第五讲 排队论模型

【修理工录用问题】工厂平均每天有一台机器发生故障而需要修理，机器的故障数服从泊松分布。修理一台机器平均花费 20 元。现有技术水平不同的修理工人 A 和 B，A 种修理工平均每天能修理 1.2 台机器，每天工资 3 元；B 种修理工平均每天能修理 1.5 台机器，每天工资 5 元，两种修理工修理机器的时间为负指数分布。问工厂录用哪种工人较合算？

本讲主要内容

1. 排队论的基本概念
2. 单服务台的排队模型
3. 多服务台的排队模型
4. 排队系统的最优化问题
5. 数学建模实例：校园网的设计和调节收费问题

5.1 排队论的基本概念

5.1.1 什么是排队系统

排队论也称随机服务系统理论，它是 20 世纪初由丹麦数学家 Erlang 应用数学方法在研究电话话务理论过程中而发展起来的一门学科，在实际中有广泛的应用。它涉及的是建立一些数学模型，藉以对随机发生的需求提供服务的系统预测其行为。现实世界中排队的现象比比皆是，如到商店购货、轮船进港、病人就诊、机器等待修理等等。排队的内容虽然不同，但有如下共同特征：

（1）有请求服务的人或物，如候诊的病人、请求着陆的飞机等，我们将此称为“顾客”。

（2）有为顾客提供服务的人或物，如医生、飞机跑道等，我们称此为“服务员”。由顾客和服务员就组成服务系统。

（3）顾客随机地一个一个（或者一批一批）来到服务系统，每位顾客需要服务的时间不一定是确定的，服务过程的这种随机性造成某个阶段顾客排长队，而某些时候服务员又空闲无事。

为了叙述一个给定的排队系统，必须规定系统的下列组成部分：

1.输入过程　即顾客来到服务台的概率分布。排队问题首先要根据原始资料，由顾客到达的规律、作出经验分布，然后按照统计学的方法（如卡方检验法）确定服从哪种理论分布，并估计它的参数值。我们主要讨论顾客来到服务台的概率分布服从泊松分布，且顾客的达到是相互独立的、平稳的输入过程。所谓“平稳”是指分布的期望值和方差参数都不受时间的影响。

2.排队规则　即顾客排队和等待的规则。排队规则一般有即时制和等待制两种。所谓即时制就是服务台被占用时顾客便随即离去；等待制就是服务台被占用时，顾客便排队等候服务。等待制服务的次序规则有先到先服务、随机服务、有优先权的先服务等，我们主要讨论先到先服务的系统。

3.服务机构　服务机构可以是没有服务员的，也可以是一个或多个服务员的；可以对单

独顾客进行服务，也可以对成批顾客进行服务。和输入过程一样，多数的服务时间都是随机的，且我们总是假定服务时间的分布是平稳的。若以 ξ_n 表示服务员为第 n 个顾客提供服务所需的时间，则服务时间所构成的序列{ ξ_n}，n=1，2，…所服从的概率分布表达了排队系统的服务机制，一般假定，相继的服务时间 ξ_1， ξ_2，…是独立同分布的，并且任意两个顾客到来的时间间隔序列{T_n}也是独立的。

如果按服务系统的以上三个特征的各种可能情形来对服务系统进行分类，那么分类就太多了。因此，现在已被广泛采用的是按顾客相继到达时间间隔的分布、服务时间的分布和服务台的个数进行分类。

排队论主要是对服务系统建立数学模型，研究如下内容：

（1）排队系统的概率分布问题，主要是研究队长分布、等待时间分布和忙期分布等；

（2）最优化问题：分为静态最优化和动态最优化，即为系统的最优设计和系统的最优运行问题；

（3）排队系统的统计推断：判断一个给定的排队系统符合哪种模型，以便于根据排队理论进行分析研究。

5.1.2 排队模型的标准形式

排队模型的标准形式为 $X/Y/Z/A/B/C$，其中：

X　表示顾客来到时间间隔的分布类型；

Y　表示服务时间的分布类型；

Z　表示服务员个数；

A　系统容量；

B　顾客源个数；

C　服务规则.

