第 9 章　排队论

排队论是我们每个人都很熟悉的现象。因为人或物或是信息为了得到某种服务必须排队。有一类排队是有形的，例如在售票处等待买票的排队，加油站前汽车等待加油的排队等；还有一类排队是无形的，例如电话交换机接到的电话呼叫信号的排队，等待计算机中心处理机处理的信息的排队等。为了叙述的方便，排队者无论是人、物、或信息，以后统称为“ 顾客 ”。服务者无论是人，或事物，例如一台电子计算机也可以是排队系统中的服务者，我们以后统称为“ 服务员 ”。

排队现象是我们不希望出现的现象，因为人的排队意味着至少是浪费时间；物的排队则说明了物资的积压。但是排队现象却无法完全消失，这是一种随即现象。由于顾客到达间隔时间的随机性和为顾客服务时间的随机性是排队现象产生的原因。如果上述的两个时间是固定的，我们就可以通过妥善安排来完全消除排队现象。

排队论是研究排队系统在不同的条件下（最主要的是顾客到达的随机规律和服务时间的随机规律）产生的排队现象的随机规律性。也就是要建立反映这种随机性的数学模型。研究的最终目的是为了运用这些规律，对实际的排队系统的设计与运行做出最优的决策。

排队论中的数学模型是根据概率和随机过程的理论建立起来的，我们先来讨论泊松过程和生灭过程，然后,再此基础上研究排队系统的结构及其主要的数学模型，最后研究排队系统的优化问题。

9.1 泊松过程和生灭过程

9.1.1 泊松过程

如果用 $N(t)$ 表示在 $[0,t]$ 时间内顾客到达的总数，则对于每个给定的时刻 t，$N(t)$ 都是一个随机变量。随即变量族 $\{N(t)\big|t\in[0,T]\}$ 称作是一个随机过程。

若对　$t_1<t_2<\cdots t_n<t_{n+1}$，有

$$P(N(t_{n+1})=i_{n+1}\big|N(t_1)=i_1,N(t_2)=i_2,\cdots,N(t_n)=i_n$$

$$=P(N(t_{n+1})=i_{n+1}\big|N(t_n)=i_n) \tag{9-1}$$

则称随即过程 $\{N(t)\big|t\in[0,T)\}$ 为马尔柯夫过程。公式（9-1）所标示的性质称为“ 无后效性 ”。它的实际意义是说：如果用 t_n 表示现在时刻，t_{n+1} 表示未来时刻，$t_1,t_2,\cdots,t_{n-1}$ 表示过去的一系列时刻，则顾客到来的过程在 t_n 以前所处的状态

（即顾客到达数）对预言过程在 t_n 以后的状态不起直接作用。

若随即过程 $\{N(t)\big|t\in[0,T]\}$ 就有“独立增量性”，即对任一组 $t_1<t_2,\cdots,t_n(n\geq 3)$，随即变量 $N(t_2)-N(t_1),N(t_3)-N(t_2),\cdots,N(t_n)-N(t_{n-1})$ 相互独立，且对任意 $t\in[0,T]$, 有

$$P(N(t)=k)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}\qquad k=0,1,2,\cdots\cdots \tag{9-2}$$

其中参数 $\lambda>0$，则称这个过程为泊松过程。

独立增量性说明在互不相交的时间区间 $[t_1,t_2),[t_2,t_3),\cdots\cdots,[t_{n-1},t_n)$ 内顾客到达情况是相互独立的。由于 $\sum\limits_{k=0}^{\infty}\frac{(\lambda t)^k}{k!}e^{-\lambda t}=1$

所以 $N(t)$ 的期望值为

$$E(N(t))=\sum_{k=0}^{\infty}k\frac{(\lambda t)^k}{k!}e^{-\lambda t}=\lambda t\sum_{k=1}^{\infty}\frac{(\lambda t)^{k-1}}{(k-1)!}\cdot e^{-\lambda t}$$

$$=\lambda t\sum_{i=0}^{\infty}\frac{(\lambda t)^i}{i}e^{-\lambda t}=\lambda t \tag{9-3}$$

$$\lambda=\frac{E(N(t))}{t} \tag{9-4}$$

因此，参数 λ 就是单位时间间隔内到达顾客的平均数，同时，我们还可以求得随机变量 $N(t)$ 的方差为

$$D(N(t))=\lambda t \tag{9-5}$$

例 9.1 某天上午，从 10 点 30 分到 11 点 47 分，每隔 20 秒钟统计一次来到某汽车站的乘客数，共得 230 个记录数据，整理后得到如下的统计结果：

表 9-1

乘客数目	0	1	2	3	4
频　　数	100	81	34	9	6

试用一个泊松过程来描述此车站乘客的到达过程，并具体写出它的概率分布。

解：要写出其概率分布，只需确定公式（9-2）中的参数 λ 即可。根据 λ 的意义，只要先求出每 20 秒钟的平均数

$$\bar{\lambda}=\frac{1}{230}(0\times100+1\times81+2\times34+3\times9+4\times6)=0.87$$

因此可知每分钟平均到达的顾客数为

$$\lambda = 3\times 0.87 = 2.61 \text{（人/分钟）}$$

故所求的乘客到达过程所满足的概率分布为

$$P(N(t)=k)=\frac{(2.61t)^k}{k!}e^{-2.61t}$$

一般地，有如下结论：

定理 1 若随机过程$\{N(t)\big|t\in[0,T)\}$满足下列三个条件：

(1) 独立增量性：对任一组$t_1<t_2<\cdots\cdots<t_n(n\geq 3)$，随机变量$N(t_2)-N(t_1),N(t_3)-N(t_2),\cdots\cdots,N(t_n)-N(t_{n-1})$相互独立；

(2)平稳性：对于$[s,s+t]\subset[0,T]$，总有

$$P[N(s+t)-N(s)=k]=P[N(t)-N(0)=k]$$

其中$P(N(0)=0)=1,\sum_{k=0}^{\infty}P(N(t)=k)=1$；

(3)普遍性：令$\varphi(t)=\sum_{k=2}^{\infty}P(N(t)=k)$，有

$$\lim_{t=0}\frac{\varphi(t)}{t}=0$$

则$\{N(t)\big|t\in[0,T]\}$是一个泊松过程。

独立增量性说明在$[0,T]$中的任意区间$[s,s+t]$内来到k个顾客这一事件与区间$[0,s]$内来到顾客的情况相互独立，即在$[0,s]$内顾客来到的情况所作的任何假定下，计算出来的在$[s,s+t]$内来到k个顾客的条件概率均相等。同时可知，具有独立增量性的过程必然具有无后效性。

平稳性说明在$[s,s+t]$内来到的数值与区间长度t有关，而与时间起点s无关。也就是说，过程的统计规律不随时间的推移而改变，在同样长度的时间间隔内来到k个顾客的概率是一个常数。

普遍性表明，在同一瞬间来到两个或两个以上顾客实际上是不可能的。即在充分小的时间间隔中，最多来到一个顾客。

在排队论里，常把泊松过程称为泊松流或最简单流，参数λ称为最简单流的强度。顺便说一下，泊松过程还具有可知性。即如果$\{N_1(t)\}$和$\{N_2(t)\}$是两个泊松过程，到达强度分别为λ_1和λ_2，且两个过程相互独立，则$\{N_1(t)\}+\{N_2(t)\}$仍为一泊松过程，其到达强度为$\lambda_1+\lambda_2$，推广此结论，则更多个独立的泊松过程合并后仍为一泊松过程，其到达强度为各过程到达强度之和。

泊松过程在排队论中起着重要作用。因为泊松流或近似于泊松流的实际情况经常会遇到，并且泊松流的数学处理很简单。

9.1.2 负指数分布和爱尔朗分布

若用T_n表示第n位顾客所需的服务时间，则$\{T_n\}$是一族连续性随机变量。如

果$\{\tau_n\}$中各个随机变量相互独立，且服从相同的负指数分布：

$$P(\tau_n \le t)=\begin{cases}1-e^{-\mu t} & t\ge 0\\ 0 & t<0\end{cases} \tag{9-6}$$

（其中参数$\mu>0$）其概率密度函数为

$$p(t)=\begin{cases}\mu\cdot e^{-\mu t} & t\ge 0\\ 0 & t<0\end{cases} \tag{9-7}$$

则服务时间τ_n的期望值为：

$$E(\tau_n)=\int_{-\infty}^{+\infty} tp(t)dt=\int_0^{+\infty}\mu te^{-\mu t}dt$$

$$=-te^{-\mu t}\Big|_0^{+\infty}+\int_0^{+\infty}e^{-\mu t}dt=\frac{1}{\mu} \tag{9-8}$$

