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

第三篇 实践的原则

In matters of style, swim with the current;
In maters of principle, stand like a rock.
——托马斯·杰斐逊

第十三章 拉式计划体系

We think in generalities, we live in detail.
——AN 怀特海

13.1 引言

回忆起我们在本书的开始指出运营教育的三大基本要素是

1. 基础知识（Basics）
2. 直觉（Intuition）
3. 综合（Synthesis）

第一篇和第二篇中，我们几乎全部在讨论前两项。例如，第一篇中引进的工具和术语（如，EOQ、（Q，r）、BOM、MPS）与第二篇中变动性的量度（如，变异系数）和基本的排队概念都是对于制造经理极为重要的基础知识。传统库存模型提供的洞察力，MRP，在第一篇中观察的 JIT，产出、WIP 与周期时间的工厂物理学关系以及在第二篇中开发的变动性原理都是制定良好运营决策所需要的强有力直觉的重要组成部分。

然而，除了运营与第十一章的行为科学的比较，以及第十二章的质量的广泛方面有一些综合之外，我们几乎没有涉及过第三项，**综合（synthesis）**。现在我们要通过建立一个将第一篇和第二篇中发展的原则应用到实际制造问题的框架来填补这个重要的空白。

我们的方法建立在两个前提之上：

1. 组织不同层级处的问题需要不同水平的细节、建模建设和计划频率。
2. 不同层级之间的计划和分析工具必须一致。

第一个前提促使我们针对具体的问题使用不同的工具。不幸的是，在体系内使用不同的工具和程序容易与第二个前提冲突。由于这种潜在的不一致性，常常可以看到工业中的某些计划工具被扩展应用到了它们不适合的场所。例如，我们曾工作过的一家工厂使用的排配工具计算各台机器处细节的、一分钟一分钟的（*minute-by-minute*）产量来产生两年的集结生产计划。这个工具对于短期计划（如，一条或一周）可能还是合适的，但用于长期计划就太麻烦了（输入数据和调试就用了一周时间！）。此外，向前几周之后它就非常不精确了，以致

千辛万苦得到的排配实际上根本没有在车间内实施。（408|409）

为了开发既适合于具体应用又在应用中彼此一致的方法，我们介绍以下的计划体系开发步骤：

1. **适当地将划分系统（Divide the overall system appropriately）。**针对流程的不同部分、不同的产品类别、不同的计划展望期、不同的转产（shift）等等，可以使用不同的计划方法。关键是寻找一套区划（divisions），使得每一部分可以管理，而各个部分又能够整合。

2. **确定各区划之间的联系（Identify links between the divisions）。**例如，如果共享同一加工中心的两种产品分别制定生产计划，它们应当通过共享的加工中心的产能联系起来。如果使用不同的工具计划不同展望期的生产需求，就应当确保各个计划与其对产能、产品组合、人力等等的假定相调和。

3. **使用反馈来加强协调（Use feedback to enforce consistency）。**所有的分析、计划和控制工具都使用估计的参数（如，产能、机器速率、产出、失效与修复速率、需求速率以及其他许多的）。当系统运行时，我们应当持续更新对这些数值的知识。应当明确地使用已更新的知识来促使系统使用实时的、一致的信息，而不是以非正式、不匹配的方式来估计各种工具的输入。

在本章余下的部分，我们来预览一个与这些原则以及先前给出的工厂物理学原则一致的计划体系。我们并不认为它是与这些原则一致的唯一体系。相反地，我们将它看作是途径之一，并致力于从充分宽广的视野提出各个层级的议题从而为各种具体制造环境的定制留下余地。第三篇中的余下章节将会更加详细地讨论这个体系的主要组块。

13.2 拆分（Disaggregation）

建立计划结构的第一步是将各种决策问题分解成可管理的子问题。如我们将要讨论的，这一点可以通过建立规范计划层级来明确地实现。它也可以通过使用不同的模型和假定逐个地解决各种决策问题来模糊地实现。不管远见的水平如何，某种形式的拆分都是必须的，因为现实世界所有的生产系统都太复杂而不能用单一模型来解决。（409|410）

13.2.1 生产计划的时间尺度

制造系统的典型拆分沿用的重要维度之一就是时间。其首要原因是，制造决策因其各自结果的持续时间而大为不同。例如，新厂的建设可能在几年或几十年内对企业的位置有影响，而在某一工站处选用某一部件来加工的影响在几小时或几分钟内就消失了。这使得在决策制定过程中使用不同的**计划展望期（planning horizon）**尤为重要。因为建设新厂的决策将在数年的时间内影响运营，所以我们必须向前数年预测这些影响从而得出合理的决策。故而，这个问题的计划展望期应该很长。显然地，我们不必为评估在工站处加工什么而向前想得如此长远，故而它的计划展望期较短。

计划展望期的合适长度也因产业和组织的层级而不同。某些产业，如石油和长途电话，往往使用长达数十年的计划展望期，因为它们业务决策的后果将会持续那么长的时间。对于给定的公司，较长时间的展望期一般用在负责长期业务计划的企业办公室里，而不是用在制定每日任务的工厂里。

在本书中，我们主要关注与经营工厂相关的决策，并在这种情境下将计划展望期划分为**长期（long）**、**中期（intermediate）**和**短期（short）**。在工厂层级，长期的计划展望期可能

是从一年到五年，以两年为典型。中期的计划展望期可能是从一周到一年，一以月为典型。短期的计划展望期可能是从一小时到一周，以一天为典型。（410|411）

表 13.1 列举了制定于长期、中期与短期的计划展望期的制造决策。注意到在一般情况下，长期决策通过考虑生产什么、如何生产、如何融资、如何销售、在哪生产、从哪获取物料以及运营这个系统的一般原则等等问题指出**战略（strategy）**。中期决策通过决定加工什么、谁来加工、设备维护采取什么措施、什么产品由销售推动等等问题指出**战术（tactics）**。这些战术决策必须在战略性的长期决策设立的物理或逻辑约束内制定。最后，短期决策通过移动物料与工人、调整制程和设备以及采取为使系统持续朝向目标运行而需要的任何措施指出**控制（control）**。长期战略和中期战术决策建立了控制决策制定时必须遵守的约束。

表 13.1 战略、战术和控制决策

时间范围	长度	典型决策
长期（战略）	一年到几十年	财务决策
		营销策略
		产品设计
		加工技术决策
		资本决策
		设施选址
		供应商合约
		个人发展项目
		工厂控制政策
		品质保证方针
中期（战术）	一周到一年	作业排程
		人员分配
		预防性检修
		促销
		采购决策
短期（控制）	一小时到一周	物料流动控制
		操作员指派
		机器换模决策
		流程控制
		品质一致性决策
		事故设备修复

不同的计划展望期意味着不同的**更新频率（regeneration frequency）**。基于未来数年的信息的长期决策不需要经常重新审议，因为对如此之久的未来将发生的事的估计不会很快改变。例如，工厂重新评价它该生产什么产品是件好事，却不是每周都该做的事。一般地，长期问题一季度至一年评议一次，非常长期的议题（如，我们应当位于何种业务领域？）则更不频繁地提起。中期问题通常一周至一月评议一次。短期问题实时或一天评议一次。当然了，这些仅仅是典型的数值，它们在不同的企业和决策问题中变化很大。

除了更新频率不同之外，有着不同计划展望期的问题在细节的水平（*level of detail*）上也不同。一般地，计划展望期越短，建模和数据收集所需的细节就越多。例如，如果制定关于即将建立的工厂的规模的长期战略性产能决策，我们不需要对部件要经过的工艺路线知道

太多。对各部件在各制程处所需的时间有个粗略估计就足够用来估计产能需求了。最后，在短期的控制层级，我们需要清楚地知道部件的工艺路线，包括某一部件是否需要重工和特殊的关注，从而引导部件通过系统。

与这种战略/战术/控制的区分很相似的例子是绘图法（mapmaking）。长期问题就像长途旅行。我们需要覆盖很远距离却不很详细的地图。显示了主要的公路，可能就足以满足需求。类似地，长期决策问题需要覆盖很长的时间（即，长的计划展望期）却不很详细的工具。与之相反，短期问题就像短途旅行。我们需要没有显示太远距离却详细展示其内容的地图。显示城市的街道甚至是单个建筑的地图，可能才合适。类似地，对于短期决策，我们需要没有覆盖很长时间（即，短的计划展望期）却对其内容给出许多细节的工具。

13.2.2 拆分的其他维度

除了时间，生产计划与控制问题一般还有其他几种拆分的维度。由于现代工厂很大很复杂，常常不可能在做出具体决策时将其视为一个整体。以下的三种维度可以用于将工厂分解成较易于分析和管理的组块：（411|412）

1. **制程（Process）。**传统上，许多工厂按物理制造过程组织。铸造、研磨、钻孔和热处理等等业在位于不同地点、接受不同管理的独立部门执行。在 JIT 运动的尾流中这种制程式组织没有以前流行了；但是通过它的流动导向（flow-oriented）的单元布置，制程的区分仍然存在。例如，钢铁厂内铸造与轧制在操作上区别很大，并且有时物理布置离得很远。半导体制造业也有类似的例子。（Likewise,mass lamination of copper and fiberglass cores in large process is distinct- physically, operationally, and logistically- from the circuitizing process in which circuitry is etched into the copper in a photo-optical/chemical flow line process.） 在此类情形下，常常为不同的制程安排各自的经理。使用不同的计划、排配和控制工具也是合理的。

2. **产品（Product）。**尽管存在着专注于单一产品的工厂（如，一家聚苯乙烯工厂），现在的大多数工厂生产多种产品。确实，通过种类与定制来竞争的压力很可能已经提高了一般工厂产品种类的平均数量。例如，有 20,000 种不同料号（即，合计制成品和子组件）的工厂并不罕见。由于在这种情况下单独考虑每种料号非常困难，许多制造工厂就将料号集结归入大类，以实现计划和管理的目的。

一种集结的形式是将相同工艺路线的部件结合起来。一般来说，工艺路线比料号少得多。例如，在一个生产几千种不同电路板的印制电路板工厂，可能只有两种**基本工艺路线（basic routings）**（如，小型板线和大型板线）。然而常常也会这样，如果考虑工艺的微小变更（如，额外的测试步骤、个别作业外包（vendoring of individual operations）、接触面镀金），实际的工艺路线数目可能比基本路线多出许多。对于计划，一般可取的是通过忽略微小变更来保持“官方”工艺路线最少。

在有显著换模时间的系统中，按工艺路线集结将不合时宜（may be going too far）。例如，电路板产线有一条特定的工艺路线可能产出 1,000 种不同的电路板，却可能只有四种不同厚度的铜片。由于传送带的速度随铜片厚度变化（保证蚀刻到位），当产线切换铜片时必须重做生产准备，造成产能损失。另外，这 1,000 种电路板需要三种冲模在板上打孔。当产线在需要不同冲模的电路板之间切换时，换模再次发生。如果所有可能的铜片厚度与所需冲模的组合出现在 1,000 种电路板生产过程中，那么工艺路线中将会有 4×3＝12 种不同的**产品族（product families）**。族（*family*）的定义确保族的内部没有明显的换模但不同的族之间可能有换模。正如我们将在第十五中章讨论的，换模将给排配带来重大的分枝结果。由于这个原因，通过族来集结产品常常可以简化计划过程，而不会使其过度单纯化。（412|413）

3. **人员（People）。**有一系列的方法可将工厂劳动力分解：作业员 vs.管理人员、工会成

员 vs.非工会成员、生产现场人员 vs.支持人员、正式工 vs.临时工、部门（如，制造、生产控制、工务、人力资源）、班次等等。在一个大型工厂里，人事组织工作计划可能与机器同样复杂。尽管详细地讨论人事组织很大程度上在本书范围之外——我们已在第十一章接触其中一些议题——我们感到指出这些组织之间的逻辑牵连是重要的。例如，将经理人员分配到不同的流程或班次可能导致缺乏协调。依靠临时工形成流动性的劳动队伍可能削弱组织的制度记忆，并可能危及整体技术水平。严格坚持作业规范可能排除系统内多能培训和产生柔性的机会。如我们在第十一章中强调的，制造系统的效力（effectiveness）很大程度上是其劳动力的函数。尽管一直都有必要将员工分为不同的类来培训、发放福利与沟通，记住我们不再被强制去执行过去的程序也是重要的。通过采取一种对后勤和人员敏感的视角，好的经理人员将找到支持两者的有效人事政策。

13.2.3 协调

上述将决策问题分别沿时间、制程、产品和人员等维度分解的讨论并没有什么革命性内容。例如，事实上世界上每个制造过程都进行某种长期、中期与短期的决策制定。区分系统好坏的不是是否做出样的分解，而是作为它们结果的子问题如何解决，尤其是它们之间的协调程度。我们将在第三篇的剩余章节从细节检验这些问题。从现在开始，我们用一个例证来说明协调的问题。

何时生产何种部件的问题在长期、中期与短期水平上来说明。在长期，我们必须考虑大致的产量与产品系列来为能力与人力作计划。在中期，我们必须开发一个有些更加详细的生产计划，来采购原材料、整理供应商信息并理性地与顾客谈判签约。在短期，我们必须建立并且执行一个控制每个工站状况的详尽工作排配。所有三个问题的本质是一样的；不同的仅仅是时限。因此，显然三个不同层次做出的决策，至少在期望上，应当是相互调和的。但人们可以想见，说比做难。

给定各种时限内（通常是月度或季度）每种部件的产量时，当产生一个长期生产计划，我们不能足够详细地考虑生产流程来确定所需机器换模的准确次数。然而，当发展出一个中期生产排配，我们必须计算所需换模次数，否则不能确定排配对于产能是否可行。因此，为了使长期计划与中期计划协调，我们必须确保长期计划工具减去每个工站反应预期平均次数换模的能力损失。在时间的流逝中达到这项保证，我们应当追踪实际换模次数，并据此调整长期计划。（413|414）