例如先来先服务的等待排队模型主要由三参数法即 $X/Y/Z$，“$M/M/1/k/\infty$/FCFS”表示顾客到达间隔时间和服务时间均服从负指数分布，一个服务台，系统至多容纳 k 个顾客潜在的顾客数不限，先来先服务的排队系统。

“$M/M/$c”即 Poisson 输入，负指数服务时间分布，c 个服务台的等待制排队模型。

“$M/G/1$”即 Poisson 输入，一般服务时间分布，单个服务台的等待制排队模型。

5.1.3 排队系统的运行指标

研究排队问题的目的，是研究排队系统的运行效率，估计服务质量，确定系统参数的最优值，以决定系统的结构是否合理，设计改进措施等。所以，必须确定用来判断系统运行优劣的基本数量指标，这些数量指标通常是：

（1）**队长**　指排队系统中的顾客数，它的期望值记为 L_s；**排队长**，指在排队系统中排队等待服务的顾客数，其期望值记为 L_q。

系统中的顾客数 = 等待服务的顾客数 + 正被服务的顾客数

所以 L_q（或 L_s）越大，说明服务效率越低。

（2）**逗留时间**　指一个顾客在排队系统中的停留时间，即顾客从进入服务系统到服务完毕的整个时间。其期望值记为 W_s。**等待时间**，指一个顾客在排队系统中等待服务的时间，其期望值记为 W_q。

逗留时间 = 等待时间 + 服务时间

（3）**忙期**　指从顾客到达空闲服务机构起到服务机构再次为空闲这段时间长度，即服务机构连续工作的时间长度。它关系到服务员的工作长度，即服务机构连续工作的时间长度。

它关系到服务员的工作强度、忙期的长度和一个忙期中平均完成服务的顾客数，这些都是衡量服务效率的指标。

要计算以上这些指标必须知道系统状态的概率，所谓系统状态即时刻 t 时排队系统中的顾客数。如果时刻 t 时排队系统中有 n 个顾客，就说系统的状态是 n，其概率一般用 $P_n(t)$表示。求 $P_n(t)$的方法，首先要建立含 $P_n(t)$的关系式，因 t 为连续变量而 n 只取非负整数，所以建立的 $P_n(t)$的关系式一般是微分差分方程，这时要求方程的解是不容易的，有时即使求出也很难利用。因此，往往只求稳态解 P_n，求 P_n 并不一定求 $t\to\infty$时的 $P_n(t)$极限，而只需由 $P_n'(t)=0$，用 P_n 代替 $P_n(t)$即可。

5.2 单服务台的排队模型

设系统的输入过程服从泊松分布，服务时间服从负指数分布，单服务台的排队系统有以下三种情形：

（1）标准型：$M/M/1(M/M/1/\infty/\infty)$;

（2）系统容量有限制：$M/M/1/N/\infty$;

（3）顾客源为有限的：$M/M/1/\infty/m$.

5.2.1 标准型：M/M/1

$M/M/1$ 模型是指顾客源为无限，顾客到达相互独立，到达过程是平稳的，到达率数服从参数为 l 的泊松分布：单服务台、队长无限、先到先服务；各顾客的服务时间服从参数为 μ 的负指数分布，且相互独立。

首先求出排队系统在任意时刻 t 的、状态为 n 的概率 $P_n(t)$，已知顾客到达率服从参数为 λ 的泊松分布，服务时间服从参数为 μ 的负指数分布，由此决定了$[t，t+\triangle t]$时间间隔内：

（1）有 1 个顾客到达的概率为 $\lambda\triangle t+o(\triangle t)$，没有顾客到达的概率是 $1-\lambda\triangle t+o(\triangle t)$。

（2）当有顾客在接受服务时，1 个顾客被服务完了的概率是 $\mu\triangle t+o(\triangle t)$，没有服务完的概率是 $1-\mu\triangle t+o(\triangle t)$。

（3）多于一个顾客到达或服务完的概率为 $o(\triangle t)$，均可忽略。

注 1：因为单位时间内顾客到达数 $X\sim P(\lambda)$，所以 Δt 时间间隔内顾客到达数 $Y\sim P(\lambda\Delta t)$，因而在 Δt 时间间隔内有 1 个顾客到达的概率为：$P\{Y=1\}=\lambda\Delta t\cdot e^{-\lambda\Delta t}=\lambda\Delta t+o(\Delta t)$，没有顾客到达的概率为 $P\{Y=0\}=e^{-\lambda\Delta t}=1-\lambda\Delta t+o(\Delta t)$。

