第六章　排队论模型

排队论起源于1909年丹麦电话工程师A. K. 爱尔朗的工作，他对电话通话拥挤问题进行了研究。1917 年，爱尔朗发表了他的著名的文章—“自动电话交换中的概率理论的几个问题的解决”。排队论已广泛应用于解决军事、运输、维修、生产、服务、库存、医疗卫生、教育、水利灌溉之类的排队系统的问题，显示了强大的生命力。

排队是在日常生活中经常遇到的现象，如顾客到商店购买物品、病人到医院看病常常要排队。此时要求服务的数量超过服务机构（服务台、服务员等）的容量。也就是说，到达的顾客不能立即得到服务，因而出现了排队现象。这种现象不仅在个人日常生活中出现，电话局的占线问题，车站、码头等交通枢纽的车船堵塞和疏导，故障机器的停机待修，水库的存贮调节等都是有形或无形的排队现象。由于顾客到达和服务时间的随机性。可以说排队现象几乎是不可避免的。

排队论（Queuing Theory）也称**随机服务系统理论**，就是为解决上述问题而发展的一门学科。它研究的内容有下列三部分：

（i）性态问题，即研究各种排队系统的概率规律性，主要是研究队长分布、等待时间分布和忙期分布等，包括了瞬态和稳态两种情形。

（ii）最优化问题，又分静态最优和动态最优，前者指最优设计。后者指现有排队系统的最优运营。

（iii）排队系统的统计推断，即判断一个给定的排队系统符合于哪种模型，以便根据排队理论进行分析研究。

这里将介绍排队论的一些基本知识，分析几个常见的排队模型。

§1　基本概念

1.1　排队过程的一般表示

下图是排队论的一般模型。

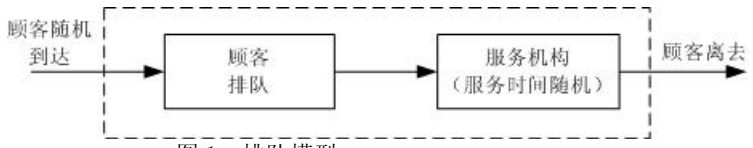

图1　排队模型

图中虚线所包含的部分为排队系统。各个顾客从顾客源出发，随机地来到服务机构，按一定的排队规则等待服务，直到按一定的服务规则接受完服务后离开排队系统。

凡要求服务的对象统称为**顾客**，为顾客服务的人或物称为**服务员**，由顾客和服务员组成服务系统。对于一个服务系统来说，如果服务机构过小，以致不能满足要求服务的众多顾客的需要，那么就会产生拥挤现象而使服务质量降低。 因此，顾客总希望服务机构越大越好，但是，如果服务机构过大，人力和物力方面的开支也就相应增加，从而会造成浪费，因此研究排队模型的目的就是要在顾客需要和服务机构的规模之间进行权衡决策，使其达到合理的平衡。

1.2　排队系统的组成和特征

一般的排队过程都由输入过程、排队规则、服务过程三部分组成，现分述如下：

1.2.1　输入过程

输入过程是指顾客到来时间的规律性，可能有下列不同情况：

（i）顾客的组成可能是有限的，也可能是无限的。

（ii）顾客到达的方式可能是一个一个的，也可能是成批的。

（iii）顾客到达可以是相互独立的，即以前的到达情况对以后的到达没有影响；否则是相关的。

（iv）输入过程可以是平稳的，即相继到达的间隔时间分布及其数学期望、方差等数字特征都与时间无关，否则是非平稳的。

1.2.2 排队规则

排队规则指到达排队系统的顾客按怎样的规则排队等待，可分为损失制，等待制和混合制三种。

（i）损失制（消失制）。当顾客到达时，所有的服务台均被占用，顾客随即离去。

（ii）等待制。当顾客到达时，所有的服务台均被占用，顾客就排队等待，直到接受完服务才离去。例如出故障的机器排队等待维修就是这种情况。

（iii）混合制。介于损失制和等待制之间的是混合制，即既有等待又有损失。有队列长度有限和排队等待时间有限两种情况，在限度以内就排队等待，超过一定限度就离去。

排队方式还分为单列、多列和循环队列。

1.2.3 服务过程

（i）服务机构。主要有以下几种类型：单服务台；多服务台并联（每个服务台同时为不同顾客服务）；多服务台串联（多服务台依次为同一顾客服务）；混合型。

（ii）服务规则。按为顾客服务的次序采用以下几种规则：

①先到先服务，这是通常的情形。

②后到先服务，如情报系统中，最后到的情报信息往往最有价值，因而常被优先处理。

③随机服务，服务台从等待的顾客中随机地取其一进行服务，而不管到达的先后。

④优先服务，如医疗系统对病情严重的病人给予优先治疗。

1.3 排队模型的符号表示

排队模型用六个符号表示，在符号之间用斜线隔开，即 $X/Y/Z/A/B/C$。第一个符号 X 表示顾客到达流或顾客到达间隔时间的分布；第二个符号 Y 表示服务时间的分布；第三个符号 Z 表示服务台数目；第四个符号 A 是系统容量限制；第五个符号 B 是顾客源数目；第六个符号 C 是服务规则，如先到先服务 FCFS，后到先服务 LCFS 等。并约定，如略去后三项，即指 $X/Y/Z/\infty/\infty/\mathrm{FCFS}$ 的情形。我们只讨论先到先服务 FCFS 的情形，所以略去第六项。

表示顾客到达间隔时间和服务时间的分布的约定符号为：

M —指数分布（M 是 Markov 的字头，因为指数分布具有无记忆性，即 Markov 性）；

D —确定型（Deterministic）；

E_k —k 阶爱尔朗(Erlang)分布；

G —一般（general）服务时间的分布；

GI —一般相互独立（General Independent）的时间间隔的分布。

例如，$M/M/1$ 表示相继到达间隔时间为指数分布、服务时间为指数分布、单服务台、等待制系统。$D/M/c$ 表示确定的到达时间、服务时间为指数分布、c 个平行服务台（但顾客是一队）的模型。

1.4 排队系统的运行指标

为了研究排队系统运行的效率，估计其服务质量，确定系统的最优参数，评价系统的结构是否合理并研究其改进的措施，必须确定用以判断系统运行优劣的基本数量指

标，这些数量指标通常是：

(i)**平均队长**：指系统内顾客数（包括正被服务的顾客与排队等待服务的顾客）的数学期望，记作L_s。

(ii)**平均排队长**:指系统内等待服务的顾客数的数学期望，记作L_q。

(iii)**平均逗留时间**：顾客在系统内逗留时间（包括排队等待的时间和接受服务的时间）的数学期望，记作W_s。

（iv）**平均等待时间**：指一个顾客在排队系统中排队等待时间的数学期望，记作W_q。

（v）**平均忙期**：指服务机构连续繁忙时间（顾客到达空闲服务机构起，到服务机构再次空闲止的时间）长度的数学期望，记为T_b。

还有由于顾客被拒绝而使企业受到损失的**损失率**以及以后经常遇到的**服务强度**等，这些都是很重要的指标。

计算这些指标的基础是表达系统状态的概率。所谓**系统的状态**即指系统中顾客数，如果系统中有n个顾客就说系统的状态是n，它的可能值是

（i）队长没有限制时，$n=0,1,2,\cdots$，

（ii）队长有限制，最大数为N时，$n=0,1,\cdots,N$，

（iii）损失制，服务台个数是c时，$n=0,1,\cdots,c$。

这些状态的概率一般是随时刻t而变化，所以在时刻t、系统状态为n的概率用$P_n(t)$表示。稳态时系统状态为n的概率用P_n表示。

§2 输入过程与服务时间的分布

排队系统中的事件流包括顾客到达流和服务时间流。由于顾客到达的间隔时间和服务时间不可能是负值，因此，它的分布是非负随机变量的分布。最常用的分布有泊松分布、确定型分布，指数分布和爱尔朗分布。

2.1 泊松流与指数分布

设$N(t)$表示在时间区间$[0,t)$内到达的顾客数（$t>0$），令$P_n(t_1,t_2)$表示在时间区间$[t_1,t_2)(t_2>t_1)$内有$n(\geq 0)$个顾客到达的概率，即

$$P_n(t_1,t_2)=P\{N(t_2)-N(t_1)=n\}\quad (t_2>t_1,n\geq 0)$$

当$P_n(t_1,t_2)$合于下列三个条件时，我们说顾客的到达形成泊松流。这三个条件是：

1° 在不相重叠的时间区间内顾客到达数是相互独立的，我们称这性质为无后效性。

2° 对充分小的Δt，在时间区间$[t,t+\Delta t)$内有一个顾客到达的概率与t无关，而约与区间长Δt成正比，即

$$P_1(t,t+\Delta t)=\lambda\Delta t+o(\Delta t) \qquad (1)$$

其中$o(\Delta t)$，当$\Delta t\to 0$时，是关于Δt的高阶无穷小。$\lambda>0$是常数，它表示单位时间有一个顾客到达的概率，称为概率强度。

3° 对于充分小的Δt，在时间区间$[t,t+\Delta t)$内有两个或两个以上顾客到达的概率极小，以致可以忽略，即

$$\sum_{n=2}^{\infty}P_n(t,t+\Delta t)=o(\Delta t) \qquad (2)$$

在上述条件下，我们研究顾客到达数 n 的概率分布。

由条件 2°，我们总可以取时间由 0 算起，并简记 $P_n(0,t)=P_n(t)$。

由条件 1°和 2°，有

$$P_0(t+\Delta t)=P_0(t)P_0(\Delta t)$$

$$P_n(t+\Delta t)=\sum_{k=0}^{n}P_{n-k}(t)P_k(\Delta t),\quad n=1,2,\cdots$$

由条件 2°和 3°得

$$P_0(\Delta t)=1-\lambda\Delta t+o(\Delta t)$$

因而有

$$\frac{P_0(t+\Delta t)-P_0(t)}{\Delta t}=-\lambda P_0(t)+\frac{o(\Delta t)}{\Delta t},$$

$$\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t}=-\lambda P_n(t)+\lambda P_{n-1}(t)+\frac{o(\Delta t)}{\Delta t}.$$

在以上两式中，取 Δt 趋于零的极限，当假设所涉及的函数可导时，得到以下微分方程组：

$$\frac{dP_0(t)}{dt}=-\lambda P_0(t),$$

$$\frac{dP_n(t)}{dt}=-\lambda P_n(t)+\lambda P_{n-1}(t),\quad n=1,2,\cdots.$$

取初值 $P_0(0)=1$，$P_n(0)=0(n=1,2,\cdots)$，容易解出 $P_0(t)=e^{-\lambda t}$；再令 $P_n(t)=U_n(t)e^{-\lambda t}$，可以得到 $U_0(t)$ 及其它 $U_n(t)$ 所满足的微分方程组，即

$$\frac{dU_n(t)}{dt}=\lambda U_{n-1}(t),\quad n=1,2,\cdots,$$

$$U_0(t)=1,\ U_n(t)=0.$$

由此容易解得

$$P_n(t)=\frac{(\lambda t)^n}{n!}e^{-\lambda t},\quad n=1,2,\cdots.$$

正如在概率论中所学过的，我们说随机变量 $\{N(t)=N(s+t)-N(s)\}$ 服从泊松分布。它的数学期望和方差分别是

$$E[N(t)]=\lambda t;\ \mathrm{Var}[N(t)]=\lambda t。$$

当输入过程是泊松流时，那么顾客相继到达的时间间隔 T 必服从指数分布。这是由于

$$P\{T>t\}=P\{[0,t)\text{内呼叫次数为零}\}=P_0(t)=e^{-\lambda t}$$

那么，以 $F(t)$ 表示 T 的分布函数，则有

$$P\{T\le t\}=F(t)=\begin{cases}1-e^{-\lambda t}, & t\ge 0\\ 0, & t<0\end{cases}$$

