第 5 章　排队论

5.1　排队服务系统的基本概念

排队论 (queueing theory), 又称为随机服务系统, 是通过研究各种服务系统在排队等待现象中的概率特性, 从而解决服务系统最优设计与最优控制的一门学科, 是运筹学的一个重要分支. 排队论的最早发展是与电话、电讯中的问题联系在一起的. 1909 年, 丹麦电话工程师 A. K. Erlang 发表著名的论文 “概率与电话通话理论”, 开创了排队论研究的历史. 最近几十年来, 排队论已广泛应用在计算机网络、陆空交通、机场管理、通讯及其他公用事业等领域.

5.1.1　引　言

在日常生活中, 经常可以碰到一个服务系统在工作过程中产生的排队等待现象. 如进入机场上空的飞机等候降落、发生故障的机器等候工人修理等等. 如果进一步把服务系统的含义扩展, 则进入水库的流水等待开闸泄洪、通信系统的报文在缓冲器上等候传送也都可看做是排队等待的现象. 将具有排队等待现象的服务系统称为排队服务系统.

一个排队服务系统是由顾客和服务设施 (或服务台) 两个要素构成的. “顾客”是对要求得到服务的人或物的总称. 如进入机场上空要求降落的飞机、要求传送的通讯系统的报文等都是顾客; 凡是给予顾客服务的人或物统称为 “服务设施” (或服务台), 如供飞机降落使用的跑道、负责修理发生故障机器的工人等都是服务设施或服务台.

下面举一个例子说明排队论要研究的问题. 假设某火车售票处有一个售票员为陆续到达的乘客按先后次序进行售票服务. 显然顾客最关心的是排的队长不长, 等待时间要多久, 他们希望售票员人数能增加, 售票速度越快越好; 售票员则关心他需要连续工作多长时间才能使队伍不再存在而得到休息; 而火车站则不但要考虑顾客的要求, 还要考虑到车站管理在经济上的合理性. 因此, 在这样一个服务系统中, 不但需要研究队长、等待时间、忙期等数量指标的变化规

律, 而且要在满足顾客服务基本要求的条件下, 研究如何使机构运行更为经济的问题, 这就是排队论要研究的内容. 关于排队论方面的文献很多, 进一步的阅读可参见文献 [5, 19, 26, 29, 30] 等.

5.1.2 排队模型的描述

任何一个顾客通过排队服务系统总要经过如下过程: 顾客到达、排队等待、接受服务、离去 (见图 5.1). 于是, 任何排队服务系统可以描述为以下三个方面, 这就是: (1) 顾客到达规律; (2) 顾客排队与接受服务的规则; (3) 服务机构的结构形式、服务台的个数与服务速率. 下面将就以上三个因素分别进行介绍. 可参见文献 [30].

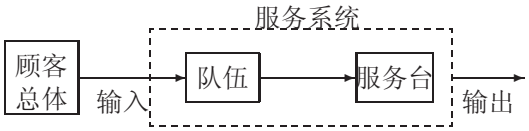

图 5.1 服务系统描述

(1) 输入过程

输入过程描述要求服务的顾客按怎样的规律到达系统, 可以从如下三个方面来刻画一个输入过程.

① 顾客总体 (或顾客源) 数可以是有限的或无限的. 如河流上游流入水库的水量可认为是无限的, 车间内停机待修的机器显然是有限的.

② 顾客到达方式可以是单个到达或是成批到达. 如在库存问题中, 若把进来的货看成顾客, 则为成批到达的例子.

③ 顾客相继到达时间的间隔, 这是刻画输入过程的最重要的内容. 令 T_n 为第 n 个顾客到达的时刻 ($n = 1, 2, \cdots$), 则有 $0 = T_0 \leqslant T_1 \leqslant \cdots \leqslant T_n \leqslant \cdots$. 记 $X_n = T_n - T_{n-1}$, 则 X_n 是第 n 个顾客与第 $n-1$ 个顾客到达的时间间隔. 通常假定 $\{X_n\}$ 是独立同分布的. 关于 $\{X_n\}$ 的分布, 排队论中常用的有以下几种:

(a) 定长输入 (D): 顾客相继到达时间的间隔为一确定的常数. 如产品通过传送带进入包装箱.

(b) 最简单流输入 (M) (或称 Poisson 流、Poisson 过程): 顾客相继到达时间的间隔 $\{X_n\}$ 独立同负指数分布, 其密度函数为

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

(c) Erlang 输入 (E_k): 顾客相继到达时间间隔 $\{X_n\}$ 相互独立, 并具有相同的 Erlang 分布密度函数

$$a(t)=\frac{\lambda(\lambda t)^{k-1}}{(k-1)!}e^{-\lambda t}\quad(t\geqslant 0)$$

其中 $\lambda>0$ 为一常数, k 为整数, 称为 Erlang 分布的阶. 容易算出此分布的均值, 即平均到达间隔为 $\frac{k}{\lambda}$.

(d) 一般独立输入 (GI): 顾客相继到达时间的间隔相互独立且同分布.

上面所有输入都是一般独立输入的特例.

(2) 排队规则

排队规则主要描述服务机构是否允许顾客排队, 顾客对排队长度、时间的容忍程度, 及在排队队列中等待服务的顺序. 常见的排队规则有如下几种情形:

① 损失制排队系统. 这种排队系统的排队空间为零, 即不允许排队. 当顾客到达系统时, 如果所有服务台均被占用, 则自动离去, 并假定不再回来. 如损失制电话系统.

② 等待制排队系统. 当顾客到达时, 若所有服务台都被占用且又允许排队, 则该顾客将进入队列等待. 服务台对顾客进行服务所遵循的规则通常有:

(a) 先来先服务 (FCFS): 按顾客到达的先后对顾客进行服务.

(b) 后来先服务 (LCFS): 在许多库存系统中会出现这种情形. 如钢板存入仓库后, 需要时总是从最上面的取出; 如在情报系统中, 后来到达的信息往往更加重要, 应首先加以分析和利用.

(c) 带优先服务权 (PS): 服务设施优先对重要性级别高的顾客服务, 在级别相同的顾客中按到达先后次序排队. 如病危的患者应优先治疗, 加急的电报电话应优先处理等.

(d) 随机服务 (SIRO): 到达服务系统的顾客不形成队伍, 当服务设施有空时, 随机选取一名顾客服务, 每一名等待顾客被选取的概率相等.

③ 混合制排队系统. 该系统是等待制和损失制系统的结合, 一般是指允许排队, 但又不允许队列无限长下去. 具体说来, 大致有三种:

(a) 队长有限, 即系统的等待空间是有限的. 例如最多只能容纳 K 个顾客在系统中, 当新顾客到达时, 若系统中的顾客数 (又称为队长) 小于 K, 则可进入系统排队或接受服务; 否则, 便离开系统并不再回来. 如旅馆的床位是有限的.

(b) 等待时间有限, 即顾客在系统中的等待时间不超过某一给定的长度 T, 当等待时间超过 T 时, 顾客将自动离去并不再回来. 如易损坏的电子元器件的库存问题, 超过一定存储时间的元器件被自动认为失效.

(c) 逗留时间 (等待时间与服务时间之和) 有限. 如用高射炮射击敌机, 当敌机飞越高射炮射击有效区域的时间为 t 时, 若在这个时间内未被击落, 也就不可能再被击落了.

不难注意到, 损失制和等待制可看成是混合制的特殊情形, 如记 s 为系统中服务台的个数, 则当 $K = s$ 时, 混合制即成为损失制; 当 $K = \infty$ 时, 即成为等待制.

(3) 服务机构

服务机构主要包括服务设施的数量、连接形式、服务方式及服务时间分布等. 服务设施的数量可以是一个或多个之分, 分别称为单服务台排队系统与多服务台排队系统; 多服务台排队系统的连接方式有串联、并联、混联和网络等; 服务方式分为单个或成批服务, 如公共汽车就一次装载大批乘客. 在这些因素中, 服务时间的分布更为重要一些, 故进一步说明如下: 记某服务台的服务时间为 V, 其分布函数为 $B(t)$, 密度函数为 $b(t)$, 则常见的分布有:

① 定长分布 (D): 每个顾客接受服务的时间是一个确定的常数.

② 负指数分布 (M): 每个顾客接受服务的时间相互独立, 具有相同的负指数分布

$$b(t) = \begin{cases} \mu e^{-\mu t}, & t \geqslant 0 \\ 0, & t < 0 \end{cases}$$

其中 $\mu > 0$ 为一常数, 易知 $\dfrac{1}{\mu}$ 为平均服务时间.

③ k 阶 Erlang 分布 (E_k): 每个顾客接受服务的时间服从 k 阶 Erlang 分布, 其密度函数为

$$b(t) = \frac{k\mu(k\mu t)^{k-1}}{(k-1)!} e^{-k\mu t} \quad (t \geqslant 0)$$

其中 $\mu > 0$ 为一常数, k 为正整数, 称之为 Erlang 分布的阶. 容易算出此分布的均值, 即平均服务时间为 $\dfrac{1}{\mu}$.

④ 一般服务分布 (G): 所有顾客的服务时间相互独立且有相同的一般分布函数 $B(t)$. 前面介绍的各种分布是一般分布的特例.

5.1.3 排队模型的符号表示

表示排队模型的记号是 20 世纪 50 年代初由 D.G.Kendall 引入的, 它大大简化了排队系统的描述. 采用如下记法:

$$\mathrm{A/B/C}/n$$

这里 A 记输入过程, B 记服务时间, C 记服务台数目, n 记等待空间数. 若 $n=\infty$, 即等待制时, 省去 n 而只用 A/B/C 记一个排队系统. 若无进一步的说明, 约定顾客源无限, 服务是按到达先后次序进行, 服务过程与输入过程独立. 如 M/M/n/K 代表顾客输入为 Poisson 流, 服务时间为负指数分布, 有 n 个并联服务站, 等待空间为 K 个的排队服务系统; D/G/1 代表定长输入, 一般服务时间, 单个服务站的排队服务系统; GI/E_k/1 代表一般独立输入, Erlang 服务时间分布, 单个服务台的排队服务系统, 如此等等.

5.1.4 排队系统的主要数量指标和记号

下面, 给出上述一些主要数量指标的常用记号:

$N(t)$: t 时刻系统中的顾客数 (又称为系统的状态), 即队长;

$N_q(t)$: t 时刻系统中排队的顾客数, 即排队长;

$w(t)$: t 时刻到达系统的顾客在系统中的逗留时间;

$w_q(t)$: t 时刻到达系统的顾客在系统中的等待时间.

上面给出的这些数量指标一般都是和系统运行的时间有关的随机变量, 求这些随机变量的瞬时分布一般是很困难的. 然而, 在许多情形下, 系统运行足够长的时间后将趋于统计平衡. 在统计平衡状态下, 队长的分布、等待时间的分布等都和系统所处的时刻无关, 而且系统的初始状态的影响也会消失. 因此, 在本章中将主要讨论与系统所处时刻无关的性质, 即统计平衡性质.

记 $P_n(t)$ 为时刻 t 时系统处于状态 n 的概率, 即系统的瞬时分布. 记 P_n 为系统达到统计平衡时处于状态 n 的概率. 又记

N: 系统处于平稳状态时的队长, 记均值 $L=E(N)$, 称为平均队长;

N_q: 系统处于平稳状态时的排队长, 记均值 $L_q=E(N_q)$, 称为平均排队长;

w: 系统处于平稳状态时顾客的逗留时间, 记均值 $W=E(w)$, 称为平均逗留时间;

w_q: 系统处于平稳状态时顾客的等待时间, 记均值 $W_q=E(w_q)$, 称为平均等待时间;

λ_n: 当系统处于状态 n 时新来顾客的平均到达率 (即单位时间内来到系统的平均顾客数);

μ_n: 当系统处于状态 n 时整个系统的平均服务率 (即单位时间内可以服务完的平均顾客数);

当系统中顾客的平均到达率 λ_n 为常数时, 记 $\lambda_n=\lambda$, 当系统中每个服务台的平均服务率为常数 μ, 则当 $n\geqslant s$ 时, 有 $\mu_n=s\mu$. 因此, 顾客相继到达的平均

时间间隔为 $\frac{1}{\lambda}$, 平均服务时间为 $\frac{1}{\mu}$. 令 $\rho = \frac{\lambda}{s\mu}$, 则 ρ 为系统的服务强度 (其中 s 为系统中并行的服务台数).

衡量一个排队系统工作状况的主要指标有:

(1) 系统中平均顾客数 (L) 或平均队长 (L_q). 这是顾客和服务机构都关心的指标, 在设计排队服务系统时也很重要, 因为它涉及到系统需要的空间大小.

(2) 顾客从进入到服务完毕离去的平均逗留时间 W (或顾客排队等待服务的平均等待时间 W_q). 每个顾客都希望这段时间越短越好.

(3) 忙期和闲期. 忙期定义为从顾客到达空闲服务机构开始到服务机构再一次变成空闲状态为止的时间. 它是衡量服务机构工作强度和利用效率的指标. 与忙期相对的是闲期, 闲期为服务机构空闲的时间长度. 在服务过程中, 忙期和闲期是相互交替出现的.

上述指标实际上反映了排队服务系统工作状态的几个侧面, 它们之间是互相联系、互相转换的. 设以 λ 表示单位时间内顾客的平均到达数, μ 表示单位时间内被服务完毕离去的平均顾客数, 则 $\frac{1}{\lambda}$ 表示相邻两个顾客到达的平均间隔时间, $\frac{1}{\mu}$ 表示对每个顾客的平均服务时间, 有

$$\begin{aligned} L &= \lambda W \qquad \text{或 } W = \frac{L}{\lambda} \\ L_q &= \lambda W_q \qquad \text{或 } W_q = \frac{L_q}{\lambda} \\ W &= W_q + \frac{1}{\mu} \end{aligned}$$

称 $L = \lambda W$ 和 $L_q = \lambda W_q$ 为 Little 公式, 是排队论中的一个非常重要的公式.

将前两式代入最后一式得到

$$L = L_q + \frac{\lambda}{\mu}$$

又由于

$$L = \sum_{n=0}^{\infty} nP_n, \quad L_q = \sum_{n=s+1}^{\infty} (n-s)P_n$$

因此只要求得 P_n 的值即可得 L, L_q 及 W 和 W_q. 当 $n = 0$ 时 P_n 值即为 P_0, 当 $s = 1$ 时, $(1 - P_0)$ 即是服务系统的忙期.

5.2 几种常见的分布函数

在组成一个排队服务系统的要素中, 由于输入与服务时间是随机的, 比较复杂, 因此抽出来单独研究.

5.2.1 Poisson 过程

Poisson 过程 (又称为 Poisson 流、最简单流) 是排队论中一种常用来描述顾客到达规律的特殊的随机过程, 与概率论中的 Poisson 分布有着密切联系.

所谓 Poisson 过程, 需同时满足如下四个条件:

(1) 平稳性. 指在一定时间间隔内, 来到服务系统有 k 个顾客的概率 $P_k(t)$ 仅与这段时间区间隔的长短有关, 而与这段时间的起始时刻无关. 即在时间区间 $[0,t]$ 或 $[a,a+t]$ 内, $P_k(t)$ 的值是一样的.

(2) 无后效性. 即在不相交的时间区间内到达的顾客数是相互独立的, 或者说在时间区间 $[a,a+t]$ 内来到 k 个顾客的概率与时刻 a 之前来到多少个顾客无关.