中期与短期计划之间也需要类似的连接。当产生一个中期生产排配，我们不能期望物料流动中的变异不会出现在实际生产过程。机器可能出故障，操作员可能请病假，流程或质量问题出现——没有一个能预知。然而，在短期，当我们逐分钟地做工作计划，必须考虑什么机器出故障，什么工人离岗以及许多其他影响工厂当前状态的因素。结果将是实际生产活动永远不会与计划完全匹配。因此，为了使短期生产活动能产出一致性的结果，至少在平均上，依据计划的需求，中期计划工具必须包含某种形式的缓冲产能或缓冲提前期来适应随机变动性。缓冲能力可能以我们在第四章 JIT 中讨论过的“两班制（two-shifting）”的形式提供。缓冲提前期是我们简单加到顾客允许的应对意外延迟上的时间。

下面我们在具体问题的情景下讨论计划水平之间的其他连接。但是，读者一定遇到过不同于本书论述的计划工具与程序，而我们也提出建立作为一般原则的连接的问题。关键是各个水平能够并且应当用不同的工具和假设来说明，而不是通过像前述的简单机理来连接。

13.3 预测

事实上所有生产计划系统的起点都是预测。这是因为制造计划决策的后果几乎总是依赖于未来事件的。现在看起来很好的决策也许稍后就会变得很糟糕。但由于没有人拥有能预测未来的水晶球，我们最好能做的就是利用当前可得的所有信息来选出我们预测能在将来取得成功的政策。

显然地，并不只有制造决策依赖于将来。政府政策的成功或失败在很大程度上受利率、经济增长、通货膨胀和实业等未来参数的影响。保险公司的盈利性取决于将来的债务，而它们又是自然灾害等不可预测事件的函数。石油公司的现金流受钻井投资的成功前景的支配。在与这些当前决策的效力取决于将来不确定结果类似的例子中，决策制定者通常依赖于某种类型的**预测（forecasting）**来获得对未来的期望从而评估备择政策。

有许多可以预测未来的方法，因此预测是一个庞大而繁杂的领域。这些方法的一种基本分类是

1. 定性预测（Qualitative forecasting）
2. 定量预测（Quantitative forecasting）

定性预测方法（Qualitative forecasting methods）力图利用人们的经验而非精确的数学模型，来描述可能的将来情形。一种从专家意见得出预测结果的结构化方法是**德尔菲法（Delphi）。**在 Delphi 法中，询问专家一些关于未来的问题，如一项新技术可能的引入时间等。通常以书面形式进行，也可以采用口头形式。收集的回答被排成表格并返回给座谈的专家，再由他们重新考虑和回答原始的以及可能还有一些新提出的问题。这个过程被多次重复，直到达成共识或形成稳定的意见不再改变。Delphi 法和与之类似的技术对那些未来以复杂的方式依赖于过去的长期预测很有用。以预测高度不确定性突破为核心内容的技术预测，就常常使用这种方法。Martino（1983）总结了这种情境下一系列的定性预测方法。（4134|415）

定量预测方法（Quantitative forecasting methods）基于这样的假设，未来能通过某种数学模型从而用过去的数值量度来预测。有两种基本的定量预测模型：

1. **因果模型（Causal models）**将未来参数（如，产品需求）视为其他参数（如，利率、GDP 增长率、新建房屋（housing start））的函数。
2. **时间序列模型（Time series models）**将未来参数（如，产品需求）视为该参数（如，历史需求）过去数值的函数。

我们无意提供关于预测的全面的总结，所以注意力将集中于与运营管理（OM）有最大关联的那些技术。特别地，由于运营决策基本上关注于计划展望期短于两年的问题，长期的定性预测技术并未广泛应用到运营管理中。故而，我们将聚焦于定量方法。更进一步地，由于时间序列模型简单易用并可以直接应用（在非预测情境）到生产追踪模块，我们将把大部分的注意力放在它上面。

在研究具体的方法之前，我们先介绍以下三条著名的预测定律：

预测第一定律：预测结果总是错误的！（*Forecasts are always wrong !*）

预测第三定律：细节预测比集结预测糟糕！（*Detailed forecasts are worse than aggregate forecasts !*）

预测第三定律：越是远期的预测，可靠性越差！（*The further into the future， the less reliable the forecast will be !*）

不管专家的资历有多深或者是模型有多精巧，对未来的完美预测都是不可能的；这就是第一定律。更进一步地，由变动性汇聚（variability pooling）的概念可知，集结的预测（如，对于产品族）将比细节的预测（如，对于某一种产品）表现较少的变动性；这就是第二定律。最后，向前越久，质变（如，竞争对手引入重要的新产品）的可能性越大，使得我们使用的任何预测方法都完全失效；这就是第三定律。

我们并不是要用这些定律来一概贬损预测的想法。与之相反，计划层级的所有观点都建立在预测的前提之上。如果对未来的需求没有某种预测，设置多大的产能、维持多少劳动力或者是保有多少存货这些问题都没有办法得到很好的解决。但是由于预测最好也不过是近似的，我们应当致力于使这些决策对于预测结果的错误尽量稳健。例如，使用能促进新产品适应性、产量变更以及产品族转换，有时被称为**敏捷制造**（agile manufacturing），的设备和工厂布置能极大地降低预测错误的后果。类似地，工人的交叉培训以及适应性的劳动力排配政策可以充分地提升柔性。最后，如我们在第二篇中提到的，压缩制造周期时间可以降低对预测的依赖性。

13.3.1 因果预测

在因果预测中，我们力图用其他可观测或者至少是较易预测的参数来解释一个不确定的未来参数的行为。例如，如果要评估在某一地点开设一家新的快餐店的前景，我们需要对客户需求的预测。需求的可能预测因素包括离这个地点一定距离范围内的人口和竞争者快餐店的数量。通过收集需求、人口和现存可比较餐馆的竞争状况的数据，我们可以在模型中使用统计学来估计常量的值。（415|416）

最常使用的是简单线性回归模型，其形式是

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m \qquad (13.1)$$

其中 Y 表示要预测是参数（需求），变量 X_i 是用于预测的参数（人口和竞争状况）。b_i 是要从数据统计性地估计得出的常量。

这种使方程拟合于数据的技术称为**回归分析（regression analysis）**；许多计算机软件包，包括所有主要的电子表格项目，都可用于执行必要的计算。下面的例子简要地展示了如何将回归分析作为工具应用到因果预测中。

例子：Mr. Forest's Cookies

一家新兴的饼干特许商店正在评估未来扩张的选址。高级管理层推测说，商店的成功受其周边五英里范围内居民人数的强烈影响。分析人员收集了 12 家现存特许经营商店的周边人口数据和年度销售数据，汇总在表 13.2 中。

表 13.2 *Mr. Forest's Cookies* 特许商店的数据

商店	人口数（000）	销售额（$000）
1	50	200
2	25	50
3	14	210
4	76	240
5	88	400
6	35	200
7	85	410
8	110	500
9	95	610
10	21	120
11	30	190
12	44	180

为了开发从一家新的特许商店五英里半径范围内人口数量预测其销售额的模型，分析人员使用了回归分析，而它正是一种寻找贯穿数据的“最佳拟合”直线的工具。他们选择了 Excel 的**回归（Regression）**功能，得到如图 13.1 所示的输出。以黑体标示的三个关键数值是：

1. **截距（Intercept coefficient）**，即（13.1）式中 b_0 的估计值，或是本例中的 50.30（圆整至两位小数）。这个系数表示数据的拟合直线在 Y 轴的截距。

2. X_1 **系数（** X_1 **coefficient）**，即（13.1）式中 b_1 的估计值，或是本例中的 4.17。这个系数表示数据的拟合直线的斜率。在图 13.1 中显示为“人口（000）”项。

3. **R 平方（R square）**，表示有回归直线解释的数据变异的比例。如果数据完全在拟合直线上，R 平方将等于一。R 平方越小，数据与回归直线的拟合优度越差。本例中，R 平方为 0.77441441，它意味着拟合良好但说不上是完美。Excel 也生成了如图 13.2 所示的数据和回归直线的绘图，据此我们可以直观地检验模型在多大程度上与数据拟合。（416|417）

输出摘要

回归统计	
倍数 R	0.880008188
R 平方	0.774414411
调整后的R 平方	0.751855852
标准误差	77.79635826
观测值	12

方差分析

	df	*SS*	*MS*	*F*	*F* 检验量
回归分析	1	207768.9331	207758.9331	34.32907286	0.000159631
残差	10	60522.73358	6052.273358		
总计	11	268291.6667			

	系数	标准误差	t 值	P 值	上限 95%	下限 95%
截距	**50.30456039**	45.79857723	1.098386968	0.297777155	-51.74104657	152.3501673
X 自变量 1	**4.169903827**	0.711696781	5.859101711	0.000159631	2.584144304	5.755663349

图 13.1 *Excel* 回归分析的输出

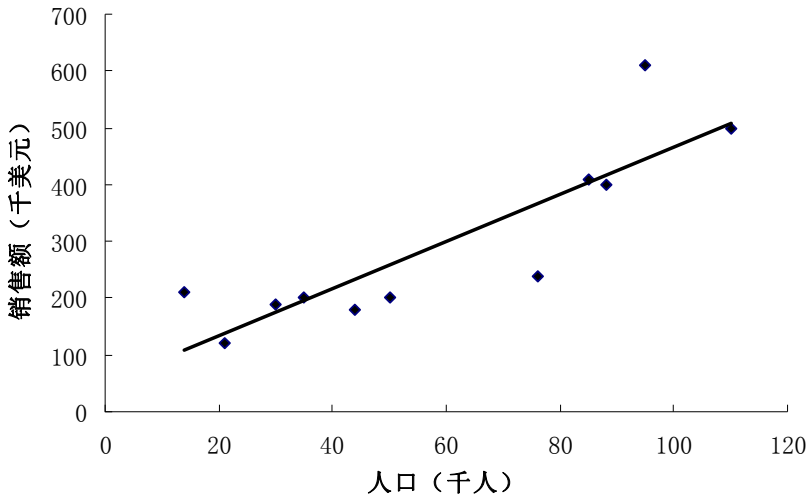

图 13.2 回归直线对 *Mr. Forest's* 的数据的拟合程度

从图 13.1、13.2 的结果来看，假定新的特许商店的服务人口是 15,000 到 110,000 之间时，模型看起来适合用于做出大略的预测。由于原始数据集并未包括这个范围之外的人口，我们没有依据对少于 15,000 或多于 110,000 的人口做预测。

如果为 Mr. Forest 服务的分析人员想要开发出更加精确的模型，他们还应当考虑加入其他的用于预测的变量，如五英里半径范围内人口的平均收入，在提议的选址特定距离之内的饼干商店的数量，在提议的选址步行距离之内的其他零售设施的数量等等。（13.1）式的一般模型，称为**多元回归模型（multiple regression model）**（与只包含一个用于预测的变量的**简单回归模型（simple regression model）**相对），以及执行此类计算的计算机软件包都允许多元的预测因子。

如 Excel 这样的软件包使回归方法的技巧变得简单。但是对于结果的充分理解就需要统计学的知识了。由于统计学和回归分析在商业上的广泛应用——用于市场分析、产品设计、人员评估、预测、质量控制和过程控制——它们成为现代经理技能集合的本质基础。任何好的商业统计学课本都会提供这些重要主题的必要背景。

尽管因果模型通常很有用，它本身并不总是能使我们对未来做出预测。例如，如果下个月对屋面材料的需求，如制造商所见，取决于上个月的房屋开工量（由于房屋开工与建筑商向供应商下达的采购订单完成之间的时间延迟），则此时模型需要的只是可见的输入并可以直接做出预测。于此相反，如果下个月对空调的需求取决于下个月的日平均气温，则在预测需求之前必须预测下个月的气温。（即使具备高精度的长期天气预报，因果模型也不见的有太大帮助。）

13.3.2 时间序列预测

在预测过去结果是未来行为良好指示器的参数，却没有强的因素-效果关系可用于构建因果模型时，常常使用**时间序列模型（time series model）**。对产品的需求的常常属于此类，故而需求预测是这种技术最常见的应用之一。其原因在于，需求是诸如客户诉求（customer appeal）、营销效力（marketing effectiveness）与竞争等因素的函数。尽管这些因素难以明确地构模，它们确实趋向于持续作用，所以过去的需求常常是未来需求的良好指示器。时间序列模型要做的，就是力图获取过去的趋势并将它们用于未来的推断。

不同的时间序列模型有许多，但基本的程序都是一致的。我们把时间分成若干时期（periods）（如，月份），标记为 i = 1, 2, …, t，其中时期 t 是用于预测的最近一期的数据。我们用 $A(i)$ 标记实际观测值，用 $f(t+\tau)$ 标记对时期 $t+\tau$，τ = 1, 2, …的预测值。如图 13.3 所示，时间序列模型将过去的观测值 $A(i)$，i = 1, …, t（如，$A(i)$ 可以代表份 i 的需求，其中 t 代表可得数据的最后一期）作为输入，并生成对未来数值的预测 $f(t+\tau)$，τ = 1, 2, …（如，$f(t+\tau)$ 代表对月份 $t+\tau$ 的需求预测，它向前滚动了 τ 月）。这样，包括在此处讨论的一些模型，就计算表示对所考虑过程当前位置的估计的**平滑估计值（smoothed estimate）** $F(t)$，以及表示对过程当前趋势的估计的**平滑趋势（smoothed trend）** $T(t)$。（418|419）

