XJTU-IE, 2007.9-2008.9, M: xuchen.xuchen@163.com(徐琛), lucifer_tcl@126.com(李慰祖)

第二章 库存控制：从 EOQ 到 ROP

当你的药吃到只剩四片的时候
就要再次购买了
——佚名 选自 Hadley 和 Whitin（1963）

2.1 引言

科学管理（Scientific management）的诞生使得运作管理（OM）这一现代学科的建立成为了可能。科学管理不仅使得管理学成为一个值得研究的学科，还因为它注重数量上的精确，从而使数学第一次成为一种管理工具。泰勒原初的工作公式是后来许多数学模型的先驱，设计这些数学模型是用来帮助工厂设计和控制人员在各个层面上进行决策的。这些模型都成为了商学和工学课程中的标准科目，整个学术研究的各个学科都是围绕着几个运作管理问题领域而纷纷建立起来的，包括库存控制、生产排程、能力计划、需求预测、质量控制和设备维护。这些模型以及促使这些模型建立的运作管理的焦点问题，现在已是商学的标准语言的一部分了。

在那些产生数学模型的诸多运作管理学科分支中，对工厂管理来说没有什么比库存控制更为核心的了，对运作管理来说也没有什么比库存控制更能体现美国式方法的了。在这一章中，我们来追溯美国库存控制所使用的数学模型方法的历史。我们这样做的原因有以下这么几个：

1. 我们所要讨论的库存控制模型是运作管理领域里最早的成果之一，并且现在仍被广泛的使用和借鉴。同时，它们是制造管理语言中的基本组成部分。
2. 库存几乎在所有的制造系统的物流工作中扮演着关键性的角色。这些历史上的模型中介绍的概念会在本书第二篇工厂物理学和第十七章库存管理中再次出现。
3. 这些经典库存理论的结果是其他许多现代制造管理技术的核心，例如物料需求计划（MRP）、精益生产（JIT）以及基于时间的竞争（TBC），它们因此将作为本书第一篇剩余章节的重要基础。（48|49）

我们从最早最简单的模型——经济订货批量（EOQ）模型开始，逐渐延伸到更为复杂的再订货点（ROP）模型。对于每个模型，我们都给出了一个启发性的例子，一个关于其发展的介绍和关于它所蕴含思想的讨论。

2.2 经济订货批量模型

对工厂管理来说，最早的运用到数学的是 Ford W. Harris（1913）对制造批量设置问题的研究。尽管这篇文章本身显然是被错误地引用了很多年（见 Erlenkotter 1989，1990），但是 Harris 的经济批量模型已经被广泛地研究，并且这几乎成为每一本介绍性质的生产与运作管理课本的必选素材。

2.2.1 动因（Motivation）

我们从一家叫做 MedEquip 的小型制造商的情况出发，这家企业生产的是手术室监视器和诊断设备，它的生产方式是通过在标准金属架上组装电子原件来生产各种最终产品。这些金属架是从一个当地的金属加工厂购买的，当每次需要生产一批架子时都必须布置一次设备（冲压机，加工中心和焊接中心）。由于每次布置工站都会浪费时间，因此如果这个加工厂一次大量采购这些架子，那么就可以更便宜地进行生产（和销售）。然而，因为 MedEquip 不想花太多宝贵的现金在那些金属架库存上，所以它并不愿意大批地购买。

这种矛盾正是 Harris 在他的文章《每次生产多少零件》里所研究的。他这样写道：

与工资、原材料成本和日常管理费用密切相关的资本利润率决定了生产零件的（可获利的）最大批量；加工的准备成本则决定生产的最小批量。管理者可以通过经验来确定经济批量的大小。（Harris 1913）

Harris 考虑的问题是关于一个生产多种产品并且必须承担高昂准备成本的工厂。作为一个例子，他描述了一家生产铜连接器的金属加工厂。每次这家工厂切换生产的连接器类型时，机器都必须重新校准，此外还有各种必须完成的文职工作，并且还可能要浪费一些原材料（例如在调试阶段被用于生产测试件的铜）。Harris 把准备好生产一种产品所必需的人工和原材料成本的总和定义为准备成本。（注意如果连接器是外购而不是自制的话，那么这个问题还是类似的，只是准备成本相应变成了订单采购成本。）

在 MedEuip 的例子和 Harris 的铜连接器案例中，基本的权衡是一样的。大批量生产因需要较少的生产切换而减少了准备成本。而小批量生产通过更加及时的生产来减少库存。经济批量模型是 Harris 在这两个关注焦点之间找到的一种平衡的系统方法。（49|50）

2.2.2 模型

尽管Harris声称EOQ是从实际经验出发的，但他还是脱离不了他那个重视用精确数学方法进行工厂管理的时代背景。为了得到一个计算批量大小的规则，他对制造系统做了如下假设：[1]

1. 生产是瞬间完成的。没有能力约束，并且整个批量是被同时生产出来的。
2. 运输是即时的。在生产和（可以）满足需求之间没有时间延迟。
3. 需求是确定的。需求的大小和时间都不存在变动性。
4. 需求在时间上是常量。事实上，它可以用一条直线表示，因此如果年需求是每年 365 单位的话，那么转化为日需求就是每天 1 个单位。
5. 每次生产切换产生一个固定的准备成本。不管批量是多大或者工厂处于什么状态，准备成本是相同的。
6. 各种产品都可被单独分析。要么是只生产单一产品，要么就是生产的产品之间没有相互影响（例如共用一台设备）。

在这些假设下，我们可以使用 Harris 的符号（为了表达简便做了少许更改）来构建计算最优生产批量的经济批量模型。所需要用到的符号如下：

[1] 读者需要谨记所有的模型都是基于各种简化的假设的。现实世界太过复杂而不能直接分析。好的建模假设就是那些在捕获现实问题的本质时能够使分析变得容易的条件。为了允许读者自己估量他们的合理性，我们将清楚地列示下面我们所讨论模型的假设。

D = 需求率（单位每年）

c = 单位生产成本，没有计算准备成本和库存成本（美元每单位）

A = 生产（采购）一批产品的固定准备（采购）成本（美元）

h = 持有成本（美元每单位每年）；如果持有成本完全由在库存中被占用的资金的利率决定，那么 $h = ic$，i 是年利率

Q = 批量大小（单位），是决策变量

为了便于模型化，Harris 把时间和产品都看成连续量。因为他假设的确定不变的需求，每当库存量达到零时就订购 Q 个单位，这样平均的库存水平就是 $Q/2$（如图 2.1）。相应的，与库存相关的持有成本就是每年 $hQ/2$ 美元。准备成本每单为 A 元，或者可以表示为每年 AD/Q 美元，这是因为我们每年必须下 D/Q 个订单来满足需求。生产成本是每单位 c 美元，或者每年 cD 美元。因而，每年的总（库存、准备和生产）成本可以表示为（50|51）

$$Y(Q) = \frac{hQ}{2} + \frac{AD}{Q} + cD \qquad (2.1)$$

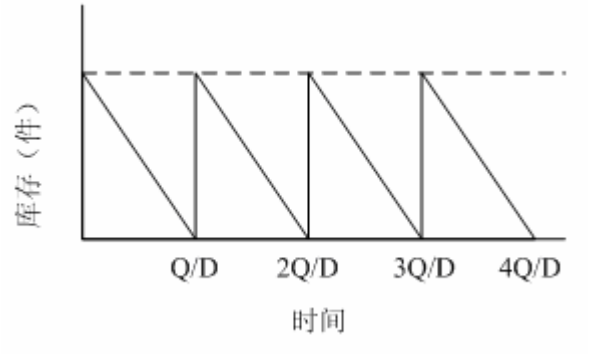

图 2.1 *EOQ* 模型中的库存与时间关系

示例：

为了阐明 $Y(Q)$ 的性质，让我们回到 MedEquip 的例子。假设它对金属架的需求是稳定并可预测的，每年预计数量为 D =1000 件。金属架的单位成本是 c =250 美元，但每个订单都要花费一个固定成本 A =500 美元，这是为下一班生产之前布置设备而停机所引起的成本。MedEquip 公司估计资金的机会成本或者叫做必要收益率（hurdle rate）为每年 10%，贮藏一件金属架所需要的仓储空间每年大约需要 10 美元的费用。因此，每个金属架的年持有成本为 h =（0.1）（250）+10 = 35 美元。把这些值代入式（2.1）可生成图 2.2 中所示的曲线。

我们通过图 2.2 可以对成本函数 $Y(Q)$ 进行观察而得到以下几个方面的结论：

1. 持有成本 hQ/D 的大小随批量 Q 线性增加，最终当 Q 很大时它会成为年总成本的主要组成部分。

2. 准备成本 AD/Q 的大小随 Q 的增大急剧变小，图中显示当批量最初开始增长时将会显著减少准备成本，但减少速率是随批量的增大而迅速减小的。

3. 单位成本 cD 不受批量的影响，因为它的算式里不包含 Q。

4. 年总成本 $Y(Q)$ 在批量 Q 取某个特定值时达到最小值。有趣的是，这个最小值所对应的 Q 值恰好使得持有成本和准备成本相等（即，持有成本和准备成本曲线的交点）。（51|52）

Harris 在他的文章里写到要找到使 Y（Q）最小的 Q 值涉及到“高等数学”的知识，并且他在没有进一步推导的情况下简单地给出了答案。他所提到的（微积分）数学在今天看来似乎也并不算多么高等，所以我们在下面的技术性注释里填补了一些他所忽略的细节。如果对这些细节不感兴趣的话可以跳过这部分，跳过本书中的技术性注释不会对本书的连贯性造成任何损失。

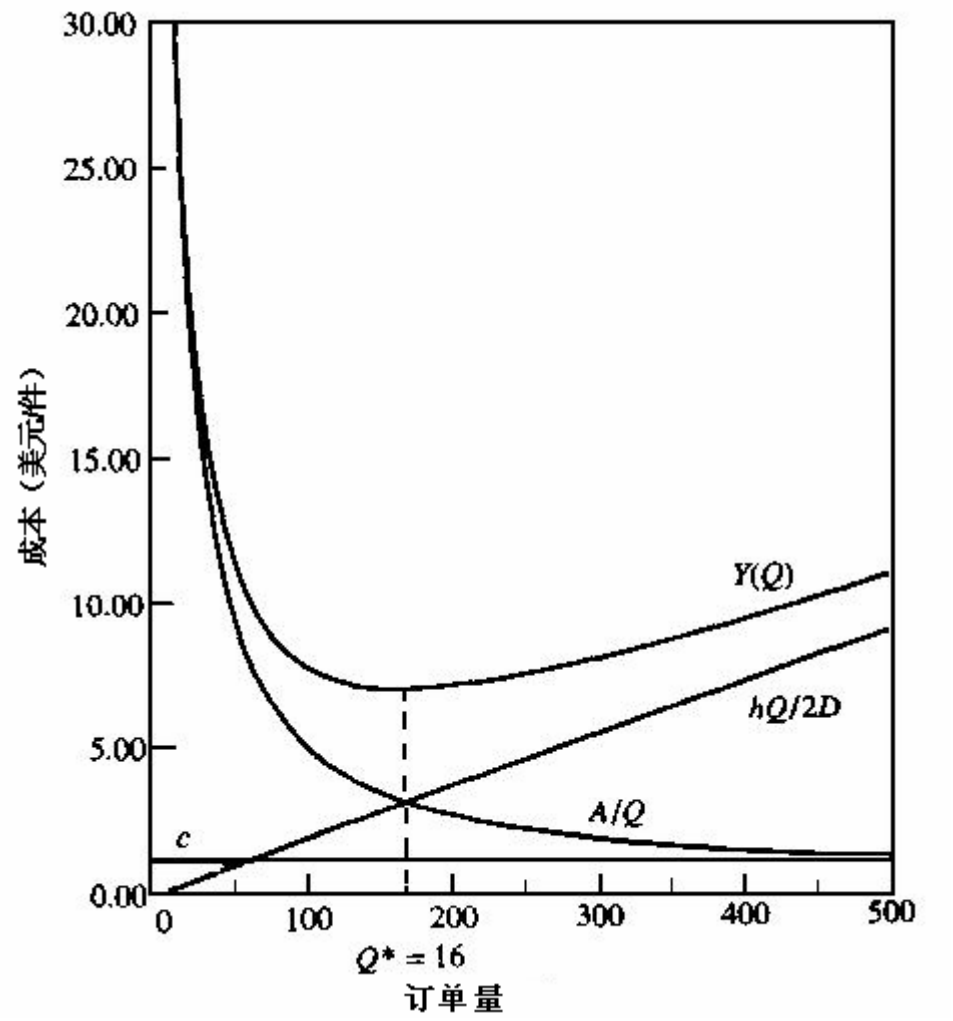

图 2.2 *EOQ* 模型中的成本曲线

技术性注释

求解 $Y(Q)$ 这样的无约束函数最小值的标准方法是对 Y 求 Q 的导数，并令 $Y'(Q)=0$，解此等式可得出结果 Q^*。这样可以找到斜率为零的点（即函数曲线处于水平的点）。如果函数是凸函数（我们在下面来验证这一点），零斜率点是唯一的，并对应着 $Y(Q)$ 的最小值。对 $Y(Q)$ 求导数并令其为零得

$$\frac{dY(Q)}{dQ}=\frac{h}{2}-\frac{AD}{Q^2}=0 \qquad (2.2)$$

这个等式表示的是得到最小值点的第一个条件，确保这个零斜率点对应着一个最小值（即不是最大值点或鞍点）的第二个条件则是检验 $Y(Q)$ 的二阶导数：

$$\frac{d^2Y(Q)}{dQ^2} = 2\frac{AD}{Q^3} \tag{2.3}$$

因二阶导数对任何正 Q 都大于零（即，Y（Q）是凸函数），接下来求解（2.2）Q^*（即式 2.4）就能得到 Y（Q）的最小值。

使成本函数（2.1）式中 Y（Q）最小的批量是

$$Q^* = \sqrt{\frac{2AD}{h}} \tag{2.4}$$

这个平方根公式就是著名的**经济订货批量（EOQ）**，也被称为**经济批量（Ecomnomic lot size）**。把这个公式应用到图 2.2 的例子中，我们就能得到

$$Q^* = \sqrt{\frac{2AD}{h}} = \sqrt{\frac{2(500)(1{,}000)}{35}} = 169$$

这个结果背后的直观意思就是：由于下订单相关的大额固定成本（$500），使得对于 MedEquip 来说值得大批量（169）订购架子。

2.2.3 经济订货批量的关键原理

上面结果中很明显的含义就是最优订货批量随着准备成本或需求率的平方根的增加而增加，随着持有成本的平方根的增加而减少。然而，Harris 的文章里一个更为基本的认识是他在他的摘要里所说的，即

