T}D_t\right]\\
&= \lim_{T\to\infty} E\left(\left[\sum_{t=1}^{T}\min\{(x_{1t})^+, D_t\}/T\right]\Big/\left[\sum_{t=1}^{T}D_t\Big/T\right]\right)\\
&= (1/\mu)\lim_{T\to\infty} E\left[\sum_{t=1}^{T}\min\{(x_{1t})^+, D_t\}\right]\Big/T\\
&= E\left[\min\left\{\left(\tau-\sum_{j=1}^{L}D_j\right)^+, D_{L+1}\right\}\right]\Big/\mu\\
&= E\left\{D_{L+1}-\left[D_{L+1}-\left(\tau-\sum_{j=1}^{L}D_j\right)^+\right]^+\right\}\Big/\mu\\
&= 1-E\left\{\left[D_{L+1}-\left(\tau-\sum_{j=1}^{L}D_j\right)^+\right]^+\right\}\Big/\mu.
\end{aligned}$$

Let $\beta(\tau)$ make explicit the dependence of β on τ, and let $K(\tau)=\mu[1-\beta(\tau)]$. So,

$$\begin{aligned}
K(\tau) &= E\left\{\left[D_{L+1}-\left(\tau-\sum_{j=1}^{L}D_j\right)^+\right]^+\right\}\\
&= \int_0^\infty\int_{\tau-a}^{\tau}(a+b-\tau)\,dG^{(L)}(b)\,dG(a)\\
&\quad+\int_0^\infty a\int_\tau^\infty dG^{(L)}(b)\,dG(a)\\
&= \mu[1-G^{(L)}(\tau)]+\int_0^\infty\int_{\tau-a}^{\tau}(a+b-\tau)\,dG^{(L)}(b)\,dG(a).
\end{aligned}$$

Leibnitz Rule yields

$$\begin{aligned}
K'(\tau) &= -\int_0^\infty\int_{\tau-a}^{\tau}dG^{(L)}(b)\,dG(a)\\
&= -\int_0^\tau\int_{\tau-b}^{\infty}dG(a)\,dG^{(L)}(b)\\
&= -\int_0^\tau[1-G(\tau-b)]\,dG^{(L)}(b)\\
&= -[G^{(L)}(\tau)-G^{(L+1)}(\tau)].
\end{aligned}$$

Given that $\beta(0)=0$, $K(0)=\mu[1-\beta(0)]=\mu$. Therefore,

$$\begin{aligned}
\beta(\tau) &= 1-K(\tau)/\mu = 1-\left[K(0)+\int_0^\tau K'(a)\,da\right]\Big/\mu\\
&= \int_0^\tau\left[G^{(L)}(a)-G^{(L+1)}(a)\right]da\Big/\mu. \quad\square
\end{aligned}$$

The following simple formulas for uncapacitated single-stage systems with $L=1$ are immediate consequences of (5):

$$\begin{aligned}
\beta &= \int_0^\tau\left[G(b)-G^{(2)}(b)\right]db\Big/\mu\\
&= 1-E[(D-\tau)^+]/\mu = \left\{\tau[1-G(\tau)]+\int_0^\tau x\,dG(x)\right\}\Big/\mu.
\end{aligned}$$

5. Gamma and Normal Demand in Uncapacitated Single-Stage Systems

Gamma Demand Distribution

Let $\Gamma(j,\lambda)$ denote the sum of j i.i.d.r.v.s, each one exponential with parameter λ, i.e., $\Gamma(j,\lambda)$ is a gamma random variable with parameters j and λ. Suppose that D is $\Gamma(r,\lambda)$ for a positive integer r, so $\mu=E(D)=r/\lambda$, $\mathrm{Var}(D)=r/\lambda^2$, and $G^{(L)}$ and $G^{(L+1)}$ are $\Gamma(Lr,\lambda)$ and $\Gamma[(L+1)r,\lambda]$, respectively. So,

$$G^{(L)}(a) = 1-\sum_{j=0}^{Lr}e^{-\lambda a}(\lambda a)^j\Big/j!$$

As a result,

$$\begin{aligned}
\beta &= \int_0^\tau\left[\sum_{j=0}^{(L+1)r}\lambda e^{-\lambda a}(\lambda a)^j\Big/j! - \sum_{j=0}^{Lr}\lambda e^{-\lambda a}(\lambda a)^j\Big/j!\right]da\Big/r\\
&= \int_0^\tau\left[\sum_{j=Lr+1}^{Lr+r}\lambda e^{-\lambda a}(\lambda a)^j\Big/j!\right]da\Big/r\\
&= \sum_{j=Lr+1}^{Lr+r}\int_0^\tau\left[\lambda e^{-\lambda a}(\lambda a)^j/j!\right]da\Big/r.
\end{aligned}$$

Therefore,

$$\beta = \sum_{j=Lr+2}^{(L+1)r+1}P\{\Gamma(j,\lambda)\le\tau\}/r. \tag{6}$$

Normal Demand Distribution

In this subsection, demand (D) is normally distributed with mean μ and variance $\sigma^2 > 0$. Let $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal d.f. and density function, respectively (mean 0 and variance 1), and let $b(a, j) = (a - j\mu)/(\sigma\sqrt{j})$. Normality of demand implies that $G^{(L)}$ has mean $L\mu$ and variance $L\sigma^2$; so $G^{(L)}(a) = \Phi[b(a, L)]$ and (5) yields

$$\beta = \int_0^{\tau} \left(\Phi[b(a,L)] - \Phi[b(a,L+1)]\right) da \Big/ \mu$$
$$= \left[\int_{b(0,L)}^{b(\tau,L)} \Phi(a)\, da - \int_{b(0,L+1)}^{b(\tau,L+1)} \Phi(a)\, da\right] \Big/ \mu. \quad (7)$$

The evaluation of the integrals in (7) exploits the following equation (Hadley and Whitin 1963; Rubinstein 1971, 1976; Stein 1973; Toft 1996; Zipkin 2000, p. 459).