(3) 普通性. 指在足够小的时间区间内只能有一个顾客到达, 不可能有两个及两个以上顾客同时到达. 如用 $\phi(t)$ 表示在 $[0,t]$ 内有两个或两个以上顾客到达的概率, 则有 $\phi(t)=o(t)(t\to 0)$.

(4) 有限性. 任意有限时间内到达有限个顾客的概率为 1.

只要一个流具有平稳性、无后效性、普通性及有限性这四个性质, 就可以证明在 t 这段时间内有 $N(t)=k$ 个顾客来到服务系统的概率 $P_k(N(t))$ 服从 Poisson 分布, 即

$$P_k(N(t))=\frac{(\lambda t)^k}{k!}e^{-\lambda t}\quad (k=0,1,2,\cdots) \tag{5.1}$$

$N(t)$ 表示在时间区间 $[0,t)$ 内到达系统的顾客数. 详细的证明过程请参见相关文献 [5, 29, 31] 等.

容易求得, Poisson 分布的数学期望值为

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

特别地, 当 $t=1$ 时有 $E(N(1))=\lambda$, 因此, λ 可看成单位时间内到达顾客的平均数, 也称为到达率. 类似可求得, Poisson 分布的方差为 $Var(N(t))=\lambda t$.

另外, 容易求得 Poisson 过程有如下一些性质.

(1) 在 $[t,t+\Delta t]$ 时间内没有顾客到达的概率为

$$P_0(\Delta t)=e^{-\lambda\Delta t}=(1-\lambda\Delta t)+o(\Delta t)=1-\lambda\Delta t$$

(2) 在 $[t,t+\Delta t]$ 时间内恰好有一个顾客到达的概率为

$$P_1(\Delta t)=1-P_0(\Delta t)-\phi(\Delta t)=\lambda\Delta t$$

在实际问题中, 顾客到达系统的情况与 Poisson 过程是近似的, 比较容易处理, 因而排队论中大量研究的是 Poisson 输入的情况. 如对到达机修车间要维修的机器数可以认为是 Poisson 流; 电话局得到电话呼唤流的总和可以近似看做是 Poisson 流.

5.2.2 负指数分布

如果连续随机变量 T 服从参数为 λ 的负指数分布, 则其概率密度函数和分布函数分别为

$$f(t)=\begin{cases}\lambda e^{-\lambda t}, & t\geqslant 0\\ 0, & t<0\end{cases},\quad F(t)=\begin{cases}1-e^{-\lambda t}, & t\geqslant 0\\ 0, & t<0\end{cases}$$

负指数分布有如下性质:

(1) 当顾客的到达过程为参数 λ 的 Poisson 过程时, 则顾客相继到达的时间间隔 T 必服从负指数分布.

事实上, 如顾客到达形成 Poisson 流, 则在时间区间 $[0,t)$ 内至少有 1 个顾客到达的概率为 $1-P_0(t)=1-e^{-\lambda t}(t>0)$, 而这一事件的概率又可表示为

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

即顾客相继到达时间间隔服从负指数分布与输入过程为 Poisson 过程等价.

(2) 假设服务设施对每个顾客的服务时间服从负指数分布, 密度函数为 $f(t)=\mu e^{-\mu t}(t\geqslant 0)$, 则它对每个顾客的平均服务时间为 $\frac{1}{\mu}$.

证明 根据数学期望的定义, 可计算得

$$E(t)=\int_0^{\infty}tf(t)dt=\int_0^{\infty}t\mu e^{-\mu t}dt=-\frac{1}{\mu}\int_0^{\infty}e^{-\mu t}d(-\mu t)=\frac{1}{\mu}$$

易求得其方差为 $Var(t)=\frac{1}{\mu^2}$. 这里, 称 μ 为每个忙碌的服务台的平均服务率, 是单位时间内获得服务离开系统的顾客数的均值.

(3) 当服务设施对顾客的服务时间 t 为参数 μ 的负指数分布, 则有

① 在 $[t,t+\Delta t]$ 内没有顾客离去的概率为 $1-\mu\Delta t$;

② 在 $[t,t+\Delta t]$ 内恰好有一个顾客离去的概率为 $\mu\Delta t$;

③ 如果 Δt 足够小的话, 在 $[t,t+\Delta t]$ 内有多于两个以上顾客离去的概率为 $\phi(\Delta t)\to o(\Delta t)$.

(4) 负指数分布具有 “无记忆性”, 或称为 Markov 性. 如假设服务设施对顾客的服务时间服从负指数分布, 则不管对某一个顾客的服务已进行了多久, 剩下服务时间的概率分布仍为负指数分布. 即对任何 $t>0,\Delta t>0$ 有

$$P(T>t+\Delta t|T>\Delta t)=P(T>t)$$

证明 经计算, 有

$$\begin{aligned}P(T>t+\Delta t|T>\Delta t)&=\frac{P(T>\Delta t,T>t+\Delta t)}{P(T>\Delta t)}\\&=\frac{P(T>t+\Delta t)}{P(T>\Delta t)}=\frac{e^{-\mu(t+\Delta t)}}{e^{-\mu t}}\\&=e^{-\mu t}=P(T>t)\end{aligned}$$

在连续型分布函数中, “无记忆性”是负指数分布独有的特性.

(5) 若干独立同负指数分布的随机变量的最小值仍服从负指数分布. 设随机变量 $T_1,T_2,\cdots,T_n$ 相互独立且服从参数分别为 $\mu_1,\mu_2,\cdots,\mu_n$ 的负指数分布, 令 $U=\min\{T_1,T_2,\cdots,T_n\}$, 则随机变量 U 也服从负指数分布.

证明 对任意 $t\geqslant 0$, 有

$$\begin{aligned}P(U>t)&=P(\min\{T_1,T_2,\cdots,T_n\}>t)\\&=P(T_1>t,T_2>t,\cdots,T_n>t)\\&=P(T_1>t)P(T_2>t)\cdots P(T_n>t)=\exp\left(-\sum_{i=1}^{n}\mu_i t\right)\end{aligned}$$

即随机变量 U 服从参数为 $\mu=\sum\limits_{i=1}^{n}\mu_i$ 的负指数分布.

这个性质说明: 如果来到服务系统的有 n 类不同类型的顾客, 每类顾客来到服务台的间隔时间服从参数 μ_i 的负指数分布, 则作为总体来讲, 到达服务系统的顾客的间隔时间服从参数为 $\sum\limits_{i=1}^{n}\mu_i$ 的负指数分布; 如果一个服务系统中有 S 个并联的服务台, 且各服务台对顾客的服务时间服从参数 μ 的负指数分布, 则整个服务系统的输出就是一个具有参数 $S\mu$ 的负指数分布.

5.2.3 k 阶 Erlang 分布

若随机变量 ξ 的分布函数为

$$F(t)=\begin{cases}1-\sum\limits_{i=0}^{k-1}\dfrac{(\lambda t)^i}{i!}e^{-\lambda t}, & t\geqslant 0\\ 0, & t<0\end{cases}$$

其中 $\lambda > 0$ 为常数.

则称 ξ 服从 k 阶 Erlang 分布, 其概率密度函数为

$$f(t) = \frac{\lambda(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} \quad (t \geqslant 0) \tag{5.2}$$

易求得, 其数学期望和方差分别为 $E(\xi) = \frac{k}{\lambda}, Var(\xi) = \frac{k}{\lambda^2}$.

在 k 阶 Erlang 分布中, 若令 $E(\xi) = \frac{1}{\mu}$, 则 $\lambda = k\mu$. 此时 k 阶 Erlang 分布的密度函数为

$$f(t) = \frac{k\mu(k\mu t)^{k-1}}{(k-1)!} e^{-k\mu t} \quad (t \geqslant 0)$$

均值和方差分别为 $E(\xi) = \frac{1}{\mu}, Var(\xi) = \frac{1}{k\mu^2}$.

【定理 5.1】 设 $t_1, t_2, \cdots, t_k$ 是相互独立且服从参数 λ 的负指数分布的随机变量, 则 $\xi = t_1 + t_2 + \cdots + t_k$ 服从 k 阶 Erlang 分布, 其概率密度函数见 (5.2) 式.

定理的详细证明可参见文献 [32].

由以上定理可知, 对 k 个串联的服务台, 其服务时间 $t_1, t_2, \cdots, t_k$ 相互独立且都服从参数为 $k\mu$ 的负指数分布, 则总服务时间 $T = t_1 + t_2 + \cdots + t_k$ 服从 k 阶 Erlang 分布. 即一个顾客走完这 k 个串联的服务台 (台与台之间没有排队现象) 所需要时间服从 k 阶 Erlang 分布.

Erlang 分布比负指数分布具有更多的适应性. 当 $k = 1$ 时, Erlang 分布即为负指数分布; 当 k 增加时, Erlang 分布逐渐变为对称的. 事实上, 当 $k \geqslant 30$ 以后, Erlang 分布近似于正态分布. 当 $k \to \infty$ 时, 方差 $\frac{1}{k\mu^2}$ 将趋于零, 即为完全非随机的. 所以 k 阶 Erlang 分布可看成完全随机 ($k = 1$) 与完全非随机 ($k \to \infty$) 之间的分布, 能更广泛地适应于现实世界. 图 5.2 显示了一些 k 值对应的分布的形状.

最后, 指出要检验实际排队模型中顾客的到达或离去是否服从某一概率分布, 可采用统计学中的 χ^2 假设检验方法. 更进一步的阅读可参见文献 [33].

5.3 生灭过程

在排队论中, 如果 $N(t)$ 表示时刻 t 系统中的顾客数, 则 $\{N(t), t \geqslant 0\}$ 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, $\{N(t), t \geqslant 0\}$ 也是一类特殊的随机过程 —— 生灭过程.

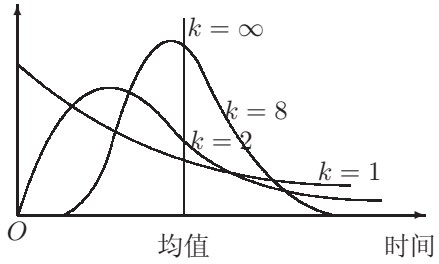

图 5.2 均值相同, 形状参数 k 不同的 Erlang 分布

【定义 5.1】 设 $\{N(t), t \geqslant 0\}$ 为一随机过程, 若 $N(t)$ 的概率分布有如下性质:

(1) 给定 $N(t)=n$, 到下一个生 (顾客到达) 的间隔时间服从参数 λ_n $(n=0,1,2,\cdots)$ 的负指数分布;

(2) 给定 $N(t)=n$, 到下一个灭 (顾客离去) 的间隔时间是服从参数 μ_n $(n=1,2,\cdots)$ 的负指数分布;

(3) 同一时刻只可能发生一个生或一个灭 (即同时只能有一个顾客到达或离去).

则称 $\{N(t), t \geqslant 0\}$ 为生灭过程.

由以上定义知, 生灭过程实际上是一特殊的连续时间 Markov 链, 即 Markov 过程. 根据上述 Poisson 分布同负指数分布的关系, λ_n 就是系统处于 $N(t)$ 时单位时间内顾客的平均到达率, μ_n 则是单位时间内顾客的平均离去率. 将上面几个假定合在一起, 则可用生灭过程的发生率来表示 (见图 5.3). 图 5.3 中箭头指明了各种系统状态发生转换的可能性. 在每个箭头边上注出了当系统处于箭头起点状态时转换的平均率.

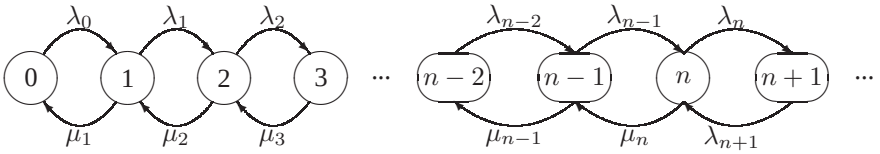

图 5.3 生灭过程的发生率图

除了少数特别情况以外, 要求出系统的瞬时状态 $N(t)$ 的概率分布是很困难的, 故下面只考虑系统处于稳定状态的情形. 先考虑系统处于某一特定状态

$N(t) = n(n = 0, 1, 2, \cdots)$. 从时刻 0 开始, 分别计算该过程进入这个状态和离开这个状态的次数,

$$E_n(t) = \text{到时刻 } t \text{ 之前进入状态 } n \text{ 的次数}$$
$$L_n(t) = \text{到时刻 } t \text{ 之前离开状态 } n \text{ 的次数}$$

因为这两个事件 (进入或离开) 是交替进行的, 因此进入和离开的次数或者相等, 或者相差一次, 即

$$|E_n(t) - L_n(t)| \leqslant 1$$

两边同时除以 t, 并令 $t \to \infty$, 则有

$$| \frac{E_n(t)}{t} - \frac{L_n(t)}{t} | \leqslant \frac{1}{t}$$

故

$$\lim_{t\to\infty} | \frac{E_n(t)}{t} - \frac{L_n(t)}{t} | = 0$$

其中, $\frac{E_n(t)}{t}, \frac{L_n(t)}{t}$ 分别表示单位时间内进入或离开的次数, 令 $t \to \infty$ 则可得单位时间内进入或离开的平均次数.

$$\lim_{t\to\infty} \frac{E_n(t)}{t} = \text{进入状态 } n \text{ 的平均率}$$
$$\lim_{t\to\infty} \frac{L_n(t)}{t} = \text{离开状态 } n \text{ 的平均率}$$

由此可以得出, 对系统的任何状态 $N(t) = n(n = 0, 1, 2, \cdots)$, 进入事件平均率 (单位时间平均到达的顾客数) 等于离去事件平均率 (单位时间平均离开的顾客数), 这就是所谓输入率等于输出率的原则. 用来表示这个原则的方程称做系统的状态平衡方程. 下面要通过建立系统的状态平衡方程来处理一些比较简单的排队模型.

先考虑 $n = 0$ 的状态. 状态 0 的输入仅仅来自状态 1. 处于状态 1 时系统的稳态概率为 P_1, 而从状态 1 进入状态 0 的平均转换率为 μ_1. 因此从状态 1 进入状态 0 的输入率为 $\mu_1 P_1$, 又从其他状态直接进入状态 0 的概率为 0, 所以状态 0 的总输入率为 $\mu_1 P_1 + 0(1 - P_1) = \mu_1 P_1$. 根据类似上面的理由, 状态 0 的总输出率为 $\lambda_0 P_0$. 于是有状态 0 的状态平衡方程

$$\mu_1 P_1 = \lambda_0 P_0$$

对其他每一个状态, 都可以建立类似的状态平衡方程, 但要注意其他状态的输入输出均有两个可能性. 表 5.1 中列出了对各个状态建立的平衡方程.