而分布密度函数为

$$f(t)=\lambda e^{-\lambda t},\quad t>0.$$

对于泊松流，λ 表示单位时间平均到达的顾客数，所以 $\frac{1}{\lambda}$ 就表示相继顾客到达平均间隔时间，而这正和 ET 的意义相符。

对一顾客的服务时间也就是在忙期相继离开系统的两顾客的间隔时间，有时也服从指数分布。这时设它的分布函数和密度函数分别是

$$G(t)=1-e^{-\mu t},\quad g(t)=\mu e^{-\mu t}$$

我们得到

$$\lim_{\Delta t\to 0}\frac{P\{T\le t+\Delta t\,|\,T>t\}}{\Delta t}=\lim_{\Delta t\to 0}\frac{P\{t<T\le t+\Delta t\}}{\Delta t P\{T>t\}}=\mu$$

这表明，在任何小的时间间隔 $[t,t+\Delta t)$ 内一个顾客被服务完了（离去）的概率是 $\mu\Delta t+o(\Delta t)$。μ 表示单位时间能被服务完成的顾客数，称为平均服务率，而 $\frac{1}{\mu}$ 表示一个顾客的平均服务时间。

2.2　常用的几种概率分布及其产生

2.2.1　常用的连续型概率分布

我们只给出这些分布的参数、记号和通常的应用范围，更详细的内容参看专门的概率论书籍。

（i）均匀分布

区间 (a,b) 内的**均匀分布**记作 $U(a,b)$。服从 $U(0,1)$ 分布的随机变量又称为随机数，它是产生其它随机变量的基础。如若 X 为 $U(0,1)$ 分布，则 $Y=a+(b-a)X$ 服从 $U(a,b)$。

（ii）正态分布

以 μ 为期望，σ^2 为方差的**正态分布**记作 $N(\mu,\sigma^2)$。正态分布的应用十分广泛。正态分布还可以作为二项分布一定条件下的近似。

（iii）指数分布

指数分布是单参数 λ 的非对称分布，记作 $\mathrm{Exp}(\lambda)$，概率密度函数为：

$$f(t)=\begin{cases}\lambda e^{-\lambda t}, & t\ge 0\\ 0, & t<0\end{cases}$$

它的数学期望为 $\frac{1}{\lambda}$，方差为 $\frac{1}{\lambda^2}$。指数分布是唯一具有无记忆性的连续型随机变量，即有 $P(X>t+s\,|\,X>t)=P(X>s)$，在排队论、可靠性分析中有广泛应用。

（iv）Gamma 分布

Gamma 分布是双参数 α,β 的非对称分布，记作 $G(\alpha,\beta)$，期望是 $\alpha\beta$。$\alpha=1$ 时蜕化为指数分布。n 个相互独立、同分布（参数 λ）的指数分布之和是 Gamma 分布（$\alpha=n,\beta=\lambda$）。Gamma 分布可用于服务时间，零件寿命等。

Gamma 分布又称爱尔朗分布。

（v）Weibull 分布

Weibull 分布是双参数 α,β 的非对称分布，记作 $W(\alpha,\beta)$。$\alpha=1$ 时蜕化为指数分布。作为设备、零件的寿命分布在可靠性分析中有着非常广泛的应用。

（vi）Beta 分布

Beta 分布是区间 $(0,1)$ 内的双参数、非均匀分布，记作 $B(\alpha,\beta)$。

2.2.2　常用的离散型概率分布

（i）离散均匀分布

（ii）Bernoulli 分布（两点分布）

Bernoulli 分布是 $x=1,0$ 处取值的概率分别是 p 和 $1-p$ 的两点分布，记作 $\text{Bern}(p)$。用于基本的离散模型。

（iii）泊松（Poisson）分布

泊松分布与指数分布有密切的关系。当顾客平均到达率为常数 λ 的到达间隔服从指数分布时，单位时间内到达的顾客数 K 服从泊松分布，即单位时间内到达 k 位顾客的概率为

$$P_k=\frac{\lambda^k e^{-\lambda}}{k!},\quad k=0,1,2,\cdots$$

记作 $\text{Poisson}(\lambda)$。泊松分布在排队服务、产品检验、生物与医学统计、天文、物理等领域都有广泛应用。

（iv）二项分布

在独立进行的每次试验中，某事件发生的概率为 p，则 n 次试验中该事件发生的次数 K 服从二项分布，即发生 k 次的概率为

$$P_k=C_n^k p^k(1-p)^{n-k},\quad k=0,1,\cdots,n.$$

记作 $B(n,p)$。二项分布是 n 个独立的 Bernoulli 分布之和。它在产品检验、保险、生物和医学统计等领域有着广泛的应用。

当 n,k 很大时，$B(n,p)$ 近似于正态分布 $N(np,np(1-p))$；当 n 很大、p 很小，且 np 约为常数 λ 时，$B(n,p)$ 近似于 $\text{Poisson}(\lambda)$。

§3　生灭过程

一类非常重要且广泛存在的排队系统是生灭过程排队系统。生灭过程是一类特殊的随机过程，在生物学、物理学、运筹学中有广泛的应用。在排队论中，如果 $N(t)$ 表示时刻 t 系统中的顾客数，则 $\{N(t),t\geq 0\}$ 就构成了一个随机过程。如果用“生”表示顾客的到达，“灭”表示顾客的离去，则对许多排队过程来说，$\{N(t),t\geq 0\}$ 就是一类特殊的随机过程－生灭过程。下面结合排队论的术语给出生灭过程的定义。

定义 1　设 $\{N(t),t\geq 0\}$ 为一个随机过程。若 $N(t)$ 的概率分布具有以下性质：

（1）假设 $N(t)=n$，则从时刻 t 起到下一个顾客到达时刻止的时间服从参数为 λ_n 的负指数分布，$n=0,1,2,\cdots$。

（2）假设 $N(t)=n$，则从时刻 t 起到下一个顾客离去时刻止的时间服从参数为 μ_n 的负指数分别，$n=0,1,2,\cdots$。

（3）同一时刻只有一个顾客到达或离去。

则称 $\{N(t),t\geq 0\}$ 为一个生灭过程。

一般来说，得到 $N(t)$ 的分布 $p_n(t)=P\{N(t)=n\}$（$n=0,1,2,\cdots$）是比较困难的，因此通常是求当系统到达平衡后的状态分布，记为 $p_n,n=0,1,2,\cdots$。

为求平稳分布，考虑系统可能处的任一状态 n。假设记录了一段时间内系统进入状态 n 和离开状态 n 的次数，则因为“进入”和“离开”是交替发生的，所以这两个数要

么相等，要么相差为 1。但就这两种事件的平均发生率来说，可以认为是相等的。即当系统运行相当时间而到达平衡状态后，对任一状态 n 来说，单位时间内进入该状态的平均次数和单位时间内离开该状态的平均次数应该相等，这就是系统在统计平衡下的“流入＝流出”原理。根据这一原理，可得到任一状态下的平衡方程如下：

$$\begin{array}{ll} 0 & \mu_1 p_1 = \lambda_0 p_0 \\ 1 & \lambda_0 p_0 + \mu_2 p_2 = (\lambda_1 + \mu_1) p_1 \\ 2 & \lambda_1 p_1 + \mu_3 p_3 = (\lambda_2 + \mu_2) p_2 \\ \vdots & \vdots \\ n & \lambda_{n-1} p_{n-1} + \mu_{n+1} p_{n+1} = (\lambda_n + \mu_n) p_n \\ \vdots & \vdots \end{array} \tag{3}$$

由上述平衡方程，可求得

$$0: \quad p_1 = \frac{\lambda_0}{\mu_1} p_0$$

$$1: \quad p_2 = \frac{\lambda_1}{\mu_2} p_1 + \frac{1}{\mu_2}(\mu_1 p_1 - \lambda_0 p_0) = \frac{\lambda_1}{\mu_2} p_1 = \frac{\lambda_1 \lambda_0}{\mu_2 \mu_1} p_0$$

$$2: \quad p_3 = \frac{\lambda_2}{\mu_3} p_2 + \frac{1}{\mu_3}(\mu_2 p_2 - \lambda_1 p_1) = \frac{\lambda_2}{\mu_3} p_2 = \frac{\lambda_2 \lambda_1 \lambda_0}{\mu_3 \mu_2 \mu_1} p_0$$

$$\vdots \qquad \vdots$$

$$n: \quad p_{n+1} = \frac{\lambda_n}{\mu_{n+1}} p_n + \frac{1}{\mu_{n+1}}(\mu_n p_n - \lambda_{n-1} p_{n-1}) = \frac{\lambda_n}{\mu_{n+1}} p_n = \frac{\lambda_n \lambda_{n-1} \cdots \lambda_0}{\mu_{n+1} \mu_n \cdots \mu_1} p_0$$

$$\vdots \qquad \vdots$$

记

$$C_n = \frac{\lambda_{n-1} \lambda_{n-2} \cdots \lambda_0}{\mu_n \mu_{n-1} \cdots \mu_1}, \quad n = 1,2,\cdots \tag{4}$$

则平稳状态的分布为

$$p_n = C_n p_0, \quad n = 1,2,\cdots \tag{5}$$

由概率分布的要求

$$\sum_{n=0}^{\infty} p_n = 1$$

有

$$\left[1 + \sum_{n=1}^{\infty} C_n\right] p_0 = 1$$

于是

$$p_0 = \frac{1}{1 + \sum_{n=1}^{\infty} C_n} \tag{6}$$

注意：（6）只有当级数 $\sum_{n=1}^{\infty} C_n$ 收敛时才有意义，即当 $\sum_{n=1}^{\infty} C_n < \infty$ 时，才能由上

述公式得到平稳状态的概率分布。

§4 $M/M/s$ 等待制排队模型

4.1 单服务台模型

单服务台等待制模型$M/M/1/\infty$是指：顾客的相继到达时间服从参数为λ的负指数分布，服务台个数为 1，服务时间V服从参数为μ的负指数分布，系统空间无限，允许无限排队，这是一类最简单的排队系统。

4.1.1 队长的分布

记$p_n = P\{N=n\}$（$n=0,1,2,\cdots$）为系统达到平衡状态后队长N的概率分布，则由式（4）～（6），并注意到$\lambda_n=\lambda, n=0,1,2,\cdots$和$\mu_n=\mu, n=0,1,2,\cdots$。记

$$\rho=\frac{\lambda}{\mu}$$

并设$\rho<1$（否则队列将排至无限远），则

$$C_n=\left(\frac{\lambda}{\mu}\right)^n,\quad n=1,2,\cdots$$

故

$$p_n=\rho^n p_0,\quad n=1,2,\cdots$$

其中

$$p_0=\frac{1}{1+\sum_{n=1}^{\infty}\rho^n}=\left(\sum_{n=0}^{\infty}\rho^n\right)^{-1}=\left(\frac{1}{1-\rho}\right)^{-1}=1-\rho \qquad (7)$$

因此

$$p_n=(1-\rho)\rho^n,\quad n=1,2,\cdots \qquad (8)$$

公式（7）和（8）给出了在平衡条件下系统中顾客数为n的概率。由式（7）不难看出，ρ是系统中至少有一个顾客的概率，也就是服务台处于忙的状态的概率，因而也称ρ为服务强度，它反映了系统繁忙的程度。此外，（8）式只有在$\rho=\frac{\lambda}{\mu}<1$的条件下才能得到，即要求顾客的平均到达率小于系统的平均服务率，才能使系统达到统计平衡。

4.1.2 几个主要数量指标

对单服务台等待制排队系统，由已得到的平稳状态下队长的分布，可以得到平均队长

$$\begin{aligned}L_s&=\sum_{n=0}^{\infty}np_n=\sum_{n=1}^{\infty}n(1-\rho)\rho^n\\&=(\rho+2\rho^2+3\rho^3+\cdots)-(\rho^2+2\rho^3+3\rho^4+\cdots) \qquad (9)\\&=\rho+\rho^2+\rho^3+\cdots=\frac{\rho}{1-\rho}=\frac{\lambda}{\mu-\lambda}\end{aligned}$$

