COLUMBIA
BUSINESS
SCHOOL

# Analyzing Inventory Cost and Service in Supply Chains

Garrett J. van Ryzin

April 2001

# Contents

# 1 Introduction

The term *supply chain* refers to the complex sequence of activities, information and material flows involved in producing and distributing a firm's outputs. Supply chains consume vast amounts of capital - in the form of plant, equipment and inventories - and are responsible for most of a firm's cost-of-goods and operating expenses. Supply chains create significant value and ultimately determine a firm's ability to satisfy the demands of its customers. As a result, effective supply chain management is a major strategic challenge for most firms.

But formulating effective strategy requires a good understanding of what drives cost and service in a supply chain. This note introduces an inventory model that will enable you to quantifying the often subtle impact of both operational and structural changes in a supply chain. In addition - and perhaps more importantly - the model can help sharpen your intuition about what factors affect the operating performance of a supply chain.

# 2 What are inventories and why do they occur?

At a very basic level, inventories are merely the byproduct of a universal "accounting law" of material flow:

**inventory = cumulative supply - cumulative demand**.

In other words, if we draw a "black box" around a piece of our supply chain and measure the total amount of product that flows into the box (the supply) and then subtract the total amount that we observe leaving the box (the demand), the result (assuming no product is destroyed along the way) is the inventory in the box.

To put it mathematically, let $I(t)$ denote the inventory in the "box" at time $t$, and assume we start operations at time 0 with the box empty, i.e. $I(0) = 0$. Let $S(0, t]$ denote the cumulative supply (flow into the "box") up to time $t$ and $D(0, t]$ denote the cumulative demand (flow out of the "box") up to time $t$. [1]

Then

$$I(t) = S(0, t] - D(0, t]. \tag{1}$$

While one normally thinks of inventory as a positive quantity (stuff sitting somewhere), we can easily extend the above definition to allow for *negative* inventory, i.e. $I(t) < 0$.

What does a negative unit of inventory look like? Well, a lot like a *back-order*. If consumers of our output are willing to order without receiving product, then we can allow demand to exceed supply and $I(t)$ can become negative, in which case, $-I(t)$ represents the total quantity on back-order. Think of a positive inventory as a pile of goods waiting for orders and a negative inventory as a pile of orders waiting for goods.

# 3 Causes of supply/demand imbalances in a supply chain

Now, the clever reader has no doubt already spotted the solution to eliminating all inventories (and back-orders for that matter): simply set $S(0, t] = D(0, t]$ at all times $t$ (keep supply equal to demand) throughout the supply chain, then - poof! - $I(t) = 0$ forever! Alas, in real life things are not quite that simple. Indeed, the real key to understanding inventory cost and service is to understand what causes imbalances in supply and demand in the first place.

## 3.1 Planned imbalances

Here's a dirty little secret about operations managers: they often deliberately create supply/demand imbalances. Why? For starters, short-run demand often exceeds the capacity of the supply process. For example, demand for candy in the week preceding Hal-

[1] The notation $D(a, b]$ is used to denote the total demand over the interval of time $(a, b]$, where $a$ $b$. Similarly, $S(a, b]$ is the total supply over the interval $(a, b]$, etc. The difference in the types of brackets indicates that the end-point $a$ is excluded. So for example an order arriving at time $a$ is included in $D(0, a]$ but is not counted in $D(a, b]$. Note by this convention that $D(0, b] = D(0, a] + D(a, b]$, and we avoid double counting the arrival at time $a$.

loween vastly outstrips the weekly capacity of most candy manufactures. To meet this peak demand, manufacturer's begin building inventories months in advance. In this way, they ensure their cumulative output is sufficient to meet cumulative demand.

A second reason for deliberately creating imbalances is to take advantage of changes (either real or anticipated) in the cost of materials or products. For example, grocery chains will "forward buy" large quantities of consumer products when manufactures offer trade discounts (just as New York City commuters buy large quantities of subway tokens prior to a fare increase). In terms of analysis, such speculative purchases are more akin to futures contract trades than to normal operating decisions.

## 3.2 Time and distance

Supply and demand imbalances also occur when the input and output of some portion of the supply chain are separated by time and/or distance. For example, suppose we are supplying an overseas market from a domestic distribution center. Assume we use container ships and the total one-way transportation time is 5 weeks. With great fan-fare, we start up operations, begin shipping product to our overseas market at a rate of 3,000 units per week and hold our breath!

What happens? Well, not much – at least not for the first 5 weeks anyway. If we were to draw the outline of our black box around the overseas transportation link, in the first 5 weeks of operation we would see the input to the box humming along at 3,000 units per week while our overseas partners twiddled their thumbs and waited. In week 6, our overseas partners would receive product shipped in week 1; in the seventh week they would receive product shipped in week 2, etc. In other words, the output is the input delayed by 5 weeks: $D(0,t] = S(0,t-5]$. Hence, $I(t) = S(0,t] - S(0,t-5]$, i.e. the inventory "on the ocean" is simply the cumulative shipments made over the last five weeks. In our case, since we are shipping exactly 3,000 units per week, the inventory is a constant $5 \times 3,000 = 15,000$ units.

In general, in a transportation link with a delay of $l$ time units, the inventory in transit is given by

$$I(t) = S(0,t] - S(0,t-l]. \qquad (2)$$

Such transportation related inventories are called *pipeline stocks*. Of course, transportation is not the only cause of delay between points in a supply chain. Lead times for production, communication and order fulfillment can also introduce significant delays. Anytime such delays are present, they may create inventories. If the delay is due to production, the inventory is typically called a *work-in-process (WIP)* inventory.

An important – and potentially confusing – issue surrounding pipeline stocks concerns the timing of payments. The physical inventory is, of course, always given by (2), but the question remains: Whose inventory is it?