由表 5.1, 有

$$
\begin{aligned}
P_1 &= \frac{\lambda_0}{\mu_1}P_0 \\
P_2 &= \frac{\lambda_1}{\mu_2}P_1 + \frac{1}{\mu_2}(\mu_1 P_1 - \lambda_0 P_0) = \frac{\lambda_1}{\mu_2}P_1 = \frac{\lambda_1\lambda_0}{\mu_2\mu_1}P_0 \\
P_3 &= \frac{\lambda_2}{\mu_3}P_2 + \frac{1}{\mu_3}(\mu_2 P_2 - \lambda_1 P_1) = \frac{\lambda_2}{\mu_3}P_2 = \frac{\lambda_2\lambda_1\lambda_0}{\mu_3\mu_2\mu_1}P_0 \\
&\vdots \\
P_n &= \frac{\lambda_{n-1}}{\mu_n}P_{n-1} + \frac{1}{\mu_n}(\mu_{n-1}P_{n-1} - \lambda_{n-2}P_{n-2}) \\
&= \frac{\lambda_{n-1}}{\mu_n}P_{n-1} = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1}P_0 \\
&\vdots
\end{aligned}
$$

表 5.1 生灭过程的状态平衡方程

状态	输入率 = 输出率
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	$\lambda_1 P_1 + \mu_3 P_3 = (\lambda_2 + \mu_2)P_2$
$\vdots$	$\vdots$
$n-1$	$\lambda_{n-2}P_{n-2} + \mu_n P_n = (\lambda_{n-1} + \mu_{n-1})P_{n-1}$
n	$\lambda_{n-1}P_{n-1} + \mu_{n+1}P_{n+1} = (\lambda_n + \mu_n)P_n$
$\vdots$	$\vdots$

如果令

$$
C_n = \frac{\lambda_{n-1}\lambda_{n-2}\cdots\lambda_0}{\mu_n\mu_{n-1}\cdots\mu_1} \quad (n = 1, 2, \cdots) \tag{5.3}
$$

且定义$C_0 = 1$, 则各稳态概率公式可以写为

$$
P_n = C_n P_0 \quad (n = 0, 1, 2, \cdots)
$$

因为

$$\sum_{n=0}^{\infty} P_n = \sum_{n=0}^{\infty} C_n P_0 = 1$$

所以, 有

$$P_0 = \Big[\sum_{n=0}^{\infty} C_n\Big]^{-1}$$

求得 P_0 后可以推出 P_n, 再根据本章第 5.1 节公式求出排队系统的各项指标, 即 L, L_q, W, W_q.

$$L = \sum_{n=0}^{\infty} nP_n, \quad L_q = \sum_{n=s}^{\infty}(n-s)P_n$$
$$W = \frac{L}{\lambda_e}, \qquad W_q = \frac{L_q}{\lambda_e}$$

其中, λ_e 是整体平均到达率. 因为 λ_n 是系统处于状态 n $(n = 0, 1, 2, \cdots)$ 时的平均到达率, 且 P_n 为相应系统处于状态 n 的概率, 于是, 有

$$\lambda_e = \sum_{n=0}^{\infty} \lambda_n P_n$$

以上结论是当参数 λ_n, μ_n 给定, 且该过程可以达到稳态的条件下推出的. 当 $\sum\limits_{n=0}^{\infty} C_n = \infty$ 时不再成立.

5.4 基于生灭过程的排队模型

基于生灭过程的排队模型都假设有 Poisson 输入流和负指数服务时间, 不同的是 λ_n 和 μ_n. 下面将依次介绍各模型.

5.4.1 M/M/s 等待制排队模型

M/M/s 等待制排队模型假设:

(1) 顾客到达系统的相继到达时间间隔独立, 且服从参数为 λ 的负指数分布 (即输入过程为 Poisson 过程);

(2) 服务台的服务时间也独立同分布, 且服从参数为 μ 的负指数分布;

(3) 系统空间无限, 允许永远排队.

这是一类最简单的排队系统, 是生灭过程的特例. 其中该排队系统的平均到达率和工作状态的服务台的平均服务率分别为与状态无关的常数 λ, μ. 当只有一个服务台, 即 $s = 1$ 时, 有 $\lambda_n = \lambda\ (n = 0, 1, 2, \cdots), \mu_n = \mu\ (n = 1, 2, \cdots)$; 当有多个服务台, 即 $s > 1$ 时, 有

$$\mu_n = \begin{cases} n\mu, & n < s \\ s\mu, & n \geqslant s \end{cases}$$

设 $\rho = \dfrac{\lambda}{s\mu} < 1$, 排队系统最终能达到稳定状态, 于是, 可直接应用生灭过程的有关结论.

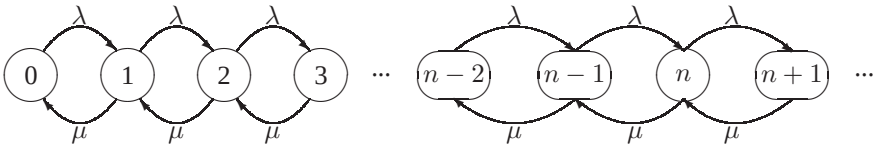

图 5.4 M/M/1/∞ 模型发生率图

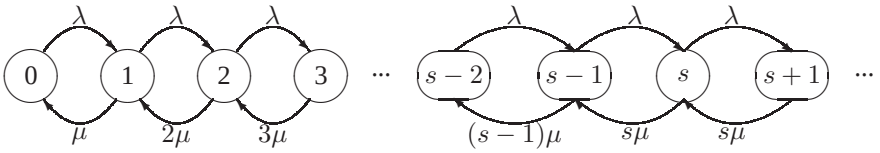

图 5.5 M/M/s/∞ 模型发生率图

5.4.1.1 单服务台 (M/M/1) 的结论

对于单服务台 $s = 1$ 的情形, 由 (5.3) 式, 有

$$C_n = \left(\frac{\lambda}{\mu}\right)^n = \rho^n \quad (n = 0, 1, 2, \cdots)$$

故

$$P_n = \rho^n P_0 \quad (n = 0, 1, 2, \cdots)$$

由于

$$P_0 = \left(\sum_{n=0}^{\infty} C_n\right)^{-1} = \left(\sum_{n=0}^{\infty} \rho^n\right)^{-1} = \left(\frac{1}{1-\rho}\right)^{-1} = 1 - \rho \tag{5.4}$$

因此, 有

$$P_n = (1-\rho)\rho^n \quad (n = 0, 1, 2, \cdots) \tag{5.5}$$

在单服务台系统中, $\rho = \dfrac{\lambda}{\mu}$, 它是单位时间顾客平均到达率与服务率的比值, 反映了服务机构的忙碌或利用的程度. 而另一方面, 由于服务机构的忙期为 $(1 - P_0) = 1 - (1 - \rho) = \rho$, 这与直观理解是完全一致的.

进一步, 可求出其他数量指标

$$\begin{aligned} L &= \sum_{n=0}^{\infty} nP_n = \sum_{n=0}^{\infty} n(1-\rho)\rho^n = (1-\rho)\rho\frac{d}{d\rho}(\sum_{n=0}^{\infty}\rho^n) \\ &= (1-\rho)\rho\frac{d}{d\rho}(\frac{1}{1-\rho}) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda} \end{aligned} \tag{5.6}$$

$$L_q = \sum_{n=1}^{\infty}(n-1)P_n = L - 1(1-P_0) = \frac{\lambda^2}{\mu(\mu-\lambda)} \tag{5.7}$$

$$W = \frac{L}{\lambda} = \frac{1}{\mu-\lambda}, \quad W_q = \frac{L_q}{\lambda} = \frac{\lambda}{\mu(\mu-\lambda)} \tag{5.8}$$

若 $\lambda \geqslant \mu$ 时, 则有 $\rho \geqslant 1$, 即平均到达率超过平均服务率, 上述结果不再适用. 在这种情况下, 排队队长会增加至无限.

假定 $\lambda < \mu$, 下面说明顾客在系统中的逗留时间 w 服从参数为 $\mu(1-\rho)$ 的负指数分布.

设一顾客到达时, 系统中已有 n 个顾客, 按先来先服务的规则, 这个顾客的逗留时间 S_{n+1} 是原有各顾客的服务时间 T_i 和这个顾客的服务时间 T_{n+1} 之和.

$$S_{n+1} = T_1' + T_2 + \cdots + T_n + T_{n+1} \quad (n = 0, 1, 2, \cdots)$$

其中, T_1' 表示这个顾客到达系统时正在接受服务的顾客仍需接受服务的时间. 若 T_i $(i = 1, 2, \cdots, n+1)$ 均服从参数为 μ 负指数分布, 根据负指数分布的无记忆性, T_1' 也服从参数为 μ 的负指数分布, 因此 S_{n+1} 服从 Erlang 分布, 有

$$f(S_{n+1}) = \frac{\mu(\mu t)^n}{n!}e^{-\mu t}, \quad P(S_{n+1} \leqslant t) = \int_0^t \frac{\mu(\mu t)^n}{n!}e^{-\mu t}\,dt$$

顾客在系统中逗留时间小于 t 的概率为

$$\begin{aligned} P(w \leqslant t) &= \sum_{n=0}^{\infty} P_n P(S_{n+1} \leqslant t) \\ &= \sum_{n=0}^{\infty}(1-\rho)\rho^n \int_0^t \frac{\mu}{n!}(\mu t)^n e^{-\mu t}\,dt = 1 - e^{-\mu(1-\rho)t} \end{aligned}$$

即顾客在系统中逗留时间大于 t 的概率为 $P(w>t)=e^{-\mu(1-\rho)}$.

下面讨论顾客在系统中等待时间 w_q 的概率情况. 当这个顾客到来时系统中没有顾客时, 他立即接受服务, 于是 $P(w_q=0)=P_0=1-\rho$; 若这个顾客到来时系统中已有 $n(\geqslant 1)$ 个顾客在等待, 则该顾客需要等待, 设等待时间为 S_n, 有

$$\begin{aligned}P(w_q>t)&=\sum_{n=1}^{\infty}P_nP(S_n>t)=\sum_{n=1}^{\infty}(1-\rho)\rho^nP(S_n>t)\\&=\rho\sum_{n=0}^{\infty}P_nP(S_{n+1}>t)=\rho P(w>t)=\rho e^{-\mu(1-\rho)t}\end{aligned}$$

于是, 由以上可知 w_q 不再服从负指数分布.

根据 (5.8) 式可知, $W_q=E(w_q)=\dfrac{\lambda}{\mu(\mu-\lambda)}$.

然而, 在 $w_q>0$ 的条件下, w_q 的条件分布确是服从参数为 $\mu(1-\rho)$ 的负指数分布的, 有

$$P(w_q>t|w_q>0)=\frac{P(w_q>t)}{P(w_q>0)}=e^{-\mu(1-\rho)t}\quad(t\geqslant 0)$$

由于

$$E(w_q|w_q>0)=\frac{W_q}{1-P_0}=\frac{\lambda}{\mu(\mu-\lambda)}\cdot\frac{\mu}{\lambda}=\frac{1}{\mu-\lambda}$$

于是, 在已经有人等待的情况下还要等待的平均时间为 $\dfrac{1}{\mu-\lambda}$.

【例 5.1】 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/小时; 修理时间服从负指数分布, 平均需要 6 分钟. 试求: (1) 修理店空闲的概率; (2) 店内恰好有 3 个顾客的概率; (3) 在店内的平均顾客数; (4) 每位顾客在店内的平均逗留时间; (5) 等待服务的平均顾客数; (6) 每位顾客平均等待服务时间; (7) 顾客在店内等待服务的时间超过 10 分钟的概率.

解 本例是一个 M/M/1/∞ 排队问题, 其中 $\lambda=4$, $\mu=\dfrac{1}{0.1}=10$, $\rho=\dfrac{\lambda}{\mu}=\dfrac{2}{5}$.

(1) 修理店空闲的概率为 $P_0=1-\rho=1-\dfrac{2}{5}=0.6$.

(2) 店内恰好有 3 个顾客的概率为 $P_3=\rho^3(1-\rho)=\left(\dfrac{2}{5}\right)^3\times\left(1-\dfrac{2}{5}\right)=0.038$.

(3) 在店内的平均顾客数为 $L=\dfrac{\lambda}{\mu-\lambda}=\dfrac{4}{10-4}=0.67$ (人).

(4) 每位顾客在店内的平均逗留时间为 $W=\dfrac{L}{\lambda}=\dfrac{0.67}{4}\times 60=10$ (分钟).

(5) 等待服务的平均顾客数为 $L_q = \dfrac{\lambda^2}{\mu(\mu-\lambda)} = \dfrac{4^2}{10\times(10-4)} = 0.267$ (人).

(6) 每位顾客平均等待服务时间为 $W = \dfrac{L_q}{\lambda} = \dfrac{0.267}{4}\times 60 = 4$ (分钟).

(7) 顾客在店内逗留时间超过 10 分钟的概率为 $P(T>10) = e^{-10(\frac{1}{6}-\frac{1}{15})} = e^{-1} = 0.3679$.

5.4.1.2 多服务台 (M/M/s) 的结论

对有 s 个服务台的服务系统, 由假设有 $\lambda_n = \lambda$ 及

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

因此

$$C_n = \begin{cases} \dfrac{(\frac{\lambda}{\mu})^n}{n!}, & n = 1,2,\cdots,s \\ \dfrac{(\frac{\lambda}{\mu})^s}{s!}(\dfrac{\lambda}{s\mu})^{n-s} = \dfrac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}, & n = s, s+1,\cdots \end{cases}$$

由此, 利用 (5.4) 式可知

$$\begin{aligned} P_0 &= \left[1 + \sum_{n=1}^{s-1}\frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!}\sum_{n=s}^{\infty}(\frac{\lambda}{s\mu})^{n-s}\right]^{-1} \\ &= \left[\sum_{n=0}^{s-1}\frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!}\frac{1}{1-\rho}\right]^{-1} \end{aligned} \tag{5.9}$$

其中, $\rho = \dfrac{\lambda}{s\mu}$. 而且

$$P_n = \begin{cases} \dfrac{(\frac{\lambda}{\mu})^n}{n!}P_0, & n = 0,1,2,\cdots,s \\ \dfrac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}P_0, & n = s, s+1,\cdots \end{cases} \tag{5.10}$$

当 $n \geqslant s$ 时, 即系统中顾客数不少于服务台个数, 这时再来的顾客必须等待, 且必须等待的概率为

$$\sum_{n=s}^{\infty} P_n = \sum_{n=s}^{\infty} \frac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}} P_0 = \frac{1}{s!}(\frac{\lambda}{\mu})^s \sum_{k=0}^{\infty} (\frac{\lambda}{s\mu})^k P_0 = \frac{(\frac{\lambda}{\mu})^s}{s!(1-\rho)} P_0 \tag{5.11}$$

上式称为 Erlang 等待公式.

进一步, 可求出其他的数量指标

$$\begin{aligned} L_q &= \sum_{n=s}^{\infty} (n-s) P_n = \sum_{j=0}^{\infty} j P_{s+j} = \sum_{j=0}^{\infty} j \frac{(\frac{\lambda}{\mu})^s}{s!} \rho^j P_0 \\ &= P_0 \frac{(\frac{\lambda}{\mu})^s}{s!} \rho \sum_{j=0}^{\infty} \frac{d}{d\rho}(\rho^j) = P_0 \frac{(\frac{\lambda}{\mu})^s}{s!} \rho \frac{d}{d\rho}\Big(\frac{1}{1-\rho}\Big) = \frac{P_0 (\frac{\lambda}{\mu})^s \rho}{s!(1-\rho)^2} \end{aligned} \tag{5.12}$$

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

$$\begin{aligned} \overline{s} &= \sum_{n=0}^{s-1} n P_n + s \sum_{n=s}^{\infty} P_n = \sum_{n=0}^{s-1} \frac{n(\frac{\lambda}{\mu})^n}{n!} P_0 + s \frac{(\frac{\lambda}{\mu})^s}{s!(1-\rho)} P_0 \\ &= \frac{\lambda}{\mu} P_0 \left[\sum_{n=1}^{s-1} \frac{(\frac{\lambda}{\mu})^{n-1}}{(n-1)!} + \frac{(\frac{\lambda}{\mu})^{s-1}}{(s-1)!(1-\rho)} \right] = \rho \end{aligned}$$

上式说明,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. 由此, 可得到平均队长 L, 即

$$\begin{aligned} L &= \text{平均排队长} + \text{正在接受服务的顾客平均数} \\ &= L_q + \rho \end{aligned} \tag{5.13}$$

对多服务台系统, Little 公式仍然成立, 于是, 有

$$W_q = \frac{L_q}{\lambda}, \quad W = W_q + \frac{1}{\mu}$$

对单服务台而言, 顾客在排队系统中等待时间的概率分布也可扩展到多服务台的情形, 可参见文献 [34]. 假设 $\lambda < s\mu$, 即 $\rho < 1$, 对任意 $t \geqslant 0$, 有

$$P(w > t) = e^{-\mu t}\left[1 + \frac{P_0\left(\frac{\lambda}{\mu}\right)^s}{s!(1-\rho)}\left(\frac{1 - e^{-\mu t(s-1-\frac{\lambda}{\mu})}}{s - 1 - \frac{\lambda}{\mu}}\right)\right]$$

且

$$P(w_q > t) = [1 - P(w_q = 0)]e^{-s\mu(1-\rho)t}$$

其中 $P(w_q = 0) = \sum\limits_{n=0}^{s-1} P_n$.