则有

$$\mu=\frac{1}{E(\tau_n)} \tag{9-9}$$

于是，$\frac{1}{\mu}$就是每位顾客所需要的平均服务时间，而μ表示单位时间内能被服务完的顾客平均数。在排队论中通常用“平均”来表示概率论中的数学期望。

同时可求得τ_n的方差为

$$D(\tau_n)=\frac{1}{\mu^2} \tag{9-10}$$

设$\{N(t)\}$是描述顾客到达情况的随机过程，以t_n表示第n个顾客到达的时刻，则 $T_n=t_n-t_{n-1}$为第n位顾客与他的前一位顾客（第$n-1$位顾客）到达时间的时间间隔。显然$\{T_n\}$也是一族随机变量。关于到达间隔时间$\{T_n\}$与顾客到达过程$\{N(t)\}$之间的关系，可证得如下的结论：

定理 2 顾客到达过程$\{N(t)\}$是一个参数为λ的泊松过程的充分必要条件为：相应的顾客到达间隔$\{T_n\}$是一族相互独立的随机变量，且每个随机变量都服从下面的负指数分布：

$$P(T_n\le t)=\begin{cases}1-e^{-\mu t} & t\ge 0\\ 0 & t<0\end{cases} \tag{9-11}$$

此定理说明，“顾客流是最简单流”与“顾客到达间隔相互独立且服从相同的负指数分布”是等价的两种描述方法。由于上面的负指数分布的数学期望$E(T_n)=\frac{1}{\lambda}$，所以在最简单流中，顾客到达时间间隔的平均值为$\frac{1}{\lambda}$。

例 9.2 在某座大桥桥口，观察到 26 辆到达桥口要过桥的汽车，其到达时刻记录如下（开始观察时刻为 0，单位为秒）：

0	15	17	23	24	25	31	39	55	58	62	63	65
68	80	82	85	89	97	99	103	111	121	122	123	133

试用一个泊松过程描述这个到达过程，并写出具体的概率模型。

解：因为泊松流中，顾客到达的时间间隔的平均值为$\frac{1}{\lambda}$，所以可着手求得λ，再定出概率模型。汽车到达的时间间隔依次为：

15	2	6	1	1	6	8	16	3	4	1	2	3
12	2	3	4	8	2	4	8	10	1	1	10	

因此，到达间隔时间的平均值为

$$\frac{15+2+\cdots+10}{25}=\frac{133}{25}=5.23\ \text{（秒）}$$

就是平均每隔 5.32 秒钟到达一辆汽车。因为顾客到达间隔平均值为$\frac{1}{\lambda}$，而λ就是泊松流的概率模型中参数，由$\frac{1}{\lambda}=5.32$可得$\lambda=\frac{1}{5.32}=0.188$，即每秒钟平均到达的汽车数约为 0.188。

于是可用如下的泊松分布来描述到达过程$\{N(t)\}$：

$$P(N(t)=k)=\frac{(0.188t)^k}{k!}e^{-0.188t}\qquad k=0,1,2,\cdots$$

也可用如下的负指数分布来描述其到达间隔$\{T_n\}$：

$$P(T_n\le t)=\begin{cases}1-e^{-0.188t} & t\ge 0\\ 0 & t<0\end{cases}$$

关于负指数分布，需要强调一个它所具有的“ 无记忆性 ”。先看一个问题，如果顾客到达间隔时间服从负指数分布，平均间隔时间 10 秒，又假设在某一时刻（任一时刻）来考察这个到达过程，发现最后一位顾客已到达了 7 秒，那么下一位顾客平均还需多长时间才会到达呢？回答是出人意料之外的：还需要 10 秒。看来已经过去了的 7 秒被遗忘了，故称“ 无记忆性 ”。下面来证明这一特性：

设到达间隔T服从负指数分布如公式（9-11）所示，而s为任一时刻，则到达间隔T大于等于s的概率为：

$$P(T\ge s)=1-P(T\le s)=\int_s^{+\infty}\lambda e^{-\lambda t}dt=e^{-\lambda s}$$

又因为$T\ge s$与$T\ge s+t(t>0)$同时发生的概率等于$T\le s+t$的概率，即

$$P(T\ge s,T\ge s+t)=P(T\le s+t)=\int_{s+t}^{+\infty}\lambda e^{-\lambda t}dt=e^{-\lambda(s+t)}$$

当给定条件$T\ge s$时，讨论$T\ge s+t$的条件概率，发现

$$P(T \geq s+t \mid T \geq s) = \frac{P(T \geq s, T \geq s+t)}{P(T \geq s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(T \geq t)$$

由结果知，此条件概率与 s 无关。这说明无论在泊松过程的什么时刻，即无论取哪一时刻为起点，考察至下一位顾客到达所经过的时间，其概率（即 $P(T \geq s+t \mid T \geq s)$）与此时刻之前的最后一位顾客是什么时候到达的无关。它和顾客相继到达的间隔时间都服从相同参数的负指数分布。

还可以指出的是，能满足无记忆性 $P(T \geq s+t \mid T \geq s) = P(T \geq t)$ 的分布，也只能是负指数分布。

下面讨论爱尔朗（Erlang）分布。

爱尔朗分布的密度函数是

$$p(t) = \begin{cases} \dfrac{\mu \cdot (\mu t)^{k-1}}{(k-1)!} e^{-\mu t} & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{9-12}$$

其中参数 $\mu > 0$，k 称为阶数 $(k = 0,1,2,\cdots)$。

当 $k = 1$ 时，则爱尔朗分布也就是负指数分布。若随机变量 τ 服从 k 阶爱尔朗分布，则 τ 的期望值和方差分别为

$$E(\tau) = \frac{k}{\mu}, D(\tau) = \frac{k}{\mu^2} \tag{9-13}$$

可以证明，如果 $\xi_1, \xi_2, \cdots, \xi_k$ 是 k 个相互独立具有相同的负指数分布（参数为 μ）的随机变量，则随机变量 $\tau = \xi_1 + \xi_2 + \cdots + \xi_k$ 服从 k 阶爱尔朗分布。

在排队论中，常把顾客到达间隔时间及接受服务时间与爱尔朗分布联系起来。如果顾客要连续接受串联的 k 个服务台服务，各服务台的服务时间相互独立，且服从相同的负指数分布（参数为 μ），那么顾客被这 k 个服务台服务完总共所需的时间就服从爱尔朗分布。这里应说明的是在顾客接受连续的服务时，只有当 k 个服务台都完成了对某个顾客的服务之后，下一个顾客才能进入这个串联服务系统。

9.1.3 生灭过程

生灭过程也是一种马尔柯夫过程。在排队论中很多排队过程和这个过程相仿。我们特别关心生灭过程在统计平衡时反映出来的稳态概率，并直接把这种稳态概率应用于建立各种排队模型。

1. 生灭过程的定义

一堆细菌，随时间推移，有的分裂为两个，有的死亡，经过一段时间之后，细菌变为多少？这种细菌的分裂与死亡的过程就是典型的生灭过程的例子。设每个细菌在 $\triangle t$ 时间内分裂成两个细菌的概率为 $\lambda \triangle t + 0(\triangle t)$；而在 $\triangle t$ 时间内死亡

的概率为 $\mu\triangle t+0(\triangle t)$，各个细菌在任一时间内分裂或死亡都是相互独立的。若将细菌的分裂或死亡都看成是随机事件，那么在 $\triangle t$ 时间内发生两个时间的概率等于下面三者之一：$(\lambda\triangle t+0(\triangle t))^2,(\mu\triangle t+0(\triangle t))^2,(\lambda\triangle t+0(\lambda))\cdot(\mu\triangle t+0(\triangle t))$。所以在 $\triangle t$ 时间内发生两个或两个以上事件的概率为 $0(\triangle t)$。又设在时刻 t 有 i 个细菌，则在时刻 $t+\triangle t$ 有 $i+1$ 个细菌的概率为 $\lambda_i\triangle t+0(\triangle t)$，在时刻 $t+\triangle t$ 内有 $i-1$ 个细菌的概率为 $\mu_i\triangle t+0(\triangle t)$。用 $\xi(t)$ 表示这堆细菌在时刻 t 的个数，并考虑在 $[0,T)$ 时间内细菌数的变化情况。因为对每个固定的 t，$\xi(t)$ 都是随机变量，故 $\{\xi(t)\big|t\in[0,T)\}$ 为一个随机过程，又因为它具有无后效性，故也是一个马尔柯夫过程，把上述细菌分裂和死亡的过程称为生灭过程。