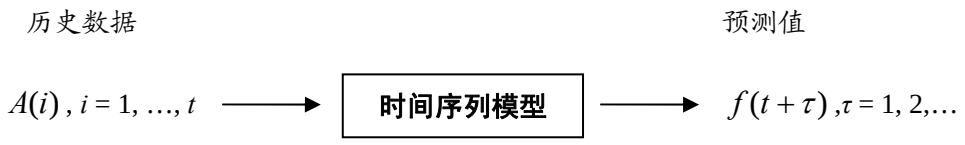

图 13.3 时间序列模型的基本结构

许多不同的模型都可以完成预测的这个基本功能；最合适的取决于具体应用场景。这里我们提供四种最简单、最通用的方法。**移动平均（moving-average）**模型计算过去 m 期观测值（m 的值由使用者决定）的均值作为下期（与其后）的预测值。**指数平滑（exponential**

smoothing）计算最近的观测值和先前的平滑估计值的加权平均数作为当前的平滑估计值。与移动平均模型类似，简单的指数平滑假定数据无趋势（即，上升或下降）并因此使用平滑估计值作为未来所有时期的预测值。**有线性趋势的指数平滑（exponential smoothing with a linear trend**）用类似于指数平滑的方式计算平滑估计值，但也计算数据中的平滑趋势或是斜率。最后，**Winter 法（Winter's method**）在有线性趋势的指数平滑模型中加入季节乘数，以表示需求显示出季节特征的情形。

移动平均。将实际观测值转变成预测值最简单的办法是简单地取其平均值。如果这样做，我们就含蓄地假定数据无趋势，即对于所有的 t 都有 $T(t) = 0$。然后计算简单的平均值作为平滑估计值，并将这个平均值用于未来各期的预测，故而（419|420）

$$F(t) = \frac{\sum_{i=1}^{t} A(i)}{t}$$

$$f(t+\tau) = F(t) \qquad \tau = 1, 2, \ldots$$

这种方法的一个潜在问题是，它给所有的过去数据赋予相同的权重，而不管它们离现在的远近。但是，三年前的需求数据可能对未来预测已经没有代表性了。为了获取使近期数据比远期数据与未来结果更紧密关联的趋势，事实上所有的时间序列模型都包含了对陈旧数据折扣的机制。最简单的实现程序是丢掉过去某一时点之外的数据。这样做的时间序列模型称为**移动平均**模型；除了仅仅是最近 m 期（m 是由使用者取值的参数）的数据用于均值计算，它与简单平均的原理相同。再一次地，趋势被假定为零，故 $T(t) = 0$，并且所有的未来预测都被假定为当前的平滑估计值：

$$F(t) = \frac{\sum_{i=t-m+1}^{t} A(i)}{m} \qquad (13.3)$$

$$f(t+\tau) = F(t) \qquad \tau = 1, 2, \ldots \qquad (13.4)$$

注意到 m 的取值将使移动平均法的表现不同。选择适合于特定情形的值的一种方法是，尝试多个数值并观察它们对已知数据预测得怎样。例如，假设我们有某种产品过去 20 个月的需求数据，如表 13.3 所示。在任何时期，我们都假想拥有的数据只截止到该期，并使用移动平均来生成预测。如果设定 $m = 3$，则在时期 $t = 3$ 我们能计算前三期的平均值作为平滑估计值，或是

$$F(3) = \frac{10+12+12}{3} = 11.33$$

在 $t = 3$ 时刻，我们对时期 4（及以后的，因为无趋势）的预测是 $f(4) = F(3) = 11.33$。然而，一旦我们真正转移到时期 4 并查看实际需求，估计值变成第二、三、四期数据的平均值，或是

$$F(4) = \frac{12+12+11}{3} = 11.67$$

现在我们对时期 5（及以后的）的预测是 $f(5) = F(4) = 11.67$。以这种方式继续，可以计算

得出我们对 $t = 4, \ldots, 20$ 本应有的预测，如表 13.3 所示。不能对时期 1、2、3 做预测，是因为在计算三期的移动平均之前首先需要三期的数据。

如果在移动平均中将时期换成 $m = 5$，也可以计算时期 6, …, 20 的平滑估计值从而得到预测值，如表 13.3 所示。（420|421）

$m = 3$ 和 $m = 5$，哪一个较好？从表 13.3 来看，难以回答。可是，如果绘制 $A(t)$、$f(t)$ 的图形，就可以看出那种模型预测更接近于实际观测值。如图 13.4 所示，两种模型结果都趋于低估需求，而 $m = 5$ 的表现更差一些。低估的原因是移动平均模型假定数据没有上升或下降的趋势。但是我们可以从绘图中看到这些数据明显存在着上升趋势。这样，对过去需求的移动平均趋于小于未来需求。由于 $m = 5$ 的模型更强地与过去的需求联系（它包含较多的，也因而较旧的数据），它在更大程度上遭受这个趋势的影响。

表 13.3 $m = 3$ 和 $m = 5$ 时的移动平均

月份 t	需求 $A(t)$	预测值 $f(t)$	
		$m = 3$	$m = 5$
1	10	-	-
2	12	-	-
3	12	-	-
4	11	11.33	-
5	15	11.67	-
6	14	12.67	12.0
7	18	13.33	12.8
8	22	15.67	14.0
9	18	18.00	16.0
10	28	19.33	17.4
11	33	22.67	20.0
12	31	26.33	23.8
13	31	30.67	26.4
14	37	31.67	28.2
15	40	33.00	32.0
16	33	36.00	34.4
17	50	36.67	34.4
18	45	41.00	38.2
19	55	42.67	41.0
20	60	50.00	44.6

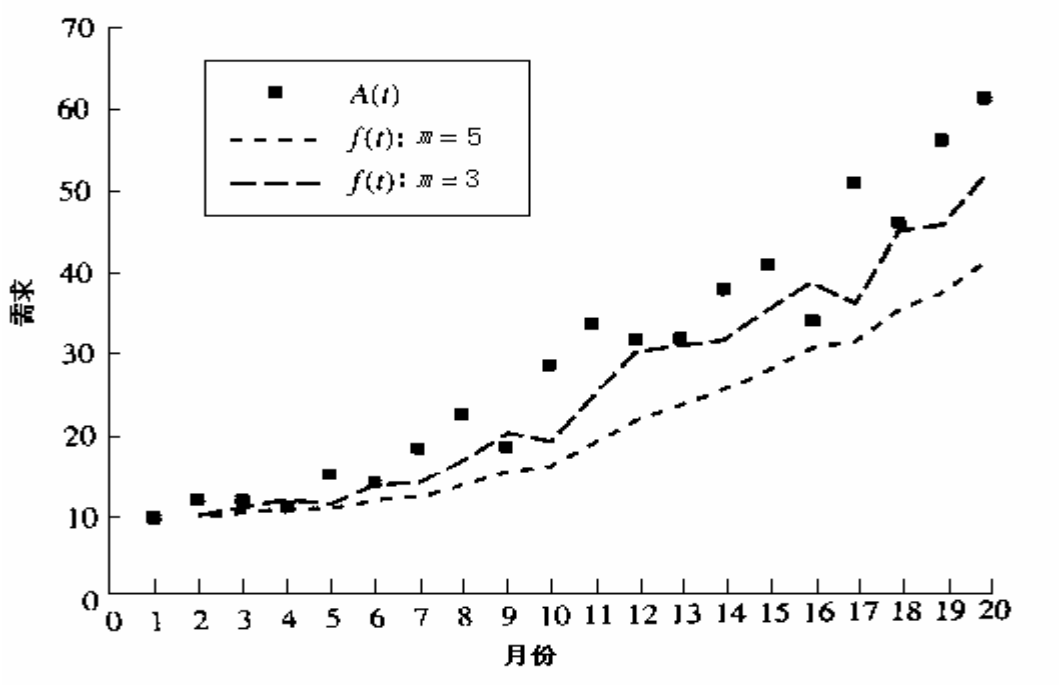

图 13.4 m = 3 和 m = 5 时的移动平均

这个例子阐明了关于移动平均模型的以下的一般结论：

1. m 值越大，模型的稳定性越好，但对将预测过程的变化的响应性越差。
2. 模型往往会低估有上升趋势的参数，高估有下降趋势的参数。

我们可以在移动平均模型的情境下解决追踪趋势的问题。对于那些与回归分析类似的情形，处理方式是先通过线性回归估计最后 m 个数据的斜率，再计算平滑估计值与这个线性趋势推断值之和作为预测值。然而，还有一种向不同的时间序列模型引入线性趋势的更简单的方法。接下来，我们将在介绍以下另一个无趋势模型后继续讨论这种方法。

指数平滑。注意到移动平均法赋予最近 m 个观测值相同的权重，而那些远于 m 期的则没有被赋予权重。另一种折扣陈旧数据的方法是，对当前的平滑估计值和最近一期的数据作平均。这样做的结果将是，数据越旧，在决定预测中占的权重就越小。我们称之为**指数平滑**（**exponential smoothing**），它的原理如下。首先，我们现在假定趋势总是零，故 $T(t)=0$。然后，计算 t 时刻的平滑估计值和预测值

$$F(t)=\alpha A(t)+(1-\alpha)F(t-1) \qquad (13.5)$$

$$f(t+\tau)=F(t) \quad \tau=1,2,\ldots \qquad (13.6)$$

其中 α 是由使用者选取的介于 0 和 1 的平滑常数。它的最优值取决于具体的数据。（421|422）

表 13.4 以曾用于移动平均的数据说明了指数平滑法。如果开始时没有历史数据 $F(0)$，则不能对时期 1 做出预测。初始化模型的方法很多（如，通过对过去某段时间的观测值取平均值），但 $F(0)$ 的选择将随着时间的延伸而淡化。因此，我们选用最简单的可行初始化方

法并设定 $F(1) = A(1) = 10$ 从而启动运算过程。在 $t = 1$ 时刻，对于时期 2（及以后）的预测是 $f(2) = F(1) = 10$。接下来到达时期 2 并看到 $A(2) = 12$，我们采用如下的方式更新平滑估计值：

$$F(t) = \alpha A(t) + (1-\alpha)F(t-1) = (0.2)(12) + (1\text{-}0.2)(10) = 10.40$$

对时期 3 及以后的预测现在是 $f(3) = F(2) = 10.40$。继续以这种方式，可以得到余下的 $f(t)$ 值，如表 13.4 所示

在表 13.4 中我们注意到，用 $\alpha = 0.6$ 替代 $\alpha = 0.2$ 时预测值对每个新的数据点更敏感。例如，在时期 2，当需求从 10 上升到 12 时，使用 $\alpha = 0.2$ 得到的预测值仅上升到 10.40，而使用 $\alpha = 0.2$ 得到的预测值上升到 11.20。当模型在追踪数据的实际趋势时，这种强化的敏感性是好的；而当它是对反常观测值的过度反应时，这种敏感性就是坏的。因此，类似于对移动平均法的观察，我们得出一次指数平滑的要点如下：

1. α 值越小，模型的稳定性越好，但对将预测过程的变化的响应性越差。
2. 模型往往会低估有上升趋势的参数，高估有下降趋势的参数。（422|423）

表 13.4 $\alpha =$ 0.2 和 $\alpha =$ 0.6 时的指数平滑

月份 *t*	需求 *A(t)*	预测值 *f(t)*	
		α= 0.2	**α = 0.6**
1	10	-	-
2	12	10.00	10.00
3	12	10.40	11.20
4	11	10.72	11.68
5	15	10.78	11.27
6	14	11.62	13.51
7	18	12.10	13.80
8	22	13.28	16.32
9	18	15.02	19.73
10	28	15.62	18.69
11	33	18.09	24.28
12	31	21.08	29.51
13	31	23.06	30.40
14	37	24.65	30.76
15	40	27.12	34.50
16	33	29.69	37.80
17	50	30.36	34.92
18	45	24.28	43.97
19	55	36.43	44.59
20	60	40.14	50.83

为指数平滑选择合适的 α 值，就像为移动平均法选择合适的 m 值一样，需要一些反复尝试。一般而言，我们最好能做的就是尝试多个不同的 α 值并看哪一个能产生最符合历史数据的预测结果。例如，图 13.5 绘出了实际值 $A(t)$，以及使用 α = 0.2 和 α = 0.6 得到的指数平滑结果 $f(t)$。图形清楚地显示出，使用 α = 0.6 的结果比 α = 0.2 的结果更接近实际数据点。较大的 α 值引起的强化的敏感性使模型能够追踪数据的明显上升趋势。然而，由于一次指数平滑模型并没有明确假定数据存在趋势，以上两种预测结果都趋向滞后于实际数据。

有线性趋势的指数平滑。我们现在转到一种特别设计来追踪有上升或下降趋势数据的模型。简单地说，这种模型假定数据的趋势是线性的。即，我们由当前向未来的预测将追随一条直线。当然了，每次接收一个新的观测值，我们都将更新这条直线的斜率，所以这种方法可以追踪以非线性方式变化的数据，尽管其精度没有一般地线性变化的数据那样高。

基本方法是每得到一个新的观测值，都更新平滑估计值 $F(t)$ 和平滑趋势 $T(t)$。这样，向前 τ 期的预测值，标记为 $f(t+\tau)$，就可以由平滑估计值加上 τ 与平滑趋势的乘积得到。执行这些计算的等式如下：

$$F(t) = \alpha A(t) + (1-\alpha)[F(t-1) + T(t-1)] \qquad (13.7)$$

$$T(t) = \beta[F(t) - F(t-1)] + (1-\beta)T(t-1) \qquad (13.8)$$

$$f(t+\tau) = F(t) + \tau T(t) \qquad (13.9)$$

其中 α 和 β 是由使用者选取的介于 0 和 1 的平滑常数。

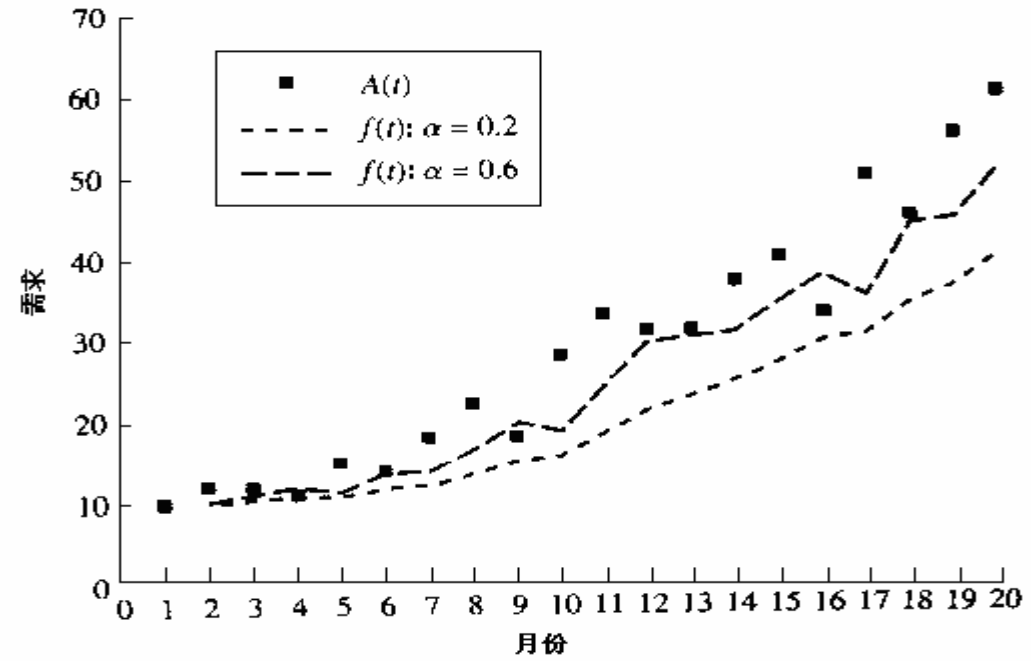

图 13.5 α = 0.2 和 α = 0.6 时的指数平滑

注意到计算 $F(t)$ 的等式与无线性趋势的指数平滑稍微有点不同。原因在于，在 $t-1$ 期对 t 期预测为 $F(t-1)+T(t-1)$（即，需要加上一期的趋势）。因此，当计算 $A(t)$ 和当前预测的加权平均数时，我们必须将 $F(t-1)+T(t-1)$ 作为当前预测。（423|424）