在批量和库存之间存在着一个权衡。

增加批量就会增大库存的平均持有量，但是会减少订货的频率。通过用一个准备成本来惩罚频繁的补货，Harris 用清晰的经济术语明白地说明了这个权衡。（52|53）

上面的这个基本见解是无可争议的。然而，特定的数学结论（即 EOQ 平方根公式）总是依赖于模型的假设，有一些假设是我们完全可以质疑的（例如，即时生产多大程度上是可以实现的？）。不仅如此，即使是为了计算的目的，EOQ 公式的有用性也取决于输入数据的真实性。尽管 Harris 声称“准备成本是适于被普遍理解的”和“可能在一个大工厂里，每个订单的准备成本会是比一美元多一点，”但是估计准备成本实际上可能是一件困难的事情。正如我们要在第二和第三篇详细讨论的，准备成本在一个制造系统里有很多其他的影响因素（例如能力、变动性和质量），这样就把一个非常复杂的成本简化成了一个简单不变成本。而在采购系统里，这些其他的影响因素中很多就不起作用了，这时准备成本就可以被清楚地转换成采购订单的成本，EOQ 模型在这里就很有用了。

值得注意的是，我们甚至不需要借助于 Harris 的平方根公式就能使用这一结论，即批量与库存之间存在着一个权衡。因为每年平均的批数 F 为

$$F = \frac{D}{Q} \tag{2.5}$$

并且总的库存投资为

$$I = \frac{cQ}{2} = \frac{cD}{2F} \tag{2.6}$$

我们可以简单地画出库存投资 I 作为补给频率 F（批次/年）的函数曲线。我们令 D =1000、c =250 美元，在图 2.3 中画出了 MedEquip 例子的曲线。注意到这个图向我们表明当每年生产或者订购的次数从 10 次增加到 20 次时（即，将批量的大小从 100 改为 50）库存减少了一半（从 12,500 美元到 6,250 美元）。然而，如果我们每年补给次数从 20 次增加到 30 次（即把批量的大小从 50 减少到 33），库存只从 6,250 美元减少到 4,125 美元，减少 34%。

这个分析表明增加补给次数产生的效果是边际递减的。如果我们可以给生产次数或者采购次数也对应一个数值（即准备成本 A），那么我们可以像图 2.2 那样使用 EOQ 公式计算出最优批量。然而，如果成本是未知的（很有可能如此），那么图 2.3 的曲线至少让我们认识了增加补给次数对于总库存所能产生的影响。理解了这一个权衡，管理者就可以选择一个合理的切换次数或者采购次数，进而确定批量的大小。（53|54）

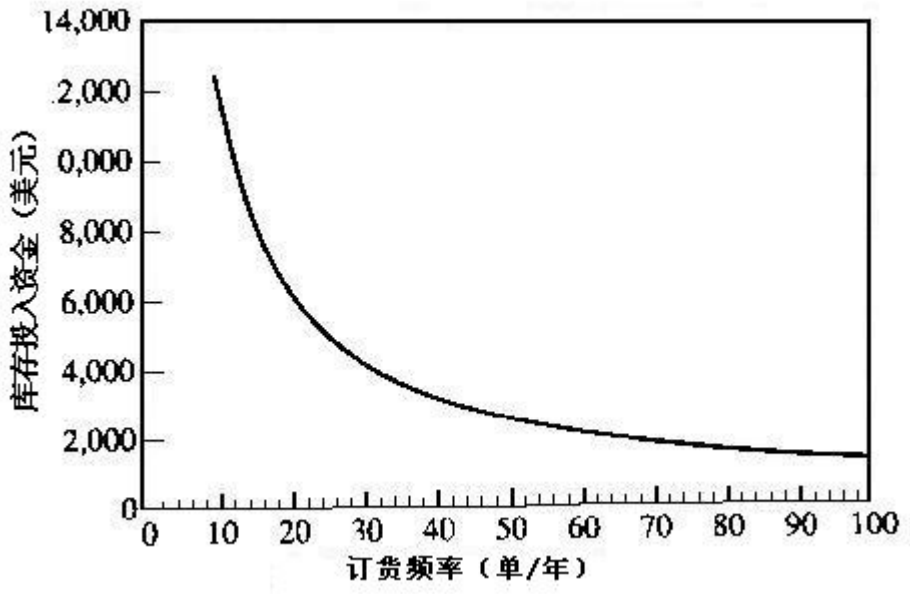

图 2.3 库存投资与每年的批次

2.2.4 灵敏度

从 EOQ 模型里得到的第二个观点是

持有成本和准备成本对于批量大小的改变很不敏感。

我们可以在图 2.2 中看出这点，当 Q 值在 96 和 306 之间变动时，总成本只在 7 和 8 之间变动。这意味着不管是什么原因，如果使用一个与 Q*稍有差别的批量时，持有成本与准备成本之和的增量不会很大。Harris 在他最初的那篇文章里定性地描述了这一特征。据我们所知，最早对其进行定量化处理的人是 Brown（1967，16）。

为了检验成本对批量大小的灵敏度，我们开始用 Q*代替 Q 代入 Y 的表达式（2.1）（但是忽略掉 c 项，因为它不受批量的影响），我们发现最小的单位持有与准备成本之和为

$$Y^* = Y(Q^*) = \frac{hQ^*}{2} + \frac{AD}{Q^*}$$

$$= \frac{h\sqrt{2AD/h}}{2} + \frac{AD}{\sqrt{2AD/h}}$$

$$= \sqrt{2ADh}$$

现在，假设不用 Q^*，我们使用其他任意的一个批量大小 Q'，它可能比 Q^*大些或小些。从 Y（Q）的表达式（2.1）我们可以看出，在取 Q'时，年持有与准备成本之和可写为

$$Y(Q') = \frac{hQ'}{2} + \frac{AD}{Q'}$$

因此，使用 Q'的年成本与最优年成本（使用 Q^*）的比率为

$$\frac{Y(Q')}{Y^*} = \frac{hQ'/2 + AD/Q'}{\sqrt{2ADh}}$$

$$= \frac{Q'}{2}\sqrt{\frac{h^2}{2ADh}} + \frac{1}{Q'}\sqrt{\frac{A^2D^2}{2ADh}}$$

$$= \frac{Q'}{2}\sqrt{\frac{h}{2AD}} + \frac{1}{2Q'}\sqrt{\frac{2AD}{h}}$$

$$= \frac{Q'}{2Q^*} + \frac{Q^*}{2Q'}$$

$$= \frac{1}{2}\left(\frac{Q'}{Q^*} + \frac{Q^*}{Q'}\right) \qquad (2.8)$$

为了解释这个表达式（2.8），假设 $Q' = 2Q^*$，意思就是使用的批量大小是最优批量的两倍。那么在这种情况下的持有与准备成本之和与最优成本之比为 1/2（2+1/2）= 1.25。即，批量上一个 100%的误差导致了成本上 25%的误差。注意到如果 $Q' = Q^*/2$，同样也可以在成本函数里得到一个 25%的误差。

由于需求是确定的，订单间隔时间完全是由订单数量所决定的，因此我们无法通过 EOQ 模型来进行更为深入的灵敏度分析了。我们可以将订单间隔时间 T 表示为（54|55）

$$T = \frac{Q}{D} \qquad (2.9)$$

这样，将式（2.4）除以 D，我们可以得到最优订单间隔时间的表达式

$$T^* = \sqrt{\frac{2A}{hD}} \qquad (2.10)$$

把式（2.9）代入到式（2.8）中，我们可以得到一个任意订单间隔时间 T'时的成本与最优成本比率的表达式：

$$\frac{T'\text{时的年成本}}{T^*\text{时的年成本}}=\frac{1}{2}\left(\frac{T'}{T^*}+\frac{T^*}{T'}\right) \tag{2.11}$$

式（2.11）在多种产品组合的情况下是很有用的，在这种情况下不同产品在相同的时间频繁地补货就是值得的（例如，便于共用配送卡车）。针对这种问题有一种在运筹学文献中被广泛提到的方法，那就是按照 2 的幂（*power-of-two*）的时间间隔来订货。也就是，令订货间隔为 1 周、2 周、4 周、8 周，依此类推。[2]结果是所有以 2^*n*周为间隔订货的订单将会被安排到与所有以 2^*k*周（*k*<*n*）为间隔订货的订单相同的时间（见图 2.4）。这将有利于共用配送卡车、减少订货的相关工作，简化配送安排等。

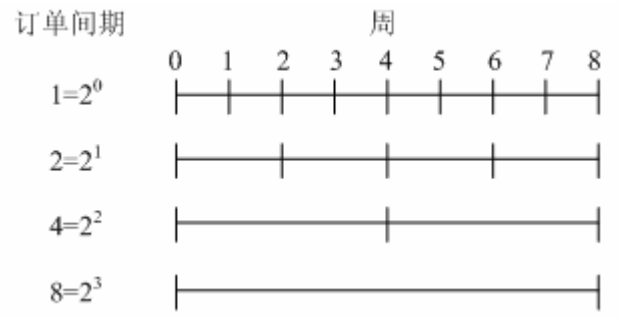

图 2.4 2 的幂的时间间隔

不仅如此，上面由 EOQ 模型所引出的灵敏度的结果表明将订单间隔限制为 2 的幂所导致的误差并不会太大。为了证明这一点，假设某种物品的最优订购间隔 T^*位于 2^m 和 2^{m+1} 之间，m 为某一特定值（如图 2.5）。那么 T^*要么位于区间[2^m，$2^m\sqrt{2}$]内，要么位于区间[$2^m\sqrt{2}$，2^{m+1}]内。所有在区间[2^m， $2^m\sqrt{2}$]内的点都不大于 2^m 的 $\sqrt{2}$ 倍。同样地，所有在区间[$2^m\sqrt{2}$， 2^{m+1}]内的点都不小于 2^{m+1} 除以 $\sqrt{2}$。例如，在图 2.5 中，2^m 在 T_1^* 乘以 $\sqrt{2}$ 和除以 $\sqrt{2}$ 之间，2^{m+1} 在 T_2^* 乘以 $1/\sqrt{2}$ 和除以 $1/\sqrt{2}$ 之间。因此，2 的幂的订购间隔 T'必定位于最优订购间隔 T^*附近的[$T^*/\sqrt{2}$，$\sqrt{2}T^*$]区间之内。这样，当 $T'=\sqrt{2}T^*$ 或者 $T'=T^*/\sqrt{2}$ 时成本的误差会达到最大。由式（2.11），$T'=\sqrt{2}T^*$ 时的误差为

$$\frac{1}{2}\left(\sqrt{2}+\frac{1}{\sqrt{2}}\right)=1.06$$

并且与 $T'=T^*/\sqrt{2}$ 时是相同的。因此，通过 2 的幂间隔来确定的最优成本与最优时间间隔的成本的误差保证不会超过 6%。Jackson、Maxwell 与 Muckstadt（1985）；Roundy（1985、1986）；Federgruen 和 Zheng（1992）给出了计算 2 的幂最优策略的算法，并且将上面的结

[2] 为了保证全面性，我们必须考虑 2 的负指数次方即 1/2 周，1/4 周，1/8 周，等等。然而，如果我们将一个足够小的时间单位作为基准（例如用天而不是周），这样在实际中就不用考虑负指数了。

果拓展到更为一般的多部件产品组合中。（55|56）

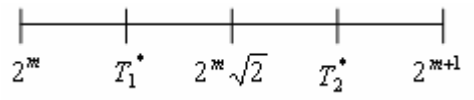

图 2.5 “平方根”间隔

作为这些概念的一个具体例子，我们再来看 MedEquip 的问题。我们算出了架子的最优订购数量为 $Q^* = 169$。这样，最优订购间隔就是 $T^* = Q^*/D = 169 / 1{,}000 = 0.169$ 年，或表示为 $0.169 \times 52 = 8.78$ 周。进一步假设 MedEquip 从同一供应商那里订购了多种其他零部件。架子的单位价格为 250 美元，这是一个送达价格，假设已经包含了平均的运输成本。然而，如果 MedEquip 将不同零件的订单组合在一起，总运输成本就会下降。如果所有的零件中最小的订购间隔为一星期，那么架子的订购间隔可以取整到最接近的一个 2 的幂级数即 $T = 8$ 周或 $8 / 52 = 0.154$ 年。订货数量 $Q = TD = 0.154$（1，000）= 154。调整后，订货数量的持有成本与订购成本之和为

$$Y(Q) = \frac{hQ}{2} + \frac{AD}{Q} = \frac{35(154)}{2} + \frac{500(1{,}000)}{154} = \$5{,}942$$

最优年成本（即用订购量 $Q^* = 169$ 得到的成本）为

$$Y^* = \sqrt{2ADh} = \sqrt{2(500)(1{,}000)(35)} = \$5{,}916$$

因此，调整后的订货数量最终增加了不到百分之一的额外成本。从该供应商处订购的其他零部件同样会引起持有加订购成本的增加——但是不会超过百分之六。如果成本的上升少于运输成本节约的话，那么 2 的幂级数的订购安排就是划算的。

2.2.5 EOQ 模型的拓展

Harris 的基本公式这些年来已经拓展到了很多方面。一个最早的拓展（Taft，1918）是针对库存补充的非即时性，即有一个有限的、恒定的生产率。这个模型被称为**经济生产批量（EPL）**模型，这个模型得到一个与 EOQ 公式相似的平方根公式。基本 EOQ 模型的其他变化形式包括延迟订单（即订货不是即时完成的，必须要等到有可得的存货），主从准备时间，数量折扣等（Johnson 和 Montgomery，1974；McClain 和 Thomas1985；Plossl 1985；Silver、Pyke 和 Peterson 1998）。

2.3 动态订货批量

正如我们在上面提到的，EOQ 公式建立在很多假设的基础之上的，具体为以下几个：

1. 即时生产
2. 即时交付
3. 确定的需求
4. 不变的需求
5. 已知的常量准备成本
6. 单一产品或相互独立的产品（56|57）

我们之前提到 Taft 松弛了即时生产这一假设。如果交付时间是已知且固定的，那么就可以直接引入延迟交付了（例如，根据 EOQ 公式计算订货数量，然后根据期望的到货时间减去交付（运输）时间来确定何时下订单）。如果交付时间是不确定的，那么就需要另一种方法了。然而，比交付时间更普遍、更重要的随机性来源是需求本身。接下来在下一节统计库存模型的讨论中将会松弛确定性需求这一假设。我们已经讨论了一种对付常量准备成本的方法（即通过考察库存与订购频率之间的权衡）。在第十七章中我们会讨论应对零部件不能被单独分析的多产品情况的方法。这样就只剩下常量需求的假设了。

2.3.1 动因

来看 RoadHog 有限公司的情况，它是一个生产摩托车配件的小型制造商。它生产带翼的消音器（翼对降低引擎噪音几乎不起作用），其生产线也可用于生产其他产品。因为建造一条生产线非常昂贵，所以 RoadHog 考虑进行批量生产。然而，顾客需求是以 10 周为一个计划期提前获得的（因为其被排进主生产计划而被“冻结”），因此也就不一定每周都有不变的需求。鉴于这违背了 EOQ 模型的关键假设，我们需要一个从根本上不同的模型来平衡准备成本和持有成本。

松弛常量需求假设最主要的研究是 Wagner-Whitin 模型（Wagner 和 Whitin，1958）。这个模型考虑的问题是当需求固定但时间变化，而其他 EOQ 模型条件也都成立的情况下如何确定生产批量。动态订货批量的重要性就在于它对生产控制的文献产生重大影响，后来它影响了物料需求计划（MRP）的产生，这一点我们将会在第三章中讨论。鉴于这些原因，我们现在先来对 Wagner-Whitin 动态批量方法进行一个介绍。