平均排队长L_q为

$$L_q=\sum_{n=1}^{\infty}(n-1)p_n=L-(1-p_0)=L-\rho=\frac{\lambda^2}{\mu(\mu-\lambda)} \tag{10}$$

关于顾客在系统中的逗留时间T，可说明它服从参数为$\mu-\lambda$的复指数分布，即

$$P\{T>t\}=e^{-(\mu-\lambda)t},\quad t\geq 0$$

因此，平均逗留时间

$$W_s=\frac{1}{\mu-\lambda} \tag{11}$$

因为，顾客在系统中的逗留时间为等待时间T_q和接受服务时间V之和，即

$$T=T_q+V$$

故由

$$W_s=E(T)=E(T_q)+E(V)=W_q+\frac{1}{\mu} \tag{12}$$

可得平均等待时间W_q为

$$W_q=W_s-\frac{1}{\mu}=\frac{\lambda}{\mu(\mu-\lambda)} \tag{13}$$

从式（9）和式（11），可发现平均队长L_s与平均逗留时间W_s具有关系

$$L_s=\lambda W_s \tag{14}$$

同样，从式（10）和式（13），可发现平均排队长L_q与平均等待时间W_q具有关系

$$L_q=\lambda W_q \tag{15}$$

式（14）和式（15）通常称为Little公式，是排队论中一个非常重要的公式。

4.1.3　忙期和闲期

在平衡状态下，忙期B和闲期I一般均为随机变量，求它们的分布是比较麻烦的。因此，我们来求一下平均忙期$\overline{B}$和平均闲期$\overline{I}$。由于忙期和闲期出现的概率分别为ρ和$1-\rho$，所以在一段时间内可以认为忙期和闲期的总长度之比为$\rho:(1-\rho)$。又因为忙期和闲期是交替出现的，所以在充分长的时间里，它们出现的平均次数应是相同的。于是，忙期的平均长度$\overline{B}$和闲期的平均长度$\overline{I}$之比也应是$\rho:(1-\rho)$，即

$$\frac{\overline{B}}{\overline{I}}=\frac{\rho}{1-\rho} \tag{16}$$

又因为在到达为Poisson流时，根据负指数分布的无记忆性和到达与服务相互独立的假设，容易证明从系统空闲时刻起到下一个顾客到达时刻止（即闲期）的时间间隔仍服从参数为λ的负指数分布，且与到达时间间隔相互独立。因此，平均闲期应为$\frac{1}{\lambda}$，这样，便求得平均忙期为

$$\overline{B}=\frac{\rho}{1-\rho}\cdot\frac{1}{\lambda}=\frac{1}{\mu-\lambda} \tag{17}$$

与式（11）比较，发现平均逗留时间（W_s）＝平均忙期（$\overline{B}$）。这一结果直观看上去是显然的，顾客在系统中逗留的时间越长，服务员连续繁忙的时间也就越长。因此，一

个顾客在系统内的平均逗留时间应等于服务员平均连续忙的时间。

4.2　与排队论模型有关的 LINGO 函数

（1）@peb(load,S)

该函数的返回值是当到达负荷为 load，服务系统中有 S 个服务台且允许排队时系统繁忙的概率，也就是顾客等待的概率。

（2）@pel（load,S）

该函数的返回值是当到达负荷为 load，服务系统中有 S 个服务台且不允许排队时系统损失概率，也就是顾客得不到服务离开的概率。

（3）@pfs(load,S,K)

该函数的返回值是当到达负荷为 load，顾客数为 K，平行服务台数量为 S 时，有限源的 Poisson 服务系统等待或返修顾客数的期望值。

例 1　某修理店只有一个修理工，来修理的顾客到达过程为 Poisson 流，平均 4 人/h；修理时间服从负指数分布，平均需要 6min。试求：（1）修理店空闲的概率；（2）店内恰有 3 个顾客的概率；（3）店内至少有 1 个顾客的概率；（4）在店内的平均顾客数；（5）每位顾客在店内的平均逗留时间；（6）等待服务的平均顾客数；（7）每位顾客平均等待服务时间；（8）顾客在店内等待时间超过 10min 的概率。

解　本例可看成一个 $M/M/1/\infty$ 排队问题，其中

$$\lambda=4,\ \mu=\frac{1}{0.1}=10,\ \rho=\frac{\lambda}{\mu}=0.4$$

（1）修理店空闲的概率

$$p_0=1-\rho=1-0.4=0.6$$

（2）店内恰有 3 个顾客的概率

$$p_3=\rho^3(1-\rho)=0.4^3\times(1-0.4)=0.38$$

（3）店内至少有 1 个顾客的概率

$$P\{N\geq 1\}=1-p_0=\rho=0.4$$

（4）在店内的平均顾客数

$$L_s=\frac{\rho}{1-\rho}=0.67\ \text{（人）}$$

（5）每位顾客在店内的平均逗留时间

$$W_s=\frac{L_s}{\lambda}=\frac{0.67}{4}(\text{h})=10(\text{min})$$

（6）等待服务的平均顾客数

$$L_q=L_s-\rho=\frac{\rho^2}{1-\rho}=\frac{0.4^2}{1-0.4}=0.267\ \text{（人）}$$

（7）每位顾客平均等待服务时间

$$W_q=\frac{L_q}{\lambda}=\frac{0.267}{4}(\text{h})=4(\text{min})$$

（8）顾客在店内逗留时间超过 10min 的概率

$$P\{T>10\}=e^{-10(\frac{1}{6}-\frac{1}{15})}=e^{-1}=0.3679$$

编写 LINGO 程序如下：

```
model:
```

```
s=1;lamda=4;mu=10;rho=lamda/mu;
Pwait=@peb(rho,s);
p0=1-Pwait;
Pt_gt_10=@exp(-1);
end
```

4.3　多服务台模型（$M/M/s/\infty$）

设顾客单个到达，相继到达时间间隔服从参数为λ的负指数分布，系统中共有s个服务台，每个服务台的服务时间相互独立，且服从参数为μ的负指数分布。当顾客到达时，若有空闲的服务台则马上接受服务，否则便排成一个队列等待，等待时间为无限。

下面来讨论这个排队系统的平稳分布。记$p_n = P\{N=n\}$（$n=0,1,2,\cdots$）为系统达到平稳状态后队长N的概率分布，注意到对个数为s的多服务台系统，有

$\lambda_n = \lambda$，　$n=0,1,2,\cdots$

和

$$\mu_n = \begin{cases} n\mu, & n=1,2,\cdots,s \\ s\mu, & n=s,s+1,\cdots \end{cases}$$

记$\rho_s = \dfrac{\rho}{s} = \dfrac{\lambda}{s\mu}$，则当$\rho_s<1$时，由式（4），式（5）和式（6），有

$$C_n = \begin{cases} \dfrac{(\lambda/\mu)^n}{n!}, & n=1,2,\cdots,s \\ \dfrac{(\lambda/\mu)^s}{s!}\left(\dfrac{\lambda}{s\mu}\right)^n = \dfrac{(\lambda/\mu)^n}{s!s^{n-s}}, & n\geq s \end{cases} \tag{18}$$

故

$$p_n = \begin{cases} \dfrac{\rho^n}{n!}p_0, & n=1,2,\cdots,s \\ \dfrac{\rho^n}{s!s^{n-s}}p_0, & n\geq s \end{cases} \tag{19}$$

其中

$$p_0 = \left[\sum_{n=0}^{s-1}\frac{\rho^n}{n!} + \frac{\rho^s}{s!(1-\rho_s)}\right]^{-1} \tag{20}$$

公式（19）和式（20）给出了在平衡条件下系统中顾客数为n的概率，当$n\geq s$时，即系统中顾客数大于或等于服务台个数，这时再来的顾客必须等待，因此记

$$c(s,\rho) = \sum_{n=s}^{\infty}p_n = \frac{\rho^s}{s!(1-\rho_s)}p_0 \tag{21}$$

式（21）称为Erlang等待公式，它给出了顾客到达系统时需要等待的概率。

对多服务台等待制排队系统，由已得到的平稳分布可得平均排队长L_q为：

$$L_q = \sum_{n=s+1}^{\infty}(n-s)p_n = \frac{p_0\rho^s}{s!}\sum_{n=s}^{\infty}(n-s)\rho_s^{n-s}$$

$$= \frac{p_0\rho^s}{s!}\frac{d}{d\rho_s}\left(\sum_{n=1}^{\infty}\rho_s^n\right) = \frac{p_0\rho^s\rho_s}{s!(1-\rho_s)^2} \tag{22}$$

或

$$L_q = \frac{c(s,\rho)\rho_s}{1-\rho_s} \tag{23}$$

记系统中正在接受服务的顾客的平均数为 $\bar{s}$，显然 $\bar{s}$ 也是正在忙的服务台的平均数，故

$$\bar{s} = \sum_{n=0}^{s-1} np_n + s\sum_{n=s}^{\infty} p_n = \sum_{n=0}^{s-1}\frac{n\rho^n}{n!}p_0 + s\frac{\rho^s}{s!(1-\rho_s)}p_0$$

$$= p_0\rho\left[\sum_{n=1}^{s-1}\frac{\rho^{n-1}}{(n-1)!} + \frac{\rho^{s-1}}{(s-1)!(1-\rho_s)}\right] = \rho \tag{24}$$

式（24）说明，平均在忙的服务台个数不依赖于服务台个数 s，这是一个有趣的结果。由式（24），可得到平均队长 L_s 为

$$L_s = \text{平均排队长} + \text{正在接受服务的顾客的平均数} = L_q + \rho \tag{25}$$

对多服务台系统，Little 公式依然成立，即有

$$W_s = \frac{L_s}{\lambda},\quad W_q = \frac{L_q}{\lambda} = W_s - \frac{1}{\mu} \tag{26}$$

例 2　某售票处有 3 个窗口，顾客的到达为 Poisson 流，平均到达率为 $\lambda = 0.9\text{人}/\min$；服务（售票）时间服从负指数分布，平均服务率 $\mu = 0.4\text{人}/\min$。现设顾客到达后排成一个队列，依次向空闲的窗口购票，这一排队系统可看成是一个 $M/M/s/\infty$ 系统，其中

$$s = 3,\quad \rho = \frac{\lambda}{\mu} = 2.25,\quad \rho_s = \frac{\lambda}{s\mu} = \frac{2.25}{3} < 1$$

由多服务台等待制系统的有关公式，可得到

（1）整个售票处空闲的概率

$$p_0 = \left[\frac{(2.25)^0}{0!} + \frac{(2.25)^1}{1!} + \frac{(2.25)^2}{2!} + \frac{(2.25)^3}{3!(1-2.25/3)}\right]^{-1} = 0.0748$$

（2）平均排队长

$$L_q = \frac{0.0748\times(2.25)^3\times 2.25/3}{3!(1-2.25/3)^2} = 1.70\ \text{（人）}$$

平均队长

$$L = L_q + \rho = 1.70 + 2.25 = 3.95\ \text{（人）}$$

（3）平均等待时间

$$W_q = \frac{L_q}{\lambda} = \frac{1.70}{0.9} = 1.89\ \text{（min）}$$

平均逗留时间

$$W_s = \frac{L_s}{\lambda} = \frac{3.95}{0.9} = 4.39\ \text{（min）}$$

（4）顾客到达时必须排队等待的概率

$$c(3,2.25)=\frac{(2.25)^3}{3!(1-2.25/3)}\times 0.0748=0.57$$

在本例中，如果顾客的排队方式变为到达售票处后可到任一窗口前排队，且入队后不再换队，即可形成 3 个队列。这时，原来的$M/M/3/\infty$系统实际上变成了由 3 个$M/M/1/\infty$子系统组成的排队系统，且每个系统的平均到达率为

$$\lambda_1=\lambda_2=\lambda_3=\frac{0.9}{3}=0.3\text{（人/min）}$$

下表给出了$M/M/3/\infty$和 3 个$M/M/1/\infty$的比较，不难看出一个$M/M/3/\infty$系统比由 3 个$M/M/1/\infty$系统组成的排队系统具有显著的优越性。即在服务台个数和服务率都不变的条件下，单队排队方式比多队排队方式要优越，这是在对排队系统进行设计和管理的时候应注意的地方。