定义：设 $\{\xi(t)\big|t\in[0,T)\}$ 为一个随机过程，随机变量 $\xi(t)$ 的取值集合为 $I=\{0,1,2,\cdots,m\}$（或可列集 $I=\{0,1,2,\cdots\cdots\}$），称此集合为状态集，设在时刻 t 时 $\xi(t)=j$，那么在时刻 $t+\triangle t$ 时，$\xi(t+\triangle t)=j+1$ 的概率为 $\lambda_j\triangle t+0(\triangle t)$，其中 $\lambda_j>0$ 为与 t 无关的常数；在时刻 $t+\triangle t$ 时，$\xi(t+\triangle t)$ 为 I 中其它元素的概率为 $0(\triangle t)$。满足上述随机过程 $\{\xi(t)\big|t\in[0,T)\}$ 称为生灭过程。

从上面的定义来看，如果把状态的变化理解为排队系统顾客的到达和离去，则在时间增量 $\triangle t$ 内，只会有一个到达，或有一个离去，或既无到达也无离去（此种情况的概率为 $1-\lambda_j\triangle t-\mu_j\triangle t-0(\triangle t)$），除此以外的其他到达、离去情况认为是不可能发生的。并且，在 $\triangle t$ 内到达一个顾客的概率和 $\triangle t$ 的长短有关，而与起始时刻 t 无关；到达一个的概率还与 t 时刻的顾客到达数 j 有关（因 λ_j 与 j 有关），而和 t 时刻以前的状态无关。同理，可进行 $\triangle t$ 时间内离开一个顾客的概率情况分析。

2. 生灭过程的稳态概率

进一步考察生灭过程中时间 $T=+\infty$ 的情况。对生灭过程我们更关心的是极限 $\lim\limits_{T=+\infty}P(\xi(t)=j)$。可以证明下面的定理。

定理 3 令

$$\pi_0=1,\pi_j=\frac{\lambda_0\lambda_1\cdots\lambda_j}{\mu_1\mu_2\cdots\mu_j}\qquad(j=1,2,\cdots)$$

并设生灭过程 $\{\xi(t)\big|t\geq0\}$ 的状态集为 $I=\{0,1,2,\cdots,m\}$ 或 $I=\{0,1,2,\cdots\}$，那么，当下列条件

$$\sum_{j=0}^{\infty}\pi_j<+\infty,\sum_{j=0}^{\infty}\frac{1}{\lambda_j\pi_j}=+\infty$$

满足时，对于任意正数 s 和任意 $j\in I,i\in I$，都有 $\lim\limits_{t\to+\infty}P(\xi(t)=j)=\lim\limits_{t\to+\infty}P(\xi(s+t)=j\big|\xi(s)=i)=P_j>0$。当状态集为有限集 $I=\{0,1,2,\cdots,m\}$ 时，

$$p_0 = (\sum_{j=0}^{\infty} \pi_j)^{-1}$$

$$p_j = \pi_j p_0 = \frac{\lambda_{j-1}}{\mu_j} p_{j-1} \qquad j = 1,2,\cdots,m$$

当状态集为无限集 $I = \{0,1,\cdots\}$ 时

$$p_0 = (\sum_{j=1}^{\infty} \pi_j)^{-1}$$

$$p_j = \pi_j p_0 = \frac{\lambda_{j-1}}{\mu_j} p_{j-1} \qquad j = 1,2,\cdots$$

我们称 $P_j (j = 0,1,2,\cdots)$ 为生灭过程在统计平衡时的概率，或称为稳态概率。

对于生灭过程，因为时刻 t 时状态为 j 的概率分布 $P(\xi(t) = j) \quad j \in I$，将随时间 t 的变化而变化，称之为瞬态解。瞬态解是很难求出的，即使求出来也难以利用。因此，在实际应用中，我们更关心的是稳态概率 $P_j, j \in I$。稳态概率的含义是说，当运行时间 t 无限长时，状态的概率分将不随时间而变化，此时状态为 j 的概率 P_j 是一个常数。需要指出，虽然在理论上由定理 3 知，生灭过程需经无限长的时间才会进入稳态。但在实际应用中，对大多数问题来说，相应的生灭过程总会很快达到稳态，不需要等到 $t \to +\infty$ 之后。

9.2 一般排队系统结构

9.2.1 排队模型结构

排队系统（或称为服务系统）可用于描述排队情况。即顾客为了获得某种服务而到达服务台；若服务员在忙，顾客不能立即获得服务而又被允许排队等待，则加入等待队列；获得服务之后则立即离开系统。整个过程如图 9-1 所示。

实际上每个具体的排队系统可能有所不同，但一般的排队系统都具有三个要素，这就是输入过程；排队规则；服务机构。由这三个要素便可以构造出一个一般排队系统的结构模型。并由此得到描述系统运行情况的数学模型。下面对这三个要素再加以具体的讨论。

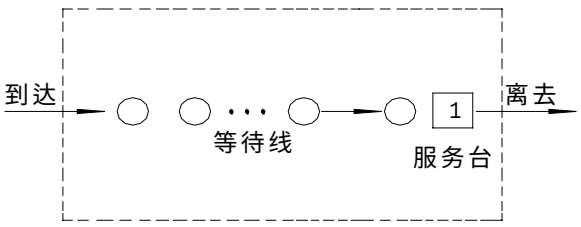

图 9-1

1. 输入过程

对于输入过程，我们需要了解的一个方面是顾客源的情况（顾客的总体可能是有限集，也可能是无限集），但主要的方面是了解顾客到达服务系统的规律，这种规律主要反映在顾客相继到达间隔时间的概率分布上。几种主要情况如下：

(1)定长输入。顾客有规律的到达，每隔时间α到达一位顾客。例如自动生产线上的装配件。顾客相继到达的间隔时间$\{T_n\}$的分布函数是：

$$P(T_n \le t)=\begin{cases}0 & t<\alpha\\ 1 & t\ge\alpha\end{cases}$$

(2)最简单流。即$\{T_n\}$中各个T_n相互独立且都服从同一负指数分布：

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

或者说，在$[0,t)$内到达的顾客数$N(t)$相应的随机过程是泊松过程，即$N(t)$的概率分布为：

$$P(N(t)=k)=\frac{(\lambda t)^k}{k!}e^{-\lambda t}\qquad k=0,1,2,\cdots$$

(3)k阶爱尔朗分布。即$\{T_n\}$中各个T_n相互独立，且都具有相同的爱尔朗分布密度：

$$p(t)=\begin{cases}\dfrac{\lambda(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda t} & t\ge 0\\ 0 & t<0\end{cases}$$

(4)一般独立分布。即中$\{T_n\}$各个T_n相互独立，且都具有相同的概率分布。

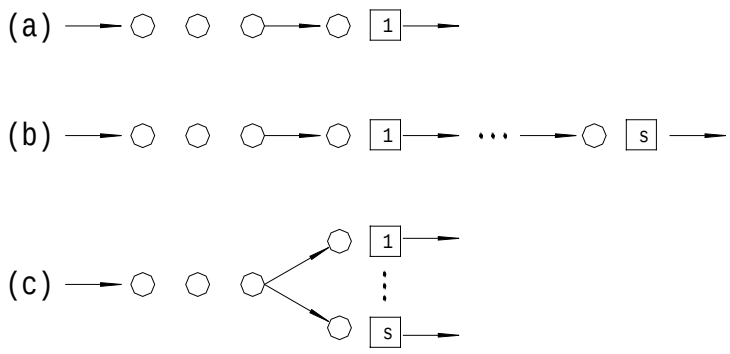

图 9-2

2. 服务机构

关于此要素需要明确的是：服务台的个数、服务台之间的串并联结构（图9-2）、服务台为每位顾客服务所需的时间τ_n的分布情况。

下面介绍几种常见的服务时间分布：

(1)定长分布　每位顾客的服务时间τ_n均为常数β,τ_n的分布函数为

$$P(\tau_n \le t)=\begin{cases}0 & \tau_n<\beta\\ 1 & \tau_n \ge \beta\end{cases}$$

(2)负指数分布　$\{\tau_n\}$中各个τ_n相互独立，且都服从相同的负指数分布

$$P(\tau_n \le t)=\begin{cases}0 & t<0\\ 1-e^{-\mu t} & t\ge 0\end{cases}$$

(3)k阶爱尔朗分布　$\{\tau_n\}$中各个τ_n相互独立，且都具有相同的k阶爱尔朗分布，其密度函数为

$$p(t)=\begin{cases}0 & t<0\\ \dfrac{\mu(\mu t)^{k-1}}{(k-1)!}e^{-\mu t} & t\ge 0\end{cases}$$