我们计算上一期平滑趋势 $T(t-1)$ 和最近一期对趋势的预测（the most recent estimate of the trend）的加权平均数，从而在（13.8）式中更新趋势值。最近一期对趋势的预测由计算最近两期平滑估计值的差值得到，或是 $F(t)-F(t-1)$。$F(t)-F(t-1)$ 就像是斜率。通过对这个斜率赋予 β（小于 1）的权重，就可以平滑对趋势的估计从而避免对数据突然变化的过度反应。

就像简一次指数平滑那样，我们必须初始化模型。可以使用历史数据来估计 $F(0)$ 和 $T(0)$。然而，最简单的初始化方法是设定 $F(1)=A(1)$ 和 $T(1)=0$。我们阐述使用这种初始化程序的有线性趋势的指数平滑，需求数据来自表 13.4，平滑常数为 α = 0.2 和 β = 0.2。例如，

$$F(2)=\alpha A(2)+(1-\alpha)[F1)+T(1)] = 0.2(12) + (1 - 0.2)(10 + 0) = 10.4$$

$$T(2)=\beta[F(2)-F(1)]+(1-\beta)T(1) = 0.2(10.4 - 10) + (1 - 0.2)(0) = 0.08$$

其余的计算在表 13.5 中给出。

图 13.6 绘出了表 13.5 中的预测值 $f(t)$ 和实际值 $A(t)$，以及 α = 0.3 和 β = 0.5 时的结果。注意到这些预测结果比移动平均或者是无线性趋势的指数平滑都更紧密地贴近数据。线性趋势使这种方法能相当有效地追踪数据的上升趋势。另外，它也显示使用 α = 0.3 和 β = 0.5 的平滑系数比使用 α = 0.2 和 β = 0.2 产生更好的预测。接下来，我们将在本节晚些时候讨论如何选择平滑常数。（424|425）

表 13.5 有线性趋势的指数平滑，$\alpha = 0.2$、$\beta = 0.2$

月份 *t*	需求 *A(t)*	平滑估计值 *F(t)*	平滑趋势 *T(t)*	预测值 *f(t)*
1	10	10.00	0.00	-
2	12	10.40	0.08	10.00
3	12	10.78	0.14	10.48
4	11	10.94	0.14	10.92
5	15	11.87	0.30	11.08
6	14	12.53	0.37	12.17
7	18	13.93	0.58	12.91
8	22	16.00	0.88	14.50
9	18	17.10	0.92	16.88
10	28	20.02	1.32	18.03
11	33	23.67	1.79	21.34
12	31	26.57	2.01	25.46
13	31	29.06	2.11	28.58
14	37	32.33	2.34	31.17
15	40	35.74	2.55	34.67
16	33	37.23	2.34	38.29
17	50	41.66	2.76	39.57
18	45	44.53	2.78	44.42
19	55	48.85	3.09	47.31
20	60	53.55	3.41	51.94

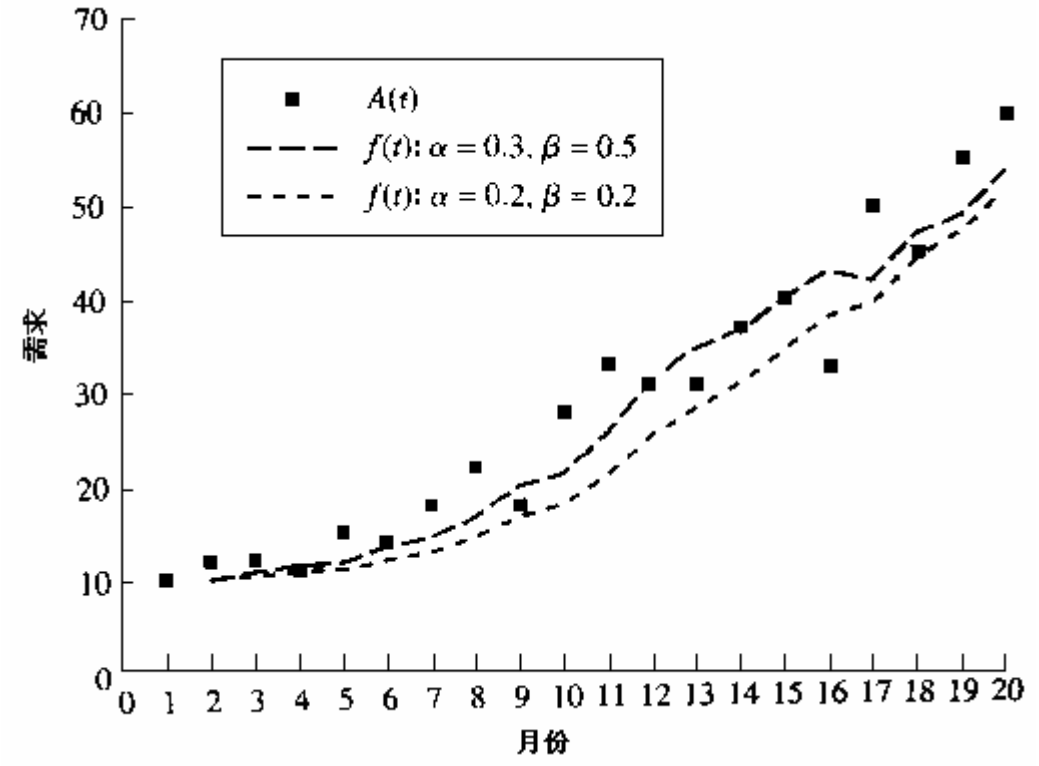

图 13.6 有线性趋势的指数平滑

应对季节性的 Winters 法。许多产品表现出季节性需求。例如，割草机、冰激凌和空调在夏季达到需求高峰，而除雪机、挡风雨条和火炉在冬季达到需求高峰。在这种情况下，前面提到的预测模型表现不佳，因为它们会将季节性增长当成持久增长以至于在萧条季节高估实际需求。类似地，它们会将萧条季节的低需求当成持久的以至于在高峰季节低估需求。

Winter（1960）提出了一种将季节性引入预测模型的自然方法。其基本思想是估计一个倍增的（multiplicative）季节因子 $c(t)$，t = 1, 2, …，来表示时期 t 的需求占该季节平均需求的比例。因此，如果该季有 N 期（如，若每季为一年而每期为一月，则 N = 12），则该季中所有 $c(t)$ 因子之和将等于 N。按季节性调整的预测等于有线性趋势的指数平滑模型的预测结果（即，$F(t)+\tau T(t)$）与合适的季节因子的乘积。执行这些计算的等式如下：

$$F(t)=\alpha\frac{A(t)}{c(t-N)}+(1-\alpha)[F(t-1)+T(t-1)] \qquad (13.10)$$

$$T(t)=\beta[F(t)-F(t-1)]+(1-\beta)T(t-1) \qquad (13.11)$$

$$c(t)=\gamma\frac{A(t)}{F(t)}+(1-\gamma)c(t-N) \qquad (13.12)$$

$$f(t+\tau)=[F(t)+\tau T(t)]c(t) \qquad (13.13)$$

其中 α、β 和 γ 是由使用者选取的介于 0 和 1 的平滑常数。注意到除了实际观测值 $A(t)$ 要除以季节因子 $c(t-N)$ 之外，（13.10）（13.11）式与有线性趋势的指数平滑模型中计算平滑估计值和平滑趋势的（13.7）、（13.8）式相同。这样就相对于均值将所有的观测值标准化了，从而以平均（非季节性）需求来表示平滑估计值和平滑趋势。（13.12）式使用指数平滑来更新季节因子 $c(t)$，其值为实际需求与平滑估计值的季节性比率 $A(t)/F(t)$ 与上季的因子 $c(t-N)$ 的加权平均数。为了使预测表现出季节性，我们用季节因子 $c(t)$ 乘以非季节性预测值 $F(t)+\tau T(t)$。（425|426）

表 13.6 应对季节性预测的 *Winters* 法

年度	月份	时期 *t*	实际需求 *A(t)*	平滑估计值 *F(t)*	平滑趋势 *T(t)*	季节因子 *c(t)*	预测值 *f(t)*
1997	Jan	1	4	-	-	0.480	
	Feb	2	2	-	-	0.240	
	Mar	3	5	-	-	0.600	
	Apr	4	8	-	-	0.960	
	May	5	11	-	-	1.320	
	Jun	6	13	-	-	1.560	
	Jul	7	18	-	-	2.160	
	Aug	8	15	-	-	1.800	
	Sep	9	9	-	-	1.080	
	Oct	10	6	-	-	0.720	
	Nov	11	5	-	-	0.600	
	Dec	12	4	8.33	0.00	0.480	
1998	Jan	13	5	8.54	0.02	0.491	4.00
	Feb	14	4	9.37	0.10	0.259	2.06
	Mar	15	7	9.69	0.12	0.612	5.68
	Apr	16	7	9.57	0.10	0.937	9.43
	May	17	15	9.83	0.12	1.341	12.76
	Jun	18	17	10.04	0.13	1.573	15.52
	Jul	19	24	10.26	0.13	2.178	21.97
	Aug	20	18	10.36	0.13	1.794	18.72
	Sep	21	12	10.55	0.14	1.086	11.33
	Oct	22	7	10.59	0.13	0.714	7.69
	Nov	23	8	10.98	0.15	0.613	6.43
	Dec	24	6	11.27	0.17	0.485	5.34

我们用表 13.6 的例子阐述 Winters 法。为了初始化运算程序，我们需要一整季的季节因子和初始的平滑估计值以及平滑趋势。最简单的方法是，使用第一季的数据计算这些初始参数，然后使用前述的等式以及后续季节的数据来更新它们。特别地，我们简单地设定平滑估计值为第一季数据的平均数

$$F(N)=\frac{\sum_{t=1}^{N}A(t)}{N} \qquad (13.14)$$

故而，在本例中，可以计算 1998 年 12 月的平滑估计值为

$$F(12)=\frac{\sum_{t=1}^{12}A(t)}{12}=\frac{4+2+\cdots+4}{12}=8.33$$

由于仅起始于一季的数据，所以没有根据来估计趋势，因此我们将在最初假定趋势为零，则 $T(N)=T(12)=0$。当加入后续的季节时，模型将很快更新这个趋势值。[1]最后，计算实际需

[1] 另外，还可以使用多季节的数据来初始化模型并从它们估计得到趋势（见 Silver、Pyke 和 Peterson 在 1998 年提出的方法）。

求与第一季平均需求的比率作为初始的季节因子：（426|427）

$$c(i)=\frac{A(i)}{\sum_{t=1}^{N}A(t)/N}=\frac{A(i)}{F(N)} \qquad (13.15)$$

例如，在本例中，一月的初始季节因子为

$$c(1)=\frac{A(1)}{F(12)}=\frac{4}{8.33}=0.480$$

一旦计算出 $F(N)$、$T(N)$ 以及 $c(1),\ldots,c(N)$ 的值，就可以启动平滑的程序。对 1998 年一月的平滑估计值为

$$F(13)=\alpha\frac{A(13)}{c(13-12)}+(1-\alpha)[F(12)+T(12)]$$

$$=0.1\left(\frac{5}{0.480}\right)+(1-0.1)(8.33+0)=8.54$$

平滑趋势为

$$T(13)=\beta[F(13)-F(12)]+(1-\beta)T(12)=0.1(8.54-8.33)+(1-0.1)(0)=0.02$$

一月份更新的季节因子为

$$c(13)=\gamma\frac{A(13)}{F(13)}+(1-\gamma)c(1)=0.1(5/8.54)+(1-0.1)(0.48)=0.491$$

计算过程依此持续，产生如表 13.6 所示的结果。我们在图 13.7 中绘出实际和预测需求。本例中，Winters 法表现良好。主要原因在于，1998 年的季节性峰形图（spike）与 1997 年的相似。也就是说，发生于某一个月份，如七月，的需求占全年总需求的比率，在各年度之间相当恒定。因此，季节因子良好地拟合于季节性行为。全年总需求在增长的实情，由模型的正数趋势所揭示，引起第二年的适度扩张的季节性峰形图。一般来说，对于季节性在各季之间变化不大的情形，Winters 法在季节性预测中表现良好。（427|428）