$$\int_r^{\infty} [1 - \Phi(a)]\, da = \phi(r) + r\Phi(r) - r.$$

This equation implies $\int_r^s \Phi(a)\, da = \phi(s) - \phi(r) + s\Phi(s) - r\Phi(r)$ whose substitution in (7) yields the following equation for the fill rate that emphasizes the role of the coefficient of variation $\nu = \sigma/\mu$:

$$\begin{aligned}\beta = \nu\big\{&\sqrt{L}\big(\phi[b(\tau,L)] - \phi[-\sqrt{L}/\nu]\big)\\ &- \sqrt{L+1}\big(\phi[b(\tau,L+1)] - \phi[-\sqrt{L+1}/\nu]\big)\big\}\\ &+ (1/\mu)(\tau - L\mu)\big(\Phi[b(\tau,L)] - \Phi[b(\tau,L+1)]\big)\\ &+ \Phi[b(\tau,L+1)] + L\Phi[-\sqrt{L}/\nu]\\ &- (L+1)\Phi[-\sqrt{L+1}/\nu]. \quad (8)\end{aligned}$$

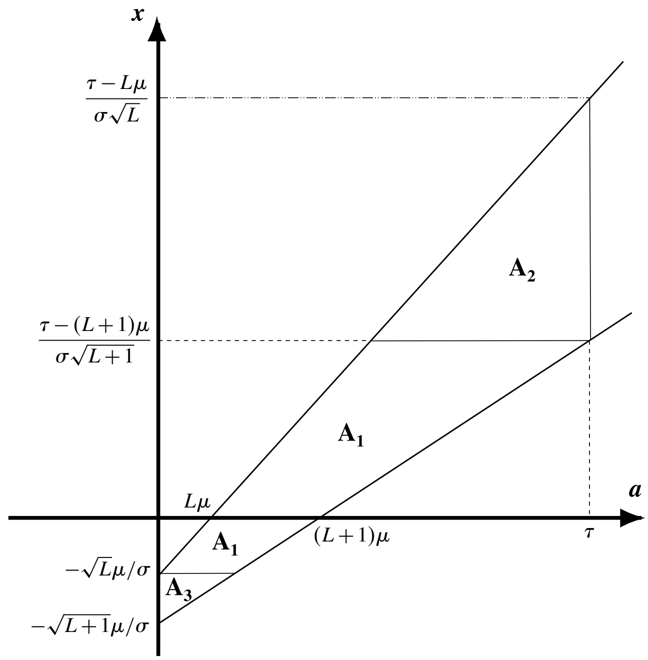
Approximations

Although (8) can be calculated easily, "...a simple approximation sometimes reveals important relationships better than an exact expression" (Zipkin 2000, p. 205). Obtaining a simple approximation begins by writing the integral in (7) as

$$\beta = \int_0^{\tau} \left[\int_{-\infty}^{b(a,L)} \phi(x)\, dx - \int_{-\infty}^{b(a,L+1)} \phi(x)\, dx\right] da \Big/ \mu$$

whose integrand in the $a - x$ plane covers the area in the northeast and southeast quadrants pictured in Figure 1 and bounded by the lines $a = 0$, $a = \tau$, $x = b(a, L)$, and $x = b(a, L+1)$. As shown in Figure 1, this area is the union of three sets

Figure 1 **The Fill Rate Integral for a Single-Stage System with Normal Demand**



in the $a - x$ plane. The first set, A_1 in Figure 1, is bounded by the lines $x = b(0, L)$, $x = b(0, L+1)$, $x = b(a, L)$, and $x = b(a, L+1)$. The second set, A_2 in Figure 1, is the triangle with vertices $(\tau, b(\tau, L+1))$, $(L\mu + \sigma\sqrt{L}b(\tau, L+1), b(\tau, L+1))$, and $(\tau, b(\tau, L))$. The third set, A_3 in Figure 1, is the triangle with vertices $(0, b(0, L+1))$, $(0, b(0, L))$, and $(\mu(L+1) - \mu\sqrt{L(L+1)}, b(0, L))$.

Employing the same reasoning that leads to (8), the integral on A_1 is

$$\int_{b(0,L)}^{b(\tau,L+1)} \left[\mu + \sigma\left(\sqrt{L+1} - \sqrt{L}\right)x\right]\phi(x)\, dx \Big/ \mu$$
$$= \Phi[b(\tau,L+1)] - \Phi[b(0,L)] + \sigma\left(\sqrt{L+1} - \sqrt{L}\right) \cdot \int_{b(0,L)}^{b(\tau,L+1)} x\phi(x)\, dx \Big/ \mu.$$

One might surmise from Figure 1 that the integral on A_1 accounts for most of the fill rate and that most of the integral on A_1 is due to the first two terms above. This heuristic reasoning provides a new approximation:

$$\beta \simeq \Phi\left(\frac{\tau - (L+1)\mu}{\sigma\sqrt{L+1}}\right) - \Phi\left(\frac{-\mu}{\sigma}\sqrt{L}\right) \qquad (L \geq 1). \quad (9)$$

It can be shown that a sufficient condition for the right side to be a lower bound on β is $\tau/\mu > L + 1 + \sqrt{L(L+1)}$. Another new approximation that uses only a table of the standard normal d.f. is obtained from (8) by treating its first three lines on the right side as approximately zero:

$$\beta \simeq \Phi\left(\frac{\tau-(L+1)\mu}{\sigma\sqrt{L+1}}\right) + L\Phi\left(\frac{-\mu}{\sigma}\sqrt{L}\right) - (L+1)\Phi\left(\frac{-\mu}{\sigma}\sqrt{L+1}\right) \qquad (L \geq 1). \tag{10}$$

Approximations (9) and (10) seem to give similar numerical results (which are close to the exact value). Also, they lead shortly to similar approximations for the base-stock level as a function of the fill rate. At high coefficients of variation of demand, both (9) and (10) are more accurate than any of the approximations evaluated by Johnson et al. (1995). Although the approximations due to Hadley and Whitin (1963) and to Johnson et al. require greater computational effort than calculating exact values via (8), they are more accurate than (9) and (10) at low coefficients of variation.

Let Φ^{-1} denote the function inverse of Φ. Then (9) and (10), respectively, correspond to the following expressions for the base-stock level:

$$\tau \simeq (L+1)\mu + \sigma\sqrt{L+1}\,\Phi^{-1}\left[\beta + \Phi\left(\frac{-\mu}{\sigma}\sqrt{L}\right)\right] \tag{11}$$

$$\tau \simeq (L+1)\mu + \sigma\sqrt{L+1}\,\Phi^{-1} \cdot \left[\beta + L\Phi\left(\frac{-\mu}{\sigma}\sqrt{L}\right) - (L+1)\Phi\left(\frac{-\mu}{\sigma}\sqrt{L+1}\right)\right]. \tag{12}$$

Johnson et al. (1995) (a) summarize the literature on approximations for the fill rate in uncapacitated single-stage systems with normal demand, (b) develop a new approximation, and (c) evaluate approximation accuracy via simulations to estimate the exact fill rate. They observe that previous approximations perform well when the fill rate is high but (i) are much too low when the actual fill rate is low and (ii) this bias is worse if demand has a high coefficient of variation. However, the exact formula (8) entails less computation than most previous approximations. Johnson et al. (1995) also consider the robustness of fill rate approximations when the demand is not normal or the lead time is random.