当 $s - 1 - \dfrac{\lambda}{\mu} = 0$ 时, 上式中的 $\dfrac{1 - e^{-\mu t(s-1-\frac{\lambda}{\mu})}}{s - 1 - \dfrac{\lambda}{\mu}}$ 替换成为 μt.

当 $\lambda \geqslant s\mu$ 时, 排队系统的队长将趋于无穷, 以上结论不再适用.

【例 5.2】 一个大型露天矿山, 正考虑修建矿石卸位的个数. 估计运矿石的车将按 Poisson 流到达, 平均每小时到 15 辆; 卸矿石时间服从负指数分布, 平均每 3 分钟卸一辆. 又知每辆运送矿石的卡车售价是 8 万元, 修建一个卸位的投资是 14 万元. 问题是建一个矿石卸位还是两个?

解 本题可用 M/M/s/∞ 排队模型分析, 其中 $\lambda = 15$, $\mu = 20$, $\dfrac{\lambda}{\mu} = 0.75$.

(1) 如果只建一个卸位, 取 $s = 1$, 由单服务台系统的 (5.6) 式及 (5.8) 式可得

$$L = \frac{\lambda}{\mu - \lambda} = \frac{15}{20 - 15} = 3 \text{ (辆)}, \quad W = \frac{L}{\lambda} = \frac{3}{15} = 0.2 \text{ (小时)} = 12 \text{ (分钟)}$$

(2) 如果建两个卸位, 取 $s = 2, \rho = \dfrac{\lambda}{s\mu} = \dfrac{0.75}{2} = 0.375$, 由多服务台系统的 (5.9) 式及 (5.12) 式可得

$$P_0 = \left[1 + 0.75 + \frac{(0.75)^2}{2! \times (1 - 0.375)}\right]^{-1} = 0.45$$

$$L = \frac{0.45 \times (0.75)^2 \times 0.375}{2! \times (1 - 0.375)^2} + 0.75 = 0.87 \text{ (辆)}$$

$$W = \frac{L}{\lambda} = \frac{0.87}{15} = 0.058 \text{ (小时)} \approx 3.5 \text{ (分钟)}$$

因此, 修建两个卸位可使在卸位处的卡车的平均数减少 $3 - 0.87 = 2.13$ 辆, 即可增加 2.13 辆卡车执行运输任务, 相当于用一个卸位的投资 14 万元, 换来了 $2.13 \times 8 = 17.04$ (万元) 的运输设备. 因此, 建造两个卸位是合算的.

【例 5.3】 某一保险公司在其一分支机构内设有 3 名理赔仲裁者. 设顾客向该公司要求赔偿的到达件数服从 Poisson 分布, 其平均到达率为每 8 小时一天为 20 件. 每位仲裁者对每件申请案所花费的时间呈负指数分布, 其平均服务时间为 40 分钟. 同时设申请案件是按顺序依次处理的.

(1) 每周每位仲裁者平均花费多少小时用于申请案?

(2) 平均而言, 一个申请案件在机构花费多少时间?

解 (1) 根据题意, 有 $s=3, \lambda=\frac{5}{2}, \mu=\frac{3}{2}, \rho=\frac{\lambda}{s\mu}=\frac{5}{9}$. 于是, 有

$$P_0=\left[1+\frac{5}{3}+\frac{1}{2}\times(\frac{5}{3})^2+\frac{1}{6}\times(\frac{5}{3})^3\times\frac{9}{4}\right]^{-1}=\frac{24}{139}$$

在任意时刻, 空闲的仲裁者的期望值数为

$$3P_0+2P_1+1P_2=3\times\frac{24}{139}+2\times\frac{40}{139}+1\times\frac{100}{3\times139}=\frac{4}{3}\text{ (人)}$$

于是, 任何一个仲裁者在某一特定时刻上为空闲的概率为 $\frac{4}{9}$; 而每周每位仲裁者花费于申请案的平均时间为 $\frac{5}{9}\times40=22.2$ 小时.

(2) 到达的案件在此系统内的平均时间, 可利用下面的方程式求得

$$W=\frac{1}{\lambda}\frac{P_0(\frac{\lambda}{\mu})^s\rho}{s!(1-\rho)^2}+\frac{1}{\mu}=\frac{2}{5}\times\frac{\frac{24}{139}\times(\frac{5}{3})^3\times\frac{5}{9}}{6\times(1-\frac{5}{9})^2}+\frac{2}{3}$$

$$\approx 0.816\text{ (小时)}\approx 49.0\text{ (分钟)}$$

5.4.2 M/M/s 混合制排队模型

在实际生活中会碰到很多这样的情况, 比如医院规定每天挂 100 个号, 那么第 101 个到达者就会自动离去; 在理发店内等待的座位都满员时, 后来的顾客就会设法另找理发店等待等等. 为了区别于 M/M/s 等待制排队模型, 这类 M/M/s 混合制排队模型用 M/M/s/K 表示.

假设在一个服务系统中可以容纳 $K(K\geqslant s)$ 个顾客 (包括被服务与等待的总数, 等待位置只有 $K-1$ 个). 假设顾客的到达率为常数 λ. 在排队系统中已有 K 个顾客的情况下, 新到的顾客将自动离去. 于是, 有

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

M/M/s/K 模型的发生率图除了在状态 K 时停止外, 其余与 M/M/s 等待制排队系统的发生率图 5.4 及图 5.5 相同.

5.4.2.1 单服务台 (M/M/1/K) 的结论

先考虑只有一个单服务台的情况. 由于 $\mu_n = \mu (n = 1, 2, \cdots, K)$, 于是, 有

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

当 $\rho = \dfrac{\lambda}{\mu} < 1$ 时, 有

$$P_0 = \left[\sum_{n=0}^{K} \left(\frac{\lambda}{\mu}\right)^n\right]^{-1} = \left[\frac{1 - \left(\frac{\lambda}{\mu}\right)^{K+1}}{1 - \frac{\lambda}{\mu}}\right]^{-1} = \frac{1-\rho}{1-\rho^{K+1}}$$

$$P_n = \frac{1-\rho}{1-\rho^{K+1}}\rho^n \quad (n = 0, 1, 2, \cdots, K)$$

由此可得, 平均队长 L 及平均排队长 L_q 为

$$\begin{aligned} L &= \sum_{n=0}^{K} nP_n = \frac{1-\rho}{1-\rho^{K+1}}\rho \sum_{n=0}^{K} \frac{d}{d\rho}(\rho^n) \\ &= \frac{1-\rho}{1-\rho^{K+1}}\rho \frac{d}{d\rho}\left(\frac{1-\rho^{K+1}}{1-\rho}\right) = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} \end{aligned} \tag{5.14}$$

$$L_q = \sum_{n=1}^{K} (n-1)P_n = L - (1 - P_0) \tag{5.15}$$

在 $\rho < 1$ 的条件下, 当 $K \to \infty$ 时, (5.14) 式的后一项值将趋于零. 这与前面的 (5.6) 式相同, 即 M/M/1/∞ 排队模型实际上是 M/M/1/K 混合制排队模型的特例.

由于排队系统的容量有限, 只有 $K - 1$ 个排队位置. 设顾客的平均到达率为 λ. 当系统处于状态 K 时, 新来的顾客将不能再进入系统, 即顾客可进入系统概率是 $1 - P_K$. 因此, 单位时间内实际进入系统的顾客平均数为

$$\lambda_e = \sum_{n=0}^{\infty} \lambda_n P_n = \sum_{n=0}^{K-1} \lambda P_n = \lambda(1 - P_K) = \mu(1 - P_0)$$

称 λ_e 为有效到达率. P_K 称为顾客损失率, 它反映了在所有来到系统的顾客数中不能进入系统的顾客比例.

在有限排队的情况下, Little 公式仍然成立. 但需要注意的是, 必须将 λ 换成有效到达率 λ_e. 于是, 可求出平均逗留时间及平均等待时间为

$$W = \frac{L}{\lambda_e} = \frac{L}{\lambda(1-P_K)}, \quad W_q = \frac{L_q}{\lambda_e} = \frac{L_q}{\lambda(1-P_K)}$$

且 $W = W_q + \dfrac{1}{\mu}$ 仍然是成立的. 需要注意的是, 以上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.

对队长受限制的排队模型, 当系统中有 K 个顾客时, 新到的顾客会自动离去, 为使系统达到稳态不一定要求有 $\rho < 1$ 成立. 因为当 $\rho = 1$ 时, 有 $P_n = \rho^n P_0 = P_0 (n = 1, 2, \cdots, K)$, 于是, 有

$$P_0 = P_1 = \cdots = P_K = \frac{1}{K+1}$$

$$L = \sum_{n=0}^{\infty} nP_n = \frac{1}{K+1}\sum_{n=0}^{\infty} n = \frac{K}{2}$$

【例 5.4】 某美容屋系私人开办并自理业务, 由于屋内面积有限, 只能安置 3 个座位供顾客等候, 一旦满座后则后来者不再进屋等候. 已知顾客到达间隔与美容时间均为负指数分布, 平均到达间隔 80 分钟, 平均美容时间为 50 分钟. 试求任一顾客期望等候时间及该美容屋潜在顾客的损失率.

解 这是一个 M/M/1/4 系统. 由题意知, $\dfrac{1}{\lambda} = 80$ 分钟/人, $\dfrac{1}{\mu} = 50$ 分钟/人, 故服务强度 $\rho = \dfrac{\lambda}{\mu} = \dfrac{5}{8} = 0.625$. 于是, 可求得

$$P_0 = \frac{1-\rho}{1-\rho^{K+1}} = \frac{1-0.625}{1-(0.625)^5} \approx 0.4145$$

$$L = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} = \frac{0.625}{1-0.625} - \frac{5\times(0.625)^5}{1-(0.625)^5} \approx 1.1396 \text{ (人)}$$

$$L_q = L - (1-P_0) = 1.1396 - (1-0.4145) \approx 0.5541 \text{ (人)}$$

$$\lambda_e = \mu(1-P_0) = \frac{1}{50} \times (1-0.4145) = 0.01171$$

故任一顾客期望等待时间为 $W_q = \dfrac{L_q}{\lambda_e} = \dfrac{0.5541}{0.01171} \approx 47$ (分钟).

美容屋潜在顾客的损失率, 即系统满员的概率为

$$P_4 = \rho^4 P_0 = (0.625)^4 \times 0.4145 \approx 0.06 = 6\%$$

或者也可按照下式进行计算, 有

$$P_4 = 1 - \frac{\lambda_e}{\lambda} = 1 - 80 \times 0.01171 \approx 0.06 = 6\%$$

5.4.2.2 多服务台 (M/M/s/K) 的结论

对多服务台的 M/M/s/K 混合制排队模型来说, 要求系统的空间有 $K \geqslant s$. 在本模型中, 由于

$$\mu_n = \begin{cases} n\mu, & n = 0, 1, 2, \cdots, s \\ s\mu, & n = s, s+1, \cdots, K \end{cases}$$

于是, 有

$$C_n = \begin{cases} \dfrac{(\frac{\lambda}{\mu})^n}{n!}, & n = 0, 1, 2, \cdots, s \\ \dfrac{(\frac{\lambda}{\mu})^s}{s!}(\dfrac{\lambda}{s\mu})^{n-s} = \dfrac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}, & n = s, s+1, \cdots, K \\ 0, & n > K \end{cases}$$

可求得

$$P_n = \begin{cases} \dfrac{(\frac{\lambda}{\mu})^n}{n!}P_0, & n = 0, 1, 2, \cdots, s \\ \dfrac{(\frac{\lambda}{\mu})^n}{s!s^{n-s}}P_0, & n = s, s+1, \cdots, K \\ 0, & n > K \end{cases} \tag{5.16}$$

其中

$$P_0 = \begin{cases} \left[\displaystyle\sum_{n=0}^{s} \frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!} \sum_{n=s+1}^{K} (\frac{\lambda}{s\mu})^{n-s}\right]^{-1}, & \text{当 } \rho \neq 1 \\ \left[\displaystyle\sum_{n=0}^{s} \frac{(\frac{\lambda}{\mu})^n}{n!} + \frac{(\frac{\lambda}{\mu})^s}{s!}(K-s)\right]^{-1}, & \text{当 } \rho = 1 \end{cases} \tag{5.17}$$

由平稳分布 $P_n(n=0,1,2,\cdots,K)$, 可得平均排队长 L_q 为

$$
\begin{aligned}
L_q &= \sum_{n=s}^{K}(n-s)P_n \\
&= \begin{cases}
\dfrac{P_0\left(\dfrac{\lambda}{\mu}\right)^s\rho}{s!(1-\rho)^2}\left[1-\rho^{K-s}-(K-s)\rho^{K-s}(1-\rho)\right], & \rho\neq 1 \\
\dfrac{P_0\left(\dfrac{\lambda}{\mu}\right)^s(K-s)(K-s+1)}{2s!}, & \rho=1
\end{cases}
\end{aligned}
$$

其中, $\rho=\dfrac{\lambda}{s\mu}$.

为求平均队长, 由

$$
\begin{aligned}
L_q &= \sum_{n=s}^{K}(n-s)P_n = \sum_{n=s}^{K}nP_n - s\sum_{n=s}^{K}P_n = \sum_{n=0}^{K}nP_n \\
&\quad - \sum_{n=0}^{s-1}nP_n - s\left(1-\sum_{n=0}^{s-1}P_n\right) = L - \sum_{n=0}^{s-1}(n-s)P_n - s
\end{aligned}
$$

可得

$$
L = L_q + s + \sum_{n=0}^{s-1}(n-s)P_n = L_q + s + P_0\sum_{n=0}^{s-1}\frac{(n-s)\left(\dfrac{\lambda}{\mu}\right)^n}{n!}
$$

同多服务台 M/M/s 等待制排队模型一样, 在利用 Little 公式时, 需要将 λ 换成有效到达率 λ_e, 可求得

$$
W = \frac{L}{\lambda_e}, \quad W_q = \frac{L_q}{\lambda_e} = W - \frac{1}{\mu}
$$

下面, 从另一个角度来推出平均队长 L. 由于平均被占用的服务台数, 也即正在接受服务的顾客数 $\overline{s}$ 为

$$
\begin{aligned}
\overline{s} &= \sum_{n=0}^{s}nP_n + \sum_{n=s+1}^{K}sP_n \\
&= P_0\left[\sum_{n=0}^{s}\frac{n\left(\dfrac{\lambda}{\mu}\right)^n}{n!} + s\sum_{n=s+1}^{K}\frac{\left(\dfrac{\lambda}{\mu}\right)^n}{s!s^{n-s}}\right] = \frac{\lambda}{\mu}(1-P_K)
\end{aligned}
$$

于是, 有 $L = L_q + \overline{s} = L_q + \dfrac{\lambda}{\mu}(1 - P_K)$.