注 2：由于服务时间 $T\sim E(\mu)$，故在有顾客接受服务时，1 个顾客被服务完的概率为 $P\{T\leqslant\Delta t\}=1-e^{-\mu\Delta t}=\mu\Delta t+o(\Delta t)$，没有被服务完的概率为 $1-\mu\Delta t+o(\Delta t)$。

在 $t+\triangle t$ 时刻，系统中有 n 个顾客的状态由 t 时刻的以下状态转化而来：

①t 时刻系统中有 n 个顾客，没有顾客到达且没有顾客服务完毕，其概率为：$[1-\lambda\triangle t+o(\triangle t)][1-\mu\triangle t+o(\triangle t)]=(1-\lambda\triangle t-\mu\triangle t)+o(\triangle t)$;

②t 时刻系统中有 $n+1$ 个顾客，没有顾客到达且有 1 个顾客服务完毕，其概率为：$[1-\lambda\triangle t+o(\triangle t)][\mu\triangle t+o(\triangle t)]=\mu\triangle t+o(\triangle t)$;

③ t 时刻系统中有 $n-1$ 个顾客，有 1 个顾客到达且没有顾客服务完毕，其概率为：$[\lambda\triangle t+o(\triangle t)][1-\mu\triangle t+o(\triangle t)]=\lambda\triangle t+o(\triangle t)$;

④其他状态的概率为 $o(\triangle t)$。

因此，在 $t+\triangle t$ 时刻，系统中有 n 个顾客的概率 $P_n(t+\triangle t)$满足：

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

移项整理，两边同除以$\triangle t$，得

$$\frac{P_n(t+\Delta t)-P_n(t)}{\Delta t}=lP_{n-1}(t)+mP_{n+1}(t)-(l+m)P_n(t)+\frac{o(\Delta t)}{\Delta t}.$$

令△t→0，得

$$\frac{dP_n(t)}{dt}=lP_{n-1}(t)+mP_{n+1}(t)-(l+m)P_n(t)\quad n=1,2\mathbf{L}.$$

当 n=0 时，因为

$$P_0(t+\Delta t)=P_0(t)(1-l\Delta t)+P_1(t)(1-l\Delta t)m\Delta t+o(\Delta t)$$

所以有

$$\frac{dP_0(t)}{dt}=-lP_0(t)+mP_1(t).$$

对于稳态情形，与 t 无关，其导数为零。因此，得到

$$\begin{cases}lP_{n-1}+mP_{n+1}-(l+m)P_n=0, n>1\\ -lP_0+mP_1=0\end{cases}$$

这是关于 P_n 的差分方程，也反映出了系统状态的转移关系，即每一状态都是平衡的，求解得 $P_1=(l/m)P_0$，递推可得 $P_n=(l/m)^nP_0(n\geq 1)$.

由概率的性质知 $\sum_{n=0}^{\infty}P_n=1$，将上式代入 λ/μ<1 时可得到

$$P_0=1-l/m$$

$$P_n=(1-l/m)(l/m)^n.$$

因为顾客到达规律服从参数为 λ 的泊松分布，服务时间服从参数为 μ 的负指数分布，其期望值就分别为 λ，1/μ。所以 λ 表示单位时间内平均到达的顾客数，μ 表示单位时间内能服务完的顾客数。如果令 $\rho=\lambda/\mu$，这时 ρ 就表示相同时间内顾客到达的平均数与能被服务的平均数之比，它是刻画服务效率和服务机构利用程度的重要标志，称 ρ 为**服务强度**。上面在 ρ<1 的条件下得到了稳定状态下的概率 P_n，n=0，1，2，…。其实，如果 ρ>1，可以证明排队长度将是无限增加的，即使 ρ=1 的情况下，$P_0(t)$也是随时间而变化的，系统达不到稳定状态. 因此，这里只讨论 ρ<1 时情况，从上面的推导知

$$P_n=(1-\rho)\rho^n \qquad n=0,1,2,\cdots.$$

下面计算出系统的运行指标.