6. Base-Stock-Level Policy and Echelon Variables

Imposing the following cost structure on the uncapacitated multistage model leads to the optimality of a base-stock-level policy (Clark and Scarf 1960, Federgruen and Zipkin 1984, Rosling 1989, Sinha and Sobel 1992, Chen and Zheng 1994). At each stage the cost of processing is proportional to the batch size, the cost of work-in-process inventory is proportional to the inventory level, and the expected net cost of finished-goods inventory (costs of storage and penalties minus revenues) is a convex function of the inventory level. The criterion is either the long-run average cost per time period or the expected present value of the time stream of costs.

Although base-stock-level policies are optimal in capacitated single-stage models (Federgruen and Zipkin 1986), Speck and van der Wal (1991) show that they are not generally optimal for multistage capacitated models. However, Parker and Kapuściński (2002) show that a modified echelon base-stock-level policy is optimal for a capacitated two-echelon model when the downstream facility has a smaller capacity than the upstream facility.

To describe a base-stock-level policy succinctly, it is useful to define the following *echelon variables* (cf. Clark and Scarf 1960):

$$s_{jt} = \sum_{k=1}^{j} x_{kt} \quad \text{and} \quad a_{jt} = s_{jt} + z_{jt} = \sum_{k=1}^{j} y_{kt} \qquad (j = 1, 2, \ldots, N). \tag{13}$$

The formulation in *installation variables* x_{jt}, y_{jt}, and z_{jt} is equivalent to a formulation in echelon variables because $x_{jt} = s_{jt} - s_{j-1,\,t}$, $y_{jt} = a_{jt} - a_{j-1,\,t}$ (let $s_{0t} = a_{0t} \equiv 0$), and $z_{jt} = a_{jt} - s_{jt}$. Because $a_{1t} = y_{1t}$, definition (3) can be written as

$$\beta = \lim_{T\to\infty} E\left[\sum_{t=1}^{T} \min\{(a_{1t})^+, D_t\} \Big/ \sum_{t=1}^{T} D_t\right]. \tag{14}$$

The constraints on batch sizes (1) correspond to

$$s_{1t} \leq a_{1t} \leq s_{2t} \leq \cdots \leq s_{Nt} \leq a_{Nt} \quad \text{and} \quad a_{jt} \leq s_{jt} + c_j, \qquad 1 \leq j < N. \tag{15}$$

The dynamics (2) correspond to

$$s_{j,\,t+1} = a_{jt} - D_t. \tag{16}$$

A base-stock-level policy depends on echelon base-stock levels $\tau_1, \ldots, \tau_N$ and can be specified as $a_{Nt} = \max\{\tau_N, s_{Nt}\}$ and for $j < N$ as follows:

$$a_{jt} = \begin{cases} s_{jt} & \text{if } \tau_j < s_{jt} \\ \min\{\tau_j, s_{jt} + c_j\} & \text{if } s_{jt} \le \tau_j \le s_{j+1,t} \\ \min\{s_{j+1,t}, s_{jt} + c_j\} & \text{if } s_{j+1,t} < \tau_j. \end{cases} \tag{17}$$

The batch size specified by (17) is $z_{jt} = \min\{(\tau_j - s_{jt})^+, c_j\}$ if this quantity is feasible in constraint (1). If $\tau_j - s_{jt}$ exceeds the upstream inventory $x_{j+1,t}$, then (17) specifies $z_{jt} = \min\{s_{j+1,t} - s_{jt}, c_j\}$.

Suppose that there is a stage j with $c_j = \infty$ where $1 + L_j$ periods are needed to process material at stage j and L_j is a positive integer (contrary to the assumption $L_j = 0$ made in §§7–10). If material is processed at stage j beginning in period t, it enters the stage j inventory during period $t + L_j$ and is available for processing at stage $j-1$ in period $t + L_j + 1$. There is an equivalent model in which stage j is replaced by $L_j + 1$ pseudo-stages $(j,1), (j,2), \ldots, (j, 1+L_j)$ where $(j, k+1)$ is upstream from (j,k) and each pseudo-stage (j,k) has $L_{jk} = 0$. It follows from (17) that if the base-stock levels τ_{jk} are set sufficiently high at the pseudo-stages, then $z_{(j,k),t} = s_{(j,k+1),t} - s_{(j,k),t}$ for every t and k. So it always takes exactly $1 + L_j$ periods for material to pass through all of the pseudo-stages. This claim remains true if $c_j < \infty$, but it is not apparent that there is an assignment of capacities to the pseudo-stages that is equivalent to the capacity constraint in the original model. Henceforth, $L_j = 0$ for each j.

7. Fill Rate in Multistage Systems

Most of the formulas are derived under the assumption that the initial echelon inventory levels are no higher than the echelon base-stock levels. The assumption is justified by the fact that, regardless of the initial inventory levels, eventually all the echelon inventory levels are bounded above by the echelon base-stock levels. The proof of this fact is brief, straightforward, and deleted.

Lemma 1. (a) *With probability* 1 *there is a period* $t^* < \infty$ *such that* $s_{jt} < \tau_j$, $j = 1, \ldots, N$ *for all* $t > t^*$. (b) *If* $s_{j1} < \tau_j$, $j = 1, \ldots, N$, *then* $s_{jt} < \tau_j$, $j = 1, \ldots, N$ *for all* t.