Luckily, we can narrow the answer down to two possibilities: the shipper or the receiver. In some cases, ownership changes hands when the shipment is initiated. From an accounting standpoint, the inventory then belongs to the receiver and becomes an entry in the accounts receivable ledger of the shipper. In other cases, ownership does not change hands until the goods are delivered, in which case the inventory stays on the books of the shipper whilst in-transit. The resulting cost differences can be significant when shipping delays are long.

These terms of trade have been carefully defined and standardized by the International Chamber of Commerce (ICC). [2] These ICC "Incoterms" define ownership, cost, liability and tariff responsibilities between shipper and receiver. For example, FOB (free on board) means the shipper has fulfilled his obligation once the goods have "passed over the ship's rail." In contrast, under FAS (free alongside ship) terms, the shipper's responsibility is complete once the goods have been placed alongside the vessel. (This difference assumes particular significance should someone happen to drop your container whilst loading it onto the ship!) DDP (delivered duty paid), to take another example, means the shipper is responsible for the goods - including paying, freight and duty - up to the point they are delivered to the receiver. And so on.

[2] *Incoterms 2000*, ICC Publishing, New York. ISBN 92-842-1199-9

## 3.3 Economies of scale

Managers may also allow an accumulation of inventory to achieve economies of scale, either in production or transportation. For example, set-up costs in a given stage of production can drive a manufacturer to produce in large lots. If demand does not occur in similarly large lots, supply/demand imbalances – and inventories – are created. That is, a "clump" of supply is added to the inventory which is only slowly depleted by a more-or-less constant demand, followed by the arrival of another "clump" of supply, etc.. These "waves" of inventory, or *cycle stocks*, can be a major contributor to the total inventory of a supply chain.

Transportation economies of scale can also cause cycle stocks. Suppose in our example above each container holds $12,000$ units of our product, and that we want to ship full containers to minimize transportation cost. Since we are producing $3,000$ units per week, it takes 4 weeks to produce enough product to fill one container.

What is the effect on total inventory? Observe that a new container starts empty and over a 4-week period builds up to $12,000$ units, at which point it is shipped and the process of filling a new container begins. It is not hard to see that the average inventory during one of these 4-week cycles is $6,000$ units (half the peak value of $12,000$ units). Thus, a cycle stock of $6,000$ units is added to the pipeline stock of $5 \times 3,000 = 15,000$ units, for a total of $21,000$ units.

## 3.4 Supply/demand uncertainties

A final cause of supply/demand imbalances is uncertainty. We will focus on demand uncertainty since it is arguably the dominant form of uncertainty in most supply chains, but similar ideas apply to supply uncertainty.

Why does demand uncertainty cause a problem? Consider our basic inventory equation (1), and imagine we are uncertain about what the demand process, $D(0,t]$, will be in the future. If we can *continuously* monitor $D(0,t]$ and adjust $S(0,t]$ *instantly* to its variations, then we can still keep the inventory at zero.

The catch, of course, is *continuous* review and *instant* resupply; without both of these components, our ability to exactly match supply to demand is diminished.

For example, suppose we only review and replenish our inventory once a week. (This practice is called *periodic review*, and the elapsed time between ordering is called the *review period*.) Because the demand process is uncertain, the demand on the inventory during the review period is also uncertain. If demand is weaker than expected, we can end the period with excess inventory; if demand runs unexpectedly high, we may end the period with a significant number of back-orders.

Note that periodic review introduces cycle stocks, since we must order, on average, enough product to satisfy average demand in a period. In our example, we would bring a week's worth of stock in at the beginning of the week and it would be depleted (on average) by the end of the week. Because of the demand uncertainty, however, we may want to start the week with more than this average amount of inventory. This excess over the average is called *safety stock*, since it serves as a hedge against uncertain variations in demand.

The presence of a delay (*lead time*) in the supply process has a similar effect. To see why, suppose we continuously monitor the inventory and suddenly notice that a surge of orders has just come in. We begin increasing the supply to make up for the loss, but the new supplies take some time to arrive. Since the supply cannot exactly track demand, the inventory drops (or perhaps back-orders rise). Alternatively, if suddenly demand drops, we may stop ordering. However, past orders in the pipeline will still arrive, causing the inventory to surge.

This tendency to under and over-shoot in our ordering increases as demand becomes more erratic and/or lead times in the supply process lengthen. In effect, an "inertia" is created by large pipeline stocks. Think of it as trying to negotiate the tight curves of a windy road while driving a heavy truck; the sharper and more unpredictable the curves (high demand variability) and the heavier the truck (long lead times), the more difficult it is to stay on the road!

# 4 Inventory cost accounting

In almost any business analysis involving inventory, physical inventory levels must be converted to inventory costs. The exact determination of the cost rate to apply is really a cost accounting matter, but here are the major components:

1. *Capital Cost* - This is usually an internal cost of funds rate multiplied by the value of the product. Because value (materials, labor, transportation, etc.) is added to the product as it moves along the supply chain, this cost tends to increase as product moves downstream.

2. *Storage Cost* - Units in inventory take up physical space, and may incur costs for heating, refrigeration, insurance, etc. An activities based cost (ABC) analysis is usually needed to determine which components of these costs are actually driven by inventory levels and which can be considered more-or-less fixed. The answer will depend on the magnitude of the inventory change you are analyzing.

3. *Obsolescence Cost* - A somewhat harder cost component to pin down is obsolescence cost. A technology or fashion shift may make your current products obsolete and severely deflate their value. The more inventory you have, the higher your exposure to this sort of loss.