(4)一般独立分布　$\{\tau_n\}$中各个τ_n相互独立，且都具有相同的概率分布。

3. 排队规则

排队规则可描述到达的顾客按照怎样的顺序接受服务。在不同的实际问题中，排队规则是多样的。一般可分为损失制、等待制、混合制三类。

当一位顾客到达时，若所有的服务台均被占用，该顾客自动消失，具有这种特点的排队规则称为损失制。例如，一位顾客到达某一旅馆，如果已经客满，他就会离开这旅馆到别处投宿。

顾客到达时，若所有的服务台均被占用，顾客将排成队伍等待服务，具有这种特点的排队规则称为等待制。接受服务的次序一般采用先到先服务规则。也可以有其他规则，例如后到先服务、优先权先服务、随机服务（选取等待队列中任一顾客进行服务）等等。后面讨论的排队模型都采用“ 先到先服务 ”的规则。

混合制是损失制和等待制兼而有之的情况。假定服务系统的容量有限，最多只能容纳k个顾客（包括等待和正在接受服务的顾客），那么当顾客到达时，发现服务系统已客满，该顾客将自动消失，否则就进入服务系统，这是一种情况。此外，还可以有顾客等待服务时间有限的情况，即当超过一定时间时，顾客将自动消失。

9.2.2 排队系统的数量指标

一个服务系统，一方面是如果服务机构过小而不能满足顾客的需要。就会产生拥挤现象并造成服务质量的下降。因此顾客希望机构大些好。另一方面，如果服务机构过大，则人力、物力等方面的开支要增加，并有可能造成资源的浪费，从这方面的分析来看，设置的服务机构过大未必能收到好的效果，因此希望机构小些。研究排队系统的目的就是要在顾客的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。

排队论研究的问题大体分为连两类。第一类是在服务机构设置之前，根据顾客输入过程与服务过程的要求，通过对排队系统从定性到定量的分析，确定系统的结构模型，建立估计实际系统性能的数量指标，在经综合分析后作出权衡决策，定出服务机构（如电信局、港口码头、机场跑道等）的规模；第二类问题是对已有的服务系统进行定性、定量及综合分析，进行追踪决策，研究对系统进行最优控制的策略，提高系统的服务质量，并使得系统在达到规定质量指标的前提下取得的经济效益最大。

研究上述的两类问题时，必须对系统进行定量的分析。在系统中能够反映系统结构及性能的可用于定量的主要数量指标如下：

队长——指系统中的顾客数（包括等待服务的顾客和正在接受服务的顾客）；

等待队长——系统中排队等待的顾客数。

逗留时间——顾客在系统中的排队等待时间与接受服务所用时间之和；

等待时间——顾客在排队等时所用时间。这些指标均为随机变量，且依赖于时刻　或顾客数。如前所述，我们主要不是研究瞬态过程的数量指标，只对统计平稳状态下的稳态数量指标感兴趣，所以在定量分析排队系统的性能时，通常是指分析系统在平衡状态下的性能。

为了定量分析一般的系统，我们又引入了下列表示排队模型的符号，用他们建立上述可数量化的指标的数学模型。

这些符号及其定义是：

λ——单位时间内平均到达的顾客数（平均到达率）。

$\frac{1}{\lambda}$——平均到达间隔时间。

μ——单位时间内受到服务的顾客平均数（平均服务率）。

$\frac{1}{\mu}$——每位顾客的平均服务时间。

S——服务台个数。

ρ——每个服务台的服务强度。

P_j——在统计平衡时，系统中有 j 个顾客的概率。

L——队长的期望值。

L_q——等待队长的期望值。

W——逗留时间的期望值。

W_q ——等待时间的期望值。

不同的排队系统具有不同的排队模型，因此需要用一些符号表达不同类型模型的主要特征。这些特征可列为 6 项：输入过程（顾客到达间隔时间的分布）：服务时间的分布；服务台的个数（多个服务台时，假设各个服务台是并联的，每个服务台只对单个顾客进行服务）；系统容量（服务台个数加上可容纳的等待顾客数）；顾客来源的总体数；排队规则。因此，我们可用符号表示不同的排队模型，例如，可将上述 6 个特征按顺序用各自的符号列出，并用“ / ”隔开，即

输入过程/服务时间分布/服务台个数/系统容量/顾客源数/排队规则。

由于本章讨论的问题都采用先到先服务规则，所以在模型的符号表示中不必再列出排队规则的符号。当系统容量或顾客源无限时，也将它们从模型的符号中省去。用以表示顾客到达时间间隔分布和服务分布的常用符号有：

M ——输入过程为最简单流，或服务时间为负指数分布。

D ——定长分布。

E_k —— k 阶爱尔朗分布。

G ——服务时间为一般独立分布。

例如 $M/M/1/1$ 表示输入过程为最简单流、服务时间为负指数分布、单服务台、系统容量为 1 的损失制排队模型；$M/G/S/k$ 表示输入过程为泊松流、服务时间为一般独立分布、 s 个服务台、系统容量为 k 的混合制排队模型。

对于损失制和混合制的排队系统，顾客在到达排队系统时，发现服务台无空或者系统容量已满，就自动消失而再不进入系统。因此，到达的顾客不一定全进入系统。为此，引入有效到达率 λ_e，它是指单位时间内平均进入系统的顾客人数，而到达率 λ 是指单位时间内平均来到系统的顾客人数(参看图 9-3)。

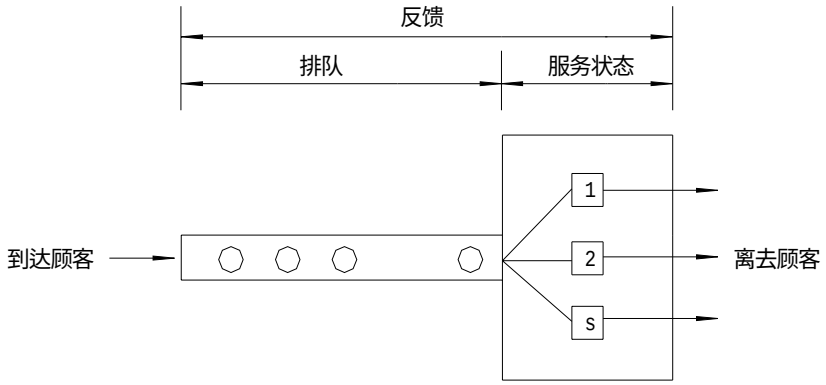

图 9-3

此处，可以指出，当排队模型处于平衡状态时，有下面的李特尔公式：

$$L_q = \lambda_e W_q \tag{9-14}$$

$$L = \lambda_e W \tag{9-15}$$

9.3 几个排队模型

本节将讨论几个具体的排队系统，这些系统的状态变化过程都属于生灭过程。本节将给出这些系统所对应的排队模型。这些模型的输入过程均为泊松流、服务时间为负指数分布，这些是最常见的排队模型，即单队、并列的多服务台模型的标准情况和特殊情况，利用生灭过程的稳态概率，可得到用以表达这些模型特征的数量指标。

9.3.1 $M/M/S$ 排队模型

本模型指的是这样的排队系统；顾客到达服从泊松分布（或到达间隔时间服从负指数分布）；服务时间服从负指数分布；并列的服务台的个数为 $S(S \geq 1)$ 个；系统容量和顾客源数无限；排队规则为先到先服务的等待制排队系统，所以 $M/M/S/\infty/\infty$ （先到先服务）模型的简写。

当 $S=1$ 时，相应的 $M/M/1$ 可看作 $M/M/S$ 的特殊情况，即 $M/M/1$ 是单服务台的等待制模型，如图 9-2（a）所示。当 $S>1$ 时，则 $M/M/S$ 为并列多服务台的排队模型，如图 9-2（c）及图 9-3 所示。

上面关于系统特点的描述很重要，不具备这些特点的模型不称为 $M/M/S$ 模型。例如当 S 个服务台仍为并联，但在每个服务台前都形成一个等待队列的排队系统，如图 9-4 所示，实际上是 S 个 $M/M/1$ 排队系统，而不是标准的 $M/M/S$ 模型。