It is convenient to write τ for τ_1, define $\mathcal{J} = \{0, 1, \ldots\}$, and for each $j < N$ let

$$C_j = \min\{c_k : 1 \le k \le j\}$$

and

$$Q_j = \max_{r \in \mathcal{J}} \left\{ \sum_{k=1}^{r} (D_{N+j+k} - C_j) \right\},$$

let $Q_N = \max_{r \in \mathcal{J}} \{\sum_{k=1}^{r} (D_{2N+k} - C_{N-1})\}$ and $\mathcal{M}_t = \tau + \max\{\sum_{k=t-j}^{t-1} D_k - \tau_{j+1} : 0 < j < N-1\}$ for $t \ge N$, and let

$$M = \tau + \max\left\{ Q_j + \sum_{k=1}^{j-1} D_k - \tau_j : j = 1, \ldots, N \right\}. \tag{18}$$

That is, C_j is the smallest capacity among the first j stages, and $\mathcal{M}_t$ is the maximum of the N "ladder heights" of nonstationary random walk $\{\sum_{k=t-j}^{t-1} D_k - (\tau_{j+1} - \tau_j)\}$, Q_j is the maximum of denumerably many ladder heights of a stationary random walk, and M is the maximal sum of ladder height maxima. It will be apparent in (19) that M is the delivery shortfall at the lowest echelon.

The mutual independence of $\{D_t\}$ implies that Q_j and $\sum_{k=1}^{j-1} D_k + \tau - \tau_j$ are independent for each j. Because $\mu < c_j$ for all $j = 1, \ldots, N-1$, Q_j is the maximum of a random walk with negative drift; therefore, it is a finite nonnegative random variable for each j.

Lemma 2.

$$\textit{As } t \to \infty,\ a_{1t} \textit{ converges in distribution to } \tau - M. \tag{19}$$

In the uncapacitated model with $s_{j1} < \tau_j$ *for each* $j = 1, \ldots, N$, *for* $t \ge N$,

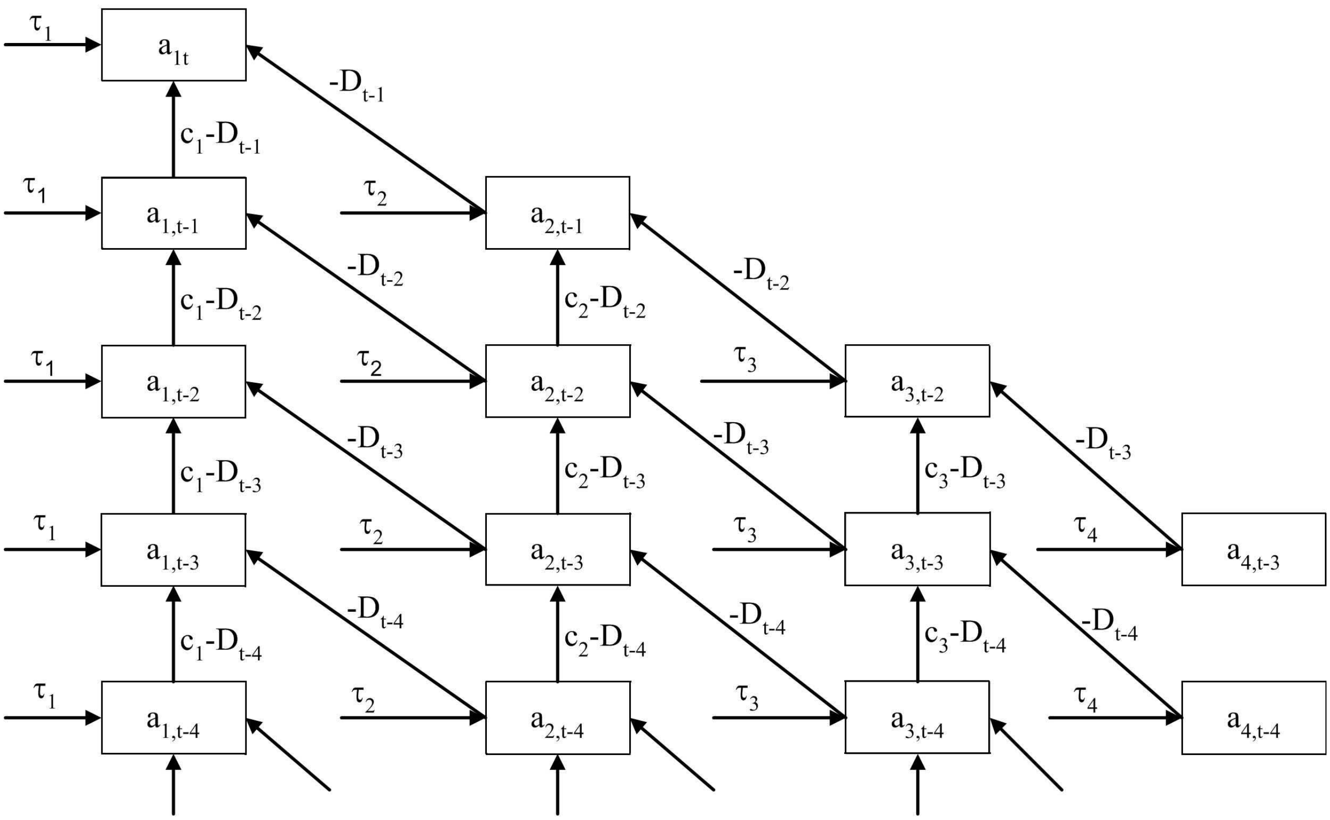
$$a_{1t} = \tau - \mathcal{M}_t \sim \tau_1 - M = \min\left\{ \tau_{j+1} - \sum_{k=1}^{j} D_k : 0 \le j \le N-1 \right\}.$$

Proof. If the model is uncapacitated with $s_{j1} < \tau_j$ for all j, Lemma 1, (16), and (17) imply

$$\begin{aligned} a_{1t} &= \min\{\tau, s_{2t}\} = \min\{\tau, a_{2,t-1} - D_{t-1}\} \\ &= \min\{\tau, \min\{\tau_2, s_{3,t-1}\} - D_{t-1}\} \\ &= \min\{\tau, \tau_2 - D_{t-1}, s_{3,t-1} - D_{t-1}\} \\ &= \min\{\tau, \tau_2 - D_{t-1}, a_{3,t-2} - D_{t-1} - D_{t-2}\} = \cdots \\ &= \min\left\{\tau, \tau_2 - D_{t-1}, \tau_3 - D_{t-1} - D_{t-2}, \ldots, \tau_N - \sum_{k=1}^{N-1} D_{t-k}\right\} \\ &= \tau + \min\left\{0, \tau_2 - \tau - D_{t-1}, \ldots, \sum_{k=1}^{N-1} (\tau_{k+1} - \tau_k - D_{t-k})\right\} = \tau - \mathcal{M}_t. \end{aligned}$$

Because $\{D_t\}$ are i.i.d.r.v.s, $\mathcal{M}_t \sim M$.