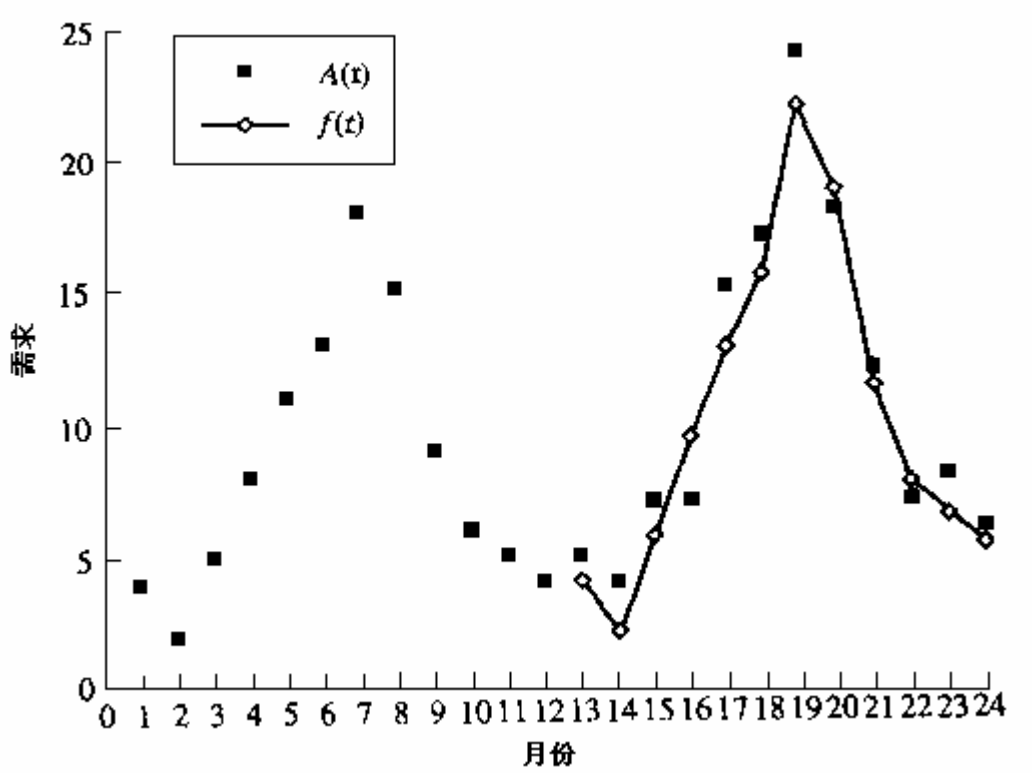

图 13.7 *Winters* 法，$\alpha = 0.2$、$\beta = 0.2$、$\gamma = 0.1$

调整预测参数。已经讨论的所有时间序列模型都包含可调整的系数（如，移动平均模型中的 m 和指数平滑模型中的 α），它们都要被“转化”成数据从而生成合适的预测模型。实际上，从图 13.6 可以看到，调整模型的平滑系数会显著地影响模型的精确性。我们现在转向如何寻找适合于给定预测情形的良好系数的问题上。

开发预测模型的第一步是绘出数据点图。它将帮助我们决定，数据是否具备可预测性，是否显现出趋势，或者是否存在季节性。一旦选定一种模型，可以绘出各种参数集下的预测值和过去实际数据的比较，从而观察模型的表现。然而，为了寻找良好的系数集，更精确地衡量模型的精度是很有帮助的。

评估预测模型最常用的三个量化指标是平均绝对偏差（*mean absolute deviation*，*MAD*）、平均平方偏差（*mean squared deviation*，*MSD*）和偏差（*bias*，*BIAS*）。每个指标都衡量预测值与实际值之间的差值，$f(t) - A(t)$，并计算一个数值。具体的公式是如下这些：

$$MAD = \frac{\sum_{t=1}^{n} |f(t) - A(t)|}{n} \qquad (13.16)$$

$$MSD = \frac{\sum_{t=1}^{n} [f(t) - A(t)]^2}{n} \qquad (13.17)$$

$$BIAS = \frac{\sum_{t=1}^{n} f(t) - A(t)}{n} \qquad (13.18)$$

MAD 和 MSD 只可能是正值，所以目标是寻找模型系数来使它们尽可能地小。BIAS 可能是正值，显示预测趋于高估实际值；或者负值，显示预测趋于低估实际值。这样，其目标

就是寻找系数来使 BIAS 接近零。然而，注意到零 BIAS 并不意味着预测是精确的，它仅仅表示过高与过低的偏差得到平衡。因此，绝不能只用 BIAS 来评价预测模型。

为了说明这些量度如何用于选取模型系数，我们回到使用表 13.5 中数据的有线性趋势的指数平滑模型。表 13.7 汇集了 α、β 的各种不同组合下的 MAD、MSD 和 BIAS。从表中可以看到，$\alpha = 0.3$、$\beta = 0.5$ 的组合在最小化 MAD 和 MSD 上做得最好，而 $\alpha = 0.6$、$\beta = 0.6$ 的组合在最小化 BIAS 上做得最好。一般来说，不大可能有一个系数集在所有三项效力指标上都取得最优值。在这种特殊的情形下，如图 13.6 所示，实际数据不仅有上升趋势，而且还趋于以非直线（即，抛物线）方式增长。这种非线性方法使有线性趋势的模型稍稍滞后于数据，引起负的 BIAS。较大的 α 和 β 值赋予新观测值较大的权重，因而使较快地追踪这种上升的摆动。这样就降低了 BIAS。然而，它也会在数据偶尔向下倾斜时产生高估，提高 MAD 和 MSD。

表 13.7 不同的 α、β 下的有线性趋势的指数平滑

α	β	MAD	MSD	BIAS	α	β	MAD	MSD	BIAS
0.1	0.1	10.23	146.94	-10.23	0.4	0.1	4.30	30.14	-3.45
0.1	0.2	8.27	95.31	-8.27	0.4	0.2	3.89	23.78	-2.34
0.1	0.3	6.83	64.91	-6.69	0.4	0.3	3.77	22.25	-1.77
0.1	0.4	5.83	47.17	-5.43	0.4	0.4	3.75	22.11	-1.46
0.1	0.5	5.16	36.88	-4.42	0.4	0.5	3.76	22.36	-1.29
0.1	0.6	4.69	30.91	-3.62	0.4	0.6	3.79	22.67	-1.18
0.2	0.1	6.48	60.55	-6.29	0.5	0.1	4.13	27.40	-2.84
0.2	0.2	5.04	37.04	-4.49	0.5	0.2	3.91	23.61	-1.94
0.2	0.3	4.26	27.56	-3.29	0.5	0.3	3.88	23.02	-1.49
0.2	0.4	3.90	23.75	-2.51	0.5	0.4	3.90	23.26	-1.25
0.2	0.5	3.73	22.32	-2.02	0.5	0.5	3.94	23.73	-1.10
0.2	0.6	3.65	21.94	-1.71	0.5	0.6	3.97	24.27	-1.00
0.3	0.1	4.98	37.81	-4.45	0.6	0.1	4.12	26.85	-2.42
0.3	0.2	4.11	26.30	-3.03	0.6	0.2	4.03	24.63	-1.66
0.3	0.3	3.82	22.74	-2.23	0.6	0.3	4.04	24.69	-1.29
0.3	0.4	3.66	21.81	-1.77	0.6	0.4	4.09	25.35	-1.08
0.3	0.5	3.65	21.78	-1.52	0.6	0.5	4.14	26.25	-0.95
0.3	0.6	3.68	22.06	-1.38	0.6	0.6	4.21	27.29	-0.84

表 13.7 显示，使用 $\alpha = 0.3$、$\beta = 0.5$ 的模型的 MSD 比我们 $\alpha = 0.2$、$\beta = 0.2$ 的初始选择的模型明显要小。这就意味着它更紧地拟合于过去的数据，如图 13.6 所示。由于使用时间序列模型的基本假设就是未来数据表现出与过去数据相似的行为，我们应当设定能与过去数据良好拟合的系数，然后将其用于预测的未来的目的。（428|429）

在这里给出表 13.7 中列举的数据是为了阐释改变平滑系数的影响。然而，实际上我们不需要使用试错法来搜索良好的平滑系数集。与之相替代，我们可以使用 Excel 内部优化工具，Solver，来执行这种搜索（见第十六章讨论 Solver 的细节）。如果我们设置 Solver 来寻找（1）零与一之间（2）最小化前例中 MSD 的 α 和 β，将得到 $\alpha = 0.284$、$\beta = 0.467$ 的解答，它使 MSD 达到 21.73。这样的结果稍微优于比我们费力搜索到的 $\alpha = 0.3$、$\beta = 0.5$，而且速度还快得多。

注意到在对选择平滑系数的讨论中，我们已经比较了向前一期的预测（即，lag-1 预测）与实际值。然而，实际中我们常常需要预测未来的未来。例如，如果使用需求预测来决定要采购多少原材料，可能要向前预测数月（如，可能需要 lag-τ 预测）。在这种情形下，我们应当使用公式来计算τ期之后的预测值$f(t+\tau)$，并且在$A(t+\tau)$发生时将二者做比较。因此，，模型参数应当以最小化$f(t+\tau)$与$A(t+\tau)$偏差为目的选取，并相应地定义 MAD、MSD 和 BIAS。

13.3 预测的艺术

用于因果预测的回归模型和四种时间序列模型，是可用于辅助预测功能的为数众多的量化工具的代表。还有许多种其他的（见 Box 与 Jenkins（1970）对更精巧的时间序列模型的综览）。显然，预测只是量化模型可以显出巨大价值的一个领域。（429|430）

然而，预测并不是选择一个模型加上胡乱地修补其参数来使其尽可能有效。没有哪种模型能将所有与预期未来相关的因素全部纳入其中。因此，在任何预测环境，都会出现预测者根据定性的信息而凌驾定量的模型的情形。例如，如果有理由期望即将发生的需求跳跃（如，由于竞争者的工厂要关闭了），预测者需要据此信息在量化模型的基础上增大（augment）预测值。尽管经验和见识无法替代，我们还是可以常常回顾原先的预测经验来看看什么信息本该用于改善预测。也许我们不能精确地预测未来，但至少能够避免一些未来的严重失误。

13.4 拉式计划

将**生产计划与控制（production planning and control）**问题分解为可管理的部分的一种逻辑和习惯做法是建立层级计划体系。我们在图 3.2 中介绍了典型的 MRP Ⅱ层级。然而，那种体系基于 MRP 任务投放的推式机制。如我们在第四章对 JIT 的讨论和第十章对推式和拉式的比较，拉式系统比推式系统有着更多的可能收益。简单地将，拉式系统

1. **更有效率（More efficient）**，在于它们能用较少的平均 WIP 到达与推式系统相同的产出。
2. **更容易控制（Easier to control）**，因为它们的控制依赖于设定（容易观察）WIP 水平，而推式系统依赖于投放速率。
3. **更稳健（More robust）**，如果 WIP 水平和投放速率发生同等百分比的错误，拉式系统绩效降低地较少。
4. **更多地支持质量改善（More supportive of improving quality）**，拉式系统的低 WIP 水平需要高质量（来防止中断（disruption））并促进高质量（通过缩减排队和加速缺陷的检测）。

这些益处促使我们将拉的各个方面并入我们的制造控制系统。不幸的是，从计划的角度看，拉式体系有个缺陷——它天生在我们设定的 WIP 水平上*速率驱动*（*rate-driven*）地运行。能力缓冲（如，可用于转换之间的加班的预防性维护时间）被用于促进非常稳定的节奏，它反过来需要非常稳定的需求。为了达到这个要求，JIT 文献高度强调生产平滑。

速率驱动的体系看起来很吸引人，却不适合于计划。拉式体系中没有与客户交期的自然连接。客户“拉出”它们需要的，信号（卡片或其他的）触发补给（replenishment）。但是

在实际发生之前，系统不会提供关于需求的任何信息。因此，拉式体系不能提供计划原材料的获得、补给以及机器维护的时机等等的内在机制。

与之相反，如在第五章所见，推式体系几乎是噩梦，但它们极其适合于计划。客户交期和订单投放到系统之间有简单而直接的连接。例如，在因需定量（lot-for-lot）MRP 体系，计划投入量（planned order release）就是客户需求（仅仅依据生产提前期向前滚动）。只要 MRP 的无限产能假设没有把这些提前期做到大得离谱，我们就可以使用它们来驱动各种计划模块。事实上，这正是 MRP Ⅱ体系所做的。（430|431）

这时候的问题就是，我们能否既获得推式的好处又开发出具备一致性的计划结构？我们认为答案是可以。可是，联系速率驱动的拉式体系与交期的机制一定比 MRP 的简单时间滚动（time phasing）复杂。我们所知的最简单的连接即为**传送带模型（conveyor model）**，如图 13.8 所示。在接下来的章节里，我们要广泛地依赖它。

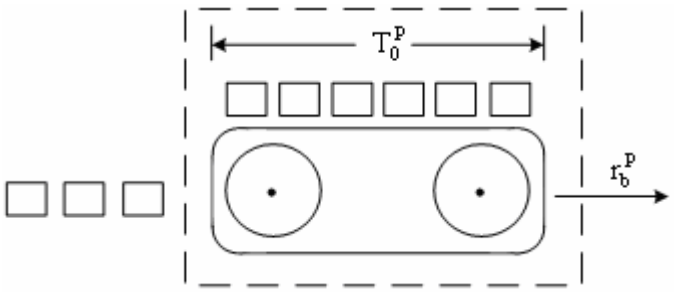

图 13.8 产线的传送带模型

传送带模型建立在这样的观察之上，即拉式体系维持相当稳定的 WIP 水平，所以产线的速度和作业通过产线的时间恒定相对。它使得我们可以用两个参数刻画一条产线：**实际生产速率（practical production rate）** r_b^P 和**最短实际提前期（minimum practical lead time）** T_0^P。它们与第七章定义的**瓶颈速率（bottleneck rate）** r_b 和**原始加工时间（raw process time）** T_0 及第九章引入的这二者的理想值 r_b^* 和 T_0^*，除些许区别之外，作用基本相同。不同于瓶颈速率，实际产出速率是产线的期望产出。这个速率也可以依据部件的复杂性来进行标准化（如，可以用在瓶颈作业处消耗的工时数来计量部件）。因此，由于 r_b 为产线的产能，在利用率 $u = r_b^P / r_b$ 时将有 $r_b^P < r_b$。类似地，T_0^P 是通过产线的实际最短的（即，无排队）可能时间。这将包括引起短期中断的扰动，如换模、路线上机器失效以及路线上等待转运等等其他不包括排队的延迟。从而，$T_0^P > T_0$。