下面先讨论与 $M/M/S$ 系统相应的生灭过程以及此生灭过程的稳态概率，这个概率是所需求出的数量指标之一。

设到达间隔时间 $\{T_n\}$ 服从的负指数分布为

$$P(T_n \leq t) = \begin{cases} 0 & t<0 \\ 1-e^{-\lambda t} & t \geq 0 \end{cases}$$

服务时间 $\{\tau_n\}$ 服从的负指数分布为

$$P(\tau_n \leq t) = \begin{cases} 0 & t<0 \\ 1-e^{-\mu t} & t \geq 0 \end{cases}$$

其中参数 λ 和 μ 如前所述，就是平均到达率和平均服务率。

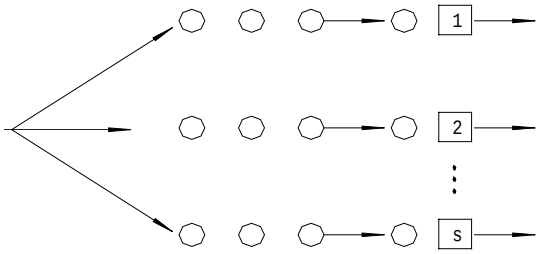

图　9.4

用$\xi(t)$表示在时刻t系统内具有的顾客数，它是随机变量，故$\{\xi(t)\}$是一个随机过程，其状态集为$I=\{0,1,2,\cdots\}$。这个随即过程可看作是一个生灭过程（视顾客到达为“生”，离去为“灭”）。因此输入过程为λ的最简单流（具有独立增量性和平稳性），故在$\triangle t$时间内到达一位顾客的概率可写成

$$P(\xi(\triangle t)=1)=\frac{\lambda\triangle t}{1!}e^{-\lambda\triangle t}=\lambda\triangle t(1-\lambda\triangle t+0(\triangle t))$$

$$=\lambda\triangle t+0(\triangle t)$$

在输入过程中负指数分布和泊松分布的等价性，对于具有负指数分布的服务时间τ_n及相应的服务过程也具有此种性质。因此，在$\triangle t$时间内一个服务台对一个顾客结束服务的概率为$\mu\triangle t+0(\triangle t)$。因为系统有$S$个服务台，所以对整个系统来说，在$\triangle t$时间内结束对一位顾客服务的概率为$S\mu\triangle t+0(\triangle t)$。如果再作进一步的分析，可推知此生灭过程有如下结果。

$$\lambda_j=\lambda \qquad j=0,1,2,\cdots$$

$$\mu_j=\begin{cases}j\mu & j=1,2,\cdots,S\\ S\mu & j=S+1,S+2,\cdots\end{cases}$$

我们关心的是此生灭过程的稳态概率P_j。现在考虑定理 3 的条件。

因此

$$\pi_j=\begin{cases}\dfrac{S_j}{j!}\rho^j & 1\le j\le S\\ \dfrac{S^s}{S!}\rho^j & j>S\end{cases} \tag{9-16}$$

其中

$$\rho=\frac{\lambda}{S\mu} \tag{9-17}$$

因此

$$\sum_{j=0}^{\infty}\pi_j=\sum_{j=0}^{S}\frac{S^j}{j!}\rho^j+\sum_{j=S+1}^{s}\frac{S^s}{S!}\rho^j$$

$$\sum_{j=0}^{\infty}\frac{1}{\lambda_j\pi_j}=\sum_{j=0}^{s}\frac{1}{\lambda\pi_j}+\sum_{j=S+1}^{\infty}\frac{1}{\lambda\pi_j}$$

$$=\frac{1}{\lambda}\sum_{j=0}^{s}\frac{j!}{\rho^j S^j}+\frac{S!}{\lambda S^S}\sum_{j=S+1}^{\infty}\frac{1}{\rho^j}$$

故当$\rho<1$时，$\sum_{j=0}^{\infty}\pi_j<+\infty,\sum_{j=0}^{\infty}\frac{1}{\lambda_j\pi_j}=+\infty$

定理 3 的条件得到满足，于是当$\rho<1$时，由定理 3 的结论便得到了此系统的稳态概率

$$P_0=(\sum_{j=0}^{\infty}\pi_j)^{-1}=(\sum_{j=0}^{S-1}\frac{(S\rho)^j}{j!}+\frac{(S\rho)^S}{S!}\cdot\frac{1}{1-\rho})^{-1} \tag{9-18}$$

$$P_j=\pi_j P_0=\begin{cases}\dfrac{(S\rho)^j}{j!}P_0 & 1\le j\le S\\ \dfrac{S^S\rho^j}{S!}P_0 & j>S\end{cases} \tag{9-19}$$

有了$M/M/S$模型的稳态概率$P_j(j=0,1.2,\cdots)$这个指标，便可得到系统在平衡状态下的队长期望值L，等待队长期望值L_q以及W,W_q等指标的计算公式。

由L及L_q的定义，可得

$$L=\sum_{j=0}^{\infty}j\cdot p_j=S\rho+\frac{\rho}{(1-\rho)^2}\cdot p_s \tag{9-20}$$

$$L_q=\sum_{j=0}^{\infty}jp_{S+j}=\frac{\rho}{(1-\rho)^2}p_S \tag{9-21}$$

另两个数量指标是W和W_q。因为排队系统处于平衡状态，由李特尔公式（9.14），（9.15）以及模型的等待制，知到达系统的顾客将全部进入系统。即$\lambda_e=\lambda$，所以

$$L=\lambda_e\cdot W=\lambda\cdot W$$
$$L_q=\lambda_e\cdot W_q=\lambda\cdot W_q$$

故
$$W=\frac{1}{\lambda}L=\frac{1}{\mu}+\frac{\rho}{\lambda(1-\rho)^2}p_s \tag{9-22}$$

$$W_q=\frac{1}{\lambda}L_q=\frac{\rho}{\lambda(1-\rho)^2}\cdot p_S \tag{9-23}$$

特别的，当$S=1$时，对于$M/M/1$模型的指标，当$\rho=\frac{\lambda}{\mu}<1$时，由公式（9-8）

得 $p_0=(1+\frac{\rho}{1-\rho})^{-1}=1-\rho$ 故有

$$\begin{cases} p_0=1-\rho \\ p_j=(1-\rho)\cdot\rho^j \qquad j=1,2,\cdots \end{cases} \tag{9-24}$$

且

$$L_q=\frac{\lambda^2}{\mu(\mu-\lambda)}=\frac{\rho^2}{1-\rho} \tag{9-25}$$

$$L=\frac{\lambda}{\mu-\lambda}=\frac{\rho}{1-\rho} \tag{9-26}$$

$$W_q=\frac{\lambda}{\mu(\mu-\lambda)}=\frac{\rho}{\mu(1-\rho)} \tag{9-27}$$

$$W=\frac{1}{\mu-\lambda}=\frac{1}{\mu(1-\rho)} \tag{9-28}$$

例 9.3 下列数据是到达邮局的顾客数和对顾客服务时间的统计结果如表 2。以每 3 分钟为一个时段，统计了 100 个时段中顾客到达的情况以及对 100 位顾客服务的时间。

表 2

到达数	0	1	2	3	4	5	6
时段数	14	27	27	18	9	4	1
服务时间（秒）	0～12	12～24	24～36	36～48	48～60	60～72	72～84
顾客人数	33	22	15	10	6	4	3
服务时间（秒）	84～96	96～108	108～120	120～150	150～180	180～200	
顾客人数	2	1	1	1	1	1	

设此服务系统是一个 $M/M/1$ 模型排队，试求顾客到达邮局后，需要排队等待服务的概率、不需等待的概率以及其它数量指标。

解：先求出每时段内到达顾客的平均数

$$\frac{0\times14+1\times27+2\times27+3\times18+4\times9+5\times4+6\times1}{100}=1.97$$

因每 3 分钟为一个时段，所以顾客平均到达率为 $\lambda=\frac{1.97}{3}\doteq0.657$（顾客/分）再计算每位顾客所需的平均服务时间 $\frac{1}{\mu}$，表格中服务时间采用组中值，可得

$$\frac{1}{\mu}=\frac{1}{100}(6\times33+18\times22+30\times15+42\times10+54\times6+66\times4+78\times3+90\times2+102\times1$$

$$+114\times1+135\times1+165\times1+190\times1)$$

$=31.72$ （秒）$\doteq0.529$ （分）

故此排队系统的平均服务率和服务台的服务强度分别为

$$\mu=1.89 \text{ （顾客/分）} \qquad \rho=\frac{\lambda}{\mu}=0.348$$

利用公式（9.24）~（9.28）可得系统处于统计平衡状态时的下列指标：

系统中有零位顾客的概率

$$P_0=1-\rho=1-0.348=0.652$$

P_0 也就是顾客到达后不需等待的概率。所以顾客到达后需排队等待服务的概率为

$$1-P_0=P=0.348$$

系统中期望的顾客数（等待队长的期望值）

$$L_q=\frac{\lambda^2}{\mu(\mu-\lambda)}=\frac{\rho^2}{1-\rho}=0.186 \text{ （位顾客）}$$