表 1　排队系统的指标值

项　　目	$M/M/3/\infty$	3 个$M/M/1/\infty$
空闲的概率	0.0748	0.25（每个子系统）
顾客必须等待的概率	0.57	0．75
平均队长	3.95	9（整个系统）
平均排队长	1.70	2.25（每个子系统）
平均逗留时间	4.39（min）	10（min）
平均等待时间	1.89	7.5（min）

求解的 LINGO 程序如下：

```
model:
s=3;lamda=0.9;mu=0.4;rho=lamda/mu;rho_s=rho/s;
P_wait=@peb(rho,s);
p0=6*(1-rho_s)/rho^3*P_wait;
L_q=P_wait*rho_s/(1-rho_s);
L_s=L_q+rho;
W_q=L_q/lamda;
W_s=L_s/lamda;
end
```

§5　$M/M/s/s$ 损失制排队模型

当s个服务台被占用后，顾客自动离去。

这里我们着重介绍如何使用 LINGO 软件中的相关函数。

5.1　损失制排队模型的基本参数

对于损失制排队模型，其模型的基本参数与等待制排队模型有些不同，我们关心如下指标。

（1）系统损失的概率

P_{lost}=@pel(rho,s)

其中 rho 是系统到达负荷$\frac{\lambda}{\mu}$，s 是服务台或服务员的个数。

（2）单位时间内平均进入系统的顾客数（λ_e）

$$\lambda_e=\lambda(1-P_{\text{lost}})$$

（3）系统的相对通过能力（Q）与绝对通过能力（A）

$$Q=1-P_{\text{lost}}$$

$$A=\lambda_e Q=\lambda(1-P_{\text{lost}})^2$$

（4）系统在单位时间内占用服务台（或服务员）的均值（即L_s）

$$L_s=\lambda_e/\mu$$

注意：在损失制排队系统中，$L_q=0$，即等待队长为0。

（5）系统服务台（或服务员）的效率

$$\eta=L_s/s$$

（6）顾客在系统内平均逗留时间（即W_s）

$$W_s=1/\mu$$

注意：在损失制排队系统中，$W_q=0$，即等待时间为0。

在上述公式中，引入λ_e是十分重要的，因为尽管顾客以平均λ的速率到达服务系统，但当系统被占满后，有一部分顾客会自动离去，因此，真正进入系统的顾客输入率是λ_e，它小于λ。

5.2　损失制排队模型计算实例

5.2.1　$s=1$的情况（$M/M/1/1$）

例3　设某条电话线，平均每分钟有0.6次呼唤，若每次通话时间平均为1.25min，求系统相应的参数指标。

解　其参数为S=1，$\lambda=0.6$，$\mu=\dfrac{1}{1.25}$。编写LINGO程序如下：

```
model:
s=1;lamda=0.6;mu=1/1.25;rho=lamda/mu;
Plost=@pel(rho,s);
Q=1-Plost;
lamda_e=Q*lamda;A=Q*lamda_e;
L_s=lamda_e/mu;
eta=L_s/s;
end
```

求得系统的顾客损失率为43%，即43%的电话没有接通，有57%的电话得到了服务，通话率为平均每分钟有0.195次，系统的服务效率为43%。对于一个服务台的损失制系统，系统的服务效率等于系统的顾客损失率，这一点在理论上也是正确的。

5.2.2　$s>1$的情况（$M/M/s/s$）

例4　某单位电话交换台有一台200门内线的总机，已知在上班8h的时间内，有20%的内线分机平均每40min要一次外线电话，80%的分机平均隔120min要一次外线。又知外线打入内线的电话平均每分钟1次。假设与外线通话的时间平均为3min，并且上述时间均服从负指数分布，如果要求电话的通话率为95%，问该交换台应设置多少条外线？

解　（1）电话交换台的服务分成两类，第一类内线打外线，其强度为

$$\lambda_1=\left(\frac{60}{40}\times 0.2+\frac{60}{120}\times 0.8\right)\times 200=140$$

第二类是外线打内线，其强度为

$$\lambda_2=1\times 60=60$$

因此，总强度为

$$\lambda = \lambda_1 + \lambda_2 = 140 + 60 = 200$$

（2）这是损失制服务系统，按题目要求，系统损失的概率不能超过5%，即

$$P_{\text{lost}} \le 0.05$$

（3）外线是整数，在满足条件下，条数越小越好。

由上述三条，写出相应的LINGO程序如下：

```
model:
lamda=200;
mu=60/3;rho=lamda/mu;
Plost=@pel(rho,s);Plost<0.05;
Q=1-Plost;
lamda_e=Q*lamda;A=Q*lamda_e;
L_s=lamda_e/mu;
eta=L_s/s;
min=s;@gin(s);
end
```

求得需要15条外线。在此条件下，交换台的顾客损失率为3.65%，有96.35%的电话得到了服务，通话率为平均每小时185.67次，交换台每条外线的服务效率为64.23%。

求解时，尽量选用简单的模型让LINGO软件求解，而上述程序是解非线性整数规划（尽管是一维的），但计算时间可能会较长，因此，我们选用下面的处理方法，分两步处理。

第一步，求出概率为5%的服务台的个数，尽管要求服务台的个数是整数，但@pel给出的是实数解。

编写LINGO程序：

```
model:
lamda=200;
mu=60/3;rho=lamda/mu;
@pel(rho,s)=0.05;
end
```

求得 $s = 14.33555$。

第二步，注意到@pel(rho,s)是s的单调递减函数，因此，对s取整数（采用只入不舍原则）就是满足条件的最小服务台数，然后再计算出其它的参数指标。

编写LINGO程序如下：

```
model:
lamda=200;
mu=60/3;rho=lamda/mu;
s=15;Plost=@pel(rho,s);
Q=1-Plost;
lamda_e=Q*lamda;A=Q*lamda_e;
L_s=lamda_e/mu;
eta=L_s/s;
end
```

比较上面两种方法的计算结果，其答案是相同的，但第二种方法比第一种方法在计算时间上要少许多。

§6 $M/M/s$ 混合制排队模型

6.1 单服务台混合制模型

单服务台混合制模型 $M/M/1/K$ 是指：顾客的相继到达时间服从参数为 λ 的负指数分布，服务台个数为1，服务时间 V 服从参数为 μ 的负指数分布，系统的空间为 K，当 K 个位置已被顾客占用时，新到的顾客自动离去，当系统中有空位置时，新到的顾客进入系

统排队等待。

首先，仍来求平稳状态下队长 N 的分布 $p_n = P\{N = n\}$，$n = 0,1,2,\cdots$。由于所考虑的排队系统中最多只能容纳 K 个顾客（等待位置只有 $K-1$ 个），因而有

$$\lambda_n = \begin{cases} \lambda, & n = 0,1,2,\cdots,K-1 \\ 0, & n \geq K \end{cases}$$

$$\mu_n = \mu，\ n = 1,2,\cdots,K$$

由式（4），式（5）和式（6），有

$$C_n = \begin{cases} \left(\frac{\lambda}{\mu}\right)^n = \rho^n, & n = 1,2,\cdots,K \\ 0, & n > K \end{cases} \tag{27}$$

故

$$p_n = \rho^n p_0，\ n = 1,2,\cdots,K$$

其中

$$p_0 = \frac{1}{1+\sum_{n=1}^{K}\rho^n} = \begin{cases} \frac{1-\rho}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{1}{K+1}, & \rho = 1 \end{cases} \tag{28}$$

由已得到的单服务台混合制排队系统平稳状态下队长的分布，可知当 $\rho \neq 1$ 时，平均队长 L_s 为：

$$L_s = \sum_{n=0}^{K} n p_n = p_0 \rho \sum_{n=1}^{K} n\rho^{n-1}$$

$$= \frac{p_0\rho}{(1-\rho)^2}[1-\rho^K-(1-\rho)K\rho^K] = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \tag{29}$$

当 $\rho = 1$ 时，

$$L_s = \sum_{n=0}^{K} n p_n = \sum_{n=1}^{K} n\rho^n p_0 = \frac{1}{K+1}\sum_{n=1}^{K} n = \frac{K}{2} \tag{30}$$

类似地可得到平均排队长 L_q 为

$$L_q = \sum_{n=1}^{K}(n-1)p_n = L_s - (1-p_0) \tag{31}$$

或

$$L_q = \begin{cases} \frac{\rho}{1-\rho} - \frac{\rho(1+K\rho^K)}{1-\rho^{K+1}}, & \rho \neq 1 \\ \frac{K(K-1)}{2(K+1)}, & \rho = 1 \end{cases} \tag{32}$$

由于排队系统的容量有限，只有 $K-1$ 个排队位置，因此，当系统空间被占满时，再来的顾客将不能进入系统排队，也就是说不能保证所有到达的顾客都能进入系统等待

服务。假设顾客的到达率（单位时间内来到系统的顾客的平均数）为λ，则当系统处于状态K时，顾客不能进入系统，即顾客可进入系统的概率是$1-p_K$。因此，单位时间内实际可进入系统的顾客的平均数为：

$$\lambda_e=\lambda(1-p_K)=\mu(1-p_0) \tag{33}$$

称λ_e为有效到达率，而p_K也被称为顾客损失率，它表示了在来到系统的所有顾客中不能进入系统的顾客的比例。下面根据 Little 公式，可得

平均逗留时间

$$W_s=\frac{L_s}{\lambda_e}=\frac{L_s}{\lambda(1-p_k)} \tag{34}$$

平均等待时间

$$W_q=\frac{L_q}{\lambda_e}=\frac{L_q}{\lambda(1-p_K)} \tag{35}$$

且仍有

$$W_s=W_q+\frac{1}{\mu} \tag{36}$$

注意：这里的平均逗留时间和平均等待时间都是针对能够进入系统的顾客而言的。

特别，当$K=1$时，$M/M/1/1$为单服务台损失系统，在上述有关结果中令$K=1$，可得到：

$$p_0=\frac{1}{1+\rho},\quad p_1=\frac{\rho}{1+\rho} \tag{37}$$

$$L_s=p_1=\frac{\rho}{1+\rho}, \tag{38}$$

$$\lambda_e=\lambda(1-p_1)=\lambda p_0=\frac{\lambda}{1+\rho} \tag{39}$$

$$W_s=\frac{L_s}{\lambda_e}=\frac{\rho}{\lambda}=\frac{1}{\mu} \tag{40}$$

$$L_q=0,\ W_q=0 \tag{41}$$

例 5　某修理站只有一个修理工，且站内最多只能停放 4 台待修的机器。设待修机器按 Poisson 流到达修理站，平均每分钟到达 1 台；修理时间服从负指数分布，平均每 1.25 分钟可修理 1 台，试求该系统的有关指标。

解　该系统可看成是一个$M/M/1/4$排队系统，其中

$$\lambda=1,\ \mu=\frac{1}{1.25}=0.8,\ \rho=\frac{\lambda}{\mu}=1.25,\ K=4$$

由式（28），

$$p_0=\frac{1-\rho}{1-\rho^5}=\frac{1-1.25}{1-1.25^5}=0.122$$

因而，顾客损失率为：

$$p_4=\rho^4 p_0=1.25^4\times 0.122=0.298$$

有效到达率为：

$$\lambda_e = \lambda(1-p_4) = 1\times(1-0.298) = 0.702$$

平均队长

$$L_s = \frac{1.25}{1-1.25} - \frac{(4+1)\times 1.25^5}{1-1.25^5} = 2.44 \text{（台）}$$