显然, 当 $K \to \infty$ 时, 以上各结果同队长不受限制时结果一样, 即 M/M/s/∞ 模型是 M/M/s/K 模型的特例.

最后, 需要指出的是, 当 $K = s$ 时, M/M/s/K 混合制排队模型就是损失制的服务系统. 只要在 (5.16) 式和 (5.17) 式中令 $K = s$, 就可得到如下损失制排队服务系统的基本公式:

$$P_0 = \left[\sum_{n=0}^{s} \frac{(\frac{\lambda}{\mu})^n}{n!}\right]^{-1} \tag{5.18}$$

$$P_n = \frac{(\frac{\lambda}{\mu})^n}{n!} P_0 = \frac{\frac{(\frac{\lambda}{\mu})^n}{n!}}{\sum\limits_{n=0}^{s} \frac{(\frac{\lambda}{\mu})^n}{n!}} \quad (n = 0, 1, \cdots, s) \tag{5.19}$$

$$L_q = 0, \quad W_q = 0, \quad W = \frac{1}{\mu} \tag{5.20}$$

$$L = \sum_{n=0}^{s} nP_n = \frac{\sum\limits_{n=0}^{s} \frac{n(\frac{\lambda}{\mu})^n}{n!}}{\sum\limits_{n=0}^{s} \frac{(\frac{\lambda}{\mu})^n}{n!}} = \frac{\lambda}{\mu}(1 - P_K) \tag{5.21}$$

【例 5.5】 某汽车加油站设有两个加油站, 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 可看做一个 M/M/2/5 排队系统, 其中 $\lambda = 2$, $\mu = 0.5$, $\dfrac{\lambda}{\mu} = 4$, $s = 2$, $K = 5$.

(1) 系统空闲的概率为

$$P_0 = \left[1 + 4 + \frac{4^2 \times \left[1 - (\frac{4}{2})^{5-2+1}\right]}{2! \times (1 - \frac{4}{2})}\right]^{-1} = 0.008$$

(2) 顾客损失率为

$$P_5 = \frac{4^5 \times 0.008}{2! \times 2^{5-2}} = 0.512$$

(3) 加油站内等待的平均汽车数为

$$L_q = \frac{0.008 \times 4^2 \times \frac{4}{2}}{2! \times (1-\frac{4}{2})^2} \times \Big[1 - \big(\frac{4}{2}\big)^{5-2} - (5-2)\big(\frac{4}{2}\big)^{5-2}\big(1-\frac{4}{2}\big)\Big] = 2.18 \text{ (辆)}$$

加油站内汽车的平均数为

$$L = L_q + \frac{\lambda}{\mu}(1-P_5) = 2.18 + \frac{2}{0.5} \times (1-0.512) = 4.13 \text{ (辆)}$$

(4) 汽车在加油站内平均逗留时间为

$$W = \frac{L}{\lambda(1-P_5)} = \frac{4.13}{2 \times (1-0.512)} = 4.23 \text{ (分钟)}$$

汽车在加油站内平均等待时间为

$$W_q = W - \frac{1}{\mu} = 4.13 - 2 = 2.23 \text{ (分钟)}$$

(5) 被占用的加油机的平均数为

$$\overline{s} = L - L_q = 4.13 - 2.18 = 1.95 \text{ (辆)}$$

【例 5.6】 某单位电话交换台有一台 200 门内线的总机. 已知在上班的 8 小时内, 有 20% 的内线分机平均每 40 分钟要一次外线电话, 80% 的分机平均隔两个小时要一次外线电话, 又知从外单位打来的电话呼唤率平均每分钟一次, 设外线通话时间平均为 3 分钟, 以上两个时间均属负指数分布. 如果要求电话接通率为 95%, 问该交换台应设置多少外线?

解 (1) 来到电话交换台的呼唤有两类: 一是各分机往外打的电话, 二是从外单位打进来的电话. 前一类 $\lambda_1 = \big(\frac{60}{40} \times 0.2 + \frac{1}{2} \times 0.8\big) \times 200 = 140$, 后一类 $\lambda_2 = 60$, 根据 Poisson 分布性质, 来到交换台的总呼唤流仍为 Poisson 分布, 其参数为 $\lambda = \lambda_1 + \lambda_2 = 200$.

(2) 这是一个具有多个服务台带损失制的服务系统, 根据 (5.19) 式, 要使电话接通率为 95%, 就是要使损失率低于 5%, 也即

$$P_n=\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}P_0=\frac{\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}}{\sum\limits_{n=0}^{s}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}}\leqslant 0.05$$

本例中 $\mu=20,\dfrac{\lambda}{\mu}=10$, 可以用表 5.2 进行计算来求 s.

表 5.2 某多服务台带损失制电话服务系统的计算过程

s	$\frac{\left(\frac{\lambda}{\mu}\right)^s}{s!}$	$\sum\limits_{n=0}^{s}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}$	P_s	s	$\frac{\left(\frac{\lambda}{\mu}\right)^s}{s!}$	$\sum\limits_{n=0}^{s}\frac{\left(\frac{\lambda}{\mu}\right)^n}{n!}$	P_s
0	1.0	1.0	1.0	8	2480.2	7330.9	0.338
1	10.0	11.0	0.909	9	2755.7	10086.6	0.273
2	50.0	61.0	0.820	10	2755.7	12842.3	0.215
3	166.7	227.2	0.732	11	2505.2	15347.5	0.163
4	416.7	644.4	0.647	12	2087.7	17435.2	0.120
5	833.3	1477.7	0.564	13	1605.9	19041.1	0.084
6	1388.9	2866.6	0.485	14	1147.1	20188.2	0.056
7	1984.1	4850.7	0.409	15	764.7	20952.9	0.036

根据计算看出, 为了外线接通率达到 0.95%, 应不少于 15 条外线.

说明: (1) 计算中没有考虑外单位打来的电话时, 内线是否占用, 也没有考虑分机打外线时对方是否占用; (2) 当电话一次打不通时, 就要打两次、三次 $\cdots$, 因此, 实际上呼唤次数要远远高于计算次数, 实际接通率也比 95% 低得多.

5.4.3 有限源排队模型

基于 M/M/s/K 混合制排队模型, 现假定顾客源有限, 不妨设只有 N 个顾客. 每个顾客来到系统中接受服务后回到原来的总体, 还有可能再来. 当排队系统中有 $n(n=0,1,\cdots,N)$ 个顾客时, 则只剩下 $N-n$ 个潜在的顾客在排队系统外. 这类有限源排队系统的典型例子有: (1) s 个工人共同负责 N 台机器的维修. (2) N 个打字员共用一台打字机等. 如图 5.6 所示.

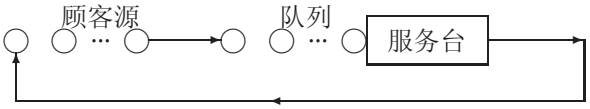

图 5.6 有限源排队系统

下面, 以机器维修问题为例进行说明. 每一台机器交替出现在排队系统中和排队系统外. 类似于 M/M/s 模型, 假定每一个顾客 (机器) 的相继到达间隔时间 (即从离开系统到再次进入排队系统的时间) 服从参数 λ 的负指数分布, 假设当前系统中有 n 个顾客, 即有 $N-n$ 个在系统外. 对排队系统而言, 再次有顾客进入排队系统的相继到达时间间隔服从参数为 $\lambda_n=(N-n)\lambda$ 的负指数分布 (由负指数分布的性质可知). 于是该模型仍可作为生灭过程的一种特殊形式. 且当 $\lambda_n=0, n=N$ 时, 该模型最终会达到平稳状态.

单服务台有限源排队模型及多服务台有限源排队模型发生率图如图 5.7 与图 5.8 所示.

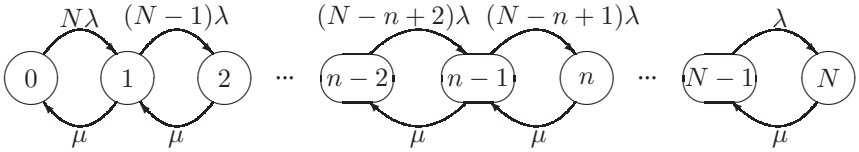

图 5.7 单服务台有限源排队模型发生率图

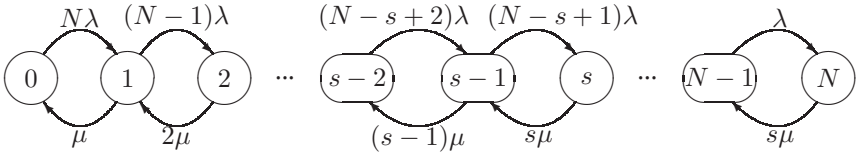

图 5.8 多服务台有限源排队模型发生率图

5.4.3.1 单服务台模型 ($s=1$)

当 $s=1$ 时, 基于生灭过程的 C_n 为

$$C_n=\begin{cases} N(N-1)\cdots(N-n+1)\big(\frac{\lambda}{\mu}\big)^n=\frac{N!}{(N-n)!}\big(\frac{\lambda}{\mu}\big)^n, & n\leqslant N \\ 0, & n>N \end{cases}$$

于是, 有

$$\begin{aligned} P_0&=\sum_{n=0}^{N}\left[\frac{N!}{(N-n)!}\big(\frac{\lambda}{\mu}\big)^n\right]^{-1} \\ P_n&=\frac{N!}{(N-n)!}\big(\frac{\lambda}{\mu}\big)^nP_0\quad(n=1,2,\cdots N) \\ L_q&=\sum_{n=1}^{N}(n-1)P_n=N-\frac{\lambda+\mu}{\lambda}(1-P_0) \\ L&=\sum_{n=0}^{N}nP_n=L_q+1-P_0=N-\frac{\mu}{\lambda}(1-P_0) \end{aligned}$$

由于顾客输入率 λ_n 随系统状态而变化, 因此, 有效到达率 λ_e 按下式计算

$$\lambda_e=\sum_{n=0}^{N}\lambda_nP_n=\sum_{n=0}^{N}(N-n)\lambda P_n=\lambda(N-L)$$

再利用 Little 公式, 有 $W=\dfrac{L}{\lambda_e},W_q=\dfrac{L_q}{\lambda_e}$.

5.4.3.2 多服务台模型

设 $N\geqslant s>1$, 有

$$C_n=\begin{cases} \frac{N!}{(N-n)!n!}\big(\frac{\lambda}{\mu}\big)^n, & 0\leqslant n\leqslant s \\ \frac{N!}{(N-n)!s!s^{n-s}}\big(\frac{\lambda}{\mu}\big)^n, & n=s,s+1,\cdots,N \\ 0, & n>N \end{cases}$$

于是, 有

$$P_n=\begin{cases} \frac{N!}{(N-n)!n!}\big(\frac{\lambda}{\mu}\big)^nP_0, & 0\leqslant n\leqslant s \\ \frac{N!}{(N-n)!s!s^{n-s}}\big(\frac{\lambda}{\mu}\big)^nP_0, & n=s,s+1,\cdots,N \\ 0, & n>N \end{cases}$$

其中

$$P_0=\left[\sum_{n=0}^{s-1}\frac{N!}{(N-n)!n!}\left(\frac{\lambda}{\mu}\right)^n+\sum_{n=s}^{N}\frac{N!}{(N-n)!s!s^{n-s}}\left(\frac{\lambda}{\mu}\right)^n\right]^{-1}$$

进一步, 可求得其他各指标

$$\begin{aligned}L_q&=\sum_{n=s}^{N}(n-s)P_n\\ L&=\sum_{n=0}^{N}nP_n=\sum_{n=0}^{s}nP_n+s\sum_{n=s+1}^{N}P_n+\sum_{n=s+1}^{N}(n-s)P_n\\ &=\sum_{n=0}^{s-1}nP_n+L_q+s\Big(1-\sum_{n=0}^{s-1}P_n\Big)\end{aligned}$$

或

$$L=L_q+\frac{\lambda_e}{\mu}=L_q+\frac{\lambda}{\mu}(N-L),\quad W=\frac{L}{\lambda_e},\quad W_q=\frac{L_q}{\lambda_e}$$

【例 5.7】 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. 当发生故障以后, 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 $\lambda=\frac{1}{15},\mu=\frac{1}{12},\rho=\frac{\lambda}{\mu}=0.8,N=5.$ 于是有,

(1) 修理工人空闲的概率为

$$\begin{aligned}P_0&=\left[\frac{5!}{5!}\times(0.8)^0+\frac{5!}{4!}\times(0.8)^1+\frac{5!}{3!}\times(0.8)^2+\frac{5!}{1!}\times(0.8)^4+\frac{5!}{0!}\times(0.8)^5\right]^{-1}\\ &=0.0073\end{aligned}$$

(2) 5 台机器都出故障的概率为 $P_5=\frac{5!}{0!}\times(0.8)^5P_0=0.287.$

(3) 出故障机器的平均数为 $L=5-\frac{1}{0.8}\times(1-0.0073)=3.76$ (台).

(4) 等待修理机器的平均数为 $L_q=3.76-(1-0.0073)=2.77$ (台).

(5) 每台机器平均停工时间为 $W=\dfrac{5}{1\times\dfrac{(1-0.0073)}{12}}-15=46$ (分钟).

(6) 每台机器平均待修时间为 $W_q=46-12=34$ (分钟).

(7) 工人的维修能力为 $A=\frac{1}{12}\times(1-0.0073)\times 60=4.96$ (台). 即该工人每小时可修理机器的平均台数为 4.96 台.

上述结果表明,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.

5.4.4 服务率或到达率依赖状态的排队模型

在前面讨论的各类排队模型中, 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, 服务率或到达率可能是随系统状态的变化而变化的. 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该系统; 而此时服务员的服务率也可能会提高.

对单服务台系统, 可分别假设实际的服务率和到达率 (它们均依赖于系统所处的状态 n) 为

$$\mu_n=n^a\mu_1\quad(n=1,2,\cdots)$$
$$\lambda_n=(n+1)^{-b}\lambda_0\quad(n=0,1,2,\cdots)$$

其中, λ_n,μ_n 表示系统中处于状态 n 的平均到达率和服务率; a,b 可称为压力系数, 且为给定正数.

下面以 $\mu_n=n^a\mu_1$ 为例来说明正数 a 的含义. 如果取 $a=1$, 则表示假设平均服务率与 n 成正比; 若取 $a=\frac{1}{2}$, 则假设平均服务率与 $\sqrt{n}$ 成正比. 在前面的各模型中, 均假设压力系数 $a=0$. 类似可解释正数 b. 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.

对多服务台系统, 可假设实际的平均到达率和平均服务率为

$$\mu_n=\begin{cases}n\mu_1, & n\leqslant s\\ (\frac{n}{s})^a s\mu_1, & n\geqslant s\end{cases}$$

$$\lambda_n=\begin{cases}\lambda_0, & n\leqslant s-1\\ (\frac{s}{n+1})^b\lambda_0, & n\geqslant s-1\end{cases}$$

于是, 对多服务台系统, 有

$$C_n=\begin{cases}\dfrac{(\frac{\lambda_0}{\mu_1})^n}{n!}, & n=0,1,2,\cdots,s\\ \dfrac{(\frac{\lambda_0}{\mu_1})^n}{s!(\frac{n!}{s!})^c s(1-c)(n-s)}, & n=s,s+1,\cdots\end{cases}$$

其中, $c=a+b$.