顾客在系统中期望逗留时间

$$W=\frac{1}{\mu-\lambda}=0.811 \text{ （分）}$$

顾客在系统中期望等待服务时间

$$W_q=\frac{\lambda}{\mu(\mu-\lambda)}=0.282 \text{ （分）}$$

例 9.4 设某医院的外科换药室有 2 位当班护士，换药者的到达过程可认为是一个泊松过程，换药时间服从负指数分布。如果平均每小时有 20 位病人来换药，每位病人的换药时间平均为 5 分钟。试求每一位病人到换药平均花费时间，并问是否应增加一位当班护士。

解：现在的排队系统为 $M/M/2$ 模型，需求出逗留时间的期望值 W。由题意可知

$$\lambda=\frac{20}{60}=\frac{1}{3}\text{（顾客/分）}\quad\text{而}\frac{1}{\mu}=5\text{（分钟）}$$

故 $$\mu=\frac{1}{5}\text{（顾客/分）}$$

而 $$\rho=\frac{\lambda}{S\cdot\mu}=\frac{\lambda}{2\cdot\mu}=\frac{5}{6}$$

计算W需先求出$P_S=P_2$，由公式（9.18）知

$$P_0=(\sum_{j=0}^{2-1}\frac{(2\rho)^j}{j!}+\frac{(2\rho)^2}{2!}\cdot\frac{1}{1-\rho})^{-1}$$

$$=(1+\frac{10}{6}+\frac{1}{2}\cdot(\frac{10}{6})^2\cdot 6)^{-1}=\frac{1}{11}$$

代入公式（9.19）,可得

$$P_j=P_s=P_2=\frac{(2\rho)^2}{2!}P_0=\frac{1}{2}(\frac{10}{6})^2\cdot\frac{1}{11}\doteq 0.126$$

将P_2代入公式（9.22），可得

$$W=\frac{1}{\mu}+\frac{\rho}{\lambda(1-\rho)^2}\cdot p_2=5+3\times 6^2\times\frac{5}{6}\times 0.126$$

$$=16.34\text{（分）}$$

如果增加一位护士，则系统为$M/M/3$模型，同理可求得

$$P_0=0.045$$
$$P_s=P_3=0.117$$

$$W=15.53\text{（分）}$$

由结果看出，增加一位护士后，并没有明显缩短一位病人在换药平均花费时间，所以不需要增加一位护士。

下面对$M/M/S$排队模型$(S\geq 2)$与S个$M/M/1$模型比较。

例 9.5 不妨设$S=2$时，比较$M/M/2(\lambda=5,\mu=4)$与两个$M/M/1$模型。也就是要比较排一个队列及排两个队列两者服务效果的优劣。

解：排一个队列的情况指一个$M/M/2$模型。关于此模型的各数量指标可由公式（9.18）~（9.23）求得，如表 9-3 所示。如果每个服务台前都各排一队，且进入队列后便坚持不换，到达的顾客以各$\frac{1}{2}$的概率加入到每个队中，这就是排两个队的情况。这时，原来的$M/M/2$系统变化成了两个独立的$M/M/1$系统，并且每个队列的平均到达率λ_1、λ_2为$\lambda_1=\lambda_2=\frac{\lambda}{2}=\frac{5}{2}$。对$M/M/1$模型的主要指标

可由公式（9.25）~（9.28）求得，如表 9-3 所示。

比较表 9-3 中各指标，除L外，其余的指标均显示出排一个队（1）比排两个队（2）效果要好。在所列的 4 个指标中，顾客在系统中的平均逗留时间W，是衡量服务质量优劣的一个重要指标。在两种排队方式下，排两个队时的每个$M/M/1$的W为 0.6667 比排一个队时的$M/M/2$的$W=0.4017$要大。

表 9-3

模型	λ	μ	L	L_p	W	W_p
(1) $M/M/2$	5	4	2.0086	0.7586	0.4017	0.1517
(2) $M/M/1$	2.5	4	1.6667	1.0417	0.6667	0.4167

比较的结果说明，尽管都是设置两个服务台，但由于采用不同的排队模型，其效果是不一样的。采用集中使用的方案（从而形成多服务台排队系统）要优于采用分散使用的方案（包括形式上在一起，实际上是分散使用的方案），在经济术语中称之为“规模收益”。在考虑服务设施的布局与使用时，需要注意这一因素。

9.3.2 $M/M/S/k$ 排队模型

$M/M/S/k$模型为顾客到达间隔时间和服务时间均服从负指数分布、S个服务台、系统容量为k（有限个）的排队系统。当$k=S$时为损失制排队系统；当$k>S$时为混合制排队系统。如果顾客到达排队系统时，系统内已有k个顾客，那么这位顾客就自动消失。下面确定此系统数量指标的关系式。

用$\xi(t)$表示在时间t系统内的顾客数，则$\{\xi(t)\}$是一个随机过程，其状态集是有限集$I=\{0,1,2,\cdots,k\}$。，可证明$\{\xi(t)\}$也是一个生灭过程，并有

$$\lambda_j=\lambda \qquad j=0,1,2,\cdots,k-1$$

$$\mu_j=\begin{cases} j\mu & j=1,2,\cdots,s \\ S\mu & j=S+1,s+2,\cdots,k \end{cases}$$

λ和μ的意义同前，与$M/M/S$类似，由定理（9.3）便可得到系统的稳态概率

$$p_0=\left(\sum_{i-0}^{S-1}\frac{(SP)^i}{i!}+\sum_{i=S}^{k}\frac{S^3\rho^i}{S!}\right)^{-1} \tag{9-29}$$

$$p_j=\begin{cases} \dfrac{(SP)^j}{j}p_0 & j=1,2,\cdots,S \\ \dfrac{S^s\rho^j}{S!}p_0 & j=S+1,S+2,\cdots,k \end{cases} \tag{9-30}$$

其中$\rho=\dfrac{\lambda}{S\mu}$，对于容量有限的模型不必有$\rho<1$的限制，因为不可能出现无限长的排队情况，$L$和$L_q$可由其含义直接计算，有

$$L=\sum_{j=0}^{k}jp_j \tag{9-31}$$

$$L_q=\sum_{j=0}^{k-s}jp_{S+j} \tag{9-32}$$

顾客到达而能进入系统的概率为$1-P_k$，故系统的有效到达率为

$$\lambda_e=\lambda(1-P_k) \tag{9-33}$$

再根据李特尔公式便可求得W和W_q。

特别的，当$S=1$时，可得$M/M/1/k$模型的数量指标如下：

$$P_0=\begin{cases}\dfrac{1-\rho}{1-\rho^{k+1}} & \rho\neq 1\\ \dfrac{1}{k+1} & \rho=1\end{cases} \tag{9-34}$$

$$P_j=\begin{cases}\dfrac{\rho^j(1-\rho)}{1-\rho^{k+1}} & \rho\neq 1\\ \dfrac{1}{k+1} & \rho=1\end{cases} \tag{9-35}$$

$$L=\begin{cases}\dfrac{\rho(1-(k+1)\cdot\rho^k)+k\rho^{k+1}}{(1-\rho)(1-\rho^{k+1})} & \rho\neq 1\\ \dfrac{k}{2} & \rho=1\end{cases} \tag{9-36}$$

$$L_q=\begin{cases}\dfrac{\rho^2}{1-\rho}-\dfrac{(k+\rho)\rho^{k+1}}{1-\rho^{k+1}} & \rho\neq 1\\ \dfrac{k(k-1)}{2(k+1)} & \rho=1\end{cases} \tag{9-37}$$

例 9.6　某电话问讯处设备有 6 部电话，平均每分钟有 4 次问题讯电话（包括接通的和未接通的），问讯到达服从泊松分布，每次通话的平均时间为 0.5 分钟，通话时间服从负指数分布。试问打到问讯处的电话能接通的概率为多少？