2.3.2 问题模型化

当需求随时间变化时，像 EOQ 模型这样的连续时间模型就无法说明问题了。因此，针对这一问题，我们将需求聚集到离散的周期内，这个周期根据生产系统的不同，可以为天，星期或月份。对于一个大批量、需求快速变化的生产系统，往往就需要日排程，而小批量、需求变化较慢的生产系统则可能更适合用月排程。

为了详细说明这个问题和模型，我们利用以下这些符号，它们可以被看成是与 EOQ 模型中所使用的静态符号相对应的动态符号：

t = 时间周期（天、星期、月），我们令 t =1, ... , T ，这里 T 表示计划周期

D_t = 第 t 期的需求（以单位计）

c_t = 单位生产成本（美元），不包括 t 期的准备成本及库存成本

A_t = 在第 t 期生产（订购）一批产品所需的准备（订货）成本（以美元计）

h_t = 从第 t 期到第 t+1 期的单位库存持有成本（美元每单位每周期）。例如，如果库存持有成本包括整个库存占用资金的利息的话，令 i 为每年的利息率，周期为星期，那么 $h_t = i\,c_t/52$ （57|58）

I_t = 第 t 期末的库存

Q_t = 第 t 期的批量，共有 T 个这样的决策变量，每个周期都有一个

示例：

利用这些符号，我们就可以将 RoadHog 的问题精确地定义清楚。我们假设接下来 10 周的数据如表 2.1 所示。这里需要注意的是，为了把问题简化，尽管 Wagner-Whitin 模型不要求准备成本 A_t，生产成本 c_t，持有成本 h_t 为常量，但是我们还是假设这些成本随时间保持不变。最基本的问题是在付出最小成本（即生产成本加上准备成本加上持有成本）的情况下满足所有需求。唯一可以控制的就是生产数量 Q_t。然而，由于所有需求必须被满足，那么就只有生产的时间安排是可以由我们来选择的，生产的总数是不能改变的。因此如果单位生产成本是常量（即 c_t 不随时间 t 变化）,那么无论考不考虑生产时间安排，生产成本都是一样的，因此整个就可以忽略掉。

人们可能想到的最简单的确定批量的方法就是每个周期只生产该周期所需要的量。这就是**批对批法（lot-for-lot rule）**，在第三章中我们将看到，在某些情况下这是可行的。然而，对于目前这个问题，批对批法意味着我们必须在每个周期都进行生产并且都要都要付出准备成本。表 2.2 显示了这种策略下的生产计划以及最终成本。既然期间不持有任何库存，这样总成本就是 10 次的准备成本，即 1,000 美元。

表 2.1 *RoadHog* 例子中固定订货批量的解

t	1	2	3	4	5	6	7	8	9	10
D_t	20	50	10	50	50	10	20	40	20	30
c_t	10	10	10	10	10	10	10	10	10	10
A_t	100	100	100	100	100	100	100	100	100	100
h_t	1	1	1	1	1	1	1	1	1	1

表 2.2 *RoadHog* 例子的批对批解法

t	1	2	3	4	5	6	7	8	9	10	总计
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	20	50	10	50	50	10	20	40	20	30	300
I_t	0	0	0	0	0	0	0	0	0	0	0
A_t	100	100	100	100	100	100	100	100	100	100	1,000
h_t	0	0	0	0	0	0	0	0	0	0	0
总成本	100	100	100	100	100	100	100	100	100	100	1,000

表 2.3 *RoadHog* 例子的固定批量解法

t	1	2	3	4	5	6	7	8	9	10	总计
D_t	20	50	10	50	50	10	20	40	20	30	300
Q_t	100	0	0	100	0	0	100	0	0	0	300
I_t	80	30	20	70	20	10	90	50	30	0	0
A_t	100	0	0	100	0	0	100	0	0	0	300
h_t	80	30	20	70	20	10	90	50	30	0	400
总成本	180	30	20	170	20	10	190	50	30	0	700

另一个似乎也可行的策略是每付出一次准备成本，就生产一批固定数量的产品。这被称为**固定批量法（fixed order quantity）**，既然一共有 300 单位的产品要生产，那么假设一个固定订购批量为 100 单位。这就要求我们进行三次生产，支付三笔准备成本，并且在第 10 期末不会留下库存。表 2.3 显示了这种策略下的生产计划和最终成本。注意到在这种策略下，我们每次生产的量都超出了当期需求，因此也就必须支付库存持有成本。然而，总的库存持有成本只有 400 美元，再加上 300 美元的准备成本，就有 700 美元的总成本。这比批对批策略下的成本要低。但是我们还可以做的更好吗？要知道这一点，下面我们可以建立一个保证可以得到准备成本加库存成本之和为最小的方法。（58|59）

2.3.3 Wagner-Whitin 方法

解决动态批量问题所需要的一个关键点就在于如果我们在第 t 期进行生产（相应产生准备成本）从而满足第 *t*+1 期的需求，那么在第 *t*+1 期进行生产（相应产生准备成本）则显然是不经济的。不管是第 *t* 期生产第 *t*+1 期的所有需求更便宜，还是在 *t*+1 期生产所有当期的需求更便宜；都绝对要比各期都生产一些要便宜。（注意，我们在表 2.3 中给出的固定订货批量方案中违反了这一性质），我们可以用更为一般的语言表述如下：

Wagner-Whitin 性质

在一个最优批量策略下，要么从前一期转入第 t+1 期的库存为零，要么第 t+1 期的生产量为零。

正如我们将看到的，这个结果可以极大地方便对最优生产批量的求解。[3]

Wagner-Whitin 性质意味着要么 $Q_t = 0$ 要么 Q_t 就是下面这些情况中的一个：D_t，$D_t + D_{t+1}$，$D_t + D_{t+1} + D_{t+2}$，...，$D_t + D_{t+1} + \cdots + D_T$。也就是说，我们要么不生产，要么就生产恰好能够满足当期加上未来若干期的需求总和。我们可以通过列举所有可能的生产期组合来计算最小成本的生产排程。然而，因为在每个时期都可以选择生产或者不生产，这样

[3] 一些权威的评论者已经指出，虽然在数学上很有用，但在现实的生产系统中 Wagner-Whitin 性质要么就是非常显而易见，要么就是荒谬可笑。实际上，它规定了我们在库存降到零之前不应该进行生产。如果有人真地接受了模型的所有假设，尤其是那些关于已知的确定需求和明确定义的固定准备成本，那么这个性质几乎是句废话。然而，在充满不确定的复杂事物的现实系统中，人们几乎总是在库存耗尽之前就开始生产了（即为了防止出现由随机扰动引起的缺货）。

组合的数量就有 2^{N-1}，如果考虑的期数很多时这个数字就会很大的。为了提高效率，Wagner-Whitin 提出了一个十分适合计算机执行的算法。我们会通过 RaadHog 的例子来说明这一算法。（59|60）

Wagner-Whitin 算法是按时间顺序向前执行的，从第 1 期开始到第 N 期结束。根据 Wagner-Whitin 性质，可以知道只有在当期期初库存为零的那一期才进行生产。在这种情况下，那么我们决策所需要考虑的就是一次生产的量要满足多少期的需求。例如，在一个六期的问题中，在第一期我们有 6 种可以选择产量，即，D_1，D_1+D_2，$D_1+D_2+D_3$，...，$D_1+D_2+D_3+D_4+\cdots+D_6$。如果我们选择生产 D_1+D_2，那么库存就会在第 3 期耗尽，所以在那一期我们必须再次生产。而在第 3 期时，我们有以下选择：生产第 3 期需求的量、生产第 3 和第 4 期需求之和的量、生产第 3、4、5 期需求之和的量，或者是第 3、4、5、6 期需求之和的量。

第一步

我们先从一期问题开始推演这个算法。可以这样来看，我们假装认为在经过一期之后世界就毁灭了。对这个问题提最优策略好像没什么意思，我们生产 20 单位来满足第 1 期的需求，结果我们当然做到了。由于没有跨期的库存，并且我们忽略了生产成本，因此在一期问题中的最小成本 Z_1^* 为

$$Z_1^* = A_1 = 100$$

正如我们将要看到的，随着这个算法的展开，追踪最后一次进行生产的那一期对于任何一个问题都是非常有用的。在这里，生产显然只发生在第 1 期，所以在一期问题中的最后生产期 j_1^* 为

$$j_1^* = 1$$

第二步

在这个算法的第二步里，我们增加一期的时间来考虑二期问题。现在对于第 2 期生产我们有两个选择；我们可以选择用第 1 期还是第 2 期生产的产品来满足第 2 期的需求。如果我们在第 1 期生产，我们会造成一个对应于从第 1 期保有库存到第 2 期的库存持有成本。如果我们在第 2 期生产，就会在第 2 期产生一个额外的准备成本。考虑到如果我们在第 2 期生产，那么满足先前需求（即第 1 期的需求）的成本就可以用 Z_1^* 来表示。由于我们是要让成本最小，那么最优的策略就是选择总成本较低的那一期，即

$$Z_2^* = \min\begin{Bmatrix} A_1 + h_1 D_2 & \text{produce in period 1} \\ Z_1^* + A_2 & \text{produce in period 2} \end{Bmatrix}$$

$$= \min\begin{Bmatrix} 100 + 1(50) = 150 \\ 100 + 100 = 200 \end{Bmatrix}$$

$$= 50$$

最优决策是在第 1 期同时生产第 1 期和第 2 期所需的产品。所以，在最优的二期策略中的最后生产期是（60|61）

$$j_2^* = 1$$

第三步

现在，我们继续讨论三期问题。这里通常需要考虑四个可能的生产排程：只在第 1 期生产，在第 1、2 期生产，在第 1、3 期生产，或者在第 1、2、3 期生产。然而，只需考虑其中的三种：只在第 1 期生产，在第 1、2 期生产，在第 1、3 期生产。这是因为我们已经解决了二期和一期问题，现在只需考虑什么时候生产第 3 期的需求。需要指出的是随着期数的增长这种迭代讨论产生的效率也会急剧提高。如对于 10 期问题，我们必须考虑的排程数量从 512 个减少到 10 个。在下面的讨论中可以看到，利用“计划期”可以将这个数目减到更小。[4]

如果我们决定在第 3 期进行生产，那么根据两期问题的解，应该在第 1 期生产第 1、2 期的需求。

$$Z_3^* = \min\begin{Bmatrix} A_1 + h_1 D_2 + (h_1 + h_2) D_3 & \text{produce in period 1} \\ Z_1^* + A_2 + h_2 D_3 & \text{produce in period 2} \\ Z_2^* + A_3 & \text{produce in period 3} \end{Bmatrix}$$

$$= \min\begin{Bmatrix} 100 + 1(50) + (1+1)(10) = 170 \\ 100 + 100 + 1(10) \quad = 210 \\ 150 + 100 \quad = 250 \end{Bmatrix}$$

$$= 170$$

最优结果仍然是在第 1 期生产所有的需求总和的量，因此

$$j_3^* = 1$$

第四步

当我们进行四期问题的这一步时情况发生了变化。对于第 4 期的需求量在什么时间进行生产有四个选项，即，第 1 期到第 4 期中的一期：

$$Z_3^* = \min\begin{Bmatrix} A_1 + h_1 D_2 + (h_1 + h_2) D_3 + (h_1 + h_2 + h_3) D_4 & \text{produce in period 1} \\ Z_1^* + A_2 + h_2 D_3 + (h_2 + h_3) D_4 & \text{produce in period 2} \\ Z_2^* + A_3 + h_3 D_4 & \text{produce in period 3} \\ Z_3^* + A_4 & \text{produce in period 4} \end{Bmatrix}$$

$$= \min\begin{Bmatrix} 100 + 1(50) + (1+1)(10) + (1+1+1)(50) = 320 \\ 100 + 100 + 1(10) + (1+1)(50) \quad = 310 \\ 150 + 100 + 1(50) \quad = 300 \\ 170 + 100 \quad = 270 \end{Bmatrix}$$

$$= 270$$

[4] 这种解决连续多期问题、在每一步里通过使用先前步骤的解来减少计算量的技术被称为*动态规划*。动态规划是*隐枚举法*（*inplicit enumeration*）的一种形式，它使得我们可以考察所有可能的解而不必具体计算每一个的成本。

这次的最优结果不再是由第 1 期生产最后一期的需求量，而是由第 4 期生产来满足其自身的需求。因此，

$$j_4^* = 4$$

如果我们的计划期只有 4 期，那么就可以到此为止了。我们将通过倒序的 j_i^* 的值来确定批量政策。$j_4^* = 4$ 意味着我们要在第 4 期生产 $D_4 = 50$ 个单位的产品。这样一来问题就变成了一个三期问题。因为 $j_3^* = 1$，所以在第 1 期生产 $D_1 + D_2 + D_3 = 80$ 单位的产品，这个结果是最优的。（61|62）

第五步到第十步

但是实际上我们的计划期不是 4 期；而是 10 期。因此，我们必须继续推演这个算法。然而在此之前，我们通过下面的分析可以大大减少所需要的计算量。值得注意的是，到目前为止，每多做一步，算法中针对最后一期的生产所需要考虑的期数都会增加。这样，到第 4 步的时候，我们就必须在第 1 期到第 4 期的每一期中都要考虑是否生产第 4 期的需求量。其实这并不总是必要的。

注意在四期问题中在第 4 期生产当期的需求量是最优的。这意味着第 4 期产生的准备成本要少于前面三期的准备成本和把库存保留到第 4 期的成本。若不是如此，我们就会在这三期里选一期来进行生产了。现在来考虑这对于第 5 期到底意味着什么。例如，在第 3 期生产第 5 期的量会比在第 4 期生产更便宜吗？无论是在第 3 期还是在第 4 期生产，这些产品库存都必须从第 4 期保留到第 5 期，在这段时间里所引起的持有成本是一样的。这样剩下的问题就是在第 3 期进行生产然后将其作为库存保留到第 4 期是否会比直接在第 4 期生产更为便宜。其实我们已经知道了这个问题的答案。$j_4^* = 4$ 就已经告诉我们在第 4 期准备生产更便宜。所以，没有必要在第 1、2、3 期考虑生产第 5 期的需求。我们只需要考虑第 4 期和第 5 期。这个道理可以更一般地表述如下：

计划期性质

如果 $j_4^* = \bar{t}$，那么在 t+1 期的最优策略里最后生产期必定属于集合 $\bar{t}$，$\bar{t}$+1，…，t+1。利用这个性质，求解五期问题最小成本所需的计算式为：

$$Z_5^* = \min\begin{Bmatrix} Z_3^* + A_4 + h_4 D_5 & \text{produce in period 4} \\ Z_4^* + A_5 & \text{produce in period 5} \end{Bmatrix}$$

$$= \min\begin{Bmatrix} 170 + 100 + 1(50) = 320 \\ 270 + 100 \qquad\quad = 370 \end{Bmatrix}$$

$$= 320$$

既然我们一定会在第 4 期进行生产，那么从第 4 期保留库存到第 5 期就会比在第 5 期再次准备生产要更为便宜。因此，

$$j_5^* = 4$$

表 2.4 *Wagner-Whitin* 例子的解

上一个生产期	计划期 t									
	1	2	3	4	5	6	7	8	9	10
1	100	150	170	320						
2		200	210	310						
3			250	300						
4				270	320	340	400	560		
5					370	380	420	540		
6						420	440	520		
7							440	480	520	610
8								500	520	580
9									580	610
10										620
Z_t^*	100	150	170	270	320	340	400	480	520	580
j_t^*	1	1	1	4	4	4	4	7	7 或 8	8