(1) 队长（平均顾客数）：由于系统的状态为 n 时即系统中有 n 个顾客，由期望的定义得

$$L_s=\sum_{n=0}^{\infty}np_n=\sum_{n=1}^{\infty}n(1-r)r^n=r/(1-r)=l/(m-l).$$

(2) 排队长：（等待的平均顾客数）

$$L_q = \sum_{n=1}^{\infty}(n-1)p_n = \sum_{n=1}^{\infty}(n-1)r^n(1-r)$$

$$= r^2/(1-r)$$

$$= r\,l\,/(m-l).$$

可以证明，顾客在系统中逗留时间服从参数为 μ-l 的负指数分布。因此，有

(3) 系统中顾客的平均逗留时间：$W_s = 1/(m-l)$.

(4) 系统中顾客的平均等待时间：$W_q = W_s - \frac{1}{m} = r/(m-l)$

由以上结论可以看出，各指标之间有如下关系：

$$L_s = l\,W_s;\quad L_q = l\,W_q;$$

$$W_s = W_q + 1/m, L_s = L_q + l/m\ ,$$

这些关系式称为 Little 公式.在指标的计算过程中，一般只要计算其中一个，其它的指标便可随之导出。

5.2.2 系统容量有限制：M/M/1/N/∞

因为是单服务台，排队系统的容量为 N，即是排队等待的顾客最多为 N-1，在某时刻一顾客到达时，如系统中已有 N 个顾客，那么这个顾客就被拒绝进入系统。假设顾客平均到达率为 l，平均服务率为 μ，在研究系统中有 n 个顾客的概率 $P_n(t)$时，和标准型 $M/M/1$ 研究方法相同，当 $n=N$ 时有

$$P_N^{'}(t) = l\,P_{N-1}(t) - mP_N(t)$$

在稳态情形下，令 $r = \frac{l}{m}$，得

$$\begin{cases} P_1 = rP_0 \\ P_{n+1} + rP_{n-1} = (1+r)P_n, n = 1,2,\mathbf{L}\,, N-1 \\ P_N = rP_{N-1} \end{cases}$$

在条件 $\sum_{i=0}^{N} P_i = 1$ 下解上式得到

$$\begin{cases} P_0 = \dfrac{1-r}{1-r^{N+1}}, r \neq 1 \\ P_n = \dfrac{1-r}{1-r^{N+1}}\,r^n \qquad 1 \le n \le N. \end{cases}$$

注：

（1）如果 ρ=1（即 $l=\mu$），由 $P_0 = P_1 = \mathbf{L} = P_N = \frac{1}{N+1}$，即到达率和服务率相等，在稳态情况下系统不会出现排队等待现象；

（2）这里因为是有限项的和，所以不要求 ρ<1，但当 ρ>1（l>μ）时，表示单位时间内到达率大于服务率，系统的损失率增加，即被拒绝排队的数量增大.

下面给出系统的各种指标的计算结果：

（1） 队长：

$$L_s = \sum_{n=0}^{N} nP_n = \sum_{n=0}^{N} \frac{n}{N+1} = \frac{N}{2}, r = 1$$

$$L_s = \sum_{n=1}^{N} np_n = \sum_{n=0}^{N} \frac{n(1-r)r^n}{1-r^{N+1}} = \frac{r}{1-r} - \frac{(N+1)r^{N+1}}{1-r^{N+1}}, r \neq 1$$

（2）排队长：$L_q = \sum_{n=1}^{N} (n-1)P_n = L_s - (1-P_0)$

$$= \begin{cases} \frac{N}{2} - \frac{N}{N+1}, & r = 1 \\ \frac{r}{1-r} - \frac{Nr^{N+1} - r}{1-r^{N+1}}, & r \neq 1 \end{cases}$$

（3）逗留时间： $W_s = L_s / [m(1-P_0)]$；

（4）等待时间：$W_q = W_s - \frac{1}{m}$.