4. *Quality Cost* - High levels of inventory usually increase the chance of product damage and create slower feed-back loops between supply chain partners. The result: lower levels of quality and a rise in the myriad costs associated with low quality. Again, these costs are difficult to quantify precisely, but the current consensus is that they can be quite significant.

Typically, all these costs are rolled together into a single *inventory cost rate*, expressed as a percentage of the value of the product or material per unit time (e.g. 20% per year). Other equivalent terms for this same cost rate are *inventory holding cost rate* and *inventory carrying cost rate*.

Needless-to-say, there are a host of problems with this approach. For one, the value of a product is not the sole driver of inventory costs. Other product attributes, such as size, the need for refrigeration, obsolescence risk, etc., determine major components of inventory cost. Applying a single cost rate to all products at all stages of production/distribution can be a gross oversimplification.

Secondly, in relying on an inventory cost rate in an analysis, one is implicitly assuming that only *marginal* changes in inventory will occur. A major structural change in the supply chain may eliminate whole categories of expenses that were considered “fixed” in the original ABC analysis of the inventory cost rate. For example, a major reduction in inventory may eliminate the need for an entire warehouse, the operating cost of which may have been considered fixed when the inventory cost rate was determined.

A more reliable approach is to use the *with-without principle*. That is, analyze the relevant costs under the current system (*without* any change) and then analyze the same costs *with* the proposed change. The difference is a more accurate gauge of cost impact than what one gets by multiplying inventory levels by a marginal cost rate.

But even lower *cost*, itself, may not be the most important benefit of a reduction in inventory. Inventory reductions can free up considerable quantities of cash, which are often critical for a rapidly expanding business or a business on the verge of insolvency. Reductions in inventory also lower the asset base of an operation, providing a higher return on assets (ROA). Therefore, justifying an operating change will depend on a firm’s financial objectives.

The point, quite simply, is that your own business judgment and skill must be applied to accurately capture the right costs and benefits of inventory. Inventory models will only tell you how inventory levels change; you bear the responsibility of translating these physical changes into changes in the relevant financial components.

# 5 Models of inventory cost and service

The preceding discussion highlights some of the complexity involved in understanding the drivers of inventory cost and service. To understand these effects in more detail and to quantify cost/service measures, we need a formal model of the inventory process. Yet real supply chains are quite complex, consisting of an entire network of production and distribution facilities connected by transportation links, each potentially handling thousands of different products. How, then, are we to make sense of all this complexity?

One approach is to develop an integrated model of the entire supply chain network, accounting for all cost and service interrelationships. While such models do exist and are continually being refined by researchers, they are really the domain of operations management specialists.

An alternative approach - which we shall adopt here - is to decompose the problem. That is, we can think of breaking our complex supply chain network into a series of much simpler pieces, each consisting of one source, supplying one destination with one product. For example, we might look at how our domestic plant supplies Product A to our Western distribution center - the plant is the source, the distribution center is the destination and only Product A is considered. We could then look at how the Western distribution center itself supplies Product A to a customer's retail outlet. Then, we could repeat the analysis for Product B, or repeat the analysis for Product A at the Eastern distribution center, etc..

In reality, a host of factors can confound such a simplified approach. Products are often ordered or shipped jointly; customer orders may be positively or negatively correlated, etc. However, decomposing the problem serves as a useful first-cut analysis. It also makes the problem much easier to understand, which in turn helps build valuable intuition.

## 5.1 Types of inventory control policies

There are basic categories of policies for controlling inventories: fixed order quantity policies and fixed time period policies. In the first, as the name implies, the order quantity is always the same but the time between orders will vary depending on demand and the current inventory levels. Specifically, inventory levels are continuously monitored and an order is placed whenever the inventory level drops below a predetermined *reorder point*. For this reason, this type of policy is also called a *continuous review* policy.

In fixed time period policies, the time between orders is constant but the quantity ordered each time varies with demand and the current level of inventory. Because ordering follows a fixed cycle, these policies are also called *periodic-review* policies. We will focus on periodic-review policies in this note, but the major insights and analyses are very similar to those for continuous review policies.

## 5.2 A basic model

In our basic model, shown in Figure 2, an inventory of one product is supplied by a transportation (or production) pipeline that has a constant lead time $l$. $Q(0,t]$ denotes the cumulative quantity ordered up to time $t$, and $D(0,t]$ denotes the cumulative demand on the inventory up to time $t$. Since the lead time is a constant $l$ units, the cumulative supply flowing into the inventory, $S(0,t]$, is simply given by

$$S(0,t] = Q(0,t-l], \tag{3}$$

so by (1)

$$I(t) = Q(0,t-l] - D(0,t] \tag{4}$$

We allow demand to be backordered, so $I(t)$ can be negative. To distinguish it from inventory in the transportation pipeline, we will henceforth refer to $I(t)$ as the *on-hand* inventory.

### 5.2.1 The demand process

Cumulative demand, $D(0,t]$, is assumed to be random and normally distributed with

$$\begin{aligned} E(D(0,t]) &= \lambda t && (5)\\ \mathrm{Var}(D(0,t]) &= \sigma^2 t. && (6) \end{aligned}$$

This model is usually a reasonable approximation in practice, since we can often view cumulative demand as the sum of demands from a large number of smaller time periods (e.g. monthly demand is the sum of 30 daily demands). Then, regardless of the distribution of the smaller demands, the sum is approximately

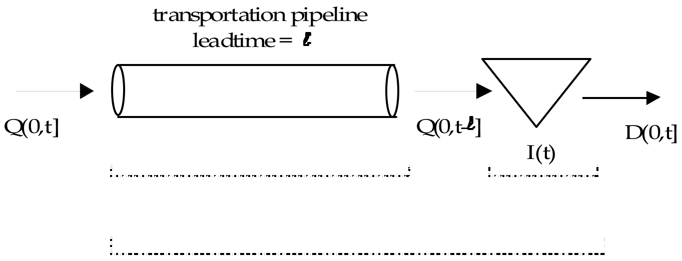

Figure 1: Diagram of Basic Inventory Model

normally distributed. In cases where this normal approximation is not perfect, it still serves as a useful base case for analysis.