下面来看一个简单的特例, 考虑一个顾客到达率依赖状态的单服务台等待制系统 M/M/1/∞, 其参数为

$$\lambda_n=\frac{\lambda}{n+1}\quad(n=0,1,2,\cdots)$$
$$\mu_n=\mu\qquad(n=1,2,\cdots)$$

于是, 有

$$C_n=\frac{\lambda\cdot(\frac{\lambda}{2})\cdot(\frac{\lambda}{3})\cdots(\frac{\lambda}{n})}{\mu^n}=\frac{\lambda^n}{n!\mu^n}$$

设 $\frac{\lambda}{\mu}<1$, 有

$$P_0=\left[\sum_{n=0}^{\infty}\frac{1}{n!}(\frac{\lambda}{\mu})^n\right]^{-1}=e^{-\frac{\lambda}{\mu}}$$
$$P_n=\frac{\lambda^n}{n!\mu^n}P_0\quad(n=1,2,\cdots)$$
$$L=\sum_{n=0}^{\infty}nP_n=\sum_{n=0}^{\infty}\frac{n(\frac{\lambda}{\mu})^n}{n!}P_0=\frac{\lambda}{\mu}$$
$$L_q=\sum_{n=0}^{\infty}(n-1)P_n=L-(1-P_0)=\frac{\lambda}{\mu}+e^{-\frac{\lambda}{\mu}}-1$$
$$\lambda_e=\sum_{n=0}^{\infty}\frac{\lambda}{n+1}P_n=\mu(1-e^{-\frac{\lambda}{\mu}})$$
$$W=\frac{L}{\lambda_e}=\frac{\frac{\lambda}{\mu}}{\mu(1-e^{-\frac{\lambda}{\mu}})},\quad W_q=\frac{L_q}{\lambda_e}=W-\frac{1}{\mu}$$

5.5 非生灭过程排队模型

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. 而对于非生灭过程的排队模型分析是非常困难的, 下面仅就几种特殊情形给出有关的结果.

5.5.1 M/G/1 排队模型

M/G/1 排队模型是指顾客的到达为 Poisson 流 (或相继到达间隔时间服从负指数分布), 服务时间服从一般独立分布的单服务台排队模型.

设顾客的平均到达率为 λ, 服务时间的均值 $\frac{1}{\mu} < \infty$, 方差 $\sigma^2 < \infty$. 可以证明, 只要 $\rho = \frac{\lambda}{\mu} < 1$, 系统就可以达到平稳状态, 并有稳态概率 $P_0 = 1 - \rho$. 且根据 Pollaczek-Khintchine (P-K) 公式, 有

$$L_q = \frac{\lambda^2\sigma^2 + \rho^2}{2(1-\rho)}, \quad L = \rho + L_q$$

由此, 进一步可求出 $W_q = \frac{L_q}{\lambda}, W = W_q + \frac{1}{\mu}$.

由以上公式可以看出, L_q, L, W, W_q 都仅仅依赖于 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.

而且, 从上式还不难发现, 当服务率 μ 给定, 方差 σ^2 减少时, 平均队长和等待时间等都将减少. 于是, 可通过改变服务时间的方差来缩短平均队长. 当 $\sigma^2 = 0$ 时, 即服务时间为定长时, 平均队长及等待时间都将减到最少水平.

【例 5.8】 某储蓄所有一个服务窗口, 顾客按 Poisson 流平均每小时到达 10 人. 设为任一顾客办理存款、取款等业务的时间为 V, 根据过去的经验表明, 有 $V \sim N(0.05, 0.01^2)$. 试求该储蓄所空闲的概率及其主要工作指标.

解 本例中, $\lambda = 10, \frac{1}{\mu} = 0.05, \sigma^2 = 0.01^2, \rho = \frac{\lambda}{\mu} = 10 \times 0.05 = 0.5$.

于是, 可得如下结果:

$$\begin{aligned}
P_0 &= 1 - \rho = 1 - 0.5 = 0.5 \\
L_q &= \frac{0.5^2 + 10^2 \times (0.01)^2}{2 \times (1 - 0.5)} = 0.26 \text{ (人)} \\
L &= L_q + \rho = 0.26 + 0.5 = 0.76 \text{ (人)} \\
W &= \frac{L}{\lambda} = \frac{0.76}{10} = 0.076 \text{ (小时)} \approx 5 \text{ (分钟)} \\
W_q &= \frac{L}{\lambda} = \frac{0.26}{10} = 0.026 \text{ (小时)} \approx 2 \text{ (分钟)}
\end{aligned}$$

5.5.2 M/D/1 排队模型

正如前面所叙述的, 对定长服务时间的 M/D/1/∞ 模型, 有 $E(V)=\dfrac{1}{\mu}, \sigma^2=0$. 由 Pollaczek-Khintchine 公式有

$$L_q=\frac{\rho^2}{2(1-\rho)}=\frac{\lambda^2}{2\mu(\mu-\lambda)},\quad L=L_q+\rho=\frac{\lambda(2\mu-\lambda)}{2\mu(\mu-\lambda)}$$

由此, 可得 $W_q=\dfrac{\rho^2}{2\lambda(1-\rho)}=\dfrac{\lambda}{2\mu(\mu-\lambda)}$. 不难发现, 在服务时间服从负指数分布的条件下的等待时间 W_q 正好是定长服务时间条件下等待时间的 2 倍.

5.5.3 M/E$_k$/1 排队模型

设系统对任一顾客的服务时间 V 服从 k 阶 Erlang 分布, 其密度函数为

$$f(t)=\frac{k\mu(k\mu t)^{k-1}}{(k-1)!}e^{-k\mu t}\quad (t\geqslant 0)$$

已知 $E(V)=\dfrac{1}{\mu}, \sigma^2=\dfrac{1}{k\mu^2}$, 且 M/E$_k$/1 模型可作为 M/G/1 模型的一个特例, 于是, 根据 Pollaczek-Khintchine 公式, 可得

$$\begin{aligned}L_q&=\frac{\dfrac{\lambda^2}{k\mu^2}+\rho^2}{2(1-\rho)}=\frac{1+k}{2k}\cdot\frac{\lambda^2}{\mu(\mu-\lambda)}\\ W_q&=\frac{1+k}{2k}\cdot\frac{\lambda}{\mu(\mu-\lambda)}\\ W&=W_q+\frac{1}{\mu},\quad L=\lambda W\end{aligned}$$

【例 5.9】 一个质量检查员平均每小时收到两件送来检查的样品, 每件样品要依次完成 5 项检验才能判定是否合格. 据统计, 每项检验所需时间的期望值都是 4 分钟, 每项检验的时间和送检产品的到达间隔都为负指数分布. 问一件样品从送到至检查完毕预期需要多少时间?

解 分析题意可知, 该系统为 M/E$_k$/1 模型, 且有 $\lambda=2$ 件/小时, $k=5$. 设 $V_i(i=1,2,\cdots,5)$ 为任一样品第 i 项检验的时间, 由 Erlang 分布的性质可知

$$E(V_i)=\frac{1}{k\mu}=\frac{1}{5\mu}=\frac{1}{15}\text{ (件/小时)}\quad (i=1,2,\cdots,5)$$

于是, 由 $\frac{1}{\mu}=\frac{1}{3}$ 件/小时, 有 $\mu=3$ 件/小时, 即 $\rho=\frac{\lambda}{\mu}=\frac{2}{3}$. 将 ρ 及 k 代入有关公式中, 可得

$$L_q=\frac{(5+1)\times\frac{2}{3}}{2\times5\times(1-\frac{2}{3})}=\frac{6}{5}\text{ (件)}$$

$$W_q=\frac{L_q}{\lambda}=\frac{6}{2\times5}=\frac{3}{5}\text{ (小时)}$$

$$W=W_q+\frac{1}{\mu}=\frac{3}{5}+\frac{1}{3}=\frac{14}{15}\text{ (小时)}=56\text{ (分钟)}$$

即每一件样品从送到至检查完毕预期需要 44 分钟.

5.6 服务机构串连的排队系统

这类模型在生产中碰到的较多: 产品的生产要经过若干工艺阶段; 零件的加工要经过好几道工序, 在一条流水生产线或装配线上, 零件按一定的节拍从上一道工序传到下一道工序, 由于工序时间的波动或工序间库存位置的不足, 都可能造成生产的混乱或阻塞 (见图 5.9), 因此这类模型中, 要研究服务站工序时间波动及库存位置变动情况下, 造成生产混乱或阻塞的概率.

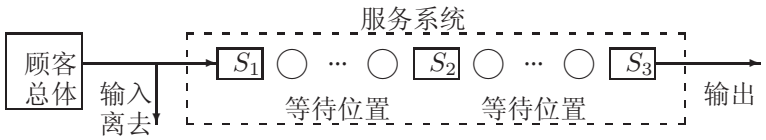

图 5.9 服务机构串连的排队系统

假设顾客的到达服从参数为 λ 的 Poisson 过程, 每个服务台对顾客的服务时间均服从参数为 μ 的负指数分布. 各服务台前允许顾客排队等待的位置有三种情况: (1) 无位置; (2) 有限的位置; (3) 无限的位置. 由于第三种情况相当于每个服务台都是一个独立的排队系统, 所以只需考虑前面两种情况.

先考虑有两个服务站且工序间无排队位置的情况. 即顾客在第一个服务站 (S_1) 服务完毕, 如第二个服务站 (S_2) 空闲, 立即转入 S_2; 否则仍停留在 S_1. 这样 S_1 可能有三种状态: (1) 无顾客 (记 $i=0$); (2) 有一个顾客正得到服务 (记作 $i=1$); (3) 有一个顾客服务已完毕, 但由于第二个服务台无空闲, 顾客仍留在第

一个服务站 (记作 $i = b$). 第二个服务台 S_2 可能有两种状态: (1) 空闲无顾客 (记作 $j = 0$); (2) 有一个顾客正得到服务 (记作 $j = 1$). 于是, 整个系统就可能有五种状态: $(0,0),(0,1),(1,0),(1,1),(b,1)$. 为了分别找出处于五种状态下的概率 P_{ij}, 画出其生灭过程发生率图 (如图 5.10 所示). 当 $\mu_1 = \mu_2 = \mu$ 时, 写出各

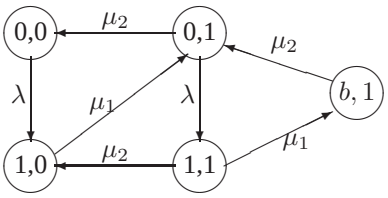

图 5.10 串连系统生灭过程发生图 1

状态的平衡方程如下:

$$\begin{cases} \mu P_{01} = \lambda P_{00} \\ \mu P_{10} + \mu P_{b1} = (\lambda + \mu) P_{01} \\ \lambda P_{00} + \mu P_{11} = \mu P_{10} \\ \lambda P_{01} = 2\mu P_{11} \\ \mu P_{11} = \mu P_{b1} \end{cases}$$

又因为有 $P_{00} + P_{01} + P_{10} + P_{11} + P_{b1} = 1$, 联立求解得, $P_{00} = \frac{2}{H}$, $P_{01} = \frac{2\rho}{H}$, $P_{10} = \frac{\rho^2 + 2\rho}{H}$, $P_{11} = \frac{\rho^2}{H}$. 其中 $H = 3\rho^2 + 4\rho + 2$.

由此得到系统中顾客的平均数为

$$L = \sum_i \sum_j n P_{ij} = 0P_{00} + 1P_{01} + 1P_{10} + 2P_{11} + 2P_{b1} = \frac{4\rho + 5\rho^2}{H}$$

服务机构的忙期为 $1 - P_{00} = 1 - \frac{2}{H}$.

顾客的有效输入率为 $\lambda_e = \lambda(P_{00} + P_{01}) = \big(\frac{2 + 2\rho}{H}\big)\lambda$.

下面仍假设有两个服务站, 但中间有一个等待位置. 仍用 i 记录第一个服务站 S_1 所处的状态, 这时 S_1 有三个状态 ($i = 0, 1, b$); 第二个服务站 S_2 也有三个状态, $j = 0$ (S_2 空闲); $j = 1$ (S_2 有一个顾客正得到服务, 无顾客等待); $j = 2$ (S_2 有一个顾客正得到服务, 有一个顾客正在等待). 这时整个系统就可能

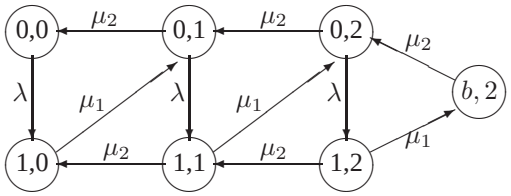

图 5.11 串连系统生灭过程发生图 2

有七种状态: $(0,0),(0,1),(0,2),(1,0),(1,1),(1,2),(b,2)$. 同前面的情况类似, 画出其生灭过程发生率图, 如图 5.11 所示.

当 $\mu_1=\mu_2=\mu$ 时, 写出各状态的平衡方程为

$$\begin{cases}\mu P_{01}-\lambda P_{00}=0\\ \mu P_{02}+\mu P_{10}-(\lambda+\mu)P_{01}=0\\ \mu P_{11}+\mu P_{b2}-(\lambda+\mu)P_{02}=0\\ \lambda P_{00}+\mu P_{11}-\mu P_{10}=0\\ \lambda P_{01}+\mu P_{12}-2\mu P_{11}=0\\ \lambda P_{02}-2\mu P_{12}=0\\ \mu P_{12}-\mu P_{b2}=0\end{cases}$$

又因为有 $P_{00}+P_{01}+P_{02}+P_{10}+P_{11}+P_{12}+P_{b2}=1$, 求解得

$$P_{00}=\frac{\rho+4}{H_1},\qquad P_{01}=\frac{(\rho^2+4\rho)}{H_1},\quad P_{02}=\frac{2\rho^2}{H_1}$$
$$P_{10}=\frac{\rho^3+3\rho^2+4\rho}{H_1},\quad P_{11}=\frac{\rho^3+2\rho^2}{H_1},\quad P_{12}=P_{b2}=\frac{\rho^3}{H_1}$$

其中, $H_1=4\rho^3+8\rho^2+9\rho+4$.

系统内顾客的平均数为

$$L=1P_{00}+1P_{01}+2P_{02}+2P_{11}+3P_{12}+3P_{b2}=\frac{9\rho^3+12\rho^2+8\rho}{H_1}$$

系统的忙期为 $1-P_{00}=1-\dfrac{\rho+4}{H_1}$.

顾客的有效输入率为 $\lambda_e=\lambda(p_{00}+p_{01}+p_{02})=\left(\dfrac{4+5\rho+3\rho^2}{H_1}\right)\lambda$.

5.7 具有优先服务权的排队模型

这类模型中不再按照先来先服务的原则进行服务. 级别较高的顾客比级别较低的顾客享有优先服务权, 在同一级别的顾客中则按照先来先服务的原则进行服务. 如打电报分加急和一般; 到医院治病有急诊与普通门诊等.

假定在一个排队系统中, 顾客可划分为 N 个等级, 第一级享有最高的优先权, 第 N 级享有最低级别的优先权. 假设第 i 级优先权顾客的到达服从参数为 $\lambda_i(i=1,2,\cdots,N)$ 的 Poisson 分布, 同时, 系统对任何级别顾客的服务时间均服从参数为 μ 的负指数分布, 即服务台对任何级别顾客的平均服务时间为 $\frac{1}{\mu}$. 假设当一个具有较高级别优先权的顾客到来时, 若正被服务的顾客是一个具有较低级别优先权的顾客, 则该顾客将被中断服务, 回到排队系统中等待重新得到服务.