Figure 2 **Sample Paths Culminating in a_{1t} when $N = 4$**



To see the general structure of a_{1t} in the capacitated model, consult Figure 2, which maps the sample paths that culminate in a_{1t} when $N = 4$. For example, three arcs enter node $a_{2,t-3}$ because there are three possibilities in $a_{2,t-3} = \min\{\tau_2, s_{2,t-3} + c_2, s_{3,t-3}\}$. The arc to node $a_{2,t-3}$ from node $a_{2,t-4}$ corresponds to $a_{2,t-3} = s_{2,t-3} + c_2 = a_{3,t-4} - D_{t-4} + c_2$; so this arc is labelled $c_2 - D_{t-4}$. The arc to node $a_{2,t-3}$ from node $a_{3,t-4}$ corresponds to $a_{2,t-3} = s_{3,t-3}$ because $s_{3,t-3} = a_{3,t-4} - D_{t-4}$; so this arc is labelled $-D_{t-4}$. The third arc to node $a_{2,t-3}$ corresponds to $a_{2,t-3} = \tau_2$; so it carries the label τ_2.

Tracing a path to node a_{1t} and summing the labels on the path's arcs yields a sample path value of a_{1t}. For example, there are six paths to a_{1t} that begin with the arc to node $a_{3,t-4}$ that is labelled τ_3. The minimum of their label sums is

$$\tau_3 - \sum_{i=1}^{4} D_{t-i} + \min\{2c_1, 2c_2, 2c_3, c_1 + c_2, c_1 + c_3, c_2 + c_3\} = \tau_3 - \sum_{i=1}^{4} D_{t-i} + 2c_3.$$

There are three paths to a_{1t} that originate with the arc to node $a_{3,t-3}$, and the minimum of their label sums is

$$\tau_3 - \sum_{i=1}^{3} D_{t-i} + \min\{c_1, c_2, c_3\} = \tau_3 - \sum_{i=1}^{3} D_{t-i} + C_3.$$

Generally, starting from the arc to node $a_{j,t-k}$ that is labelled τ_j $(j-1 \le k \le t-1)$, the minimum of the sums of labels on the paths to a_{1t} is

$$\tau_j - \sum_{i=1}^{k} D_{t-i} + (k-j+1)C_j = \tau - \sum_{i=1}^{j-1}(D_{t-i} - \Delta_i) + \sum_{i=j}^{k}(C_j - D_{t-i}).$$

Therefore, the minimum of the sums of labels on the paths to a_{1t} starting from *any* arc that is labelled τ_j is

$$\tau - \sum_{i=1}^{j-1}[D_{t-i} - (\tau_{i+1} - \tau_i)] - \max\left\{\sum_{i=j}^{k}(D_{t-i} - C_j) \colon j-1 \le k \le t-1\right\}.$$

The second term is $\sum_{k=1}^{j-1} D_k + \tau - \tau_j$, and the third term converges in distribution to Q_j as $t \to \infty$ because $\mu < C_{N-1} \le C_j$.

The proof of (19) is completed by confirming that the probability of the complement of the set of sample paths whose first arc is labelled τ_j converges to zero as $t \to \infty$. Because there are $N < \infty$ of these sets (corresponding to $\tau_1, \ldots, \tau_N$), the probability of their union converges to zero.

For any j, k, and t with $0 \le j-1 \le k \le t-1$,

$$\left\{\tau_j \ge a_{j,\,t-k} + \sum_{i=1}^{k}(c_j - D_{t-i})\right\}$$
$$= \left\{\tau_j \ge s_{j,\,t-k+1} + D_{t-k} + \sum_{i=1}^{k}(c_j - D_{t-i})\right\}$$
$$\subseteq \left\{\tau_j \ge \tau_j + D_{t-k} + \sum_{i=1}^{k}(c_j - D_{t-i})\right\}$$
$$= \left\{0 \ge D_{t-k} + \sum_{i=1}^{k}(c_j - D_{t-i})\right\}.$$

Therefore, starting from node a_{j1} (i.e., $k = t-1$) and exploiting $c_j - \mu > 0$ yields

$$P\left\{\tau_j \ge a_{j,1} + \sum_{i=1}^{t-1}(c_j - D_i)\right\}$$
$$\le P\left\{0 \ge D_1 + \sum_{i=1}^{t-1}(c_j - D_i)\right\} \to 0 (t \to \infty). \quad \square$$

In the uncapacitated model the distribution function of M (hence the distribution function of a_{1t} for $t \ge N$) can be specified recursively as follows. Let $M_N = 0$, and for $j = 1, \ldots, N-1$ let

$$M_j = \tau_j + \max\left\{\sum_{k=j}^{r} D_k - \tau_{r+1} : j \le r \le N-1\right\}. \tag{20}$$

So $M = M_1$. Let B_j be the distribution function of M_j. Then B_1 is the distribution function of M, $B_N(u) = 1\ (0)$ if $u \ge 0\ (<0)$, and $B_j(u) = 0$ if $u < 0$ (for each j).

Proposition 1. *For $j = 1, \ldots, N-1$ in the uncapacitated model, $M_j = (D_j - \tau_{j+1} + \tau_j + M_{j+1})^+$, and for $y \ge 0$*

$$B_j(y) = B_{j+1}(0) G(y + \tau_{j+1} - \tau_j) + \int_0^{y+\tau_{j+1}-\tau_j} G(y - z + \tau_{j+1} - \tau_j)\, dB_{j+1}(z). \tag{21}$$

Proof. Definition (20) yields $M_j = (D_j - \tau_{j+1} + \tau_j + M_{j+1})^+$ that implies (21) because $D_j - \tau_{j+1} + \tau_j$ and M_{j+1} are independent, D_j is nonnegative, and $P\{D_j - \tau_{j+1} + \tau_j \le y - z\} = G(y - z - \tau_{j+1} + \tau_j) = 0$ if $z > y + \Delta_j$. So

$$\int_0^\infty P\{D_j - \tau_{j+1} + \tau_j \le y - z\}\, dB_{j+1}(z)$$
$$= \int_0^{y+\tau_{j+1}-\tau_j} G(y - z + \tau_{j+1} - \tau_j)\, dB_{j+1}(z). \quad \square$$

It follows from Lemma 2 that $s_{j1} < \tau_j$ in an uncapacitated model, implying that a_{1N}, $a_{1,N+1}, \ldots$ are identically distributed. This sequence also has periodic independence. Let $\mathcal{F}_t = 1$ if $D_t = 0$ and $\mathcal{F}_t = \min\{(a_{1t})^+, D_t\}/D_t$ if $D_t > 0$. Therefore, $\beta = \lim_{t\to\infty} E(\mathcal{F}_t)$. It follows from Lemma 2 that $\mathcal{F}_t$ converges in distribution as $t \to \infty$. Recall that G is the distribution function of D.