我们用同样的方法求解了剩下的五期，并把求解的结果汇总为表 2.4。注意到表的右上角有一块空白区域，这就是我们利用计划期性质的结果。如果没有这个性质，我们就必须为这里面每一个空格计算出数值。（62|63）

2.3.4 结果说明

准备成本与持有成本之和的最小值为 Z_{10} = \$580，注意到这个数值比前面所说的批对批或固定批量策略所求得的成本都要低。最优的批量大小是由 j_t^* 的值决定的，由于 $j_{10}^* = 8$，在第 8 期生产第 8，9，10 期的需求是最优的。因此，$Q_8^* = D_8 + D_9 + D_{10} = 90$。解决了第 8、9、10 期的问题，剩下就是考虑前面的七期问题了。因为 $j_7^* = 4$，因此最优的情况就是在第 4 期生产第 4、5、6、7 期的需求量。这样，$Q_4^* = D_4 + D_5 + D_6 + D_7 = 130$。然后就剩下了一个三期问题，因为 $j_3^* = 1$，我们应该在第 1 期生产第 1、2、3 期的需求量，因此，$Q_1^* = D_1 + D_2 + D_3 = 80$。

2.3.5 忠告

表 2.4 所包含的计算过程由手工来做确实很冗繁，但是对于电脑来说就不一样了。即便如此，让人觉得奇怪的是许多生产运作管理的教材都忽略了 Wagner-Whitin 算法，却选择使用相对简单但大多不能给出最优解的探索发现法（heuristics）。也许是因为“简单”意味着既减少了计算的繁琐工作又更容易解释。既然算法仅仅被用在那些生产计划本身就是计算机自动处理的情况下，那么所谓的有计算负担也就没有什么说服力了。此外，算法背后所包含的概念并不算困难——当然也不会困难到技术人员都不愿把它整合到商业软件中去！

然而不管是使用 Wagner-Whitin 算法还是任何探索发现法，都必须注意到对于“最优”批量的整个概念有一些更为重要的关注点。

1. 像经济批量模型一样，Wagner-Whitin 模型假设准备成本在批量确定之前是已知的。但是，像我们之前指出的，在制造系统中估计准备成本是非常困难的。不仅如此，准备的真实成本是受能力影响的。例如，当生产接近能力极限的时候，换模时的停机所造成的产能损失的代价是很大的；然而如果有很多额外能力的话，代价就没那么大了。任何假设独立准备成本的模型都无法准确表述这个问题。这样，Wagner-Whitin 模型就会像经济订货批量模型一样更适用于采购系统而非生产系统。（63|64）

2. 和经济批量模型一样，Wagner-Whitin 模型假设具有确定的需求和确定的产出。像订单取消、产出损失（yeild loss）、和配送计划变动这样的不确定性都不考虑。这样的结果就是 Wagner-Whitin 算法所给出的“最优”生产排程都必须经过修正才能够满足现实情况（例如，为了在出现订单取消的情况下适应（accommodate）剩余库存而减少产量或者因为预期的产出损失而增加产量）。由于需要针对每一个特殊情况做出调整，再加上准备成本的复杂性（speculative），使这个理论上的最优排程在现实中难以发挥作用。

3. 另一个关键假设就是独立生产，即不同产品的生产不共用资源。这个假设在很多情况下都显然是不能满足的。如果某些资源是利用率很高，那么这一点就十分重要了。

4. Wagner-Whitin 性质所给我们带来的结论就是对于一个生产周期，我们要么不生产，要么就生产未来若干期的需求总和。这个性质基于这样两个前提：（1）在每次生产的时候产生一个固定的准备成本；（2）没有能力限制。在现实世界里，准备成本会产生更多微妙的效果，并且能力也是有限制的，一个真正明智的生产计划可能会与所谓的最优计划大不相同。例如很有可能会依照一个平准生产计划进行生产（即每期都生产大致相同的量），从而可以在产线上实现一定程度的速度和节奏。而 Wagner-Whitin 方法则由于过于关注固定成本和持有成本之间权衡，却很可能会让我们的直觉脱离现实。

2.4 静态库存模型

以上讨论的所有模型都建立在这样一个假设之上：需求是确定并且已知的。尽管在有些情况下这个假设确实可能是接近实际情况的（例如排程在计划期内被严格的冻结），但是更多的时候往往不是这样。当需求不确定的时候，可以采用两种基本方法：

1. 把需求模型化，建立确定性模型，然后修改模型的解以抵消随机性。
2. 在模型中明确包含和表示不确定性。

没有哪种方法是绝对正确或绝对错误的，关键在于哪一种方法更为有效。一般来说，这个问题的答案取决于具体的环境。当计划跨越了一个长度足以使随机变差“平均抵消”的时期时，确定性模型可能会更有用。当然，如果在一个确定性模型中加入了预测随机性的“弹性因素（fudge factor）”，并且配合适当频率的计划更新周期以纠正出现的偏差的话，确定性模型将会更加有效。然而，如果想确定这些弹性因素或者要帮助设计应对随机性在其中起关键作用的时间帧（time frames）策略时，一个明确包含随机性的模型将会更为合适。

从历史的观点来看，运作管理的文献一直都同时延用着以上两种方法。生产排程最普遍使用的确定性模型是物料需求计划（MRP），也就是第三章的主题。最普遍使用的不确定性模型是**统计再订货点**模型，这是我们在这一部分要讨论的内容。

针对生产和库存控制问题的统计模型已经不是什么新内容了，它至少可以追溯到 Wilson 在 1934 写的一篇论文，在这篇经典论文中，Wilson 把库存控制问题分解为两个部分：

1. 确定**订货数量（order quantity）**，也就是每次补货所要采购或生产的库存数量。
2. 确定**再订货点（order point）**，即触发补货订单（采购或生产）所需达到的库存水平。

在这一节，我们将把这个两段问题分解为三个阶段来讨论。（64|65）

首先，我们要考虑只关注单次补给的情况，因此唯一的问题就是要针对不确定的需求来确定适当的订货数量。这通常被称为**报童模型（newsvender model）**，因为它适用于这样的情况：某人在一天的开始购入报纸，销售掉一个随机的量，最后丢弃剩余的库存。

第二，我们要考虑随着随机（译者注：指时间上的随机，一次只产生一件）需求的产生，每次只补充一件库存的情况，这样唯一的问题就是要确定再订货点。我们为系统设定的目标库存水平被称为基准库存水平，因此这个模型就被称为**基准库存模型（base sotck model）**。

第三，我们要考虑这样的情况，即库存被持续监控，并且需求随机发生，可能是以批量形式。当库存水平达到（或低于）r 时，就产生一个数量为 Q 的订单。经过一个提前期 ℓ 之后（在这期间可能发生缺货）收到订货。这个问题就是确定适当的 Q 和 r 值。用来描述这个问题的模型被称为**（Q，r）模型**。

这些模型会用到概率论里的概念和符号。如果读者学习完这些内容已经有一段时间了，那么现在最好再精读一下附录 2A 的内容。

2.4.1 报童模型

我们先来看一个圣诞灯制造商面临的情况。它面对的需求是不可预知的，并且是在圣诞节之前如此短暂的时间内爆发出来，如果货架上没货了，那么就等于是销量的损失。因此，必须在节日季节到来之前就决定要生产多少套灯。此外，由于把没有售出的库存回收并保存到下一年的成本太高，因此没人愿意持有隔年库存。比较合适的做法就是在圣诞之后把所有剩余的灯都打折卖掉。

现在如果要确定一个适当的产量，就要考虑以下这些重要信息：（1）预期的需求；（2）生产过多或者过少的成本。为了建立一个正式的模型，我们需要做以下假设：

1. 产品是独立的。这样的话产品之间就没有交互作用（例如共享某种制造资源），因此我们可以一次考虑一种产品。
2. 一次只做一期的计划。因为当前决策对未来的影响是可以忽略的（例如因为库存不能隔期），这样决策的时候我们就可以把后期忽略掉。
3. 需求是随机的。我们可以用一个已知的概率分布来刻画需求。
4. 产品配送先于需求到达。所有已经订购或生产的库存都可直接用来满足需求而无需等待配送。
5. 产品过剩或缺货成本都是线性的。保有太多或者太少库存的费用是与过剩量或缺失量成比例的。

我们依据这些假设用下面的符号来建立模型：（65|66）

X = 需求，这是一个随机变量

$G(x)$ = $P(X \le x)$= 需求的累积分布函数；对于这个模型我们假设 G 是一个连续分

布，因为它便于求得解析解，但是结果基本上和 G 是离散的情况是一样的（即将解限于整数值）

$g(x)$ = $\frac{d}{dx}G(x)$= 需求的概率密度函数

u = 需求的均值

σ = 需求的标准差

c_0 = 需求产生后的每件过剩成本（美元）

c_s = 每件缺货成本（美元）

Q = 产量或订货量，这是决策变量。

示例：

现在加入一些数字来看圣诞灯的例子。假设一套灯的制造、配送成本为 1 美元，然后以 2 美元的价格卖出。圣诞过后剩余的灯都以 0.5 美元的折扣价卖掉。根据以上数字，可以看出单位剩余成本就是剩余的每套灯的损失，即 c_0 = 1 - 0.5 = 0.5 美元。每套灯的缺货成本就是销售中损失的利润，即 c_s = 2 - 1 = 1 美元。再进一步假设已预测到需求为 10,000 套，标准差为 1,000 套，并且以正态分布作为需求的分布函数。

该公司可以选择生产 10,000 套灯。但是注意，正态分布的对称形状（即钟型）说明了需求大于或小于 10,000 的概率是相等的。如果需求少于 10,000 套，公司过量生产的每套产品将会损失掉 c_0 = \$0.5。如果需求大于 10,000 套，公司供应不足的每套产品将损失 c_s = \$1 美元。很明显，缺货比过量生产更糟糕。这意味着公司或许应该生产多于 10,000 套的产品。但是多多少？下面建立的这个模型就是为了精确地回答这个问题。

为了建立这个模型，注意到如果我们生产 Q 套产品而实际需求为 X 套，那么过量生产的数量就是

$$\text{过剩数量} = \max\{Q - X, 0\}$$

也就是说，如果 $Q \geqslant X$，则过量产量就是 Q-X；但如果 $Q<X$，则存在缺货，因此过量数为 0。我们可以用下式计算过量数量的期望

$$E[\text{过剩数量}] = \int_0^{\infty} \max\{Q - x, 0\} g(x) dx$$

$$= \int_0^{Q} (Q - x) g(x) dx \qquad (2.12)$$

类似地，缺货的数量可以表示为

$$\text{缺货数量} = \max\{X - Q, 0\}$$

即，如果 $X \geqslant Q$，则缺货量就是 X-Q；但如果 $Q<X$，则存在过量，因此缺货量为 0。我们可以用下式计算缺货数量的期望

$$E[\text{缺货数量}]=\int_0^{\infty}\max\{\text{x-Q},0\}\text{g(x)dx}$$

$$=\int_0^{Q}\text{(x-Q)g(x)dx} \qquad (2.13)$$

利用（2.12）和（2.13），我们可以写出期望成本关于产量的函数（66|67）

$$Y(Q)=c_o\int_0^{Q}\text{(Q-x)g(x)dx}+c_S\int_Q^{\infty}\text{(x-Q)g(x)dx} \qquad (2.14)$$

在以下的技术性注释中，我们可以找到使得预期成本最小的产量 Q 的值。

技术性注释

和 EOQ 模型一样，要求得 Y（Q）的最小值，只需对其求导并令其等于 0。要做到这一点，我们需要对有上下限的积分函数求导。为此我们需要借助于莱布尼兹定理，它的公式写作

$$\frac{d}{dQ}\int_{a1(Q)}^{a_2(Q)}f(x,Q)dx=\int_{a1(Q)}^{a_2(Q)}\frac{\partial}{\partial Q}[f(x,Q)]dx+f(a_2(Q),Q)\frac{da_2(Q)}{dQ}-f(a_1(Q),Q)\frac{da_1(Q)}{dQ}$$

根据这个公式可以得到 Y（Q）的导函数并令其等于 0，可推出

$$\frac{dY(Q)}{dQ}=c_o\int_0^{Q}1g(x)dx+c_s\int_0^{\infty}(-1)g(x)dx$$

$$=c_oG(Q)-c_s[1-G(Q)]=0 \qquad (2.15)$$

求解（2.15）式（下面我们将其简化为 2.16 式）得到 Q*即是使 Y（Q）最小的最优产量（订购量）。

为了使剩余成本、缺货成本之和的期望最小，我们应该选择一个产量或订购量 Q*，并有

$$G(Q^*)=\frac{c_s}{c_o+c_s} \qquad (2.16)$$

首先，由于 G（Q*）代表需求小于等于 Q*的概率，这个结果意味着应该选择 Q*，从而使拥有足够库存满足需求的概率为 c_s/（c_0+c_s）。其次，由于 G（x）随着 x 的增加而增加（累积分布函数总是单调递增的），因此要使等式（2.16）的右边变大就必然要有更大的 Q*。这说明增大 c_s 会使 Q*增大，而增大 c_0 会使 Q*减小，这和我们猜的一样。

如果假设 G 是正态分布，我们可以进一步简化式（2.16），可以写作

$$G(Q^*)=\Phi\left(\frac{Q^*-\mu}{\sigma}\right)=\frac{c_s}{c_o+c_s}$$

此处Φ为标准正态分布的累积分布函数（cdf）[5]。这也就意味着

[5] 我们在这里利用了这样一个著名的结论，即如果 X 是均值为 μ，方差为 σ 的正态分布，那么 $(X-\mu)/\sigma$ 是

$$\frac{Q^* - \mu}{\sigma} = z$$

此处 z 为标准正态表中的值（参看本书最后的表 1），对其有 Φ（Z）=c_s/（c_0+c_s），因此有（67|68）

$$Q^* = \mu + z\sigma \qquad (2.17)$$

表达式 2.17 说明对于正态分布而言，Q*是需求均值 μ 的增函数。假如 z 是正值，Q*也会随需求标准差 σ 的增加而增加。由于 Φ（0）=0.5 并且 Φ（z）随 z 递增 ，因此只要 c_s/（c_0+c_s）大于 0.5，以上的结论就成立。然而，如果 c_s/（c_0+c_s）小于 0.5，则最优订货批量 Q*将会随着 σ 的增加而减少。

示例：

现在我们回到圣诞灯的例子。因为需求为正态分布，故可以根据式 2.17 计算出 Q*。为此首先要算出 z，即：

$$\frac{c_s}{c_0 + c_s} = \frac{1}{1+0.5} = 0.67$$

然后在标准正态表中查得 Φ（0.44）= 0.67。因此 z = 0.67 且

$$Q^* = \mu + z\sigma = 10{,}000 + （0.44）1{,}000 = 10{,}440$$

这一结果可以理解为应生产多于平均需求 0.44 倍标准差的产量。因此，如果需求的标准差为 2,000 套而不是 1,000 套，那么答案就会是比平均需求多生产 0.44×2,000 = 880 套产品，共 10,880 套。