解：由题意知，这是一个$M/M/6/6$排队系统的模型。当$k=S=6$时，这是一个损失制系统。并且平均到达率$\lambda=4$，因平均服务时间

$\frac{1}{\mu}=0.5$　　所以平均服务率μ=2

故服务强度$\rho=\frac{\lambda}{S\mu}=\frac{4}{6\times 2}=\frac{1}{3}$

将ρ及S代入公式（9.29）和（9.30）得

$$P_0=\left(\sum_{i=0}^{5}\frac{(6\times\frac{1}{3})^i}{i^{i!}}+\frac{(6\times\frac{1}{3})^6}{6!}\right)^{-1}=0.136$$

$$P_k=P_j=P_6=\frac{(S\rho)^j}{j!}P_0=\frac{(6\times\frac{1}{3})^6}{6!}\cdot P_0=0.012$$

因为当$k=S=6$时，6 个电话均被用，因此，新到达的问讯电话不能被接通，所以$P_k=P_6$就是新来到的电话未能被接通的概率。对于服务系统来说，有称作损失概率，它是损失制系统的重要指标。因此，问讯电话能被接通的概率为

$$1-P_6=0.988$$

对于损失制系统由公式（9.33）顺便还可以求出系统的有效到达率为

$$\lambda_e=\lambda(1-P_6)=4\times 0.988=3.95$$

例 9.7　某加油站上有一个加油泵，而且场地很小，最多只能停放 5 辆汽车（包括正在加油的一辆）。设汽车到达间隔和加油时间都服从负指数分布，平均每 2 分钟到达一辆汽车，每辆汽车加油平均需要 2 分钟。试问：来加油的汽车到达加油站时能立即加油的概率为多少？加油站场地有空的概率为多少？实际加油的汽车在加油站平均逗留的时间是多少？

解：此为一个$M/M/1/5$排队模型。

由题意知$k=5,S=1,\frac{1}{\lambda}=2$(分)，$\frac{1}{\mu}=2$（分）

则$\lambda=\frac{1}{2}$（辆/分），$\mu=\frac{1}{2}$（辆/分），$\rho=\frac{\lambda}{\mu}=1$

即平均服务强度为 1。

由公式（9.34）和（9.35）中 ρ=1 的情况

可知，系统的稳态概率为

$P_0=\frac{1}{k+1}=\frac{1}{6}$，另外 $P_1=P_2=P_3=P_4=P_5=\frac{1}{k+1}=\frac{1}{6}$

其中 P_0 为系统中无顾客的概率，P_5 为系统中无空位的概率。

对于混合制系统 $(k>S)$，由公式（9.33），因为有效到达率为

$$\lambda_e=\lambda(1-P_k)=\lambda(1-P_5)=\frac{1}{2}\times(1-\frac{1}{6})=\frac{5}{12}$$（辆/分钟）

再用李特尔公式及公式（9.35）中 ρ=1 的情况，便可求得实际加油的汽车在加油站平均逗留的时间

$$W=\frac{L}{\lambda_e}=\frac{\frac{k}{2}}{\lambda_e}=6$$（分钟）

所以可立即加油的概率为 P_0=0.2，发现场地有空的概率为 $1-P_5=\frac{5}{6}$，平均逗留时间为 6 分钟，

9.3.3 $M/M/S/m/m$ **排队模型**

前面讨论的排队模型中，顾客源都是无限的。现在要讨论的是一种更特殊的情况，即排队系统具有 S 个服务台，但顾客源为有限数 m，其他规定与 $M/M/S/k$ 模型的规定同样，这里 $k=m$。这种系统中如果已有了 m 个顾客 ，就不会再有新的顾客到达，除非系统中的顾客得到服务的后又返回顾客源，系统才可能有顾客继续到来。其模型如图 9-5 所示。

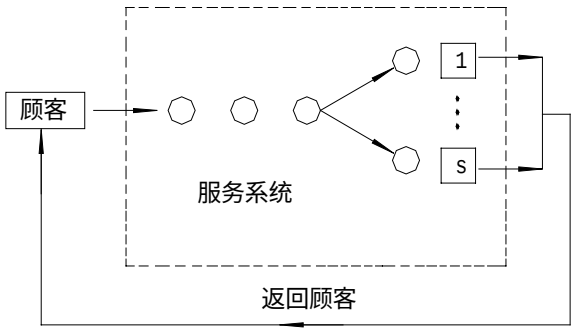

图 9-5

有限顾客源的服务系统的典型例子是机器看管问题（即机器因故障停机待修问题）。现在设有 S 个工人共同看管 m 台机器（$m\geq S,m$ 台机器就是顾客源数）。每当机器发生故障时，就立即有一位工人负责修理，使其恢复生产。所以，出故障的机器就是要求获得服务的顾客，即顾客到达就是机器出了故障；工人就是服务台。当 S 个工人分别在修理 S 台出故障的机器时间，新发生故障的机器就只能

等待工人修理。这里假定：

（1）每台机器连续正常运转的时间都服从参数为λ的负指数分布，每台机器平均连续运转时间为$\frac{1}{\lambda}$,λ就是一台机器在单位运转时间内发生故障的平均次数。

（2）每台机器的修复时间都服从参数为μ的负指数分布，工人修理机器的平均时间为$\frac{1}{\mu}$。

（3）每台机器在任何时段内连续运转的时间与工人修复时间彼此独立。

若以$\xi(t)$表示在时间t不在正常运转的机器数，它是一个随机变量，则$\{\xi(t)\}$为一随机过程，它的状态集 $I=\{0,1,\cdots,m\}$。可以证明$\{\xi(t)\}$是一个生灭过程，且有：

$$\lambda_j=(m-j)\lambda \qquad j=0,1,2,\cdots,m-1$$

$$\mu_j=\begin{cases} j\mu & j=1,2,\cdots,S \\ S\mu & j=S+1,S+2,\cdots,m \end{cases}$$

于是根据定理 3 便可以得到系统的稳态概率：

$$p_0=\left(\sum_{j=0}^{S}\binom{m}{j}\left(\frac{\lambda}{\mu}\right)^j+\sum_{j=S+1}^{m}\frac{j!}{S!S^{j-S}}\left(\frac{\lambda}{\mu}\right)^j\right)^{-1} \tag{9-38}$$

$$p_j=\begin{cases} \binom{m}{j}\left(\frac{\lambda}{\mu}\right)^j\cdot p_0 & j=1,2,\cdots,S \\ \binom{m}{j}\frac{j!}{S!S^{j-S}}\left(\frac{\lambda}{\mu}\right)^j\cdot p_0 & j=S+1,S+2,\cdots,m \end{cases} \tag{9-39}$$

再根据$L=\sum_{j=0}^{m}jP_j$和$L_q=\sum_{j=0}^{m-S}jP_{S+j}$得到$L$及$L_q$的公式。需注意的是，现在顾客平均到达率不是$\lambda$而是$\lambda_e=(m-L)$。应用李特尔公式还可导出求$W$及$W_q$的计算公式。

特别的，当$S=1$时，有下述单服务台系统的指标公式为：

$$p_0=\left(\sum_{i=0}^{m}\frac{m!}{(m-i)!}\left(\frac{\lambda}{\mu}\right)^j\right)^{-1} \tag{9-40}$$

$$p_j=\frac{m!}{(m-j)!}\left(\frac{\lambda}{\mu}\right)^j\cdot p_0 \tag{9-41}$$

$$L = m - \frac{\mu}{\lambda}(1 - p_0) \tag{9-42}$$

$$L_q = L - (1 - p_0) \tag{9-43}$$

例 9.8 4 名工人看管 10 台机器，每台机器平均每过半小时就要修理一次，每次修理需要 10 分钟。设机器连续运转时间和修理时间均服从负指数分布。试求机器发生故障不能马上得到修理的概率，以及 L, L_q, W, W_q 等数量指标。

解：按题意知，这是一个 $M/M/S/m/m$ 系统。

$S=4, m=10,\ \lambda=\frac{1}{30}, \mu=\frac{1}{10}, \rho=\frac{1}{3}$。

代入公式（9.38），（9.39）可得系统的稳态概率：

$$P_0 = \left[\binom{10}{0}\left(\frac{\lambda}{\mu}\right)^0 + \binom{10}{1}\left(\frac{\lambda}{\mu}\right)^1 + \binom{10}{2}\left(\frac{\lambda}{\mu}\right)^2 + \binom{10}{3}\left(\frac{\lambda}{\mu}\right)^3 + \binom{10}{4}\left(\frac{\lambda}{\mu}\right)^4 + \right.$$

$$\binom{10}{5}\frac{5!}{4!4}\left(\frac{\lambda}{\mu}\right)^5 + \binom{10}{6}\frac{6!}{4!4^2}\left(\frac{\lambda}{\mu}\right)^6 + \binom{10}{7}\frac{7!}{4!4^3}\left(\frac{\lambda}{\mu}\right)^7 +$$

$$\left.\binom{10}{8}\frac{8!}{4!4^4}\left(\frac{\lambda}{\mu}\right)^8 + \binom{10}{9}\frac{9!}{4!4^5}\left(\frac{\lambda}{\mu}\right)^9 + \binom{10}{10}\frac{10!}{4!4^6}\left(\frac{\lambda}{\mu}\right)^{10} \right]^{-1}$$

$$=(1+3.3333+5+4.4444+2.5926+1.2963+0.5401+0.1800+$$
$$0.0450+0.0075+0.0006)^{-1}$$
$$= (18.44)^{-1} = 0.0542$$