Note $\lambda$ is the average demand in a unit of time and $\sigma^2$ is the variance of demand in a unit of time. These values depend, of course, on the units you choose to measure time. For example, if we are measuring time in days, then $\lambda$ and $\sigma^2$ are, respectively, the mean and variance of daily demand; if time is measured in weeks, then $\lambda$ and $\sigma^2$ are, respectively, the mean and variance of weekly demand. The unit of time you use does not really matter as long as you are consistent. However - a word to the wise - using inconsistent units is a *very* common mistake.

### 5.2.2 The ordering process and inventory position

To proceed further, we need a model of how ordering takes place in response to demand - an ordering *policy*. Ideally, we would like a policy which is optimal in some sense. However, usually such policies are quite complex. Instead, we shall settle for a simple ordering policy which is "nearly optimal". To define the policy, however, we need to introduce a new concept - *inventory position*.

At first glance, it might seem our ordering decision should be driven by the on-hand inventory level, $I(t)$, alone. However, as in life, in inventory management it pays to think ahead - in this case we need to think one lead-time ahead. The reason is that future inventory levels are affected by both the on-hand inventory *and* the orders in the pipeline that are due to arrive. Inventory position captures this idea.

The *inventory position* , $P(t)$, is defined by

$$P(t) = I(t) + Q(t-l, t] \tag{7}$$

where $Q(t-l,t] = Q(0,t] - Q(0,t-l]$ is the total quantity ordered in the last $l$ units of time, which is the total quantity in the pipeline that is due to arrive by time $t+l$, Hence, inventory position is the total amount *on-hand* plus the total amount *on-order* (i.e. in the pipeline). Figure 1 shows the relationship of inventory position to the pipeline and on-hand inventory. Keeping a careful eye on the inventory position reduces the nasty tendency to overshoot and undershoot the inventory that lead times typically engender.

An important fact to recognize is that we can regulate the inventory position by simply placing an order or by holding back orders. In other words, *the inventory position is controllable.*

Unfortunately, we do not exercise the same degree of control over the on-hand inventory level. To see why, consider the on-hand inventory one lead-time ahead (still assuming $I(0) = 0$):

$$\begin{aligned} I(t+l) &= S(0,t+l] - D(0,t+l] \\ &= S(t,t+l] - D(t,t+l] + S(0,t] - D(0,t] \\ &= Q(t-1,t] - D(t,t+l] + I(t) \\ &= P(t) - D(t,t+l]. \end{aligned} \tag{8}$$

So the on-hand inventory one lead-time in the future is the current inventory position minus the demand between now and time $t + l$. The term, $D(t, t + l]$, is called the *lead time demand*. Note that while $P(t)$ is completely under our control, we have absolutely no control over the lead time demand $D(t, t + l]$. All we can do is try to control the inventory position, $P(t)$, so that the on-hand inventory, $I(t)$, behaves “reasonably”.

### 5.2.3 The periodic-review, order-up-to policy

We are now in a position to define our policy. Here it goes: We review the inventory position every $p$ units of time. At these review points, we place an order sufficient to bring the inventory position up to a fixed level $S$. (The quantity $S$ is called the *base stock* level.) We then repeat the process at the next review point, etc... That’s it!

For obvious reasons, this policy is called a *periodic-review, order-up-to policy* and, for appropriate choices of $p$ and $S$, it is a close-to-optimal way to order. In many cases $p$ is given - or at least implicitly given. For example, a firm may already have a fixed order cycle, driven perhaps by the need for a regular schedule of deliveries or by a fixed production sequence at a supplier’s plant. Alternatively, if there is a fixed ordering cost then $p$ may be chosen to balance ordering cost and holding cost using the economic order quantity (EOQ) formula. We will see how to chose $S$ shortly. For now, let’s focus on how the inventory behaves under this policy.

Figure 2 shows a sample graph of the inventory position and the on-hand inventory for a periodic-review, order-up-to policy with a lead time of $l = 3$, review period $p = 10$ and base stock level $S = 40$. In the figure, we start at time $t = 0$ with no inventory in the system and immediately place an order of size 40. Note that at every order point, the inventory position is restored exactly to $S = 40$. The on-hand inventory, on the other hand, receives no replenishments for the first $l = 3$ units of time, and thereafter receives replenishments every $p = 10$ time units as well. Notice, however, that the on-hand inventory is not restored to the same level after every replenishment.

It is important to recognize that this policy, although certainly a reasonable way to manage inventory, is really just a *model* of how ordering takes place. What we really care about in the final analysis are the cost and service characteristics of the inventory system. Provided the real inventory policy is reasonably close to the model, however, the service measures developed below will still provide good estimates of cost and service.

In a way, you can think of this policy as a model of rational operational behavior on the part of a firm in response to its specific physical constraints and service objectives. No doubt a firm can always do worse, but they would have a hard time doing significantly better than using a periodic-review, order-up-to policy given the physical and financial constraints at hand. As one operations manager put it: “The difference between intelligence and stupidity is that there are some limits on intelligence.” A periodic-review, order-up-to policy is, for our model, approximately the limit on intelligence.

Even under this simple policy, an exact expression for the on-hand inventory level is hard to come by. However, there exist some quite accurate approximations to various inventory costs and service measures, which we look at shortly.