平均排队长

$$L_q = L_s - (1-p_0) = 2.44 - (1-0.122) = 1.56 \text{（台）}$$

平均逗留时间

$$W_s = \frac{L_s}{\lambda_e} = \frac{2.44}{0.702} = 3.48 \text{（分钟）}$$

平均等待时间

$$W_q = W_s - \frac{1}{\mu} = 3.48 - \frac{1}{0.8} = 2.23 \text{（分钟）}$$

编写 LINGO 程序如下：

```
model:
sets:
state/1..4/:p;
endsets
lamda=1;mu=1/1.25;rho=lamda/mu;k=4;
lamda*p0=mu*p(1);
(lamda+mu)*p(1)=lamda*p0+mu*p(2);
@for(state(i)|i #gt#1 #and# i #lt#
k:(lamda+mu)*p(i)=lamda*p(i-1)+mu*p(i+1));
lamda*p(k-1)=mu*p(k);
p0+@sum(state:p)=1;
P_lost=p(k);lamda_e=lamda*(1-P_lost);
L_s=@sum(state(i)|i #le#k:i*p(i));
L_q=L_s-(1-p0);
W_s=L_s/lamda_e;
W_q=W_s-1/mu;
end
```

6.2　多服务台混合制模型

多服务台混合制模型 $M/M/s/K$ 是指顾客的相继到达时间服从参数为 λ 的负指数分布，服务台个数为 s，每个服务台服务时间相互独立，且服从参数为 μ 的负指数分布，系统的空间为 K。

由式（4），式（5）和式（6），并注意到在本模型中

$$\lambda_n = \begin{cases} \lambda, & n = 0,1,2,\cdots,K-1 \\ 0, & n \geq K \end{cases}$$

$$\mu_n = \begin{cases} n\mu, & 0 \leq n < s \\ s\mu, & s \leq n \leq K \end{cases}$$

于是

$$p_n=\begin{cases}\dfrac{\rho^n}{n!}p_0, & 0\le n<s\\ \dfrac{\rho^n}{s!s^{n-s}}p_0, & s\le n\le K\end{cases} \tag{42}$$

其中

$$p_0=\begin{cases}\left(\sum\limits_{n=0}^{s-1}\dfrac{\rho^n}{n!}+\dfrac{\rho^s(1-\rho_s^{K-s+1})}{s!(1-\rho_s)}\right)^{-1}, & \rho_s\ne 1\\ \left(\sum\limits_{n=0}^{s-1}\dfrac{\rho^n}{n!}+\dfrac{\rho^s}{s!}(K-s+1)\right)^{-1}, & \rho_s=1\end{cases} \tag{43}$$

由平稳分布 $p_n, n=0,1,2,\cdots,K$，可得平均排队长为

$$\begin{aligned}L_q&=\sum_{n=s}^{K}(n-s)p_n\\&=\begin{cases}\dfrac{p_0\rho^s\rho_s}{s!(1-\rho_s)^2}[1-\rho_s^{K-s+1}-(1-\rho_s)(K-s+1)\rho_s^{K-s}], & \rho_s\ne 1\\ \dfrac{p_0\rho^s(K-s)(K-s+1)}{2s!}, & \rho_s=1\end{cases}\end{aligned} \tag{44}$$

为求平均队长，由

$$\begin{aligned}L_q&=\sum_{n=s}^{K}(n-s)p_n=\sum_{n=s}^{K}np_n-s\sum_{n=s}^{K}p_n\\&=\sum_{n=0}^{K}np_n-\sum_{n=0}^{s-1}np_n-s\left(1-\sum_{n=0}^{s-1}p_n\right)=L_s-\sum_{n=0}^{s-1}(n-s)p_n-s\end{aligned}$$

得到

$$L_s=L_q+s+p_0\sum_{n=0}^{s-1}\frac{(n-s)\rho^n}{n!} \tag{45}$$

由系统空间的有限性，必须考虑顾客的有效到达率 λ_e。对多服务台系统，仍有

$$\lambda_e=\lambda(1-p_K) \tag{46}$$

再利用 Little 公式，得到

$$W_s=\frac{L_s}{\lambda_e},\quad W_q=\frac{L_q}{\lambda_e}=W_s-\frac{1}{\mu} \tag{47}$$

平均被占用的服务台数（也是正在接受服务的顾客的平均数）为

$$\begin{aligned}\bar{s}&=\sum_{n=0}^{s-1}np_n+s\sum_{n=s}^{K}p_n=p_0\left[\sum_{n=0}^{s-1}\frac{n\rho^n}{n!}+s\sum_{n=s}^{K}\frac{\rho^n}{s!s^{n-s}}\right]\\&=p_0\rho\left[\sum_{n=1}^{s-1}\frac{\rho^{n-1}}{(n-1)!}+\sum_{n=s}^{K}\frac{\rho^{n-1}}{s!s^{n-1-s}}\right]=p_0\rho\left[\sum_{n=0}^{s-1}\frac{\rho^n}{n!}+\sum_{n=s}^{K}\frac{\rho^n}{s!s^{n-s}}-\frac{\rho^K}{s!s^{K-s}}\right]\end{aligned}$$

$$= \rho(1-\frac{\rho^K}{s!s^{K-s}}p_0) = \rho(1-p_K) \quad (48)$$

因此，又有

$$L_s = L_q + \bar{s} = L_q + \rho(1-p_K) \quad (49)$$

例 6　某汽车加油站设有两个加油机，汽车按 Poisson 流到达，平均每分钟到达 2 辆；汽车加油时间服从负指数分布，平均加油时间为 2 分钟。又知加油站上最多只能停放 3 辆等待加油的汽车，汽车到达时，若已满员，则必须开到别的加油站去，试对该系统进行分析。

解　可将该系统看作一个 $M/M/2/5$ 排队系统，其中

$$\lambda = 2,\ \mu = 0.5,\ \rho = \frac{\lambda}{\mu} = 4,\ s = 2,\ K = 5$$

（1）系统空闲的概率

$$p_0 = \left\{1+4+\frac{4^2[1-(4/2)^{5-2+1}]}{2!(1-4/2)}\right\}^{-1} = 0.008$$

（2）顾客损失率

$$p_5 = \frac{4^5 \times 0.008}{2!\times 2^{5-2}} = 0.512$$

（3）加油站内在等待的平均汽车数

$$L_q = \frac{0.008\times 4^2\times(4/2)}{2!(1-4/2)^2}[1-(4/2)^{5-2+1}-(1-4/2)(5-2+1)(4/2)^{5-2}]$$

$$= 2.18 \text{（辆）}$$

加油站内汽车的平均数为

$$L_s = L_q + \rho(1-p_5) = 2.18 + 4(1-0.512) = 4.13 \text{（辆）}$$

（4）汽车在加油站内平均逗留时间为

$$W_s = \frac{L_s}{\lambda(1-p_5)} = \frac{4.13}{2(1-0.512)} = 4.23 \text{（分钟）}$$

汽车在加油站内平均等待时间为

$$W_q = W_s - \frac{1}{\mu} = 4.23 - 2 = 2.23 \text{（分钟）}$$

（5）被占用的加油机的平均数为

$$\bar{s} = L_s - L_q = 4.13 - 2.18 = 1.95 \text{（个）}$$

编写 LINGO 程序如下：

```
model:
sets:
state/1..5/:p;
endsets
lamda=2;mu=0.5;rho=lamda/mu;s=2;k=5;
lamda*p0=mu*p(1);
(lamda+mu)*p(1)=lamda*p0+2*mu*p(2);
@for(state(i)|i #gt#1 #and# i #lt# s:
(lamda+i*mu)*p(i)=lamda*p(i-1)+(i+1)*mu*p(i+1));
```

```
@for(state(i)|i #ge# s #and# i #lt# k:
(lamda+s*mu)*p(i)=lamda*p(i-1)+s*mu*p(i+1));
lamda*p(k-1)=s*mu*p(k);
p0+@sum(state:p)=1;
P_lost=p(k);lamda_e=lamda*(1-P_lost);
L_s=@sum(state(i):i*p(i));
L_q=L_s-lamda_e/mu;
W_s=L_s/lamda_e;
W_q=W_s-1/mu;
end
```

在对上述多服务台混合制排队模型 $M/M/s/K$ 的讨论中，当 $s=K$ 时，即为多服务台损失制系统。对损失制系统，有

$$p_n=\frac{\rho^n}{n!}p_0,\quad n=1,2,\cdots,s \tag{50}$$

其中

$$p_0=\left(\sum_{n=0}^{s}\frac{\rho^n}{n!}\right)^{-1} \tag{51}$$

顾客的损失率为

$$B(s,\rho)=p_s=\frac{\rho^s}{s!}\left(\sum_{n=0}^{s}\frac{\rho^n}{n!}\right)^{-1} \tag{52}$$

式（52）称为 Erlang 损失公式，$B(s,\rho)$ 亦表示了到达系统后由于系统空间已被占满而不能进入系统的顾客的百分比。

对损失制系统，平均被占用的服务台数（正在接受服务的顾客的平均数）为

$$\begin{aligned}\bar{s}&=\sum_{n=0}^{s}np_n=\sum_{n=0}^{s}\frac{n\rho^n}{n!}p_0\\&=\rho\left(\sum_{n=0}^{s}\frac{\rho^n}{n!}-\frac{\rho^s}{s!}\right)\left(\sum_{n=0}^{s}\frac{\rho^n}{n!}\right)^{-1}=\rho(1-B(s,\rho))\end{aligned} \tag{53}$$

此外，还有

平均队长 $L_s=\bar{s}=\rho(1-B(s,\rho))$ （54）

平均逗留时间 $W_s=\frac{L_s}{\lambda_e}=\frac{\rho[1-B(s,\rho)]}{\lambda[1-B(s,\rho)]}=\frac{1}{\mu}$ （55）

其中 $\lambda_e=\lambda(1-p_s)$ 为有效到达率。在损失制系统中，还经常用 $A=\lambda(1-p_s)$ 表示系统的绝对通过能力，即单位时间内系统实际可完成的服务次数；用 $Q=1-p_s$ 表示系统的相对通过能力，即被服务的顾客数与请求服务的顾客数的比值。系统的服务台利用率（或通道利用率）为

$$\eta=\frac{\bar{s}}{s} \tag{56}$$

§7 其它排队模型简介

7.1 有限源排队模型

现在，来分析一下顾客源为有限的排队问题。这类排队问题的主要特征是顾客总数是有限的，如果有 m 个顾客。每个顾客来到系统中接受服务后仍回到原来的总体，还有可能再来，这类排队问题的典型例子是机器看管问题。如一个工人同时看管 m 台机器，当机器发生故障时即停下来等待维修，修好后再投入使用，且仍然可能再发生故障。类似的例子还有 m 个终端共用一台打印机等，如图 2 所示。

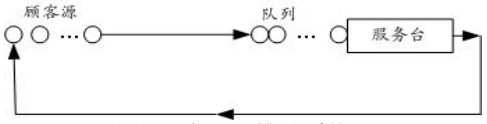

图 2　有限源排队系统

关于顾客的平均到达率，在无限源的情形中是按全体顾客来考虑的，而在有限源的情形下，必须按每一顾客来考虑。设每个顾客的到达率都是相同的，均为 λ（这里 λ 的含义是指单位时间内该顾客来到系统请求服务的次数），且每一顾客在系统外的时间均服从参数为 λ 的负指数分布。由于在系统外的顾客的平均数为 $m-L_s$，故系统的有效到达率为

$$\lambda_e=\lambda(m-L_s)$$

下面讨论平稳状态下队长 N 的分布 $p_n=P\{N=n\}$，$n=0,1,2,\cdots,m$。由于状态间的转移率为

$$\lambda_n=\lambda(m-n)，\quad n=0,1,2,\cdots,m$$

$$\mu_n=\begin{cases}n\mu, & n=1,2,\cdots,s\\ s\mu, & n=s+1,\cdots,m\end{cases}$$