报童模型和和式（2.16）中所体现的临界比率解法可以扩展各种应用场合，可以比这个圣诞灯例子中的一期问题讨论更多的期数。一种常见情况就像这样一个问题，其中

1. 公司面临相互独立且具有同分布 G（x）的周期性（如每月）需求。
2. 所有的订单均被延迟（即在最后满足）。
3. 没有与生产一个订单相关联的准备成本。

在这种情况下，“订到 Q”的策略（即在每次需求产生之后，生产足够的产品使库存水平达到 Q）是最优的。不仅如此，寻找这个库存水平最优值 Q*的问题可以被表述为一个报童模型（参见 Nahmias 1993，291-294）。结果 Q*也就满足式（2.16），其中 c_0 代表每期持有单位库存的成本，c_s 代表每期持有单位延迟交货产品（即尚未满足的订单）的成本。类似的，在相同条件下，可以预期到损失销售而不是延迟订单，最优的库存持有水平 Q*可通过

均值为 0，方差为 1 的正态分布（即标准正态分布）。

计算式（2.16）求得，其中c_0等于每期的持有成本而c_s等于单位利润（即售价减去生产成本）。

通过总结报童模型的基本思想我们可以得出以下结论：

1. 在需求不确定的环境中，合适的生产或订货量由需求的分布以及剩余或缺货的相关成本共同决定。

2. 当需求为正态分布时，如果$c_s/(c_0+c_s)>0.5$，则增加需求的变动性（即标准差）会引起产量或订货量的增加，而如果$c_s/(c_0+c_s)<0.5$则会使其减少。（68|69）

2.4.2 基准库存模型

我们来看 Superior Appliance 的情况，这是一家销售特殊型号冰箱的商店。因为它的空间有限而且生产厂商要频繁的运送其他电器，所以商店认为可以每当售出一台冰箱的时候下订单补货。实际上它拥有一套系统，只要销售完成就会自动下订单。但是由于生产厂商满足订单需要一定时间，所以为及时满足客户需求，商店需要持有一定的库存。在这种情况下，关键的问题是应该持有多少库存。

我们需要通过模型来解决这一问题。为了建立这个模型，我们需要用连续的时间帧（time framework）（像 EOQ 模型那样）和以下这些模型假设：

1. 产品可以单独分析。产品间没有相互影响（例如共享制造资源）。
2. 需求每次产生一个。没有批量订单。
3. 没有满足的需求会变成延迟订单。不会有顾客损失。
4. 补货提前期是固定且已知的。配送提前期不存在随机性。（在本章后面的部分我们将告诉大家如何松弛这个假设，考虑变化的提前期）。
5. 一次订购一个产品。没有准备成本和订货次数限制之类的会导致批量订货的因素。

我们将在下一节的（Q，r）模型中松弛这最后一个假设，那时批量订货将会变得颇具吸引力。

我们将使用以下这些符号：

Q = 产量或订货量，这是决策变量。

l = 补货提前期（天），本节假设它为常数

X = 补货提前期内的需求，它是一个随机变量

$p(x)$ = $P(X=x)$= 补货提前期等于 x 时的需求（概率质量函数）。我们假设需求是离散的（即可计数的），但有时为了方便也会把它近似看成是连续分布。当把需求看作是连续时，我们假设一个密度函数 g（x）来代替概率质量函数。

$G(x)$ = $P(X\le x)=\sum_{i=0}^{x}p(i)$= 提前期内的需求小于或等于 x 时的需求（累积分布函数）

θ = $E[x]$，提前期 ℓ 内需求的均值

σ = 提前期 ℓ 内需求的标准差
h = 持有一年单位库存的成本（美元/单位、年）
b = 持有一年单位延迟产品的成本（美元/单位、年）
r = 再订货点，表示引发补货订单的库存水平，它是一个决策变量
R = $r+1$，基准库存水平
s = $r-\theta$，安全库存水平
$S(R)$ = 满足率（通过库存满足的订单的比率），它是 R 的函数
$B(R)$ = 平均延迟订单量，它是 R 的函数
$I(R)$ = 平均库存持有水平，它是 R 的函数（69|70）

当库存量剩下 r 时我们就下单补货，而在我们等待订货到达的时间里需求的期望为 θ，那么当订货到达时，我们手头持有库存的期望就是 $r-\theta$。如果 $s=r-\theta>0$，我们把它称为系统的**安全库存（safety stock）**，因为它可以防止需求或供货的变动性所造成的缺货。因为求解 $r-\theta$ 与求解 r 是等效的（因为 θ 是常数），所以可以把问题看作是求解最优的基准库存水平（$R=r+1$）以及再订货点 r 或安全库存水平（$s=r-\theta$）。

我们有两种方法可以用来求解最优基准库存水平。其中之一是按照之前我们一直使用的方法（在 EOQ、Wagner-whitin 和报童模型中所使用的）表示出成本函数然后解出使成本最小的再订货点。或者我们也可以先确定一个必要的顾客服务水平然后再找到达到这一水平的最小再订货点。我们将在下面把这两种方法都列示出来。但是首先需要建立绩效指标 $S(R)$、$B(R)$ 和 $I(R)$ 的表达式。

我们从分析基准库存规则下库存、再订货量和延迟订单量的关系开始。为此，我们要区分两种库存量：**库存持有量（on-hand inventory）**，表示实际的库存量（永远非负）；**库存量（inventory position）**，表示库存持有量、延迟订单量和再订货数量消减后的理论库存量，即

库存量 = 库存持有量 － 延迟订单量 ＋ 再订货量　　（2.18）

在基准库存策略下，只要有需求产生，我们就进行一次订货，因此总有

库存量= R　　（2.19）

通过（2.18）和（2.19）我们可以得到绩效指标的表达式。

服务水平（service lever）。首先单独分析一个补货订单。因为提前期是常量，所以我们知道在这次订货到达之前，所有的 R-1 个商品都可以用来满足新的需求。这样，订货晚于其需求到达的唯一可能情况就是提前期内的需求大于或等于 R（即，$X\geqslant R$）。因此订货满足新需求（即不产生延迟订单）的概率为 $P(X<R)=P(X\leq R-1)=G(R-1)=G(r)$。因为所有的订单关于这一等式都是类似的，那么通过库存来满足的需求的比例就等于订货先于其需求到达的概率，即

$$S(r)=G(R-1)=G(r) \qquad (2.20)$$

因此，$G(R-1)$ 就表示需求通过库存满足的比例。它被称作满足率，而且在许多库存控制系

统中被视为顾客服务的定义。

延迟订单水平（Backorder Lever）。在任何时候，订单数都等于最近 ℓ 时间内产生的需求数。如果令 X 表示这个需求（随机）数，那么根据（2.18）和（2.19）就可得（70|71）

$$\text{库存持有量} - \text{延迟订单量} = R - X \qquad (2.21)$$

由于库存持有量和延迟订单量不可能同时为正（因为如果同时存在库存和延迟订单，那么就会用库存来满足延迟订单，最后要么库存用完，要么延迟订单被满足）。因此，当订单量为 $X = x$ 时，延迟订单量就可以表示为

$$\text{延迟订单量} = \begin{cases} 0 & \text{若} x < R \\ x - R & \text{若} x \geq R \end{cases}$$

延迟订单量的期望可以通过求 x 的可能值的平均而得到

$$B(R) = \sum_{x=R}^{\infty} (x - R)\, p(x) \qquad (2.22)$$

表达式（2.22）在库存控制理论中是一个非常重要和有用的函数。因为它衡量了未满足需求（延迟订单量）的数量，它被称为损失函数。它可以通过式（2.22）计算得到，但是通常更加方便的是将其写成累积分布函数的形式：

$$B(R) = \theta - \sum_{x=0}^{R} [1 - G(x)] \qquad (2.23)$$

这个损失函数在（Q，r）模型中会再次出现。对于需求是泊松分布或者近似于（连续）正态分布的情况，则在附录 2B 中给出了的更为简单的电子表格计算公式。

库存水平（Inventory Level）。对式（2.21）两边分别求期望，注意到 $I(R)$ 表示期望的库存持有量，$B(R)$ 表示期望的延迟订单量，以及 $E[x] = \theta$ 为提前期需求的期望，这样我们就得到

$$I(R) = R - \theta + B(R) \qquad (2.24)$$

示例：

现在我们可以来分析Superior Appliance的例子了。根据以往的经验假设我们知道冰箱的需求均值为 10 台每月，而补货提前期为一个月。那么，提前期内的需求均值就是 $\theta = 10$ 台。进一步假设我们用泊松分布[6]表示需求，特别地，对于 k 和 x 的任意整数值，我们设（71|72）

$$p(R) = P\{\text{提前期需求} = \mathrm{R}\} = \frac{\theta^{\mathrm{R}} \mathrm{e}^{-\theta}}{R!} = \frac{10^{\mathrm{R}} \mathrm{e}^{-10}}{R!}$$

以及

[6] 泊松分布对于需求是一个一个产生而不存在周期性波动的情况下是一个很好的选择。它仅仅由一个参数来表征，即均值，因此当缺乏需求变动性的信息时用它就很方便了。泊松分布的标准差等于均值的平方根。

$$G(R)=\sum_{k=0}^{R}p(k)=\sum_{k=0}^{R}\frac{10^k e^{-10}}{k!}$$

有了这些我们也可以根据附录 2B 中的公式算出 $B(r)$。我们在表 2.5 中总结了结果。如果我们想要使满足率至少达到 90%，那么就必须有一个 R 值令 $G(R-1)\geqslant 0.9$。从表 2.5 我们可以看到应当有 R-1 = 14，即 R = 15，此时满足率为 91.7%。由于补货提前期内的需求均值是 10 台，也就是说设定了一个安全库存，即 $r-\theta$= 14 - 10 = 4 台。平均延迟订单量则是 B（15）= 0.103。平均库存水平则是

$$I(R)=R-\theta+B(R)=15-10+0.103=5.103$$

如果我们要把基准库存水平从 15 提高到 16，满足率就会上升到 95.1%，延迟订单量就会下降到 0.055，而平均库存水平则会增至 6.055。是否值得以这些额外的库存投资来改善顾客服务（由满足率和延迟订单量来衡量）则是 Superior Appliance 所要考虑的问题了。解决这种矛盾问题的一种方法就是利用一个最优成本模型，这正是我们下面要讨论的。

表 2.5 满足率与 R 的各种值

R	$p(R)$	$G(R)$	$B(R)$	R	$p(R)$	$G(R)$	$B(R)$
0	0.000	0.000	10.000	12	0.095	0.792	0.531
1	0.000	0.000	9.000	13	0.073	0.864	0.322
2	0.002	0.003	8.001	14	0.052	0.917	0.187
3	0.008	0.010	7.003	15	0.035	0.951	0.103
4	0.019	0.029	6.014	16	0.022	0.973	0.055
5	0.038	0.067	5.043	17	0.013	0.986	0.028
6	0.063	0.130	4.110	18	0.007	0.993	0.013
7	0.090	0.220	3.240	19	0.004	0.997	0.006
8	0.113	0.333	2.460	20	0.002	0.998	0.003
9	0.125	0.458	1.793	21	0.001	0.999	0.001
10	0.125	0.583	1.251	22	0.000	0.999	0.000
11	0.114	0.697	0.834	23	0.000	1.000	0.000

总的来说，补货提前期内需求均值越高，达到特定满足率所需要的基准库存水平也越高。这很简单，因为再订货点 r 必须保留足够的库存来满足订货到达以前产生的需求。如果提前期的需求分布是对称的（例如是钟形曲线），那么提前期需求超过 θ 的概率就是二分之一。这样，如果要是满足率高于二分之一，r 就必须大于 θ。

除了需求均值，需求的变动性同样影响着基准库存水平的选择。补货提前期需求的标准差越大，对于一个给定的满足率也就需要更大的 r。如果在前面的例子中，我们将 G（x）近似看做均值为 θ、标准差为 σ 的正态分布，则 σ 的不同会影响到表 2.5 中的结果。如果令 $\sigma=\sqrt{\theta}$，则会得到与泊松分布（泊松分布中标准差总是均值的平方根）相似的结果。σ 越大，对于不同的 r 的满足率就越低；而 σ 越小，则满足率越高。

基准库存模型在运作管理文献中曾被广泛研究过。这其中有一部分原因是它相对比较容易分析，同时也因为它容易被推广到各种情况。例如，基准库存可以用来控制多阶段生产线

中的下料。在这样的系统中，基准库存水平即相当于生产线中的库存缓冲（例如在工站前）。每当有工件从缓冲中离开，就会自动触发一个补货订单。正如我们将在第四章中讨论的，这一点在日本人的看板系统中是非常关键的。（72|73）

最终，我们来看一个设定基准库存水平的最优化方法。为此，我们将需求近似看做是一个连续分布 G（x），它的密度函数是 g（x）。这样我们就能写出包括库存持有成本和延迟订单成本在内的成本函数，即

$$Y（R）= \text{持有成本＋延迟订单成本} \tag{2.25}$$

$$= hI(R)+bB(R)$$

$$= hI(R-\theta+B(R))+bB(R)$$

$$= hI(R-\theta)+(b+h)B(R) \tag{2.26}$$

在下面这个技术性注释中我们给出了计算使$Y(R)$取得最小值的基准库存水平 R。

技术性注释

将 R 看做一个连续变量，这样就能对其进行求导：

$$\frac{dY(R)}{dR}=h+(b+h)\frac{dB(R)}{dR}$$

式（2.22）中订单延迟函数$B(R)$的连续表达式为

$$B(R)=\int_R^{\infty}(x-R)g(x)dx \tag{2.27}$$

这样就可计算 $dB(R)/dR$

$$\frac{dB(R)}{dR}=\frac{d}{dR}\int_R^{\infty}(x-R)g(x)dx$$

$$=-\int_R^{\infty}g(x)dx$$

$$=-[1-G(R)]$$

令 $dB(R)/dR$ 等于零则有

$$\frac{dY(R)}{dR}=h-(b+h)[1-G(R)]=0 \tag{2.28}$$

解式（2.28）即得 R 的最优值。

使得持有成本加延迟成本之和最小的基准库存水平 R 可以用下式表示

$$G(R^*)=\frac{b}{b+h} \tag{2.29}$$

可以看到这个公式与我们在报童模型的解（2.16）中看到的符号比例关系是一样的。这就说明了与最优基准库存水平相对应的满足率就是$b/(b+h)$。这个结果的意思也容易理解，因为增大持有成本会引起 R^*变小；而增加延迟订单成本 b 则会引起 R^*变大。注意当延迟订单成本和持有成本相等时，满足率就是二分之一，因此 $R^*=\theta$，即补货提前期内的需求均值，这时就没有安全库存。（73|74）