Proposition 2. (a)

$$\beta = G(0) + [1 - G(0)] \lim_{t\to\infty} E(\mathcal{F}_t \mid D_t > 0). \tag{22}$$

(b) *In an uncapacitated model, if $s_{j1} < \tau_j$, $j = 1, \ldots, N$, then* (i) *a_{1t}, $a_{1,t+N-1}$, $a_{1,t+2N-2}, \ldots$ are independent random variables for all $t \ge N$, and* (ii) *$\mathcal{F}_N$, $\mathcal{F}_{N+1}, \ldots$ are identically distributed; so $\beta = E(\mathcal{F}_N \mid D_N > 0) \cdot [1 - G(0)] + G(0)$.*

Proof. Equation (22) follows from $\beta = \lim_{t\to\infty} E(\mathcal{F}_t)$ (due to (14) and Lemma 2) and the implication of Lemma 2 that $\mathcal{F}_t$ converges in distribution as $t \to \infty$. For (b), in an uncapacitated model Lemma 2 provides $a_{1t} = \tau_1 - \mathcal{M}_t$ and $\mathcal{M}_t$ depends on $D_{t-N+1}, \ldots, D_{t-1}$. So (i) follows from the dependence of $a_{1,t+N-1}$ only on $D_t, \ldots, D_{t+N-2}$. For (ii), Lemma 2 and (14) imply $\beta = E(\mathcal{F}_N)$. $\square$

Because D, D_1, $D_2, \ldots$ are i.i.d.r.v.s, it follows that D and M are independent. The following formula is a consequence of Lemma 2 and is a generalization of (4). It emphasizes that the fill rate depends not only on the base-stock level at echelon 1, but also on the shortfalls of deliveries to stage 1 due to the upstream inventories and capacity constraints.

Theorem 2.

$$\beta = 1 - E\{[D - (\tau - M)^+]^+\}/\mu. \tag{23}$$

Proof. Without loss of generality (Lemma 1), let $s_{j1} < \tau_j$, $j = 1, \ldots, N$. Employing Lemma 2,

$$\begin{aligned}
\beta &= \lim_{T\to\infty} E\left[\sum_{t=1}^{T} \min\{(a_{1t})^+, D_t\} \Big/ \sum_{t=1}^{T} D_t\right] \\
&= \lim_{T\to\infty} E\left[\sum_{t=1}^{T} \min\{(a_{1t})^+, D_t\} \Big/ T \Big/ \sum_{t=1}^{T} D_t \Big/ T\right] \\
&= (1/\mu) \lim_{T\to\infty} E\left[\sum_{t=1}^{T} \min\{(a_{1t})^+, D_t\}\right] \Big/ T \\
&= E[\min\{(\tau - M)^+, D\}]/\mu \\
&= E\{D - [D - (\tau - M)^+]^+\}/\mu \\
&= 1 - E\{[D - (\tau - M)^+]^+\}/\mu. \quad \square
\end{aligned}$$

It follows from (23) that the long-run fraction of demand that is backordered is $E\{[D - (\tau - M)^+]^+\}/\mu$.

Let A be the distribution function of $(a_{1t})^+$. For large enough t, Lemma 2 implies $A(x) = P\{\tau - M < x\}$. Let $\overline{A}(x) = 1 - A(x)$ and $\overline{G}(x) = 1 - G(x)$.

Proposition 3.

$$\beta = \left[\int_0^\infty x\overline{A}(x)\, dG(x) + \int_0^\infty x\overline{G}(x)\, dA(x)\right] \Big/ \mu. \tag{24}$$

Proof. Assume $s_{j1} < \tau_j$, $j = 1, \ldots, N$, and $t > N$. Let B be the distribution function of $\min\{(a_{1t})^+, D_t\}$. Because $(a_{1t})^+$ and D_t are independent, $B(x) = 1 - [1 - A(x)][1 - G(x)]$. Therefore, (5) implies

$$\begin{aligned}
\beta &= \int_0^\infty x\, dB(x) \Big/ \mu \\
&= \left[\int_0^\infty x\, dG(x) + \int_0^\infty x\, dA(x) - \int_0^\infty x G(x)\, dA(x) \right.\\
&\qquad \left. - \int_0^\infty x A(x)\, dG(x)\right] \Big/ \mu \\
&= \left[\int_0^\infty x[1 - A(x)]\, dG(x) \right.\\
&\qquad \left. + \int_0^\infty x[1 - G(x)]\, dA(x)\right] \Big/ \mu. \quad \square
\end{aligned}$$

It is apparent from Theorem 2 that the fill rate depends on the target echelon inventories $\tau, \tau_2, \ldots, \tau_N$ because they influence the distribution of M. The next result focuses on the effect of the finished goods target inventory (τ) on the fill rate. Because M and D are independent nonnegative random variables, the following result, a generalization of (5), implies that the fill rate is a nondecreasing function of τ and specifies trade-offs between the fill rate and the finished goods target inventory level.