## 5.3 Choosing the policy parameters

How do we choose a “good” basestock level $S$? To answer this question, note that $(p+l)\lambda$ is the average demand experienced over a review period plus a lead time. $S$ should be roughly this large. Why? Well, recall that inventory position is the total material in the system (on-hand plus on-order), and we start each review period with the same inventory position $S$. We must then wait $p$ units of time to review the inventory position again, at which time we may want to place an order. However, that order will take another $l$ units of time to arrive. Hence, the inventory in the system at the start of a period should be large enough to cover the average demand over these two intervals of time, or about $(p + l)\lambda$.

But demand is not constant. Indeed, from (6) we see its standard deviation is $\sigma\sqrt{p + l}$. Therefore, we may want to keep a little more inventory to hedge against running out before we get a chance to reorder. This additional inventory is called *safetly stock*.

It is convenient to measure safety stock in terms

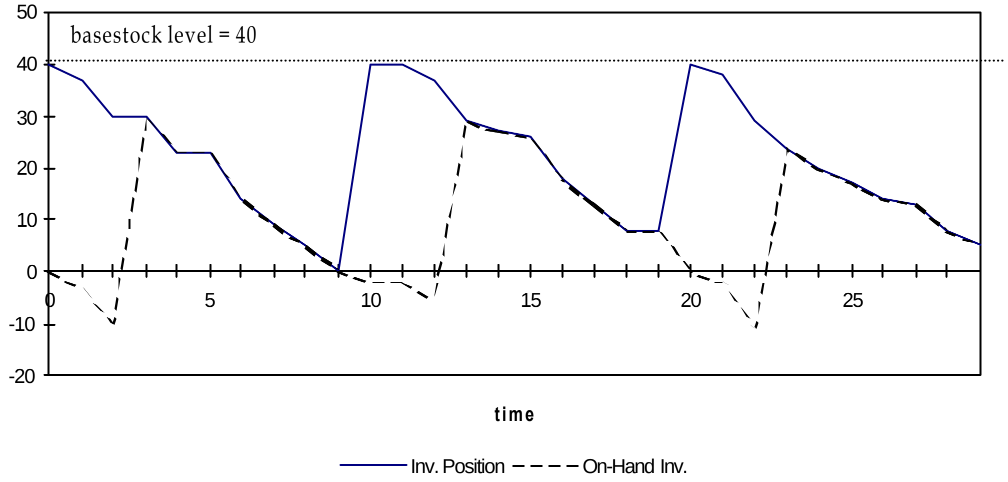

Figure 2: Examples of inventory position and on-hand inventory under a periodic-review, order-up-to policy ($l = 3$, $p = 10$ and $S = 40$)

of the number of standard deviations of demand, $z$. That is,

$$S = (p + l)\lambda + z\sigma\sqrt{p + l}, \tag{9}$$

or equivalently,

$$z = \frac{S - (p + l)\lambda}{\sigma\sqrt{p + l}}. \tag{10}$$

The more cautious we are about stockouts, the higher a $z$ value we choose. Of course, a higher $z$ means higher inventory as well. The next section looks at determining an appropriate value for $z$ to meet a specified level of customer service.

## 5.4 Cost and service measures

### Order fill rate

The primary measure of service in most inventory systems is the *fill rate*, $f$, defined as the fraction of units ordered that can be filled directly from stock. (One minus the fill rate is the fraction of units backordered.) The fill rate may be determined by balancing the costs of holding additional inventory against the cost of stocking out, if known. More commonly, it is obtained by a subjective judgement and/or competitive benchmarking.

After some analysis (the details of which we shall not delve into here) one can obtain the following estimate of the fill rate, $f$:

$$f = 1 - \frac{\sigma\sqrt{p + l}L(z)}{\lambda p}, \tag{11}$$

where $z$ is defined in (10) as before. The function $L(z)$ is called the *standard loss function*, which is used in the process of determining the average number of back orders in a cycle. The Appendix contains a table of values for the function $L(z)$. Rearranging the above relation, we obtain

$$L(z) = \frac{(1 - f)\lambda p}{\sigma\sqrt{p + l}}. \tag{12}$$

Given a target fill rate $f$, one can use (12) to determine $L(z)$ and then use the table in the Appendix to determine the resulting $z$ value. (Examples of such calculations are given at the end of the note.)

This estimate of fill rate (11) is accurate provided the level of backorders is not too high. In other words, it is a good estimate for high-fill-rate scenarios, which is usually the case we are interested in.

### Average inventory levels

Once $z$ is determined, we can evaluate the average level of inventory at various stages. The average on-hand inventory level is given by, [3]

$$\text{Avg. On-Hand Inv.} = \frac{\lambda p}{2} + z\sigma\sqrt{p+l}. \tag{13}$$

The two terms on the right-hand-side have special significance. The first is the cycle stock:

$$\text{Cycle Stock} = \frac{\lambda p}{2} \tag{14}$$

Because we order periodically, orders get "bunched up". As shown in the next section, the average order size is $\lambda p$ units. Hence, a long review period $p$ implies a large average order size. These "waves" of inventory have a peak height of about $\lambda p$ (the typical order size), so their average height is approximately $\lambda p/2$.

The second term on the right-hand-side in (13) is the *safety stock*:

$$\text{Safety Stock} = z\sigma\sqrt{p+l}. \tag{15}$$

You can think of this as an extra "layer" of inventory that is added underneath the fluctuating cycle stocks. Its purpose is to hedge against unusually "deep" cycles, i.e. cycles during which the demand exceeds its expected value. The higher $z$ is, the thicker the layer of safety stock and the less likely stockouts become.