正如我们在报童模型中所做的，我们可以将式（2.29）简化为 G 是正态分布的情况。和得出式（2.17）的道理一样，可以写出

$$R^* = \theta + z\sigma \qquad (2.30)$$

这里 z 是$\Phi(z)=b/(b+h)$的正态分布表中的值，而 μ 和 σ 分别是提前期需求的均值和标准差。注意到当 z >0 时 R^*随 θ 递增而随 σ 递减。只要$b/(b+h)$>0.5，或 $b>h$，这种情况就一定满足，因为持有单位延迟订单成本通常总是要高于单位库存持有成本，所以往往最优基准库存水平是需求变动性的增函数。

示例：

让我们回到 Superior Appliance 的例子。为了将需求近似为一个连续分布，我们假设提前期需求是均值为θ=10 台每月、标准差为$\sigma=\sqrt{\theta}=3.16$台每月的正态分布。（选择令$\sigma=\sqrt{\theta}$可以使标准差与前例中的泊松分布相同。）假设冰箱的销售总额为 750 美元，公司以 2%每月来计算库存成本，因此 $h=0.02$（750）= 15 美元每月。再假设延迟订单成本估计为 25 美元每单位每月，这是因为该公司对于缺货的产品总是以折扣的形式来保证销售。

这样，要得到最优基准库存水平就需先算出 z

$$\frac{b}{b+h}=\frac{25}{25+15}=0.625$$

然后在正态分布表中找到 Φ（0.32）= 0.625。这样，$z=0.32$

那么

$$R^* = \theta + z\sigma = 10 + 0.32(3.16) = 11.01 \approx 11$$

利用表 2.5 我们就可以计算出在这个基准库存水平下的满足率 S（R）= G（R-1）= G（10）= 0.583。（注意即使我们是用一个连续模型来得到 R^*，我们仍然要用表 2.5 中的离散公式来计算实际的满足率。因为在现实中，冰箱的需求总是离散的。）这是一个相当低的满足率，这意味着我们设的延迟订单成本太低了。

如果我们把延迟订单成本增加到b = 200 美元，这时候临界比（Critical ratio）就会增至 0.93（$z_{0.93}$= 1.48），最优库存水平会增至R^* =10 + 1.48（3.16）= 14.67≈15。这是我们之前分析中为了达到 90%的满足率而得到过的基准库存水平，记得当时得到的实际满足率为 91.7%。这里我们可以看出两点：第一，将R圆整为 15 之后通过表 2.5 用泊松分布算出的实际满足率 91.7%总是要低于式 2.29 中的临界比所得到的 93%。因为连续的需求分布总是使库存看起来要比实际情况更有效率。第二，当基准库存水平为 15 而满足率大于 90%时，延迟订单成本总是非常大（200 美元每单位每月），这意味着虽然有如此之高的满足率却并不

经济。[7]（74|75）

我们总结了简单基准库存模型的主要思想如下：

1. 再订货点通过安全库存来控制出现缺货的概率。

2. 针对一个给定的满足率所需达到的基准库存水平（同时对应有**安全库存**）是补货提前期需求均值和标准差（当单位延迟订单成本超过单位库存持有成本时）的增函数。

3. “最优”的满足率是延迟订单成本的增函数而是库存持有成本的减函数。这样，如果我们固定库存持有成本，我们就可以用服务水平约束或延迟订单成本来决定合适的基准库存水平。

4. 多阶段生产系统中的基准库存水平与看板系统非常相似，因此以上的这些观点也适用于看板系统。

2.4.3 （Q，r）模型

我们设想这样一个情景，Jack 是一个维修经理，他必须保有空余的零件以用于设备维修。零件的需求是机器停机次数的函数，因此它本质上是不可预测（即随机）的。但是，和基准库存模型中的情况不同，假定采购成本（针对从外部供应商那里购得零件）或开机成本（针对内部生产的零件）大到无法一次只补充一个零件，这样的话，维护经理就不仅要决定持有多少库存（和基准库存模型一样），还要决定每次生产或订购多少零件（和 EOQ、报童模型一样）。（Q，r）模型的核心就是同时处理这两个问题。

从建模的角度看，（Q，r）模型所包含的假设与基准库存模型是一样的，除此之外还需要以下两个假设的其中之一

1. 补货订单具有固定的成本，或
2. 每年补货订单的次数是有约束的。

这样的话大于 1 的补货数量才有意义。

（Q，r）模型的基本机理如图 2.6 所示，图中显示了某一产品的净库存水平（持有库存减去订单延迟量）和库存量（净库存加上再订购量）的连续检查值。需求是随机产生的，但是我们假设一次只产生一个需求，这就是为什么在图 2.6 中净库存总是一次下降一个单位。当库存量落到再订货点 r 时，就下一个数量为 Q 的补货订单。（注意因为订单是库存量一到 r 就产生，所以库存量立即跃升至 $r+Q$，这样库存量处于 r 水平的持续时间实际上也就是 0。）经过一个（固定的）提前期 ℓ 后订购的货物到达，在此期间则可能出现缺货。问题就是确定合适的 Q 和 r 值。（75|76）

[7] 部分原因是 b 必须大到使 R=15，这是我们将 R 取整的最接近的整数的情况。如果我们想要服务水平至少为 $b/(b+h)$，那么就应该向上取整，这样即使取 b=135 美元（仍然很大），得 $b/(b+h)$=0.9，这时 R=14.05 向上取整也可以得到 15。因为连续分布始终只是对需求的近似，因此是否有大的 b 值或者用向上取整的方法来得到最终结果都并不重要，关键是使用者是否能通过灵敏度分析来理解得到的结果及其影响。

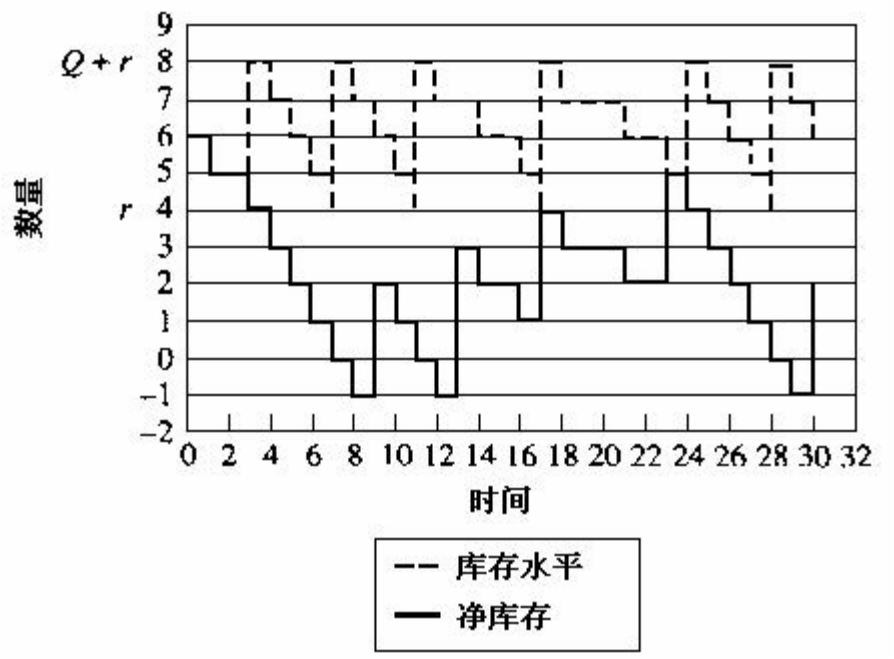

图 2.6 $Q=4$、$r=4$ 时（Q，r）模型的净库存与库存量

正如 Wilson（1943）在（Q，r）模型第一次正式发表时指出的那样，Q 和 r 这两个控制变量有着截然不同的目的。在 EOQ 模型中，订货数量 Q 影响着生产或订货频率与库存之间的权衡。数值较大 Q 会造成较小的补货次数和较高的库存量；较低的 Q 值会产生较低的库存量和较多的补货次数。与此不同，再订货点 r 则影响着出现缺货的可能性。较高的再订货点会有较高的库存但是缺货的可能性较低；较低的再订货点可以减少库存，但代价是出现缺货的可能性更大。

根据成本和顾客服务的表达形式，我们可以发现 Q 和 r 可以共同对库存、生产/订货频率或顾客服务产生影响。然而，我们必须认识到这两个参数产生的是两种根本不一样的库存，这一点是非常重要的。补货数量 Q 影响的是周期库存（即为了避免补货太过频繁而持有的库存）。再订货点 r 影响的是安全库存（即防止缺货所持有的成本）。注意到在这些定义下，EOQ 模型中所有的库存都是周期库存，而基准库存模型中所有的库存都是安全库存。从某种意义上讲，（Q，r）模型是这两个模型的综合。

为了表示基本的（Q，r）模型，我们将 EOQ 和基准库存模型中的成本结合到一起，也就是我们要决定 Q 和 r 的值以求解

$$\min_{Q,r}\{\text{固定准备成本+延迟成本+持有成本}\} \qquad (2.31)$$

或

$$\min_{Q,r}\{\text{固定准备成本+缺货成本+持有成本}\} \qquad (2.32)$$

公式（2.31）和（2.32）的区别在于顾客服务的表达方式不同。延迟订单成本（backorder cost）假设的费用是对于每个未被满足的客户订单随时间而增加的，而缺货成本则假设的是每次需求未由库存满足时所发生固定的费用（不考虑订单延误的持续时间）。我们将在下面的分析中同时使用这两种方法。（76|77）

标记。为了表示以上所讨论的各种成本，我们用到以下这些符号：

D = 每年期望需求量（件）

l = 补货提前期（天），起初假设它为常数，而在这一部分结束的时候我们将介绍如何引入变动提前期

X = 补货提前期内的需求，它是一个随机变量

θ = $E[x]=Dl/365$，提前期 ℓ 内需求的均值

σ = 提前期 ℓ 内需求的标准差

$p(x)$ = $P(X=x)$= 补货提前期等于 x 时的需求（概率密度函数）。我们假设需求是离散的（即可计数的），但有时为了方便也会把它近似看成是连续分布。当把需求看作是连续时，我们假设一个密度函数 g（x）来代替概率质量函数。

$G(x)$ = $P(X\le x)=\sum_{i=0}^{x}p(i)$= 提前期内的需求小于或等于 x 时的概率（累积分布函数）

A = 准备或采购成本（美元）

c = 单位制造成本（美元）

h = 持有一年单位库存的成本（美元/件·年）

k = （每次）缺货成本（美元）

b = 持有一年单位延迟产品的成本（美元/单位、年）；注意当出现无法使库存满足需求的情况时将通过 k 或者 b 中的一个来进行惩罚，两者不能同时使用

Q = 补货批量（件）；这是一个决策变量

r = 再订货点，表示引发补货订单的库存水平，它是一个决策变量

s = $r-\theta$，安全库存水平（件）

$F(Q,r)$ = 订货频率（每年订货次数），是 Q 和 r 的函数

$S(Q,r)$ = 满足率（由库存成功满足的订单占到的比例），是 Q 和 r 的函数

$B(Q,r)$ = 平均延迟订单数，是 Q 和 r 的函数

$I(Q,r)$ = 平均库存持有水平（件），是 Q 和 r 的函数

成本

固定准备成本（Fixed Setup Cost）。要体现订货量批大于一的必要性可以有两种基本的方法。第一种是限制每年订货的次数。因为每年订货次数是

$$F(Q,r)=\frac{Q}{D} \tag{2.33}$$

如果给定订货频率 F，我们就能根据 $Q=D/F$ 算出 Q。另一种方法是每次订货都产生一个固定成本 A，这样，年订货成本就是 $F(Q,r)A=(D/Q)A$。（77|78）

缺货成本（Stockout Cost）。正如我们之前注意到的，有两种基本的方法来惩罚低客户服务水平。一种是每当需求不能通过库存满足（即缺货）时就产生一个固定成本。另一个是根据

顾客订单等待被满足的时间产生一个与之成正比的惩罚成本。

年缺货成本与每年缺货的平均次数成正比，即 $D\left[1-S\left(Q,r\right)\right]$。根据图 2.6 中库存量只能取 r+1, r+2, … , r+Q（注意不能取 r 值，因为每当其落到 r，马上就会产生一个数量为 Q 的订单），我们可以算出 $S\left(Q,r\right)$。事实上，从长期来看，库存量可能取到这一区间内任何一个值，并且取各个值的可能性是一样的。考虑到这一点，我们就可以在下面的分析中使用基准库存模型中的结果（参见 Zipkin 1999 对此进行的严格推导）。

假设我们在系统[8]运行了很长时间之后观察到库存量是x。这意味着我们我们手头持有的加上订购的库存足以满足x个产品的需求。因此我们要问，第x+1 个需求被库存满足的概率有多大？这个问题的答案与基准库存模型的完全一样。即，因为所有已经下单的订货都会在补货提前期之内到达，第x+1 个产品的需求无法被库存满足的唯一可能情况就是在补货提前期内出现的需求大于等于x。根据我们在基准库存模型中的分析，我们知道缺货的可能性为

$$\begin{aligned} P\{X \geq x\} &= 1-P\{X < x\} \\ &= 1-P\{X \leq x-1\} \\ &= 1-G\left(x-1\right) \end{aligned}$$

这样，给定库存量 x 的情况下，满足率等于 1 减去缺货的概率，即 $G\left(x-1\right)$。由于这 Q 个可能的库存量的取值概率相同，因此整个系统的满足率就可以通过所有可能的库存量下满足率的平均值，即

$$S(Q,r)=\frac{1}{Q}\sum_{x=r+1}^{r+Q}G(x-1)=\frac{1}{Q}[G(r)+\cdots+G(r+Q-1)] \tag{2.34}$$

我们可以用式（2.34）直接算出任意给定（Q，r）所对应的满足率。然而，将这个公式变化一下可以使运算更为简便。由于基准库存延迟订单水平函数 B（R）可以写成式（2.23）那样的累计分布函数的形式，这样就可以直接写出（Q，r）模型中满足率的表达式：

$$S(Q,r)=1-\frac{1}{Q}[B(r)-B(r+Q)] \tag{2.35}$$

这个 $S\left(Q,r\right)$ 的表达式在电子表格中计算起来十分简单，尤其是用附录 2B 中给出的公式。然而，有时候使用解析表达式比较困难，因此就产生很多近似方法。有一个叫做**基准库存**或**第一类服务（type I service）**近似公式简单的表示了满足率的基准库存公式，即

$$S(Q,r)\approx G(r) \tag{2.36}$$

从式（2.34）来看，很显然 $G\left(r\right)$ 低估了实际的满足率。这是因为累积分布函数（cdf）$G\left(x\right)$ 是 x 的增函数。这样，我们取到的就是均值的最小值。然而，尽管它可能会严重低估实际的满足率，但是由于它只和 r 有关，因此用起来十分简单。以它为基准，可以使用一种非常有

[8] 这一技术被称为随机事件（即库存量的值）检验（*conditioning*），它是概率论领域非常有力的一种分析工具。

效的探索发现方法来计算出有效的（Q，r）策略，这正是我们在后面要介绍的。（78|79）

另一个近似得到满足率的方法被称**为第二类服务（type II service）**，即忽略式（2.35）中的第二项（Nahmias 1993）。这就是

$$S(Q,r) \approx 1 - \frac{B(r)}{Q} \tag{2.37}$$

同样，这个近似算法也有低估真实满足率的倾向，因为式（2.35）中的项 $B(r+Q)$ 为正。然而，由于这一近似算法仍然同时与 Q、r 有关，因此用起来也就并不比原式简单了。但是正如我们下面将看到的，对于推出一个再订货点公式，它是一个非常有用的中间变换。