由式（4），式（5）和式（6），有（记 $\rho=\dfrac{\lambda}{\mu}$）

$$C_n=\begin{cases}\dfrac{m!}{(m-n)!n!}\rho^n, & n=1,2,\cdots,s\\ \dfrac{m!}{(m-n)!s!s^{n-s}}\rho^n, & n=s,\cdots,m\end{cases}\tag{57}$$

故

$$p_n=\begin{cases}\dfrac{m!}{(m-n)!n!}\rho^n p_0, & n=1,2,\cdots,s\\ \dfrac{m!}{(m-n)!s!s^{n-s}}\rho^n p_0, & n=s,\cdots,m\end{cases}\tag{58}$$

其中

$$p_0=\left[\sum_{n=0}^{s-1}\frac{m!}{(m-n)!n!}\rho^n+\sum_{n=s}^{m}\frac{m!}{(m-n)!s!s^{n-s}}\rho^n\right]^{-1}\tag{59}$$

下面给出系统的有关运行指标

$$L_q = \sum_{n=s}^{m} (n-s) p_n \tag{60}$$

$$L_s = \sum_{n=0}^{s-1} n p_n + L_q + s(1 - \sum_{n=0}^{s-1} p_n) \tag{61}$$

或

$$L_s = L_q + \frac{\lambda_e}{\mu} = L_q + \rho(m - L_s) \tag{62}$$

$$W_s = \frac{L_s}{\lambda_e}, \quad W_q = \frac{L_q}{\lambda_e} \tag{63}$$

特别，对单服务台（$s=1$）系统，有

$$p_n = \frac{m!}{(m-n)!} \rho^n p_0, \quad n = 1,2,\cdots,m \tag{64}$$

$$p_0 = \left[\sum_{n=0}^{m} \frac{m!}{(m-n)!} \rho^n \right]^{-1} \tag{65}$$

$$L_q = \sum_{n=1}^{m} (n-1) p_n \tag{66}$$

$$L_s = L_q + (1 - p_0) \tag{67}$$

或

$$L_s = m - \frac{\mu}{\lambda}(1 - p_0) \tag{68}$$

$$W_s = \frac{L_s}{\lambda_e} = \frac{m}{\mu(1-p_0)} - \frac{1}{\lambda}, \quad W_q = W_s - \frac{1}{\mu} \tag{69}$$

系统的相对通过能力$Q=1$，绝对通过能力

$$A = \lambda_e Q = \lambda(m - L_s) = \mu(1 - p_0) \tag{70}$$

例 7　设有一工人看管 5 台机器，每台机器正常运转的时间服从负指数分布，平均为 15 分钟。当发生故障后，每次修理时间服从负指数分布，平均为 12 分钟，试求该系统的有关运行指标。

解　用有限源排队模型处理本问题。已知

$$\lambda = \frac{1}{15}, \quad \mu = \frac{1}{12}, \quad \rho = \frac{\lambda}{\mu} = 0.8, \quad m = 5$$

于是，有

（1）修理工人空闲的概率

$$p_0 = \left[\frac{5!}{5!}(0.8)^0 + \frac{5!}{4!}(0.8)^1 + \frac{5!}{3!}(0.8)^2 + \frac{5!}{2!}(0.8)^3 + \frac{5!}{1!}(0.8)^4 + \frac{5!}{0!}(0.8)^5 \right]^{-1} = 0.0073$$

（2）5 台机器都出故障的概率

$$p_5 = \frac{5!}{0!}(0.8)^5 p_0 = 0.287$$

（3）出故障机器的平均数

$$L_s = 5 - \frac{1}{0.8}(1 - 0.0073) = 3.76$$（台）

（4）等待修理机器的平均数

$$L_q = 3.76 - (1 - 0.0073) = 2.77$$（台）

（5）每台机器发生一次故障的平均停工时间

$$W_s = \frac{5}{\frac{1}{12}(1 - 0.0073)} - 15 = 46$$（分钟）

（6）每台机器平均待修时间

$$W_q = 46 - 12 = 34$$（分钟）

（7）系统绝对通过能力（即工人的维修能力）

$$A = \frac{1}{12}(1 - 0.0073) = 0.083$$（台）

即该工人每小时可修理机器的平均台数为$0.083 \times 60 = 4.96$台。

上述结果表面，机器停工时间过长，看管工人几乎没有空闲时间，应采取措施提高服务率或增加工人。

LINGO 计算程序如下

```
model:
lamda=1/15;mu=1/12;rho=lamda/mu;s=1;m=5;
load=m*rho;
L_s=@pfs(load,s,m);
p_0=1-(m-L_s)*rho;
lamda_e=lamda*(m-L_s);
p_5=@exp(@lgm(6))*0.8^5*p_0;
L_q=L_s-(1-p_0);
w_s=L_s/lamda_e;w_q=L_q/lamda_e;
end
```

7.2 服务率或到达率依赖状态的排队模型

在前面的各类排队模型的分析中，均假设顾客的到达率为常数λ，服务台的服务率也为常数μ。而在实际的排队问题中，到达率或服务率可能是随系统的状态而变化的。例如，当系统中顾客数已经比较多时，后来的顾客可能不愿意再进入系统；服务员的服务率当顾客较多时也可能会提高。因此，对单服务台系统，实际的到达率和服务率（它们均依赖于系统所处的状态n）可假设为

$$\lambda_n = \frac{\lambda_0}{(n+1)^a}, \quad n = 0,1,2,\cdots$$

$$\mu_n = n^b \mu_1, \quad n = 1,2,\cdots$$

对多服务台系统，实际到达率和服务率假设为

$$\lambda_n = \begin{cases} \lambda_0, & n \le s-1 \\ \left(\frac{s}{n+1}\right)^a \lambda_0, & n \ge s-1 \end{cases}$$

$$\mu_n=\begin{cases} n\mu_1, & n\le s \\ \left(\dfrac{n}{s}\right)^b s\mu_1, & n\ge s \end{cases}$$

其中 λ_n 和 μ_n 分别为系统处于状态 n 时的到达率和服务率。上述假设表明，到达率 λ_n 同系统中已有顾客数 n 呈反比关系；服务率 μ_n 同系统状态 n 呈正比关系。

由式（4），对多服务台系统有

$$C_n=\begin{cases} \dfrac{(\lambda_0/\mu_1)^n}{n!}, & n=1,2,\cdots,s \\ \dfrac{(\lambda_0/\mu_1)^n}{s!(n!/s!)^{a+b}s^{(1-a-b)(n-s)}}, & n=s,s+1,\cdots \end{cases} \tag{71}$$

下面看一个简单的特例，考虑一个到达依赖状态的单服务台等待制系统 $M/M/1/\infty$，其参数为

$$\lambda_n=\frac{\lambda}{n+1},\quad n=0,1,2,\cdots$$

$$\mu_n=\mu,\quad n=1,2,\cdots$$

于是由式（5），式（6），并设 $\rho=\dfrac{\lambda}{\mu}<1$，有

$$p_n=\frac{\rho^n}{n!}p_0,\quad n=1,2,\cdots \tag{72}$$

$$p_0=e^{-\rho} \tag{73}$$

平均队长

$$L_s=\sum_{n=0}^{\infty}np_n=\sum_{n=0}^{\infty}\frac{n\rho^n}{n!}p_0=\rho \tag{74}$$

平均排队长

$$L_q=\sum_{n=1}^{\infty}(n-1)p_n=L_s-(1-p_0)=\rho+e^{-\rho}-1 \tag{75}$$

有效到达率（单位时间内实际进入系统的顾客的平均数）

$$\lambda_e=\sum_{n=0}^{\infty}\frac{\lambda}{n+1}p_n=\mu(1-e^{-\rho}) \tag{76}$$

平均逗留时间为

$$W_s=\frac{L_s}{\lambda_e}=\frac{\rho}{\mu(1-e^{-\rho})} \tag{77}$$

平均等待时间

$$W_q=\frac{L_q}{\lambda_e}=W_s-\frac{1}{\mu} \tag{78}$$

7.3　非生灭过程排队模型

一个排队系统的特征是由输入过程，服务机制和排队规则决定的。本章前面所讨论的排队模型都是输入过程为 Poisson 流，服务时间服从负指数分布的生灭过程排队模型。这类排队系统的一个主要特征是马尔可夫性，而马尔可夫性的一个主要性质是由系统当前的状态可以推断未来的状态。但是，当输入过程不是 Poisson 流或服务时间不服从负指数分布时，仅知道系统内当前的顾客数，对于推断系统未来的状态是不充足的，因为正在接受服务的顾客，已经被服务了多长时间，将影响其离开系统的时间。因此，必须引入新的方法来分析具有非负指数分布的排队系统。

7.3.1　$M/G/1$排队模型

$M/G/1$系统是指顾客的到达为 Poisson 流，单个服务台，服务时间为一般分布的排队系统。现假设顾客的平均到达率为λ，服务时间的均值为$\frac{1}{\mu}$，方差为σ^2，则可证明：当$\rho=\frac{\lambda}{\mu}<1$时，系统可以达到平稳状态，而给出平稳分布的表示是比较困难的。已有的几个结果为：

$$p_0=1-\rho \tag{79}$$

$$L_q=\frac{\lambda^2\sigma^2+\rho^2}{2(1-\rho)} \tag{80}$$

$$L_s=\rho+L_q \tag{81}$$

$$W_q=\frac{L_q}{\lambda} \tag{82}$$

$$W_s=W_q+\frac{1}{\mu} \tag{83}$$

由式（80）可看出，L_q,L_s,W_s,W_q等仅依赖于ρ和服务时间的方差σ^2，而与分布的类型没有关系，这是排队论中一个非常重要且令人惊奇的结果，式（80）通常被称为 Pollaczek-Khintchine（P-K）公式。

从式（80）还不难发现，当服务率μ给定后，当方差σ^2减少时，平均队长和等待时间等都将减少。因此，可通过改变服务时间的方差来缩短平均队长，当且仅当$\sigma^2=0$，即服务时间为定长时，平均队长（包括等待时间）可减少到最少水平，这一点是符合直观的，因为服务时间越有规律，等候的时间也就越短。

例 8　有一汽车冲洗台，汽车按 Poisson 流到达，平均每小时到达 18 辆，冲洗时间V 根据过去的经验表明，有$E(V)=0.05h/$辆，$\mathrm{Var}(V)=0.01(\mathrm{h/辆})^2$，求有关运行指标，并对系统进行评价。

解　本例中，$\lambda=18$，$\rho=\lambda E(V)=18\times0.05=0.9$，$\sigma^2=0.01$，$\mu=20$，于是

$$L_q=\frac{18^2\times0.01+(0.9)^2}{2(1-0.9)}=20.25\ （辆）$$

$$L_s=20.25+0.9=21.15\ （辆）$$

$$W_s=\frac{21.15}{18}=1.175\ （\mathrm{h}）$$

$$W_q = \frac{20.25}{18} = 1.125 \text{（h）}$$

上述结果表明，这个服务机构很难令顾客满意，突出的问题是顾客的平均等待时间是服务时间的$\frac{W_q}{E(V)} = \frac{1.125}{0.05} = 22.5$倍（通常称$\frac{W_q}{E(V)}$为顾客的时间损失系数）。