应该指出，W_s，W_q 的导出过程中不是采用平均达到率 λ，而是采用有效到达率 $\lambda_{效}$。这主要是由于当系统已满时，顾客的实际到达率为零，因为正在被服务的顾客的平均数为 $1-P_0 = l_{效}/m$，于是 $l_{效} = m(1-P_0)$.

5.2.3 顾客源为有限的：M/M/1/∞/m

对该模型的顾客总体虽只有 m 个顾客，但每个顾客的到来并接受服务后，仍然回到顾客总体，即可以再次到来，所以对系统的容量是没有限制的，实际上系统中的顾客数永远不会超过 m，即与模型 $M/M/1/m/m$ 的意义相同。

与前面情况类似，假设每个顾客的到达率相同为 l，在系统外的平均顾客数为 $m-L_s$,故系统的有效到达率为 $l_e=l(m-L_s)$.考虑稳态的情况，可得系统状态概率的平衡方程为

$$\begin{cases} mP_1 = ml P_0, \\ mP_{n+1} + (m-n+1)l P_{n-1} = [(m-n)l + m]P_n, (1 \le n \le m-1) \\ mP_m = l P_{m-1}. \end{cases}$$

注意到$\sum_{n=0}^{m} P_n = 1$，由递推关系不难求得系统状态的概率为

$$\begin{cases} P_0 = \dfrac{1}{\sum_{i=0}^{m} \dfrac{m!}{(m-i)!}\left(\dfrac{l}{m}\right)^i}, \\ P_n = \dfrac{m!}{(m-n)!}\left(\dfrac{l}{m}\right)^n P_0, 1 \le n \le m. \end{cases}$$

该系统的运行指标为

$$L_s = \sum_{n=1}^{N} np_n = m - \frac{m}{l}(1-P_0), W_s = \frac{m}{m(1-P_0)} - \frac{1}{l}, W_q = W_s - \frac{1}{m},$$

$$L_q = \sum_{n=1}^{m} (n-1)P_n = m - \frac{(l+m)(1-P_0)}{l} = L_s - (1-P_0).$$

【例 5-1】病人候诊问题 某单位医院的一个科室有一位医生值班，经长期观察，每小时平均有 4 个病人，医生每小时平均可诊 5 个病人，病人的到来服从泊松分布，医生的诊病时间服从负指数分布。试分析该科室的工作状况。如果满足 99%以上的病人有座，此科室至少应设多少个座位？如果该单位每天 24h 上班，病人看病 1h 因耽误工作单位要损失 30 元，这样单位平均每天损失多少元？如果该科室提高看病速度，每小时平均可诊 6 个病人，单位每天可减少损失多多少？可减少多少个座位？

解 由题意知 λ =4， μ =5， ρ =4/5， ρ =4/5=0.8<1，从而排队系统的稳态概率为：

$$P_n = 0.2 \times 0.8^n，n=0，1，2\cdots$$

该科室平均有病人数为：$L_s = r/(1-r) = 0.8/(1-0.8) = 4$（人）

该科室内排队候诊病人的平均数为：$L_s = L_q - l/m = 4 - 0.8 = 3.2$（人）

看一次病平均所需的时间为：$W_s = L_s / l = 4/4 = 1h$

排队等候看病的平均时间为：$W_q = W_s - 1/m = 1 - 1/5 = 0.8h$

为满足 99%以上的病人有座，设科室应设 m 个座位，则 m 应满足：

P{医务室病人数≤m}≥0.99

$$\sum_{n=0}^{m} r^n(1-r) = 1 - r^{m+1} \ge 0.99$$

$$r^{m+1} \le 0.01$$

$$m \ge \frac{\ln 0.01}{\ln r} - 1 = 20$$

所以该科室至少应设 20 个座位。

如果该单位 24h 上班，则每天平均有病人 24×4=96 人，病人看病所花去的总时间为 96×1=96 h。因看病平均每天损失 30×96=2880 元。

如果医生每小时可诊 6 个病人， ρ =2/3，则

$$L_s=2\text{（人）}，\ L_q=4/3\text{（人）}$$

$$W_s=0.5\text{h}，\ W_q=1/3\text{h}，$$

这样单位每天的损失费为 96×0.5×30=1440 元，因而单位每天平均可减少损失 2880-1440=1440 元，这时为保证 99%以上的病人有座，应设座位数 m≥ln0.01/ln(2/3)-1=11 个，比原来减少了 9 个。