延迟成本（Backorder Cost）。如果不是用一个固定成本 k 来惩罚每次缺货，而是用延迟订单的等待时间来进行惩罚，那么年延迟订单成本就会与平均订单延迟水平 $B(Q,r)$ 成正比。计算 $B(Q,r)$ 的值可以用和满足率相似的方法得到，即通过对基准库存模型中所有 r+1 到 r+Q 之间的库存量对应的订单延迟水平求平均值，即

$$B(Q,r) = \frac{1}{Q}\sum_{x=r+1}^{r+Q} B(x) = \frac{1}{Q}[G(r+1) + \cdots + G(r+Q)] \tag{2.38}$$

这一表达式同样可以直接使用，或者也可以转换为更简单的形式（见附录 2B），以便于使用电子表格计算。和 $S(Q,r)$ 的表达式一样，有时候用一个较简单的不含 Q 的表达式来进行近似计算会更为方便。例如用类似第一类服务的式子和使用基准库存订单延迟公式

$$B(Q,r) \approx B(r) \tag{2.39}$$

注意如果要准确依照满足率的第一类近似算法，就应该取式（2.38）的最小项，即 B（r+1）。由于差别不大，用 B（r）会更简单一点。原因是用一个连续函数近似表示需求的时候也总是这么做的；在这一假设下，基准库存模型订单延迟的表达式真的变成了 $B(r)$（而非 $B(R)$）。

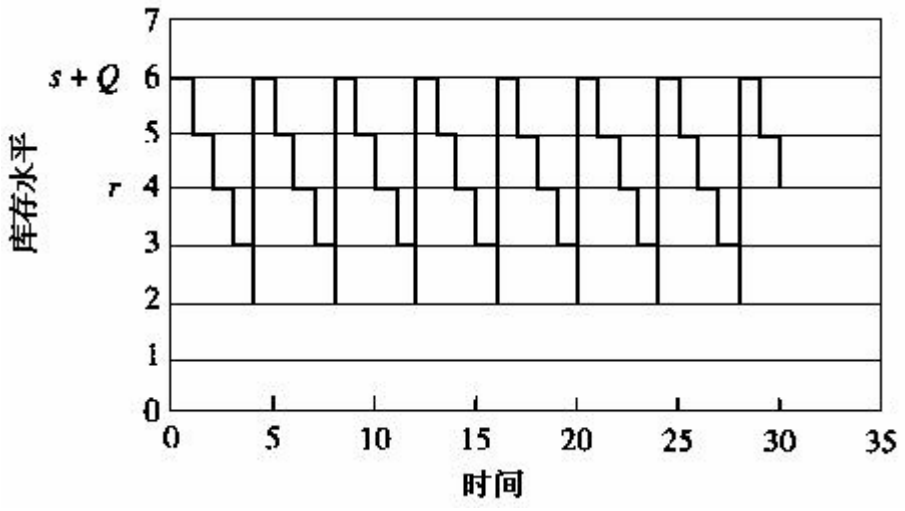

图 2.7 $Q=4$、$r=4$、$\theta=2$ 的（Q，r）的期望库存与时间

持有成本（Holding Cost）。问题（2.31）和（2.32）中的最后一项成本是库存持有成本，可以用 $hI(Q,r)$ 来表示。我们可以通过查看平均净库存然后假设需求是确定的（如图 2.7 所示，这里描述了一个 $Q=4$、$r=4$、$\ell=2$、$\theta=2$ 的系统），从而近似表示 $I(Q,r)$。需求完全确定，因此每次库存到达再订货点（$r=4$）的时候就会下一个订单，订货会在两个单位时间后到达。由于订单恰好是在补货周期最后一个需求产生时到达，所能达到的最低库存水平也就是 $r-\theta+1=s+1=3$。总而言之，在这些确定性的条件下，库存会在每个补货周期从 $Q+s$ 降至 $s+1$。这样，平均的库存就是

$$I(Q,r)\approx\frac{(Q+s)+(s+1)}{2}=\frac{Q+1}{2}+s=\frac{Q+1}{2}+r-\theta \tag{2.40}$$

然而在现实中，需求是变化的，有时会造成缺货。由于持有的库存不可能低于 0，上面确定性的近似公式也就低估了实际的平均库存，低估的量与平均延迟订单水平有关。这样，实际的表达式就是（80|81）

$$I(Q,r)=\frac{Q+1}{2}+r-\theta+B(Q,+r) \tag{2.41}$$

延迟成本法（Backorder Cost Approach）。我们现在可以把用来示意的（2.31）式转化为数学模型了。准备成本与采购成本、延迟订单成本和库存持有成本之和可以写作

$$Y(Q,r)=\frac{D}{Q}A+bB(Q,r)+hI(Q,r) \tag{2.42}$$

不幸的是，成本函数 Y（Q，r）中有两大难点。第一个是成本参数 A 和 b 在现实中是难以估计的。尤其是延迟订单成本几乎不可能明确表示，因为它涉及到诸如顾客信誉和公司声誉的无形损失。然而幸运的是，实际的目标并不是要使成本最小化，而是要在生产准备、服务和库存之间取得一个合理的平衡。利用一个成本函数可以使我们方便地利用最优化工具来推导出 Q、r 关于各个问题参数的表达式。但是所得策略的质量必须直接以绩效指标的形式来衡量，这一点我们在下个例子中就会介绍。$B(Q,r)$ 和 $I(Q,r)$ 的表达式同时与 Q、r 密切相关。因此使用这些确切的表达式无法使我们得到更为简单的 Q、r 表达式。这样，为了得到更容易处理的表达式，我们用式（2.39）来近似表示 $B(Q,r)$，而用它代替实际的 $B(Q,r)$ 表达式，$I(Q,r)$ 也是如此。通过这样的近似处理，我们的目标函数变成了

$$Y(Q,r)\approx\tilde{Y}(Q,r)=\frac{D}{Q}A+bB(r)+h\left[\frac{Q+1}{2}+r-\theta+B(r)\right] \tag{2.43}$$

我们在下面的技术性注释中来计算使 $\tilde{Y}(Q,r)$ 取得最小的 Q 和 r 的值。

技术性注释

将 Q 看作是一个连续变量，求 $\widetilde{Y}(Q,r)$ 对 Q 的微分，令结果等于 0，即（80|81）

$$\frac{\partial \widetilde{Y}(Q,r)}{\partial Q} = -\frac{DA}{Q^2} + \frac{h}{2} = 0 \tag{2.44}$$

用一个密度函数为 g（x）的连续分布近似表示提前期需求，求 $\widetilde{Y}(Q,r)$ 对 r 的微分，令结果等于 0，即

$$\frac{\partial \widetilde{Y}(Q,r)}{\partial r} = (b+h)\frac{dB(r)}{dr} + h = 0 \tag{2.45}$$

和基准库存的情况一样，由于 B（r）函数的连续模拟量为

$$B(r) = \int_r^{\infty} (x-r)g(x)dx$$

可以算出 B（r）的导数为

$$\begin{aligned}\frac{dB(r)}{dr} &= \frac{d}{dr}\int_r^{\infty}(x-r)g(x)dx \\ &= -\int_r^{\infty} g(x)dx \\ &= -[1-G(r)]\end{aligned}$$

然后将式（2.45）重写为

$$-(b+h)[1-G(r)] + h = 0 \tag{2.46}$$

这样，我们必须解出式（2.44）和式（2.46）以取得 $\widetilde{Y}(Q,r)$ 的最小值，如式（2.47）和式（2.48）所示。

最优的订货批量 Q^* 及再订货点 r^* 如下所示：

$$Q^* = \sqrt{\frac{2AD}{h}} \tag{2.47}$$

$$G(r^*) = \frac{b}{b+h} \tag{2.48}$$

注意到 Q^* 是由 EOQ 公式给出的，而 r^* 是由基准库存模型的临界比公式给出的。（后者并不奇怪，因为我们用基准库存的形式近似表示延迟订单水平。）如果我们进一步假设提前期需求是均值为 θ、标准差为 σ 的正态分布，那么我们就可以和式（2.30）中的基准库存模型一样把式（2.48）简化为

$$r^* = \theta + z\sigma \tag{2.49}$$

此处 z 为标准正态表中使 $\Phi(z) = b/(b+h)$ 的值。

记住 Q^*和 r*的值只是近似值，这一点很重要。因此我们应该用实际的公式来检查它们的表现情况（例如通过平均库存、满足率、订单频率和延迟订单水平）。如果表现不佳，那么可以调整成本参数。通常情况下，可以不管持有成本 h 而调整固定订货成本 A 和延迟订单成本 b，因为它们相对而言比较难以进行进一步的估计。注意当增大 A 会增大 Q^*而降低平均订购频率，而增大 b 则会增大 r^*而降低缺货率和平均延迟订单水平。我们将在 82 页的例子中说明这一点。

缺货成本法（Stockout Cost Approach）。作为延迟订单法的替代方法，我们可以将示意性的式（2.32）转化为数学模型，即写出年准备或采购成本、缺货成本、库存持有成本之和（81|82）

$$Y(Q,r) = \frac{D}{Q}A + kD\left[1 - S(Q,r)\right] + hI(Q,r) \tag{2.50}$$

这个成本函数包含的参数和延迟订单模型的情况一样难以界定。尤其是缺货成本 k 和延迟订单成本一样取决于无形因素（损失顾客信誉和公司声誉）。这样，这个成本函数同样也仅仅是推出 Q 和 r 的表达式以便于合理平衡生产准备、服务和库存的一种手段。它本身不是一种绩效指标。

和延迟订单模型一样，缺货模型的成本函数包含 $S(Q,r)$和 $I(Q,r)$，也就同时与 Q、r 有关，因此也就无法得到简单的表达式。所以我们将进行一个两阶段近似以产生一个关于 Q、r 的闭合解表达式（Closed-form expressions）。

首先，类似于在上面的延迟订单成本模型中所做的，我们假设 Q 对于满足率 $S(Q,r)$的影响以及库存项 $I(Q,r)$中的延迟订单修正因素 $B(Q,r)$可以忽略不计。这就得到了熟悉的订货批量 EOQ 公式

$$Q^* = \sqrt{\frac{2AD}{h}}$$

其次，为了得到再订货点的表达式，我们在式（2.50）中进行了两次近似。我们用式（2.37）代替服务水平 $S(Q,r)$的第二类近似，用基准库存式（2.39）近似表示库存项中的延迟订单修正项 $B(Q,r)$。这样就得到了下面这个近似成本函数

$$Y(Q,r) \approx \tilde{Y}(Q,r) = \frac{D}{Q}A + kD\frac{B(r)}{Q} + h\left[\frac{Q+1}{2} + r - \theta + B(r)\right] \tag{2.51}$$

采用常规的求最优化步骤（对 r 求导，令其等于 0，然后求解 r）得到下面的最优再订货点的表达式

$$G(r^*) = \frac{KD}{KD + hQ} \tag{2.52}$$

如果我们进一步假设提前期需求是均值为 θ、标准差为 σ 的正态分布，那么我们就可以把再订货点的表达式简化为

$$r^* = \theta + z\sigma \qquad (2.53)$$

其中$\Phi(z) = kD/(kD + hQ)$。

注意和式（2.49）不同，式（2.53）对 Q 是敏感的。尤其当 Q 增大时 $kD/(kD + hQ)$ 的比率就会变小，这样 r*也就变小。原因就是 Q 变大会增大满足率（因为达到再订货点的频率降低了），这样为了达到一个给定的服务水平就需要有一个更小的再订货点。

示例：

维护经理 Jack 收集的历史数据表明他保管的其中一种备用零件的年需求（D）为 14 单位每年。零件的单位成本 c 是 150 美元，因为公司使用的利息率为 20%，这样每年的持有成本 h 就是 0.2（150）= 30 美元每年。补货提前期为 45 天，因此补货提前期内的平均需求为（82|83）

$$\theta = \frac{14}{365} \times 45 = 1.726$$

这种零件是从外部供应商处采购的，Jack估计下一个采购订单所需要的时间和物料成本A为 15 美元。剩下的成本就是延迟订单成本和缺货成本的其中之一。让Jack估算这些成本一定令他很不愉快，但经过一番劝说之后，他猜测每年的缺货成本大约是b = 100 美元/每年，每次缺货的成本大约是k = 40 美元[9]。最后，Jack认为模型的需求函数应该为泊松分布[10]。

不论我们用订单延迟成本模型还是缺货成本模型，定购数量都可以用式（2.47）来计算，得到

$$Q^* = \sqrt{\frac{2AD}{h}} = \sqrt{\frac{2(15)(14)}{30}} = 3.7 \approx 4$$

我们可以用订单延迟成本模型或缺货成本模型计算再订货点。为了能够使用延迟订单模型中正态需求的表达式(2.49)，我们将泊松分布近似为均值 θ = 1.726，标准差 $\sigma = \sqrt{1.726} = 1.314$ 的正态分布。其临界比为

$$\frac{b}{b+h} = \frac{100}{100+30} = 0.769$$

从标准正态分布表中可以查到$\Phi(0.736) = 0.769$，因此 z = 0.736，而

$$r^* = \theta + z\sigma = 1.726 + 0.736(1.314) = 2.693 \approx 3$$

正如我们前面提过的，这个方案的好坏不应该由成本函数或近似的绩效指标来判断。而是应该是由订货频率、满足率、延迟订单水平和库存水平的确切计算结果来衡量。要计算这些变量，我们要先算出 p（r）、G（r）和 B（r）。我们可以利用附录 2B 中泊松分布的需求计算公式，表 2.6 中汇总了计算的结果。利用这些结果就可以计算出（Q = 4，r = 3 时）订货频率，满足率，延迟订单水平和平均库存水平：（83|84）

[9] 注意到无论是惩罚订单延迟还是缺货，都假设其成本不受机器状况的影响。当然在现实中，使用频繁的关键机器的缺货成本要远大于偶尔使用、有剩余产能的机器。

[10]当需求是产生自许多独立的来源（例如不同机器的停机）时，泊松分布是一个很好的假设。然而，如果需求是产生自更加有规律的过程，例如有日程安排的预防性维护，泊松分布就会高估变动性，因此导致保守（并且可能过量）的安全库存策略。

$$F(Q,r)=\frac{D}{Q}=\frac{14}{4}=3.5$$

$$S(Q,r)=1-\frac{1}{Q}[B(r)-B(r+Q)]$$

$$=1-\frac{1}{4}[B(3)-B(3+4)]$$

$$=1-\frac{1}{4}[0.140-0.001]$$

$$=0.965$$

$$B(Q,r)=\frac{1}{Q}\sum_{x=r+1}^{r+Q}B(x)$$

$$=\frac{1}{4}[B(4)+B(5)+B(6)+B(7)]$$

$$=\frac{1}{4}(0.042+0.011+0.003+0.001)$$

$$=0.014$$

$$I(Q,r)=\frac{Q+1}{2}+r-\theta+B(Q,r)$$

$$=\frac{4+1}{2}+3-1.726+0.014)$$

$$=3.79$$

表 2.6 $\theta=1.726$ 时的 p（r）、G（r）和 B（r）

r	$p(r)$	$G(r)$	$B(r)$
0	0.178	0.178	1.726
1	0.307	0.485	0.904
2	0.265	0.750	0.389
3	0.153	0.903	0.140
4	0.066	0.969	0.042
5	0.023	0.991	0.011
6	0.007	0.998	0.003
7	0.002	1.000	0.001
8	0.000	1.000	0.000
9	0.000	1.000	0.000
10	0.000	1.000	0.000