例 9　考虑定长服务时间$M/D/1/\infty$模型，这时，$E(V) = \frac{1}{\mu}$，$\sigma^2 = \mathrm{Var(V)} = 0$，由式（80）有

$$L_q = \frac{\rho^2}{2(1-\rho)} = \frac{\lambda^2}{2\mu(\mu-\lambda)} \tag{84}$$

$$L_s = L_q + \rho = \frac{\lambda(2\mu-\lambda)}{2\mu(\mu-\lambda)} \tag{85}$$

$$W_q = \frac{\rho^2}{2\lambda(1-\rho)} = \frac{\lambda}{2\mu(\mu-\lambda)} \tag{86}$$

$$W_s = W_q + \frac{1}{\mu} \tag{87}$$

将式（13）和式（86）比较，不难发现在服务时间服从负指数分布的条件下，等待时间正好是定长服务时间的 2 倍。可以证明，在一般服务时间分布下得到的L_q和W_q中，以定长服务时间下得到的为最小。

7.3.2　爱尔朗（Erlang）排队模型

爱尔朗分布族比负指数分布族对现实世界具有更广泛的适应性。下面介绍一个最简单的爱尔朗排队模型。

对爱尔朗排队模型研究的一般方法是根据k阶 Erlang 分布恰为k个相同负指数分布随机变量和的分布这个关系，把服务时间或到达过程假想地（实际并非如此）分为k个独立的同分布的位相（或阶段），然后利用负指数分布的性质来加以分析。如对$M/E_k/1/\infty$系统来说，服务时间是k阶 Erlang 分布，把每个顾客的服务时间假想地分为k个位相，每个位相的平均服务时间为$\frac{1}{k\mu}$，顾客先进入第k个位相，最后进入第 1 个位相。仍令N为系统达到平衡状态时的顾客数，但考虑到顾客可能处在不同位相，故系统的状态一般用(n,i)表示，其中n表示有n个顾客在系统中，i表示正在接受服务的顾客处在第i个位相，令

$$p_{ni} = P\{N = (n,i)\}$$

则可得到类似于（3）的差分方程组，从而在平稳分布存在的条件下得到平稳分布和各有关指标。由于本节已给出了$M/G/1/\infty$系统的主要结果，作为一个特例，可直接给出$M/E_k/1/\infty$的主要数量指标。

由于服务时间为k阶 Erlang 分布，其分布密度函数为

$$a(t) = \frac{\mu k(\mu k t)^{k-1}}{(k-1)!} e^{-\mu k t}, \quad t \geq 0 \tag{88}$$

故其均值和方差分别为

$$E(E_k)=\frac{1}{\mu},\quad \mathrm{Var}(E_k)=\frac{1}{k\mu^2}$$

将 $\rho=\frac{\lambda}{\mu}$，$\sigma^2=\frac{1}{k\mu^2}$ 代入式（80）～式（83），得

$$L_q=\frac{\rho^2(k+1)}{2k(1-\rho)}=\frac{\rho^2}{1-\rho}-\frac{(k-1)\rho^2}{2k(1-\rho)} \tag{89}$$

$$L_s=L_q+\rho=\frac{\rho}{1-\rho}-\frac{(k-1)\rho^2}{2k(1-\rho)} \tag{90}$$

$$W_s=\frac{1}{\mu(1-\rho)}-\frac{(k-1)\rho}{2k\mu(1-\rho)} \tag{91}$$

$$W_q=\frac{\rho}{\mu(1-\rho)}-\frac{(k-1)\rho}{2k\mu(1-\rho)} \tag{92}$$

例 10　设一电话间的顾客按 Poisson 流到达，平均每小时到达 6 人，平均通话时间为 8 分钟，方差为 16 分钟。直观上估计通话时间服从爱尔朗分布，管理人员想知道平均排队长度和顾客平均等待时间是多少？

解　设 V 为通话时间，服从 k 阶 Erlang 分布，由

$$k=\frac{[E(V)]^2}{\mathrm{Var}(V)}=\frac{8^2}{16}=4$$

可知该系统为 $M/E_4/1/\infty$ 系统，其中 $\rho=6\times\frac{8}{60}=0.8$。由式（89），有

$$L_q=\frac{(0.8)^2(4+1)}{2\times4(1-0.8)}=2\ \text{（人）}$$

$$W_q=\frac{L_q}{\lambda}=\frac{2}{6}=0.33\ \text{（h）}$$

§8　排队系统的优化

排队系统中的优化模型，一般可分为系统设计的优化和系统控制的优化。前者为静态优化，即在服务系统设置以前根据一定的质量指标，找出参数的最优值，从而使系统最为经济。后者为动态优化，即对已有的排队系统寻求使其某一目标函数达到最优的运营机制。由于对后一类问题的阐述需要较多的数学知识，所以本节着重介绍静态最优问题。

在优化问题的处理方法上，一般根据变量的类型是离散的还是连续的，相应地采用边际分析方法或经典的微分法，对较为复杂的优化问题需要用非线性规划或动态规划等方法。

8.1　$M/M/1$ 模型中的最优服务率 μ

先考虑 $M/M/1/\infty$ 模型，取目标函数 z 为单位时间服务成本与顾客在系统中逗留费用之和的期望值，即

$$z=c_s\mu+c_wL_s$$

其中 c_s 为服务一个顾客时单位时间内的服务费用，c_w 为每个顾客在系统中逗留单位时

间的费用，则由式（9），有

$$z = c_s\mu + c_w\frac{\lambda}{\mu-\lambda}$$

令

$$\frac{dz}{d\mu} = c_s - c_w\lambda\frac{1}{(\mu-\lambda)^2} = 0$$

解出最优服务率为

$$\mu^* = \lambda + \sqrt{\frac{c_w}{c_s}\lambda} \tag{93}$$

下面考虑 $M/M/1/K$ 模型，从使服务机构利润最大化来考虑。由于在平稳状态下，单位时间内到达并进入系统的平均顾客数为 $\lambda_e = \lambda(1-p_K)$，它也等于单位时间内实际服务完的平均顾客数。设每服务一个顾客服务机构的收入为 G 元，于是单位时间内收入的期望值是 $\lambda(1-p_K)G$ 元，故利润 z 为

$$z = \lambda(1-p_K)G - c_s\mu = \lambda G\frac{1-\rho^K}{1-\rho^{K+1}} - c_s\mu$$

$$= \lambda\mu G\frac{\mu^K-\lambda^K}{\mu^{K+1}-\lambda^{K+1}} - c_s\mu$$

令 $\frac{dz}{d\mu}=0$，得

$$\rho^{K+1}\frac{K-(K+1)\rho+\rho^{K+1}}{(1-\rho^{K+1})^2} = \frac{c_s}{G} \tag{94}$$

当给定 K 和 $\frac{c_s}{G}$ 后，即可由（94）式得到最优利润的 μ^*。

例 11　设某工人照管 4 台自动机床，机床运转时间（或各台机床损坏的相继时间）平均为负指数分布，假定平均每周有一台机床损坏需要维修，机床运转单位时间内平均收入 100 元，而每增加 1 单位 μ 的维修费用为 75 元。求使总利益达到最大的 μ^*。

解　该系统为 $M/M/1/K/K$ 系统，其中

$$K=4,\ \lambda=1,\ G=100,\ C_s=75$$

设 L_s 是队长，则正常运装的机器为 $K-L_s$ 部，因此目标函数为

$$f = 100(K-L_s) - 75\mu$$

题意就是在上述条件下，求目标函数 f 的最大值。

编写 LINGO 程序如下：

```
model:
s=1;k=4;lamda=1;
L_s=@pfs(k*lamda/mu,s,k);
max=100*(k-L_s)-75*mu;
end
```

求得 $\mu^* = 1.799$，最优目标值 $f^* = 31.49$。

例 12　假定有一混合制排队系统 $M/M/1/K$，其中 $K=3$，顾客的到达率为每小时 3.6 人，其到达间隔服从 Poisson 过程，系统服务一个顾客收费 2 元。又设系统的服务强度 μ（$\mu=\dfrac{1}{T}$，T 为服务时间）服从负指数分布，其服务成本为每小时 0.5μ 元。求系统为每个顾客的最佳服务时间。

解　系统的损失率为 p_K，则系统每小时服务的人数为 $\lambda(1-p_K)$，每小时运行成本为 0.5μ，因此目标函数为

$$f=2\lambda(1-p_K)-0.5\mu$$

题意就是在上述条件下，求目标函数 f 的最大值。

编写 LINGO 程序如下：

```
model:
sets:
state/1..3/:p;
endsets
lamda=3.6;k=3;
lamda*p0=p(1)/t;
(lamda+1/t)*p(1)=lamda*p0+p(2)/t;
@for(state(i)|i #gt# 1 #and# i #lt# k:
(lamda+1/t)*p(i)=lamda*p(i-1)+p(i+1)/t);
lamda*p(k-1)=p(k)/t;
p0+@sum(state:p)=1;
max=2*lamda*(1-p(k))-0.5/t;
end
```

求得系统为每位顾客最佳服务时间是0.2238h，系统每小时赢利3.70元。

8.2　$M/M/s$ 模型中的最优的服务台数 s^*

这里仅讨论 $M/M/s/\infty$ 系统，已知在平稳状态下单位时间内总费用（服务费用与等待费用）之和的平均值为

$$z=c_s's+c_wL \tag{95}$$

其中 s 为服务台数，c_s' 是每个服务台单位时间内的费用，L 是平均队长。由于 c_s'，c_w 是给定的，故唯一可变的是服务台数 s，所以可将 z 看成是 s 的函数，记为 $z=z(s)$，并求使 $z(s)$ 达到最小的 s^*。

因为 s 只取整数，$z(s)$ 不是连续函数，故不能用经典的微分法，下面采用边际分析方法。根据 $z(s^*)$ 应为最小的特点，有

$$\begin{aligned} z(s^*)&\le z(s^*-1)\\ z(s^*)&\le z(s^*+1)\end{aligned} \tag{96}$$

将式（95）代入式（96），得

$$c_s's^*+c_wL(s^*)\le c_s'(s^*-1)+c_wL(s^*-1)$$

$$c_s's^*+c_wL(s^*)\le c_s'(s^*+1)+c_wL(s^*+1)$$

化简后得到

$$L(s^*)-L(s^*+1)\le\frac{c_s^{'}}{c_w}\le L(s^*-1)-L(s^*) \qquad (97)$$

依次求当$s=1,2,3,\cdots$时L的值，并计算相邻两个L值的差。因$\frac{c_s^{'}}{c_w}$是已知数，根据其落在哪个与s有关的不等式中，即可定出最优的s^*。

例 13　某检验中心为各工厂服务，要求进行检验的工厂(顾客)的到来服从 Poisson 流，平均到达率为$\lambda=48$（次/d)；每天来检验由于停工等原因损失 6 元；服务（检验）时间服从负指数分布，平均服务率为$\mu=25$（次/d)；每设置一个检验员的服务成本为 4 元/d，其它条件均适合$M/M/s/\infty$系统。问应设几个检验员可使总费用的平均值最少？

解　已知$c_s^{'}=4$，$c_w=6$，$\lambda=48$，$\mu=25$，$\frac{\lambda}{\mu}=1.92$，设检验员数为s，由式（20）和式（25）

$$p_0=\left[\sum_{n=0}^{s-1}\frac{(1.92)^n}{n!}+\frac{(1.92)^s}{(s-1)!(s-1.92)}\right]^{-1}$$

$$L=L_q+\rho=\frac{p_0(1.92)^{s+1}}{(s-1)!(s-1.92)^2}+1.92$$

将$s=1,2,3,4,5$依次代入得到表 2。由于$\frac{c_s^{'}}{c_w}=\frac{4}{6}=0.67$落在区间$(0.582,21.845)$之间，故$s^*=3$，即当设 3 个检验员时可使总费用$z$最小，最小值为

$$z(s^*)=z(3)=27.87\text{（元）}$$

表 2

检验员数s	平均顾客数$L(s)$	$L(s)-L(s+1)\sim L(s-1)-L(s)$	总费用$z(s)$
1	∞		∞
2	24.49	21.845～∞	154.94
3	2.645	0.582～21.845	27.87
4	2.063	0.111～0.582	28.38
5	1.952		31.71