Proposition 4.

$$\beta(\tau) = \int_0^\tau [P\{M \le v\} - P\{M + D \le v\}]\, dv \Big/ \mu. \tag{25}$$

Proof. Let $K(\tau) = \mu[1 - \beta(\tau)]$; so $\beta(\tau) = 1 - K(\tau)/\mu$ and $\beta'(\tau) = -K'(\tau)/\mu$. Letting Q denote the distribution function of M, (23) can be written

$$K(\tau) = \int_0^\infty \int_{\tau - a}^{\tau} (a + b - \tau)\, dQ(b)\, dG(a) + \mu[1 - Q(\tau)].$$

So, $K'(\tau) = \int_0^\infty Q(\tau - a)\, dG(a) - Q(\tau) = P\{M + D \le \tau\} - P\{M \le \tau\}$. Given that $\beta(0) = 0$ implies $K(0) = \mu$,

$$\begin{aligned}
\beta(\tau) &= 1 - K(\tau)/\mu = 1 - K(0) + \int_0^\tau K'(a)\, da\Big] \Big/ \mu \\
&= \int_0^\tau [P\{M \le a\} - P\{M + D \le a\}]\, da \Big/ \mu. \quad \square
\end{aligned}$$

When demand is a discrete integer-valued random variable, (25) becomes

$$\beta(\tau) = \frac{1}{\mu} \sum_{k=1}^{\tau} [P\{M \le k\} - P\{M + D \le k\}]. \tag{26}$$

Therefore, it is simple to investigate the sensitivity of $\beta(\tau)$ to τ because

$$\begin{aligned}
&\beta(\tau) - \beta(\tau - 1) \\
&\quad = [P\{M \le k\} - P\{M \le k - 1\}] \\
&\qquad - [P\{M + D \le k\} - P\{M + D \le k - 1\}].
\end{aligned}$$

Similarly, if demand is a continuous random variable,

$$\beta'(\tau) = P\{M \le \tau\} - P\{M + D \le \tau\}.$$

The next result asserts that the fill rate diminishes as the number of stages increases. Therefore, it is important to evaluate the service levels of alternative system designs. Let β_k denote the fill rate of the k-stage model in which stages $k+1, \ldots, N$ have been removed and the constraints (1) are replaced by

$$0 \le z_{jt} \le x_{j+1,\,t} \qquad \text{if } 1 \le j < k; \;\; 0 \le z_{kt}.$$

Corollary 1.

$$\beta_1 > \beta_2 > \cdots > \beta_N.$$

Proof. By definition, M is stochastically increasing in N. So (19) implies $\beta_j > \beta_{j+1}$ for each j. $\square$

This result is similar in spirit to the reduction of the reliability of a series system as the number of stages increases.

8. Fill Rate Bounds

Theorem 2 yields a good lower bound. Unfortunately, the same reasoning provides a useless upper bound (see §9).

Corollary 2.

$$1 - E[(D-\tau+M)^+]/\mu \le \beta \le E[(\tau - M)^+]/\mu. \quad (27)$$

Proof. The inequalities are immediate consequences of (23) and $E\{[D-(\tau-M)^+]^+\} < E[(D-\tau+M)^+]$. □

An alternative derivation of the lower bound begins with (3) and

$$\min\{(a_{1t})^+, D_t\} = D_t - [D_t - (a_{1t})^+]^+ > D_t - (D_t - a_{1t})^+.$$

So,

$$\sum_{t=1}^{T} \min\{(a_{1t})^+, D_t\} \Big/ \sum_{t=1}^{T} D_t > 1 - \left[\sum_{t=1}^{T} (D_t - a_{1t})^+ \Big/ T\right] \Big/ \left[\sum_{t=1}^{T} D_t \Big/ T\right]. \quad (28)$$

The bound results from the application of Lemma 2 to (28) in the same manner as in the proof of Theorem 1.

The lower bound in the next result is looser than in (27) but simpler to calculate. The upper bound is obtained by treating a period's fraction filled as unity if the supply level is nonnegative, i.e., if any demand could be satisfied. Recall that G is the distribution function of D. The upper bound is better than the one in (27) but is still poor (see §9).

Proposition 5. *In the uncapacitated model,*

$$E[G(\tau - M)] \le \beta \le P\{M \le \tau\} + G(0). \quad (29)$$

Proof. For the lower bound, let $\delta(x) = 1$ if $x > 0$ and $\delta(x) = 0$ if $x < 0$. Given that $\min\{(a_{1t})^+, D_t\} > D_t\delta(a_{1t} - D_t)$, (22) yields

$$\begin{aligned}
\beta &= G(0) + \overline{G}(0) \lim_{t\to\infty} E(\{\mathcal{F}_t \mid D_t > 0) \\
&= G(0) + \overline{G}(0) \lim_{t\to\infty} E[\min\{(a_{1t})^+, D_t\}/D_t \mid D_t > 0] \\
&\ge G(0) + \overline{G}(0) \lim_{t\to\infty} E[D_t\delta(a_{1t} - D_t)/D_t \mid D_t > 0] \\
&= G(0) + \overline{G}(0) E[\delta(\tau - M - D \mid D > 0] \\
&= G(0) + E[P\{0 < D \le M\} \\
&= G(0) + (E[G(\tau - M)] - G(0)).
\end{aligned}$$

For the upper bound in (29), when $d > 0$ in (26), use $\min\{d, (\tau - M)^+\}/d \le \delta(\tau - M)$:

$$\begin{aligned}
\beta &\le G(0) + \overline{G}(0) E[\delta(\tau - M) \mid D > 0] \\
&= G(0) + \overline{G}(0) P\{M \le \tau\}/P\{D > 0\}. \quad \square
\end{aligned}$$

Table 1 Probability Distribution of Demand

i	0	1	2	3	4	5	6
$P\{D = i\}$	0.2	0.1	0.1	0.2	0.2	0.1	0.1

Table 2 Probability Distributions of Maxima

k	$P\{\mathcal{V}_2 = k\}$	$P\{\mathcal{V}_3 = k\}$	$P\{\mathcal{V}_4 = k\}$
0	0.8	0.74	0.462
1	0.1	0.10	0.128
2	0.1	0.08	0.126
3	—	0.05	0.093
4	—	0.02	0.073
5	—	0.01	0.053
6	—	—	0.033
7	—	—	0.019
8	—	—	0.009
9	—	—	0.003
10	—	—	0.001

9. Examples

A model with four stages in Sinha and Sobel (1989) yields an optimal policy consisting of $\tau = 6$, $\tau_2 = 10$, $\tau_3 = 13$, and $\tau_4 = 14$ with the distribution of demand in Table 1. It is instructive to calculate the fill rate and its bounds for a succession of four models, labelled 1 through 4, in which model n consists of stages 1 through n. Let $\mathcal{V}_n$ denote M in model n. It is straightforward to calculate the probability distributions of $\mathcal{V}_1, \ldots, \mathcal{V}_4$, and they are displayed in Table 2. These

Table 3 Fill Rates and Fill Rate Bounds, Models 1–4

n	$E[G(\tau - M)]$	$1 - E(D + M - \tau)^+/\mu$	β	$G(0) + P\{M \le \tau\}$	$E[(\tau - M)^+]/\mu$
1	1	1	1	1.2	2.14
2	0.97	0.9857	0.9857	1.2	2.04
3	0.935	0.9589	0.9589	1.2	2.10
4	0.7855	0.8089	0.8248	1.168	1.57

distributions yield Table 3, which tabulates the fill rates and bounds on fill rates for the four models.