作为延迟订单成本模型的替代方法，我们可以用缺货成本模型中的式（2.53）来计算再订货点。公式中的临界比为

$$\frac{KD}{KD+hQ}=\frac{40(14)}{40(14)+30(4)}=0.824$$

从正态分布表中可以查出 Φ（0.929）= 0.824，所以 z = 0.929，而

$$r^{*}=\theta+z\sigma \text{=1.726+0.929（1.314）=2.946}\approx 3$$

由于这个订货策略（Q = 4，r = 3）与延迟订单成本模型的结果一样，所以绩效指标也是一样的。因此实际上，Jack 选择的延迟订单成本和缺货成本是相等的。在单一产品的情况下，任何一种模型都是可以使用的——增加 b 或者 k 都会引起客户服务水平的增加并且延迟订单水平（代价是更高的存货水平）的减少。所以通过改变成本参数，这两个模型都可以得到有效的方案。但我们在十七章中将会看到，在多产品系统中，这两种模型是有区别的。

由现有成本系数产生的这个策略要求每年要补货 3.5 次，满足率处于相当高的水平（96.5），同时也会有少量的延迟订单（平均仅为 0.014），手头持有的库存平均略低于 4 个单位（3.79）。决策者看到这些数值可能会觉得这是个不错的策略。如果不是，那么就需要用灵敏度分析来找出更多的方案了。

举例来说，假设决策者觉得一年 3.5 次的补货订单太少了，在给定采购部门能力的情况下，每年 F = 7 次是可以做到的。那么我们可以算出 $Q = D/F$ =14/7 =2。但如果我们仍然设定再订货点为 r = 3，那么满足率就变为（84|85）

$$S(Q,r)=1-\frac{1}{Q}\left[B(r)-B(r+Q)\right]=1-\tfrac{1}{2}(0.140-0.011)=0.936$$

这样的满足率对于一个维修备件来说太低了。如果我们把再订购点增加到 r = 4，那么满足率将变为

$$S(Q,r)=1-\frac{1}{Q}\left[B(r)-B(r+Q)\right]=1-\tfrac{1}{2}(0.042-0.003)=0.980$$

对于这个新策略（Q = 2，r = 4），我们可以很容易的算出延迟订单水平和平均库存水平，即

$$\begin{aligned}B(r)&=\frac{1}{Q}\left[B(5)+B(6)\right]\\&=\frac{1}{2}\left[0.011+0.003\right]\\&=0.007\end{aligned}$$

$$\begin{aligned}I(Q,r)&=\frac{Q+1}{2}+r-\theta+B(Q,r)\\&=\frac{2+1}{2}+3-1.726+0.007\\&=3.78\end{aligned}$$

相对于原有策略（Q = 4，r = 3）而言，目前增加的再订货点降低了延迟订单率，而增加的订货频率减少了平均库存水平。当然，这样做的代价就是每年额外要补货 3.5 次。

灵敏度分析有一种替代的方法，即调整固定订货成本 A，直到订货频率 $F(Q,r)$ 满足要求，然后调整延迟订单成本 b 或缺货成本 k（取决于选用的是哪一个模型），直到满足率 $S(Q,r)$ 和延迟订单水平 $B(Q,r)$ 可以接受。像这样的单产品问题中，因为我们仍然是在两个

变量（即 A 和 b 或者 k）上搜索答案，因此使用这种方法并没有特别大的优势。但是正如我们在第十七章中将会看到的，这种方法在多产品问题上要有效率得多，在多产品问题中对于每一个产品都可以搜索一个对应的（A，b）或（A，k）组合。不仅如此，由于式（2.47）、式（2.49）和式（2.53）都是涉及问题数据的简单闭合解等式，他们在电子表格中计算起来会极其方便。

为提前期变动性建模（Modeling Lead-Time Variability）。在对基准库存和（Q，r）模型的讨论中，我们假设补货提前期 ℓ 是固定的。模型中的所有变动性我们都假设是由需求变化引起的。然而，在许多实际情况中，提前期也是具有变动性的。例如，某个部件供应商有时可能会推迟（或提早）发货。这个新增的变动性的最主要作用就是放大了补货提前期需求的标准差 σ。通过推出考虑提前期变动性的 σ 的公式，我们就可以很容易地把这个额外的变动性引入到基准库存和（Q，r）模型中。

为了推导这个的公式，我们必须引入一些额外的符号：

L = 补货提前期（用周期数表示），它是一个随机变量

l = $E[L]$= 补货提前期的期望（用周期数表示）（85|86）

σ_L = 提前期内需求的标准差

D_t = 第 t 日的需求，是一个随机变量。我们假设需求不变，故 D_t 独立同分布

d = $E[D_t]$= 日需求期望

σ_D = 日需求的标准差

和前面一样，我们令 X 代表补货提前期内的（随机）需求量。利用上面所示的符号，它可以表示为

$$x = \sum_{t=1}^{L} D_t \tag{2.54}$$

因为日需求独立同分布，我们可以计算出补货提前期内的需求期望为

$$E[X] = E[L]E[D_t] = ld = \theta \tag{2.55}$$

这和我们之间的结果一样。然而，可变提前期改变了补货提前期内需求的方差。利用独立同分布随机变量的和的基本公式，我们可以计算出

$$Var(X) = E[L]Var(D_t) + E[D_t]^2 Var(L) = l\sigma_d^2 + d^2\sigma_L^2 \tag{2.56}$$

虽然式（2.56）里的“单位”看起来有问题（第一项的单位是时间，而第二项的单位是时间的平方），但其实这些项实际上都是无量纲的。原因就是 L 被定义为一个随机变量，它表示的是周期的数量而不是周期本身。尽管 L 的均值和方差都不一定是整数，但是随机变量实际得到的数值却一定是整数。例如统计提前期天数的时候，我们可能观察到天数为 5 天、6 天或 3 天，因而得到它们的均值为 4.667 天。然而，如果是 5/7 星期、6/7 星期和 3/7 星期就

不行。因此，提前期需求的标准差为

$$\sigma = \sqrt{\mathrm{Var}(X)} = \sqrt{\ell\sigma_D^2 + d^2\sigma_L^2} \qquad (2.57)$$

为了对公式（2.57）的性质有更好的了解，再来考虑需求泊松分布的情况。因为泊松分布随机变量的标准差总是均值的平方根，即 $\sigma_D = \sqrt{d}$ 。将其代入式（2.57），得

$$\sigma = \sqrt{\ell d + d^2\sigma_L^2} = \sqrt{\theta + d^2\sigma_L^2} \qquad (2.58)$$

注意到如果 $\sigma_L = 0$，也就是说补货提前期是常量，那么上式就简化为 $\sigma = \sqrt{\theta}$ ，这正是我们在泊松分布的情况所得到的结果。如果 σ_L >0，式（2.58）将会放大 σ 值，使其比常量提前期情况下的值要大。

为了在库存模型中举例说明上式的使用，让我们来回顾 2.4.2 节 Superior Appliance 的案例。在那个案例中我们假设对冰箱的需求服从均值为 10 台/月的泊松分布，提前期为一个月（30 天）。所以平均日需求量为 d =10/30 = 1/3。因为我们假设需求服从泊松分布，所以我们可以用式 2.58 来计算 σ。使用和 2.4.2 节中一样的库存持有成本和延迟订单成本，即 h = 15，b = 25，从而得到临界比为 b /（h+b）= 25 /（15+25）= 0.625。因此最优基准库存水平（假定需求近似正态分布）为（86|87）

$$R^* = \theta + z\sigma = \theta + z\sqrt{\theta + d^2\sigma_L^2}$$

如果 $\sigma_L = 0$，那么就得到 R^* = 11.01，这和我们以前得到的结果一样。如果 $\sigma_L = 30$（即补货提前期的变动太大以致于标准差等于均值），则得到 R^* = 13.34。要达到同样的客户服务水平，就需要用额外的 3.33 单位库存来应对更加不确定的需求。

式（2.57）或式（2.58）同样可以用来放大（Q，r）模型中式（2.49）或（2.53）的再订货点，以应对变动的补货提前期。

基本（Q，r）模型的意义（Basic （Q，r） Insights）。撇开所有数学和建模的复杂性，（Q，r）模型背后的基本原理本质上和 EOQ 模型、基准库存模型的是一样的，即

周期库存随补货频率减少而增加。

以及

安全库存为缺货提供缓冲。

（Q，r）模型将这些原理整合到了一个统一的框架中去。

（Q，r）模型（包括基准库存模型这个特例，即 Q = 1 的（Q，r）模型）是历史上最早对需求处理过程中的变动性进行建模的尝试之一，并且就安全库存如何影响客户服务水平提供了定量化的理解。根据初步的直觉，这个模型指出：安全库存、服务水平和延迟订单水平都主要受再订购点 r 影响，而周期库存和订购频率则本质上是补货批量 Q 的函数。然而模型的数学运算表明，实际情况是甚为微妙的。正如我们上面所看到的，服务水平、延迟订单水平的表达式同时由 Q 和 r 决定。理由是如果 Q 值很大，因而零件以低频率大批量的方式

进行补货，那么库存水平落到再订购点的次数就很少，因此很少可能会缺货。另一方面，如果 Q 值很小，那么库存水平就会经常落到再订购点，因此出现缺货的可能性就更大。

除了这些定性分析，（Q，r）模型还提供了关于影响最优库存策略的各种因素的一些定量分析。从近似公式（2.47）、（2.49）和（2.53）我们可以得出以下结论：

1. 增大平均年需求 D 将增大最优订货批量 Q。

2. 增大补货提前期内的平均需求 θ 将增大最优再订货点 r。注意到无论增加年需求量 D 或补货提前期 ℓ 都将增大 θ。这说明无论是需求大还是补货提前期长都要求持有更多的库存。增大需求的变动性σ将增大最优再订货点r。[11]这里最关键的一点就是高度变化的需求比起十分稳定的需求，往往需要更多的安全库存以作为预防缺货的保护手段。

3. 增大库存持有成本 h 将减小最优补货批量 Q 和再订货点 r。注意到库存持有成本可以通过增加产品成本、存货相关利率或其他持有成本（如搬运或损坏）而增加。关键在于持有库存的成本越高，持有的数量就应该越少。（87|88）

（Q，r）模型是一个很好的方法，它既提供有力的总体思想、又提供有用的实用工具。因而它是制造经理技能中的一个基本组成部分。

表 2.7 *库存模型的分类*

建模决策	模型					
	EOQ	**EPL**	**WW**	**NV**	**BS**	**（Q，r）**
连续（C）或离散（D）时间	C	C	D	D	C	C
单产品（S）或多产品（M）	S	S	S	S	S	S
单周期（S）或多周期（M）	/	/	M	S	/	/
延迟订单（B）或损失销售（L）	/	/	/	L	B	B
换模成本与订购成本（Y/N）	Y	Y	Y	N	N	Y
确定（D）或随机（R）需求	D	D	D	R	R	R
确定（D）或随机（R）产能	D	D	D	D	D	D
常量（C）或动态（D）需求	C	C	D	/	C	C
有限（F）或无限（I）产出率	I	F	I	/	I	I
有限（F）或无限（I）计划期	I	I	F	F	I	I
单级（S）或多级（M）	S	S	S	S	S	S

2.5 结论

虽然这一章内容涉及到一系列库存建模方法，但是我们只是刚刚触及了库存——这一运作管理（OM）大分支的皮毛。库存系统的复杂性和多变性已经促使了许多模型的产生。表 2.7 总结了用以区分模型的一些维度，并且把我们在这一章看到的模型归为了五类（即 EOQ 模型、Wagner-Whitin（WW）模型、报童模型（NV）、基准库存（BS）模型和（Q，r）模型），此外再加上作为 EOQ 模型的推广所提及的经济生产批量模型（EPL）。（注意到 2.7 中的一些条目包含有斜杠，这意味着这些特定的建模决策受其他模型假设的干扰因此而不可应

[11] 注意只有当式（2.49）或式（2.53）中的临界比大于 1/2 时这一点才为真。如果比值小于 1/2，那么 z 将为负，最优再订货点将会随提前期需求标准差的增大而减小。但是这种情况只有当成本最优化为产品设定了一个相当低的满足率时才发生。所以，z 为正数的情况在实际中是非常常见。

用。）运作管理文献中有各种类型的模型，它们代表了这些维度的各种合理组合，当然也包括这些特性以外的模型（例如产品间的互相代替、备件库存与维护人员利用率之间的关联以及易腐品存货）。这本书中，我们将在第十七章继续研究库存管理中的重要问题，并且将把本章讲到的某些模型扩展到多产品、多级库存系统的重要现实环境中去。相对于我们所提供的这两章内容，如果读者想要了解更多更复杂的内容，那么可以参考 Graves、Rinnooy Kan 和 Zipkin（1993）；Hadley 和 Whitin（1963）；Johnson 和 Montgomery（1974）；McClain 和 Thomas（1985）；Nahmias（1993）；Peterson 和 Silver（1985）；Sherbrooke（1992）；和 Zipkin（1998）。

尽管这其中一些模型所需要数据可能很难或者根本无法获得，但是他们的确揭示了一些基本原理：

1. *换模（补货频率）和库存之间存在权衡*。补货越频繁，周期库存就越少。

2. *客户服务水平和库存之间存在权衡*。在随机需求的情况下，客户服务水平（即订单满足率）越高，需要的安全库存水平也就越多。

3. *变动性和库存之间存在权衡*。在给定的补货频率下，如果客户服务水平保持（在一个足够高的水平）不变，那么变动性（即需求或补货提前期的标准差）越大，必须持有的库存也越多。

尽管一些 JIT 的成果声称否认这些权衡的存在，但它们的确是制造业实实在在的事实。一些常常可以听到的警告如“库存是魔鬼”或“换模是有害的”对于引导管理者制定有些策略来说是毫无裨益的。（88|89）

与之相反，对于库存动态性、补货频率和顾客服务的理解可以使管理者正确评估应该采取哪些行动才可能产生最有效的影响。这样的直觉可以表述为这样的问题：哪项生产准备最具有破坏性？库存持有多少才算多？对客户服务水平进行一次提升的成本是多少？一个更可靠的供应商价值有大？等等。我们在第二篇会提出更多关于库存的思想，并会在第三篇的第十七章中回到库存管理的实际问题中来。

这里所讨论的库存模型和原理也为另一类问题提供了思考框架，即关于那些可以改变这些权衡性质的更高层次的行动，例如增加系统柔性，更好的供应商管理，质量提升。找到改变这些根本关系的方法是在管理中需要优先考虑的关键问题，我们将在第二、三篇更深入的对其进行探究。

附录 2A 基础概率知识