使用里特定律，可知 CONWIP 水平 W 应为

$$W = r_b^P \times T_0^P$$

我们现在可以使用传送带模型来预测加工任务何时被产线或加工中心完成。例如，假设向已有 n 件加工任务在排队等待获准进入（即，等待获得传送带的空间）的 CONWIP 产线投放一件任务。该任务直到完成加工的时间 l 为

$$l = \frac{n}{r_b^P} + T_0^P = \frac{n+W}{r_b^P} \qquad (13.19)$$

例如，假设图 13.8 描绘的是一条电路板组装线。该产线以 r_b^P = 2 件/小时的平均速率运行，每个加工任务由一个标准尺寸容器内盛装的电路板构成。一旦启动作业，每个加工任务需要 T_0^P = 8 小时的平均时间来完成。那么，一个新的加工任务前面有 n = 3 件等待投入产线（即，等待 CONWIP 的授权信号），则其完成时间 l 的均值为

$$l = \frac{n}{r_b^P} + T_0^P = 3/2 + 8 = 9.5 \text{ 小时}$$

在第十五章对传送带模型的精炼中，我们将再次提起这个例子，并在生产速率中加入变动性。（431|432）

通过估计特定加工任务的输出时间，我们可以解决许多问题：

1. 如果销售人员明了工厂的负载，他们将可以使用传送带模型来预测新订单需要多久来完成，从而能够向客户报告合适的交期。

2. 如果我们明了系统如何随时间演化（即，什么作业在产线中，什么作业在排队等待），就可以“模拟”产线的绩效。这将为分析不同的优先规则或产能决策对产出的影响的“如果那么”工具提供基础。如在第三章所见，能力需求计划（CRP）试图做这样的分析。然而，当时我们就指出，CRP 使用的无限产能模型使得资源满载或过载时的预测失效。做这种预测的更精巧的有限产能模型已经推出市场。尽管比 CRP 更加精确，有限产能模型常常需要大量的数据和复杂的计算，类似于离散事件仿真模型。传送带模型可以简化数据需求与计算，我们将在第三篇的不同章节讨论这个问题。

3. 我们可以使用传送带模型确定作业完成情况是否能满足客户交期，从而开发用于设定作业投放时间的优化模型。

通过解决这些及其他问题，传送带模型给出了拉式生产系统的计划体系的关键点。当产线简单到可以直接调用它时，它就是一个强大的整合工具。我们给出利用这种整合的体系的轮廓，并在第三篇的余下部分补充细节以及讨论将传送带模型被过度简化的情形一般化。

13.5 层级生产计划

以传送带模型预测作业的完成，我们可以为拉式生产系统开发层级**生产计划与控制（production planning and control，PPC）**体系。图 13.9 阐明了这个层级，从顶层的长期战略议题跨越到底层的短期控制议题。

图 13.9 中的每个矩形框都代表一个独立的决策问题，因而成为一个**计划模块（planning module）**。[2]圆角矩形框代表某一模块的输出，它们中的大多数用作其他模块的输入。椭圆形框代表外生于计划层级（如，通过市场调查或工程设计）的模块输入。最后，箭头指示模块之间的相互依赖。

[2] 我们用术语模块来表示用于解决计划问题的分析模型、计算工具和人脑判断的结合物。这样，它们永远不是，也不应当是，完全自动的。

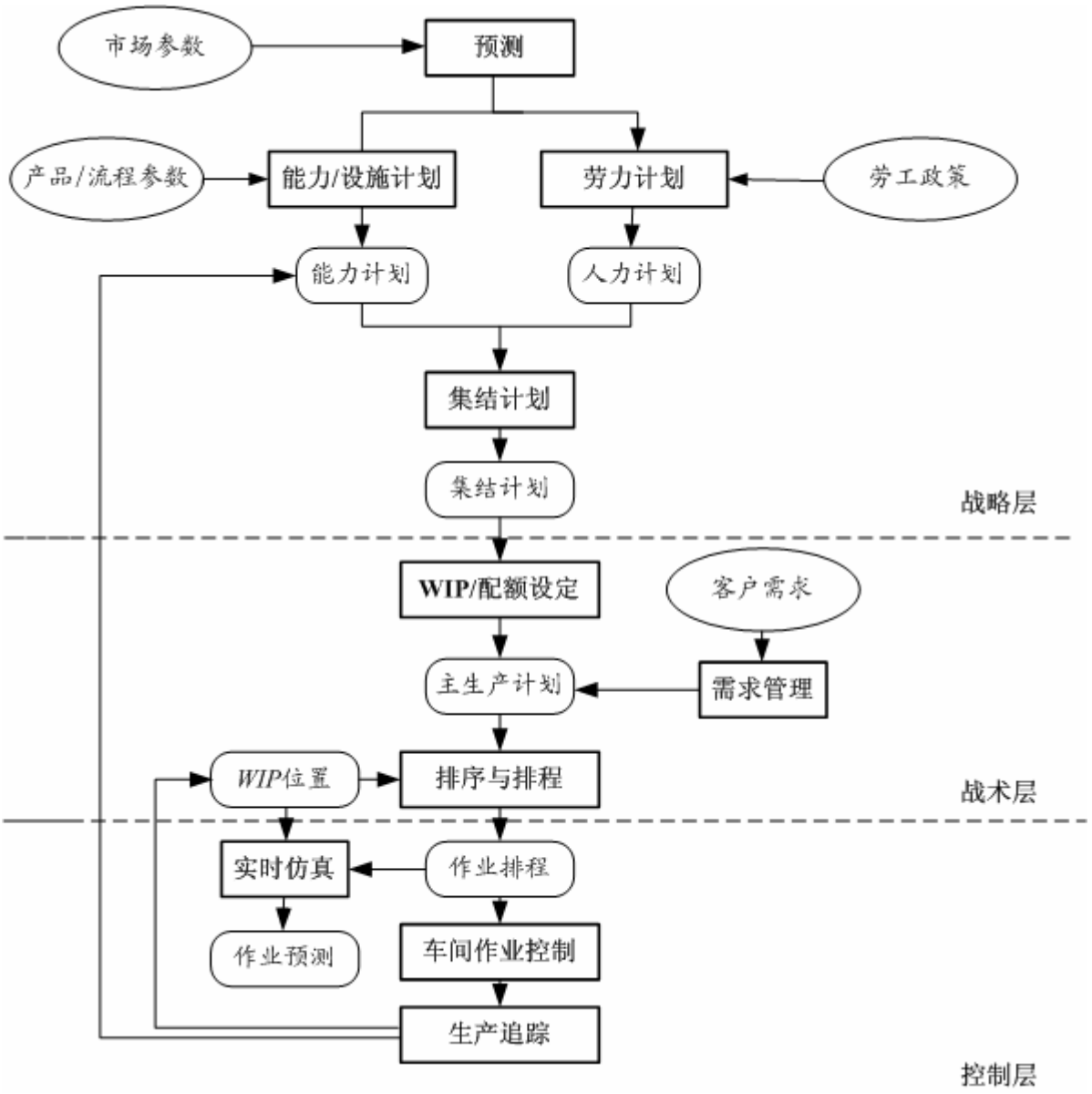

图 13.9 拉式系统的生产计划与控制层级

这个 PPC 层级被分为三个基本的水平，相应的是长期计划（战略）、中期计划（战术）和短期计划（控制）。当然了，从企业的角度看，还有一些在图 13.9 所示的之上的级别，如产品开发和业务计划等。这些当然是重要的业务战略决策，并且它们和制造职能的相互作用需要慎重的考虑。事实上，我们正希望未以制造为职业的读者也能积极创造将制造更多地纳入这些领域的机会。但是，我们还是会坚持关注于运营，并假定业务战略决策，如处于哪种业务和开发哪种产品，已经制定。因此，当提及战略时，我们指的是工厂的战略（*plant strategy*），而它仅为整体业务战略的一部分。（433|434）

图 13.9 所示的长期战略计划工具的基本功能是建立能够满足工厂整体目标的生产环境。在工厂水平，它始于输入市场信息并生成未来需求预测的**预测（forecasting）**模块，很可能用到类似于我们先前讨论的定量模型。**产能/设施计划（capacity/facility planning）**模块使用上面的需求预测和制造各种产品的工艺需求信息，来确定对实体设备的需求。类似地，**劳力计划（workforce planning）**模块使用需求预测，并依据企业的劳工政策生成雇用、解雇

与培训等等的人力计划。使用需求预测、产能/设施计划、劳力计划以及各种经济参数（物料费用、工资、采购成本（vendoring cost）等），**集结计划（aggregation planning）**生成对未来产品组合与产量的粗略预测。集结计划还可以说明其他相关问题，如哪些部件要自制而哪些部件由外部供应商代工，人力计划是否需要调整等等。

图 13.9 所示的中期战术工具接收战略层级的长期计划，结合客户订单的信息，来生成工厂为即将到来的生产做准备的行动的一般计划（通过获取物料，列示（line up）代工商等）。**WIP/配额设定（WIP/quota-setting）**模块将集结计划转化成拉式体系所需的卡片数目（card counts） 和周期性生产配额（periodic production quotas）。**主生产计划（master production schedule，MPS）**基于集结计划模块整理过的需求预测，而生产配额形成 MPS 的一部分。MPS 也包含确认的计划订单（firm customer orders），然后被需求管理（*demand management*）模块合适地平滑以用于拉式生产系统。**排序与排程（sequencing and scheduling）**模块将 MPS 转化成作业计划，规定在近期执行什么作业，如在下个月、下一天或下一班等。

图 13.9 所示的低层级工具直接控制工厂。**车间作业控制（shop floor control）**模块控制着依据排程产生的工厂内的实时物料流动，而**生产追踪（production tracking）**模块测定相对于排程的实际流程。在图 13.9 中，生产追踪模块也显示出另一种有用的功能，即为其他的计划模块提供反馈信息（如，产能数据）。最后，PPC 层级包含**实时仿真（real-time simulation）**模块，于是可以进行“如果-那么”的检查。例如，如果某些任务“加急（hot）”会怎样？

在接下来的小节，我们以综览的形式讨论各个水平的问题以及将这个 PPC 层级综合起来的哲学。我们将自下而上地展开讨论，因为这样有助于明确不同水平之间的相互作用。在接下来的章节中，还将给出如何构建各个模块的细节。那些将自上而下地进行，以强调各个计划问题与实际生产过程的联系。

13.5.1 产能/设施计划

一旦有了对未来需求的预测，以及做出努力满足它的战略决策，我们必须确保拥有足够的物理产能。这就是图 13.9 所示的**产能/设施计划（capacity/facility planning）**模块的功能。关于产能的最基本决策是要买多少以及哪种设备。自然地，它包括用于生产组件和最终产品的实际机器。但是，它也扩展到与支持这些机器相关的其他设施上，如工厂空间（factory floor space）、动力源、气/水/化学品供应、备件库存、物料搬运系统、WIP 与 FGI 存储以及人员供给水平等。（434|435）

产能/设施计划过程中考虑的问题有以下几个方面：

1. **产品生命周期（Production lifetime）**。设置哪种类型以及多大的产能，取决于我们打算在多长的时期内生产这种产品。近年来，产品生命周期已经显著缩短，甚至常常比设备的自然寿命还短。这就意味着，设备必须要么在产品的生命周期内收回成本，要么有足够的柔性从而用于制造其他的未来产品。在任何程度上预测未来的产品是什么都是困难的，所以将柔性的收益量化很不容易。但是柔性可能是设施规划最重要的方面之一，因为能迅速重组起来生产其他产品的柔性工厂可以成为一种非常有效的战略武器。

2. **外购决策（Vendoring options）**。在描述要安装的设备的特性之前，必须对制成品和子组件做出“自制或外购（make or buy）”的决策。这是一个复杂的问题，无法在这里全面介绍，所以提供一些观察结果。

a. 自制或外购的决策不应仅仅用于成本。将一种产品外包是由于看起来外包的单位成本低于（满载的）自己生产的单位成本，这样可能是有风险的。因为单位产品强烈取决于一般管理费用（overhead）的分派方式，看起来是局部理性的决策可能是会给全局带来灾

难。例如，自己生产的单位成本高于外部供应商的报价时把产品外包，可能不会将计入单位成本的一般管理费用削减多少。进而，这些成本被分摊到内部制造的其他产品上，引发单位成本上升，使它们也成为外包的候选项。许多企业陷入事实上一轮又一轮在单位成本比较基础上外包的“死亡螺旋”，以上是些具体例子。除了与外包相关的经济问题，内部制造还有其他的好处，如学习效应、掌控自己命运的能力、对排程的更紧密控制等等，都是简单的成本比较不能获得的。

b. 自制或外购决策应当关注长期。我们看到过通过一系列外包决策从制造发展到分销/服务的企业。这并不一定是坏的转变，但一定要对后果有充分了解，并对作为非制造实体在市场上的生存和发展能力有详尽考虑。

c. 自制或外购决策关系到是否要生产产品，它同时是一个产能计划决策。然而，许多制造经理都倾向将有自制能力的产品的部分份额转向代工。这种代工能增大产能，平滑工厂的负荷。由于哪种产品、多大数量要代工决定于产能和计划产量，这个问题溢出到制定长期生产计划的集结计划模块。我们将稍后在第十六章中更详细地讨论这个问题。从高层级的战略视角看，一定要记住，将业务转向承包商会使它们形成能力并在某天成为竞争对手。IBM 选择微软供应其个人电脑的操作系统，就是一个可能会发生什么的实例。（435|436）