根据以上假定, 对具有最高级别优先权的顾客来到排队系统时, 只有当具有同样最高级别的顾客正得到服务时需要等待外, 其余情况下均可以立即得到服务. 因此, 对具有第一级优先权的顾客在排队系统中得到服务的情况就如同没有其他级别的顾客时一样. 因此, 第 5.4.1 节中 (5.8) 式对最高级优先权的顾客完全适用.

现同时考虑享有第一及第二优先级的顾客. 由于他们的服务不受其他级别顾客的影响, 设 $\overline{W}_{12}$ 表示第一、第二两级综合在一起的每个顾客在系统中的平均逗留时间, 则有

$$(\lambda_1+\lambda_2)\overline{W}_{12}=\lambda_1 W_1+\lambda_2 W_2$$

其中 W_1, W_2 分别表示享有第一级和第二级优先服务权的顾客在系统中的平均逗留时间. 根据负指数分布的性质, 对于高一级顾客到达而中断服务, 重新回到队伍中的较低级别顾客的服务时间的概率分布, 不因前一段已得到服务及服务了多长时间而有所改变, 因此对 $\overline{W}_{12}$ 只需将具有第一、第二级优先级的顾客的输入率加在一起, 即 $\lambda_1+\lambda_2$, 于是按第 5.4.1 节中 (5.8) 式可以进行计算. 由此, 又可求出 W_2, 有

$$W_2=\frac{\lambda_1+\lambda_2}{\lambda_2}\overline{W}_{12}-\frac{\lambda_1}{\lambda_2}W_1$$

同理, 有

$$(\lambda_1+\lambda_2+\lambda_3)\overline{W}_{123}=\lambda_1 W_1+\lambda_2 W_2+\lambda_3 W_3$$

所以

$$W_3=\frac{\lambda_1+\lambda_2+\lambda_3}{\lambda_3}\overline{W}_{123}-\frac{\lambda_1}{\lambda_3}W_1-\frac{\lambda_2}{\lambda_3}W_2$$

依次类推, 可以求得

$$W_N = \frac{\sum\limits_{i=1}^{N}\lambda_i}{\lambda_N}\overline{W}_{12\cdots N} - \frac{\sum\limits_{i=1}^{N}\lambda_i W_i}{\lambda_N}$$

其中, 有 $\sum\limits_{i=1}^{N}\lambda_i < s\mu$.

【例 5.10】 来到某医院门诊部就诊的病人按照 $\lambda = 2$ 人/小时 的 Poisson 分布到达, 医生对每个病人的服务时间服从负指数分布, $\frac{1}{\mu} = 20$ 分钟. 假如病人中 60% 属一般病人, 30% 属重病急病, 10% 是需要抢救的病人. 该门诊部的服务规则是先治疗抢救病人, 然后重病或急病人, 最后一般病人. 属同一级别的病人, 按到达先后次序进行治疗. 当该门诊部分别有一名医生和两名医生就诊时, 试分别计算各类病人等待治疗的平均等待时间.

解 假设需要抢救的病人属于第一类, 重病急病病人属于第二类, 一般病人属于第三类, 根据条件, $\mu = 6, \lambda = 2$ 人/小时, 于是有 $\lambda_1 = 0.2, \lambda_2 = 0.6, \lambda_3 = 1.2$.

(1) 当有一名医生就诊时, 有

$$W_1 = \frac{1}{\mu - \lambda_1} = \frac{1}{3 - 0.2} = 0.357 \text{ (小时)}$$
$$\overline{W}_{12} = \frac{1}{\mu - (\lambda_1 + \lambda_2)} = \frac{1}{3 - 0.8} = 0.454 \text{ (小时)}$$
$$\overline{W}_{123} = \frac{1}{\mu - (\lambda_1 + \lambda_2 + \lambda_3)} = \frac{1}{3 - 2} = 1 \text{ (小时)}$$

由此, 可得

$$W_2 = \frac{0.6 + 0.2}{0.6} \times 0.454 - \frac{0.2}{0.6} \times 0.357 = 0.486 \text{ (小时)}$$
$$W_3 = \frac{1.2 + 0.6 + 0.2}{1.2} \times 1 - \frac{0.2}{1.2} \times 0.357 - \frac{0.6}{1.2} \times 0.454 = 1.379 \text{ (小时)}$$

所以, $w_{q1} = 0.357 - 0.333 = 0.024$ (小时), $W_{q2} = 0.486 - 0.333 = 0.153$ (小时), $W_{q3} = 1.379 - 0.333 = 1.046$ (小时).

(2) 有两名医生就诊时, 有

$$W=\frac{\dfrac{(\frac{\lambda}{\mu})^2(\frac{\lambda}{2\mu})}{2\lambda(1-\frac{\lambda}{2\mu})^2}}{1+(\frac{\lambda}{\mu})+\frac{1}{2}(\frac{\lambda}{\mu})^2\dfrac{1}{(1-\frac{\lambda}{2\mu})}}+\frac{1}{\mu}=\frac{\dfrac{\lambda^2}{\mu(2\mu-\lambda)^2}}{\left[1+\dfrac{\lambda}{\mu}+\dfrac{\lambda^2}{\mu(2\mu-\lambda)}\right]}+\frac{1}{\mu}$$

$$W_1=\frac{\dfrac{(0.2)^2}{3\times(6-0.2)^2}}{1+\dfrac{0.2}{3}+\dfrac{(0.2)^2}{3\times(6-0.2)}}+\frac{1}{3}=0.33370\ (\text{小时})$$

$$\overline{W}_{12}=\frac{\dfrac{(0.8)^2}{3\times(6-0.8)^2}}{1+\dfrac{0.8}{3}+\dfrac{(0.8)^2}{3\times(6-0.8)}}+\frac{1}{3}=0.3391\ (\text{小时})$$

$$\overline{W}_{123}=\frac{\dfrac{2^2}{3\times(6-2)^2}}{1+\dfrac{2}{3}+\dfrac{2^2}{3\times(6-2)}}+\frac{1}{3}=0.375\ (\text{小时})$$

故

$$W_2=\frac{0.6+0.2}{0.6}\times0.3391-\frac{0.2}{0.6}\times0.33370=0.341\ (\text{小时})$$

$$W_3=\frac{1.2+0.6+0.2}{1.2}\times0.375-\frac{0.2}{1.2}\times0.3370-\frac{0.6}{1.2}\times0.3391=0.3999\ (\text{小时})$$

所以, $W_{q1}=0.00037$ (小时), $W_{q2}=0.0077$ (小时), $W_{q3}=0.0666$ (小时).

5.8 排队网络

直观地讲, 网络是由节点和弧组成的. 在排队网络中, 节点处设置一个或多个服务台. 假定在节点 i 处有 $s_i(i=1,2,\cdots,m)$ 个服务台. 通常顾客可以从外界进入系统中的任一节点, 也可以从任一节点处离开系统. 进入系统的顾客在完成一个服务后可以在节点之间转移. 例如, 自动线上的各道工序, 一个加工件必须依次得到服务后才能离开生产线. 因此, 为获得排队网络整体的平均等待时间, 平均顾客数等, 必须考虑整个排队网络. 下面, 对此只作简单介绍.

以下性质是成立的: 设一服务节点有 s 个服务台, 输入过程为服从参数 λ 的 Poisson 流, 无限排队, 每一服务台服务时间服从参数为 μ 的负指数分布 (M/M/s 模型), $\lambda < s\mu$. 则该服务节点的稳态输出是参数为 λ 的 Poisson 流.

注意到以上性质对服务规则没有任何约束, 可以是先来先服务原则, 随机原则, 或是优先权原则等. 由以上性质知, 接受完节点 1 服务的顾客以 Poisson 流离开该节点. 如果该顾客必须进入另一服务节点继续接受服务, 则节点 2 的输入过程也为 Poisson 流. 如节点 2 的各服务台服务时间服从负指数分布, 则以上性质对节点 2 也成立. 依次可以类推下去.

5.8.1 串联排队网络

此处讨论的串联排队网络有 m 个串联的服务节点, 每服务节点 i 处有 $s_i(i = 1, 2, \cdots, m)$ 个服务台. 顾客以服从参数 λ 的 Poisson 过程到达第 1 个服务节点, 然后依次经过各服务节点, 最后直到在第 m 个服务节点结束服务后离去. 假设每一服务节点 i 处各服务台的服务时间都服从参数为 μ_i 的负指数分布, 且 $\lambda_i < s_i\mu_i$, 于是在以上稳态条件下可知, 各服务节点的输出过程为 Poisson 流. 因此, 基本的 M/M/s 模型可以用来单独分析网络中各服务节点.

利用 M/M/s 模型中 P_n 可以得出在某一服务节点处有 n 个顾客的概率. 于是, 节点 1 有 n_1 个, 节点 2 有 n_2 个 $\cdots$, 节点 m 有 n_m 个顾客的联合概率 $P((N_1, N_2, \cdots, N_m) = (n_1, n_2, \cdots, n_m))$ 为

$$P((N_1, N_2, \cdots, N_m) = (n_1, n_2, \cdots, n_m)) = P_{n_1} P_{n_2} \cdots P_{n_m}$$

以上形式的解称为乘积形式解. 类似地, 把各个服务节点处所求得的平均等待时间、平均顾客数相加, 可求得整个系统的平均等待时间及平均顾客数.

然而, 在串联排队网络中, 如果附加第 $2, \cdots, k$ 个节点前等待空间有限这个约束, 则此时系统会发生阻塞现象. 在自动生产线的设计中, 各道工序前的等待空间总是有限的 (受场地面积, 管理费用等限制), 因此阻塞现象总可能发生, 合理设计等待空间的大小就是一个非常重要的问题. 而且在有限排队的约束下, 以上的性质不再成立. 但这类模型的分析是很复杂的, 可参见文献 [19].

5.8.2 Jackson 网络

另外一个具有乘积形式解的重要网络是 Jackson 网络.

假定排队网络由 $i(i = 1, 2, \cdots, m)$ 个节点组成, 满足以下条件的网络为 Jackson 网络.

(1) 无限顾客源.

(2) 从外部到达节点 i 的输入是参数 a_i 的独立 Poisson 过程.

(3) 节点 i 处有 c_i 个服务台, 每个服务台对顾客的服务时间都服从参数 μ_i 的负指数分布.

(4) 在节点 i 处服务完的顾客以概率 p_{ij} (与状态无关) 转入节点 $j(j = 1, 2, \cdots, m)$, 以概率 $q_i = 1 - \sum\limits_{j=1}^{m} p_{ij}$ 离开系统.

在稳态条件下, Jackson 网络中的节点 $j(j = 1, 2, \cdots, m)$ 的到达率为

$$\lambda_j = a_j + \sum_{i=1}^{m} \lambda_i p_{ij}$$

其中 $\lambda_j < s_j \mu_j$.

【例 5.11】 考虑如下一 Jackson 网络, 假设有 3 个节点, 已知条件见表 5.3.

表 5.3 Jackson 网络

节点 j	s_j	μ_j	a_j	概率 p_{ij}		
				$i=1$	$i=2$	$i=3$
$j=1$	1	10	1	0	0.1	0.4
$j=2$	2	10	4	0.6	0	0.4
$j=3$	1	10	3	0.3	0.3	0

解 根据前面的介绍, 有

$$\begin{aligned}
\lambda_1 &= 1 + 0.1\lambda_2 + 0.4\lambda_3 \\
\lambda_2 &= 4 + 0.6\lambda_1 + 0.4\lambda_3 \\
\lambda_3 &= 3 + 0.3\lambda_1 + 0.3\lambda_2
\end{aligned}$$

解以上方程组, 得 $\lambda_1 = 5$, $\lambda_2 = 10$, $\lambda_3 = \frac{15}{2}$. 以上三个服务节点可利用 M/M/s 排队模型的结论单独进行分析. 利用 $\rho_i = \frac{\lambda_i}{s_i \mu_i}$ 可得, $\rho_1 = \frac{1}{2}$, $\rho_2 = \frac{1}{2}$, $\rho_3 = \frac{3}{4}$.

于是, 对服务节点 1, 有 $P_{n_1} = \frac{1}{2} \times (\frac{1}{2})^{n_1}$; 对服务节点 2, 当 $n_2 = 0$ 时, $P_{n_2} = \frac{1}{3}$; 当 $n_2 = 1$ 时, $P_{n_2} = \frac{1}{3}$; 当 $n_2 = 2$ 时, $P_{n_2} = \frac{1}{3} \times (\frac{1}{2})^{n_1}$; 对服务节点 3, $P_{n_3} = \frac{1}{4} \times (\frac{3}{4})^{n_3}$.

从而可得 (n_1, n_2, n_3) 的联合概率为

$$P((N_1, N_2, N_3) = (n_1, n_2, n_3)) = P_{n_1} P_{n_2} P_{n_3}$$

类似可得, 在第 i 个服务节点前等待的平均队长 L_i 有, $L_1 = 1$, $L_2 = \frac{4}{3}$, $L_3 = 3$. 于是在整个排队网络中排队等待的平均队长 $L = L_1 + L_2 + L_3 = 5\frac{1}{3}$.

接下来, 利用 Little 公式可求得整个排队网络的顾客平均逗留时间 W 为

$$W = \frac{L}{\lambda} = \frac{L}{a_1 + a_2 + a_3} = \frac{5\frac{1}{3}}{8} = \frac{2}{3}$$

5.9 经济分析 —— 系统的最优化

5.9.1 排队系统的最优化问题

所谓的排队系统的优化设计是指设计一个未来的排队系统, 使适当的利益指标函数最优化. 常用的利益指标有: 稳态系统单位时间的平均总费用或平均总利润. 稳态系统单位时间的平均总费用由服务费用和等待费用构成. 一般情况下, 提高服务水平 (数量, 质量) 自然会降低顾客的等待费用 (损失), 但却常常增加了服务机构的成本, 优化的目标之一就是使两者费用之和为最小, 并确定达到最优目标值的最优的服务水平. 如图 5.12 所示.

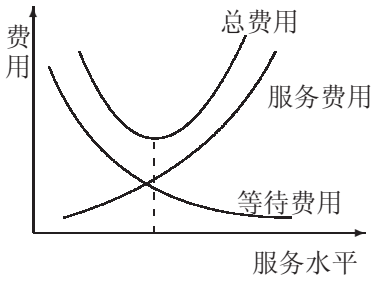

图 5.12 服务系统描述

服务水平可以由不同形式来表示, 如平均服务率 μ, 服务台的个数 s, 系统容量 K, 及服务强度 ρ 等. 于是以上平均总费用函数是关于平均服务率 μ, 服务台数 s, 系统容量 K 等决策变量的函数. 由于 μ 连续, s, K 离散, 因而决策变量类型复杂, 再加上利益指标函数的形式也很复杂, 所以这类优化稳态的求解很复杂. 通常采用数值法并需要在计算机上实现. 对于少数能够采用解析法求解的, 如对离散变量常用边际分析法, 对连续变量常用经典的微分法, 有时需要采用非线性规划或动态规划等方法.

5.9.2 M/M/1 模型中最优服务率 μ

5.9.2.1 标准模型

先考虑 M/M/1/∞ 排队模型, 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

$$z = c_s\mu + c_wL \tag{5.22}$$

其中, c_s 为当 $\mu = 1$ 时服务机构单位时间的平均费用, c_w 为每个顾客在系统中逗留单位时间的费用.