The inventory in the pipeline is given by

$$\text{Pipeline Stock} = \lambda l. \tag{16}$$

This inventory is "in transit" to our facility and is not available to customers; hence, it does not affect service levels. Moreover, notice that it is completely independent of the policy parameters $p$, $S$ and $z$. *The ordering policy simply has no affect on this inventory!*

Yet it still matters. Depending on the trading terms, pipeline inventory impacts total inventory cost, and hence it often must be included in a supply chain analysis. It will also change if there is a change in lead time, so any analysis which involves a structural change in the supply pipeline (e.g. a sourcing change, change in the mode of transportation, etc.) or change in trading terms (e.g. FOB to DDP) should factor in the resulting change in pipeline inventory costs.

### Order frequency and size

Some components of cost are driven by the frequency and size of orders. Under our periodic-review, order-up-to policy, we have

$$\text{Order Frequency} = \frac{1}{p} \tag{17}$$

and

$$\text{Average Order size} = \lambda p. \tag{18}$$

The reason for (17) is straightforward and (18) follows by conservation of material flow. That is, the average flow into the inventory must equal the average flow out over a long period of time, otherwise the inventory will steadily drift down or steadily build up forever. Since we order at a frequency of $1/p$, the average quantity ordered ($\lambda p$) times the order frequency ($1/p$) must equal the average demand rate of $\lambda$.

Both of these quantities may be significant drivers of cost. In times past, there was a significant amount of clerical work involved in processing an order. Such costs are not driven by the amount ordered (It takes only seconds more of a clerks time to add a few more zeros to the order quantity!), but rather by the frequency of orders - the order volume. Today, such costs are greatly diminished with the advent of computers and electronic trading using electronic data interchange (EDI) and electronic funds transfer (EFT). Even so, in some industries ordering costs are important.

To the extent that there are scale economies in procurement and/or transportation, cost is also affected by the order size. For example, unit transportation costs for full truck loads of product are usually substantially lower than for partial truck loads. If we order less frequently but in higher volumes, we may be able to achieve significant transportation savings. Supplier volume discounts or fixed manufacturing setup costs create similar economies of scale.

All these costs will be driven by our decision about the length of the review period $p$; the stronger the scale economies and the higher the ordering costs, the more likely we are to prefer long review periods and large order sizes.

[3] By average in this case we mean the time average value of $I(t)$ measured over a very long interval of time. The formal definition is $\lim_{h\to\infty}(1/h)\int_0^h I(t)dt$.

# 6 Examples

## 6.1 Calculating the inventory needed to meet a target fill rate

Consider the following (hypothetical) scenario: A university computer store orders from Apple computer once every two weeks. The lead time from Apple is 3 weeks. Assume weekly demand for Model X11 has a mean of $\lambda = 1.5$ and a variance of $\sigma^2 = 4.0$. The store would like to maintain a fill rate of 95% on its Apple products. For Model X11, what base-stock level should the store use and how much on-hand inventory is required to meet this service level?

To determine the answer, we see the review period $p = 2$ weeks, the lead time $l = 3$ weeks and we want a fill rate of $f = 0.95$. Therefore, using (11)

$$\begin{aligned} 0.95 &= 1 - \frac{\sigma\sqrt{p+l}L(z)}{\lambda p} \\ &= 1 - \frac{2 \times \sqrt{2+3} \times L(z)}{1.5 \times 2} \\ &= 1 - 1.49 \times L(z) \end{aligned}$$

Therefore, we must have

$$L(z) = \frac{0.05}{1.49} = 0.0335.$$

From Table 2, we see that a value between $z = 1.4$ and $z = 1.5$ should do the job. We'll use $z = 1.45$ as an approximation. Substituting this value of $z$ into (9), the required base-stock level is

$$S = 1.5 \times (2+3) + 1.45 \times 2 \times \sqrt{2+3} = 13.98$$

(which we would no doubt round up to 14), and, using (13), the average on-hand inventory is

$$\frac{1.5 \times 2}{2} + 1.45 \times 2 \times \sqrt{2+3} = 7.98.$$

This consists of a cycle stock of 1.5 units and a safety stock of 6.48 units.

## 6.2 Determining service provided by a given inventory level

Consider the same situation above, but suppose we do not know the base-stock level $S$. However, we know the store uses a periodic-review, order-up-to policy, and from company data we have determined that the average on-hand inventory is 12.6 units. What fill rate is the store providing?

In this case, we back up the calculation. Indeed, using (13) as above with $z$ unknown gives

$$12.6 = \frac{1.5 \times 2}{2} + z \times 2 \times \sqrt{2+3}.$$

Solving for $z$, we obtain

$$z = 2.48.$$

Rounding up to $z = 2.5$ and looking at Table 2, we see $L(2.5) = 0.0020$. Therefore, by (11)

$$\begin{aligned} f &= 1 - \frac{\sigma\sqrt{p+l}L(z)}{\lambda p} \\ &= 1 - \frac{2 \times \sqrt{2+3} \times 0.002}{1.5 \times 2} \\ &= 1 - 0.003 \\ &= 0.997 \end{aligned}$$

So the fill rate provided by this level of inventory is approximately 99.7%.

## 6.3 Determining policy parameters

Now suppose the store wants to reevaluate the frequency with which it places orders with a view toward minimizing its total cost. It wants to retain the target fill rate of 95%. Apple charges a fixed fee of \$25 for shipping and handling on each order, regardless of the size of the order. The university also determines that its cost to process an order is \$15. The Model X11 has a wholesale price of \$3,000 and the university's holding cost rate is estimated at 20% per year.