In Table 3, the fill rates diminish with n as asserted by Corollary 1. Both lower bounds seem to be useful with $1 - E[D + M - \tau]^+/\mu$ closer to the fill rate than $E[G(\tau - M)]$ but requiring more computational effort.

Acknowledgments

The author is grateful to Paul Zipkin and the referees for suggesting that $\int_r^\infty [1 - \Phi(a)]\, da = \phi(r) + r\Phi(r) - r$ be exploited in §5, and to Peter Ritchken for references to bivariate generalizations of this equation.

References

Ahire, S. L., C. P. Schmidt. 1996. A model for a mixed continuous-periodic review one-warehouse, N-retailer inventory system. *Eur. J. Oper. Res.* **92** 69–82.

Axsäter, S. 2003. Note: Optimal policies for serial inventory systems under fill rate constraints. *Management Sci.* **49** 247–253.

Boyaci T., G. Gallego. 2001. Serial production/distribution systems under service constraints. *Manufacturing Service Oper. Management* **3** 43–50.

Chen, F., Y.-S. Zheng. 1994. Lower bounds for multi-echelon stochastic inventory systems. *Management Sci.* **40** 1426–1443.

Clark, A. J., H. Scarf. 1960. Optimal policies for a multi-echelon inventory problem. *Management Sci.* **6** 475–490.

de Kok, T. G. 1986. An inventory model with limited production capacity and uncertain demands. *Math. Oper. Res.* **11** 193–215.

de Kok, T. G. 2003. Evaluation and optimization of strongly ideal assemble-to-order systems. Unpublished manuscript, Department of Technology Management, Technische Universiteit Eindhoven, Eindhoven, The Netherlands.

de Kok, T. G., P. Zipkin. 1984. Computational issues in an infinite-horizon multi-echelon inventory model. *Oper. Res.* **32** 818–836.

Federgruen, A., H. Groenevelt, H. C. Tijms. 1984. Coordinated replenishments with compound Poisson demands. *Management Sci.* **30** 344–357.

Glasserman, P. 1997. Bounds and asymptotics for planning critical safety stocks. *Oper. Res.* **45** 244–257.

Glasserman, P., T.-W. Liu. 1997. Corrected diffusion approximations for a multistage production-inventory system. *Math. Oper. Res.* **22** 186–201.

Glasserman, P., S. Tayur. 1994. The stability of a capacitated, multi-echelon production-inventory system under a base-stock policy. *Oper. Res.* **42** 913–925.

Hadley, G., T. Whitin. 1963. *Analysis of Inventory Systems*. Prentice-Hall, Englewood Cliffs, NJ.

Hoffman, T. 1996. Master Lock takes fear out of customer service overhaul. *Computerworld* **30**(6) 81–82.

Janssen, F., R. Heuts, T. De Kok. 1999. The impact of data collection on fill rate performance in the (R, s, Q) inventory model. *J. Oper. Res. Soc.* **50** 75–84.

Johnson, M. E., H. L. Lee, T. Davis, R. Hall. 1995. Expressions for item fill rates in periodic inventory systems. *Naval Res. Logist.* **42** 57–80.

Liu, L., K. L. Cheung. 1997. Service constrained inventory models with random lifetimes and lead times. *J. Oper. Res. Soc.* **48** 1022–1028.

Orenstein, D. 1999. MGM outsources to manage video. *Computerworld* **33**(29) 4.

Parker, R. P., R. Kapuściński. 2002. Optimal policies for a capacitated two-echelon inventory system. Unpublished manuscript, Department of Operations Management, School of Business, University of Michigan.

Plantz, B. 1997. Turning a company around. *Furniture Design Manufacturing* **69**(3) 34–41.

Platt, D. E., L. W. Robinson, R. B. Freund. 1997. Tractable (Q, R) heuristic models for constrained service levels. *Management Sci.* **43** 951–965.

Richardson, H. L. 1998. Customer service drives warehouse choices. *Transportation Distribution* **39**(12) 59–64.

Rosling, K. 1989. Optimal inventory policies for assembly systems under random demands. *Oper. Res.* **37** 565–579.

Rubinstein, M. 1971. A comparative statics analysis of risk premiums, Essay 2. *Four Essays in the Theory of Competitive Security Markets*. Ph.D. dissertation, UCLA, Los Angeles, CA.

Rubinstein, M. 1976. The valuation of uncertain income streams and the pricing of options. *Bell J. Econom.* **7** 407–425.

Shang, K. H., J. S. Song. 2001. Analysis of serial supply chains with a service constraint. Unpublished manuscript, Graduate School of Management, University of California, Irvine, CA.

Silver, E. A. 1970. A modified formula for calculating customer service under continuous inventory review. *AIIE Trans.* **2** 241–245.

Sinha, C., M. J. Sobel. 1992. Work order release rules in serial manufacturing of a single product. Unpublished manuscript, Department of Applied Mathematics and Statistics, State University of New York at Stony Brook, NY.

Song, J. S. 1998. On the order fill rate in a multi-item, base-stock inventory system. *Oper. Res.* **46** 831–845.

Speck, C. J., N. J. van der Wal. 1991. The capacitated multi-echelon inventory system with serial structure: 1. The “push ahead”-effect. Memorandum COSOR 91-39, Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, The Netherlands.

Stein, C. 1973. Estimation of the mean of a multivariate normal distribution. *Proc. Prague Sympos. Asymptotic Statist.* Universita Karlova (September 3–6).

Sweeney, J. 1996. Keeping the customer satisfied. *Comput. Reseller News* (677) 47–48.

Tijms, H. C., H. Groenevelt. 1984. Simple approximations for the reorder point in periodic and continuous review (s, S) inventory systems with service level constraints. *Eur. J. Oper. Res.* **17** 175–190.

Toft, K. B. 1996. On the mean-variance tradeoff in option replication with transactions costs. *J. Financial Quant. Anal.* **31** 233–263.

Zipkin, P. H. 2000. *Foundations of Inventory Management*. McGraw Hill, New York.