3. **定价（Pricing）。**在本书中我们力图避开定价的问题，因为工厂对它的影响甚微。然而，在产能决策中，有效的经济性分析不可能在缺少对价格的某种预测时做出。我们需要知道销售可以产生多少收入，从而决定某种设备配置是否在经济上可行。由于价格常常受巨大的不确定性的影响，敏感性分析在此领域非常重要。

4. **资金的时间价值（Time Value of Money）。**一般地，产能提升和设备改进是资本支出的结果（are made as capital requisition），并随着时间折旧。因而，利率和折旧进度对设备的选择将产生显著的影响。

5. **可靠性和可维护性（Reliability and maintainability）。**如我们在第二篇中讨论的，可靠性（如，平均失效间隔（MTTF））和可维护性（如，平均恢复间隔（MTTR））是产能的重要决定因素。记起可用率（availability）A（机器正常作业时间的比例）为

$$A = \frac{MTTF}{MTTF + MTTR}$$

显然地，其他因素不变的情况下，我们希望 MMTF 很大而 MTTR 很小。但其他因素永远不可能不变，就像我们在下面两条中指出的。

6. **瓶颈效应（Bottleneck effects）。**从第二篇的讨论就应当明白，瓶颈资源处产能的提升，比非瓶颈资源处的提升，一般对产出有更大的影响。因此，花费额外的资金购买高速或高可用率的机器似乎在瓶颈资源处最有吸引力。然而，稳定的、明显的瓶颈可能不存在；除此之外，如我们在下一条中指出的，这种过于简化的推理还有其他问题。

7. **拥堵效应（Congestion effects）。**当今美国工业的产能分析中被忽视最多的一个因素，就是变动性。如我们在第二篇中一次又一次地看到的那样，变动性降低绩效。显著地受到失效影响的机器变动性，是产出的一个重要决定因素。一旦考虑到变动性，可靠性和可维护性在瓶颈资源处是重要因素，在非瓶颈资源处也是。

我们将在第十八章更详细地讨论产能/设施分析问题。现在，我们应当有长期战略的眼光，并在某一水平上明确地考虑变动性。依据层级计划计划体系，产能计划作业的输出是对一段计划展望期内工厂物理能力的预测。这段计划展望期必须足够长从而做出集结计划——一般在两年的数量级上。

13.5.2 劳力计划

图 13.9 所示的产能/设施计划模块决定了需要什么设备，与之类似地，**劳力计划（workforce planning）**模块决定了需要什么样的劳动力来支持生产。这两个计划问题都包括长期情况，因为工厂实体或者劳动力都不能在短期内激进调整。所以，这两个计划模块都与对需求的长期预测协同进行，并致力于构建实现整个体系的目标的环境。当然了，事件的实际结果都不会精确地符合计划，所以长期产能/设施计划与劳力计划都受随时间做出的短期修正的影响。（436|437）

长期中要解决的基本劳动力问题是，多少以及何种类型的劳动力要成为可用的。这些问题要在企业劳工政策形成的约束下回答。例如，在没有工会组织的工厂里，劳动合同可能会限定谁可以被雇用或解雇，不同的劳动力类型可以被指派哪些不同的任务，以及可以在哪些时段内工作。通常地，管理层花费大量时间推敲这些与劳工的合约的细节，而用于决定为支持长期生产计划需要什么样的劳动力的时间就少得多。尽管谨慎地使用劳力计划模块也不能避免劳资冲突，它却有助于使双方关注对企业具有战略重要性的议题。

大多数长期劳力计划的基础是对工厂生产的产品所需的**标准工时（standard hours）**的一系列估计。例如，一台商用排气罩（vent hood）可能需要一位焊工用 20 分钟（三分之一小时）来组装。如果焊工一周有 36 小时可用，则他的产能为 36×3 = 108 件/周。从而，540 件排气罩的生产计划将需要五个焊工。

简单的标准工时变换可能是劳力计划模块的起始点。然而，它远远不能完整地代表劳力计划的所有问题。这些问题有：

1. **作业员可用率（Worker availability）。**标准工时估计必须要足够精巧，以至于能包含休息、休假、培训和其他降低作业员可用率的因素。许多企业设定“宽放系数（inflation factor）”来将直接需要的劳动力转换成“实际（onboard）”劳动力。例如，1.4 的乘数意味着，要雇用 14 名工人来产生 10 名工人在某一班次中始终进行作业的效果。

2. **劳动力稳定性（Workforce stability）。**生产需求可能会突然走高或走低；但一般来说，劳动力数量激进上升或下降既不可能也不值得。企业招募高质量员工的能力，以及它的整体工作氛围，可能都会受到劳动力规模变化的强烈影响。这样的“软性问题” 中有些难于整合进入模型，但它们对维持一支高产出的员工队伍绝对是重要的。

3. **雇员培训（Employee training）。**培训新招募人员要花费资金以及当前员工的时间。此外，没有经验的员工需要时间来达到全生产力。这些考虑会反对劳动力的激进大幅扩充。然而，当业务成长需要劳动力的急速膨胀，就要做出一致努力来维护企业文化（即，引起最初的成长的一切事物）。

4. **短期柔性（Short-term flexibility）。**劳动力不仅仅由数量来描述（A workforce is described by more than head count）。员工之间的交叉培训是工厂柔性（它对产品系列和产量的短期变化的响应能力）的一个重要决定因素。因而，劳力计划需要将视野超出生产计划来考虑计划外的、系统应当能应付的突然事件（紧急订单、新产品的巨大成功）。

5. **长期敏捷性（Long-term agility）。**标准工时方法简单地将劳动力视为与物料和资本设备一道输入生产因素。但员工的意义不止于此。在当今时代，产品和流程持续变更，员工是敏捷性（工厂在新产品引入后迅速改装制造系统实现高效生产的能力）的关键来源。所谓敏捷制造（agile manufacturing）就是极大的依赖其人员，经理和职工，来学习和跟随变更产生进化。（437|438）

6. **质量改进（Quality improvement）。**如在第十二章提到的那样，质量，不论是内部的还是外部的，都是一系列因素的结果，其中有许多都直接受控于员工。在质量控制方法上教育机器操作员，交叉培训员工从而使他们产生对自己行为引起的质量影响的全局性评价，缓和新员工的流入以使企业的质量意识不受破坏——所有这些都是持续改进质量的计划的重

要部分。这些因素难于明确地整合进入人力计划模型，但重要的是在整个劳力计划模块中他们得到承认。

劳力计划是一个深入、影响深远的主题，接近制造管理的核心。同样地，它远超出运营管理或工厂物理学的范围。在第十六章中我们以解析的视角再次讨论这个话题，并检视劳力计划与集结计划的关系。这是劳力计划的一个有用的起点，但我们提醒读者它也仅限于此。良好平衡的人力计划必须考虑先前列出的那些问题，并实际上需要来自制造组织的所有部门的输入。

13.5.3 集结计划

一旦估计了未来需求并决定了哪些设备和人力将是可用的，就能生成详细说明各期各种产品的产量的**集结计划（aggregate plan）**。这就是图 13.9 所示的**集结计划（aggregate plan）**模块的任务。不同的设施有着不同的优先权和运行特性，集结计划也随之不同。某些工厂中，主要问题是产品系列，所以集结计划主要是依据需求、产能和原材料可用性来决定各期各种产品的产量。另一些工厂中，至关紧要的问题是生产的时机，所以集结计划模块将寻求平衡生产成本（如，加班与劳动力规模的变化）与满足需求目标的同时仍保有库存的成本。还有其他一些工厂中，关注点将是人员增减的时机。在所有这些情况中，还会存在使用外部承包商来增加产能的可能性。

不管具体的明确表述的集结计划问题如何，能识别要建立哪些约束都是有意义的。例如，如果集结计划模块显示去年某一加工中心平均起来负荷沉重，那么我们就知道这是一处要小心管理的资源。我们将希望制定特别的运行政策，如使用流动人力，来确保这个工艺在休息和午餐期间保持运转。如果问题已经很严重，则可能有必要回到程序的开始，修改产能和人力计划，补充另外的机器和/或人力。

集结计划模块做出的决策需要相当多的预先计划。例如，如果寻求为夏季的一段高峰需求建立库存，显然我们必须考虑夏季之前的几个月的生产计划。如果希望调整人员来适应生产计划，则需要更多的预警。这就常常意味着集结计划的计划展望期必须相当长，一般是一年或以上。当然了，我们应当比展望期更频繁地更新集结计划，因为一年之久的计划在后期将是高度不可靠的。通常是每季度或每半年更新一次。

我们将在第十六章给出对具体的、有代表性的集结计划模块的明确表述。由于常常以在满足需求的条件下最小化成本的形式陈述问题，我们一般可以使用线性规划来辅助求解集结计划问题。线性规划的长处有（438|439）

1. 求解速度快，使我们能很快地解决庞大的问题。对于在如果-那么模式下使用集结计划模块，这一点尤为重要。

2. 提供了强大的敏感性分析能力，例如计算增加的产能能多大程度地影响总成本。这使我们能识别关键资源，并迅速测量各种变化的效力。

如将在第十六章所见的，线性规划还提供了表示不同的集结计划情形的极大柔性。

13.5.4 WIP 与定额计算

图 13.9 中紧邻集结计划模块工作的 **WIP/定额计算（WIP/quota-setting）**模块，被用于将集结计划转化成拉式体系的控制参数。记起拉式体系中关键控制指标是产线中的 WIP 水平，或卡片数量。另外，要将拉式体系与客户交期联系起来，我们还需要增设一项控制指标，称为生产定额。通过建立定额，再使用缓冲产能来保证有规律地实现定额，就可以使系统行

为接近前面讨论的“传送带模型”。传送带模型的可预测性使我们将系统输出与客户交期协调起来。

卡片数量（Card Counts）。将**WIP 设定（WIP setting）**，或是卡片数量设定，放在在图 13.9 所示的 PPC 层级的中层而非底层，是想提醒读者 WIP 水平不应当太频繁地调整。如第十章所讲的，WIP 是个相当迟钝的控制指标。为使产出追上需求而改变卡片数量不大可能起作用，因为系统不能足够快速地对此做出响应。因此，与这个层级体系中处于同一层次的其他决策类似，WIP 水平的重估应当不太频繁，比如说，每季度一次。

幸运的是，WIP 水平是个不敏感的控制指标的事实也使它相对来说易于设定。只要 WIP 水平足够实现期望产出并且不是太高，系统就可以良好地运行。因此，并不需要开发非常精巧的工具来计算 WIP 水平。在由推式转向拉式的系统，可能得将拉式系统的初始 WIP 水平设定为先前推式情况下的平均水平。然后，一旦系统运行平稳，就逐渐降低。如果使用了看板，产线的不同位置设定了不同的 WIP 水平，则减少那些有着永不能或很少能清空的长队列的工站处的卡片。如果使用了 CONWIP，则总体的 WIP 水平可以逐渐降低。一旦建立起可用的 WIP 水平，它们不应被频繁地调整，以保证变化的做出是为了响应长期趋势而非短期波动。

如果必须为一条按 CONWIP 方式运行的新建（或重布置）路线设定 WIP 水平，就不能依赖历史表现来获取合适的数值了。在这种情形下，以下给出一种经验做法。首先，建立一个期望且是可行的周期时间 CT，并确定实际生产速率 r_b^P（如，瓶颈速率 r_b 的一个可行比例）。然后使用里特定律求解 WIP 水平为

$$WIP = r_b^P \times CT$$

如果 r_b^P 和 CT 都是现实的，这种方法将得出 WIP 的一个合理起始点，之后再随着时间做调整。一般来说，一定要注意不能低估了可行的周期时间或实际生产速率，因为这样将会导致过低的 WIP，从而引起过低的产出。（439|440）

生产定额（Production Quotas）。除了 WIP 水平，另一个控制拉式系统的关键参数是生产定额。因此，**定额计算（quota setting）**也包含在图 13.9 所示的 PPC 层级的 WIP 设定模块中。

生产定额的基本含义是，我们为作业建立一个周期性的数量，并将（几乎）总是能在定额时期内完成。这个时期可以是一班、一天或一周。以其最严格的定义形式，生产定额意味着

1. 在该时期中，一旦达到定额，生产即停止。
2. 正常时间内完成不了的，要以期末加班来补足。

它使我们可以指望稳定的产出，并促进计划和提报交期。当然了，实际中很少有定额体系严格遵守这种协议。事实上，我们在第十章讲述的 CONWIP 的一种益处就是在环境允许时它可以先于进度表生产。然而，为了计划一个合理的周期性生产定额，按达到定额即停止生产的方式为系统构模也是可以的。

建立一个经济可行的生产定额需要同时考虑成本和产能数据。相关成本指的是与损失的产出和加班联系的成本。某一特定时段（如，一天或一周）内的产量的均值和标准差都是重要的产能参数。标准差也是需要的，是因为输出的变动性对我们制定目标生产定额有影响。

一般来说，生产过程的变动性越大，我们就越有可能达不到定额（miss the quota）。

为了说明这一点，请看图 13.10。假设我们为常规时间产量（regular time production）设定的定额为Q件。[3]如果在常规时间内做不出Q件，就必须加班（如，周六和周日）来补足落后的进度。由于存在一般性的偶然事件（机器失效、作业员旷工、产出损失，等等），在常规时间内完成的实际数量每期都不同。图 13.10 表示常规时间产量的两种概率分布，它们的均值μ相同的但是标准差σ不同。达不到定额的概率有曲线以下、Q值以左的面积表示。有着较小的标准差的曲线A下的面积较小，所以它达不到定额的概率较小。这就意味着，如果定义乐于接受的与“服务水平”形似的达不到定额的概率，就将能为曲线A设定高于曲线B的定额我们可以更近地盯住产能，因为曲线A的可预测性越高，我们就对以常规方式实现目标的能力越有信心。