【例 5-2】单人理发馆有 6 个椅子，当 6 个椅子都坐满时，后来到的顾客不进店就离开。顾客平均到达率为 3 人/h，理发平均需 15min，试分析该服务系统。

解 由题意知 N=7， λ =3 人/h， μ =4 人/h，因此，某顾客一到达就能理发的概率为：

$$P_0=(1-3/4)(1-(3/4)^8)=0.2778$$

平均需要等待的顾客数量为：$L_s=\dfrac{3/4}{1-(3/4)}-\dfrac{8(3/4)^8}{1-(3/4)^8}=2.11$ 人

$$L_q=L_s-(1-P_0)=2.11-(1-0.2778)=1.39\text{ 人}$$

有效到达率为：$l_{效}=m(1-P_0)=4(1-0.2778)=2.89$ 人/h.

顾客在理发馆平均逗留时间为：$W_s=L_s/l_{效}=\dfrac{2.11}{1.89}=0.73h=43.8\min$.

5.3 多服务台的排队模型

这里研究单队列、并列的 C 个服务台的情形，同单服务台类似，讨论如下三种模型：

（1）标准型：$M/M/\text{C}(M/M/\text{C}/\infty/\infty)$;

（2）系统容量有限制：$M/M/\text{C}/N/\infty$;

（3）顾客源为有限的：$M/M/\text{C}/\infty/m$.

5.3.1 标准型：M/M/C（M/M/C/∞/∞）

前提假设同 $M/M/1/\infty/\infty$，顾客流为泊松流，平均到达率为 l，各服务台的服务时间满足负指数分布，而各而各服务台的工作是相互独立的（不搞协作），单个服务台的平均服务率为 μ，则整个服务机构的平均服务率为 Cμ(当 $n\geq$C),或 $n\mu$(当 $n<$C),则系统的服务强度为 $r=l/C\mu$, 当$r>1$ 时，系统就会出现排队现象。

类似地，可以得到系统状态概率的平衡方程

$$\begin{cases} mP_1=lP_0, \\ (n+1)mP_{n+1}+lP_{n-1}=(l+nm)P_n,(1\leq n\leq C) \\ CmP_{n+1}+lP_{n-1}=(l+Cm)P_n,(n>C). \end{cases}$$

其中 $\sum_{n=0}^{\infty} P_n = 1$，且 $r = \frac{l}{Cm} \le 1$，由递推关系可得系统状态概率

$$P_0 = [\sum_{k=0}^{C-1} \frac{1}{k!}(\frac{l}{m})^k + \frac{1}{C!}\frac{1}{1-r}(\frac{l}{m})^C]^{-1}$$

$$P_n = \begin{cases} \frac{1}{n!}(\frac{l}{m})^n P_0, & n \le C \\ \frac{1}{C!C^{n-C}}(\frac{l}{m})^n P_0, & n > C \end{cases}$$

系统的运行指标为

$$L_s = L_q + \frac{l}{m}, \quad L_q = \sum_{n=C+1}^{\infty}(n-C)P_n = \sum_{k=1}^{\infty} kP_{k+C} = \frac{(Cr)^C r}{C!(1-r)^2}P_0,$$

$$W_q = L_q / l \ , \quad W_s = W_q + 1/m = L_s / l \ .$$

5.3.2 系统容量有限制：M/M/C/N/∞

假设系统中有 C 个服务台，系统的最大容量为 $N(N \ge C)$,其它假设同前面一样。当系统客满（即系统中有 N 个顾客）时，有 C 个接受服务，N-C 个在排队，再有顾客到来将被拒绝而离去，系统将有损失率。

当系统的状态为 n 时，每个服务台的服务率为 μ,则系统的的总服务率：当 $0<n<C$ 时为 $n\mu$；当 $n \ge C$ 时为 $C\mu$，系统的服务强度为 $r = l / Cm$.