In this case, we will first compute the EOQ based on the parameters. The holding cost rate per week for the Model X11 is

$$h = \frac{\$3,000 \times 0.20}{52\text{weeks/year}} = \$11.5.$$

The fixed cost per order is $s = \$25 + \$15 = \$40$. Therefore, from the EOQ is

$$q^* = \sqrt{\frac{2\lambda s}{h}} = \sqrt{\frac{2 \times 1.5 \times 40}{11.5}} = 3.2.$$

Since, $p = q/\lambda$ this implies an ideal order period of

$$p^* = \frac{q^*}{\lambda} = \frac{3.2}{1.5} = 2.15.$$

This is very close to the current order period of two weeks. Therefore, a two week reorder period and a base stock level of $S = 14$, which we know from the first example is the base stock level that is needed to provide a 95% fill rate, are close to optimal for these ordering and holding costs.

# 7 Appendix

## 7.1 Partial expectation, the standard loss function and average backorders (optional)

Partial expectation is a new twist on an old statistical idea. Recall, the expected value of a random quantity (random variable) is the theoretical, long-run average of the quantity. For example, if we were to roll a fair die a large number of times and record the values as we went along, the average of these values would approach 3.5. (Each integer $1, 2, ..6$ is equally likely, so the average is just halfway between 1 and 6.) If $X$ represents the value of one roll of the die, then we would say $E(X) = 3.5$.

Partial expectation answers a somewhat more complex question: On average, how much *above a given threshold value* $z$ is the random quantity? For example, suppose the threshold $z = 2.5$ and we want to know, on average, by how much our die value exceed 2.5.

Table 1 shows how we would compute such a quantity empirically. The second column shows the value $X$ of each roll of the die; the third column shows the value of $X$ minus the threshold $z = 2.5$; the last column shows the excess over $z = 2.5$, denoted $(X - z)^+$ (i.e. the positive part of the difference: $X - z$). The partial expectation is the long-run average of this last column of values, which is denoted $E(X - z)^+$.

Why is this quantity so important? In our inventory system, we frequently need to compute the average value of the inventory above zero (physical inventory) and the average value of the inventory below zero (backorders). Partial expectation is the tool that allows us to do this.

| Trial | $X$ | $X - 2.5$ | $(X - 2.5)^+$ |
|---|---|---|---|
| 1 | 3 | 0.5 | 0.5 |
| 2 | 1 | -1.5 | 0.0 |
| 3 | 5 | 2.5 | 2.5 |
| 4 | 3 | 0.5 | 0.5 |
| 5 | 6 | 3.5 | 3.5 |
| 6 | 4 | 1.5 | 1.5 |
| 7 | 2 | -0.5 | 0.0 |
| 8 | 4 | 1.5 | 1.5 |
| 9 | 2 | -0.5 | 0 |
| 10 | 3 | 0.5 | 0.5 |

Table 1: Example Inputs in Computing Partial Expectations

Recall in our inventory system we start each order period with an inventory position of $S$. Since we then wait $p$ units of time before reordering and then an additional $l$ units of time before this reordered quantity comes in, the inventory position of $S$ must cover the demand over a period of length $p + l$. (Note that all $S$ units of the on-hand and on-order inventory will be "flushed out" by the end of this period of time.)

Let $D(0, p + l]$ denote the demand over the period $p + l$. The on-hand inventory will be negative if $D(0, p + l] > S$, and therefore the total amount backordered is $(D(0, p + l] - S)^+$. Since this process is repeated once every order period, it follows that $E(D(0, p + l] - S)^+$ is the average quantity backordered each order period.

How do we compute $E(D(0, p+l] - S)^+$? If $Z$ is a standard normal random variable - i.e. a random variable that is normally distributed with a mean of zero and a variance of one - then its partial expectation, denoted $L(z)$, goes by a special name: *the standard loss function.* That is,

$$L(z) = E(Z - z)^+.$$

Table 2 provides values for $L(z)$ as well as the cumulative distribution of $Z$, denoted $F(z) = P(Z \quad z)$.

With this table in hand, it is easy to compute the partial expectation of a normally distributed random variable with any mean and variance. Indeed, if $X$ is normally distributed with mean $\lambda$ and variance $\sigma^2$, then

$$E(X-x)^+ = \sigma L(z)$$

where

$$z = \frac{x-\lambda}{\sigma}.$$

To see this, note that

$$\begin{aligned} E(X-x)^+ &= E((X-\lambda)-(x-\lambda))^+ \\ &= \sigma E\left(\frac{X-\lambda}{\sigma} - \frac{x-\lambda}{\sigma}\right)^+ \\ &= \sigma E(Z-z)^+ \\ &= \sigma L(z) \end{aligned}$$

We are now almost home. Since $D(0,p+l]$ is normally distributed with mean $(p+l)\lambda$ and variance $\sigma^2(p+l)$ (see (5) and (6)), then

$$E(D(0,p+l]-S)^+ = \sigma\sqrt{p+l}\, L(z)$$

where $z$ is given by (10). Hence, the average number of back orders in a cycle is $\sigma\sqrt{p+l}\, L(z)$ and, since we order on average $p\lambda$ units per cycle, taking a simple ratio gives us the fill rate equation (11).

## 7.2 Extension to random lead times (optional)

In many cases lead times are not constant. For example, variabilities in production throughput times and transportation times (e.g. weather/handling delays) can cause order lead times from a supplier to vary.

To extend the basic model to the case where lead times are variable, let us suppose that the actual lead time is a random variable, $L$, with

$$E(L) = l$$

and variance $\mathrm{Var}(L)$. The same basic reasoning as before works. That is, we want to look ahead one lead time and manage the inventory position. However, now the demand over (the random) lead time $L$, $D(t,t+L]$, is more complicated.

We have the same average lead time demand of

$$E(D(t,t+L]) = \lambda E(L) = \lambda l.$$

However, it turns out that now the variance of the lead time demand, which we will denote $\sigma^2_{LD}$, is given by

$$\sigma^2_{LD} = \mathrm{Var}(D(t,t+L]) = \sigma^2 l + \lambda^2 \mathrm{Var}(L).$$

In our old case with constant lead times, $\mathrm{Var}(L) = 0$, and thus $\sigma^2_{LD} = \sigma^2 l$. Note the presence of variability in the lead times increases the variance of the lead time demand, as one might have expected.