将 M/M/1/∞ 模型中 $L = \dfrac{\lambda}{\mu - \lambda}$ 代入上式, 可得

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

为了求极小, 先求 $\dfrac{dz}{d\mu}$, 然后令它为 0, 即

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

解出最优服务率为

$$\mu^* = \lambda + \sqrt{\frac{c_w}{c_s}\lambda} \tag{5.23}$$

根号前取 + 号是为了保证 $\rho < 1, \mu > \lambda$, 因为只有这样系统才会达到稳态.

于是, 最小平均总费用为

$$z^* = c_s\lambda + 2\sqrt{c_sc_w\lambda}$$

另外, 若设 c_w 为平均每个顾客在队列中等待单位时间的损失, 则需用 $L_q = \dfrac{\lambda^2}{\mu(\mu - \lambda)}$ 取代 (5.22) 式中的 L, 这时类似可得一阶条件

$$c_s\mu^4 - 2c_s\lambda\mu^3 + c_s\lambda^2\mu^2 - 2c_w\lambda^2\mu + c_w\lambda^3 = 0$$

这是一个关于 μ 的四次方程, 尽管有求根公式, 但过于复杂, 通常都采用数值法 (如 Newton 法) 确定其根 μ^*.

5.9.2.2 系统中顾客最大限制数为 K 的情形

下面针对 M/M/1/K 模型, 从使服务机构利润最大化的角度来考虑. 根据前面的讨论可知, 如系统中已有 K 个顾客, 则后来的顾客不能再进入该系统, 即 P_K 为被拒绝的概率, $1-P_K$ 即为能接受服务的概率. 在平稳状态下, 单位时间内到达并能够进入排队系统的平均顾客数为 $\lambda(1-P_K)$, 它也等于单位时间内实际服务完的平均顾客数.

设每服务一名顾客服务机构能收入 G 元. 于是, 单位时间收入的期望值为 $\lambda(1-P_K)G$ 元. 纯利润 z 为

$$\begin{aligned} z &= \lambda(1-P_K)G - c_s\mu = \lambda G \cdot \frac{1-\rho^K}{1-\rho^{K+1}} - c_s\mu \\ &= \lambda\mu G \cdot \frac{\mu^K - \lambda^K}{\mu^{k+1} - \lambda^{K+1}} - c_s\mu \end{aligned}$$

求 $\frac{dz}{d\mu}$ 并令 $\frac{dz}{d\mu}$=0, 可得如下方程:

$$\rho^{K+1} \cdot \frac{K-(K+1)\rho+\rho^{K+1}}{(1-\rho^{K+1})^2} = \frac{c_s}{G} \tag{5.24}$$

最优解 μ^* 应满足 (5.24) 式, (5.24) 式中 c_s, G, λ, K 都是给定的, 但要由上式中解出 μ^* 是很困难的. 通常是通过数值计算来求 μ^* 的, 或将上式左方 (对一定的 K) 作为 ρ 的函数作出图形, 对于给定的 $\frac{c_s}{G}$, 根据图形可求出 $\frac{\mu^*}{\lambda}$.

【例 5.12】 设有一电话亭, 其到达率为每小时 12 位顾客. 假定每一位接受服务的顾客其等待费用为每小时 5 元. 服务成本为每位顾客 2 元. 欲使总平均总费用最小化的服务率应为多少?

解 设以一小时为时间单位, 可得出 $\lambda = 12$ 人/小时, $c_w = 5$, $c_s = 2$ 元. 要求最小成本的服务率 μ^*, 可由 (5.23) 式求得, 于是, 有

$$\mu^* = \lambda + \sqrt{\frac{c_w}{c_s}\lambda} = 12 + \sqrt{\frac{5}{2} \times 12} = 17.5 \text{ (人/小时)}$$

且最小的总系统费用为 $z^* = c_s\lambda + 2\sqrt{c_s c_w \lambda} = 2 \times 12 + 2\sqrt{2 \times 5 \times 12} \approx 46$ 元.

所以, 管理当局应选取每 2 小时可服务 11 位顾客的服务设施, 且其成本为 $2 \times 5.5 = 11$ 元/每小时.

【例 5.13】 对某服务台进行实测, 得到数据如表 5.4 所示. 平均服务时间为 10 分钟, 服务一个顾客的收益为 2 元, 服务机构运行单位时间成本为 1 元, 问服务率为多少时可使单位时间平均总收益最大?

解 首先通过实测数据估计平均到达率 λ. 由于该系统为 M/M/1/3 系统, 故有 $\frac{P_n}{P_{n-1}}=\rho$. 因此, 可用下式来估计 ρ, 即

表 5.4 某服务系统中实测的顾客数据

系统中的顾客数 (n)	0	1	2	3
记录到的次数 (m_n)	161	97	53	34

$$\widehat{\rho}=\frac{1}{3}\sum_{n=1}^{3}\frac{m_n}{m_{n-1}}=\frac{1}{3}(0.60+0.55+0.64)=0.60$$

由 $\mu=6$ 人/小时, 可得 λ 的估计值为 $\widehat{\lambda}=\widehat{\rho}\mu=0.6\times 6=3.6$ 人/小时. 为求最优服务率, 根据 (5.24) 式, 取 $K=3,\frac{c_s}{G}=\frac{1}{2}=0.5$, 可求得 $\rho^*=1.21$. 故

$$\mu^*=\frac{\widehat{\lambda}}{\rho^*}=\frac{3.6}{1.21}=3\ (\text{人/小时})$$

下面进行收益分析. 当 $\mu=6$ 人/小时时, 总收益为

$$z=2\times 3.6\times\frac{1-0.6^3}{1-0.6^4}-1\times 6=0.485\ (\text{元/小时})$$

当 $\mu=3$ 人/小时时, 总收益为

$$z=2\times 3.6\times\frac{1-1.21^3}{1-1.21^4}-1\times 6=1.858\ (\text{元/小时})$$

单位时间内平均收益可增加 $1.858-0.485$=1.373 元.

【例 5.14】 考虑一个 M/M/1/K 系统, 具有 $\lambda=10$ 人/小时, $\mu=30$ 人/小时, $K=2$. 管理者想改进服务机构, 方案有两个: 方案 A 是增加一个等待空间, 即使 $K=3$; 方案 B 是提高平均服务率到 $\mu=40$ 人/小时. 设每服务一个顾客的平均收入不变, 问哪个方案将获得更大的收入? 当 λ 增加到 30人/小时时, 又将得到什么结果?

解 对方案 A, 单位时间内实际进入系统的顾客的平均数为

$$\lambda_A = \lambda(1-P_3) = \lambda\Big(\frac{1-\rho^3}{1-\rho^4}\Big) = 10 \times \frac{1-\big(\frac{1}{3}\big)^3}{1-\big(\frac{1}{3}\big)^4} = 9.75 \text{ (人/小时)}$$

对方案 B, 当 $\mu = 40$ 人/小时时, 单位时间内实际进入该系统的顾客的平均数为

$$\lambda_A = \lambda(1-P_2) = \lambda\Big(\frac{1-\rho^2}{1-\rho^3}\Big) = 10 \times \frac{1-\big(\frac{1}{4}\big)^2}{1-\big(\frac{1}{4}\big)^3} = 9.52 \text{ (人/小时)}$$

因此, 采取扩大等待空间将获得更多的利润.

当 λ 增加到 30 人/小时时, 从有关公式中关于 $\rho = 1$ 的结果, 有

$$\lambda_A = 30 \times \frac{3}{3+1} = 22.5 \text{ (人/小时)}, \quad \lambda_B = 30 \times \frac{1-\big(\frac{3}{4}\big)^2}{1-\big(\frac{3}{4}\big)^3} = 22.7 \text{ (人/小时)}$$

因此, 当 λ 增加到 30 人/小时时, 采取提高服务率到 $\mu = 40$ 人/小时将会得到更多的收益.

5.9.3 M/M/s 模型中最优的服务台数

下面仅讨论 M/M/s/∞ 模型, 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) 的期望值为

$$z = c_s' \cdot c + c_w \cdot L \tag{5.25}$$

其中, s 是服务台数, c_s' 是每个服务台单位时间内的总费用, c_w 为每个顾客在系统停留单位时间的费用, L 是系统中的顾客平均数, 也可把 L 换成是系统中等待的顾客平均数 L_q.

显然, 它们都随 s 值的不同而不同. 因为 c_s' 和 c_w 都是给定的, 惟一可能变动的是服务台数 s, 所以 z 是 s 的函数 $z(s)$, 现在是求最优解 s^* 使 $z(s^*)$ 为最小.

因为 s 只能取整数值, 于是 $z(s)$ 不是连续变量的函数. 采用边际分析法 (marginal analysis), 根据 $z(s^*)$ 是最小的特点, 有

$$z(s^*) \leqslant z(s^*-1), \quad z(s^*) \leqslant z(s^*+1)$$

将 (5.25) 式中 z 代入, 得

$$\begin{cases} c_s's^* + c_wL(s^*) \leqslant c_s'(s^*-1) + c_wL(s^*-1) \\ c_s's^* + c_wL(s^*) \leqslant c_s'(s^*+1) + c_wL(s^*+1) \end{cases} \tag{5.26}$$

上式化简后, 得

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

依次求 $s = 1, 2, 3, \cdots$ 时 L 的值, 并做两相邻的 L 值之差, 因 $\frac{c_s'}{c_w}$ 是已知数, 根据这个数落在哪个不等式的区间里就可定出 s^*.

【例 5.15】 某检验中心为各工厂服务, 要求做检验的工厂 (顾客) 的到来服从 Poisson 流, 平均到达率 λ 为每天 48 次, 每次来检验由于停工等原因损失为 6 元. 服务 (做检验) 时间服从负指数分布, 平均服务率 μ 为每天 25 次, 每设置 1 个检验员服务成本 (工资及设备损耗) 为每天 4 元. 其他条件适合标准 M/M/c 的模型, 问应设几个检验员 (及设备) 才能使总费用的期望值为最小?

解 由条件可知, $c_s' = 4$ 元/每检验员, $c_w = 6$ 元/次, $\lambda = 48, \mu = 25, \frac{\lambda}{\mu} = 1.92$. 设检验员数为 s, 令 s 依次为 1, 2, 3, 4, 5, 根据表 5.5, 求出 L. 将 L 值代入 (5.26) 式得表 5.6.

表 5.5 某检验中心期望总费用最小计算过程 1

s	1	2	3	4	5
$\frac{\lambda}{s\mu}$	1.92	0.96	0.64	0.48	0.38
查表 $w_q \cdot \mu$	–	10.2550	0.3961	0.772	0.0170
$L = \frac{\lambda}{\mu}(w_q \cdot \mu + 1)$	–	21.610	2.680	2.068	1.952

$\frac{c_s'}{c_w} = 0.66$ 落在区间 $(0.612 \sim 18.930)$ 内, 所以 $s^* = 3$. 即以设 3 个检验员使总费用为最小, 直接代入也可验证 $z(s^*) = z(3) = 27.87$ 元为最小.

5.10 分析排队系统的随机模拟法

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.

表 5.6 某检验中心期望总费用最少计算过程 2

检验员数 s	来检验顾客数 $L(s)$	$L(s)-L(s+1)$ $L(s)-L(s-1)$	总费用 (每天) $z(s)$
1	∞		∞
2	21.610	$18.930 \sim \infty$	154.94
3	2.680	$0.612 \sim 18.930$	$27.87(*)$
4	2.068	$0.116 \sim 0.612$	28.38
5	1.952		31.71

【例 5.16】 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 3 车, 若一天内到达数当超过 3 车, 那么就推迟到次日卸货, 根据表 5.7 所示的经验, 货车到达数的概率分布 (相对频率) 平均为 2.4 车/天, 求每天推迟卸货的平均车数.

表 5.7 货车到达的相对频率表

到达车数	0	1	2	3	4	5	$\geqslant 6$
概率	0.05	0.30	0.30	0.10	0.05	0.20	0.00

解 这是单服务台的排队系统, 可验证到达车数不服从 Poisson 分布, 服务时间也不服从负指数分布 (这是定长服务时间), 不能用以前的方法求解.

随机模拟法首先要求能按经验的概率分布规律出现. 为此, 可利用随机数表. 表 5.8 就是一个 2 位数的随机数表的一部分. 在进行模拟求解时, 先按到达

表 5.8 2 位数的部分随机数表

到达车数	概率	累计概率	对应随机数	到达车数	概率	累计概率	对应随机数
0	0.05	0.05	$00 \sim 04$	3	0.10	0.75	$65 \sim 74$
1	0.30	0.35	$05 \sim 34$	4	0.05	0.80	$75 \sim 79$
2	0.30	0.65	$35 \sim 64$	5	0.20	1.00	$80 \sim 99$
$\sum$	1.00						

车数的概率分别来分配随机数, 见表 5.8, 然后开始模拟, 见表 5.9. 前 3 天作为模拟预备期, 日期记为 x. 然后依次是第 1 天, 第 2 天 $\cdots$, 第 50 天. 如第 1 天的随机数是 66, 由表 5.8 可知, 到达的车应为 3; 第 2 天得到的随机数是 96, 到达

的车数应为 5 等等. 如此, 一直到第 30 天. 将每天的随机数和应到达的车数记入表 5.9 的第 2 列和第 3 列, 然后计算出第 (4), (5), (6) 列的值. 公式是

$$\text{当天到达的车数(3)} + \text{前一天推迟卸货车数(6)} = \text{当天需要卸货车数}$$

$$\text{卸货车数(5)} = \begin{cases} \text{需要卸货数(4)}, & \text{当需要卸货车数} \leqslant 3 \\ 3, & \text{当需要卸货车数} > 3 \end{cases}$$

分析结果时, 不考虑前 3 天的预备阶段的数据. 这是为了使模拟从一个稳定过程中任意点开始, 否则, 如认为开始时没有积压就失去了随机性. 表 5.9 中给出了 30 天的模拟情况, 可以看出, 在 21 天里没有发生由于推迟卸货而造成的积压, 平均到达车数为 2.27, 比期望值略低, 平均每天有 0.7 车推迟卸货. 当然, 模拟时间越长, 结果会越准确. 这种方法适用于不同方案可能产生的结果进行比较, 并可以利用计算机进行模拟, 模拟方法只能得到数字结果, 不能得出解的解析表达式.

最后需要说明, 有关排队论模型的软件实现, 可在 LINGO 软件中找到相应的工具, 详细请参见 LINGO 软件的帮助文件.

习 题

5.1 判断下列说法是否正确:

(1) 若到达排队系统的顾客为 Poisson 流, 则依次到达的两名顾客之间的间隔时间服从负指数分布;

(2) 假如到达排队系统的顾客来自两个方面, 分别服从 Poisson 分布, 则这两部分顾客合起来的顾客流仍为 Poisson 分布;

(3) 若两两顾客依次到达的间隔时间服从负指数分布, 又将顾客按到达的先后排序, 则第 $1, 3, 5, 7, \cdots$ 名顾客到达的间隔时间也服从负指数分布;

(4) 对 M/M/1 或 M/M/s 的排队系统, 服务完毕离开系统的顾客流也为 Poisson 流;

(5) 在排队系统中, 一般假定对顾客服务时间的分布为负指数分布, 这是因为通过对大量实际系统的统计研究, 这样的假定比较合理;

(6) 一个排队系统中, 不管顾客到达和服务时间的情况如何, 只要运行足够长的时间后, 系统将进入稳定状态;

(7) 排队系统中, 顾客等待时间的分布不受排队服务规则的影响;

(8) 在顾客到达及机构服务时间的分布相同的情况下, 对容量有限的排队系统, 顾客的平均等待时间将少于允许队长无限的系统;