而 $P_1 = \binom{10}{1}\left(\frac{\lambda}{\mu}\right)^1 \cdot P_0 = 3.3333 P_0 = 0.1807$

同理 $P_2 = \binom{10}{2}\left(\frac{\lambda}{\mu}\right)^2 \cdot P_0 = 5 \cdot P_0 = 0.2710$

$P_3 = 0.2409$　　$P_4 = 0.1405$

$P_5 = 0.0703$　　$P_6 = 0.0293$

$P_7 = 0.0098$　　$P_8 = 0.0024$

$P_9 = 0.0075$　　$P_0 = 0.0004$

$P_{10} = 0.0006$　　$P_0 = 0.00003$

再计算 L 和 L_q

$$L=\sum_{j=0}^{10} jP_j = 2.6261$$

$$L_q=\sum_{j=0}^{10-4} j\cdot P_{4+j} = P_5+2P_6+3P_7+4P_8+5P_9+6P_{10}$$

$$= 0.1701$$

而顾客到达率为

$$\lambda_e=\lambda(m-L)=\frac{1}{30}(10-2.6261)=0.2458$$

由李特尔公式求得W及W_q得：

$$W=\frac{L}{\lambda_e}=\frac{2.6261}{0.2458}=10.684 \text{（分）}$$

$$W_q=\frac{L_q}{\lambda_e}=\frac{0.1701}{0.2458}=0.692 \text{（分）}$$

发生故障的机器不能马上获得修理的概率为

$$1-(P_0+P_1+P_2+P_3)=0.253$$

例 9.9 某企业有某种型号机器若干台，它们的连续工作时间服从同一参数λ的负指数分布。工人修理时间服从同一参数μ的负指数分布。设$\frac{\lambda}{\mu}=0.1$。现有两个方案：方案Ⅰ为 3 名工人各自独立看管机器，每人看管 6 台机器。方案Ⅱ为 3 名工人共同看管 20 台机器，当机器需要维修时，只要 3 名工人中有人空闲，就可以得到修理。试比较两个方案的优劣。

解：方案Ⅰ显然是形成了 3 个$M/M/1/6/6$排队模型。考察一下此模型的有关数量指标。

已知$\frac{\lambda}{\mu}=0.1$，由公式（9.40）及（9.41）可求得系统的稳态概率如下：

$$P_0=0.48,\qquad p_1=0.29,\qquad P_2=0.15$$

$$P_3=0.058,\qquad P_4=0.018,\qquad P_5=0.0035$$

$$P_6=0.0003.$$

由公式（9.42）及公式（9.43）得

$$L = 0.855\qquad L_q=0.335$$

而有效到达率为

$$\lambda_e = \lambda(m - L) = \lambda(6 - 0.855) = 5.145\lambda$$

由李特尔公式得方案Ⅰ的需修机器平均修理时间为

$$W_q(I) = \frac{L_q}{\lambda_e} = \frac{0.335}{5.145\lambda}$$

再来考察方案Ⅱ，这是一个$M/M/3/20/20$排队模型。$\frac{\lambda}{\mu} = 0.1$，可求得有关数量指标如下：

系统的稳态概率是

$P_0 = 0.14$，　$P_1 = 0.27$，　$P_2 = 0.26$，　$P_3 = 0.16$

$P_4 = 0.088$，　$P_5 = 0.047$，　$P_6 = 0.023$，　$P_7 = 0.011$

$P_8 = 0.0048$，　$P_9 = 0.0019$，　$P_{10} = 0.0007$，　$P_{11} = 0.0002$

$p_{12} = 0.0001$，而$p_{13}, p_{14}, ..., p_{20}$都几乎等于零。从而可得

$$L = \sum_{j=0}^{20} jp_j = 2.13$$

$$L_q = \sum_{j=0}^{20-3} jp_{3+j} = 0.337$$

在$M/M/3/20/20$中得有效到达率为

$$\lambda_e = \lambda(20 - 2.13) = 17.87\lambda$$

所以方案Ⅱ的机器等待修理时间

$$W_q(\text{Ⅱ}) = \frac{0.337}{17.87\lambda}$$

取两个方案的平均等待修理时间之比为

$$\frac{W_q(\text{Ⅰ})}{W_q(\text{Ⅱ})} = \frac{0.335}{5.145\lambda} \div \frac{0.337}{17.87\lambda} = 3.453$$

由比值可得方案Ⅱ的机器等待修理时间比方案Ⅰ的要小，所以方案Ⅱ优于方案Ⅰ。这可看作是前面所述关于集中使用的方案由于分散使用的方案这一结论又一例证。

9.3.4 一般服务时间的$M/G/1$排队模型

前面的三个排队模型均是泊松输入过程，服务时间服从负指数分布的排列系统。现在将服务时间服从的分布形式予以扩大，不作任何限制，即服从任何一个

分布均可，无论这种分布是否能写出分布函数的关系式。

$M/G/1$模型是单服务台的等待排队系统，输入过程为泊松过程，而各顾客的服务时间是相互独立且具有相同分布的随机变量。

假定输入过程是以λ为参数的泊松过程，服务时间τ_n期望值和方差分别$E(\tau_n)=\frac{1}{\mu}$和 $D(\tau_n)$，服务强度为$\rho=\frac{\lambda}{\mu}$。当$\rho<1$时，可得稳态下的结论：

$$P_0=1-\rho \tag{9-44}$$

$$L=\rho+\frac{\rho^2+\lambda^2 D(\tau_n)}{2(1-\rho)} \tag{9-45}$$

$$L_q=\frac{\rho^2+\lambda^2 D(\tau_n)}{2(1-\rho)} \tag{9-46}$$

应用李特尔公式（9.14）及（9.15）还可得稳态下的

$$W=\frac{L}{\lambda} \qquad W_q=\frac{L_q}{\lambda}$$

例 9.11 设做每套西服需要依次经过四道工序。某服装店只有一位做西服的服装师傅。设顾客前来定制西服的过程为泊松过程，平均每周来到 7 人（每人定制一套西服，且设每周工作 6 天，每天工作 10 小时)。每道工序所需时间服从同参数的负指数分布，平均需要 2 小时。试问一位顾客从订货到做好一套西服平均需要多少时间？

解：设 $X_n^1, X_n^2, X_n^3, X_n^4$ 分别是裁缝为第n位顾客缝制西服时在四道工序上所花的时间。它们是相互独立且服从相同参数的负指数分布的随机变量。故为第n位顾客缝制一套西服所需的总时间 $\tau_n=X_n^1+X_n^2+X_n^3+X_n^4$是服从4阶爱尔朗分布的随机变量。

由题意知，随即变量$X_n^i(i=1,2,3,4)$的分布函数为

$$P(X_n^i\le t)=\begin{cases}1-e^{-1/2t} & t\ge 0\\ 0 & t<0\end{cases}$$

再由k阶爱尔朗分布中随机变量的期望值及方差公式（9.13）得

$$E(\tau_n)=\frac{4}{\frac{1}{2}}=8 \qquad D(\tau_n)=\frac{4}{\left(\frac{1}{2}\right)^2}=16$$

因此，每位顾客的平均服务时间为$\frac{1}{\mu}=E(\tau_n)=8$。另外，由题意还知

$$\lambda = \frac{7}{60} = 0.11667$$

则 $$\rho = \frac{\lambda}{\mu} = \frac{7}{60} \times 8 = 0.933$$

由公式（9.45）可得平均队长为

$$L = 0.933 + \frac{(0.933)^2 + \frac{7}{60} \times 16}{2 \times (1 - 0.933)} = 9.0544$$

最后由李特尔公式

可知每位顾客为得到一套西服平均所需时间为 $W = \frac{L}{\lambda} = \frac{9.0544}{0.1167} \doteq 77.6$ (小时)

即大约为 1.3 周。