With this new characterization of lead time demand, the analysis proceeds as before, substituting $\sigma^2_{LD}$ for $\sigma^2 l$. For example, the expression for $z$ becomes

$$z = \frac{S-(p+l)\lambda}{\sqrt{\sigma^2_{LD}+\sigma^2 p}}$$

or equivalently, expressing $S$ in terms of $z$,

$$S = (p+l)\lambda + z\sqrt{\sigma^2_{LD}+\sigma^2 p}.$$

The fill rate is then

$$f = 1 - \frac{\sqrt{\sigma^2_{LD}+\sigma^2 p}\, L(z)}{\lambda p},$$

and so on.

Because the variance of lead time demand is higher when both demand and lead times vary, both the $z$ value and the safety stock will need to be higher to meet a given fill rate, $f$, compared to the case of constant lead times. This is intuitive; lead time variability only makes it harder (and more costly) to meet service level targets. The above extension lets you quantify exactly how much cost is added by lead time variabilities.

| $z$ | $F(z)$ | $L(z)$ | $z$ | $F(z)$ | $L(z)$ |
|---|---|---|---|---|---|
| -4.0 | 0.0000 | 4.0000 | 0.0 | 0.5000 | 0.3989 |
| -3.9 | 0.0000 | 3.9000 | 0.1 | 0.5398 | 0.3509 |
| -3.8 | 0.0001 | 3.8000 | 0.2 | 0.5793 | 0.3069 |
| -3.7 | 0.0001 | 3.7000 | 0.3 | 0.6179 | 0.2668 |
| -3.6 | 0.0002 | 3.6000 | 0.4 | 0.6554 | 0.2304 |
| -3.5 | 0.0002 | 3.5001 | 0.5 | 0.6915 | 0.1978 |
| -3.4 | 0.0003 | 3.4001 | 0.6 | 0.7257 | 0.1687 |
| -3.3 | 0.0005 | 3.3001 | 0.7 | 0.7580 | 0.1429 |
| -3.2 | 0.0007 | 3.2002 | 0.8 | 0.7881 | 0.1202 |
| -3.1 | 0.0010 | 3.1003 | 0.9 | 0.8159 | 0.1004 |
| -3.0 | 0.0013 | 3.0004 | 1.0 | 0.8413 | 0.0833 |
| -2.9 | 0.0019 | 2.9005 | 1.1 | 0.8643 | 0.0686 |
| -2.8 | 0.0026 | 2.8008 | 1.2 | 0.8849 | 0.0561 |
| -2.7 | 0.0035 | 2.7011 | 1.3 | 0.9032 | 0.0455 |
| -2.6 | 0.0047 | 2.6015 | 1.4 | 0.9192 | 0.0367 |
| -2.5 | 0.0062 | 2.5020 | 1.5 | 0.9332 | 0.0293 |
| -2.4 | 0.0082 | 2.4027 | 1.6 | 0.9452 | 0.0232 |
| -2.3 | 0.0107 | 2.3037 | 1.7 | 0.9554 | 0.0183 |
| -2.2 | 0.0139 | 2.2049 | 1.8 | 0.9641 | 0.0143 |
| -2.1 | 0.0179 | 2.1065 | 1.9 | 0.9713 | 0.0111 |
| -2.0 | 0.0228 | 2.0085 | 2.0 | 0.9772 | 0.0085 |
| -1.9 | 0.0287 | 1.9111 | 2.1 | 0.9821 | 0.0065 |
| -1.8 | 0.0359 | 1.8143 | 2.2 | 0.9861 | 0.0049 |
| -1.7 | 0.0446 | 1.7183 | 2.3 | 0.9893 | 0.0037 |
| -1.6 | 0.0548 | 1.6232 | 2.4 | 0.9918 | 0.0027 |
| -1.5 | 0.0668 | 1.5293 | 2.5 | 0.9938 | 0.0020 |
| -1.4 | 0.0808 | 1.4367 | 2.6 | 0.9953 | 0.0015 |
| -1.3 | 0.0968 | 1.3455 | 2.7 | 0.9965 | 0.0011 |
| -1.2 | 0.1151 | 1.2561 | 2.8 | 0.9974 | 0.0008 |
| -1.1 | 0.1357 | 1.1686 | 2.9 | 0.9981 | 0.0005 |
| -1.0 | 0.1587 | 1.0833 | 3.0 | 0.9987 | 0.0004 |
| -0.9 | 0.1841 | 1.0004 | 3.1 | 0.9990 | 0.0003 |
| -0.8 | 0.2119 | 0.9202 | 3.2 | 0.9993 | 0.0002 |
| -0.7 | 0.2420 | 0.8429 | 3.3 | 0.9995 | 0.0001 |
| -0.6 | 0.2743 | 0.7687 | 3.4 | 0.9997 | 0.0001 |
| -0.5 | 0.3085 | 0.6978 | 3.5 | 0.9998 | 0.0001 |
| -0.4 | 0.3446 | 0.6304 | 3.6 | 0.9998 | 0.0000 |
| -0.3 | 0.3821 | 0.5668 | 3.7 | 0.9999 | 0.0000 |
| -0.2 | 0.4207 | 0.5069 | 3.8 | 0.9999 | 0.0000 |
| -0.1 | 0.4602 | 0.4509 | 3.9 | 1.0000 | 0.0000 |
| 0.0 | 0.5000 | 0.3989 | 4.0 | 1.0000 | 0.0000 |

Table 2: Standard normal distribution and standard loss function table