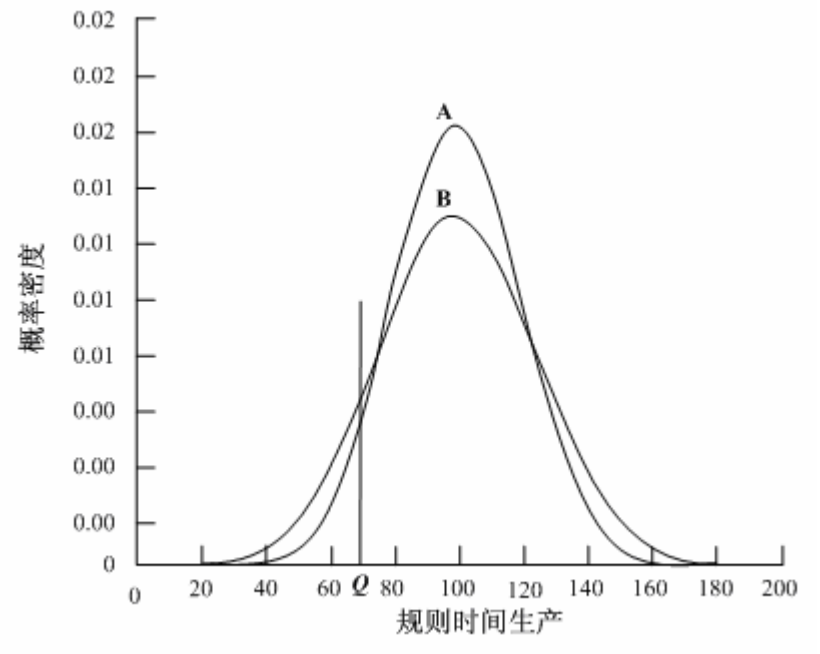

图 13.10 在不同的产量率分布下达不到定额的概率

以上的分析指出，如果知道了常规时间产量的均值μ和标准差σ，[4]一种非常简单的设置生产定额的方法是计算可以在所有时间的S百分比内实现的定额，其中S有使用者选取。如果常规时间产量X可以恰当地由正态分布近似估计，则可以通过寻找满足以下条件的Q值来计算出合适的定额（440|441）

$$\Phi\left(\frac{Q-\mu}{\sigma}\right)=1-S$$

其中$\Phi(\cdot)$表示标准正态分布的累积分布函数（cdf）。

例如，假设 $\mu = 100$，$\sigma = 10$，并且选择 $S = 85\%$作为服务水平。则定额 Q 就是满足下式的数值

[3] 例如，在单产品模型中，任务的单位就等于实体件数。在较复杂的多产品情形中，单位必须依产能做调整，如以在关键资源处所需的作业时间来计量。

[4] 我们将在第十四章中讨论由实际运营经验获取 μ 和 σ 估值的机制。

$$\Phi\left(\frac{Q-100}{10}\right) = 1 - 0.85 = 0.15$$

从标准正态分布表可以查到，$\Phi(-1.04) = 0.15$。因此，我们可以寻找到 Q 的值

$$\frac{Q-100}{10} = -1.04$$

$$Q = 89.6$$

这种简单方法的一个问题是，它仅考虑了产能，没有考虑成本。因此，它没有为所选的服务水平是否合适提供指导。较低的服务水平将导致较高的定额，它将提升产出，同时也提升加班成本。较高的服务水平将导致较低的定额，它将降低产出和加班成本。我们在附录13A 提供了一个平衡损失产出的成本和加班成本的模型，该模型更复杂的变种见 Hopp 等的著作（1993）。

13.5.5 需求管理

所有生产控制系统的效力都在很大程度上决定于其运行的环境。简单的流水线在非常简单的计划工具的管理下能运转良好，而复杂的加工车间即使有了非常精巧的工具可能还是一个管理上的噩梦。这就是生活的实情；有些工厂比另一些易于管理。它也是回忆起一条“JIT 的教训”的好理由，即环境是一种控制因素。例如，如果经理将机器分配到“单元”来生产特定组群的产品，从而使加工车间（job shop）像流水车间（flow shop），就可以极大地简化计划和控制过程。（441|442）

塑造计划层级最底层模块“所见”的环境的一个关键领域在于管理客户需求。图 13.9 所示的需求管理模块做到这一点，方法是过滤及可能地调整客户订单为某种形式以生成可管理的主生产计划。如我们在第四章所见的，平准需求或“生产平滑”正是 JIT 的本质特征。没有稳定的产量与产品系列，Ohno（1988）和其他 JIT 鼓吹者所描述的速率驱动、混合模型的生产方法无法运行。这就意味着，客户订单不能以它们到达的那种随意方式投入工厂。相反地，它们必须集中起来然后分组，从而保持工厂的相当恒定的负荷。 在需求管理模块内设置交期的定额以及建立短期 MPS 有很多种方法。如先前的讨论，如果建立起周期性生产定额，就可以使用传送带模型预测通过工厂的物流。在这些条件下，我们可以将客户交期的定额计算视为“填装传送带”。如果不需要担心机器换模并且有着缓冲的能力，就可以使用（13.19）式所刻画的传送带模型，按客户订单的接收顺序设置交期定额。然而，一旦存在变动性并且没有或少有能力缓冲，就必须以不同的程序（见第十五章）设置交期定额。类似地，如果依族类（即，共享重要的机器换模的部件）将产品成批，就可能希望用到第十五章讨论的排序技术。

尽管方法众多，要紧的问题不是用哪一种，而是有一种被使用。几乎任何一种能与排程协调一致的方法，都比几乎孤立于制造过程之外地设置交期好。

13.5.6 排序和排程

MPS仍旧是生产计划，还必须被转换成作业进度表来指导工厂的实际作业。在如图 3.2 所示的MRPⅡ层级，这项任务由MRP承担。[5]在图 13.9 所示的拉式体系生产计划与控制层级中，我们设定了与MRP类似的拉力，**排序与排程模块（sequencing/scheduling module）**。就

[5] 记起第三章中，MRP（“小 mrp”）指的是用于生成计划投入量的物流需求计划（*material requirements planning*），而 MRPⅡ（“大 MRP”）指的是合并 MRP 的总体计划体系的制造需求计划（*manufacturing resources planning*）。企业资源计划（*Enterprise resource planning，ERP*）将 MRPⅡ层级扩展到多工厂体系。

像在MRP中，这个排序与排程模块的目标是管理工单和物料的投放时机，并促进它们流过工厂的运动。

按照爱因斯坦的名言的意思，我们应该坚持将作业排程做得尽量简单，却不能简陋。目标应该是向现场人员提供足够的信息使他们做出合理的控制决策，而不是过于严格地限制他们的选择或是将排程弄得笨拙。在实际中这就意味着，不同的工厂需要不同的排程方法。在没有明显换模时间的简单流水线，工单的简单顺序，很可能是依据最早交期（EDD）做出的，可能就足够了。在这种情形下，保持其他工站处工单的先进先出（FISFO）顺序将产生高度可预测、易于管理的产出流。（442|443）

然而，在有着许多路线、机器换模、部件的装配线的极为复杂的加工车间，简单的顺序很难明确定义，更不用说运行良好了。在更复杂的情况下，MRP 也不可行了。结果就是，MRP 模块和排序/排程模块之间的迭代将成为必要。检测排程不可行与提醒补救（如，增加产能、后延交期）的程序称为**产能约束的物料需求计划（capacitated material requirements planning）**，或是 **MRP-C**，将在第十五章讨论。这个程序将需求管理、MPS 和排序/排程功能整合到一个模块中。在此种复杂的情形中，我们可能需要提供相当详细的进度表，给出作业和物料的具体投放时机以及作业到达工站的预测时间。当然了，要生成这样的进度表，其对数据的需求和系统维护的一般费用都是巨大的，这就是我们为复杂性付出的代价。

13.5.7 车间作业控制

不管排程工具是多么精确和巧妙，实际作业顺序永远不会完全遵照进度表。图 13.9 所示的**车间作业控制（shop floor control，SFC）**模块把作业进度表作为一般性指导，在可能的时候坚持它，也在必要的时候调整它。例如，如果某处的机器失效延迟了组装作业所需部件的到达，SFC 模块必须决定作业顺序该怎样变化。理论上，这可能是个极其复杂的问题，因为选择有无穷多个——我们可以等待延迟的部件，可以将排在后面的作业提前，可以打乱整个进度表，等等。但在实际中，我们必须快速决定，实时地，并因此不能指望考虑到每一种可能性。故而，SFC 模块必须将使用者的注意力约束到合理的行动的集合，并辅助做出有效、稳健的决策。

为了发挥在第十章中讨论的拉的优势，我们喜欢这样一个基于拉式机制的 SFC 模块。CONWIP 协议可能是最简单的方式，因此值得首选。为了将 CONWIP 和排序/排程模块联系协力使用，我们设定一个 WIP 上限并且在 WIP 高过这个水平时禁止向产线投料。这样将有助于在工厂落后于进度且再投料也无济于事的时候延迟投放。在一切顺利时，CONWIP 也提供超前于进度表作业的机制。如果在下一项作业计划投放之前 WIP 水平低于上限，我们就希望允许这项作业提前开始。只要不是远远地超前于进度表，或过早地赋予部件“个性”而引起柔性的损失，这种超前作业的协议将是非常有效的。

第十四章致力于解决 SFC 问题；我们将在那里讨论 CONWIP 型 SFC 模块的实施，并将识别需要更复杂的 SFC 方式的情形。

13.5.8 实时仿真

在这样一本制造管理的书中，人们容易局限于“永远不要有加急（hot）的加工任务”或“总是遵照发布的排程”。当然了，如果这些刚性规则能被遵守，工厂也就易于管理了。但是，制造工厂的终极目标不是让它的经理们的日子舒适；而是通过满足客户来挣钱。由于客户有改变主意、追逐新奇等特征，几乎所有制造环境的实情都是有时候会发生紧急情况，因此某些作业必须给予特殊待遇。人们可能会希望不会总是发生这种情况（尽管太频繁地如此，如同我们曾在一家工厂见到的 MRP 上的每件任务都被指定为“加急（hot）”）。但是，假设它发生了，还是要设计能在这些不测中生存的计划体系，甚至还能为它们提供援助。这

就是图 13.9 所示的**实时仿真（real-time simulation）**模块的任务。（443|444）

我们已经发现仿真在处理如加急任务这样的紧急情况时很有用。然而这里所说的仿真，不是指完全的有随机数发生器和数据输出分析的蒙特卡洛仿真。相反地，我们指的是非常简单的、能在短期模仿工厂行为的确定性模型。一种选择就是使用先前描述的传送带模型来代表各个加工中心的行为，输入系统中的 WIP 位置、一系列预期的投料和一系列产能数据（包含人力）来生成一系列任务输出时间。这样一个模型在短期内可以做到精确（如，在下一周内），但由于不能将机器失效这样的不可预见事件整合进去，它还不能在长期内做到精确。因此，只要我们将这种模型的用途限制在回答短期问题——如果加速任务 n，其他各种任务的交期将会发生怎样的变化？——这种工具就将非常有效。在采取紧急行动之前了解其可能的后果，可以防止严重扰乱工厂而获益了了的情况。

13.5.9 生产追踪

实际世界中常常会出现需要经理们来人为干涉的意外情况。听起来会让生产系统的设计者感到气馁，但它正是制造经理们存在的主要理由之一。好的制造经理应当争取使系统在大多数的时间内平稳运行，但也预备在失稳时采取纠偏行动。为了实时地检测出问题并明确地做出反应，经理必须在手头上有关键的数据，包括部件在厂内的位置、设备的状态（如，开机、停机、检修中）以及满足排程的进度。图 13.9 所示的**生产追踪（production tracking）**模块就是以可用的格式列表显示这种数据的。

图 13.9 所示的模块中有许多都依赖于估计的数据。特别地，产能数据对几种计划决策都是至关重要的。估计当前所有设备的产能的广泛使用的方法是，将额定的产能（如，按件/小时计）作为初始值，再按各种扰动（停机、作业员不可用、换模等等）削减这个数值。由于各种扰动都是随机出现的，这种估值可能存在严重偏差。出于这个原因，有必要使用生产追踪模块来收集和更新其他的计划模块使用的产能数据。如将在第十四章所见的，我们可以使用指数平滑技术来从预测中生成产能的平滑估值并监控随其时间的变化趋势。

13.6 结论

本章中，我们提供了与第四章、第十章中讨论的拉式生产体系一致的生产计划与控制层级的纵览。由于生产体系的构建方式有许多种，并且不同的环境很可能需要不同的体系，这个纵览有着充分的一般性。我们将在接下来的章节中对各个计划模块做详细说明。现在，我们给出与计划层级的整个结构相关的要点汇总，作为本章的结尾：（444|445）

1. *计划应当分层制定*。不可能在粗糙、随机的数据基础上使用精确、详细的模型做出全面、长期的决策。一般来说，计划展望期越短，需要的细节越多。出于这个原因，有必要将计划问题分解为长期（战略）、中期（战术）与短期（控制）问题。类似地，关于产品的信息也随时间缩短而增加，如在长期计划总产量，在中期计划产品族，在短期计划具体料号。

2. *一致性非常重要*。良好的单个模块可能因缺乏协调而被破坏。一定要使一般的产能假设与人力假设一致，并与向不同的计划模块输入的数据协调。

3. *反馈可以强化一致性和组织学习*。一些制造经理延续使用低质量的数据，不去检查它们的准确性，也不建立从实际工厂运作中收集更好的数据的体系。不管采取哪种反馈形式（如，人工或自动的），为重要参数的更新提供某种反馈都是非常重要的。更进一步地，通过提供观测与追踪流程的机制，反馈促进了持续改进的环境。

4. 不同的工厂有不同的需要。以上的原则都是一般性的；它们的实施细节必须依环境

而定。小规模、简单的工厂可以用不复杂的人工程序实现许多计划步骤。大规模、复杂的工厂可能需要精巧的自动系统。尽管我们会在第三篇的余下篇章中尽可能具体地讨论，但读者要注意不能过于纠缠于细节；它们仅仅是为了阐述问题、启发灵感而提出的，不能代替基础知识、直觉和综合的深思熟虑的应用。

附录 13A
一个定额计算模型