求解的 LINGO 程序如下：

```
model:
lamda=48;mu=25;rho=lamda/mu;
P_wait=@peb(rho,s);
L_q=P_wait*rho/(s-rho);
L_s=L_q+rho;
min=4*s+6*L_s;
@gin(s);@bnd(2,s,5);
end
```

§9　产生给定分布的随机数的方法

Matlab 可以产生常用分布的随机数。下面我们介绍按照给定的概率分布产生随机

数的一般方法，这些方法都以$U(0,1)$分布的随机变量为基础。

(i)反变换法

定理　设X是一个具有连续分布函数$F(x)$的随机变量，则$F(X)$在[0，1]上服从均匀分布。

设概率分布函数$F(x)$是严格单调增的，F的反函数记作F^{-1}。先产生$U\sim U(0,1)$，再取$X=F^{-1}(U)$即为所求，称为反变换法。

指数分布$\mathrm{Exp}(\lambda)$能够方便地用反变换法产生。由$\mathrm{Exp}(\lambda)$的分布函数$F(x)=1-e^{-\lambda x}$，可得

$$X=F^{-1}(U)=-\frac{\ln(1-U)}{\lambda}$$

思考　有的书上用$X=-\frac{\ln U}{\lambda}$代替上式，对吗，为什么？

（ii）卷积法

如果随机变量X是n个独立、同分布的另一随机变量Y之和，而Y又容易产生时，先产生n个独立的$Y_1,Y_2,\cdots,Y_n$，再令$X=Y_1+\cdots+Y_n$即可。因为X的分布函数是$Y_1,Y_2,\cdots,Y_n$分布函数的卷积，故称为卷积法。

二项分布可以用卷积法产生。因为$X\sim B(n,p)$是n个独立的$Y\sim Bern(p)$之和，而$Y\sim Bern(p)$很容易由$U\sim U(0,1)$按以下方法得到：若$U\le p$，令$Y=1$；否则令$Y=0$。

(iii)取舍法

若随机变量X在有限区间(a,b)内变化，但概率密度$f(x)$具有任意形式（甚至没有解析表达式），无法用前面的方法产生时，可用取舍法。一种比较简单的取舍法的步骤是：

1° 产生$Y\sim U(a,b)$和$U\sim U(0,1)$；

2° 记$C=\max\limits_{a\le x\le b}f(x)$，若$U\le\frac{f(Y)}{C}$，则取$X=Y$；否则，舍去，返回1°。

§10　排队模型的计算机模拟

10.1 确定随机变量概率分布的常用方法

在模拟一个带有随机因素的实际系统时，究竟用什么样的概率分布描述问题中的随机变量，是我们总是要碰到的一个问题，下面简单介绍确定分布的常用方法：

1° 根据一般知识和经验，可以假定其概率分布的形式，如顾客到达间隔服从指数分布$\mathrm{Exp}(\lambda)$；产品需求量服从正态分布$N(\mu,\sigma^2)$；订票后但未能按时前往机场登机的人数服从二项分布$B(n,p)$。然后由实际数据估计分布的参数λ,μ,σ等，参数估计可用极大似然估计、矩估计等方法。

2° 直接由大量的实际数据作直方图，得到经验分布，再通过假设检验，拟合分布函数，可用χ^2检验等方法。

3° 既缺少先验知识，又缺少数据时，对区间(a,b)内变化的随机变量，可选用Beta分布（包括均匀分布）。先根据经验确定随机变量的均值μ和频率最高时的数值（即密

度函数的最大值点）m，则 Beta 分布中的参数α_1,α_2可由以下关系求出：

$$\mu = a + \frac{\alpha_1(b-a)}{\alpha_1+\alpha_2},\quad m = a + \frac{(\alpha_1-1)(b-a)}{\alpha_1+\alpha_2-2}.$$

10.2　计算机模拟

当排队系统的到达间隔时间和服务时间的概率分布很复杂时，或不能用公式给出时，那么就不能用解析法求解。这就需用随机模拟法求解，现举例说明。

例 14　设某仓库前有一卸货场，货车一般是夜间到达，白天卸货，每天只能卸货 2 车，若一天内到达数超过 2 车，那么就推迟到次日卸货。根据表 3 所示的数据，货车到达数的概率分布（相对频率）平均为 1.5 车/天，求每天推迟卸货的平均车数。

表 3　到达车数的概率

到达车数	0	1	2	3	4	5	≥6
概　　率	0.23	0.30	0.30	0.1	0.05	0.02	0.00

解　这是单服务台的排队系统，可验证到达车数不服从泊松分布，服务时间也不服从指数分布（这是定长服务时间）。

随机模拟法首先要求事件能按历史的概率分布规律出现。模拟时产生的随机数与事件的对应关系如表 4。

表 4　到达车数的概率及其对应的随机数

到达车数	概　率	累积概率	对应的随机数
0	0.23	0.23	$0 \le x < 0.23$
1	0.30	0.53	$0.23 \le x < 0.53$
2	0.30	0.83	$0.53 \le x < 0.83$
3	0.1	0.93	$0.83 \le x < 0.93$
4	0.05	0.98	$0.93 \le x < 0.98$
5	0.02	1.00	$0.98 \le x \le 1.00$

我们用 a1 表示产生的随机数，a2 表示到达的车数，a3 表示需要卸货车数，a4 表示实际卸货车数，a5 表示推迟卸货车数。编写程序如下：

```
clear
rand('state',sum(100*clock));
n=50000;
m=2
a1=rand(n,1);
a2=a1; %a2初始化
a2(find(a1<0.23))=0;
a2(find(0.23<=a1&a1<0.53))=1;
a2(find(0.53<=a1&a1<0.83))=2;
a2(find(0.83<=a1&a1<0.93),1)=3;
a2(find(0.93<=a1&a1<0.98),1)=4;
a2(find(a1>=0.98))=5;
a3=zeros(n,1);a4=zeros(n,1);a5=zeros(n,1); %a2初始化
a3(1)=a2(1);
if a3(1)<=m
   a4(1)=a3(1);a5(1)=0;
else
```

```
   a4(1)=m;a5(1)=a2(1)-m;
end
for i=2:n
   a3(i)=a2(i)+a5(i-1);
   if a3(i)<=m
   a4(i)=a3(i);a5(i)=0;
   else
   a4(i)=m;a5(i)=a3(i)-m;
   end
end
a=[a1,a2,a3,a4,a5];
sum(a)/n
```

例 15　银行计划安置自动取款机，已知 A 型机的价格是 B 型机的 2 倍，而 A 型机的性能—平均服务率也是 B 型机的 2 倍，问应该购置 1 台 A 型机还是 2 台 B 型机。

为了通过模拟回答这类问题，作如下具体假设，顾客平均每分钟到达 1 位，A 型机的平均服务时间为 0.9 分钟，B 型机为 1.8 分钟，顾客到达间隔和服务时间都服从指数分布，2 台 B 型机采取 $M/M/2$ 模型（排一队），用前 100 名顾客（第 1 位顾客到达时取款机前为空）的平均等待时间为指标，对 A 型机和 B 型机分别作 1000 次模拟，进行比较。

理论上已经得到，A 型机和 B 型机前 100 名顾客的平均等待时间分别为 $\mu_1(100)=4.13$，$\mu_2(100)=3.70$，即 B 型机优。

对于 $M/M/1$ 模型，记第 k 位顾客的到达时刻为 c_k，离开时刻为 g_k，等待时间为 w_k，它们很容易根据已有的到达间隔 i_k 和服务时间 s_k 按照以下的递推关系得到（$w_1=0$，设 c_1, g_1 已知）：

$$c_k=c_{k-1}+i_k,\quad g_k=\max(c_k,g_{k-1})+s_k$$

$$w_k=\max(0,g_{k-1}-c_k),\quad k=2,3,\cdots$$

在模拟 A 型机时，我们用cspan表示到达间隔时间，sspan表示服务时间，ctime表示到达时间，gtime表示离开时间，wtime表示等待时间。我们总共模拟了 m 次，每次 n 个顾客。程序如下：

```
tic
rand('state',sum(100*clock));
n=100;m=1000;mu1=1;mu2=0.9;
for j=1:m
   cspan=exprnd(mu1,1,n);sspan=exprnd(mu2,1,n);
   ctime(1)=cspan(1);
   gtime(1)=ctime(1)+sspan(1);
   wtime(1)=0;
   for i=2:n
      ctime(i)=ctime(i-1)+cspan(i);
      gtime(i)=max(ctime(i),gtime(i-1))+sspan(i);
      wtime(i)=max(0,gtime(i-1)-ctime(i));
   end
   result1(j)=sum(wtime)/n;
end
result_1=sum(result1)/m
```

```
toc
```

类似地，模拟B型机的程序如下：

```
tic
rand('state',sum(100*clock));
n=100;m=1000;mu1=1;mu2=1.8;
for j=1:m
   cspan=exprnd(mu1,1,n);sspan=exprnd(mu2,1,n);
   ctime(1)=cspan(1);ctime(2)=ctime(1)+cspan(2);
   gtime(1:2)=ctime(1:2)+sspan(1:2);
   wtime(1:2)=0;flag=gtime(1:2);
   for i=3:n
      ctime(i)=ctime(i-1)+cspan(i);
      gtime(i)=max(ctime(i),min(flag))+sspan(i);
      wtime(i)=max(0,min(flag)-ctime(i));
      flag=[max(flag),gtime(i)];
   end
   result2(j)=sum(wtime)/n;
end
result_2=sum(result2)/m
toc
```

读者可以用下面的程序与上面的程序比较了解编程的效率问题。

```
tic
clear
rand('state',sum(100*clock));
n=100;m=1000;mu1=1;mu2=0.9;
for j=1:m
   ctime(1)=exprnd(mu1);
   gtime(1)=ctime(1)+exprnd(mu2);
   wtime(1)=0;
   for i=2:n
      ctime(i)=ctime(i-1)+exprnd(mu1);
      gtime(i)=max(ctime(i),gtime(i-1))+exprnd(mu2);
      wtime(i)=max(0,gtime(i-1)-ctime(i));
   end
   result(j)=sum(wtime)/n;
end
result=sum(result)/m
toc
```

习 题 六

1. 一个车间内有10台相同的机器，每台机器运行时每小时能创造4元的利润，且平均每小时损坏一次。而一个修理工修复一台机器平均需4小时。以上时间均服从指数分布。设一名修理工一小时工资为6元，试求：

（i）该车间应设多少名修理工，使总费用为最小；

（ii）若要求不能运转的机器的期望数小于4台，则应设多少名修理工；

（iii）若要求损坏机器等待修理的时间少于4小时，又应设多少名修理工。

2. 到达某铁路售票处顾客分两类：一类买南方线路票，到达率为λ_1/小时，另一类买北方线路票，到达率为λ_2/小时，以上均服从泊松分布。该售票处设两个窗口，各

窗口服务一名顾客时间均服从参数 $\mu = 10$ 的指数分布。试比较下列情况时顾客分别等待时间 W_q：（i）两个窗口分别售南方票和北方票；（ii）每个窗口两种票均出售。（分别比较 $\lambda_1 = \lambda_2 = 2,4,6,8$ 时的情形）

3. 一名修理工负责5台机器的维修，每台机器平均每2h损坏一次，又修理工修复一台机器平均需时18.75min，以上时间均服从负指数分布。试求：

（1）所有机器均正常运转的概率；

（2）等待维修的机器的期望数；

（3）假如希望做到有一半时间所有机器都正常运转，则该修理工最多看管多少台机器。

（4）假如维修工工资为8元/h，机器不能正常运转时的损失为40元/h，则该修理工看管多少台机器较为经济合理。