类似地，可以得到系统的状态概率平衡方程

$$\begin{cases} mP_1 = lP_0, \\ (n+1)mP_{n+1} + lP_{n-1} = (l+nm)P_n, (1 \le n \le C) \\ CmP_{n+1} + lP_{n-1} = (l+Cm)P_n, (C \le n < N) \\ lP_{N-1} = CmP_N. \end{cases}$$

其中 $\sum_{n=0}^{N} P_n = 1$，且 $r = \frac{l}{Cm} \le 1$，由递推关系可得系统状态概率

$$P_0 = \begin{cases} \left[\sum_{k=0}^{C-1}\frac{1}{k!}(Cr)^k + \frac{C^C}{C!}\frac{r(r^C - r^N)}{1-r}\right]^{-1}, r \ne 1, \\ \sum_{k=0}^{C-1}\frac{1}{k!}(C)^k + \frac{C^C}{C!}(N-C+1), r = 1, \end{cases}$$

$$P_n=\begin{cases}\frac{1}{n!}(Cr)^n P_0, 0\le n\le C\\ \frac{C^C}{C!}(r)^n P_0, C<n\le N.\end{cases}$$

系统的运行指标为

$$L_s=L_q+Cr(1-P_N),$$

$$L_q=\sum_{n=C+1}^{N}(n-C)P_n=\frac{(Cr)^C r}{C!(1-r)^2}P_0[1-r^{N-C}-(N-C)(1-r)r^{N-C},$$

$$W_q=\frac{L_q}{l(1-P_N)},\ W_s=W_q+1/m..$$

系统满员的损失率为 $P_{损}=P_N=\frac{C^C}{C!}r^N P_0.$

特别地，当 *N*=C 时，即 *M*/*M*/C/C/∞,此时系统为即时制服务，不允许顾客在系统内排队，亦即系统的状态概率为

$$P_0=[\sum_{k=0}^{C-1}\frac{1}{k!}(\frac{l}{m})^k]^{-1},P_n=\frac{(Cr)^n}{C!}P_0,$$

相应地运行指标为

$$L_q=W_q=0,W_s=\frac{1}{m},L_s=\sum_{n=1}^{C}nP_n=Cr(1-P_C).$$

5.3.3 顾客源为有限的：M/M/C/∞/m

假设同前面的模型相同，系统的状态概率的平衡方程为

$$\begin{cases}mP_1=ml P_0,\\ (n+1)mP_{n+1}+(m-n+1)l P_{n-1}=[(m-n)l+nm]P_n,(1\le n\le C)\\ CmP_{n+1}+(m-C+1)l P_{n-1}=[(m-C)l+Cm]P_n,(C\le n<m)\\ l P_{m-1}=CmP_m.\end{cases}$$

由递推关系可得状态概率

$$P_0=\frac{1}{m!}[\sum_{k=0}^{C}\frac{1}{k!(m-k)!}(\frac{Cr}{m})^k+\frac{C^C}{C!}\sum_{k=0}^{C}\frac{1}{(m-k)!}(\frac{r}{m})^k]^{-1},r=\frac{ml}{Cm},$$

$$P_n=\begin{cases}\frac{m!}{(m-n)!n!}(\frac{l}{m})^n P_0, 0\le n\le C\\ \frac{m!}{(m-n)!C!C^{n-C}}(\frac{l}{m})^n P_0, C<n\le m.\end{cases}$$

系统的运行指标为

$$L_s=\sum_{n=1}^{m}nP_n,,\quad L_q=\sum_{n=C+1}^{m}(n-C)P_n,W_s=\frac{L_s}{l_e}\,W_q=\frac{L_q}{l_e}.$$

有效到达率为 $l_e=l(m-L_s)$，且 $L_s=L_q+\frac{l_e}{m}=L_q+\frac{l}{m}(m-L_s)$.

类似的还有 $M/M/C/N/m$，$M/M/C/m/m$，$M/M/C/C/m$ 等情况，可作相应的讨论.

【例 5-3】某火车站售票处有三个窗口，顾客的到达服从泊松分布，平均每分钟有 0.9 人到达，服务时间服从负指数分布，平均每分钟可服务 0.4 人。现假设排成一队，依次向空闲的窗口购票，试分析该排队系统。

解　据题意知 m=3，λ=0.9，μ=0.4,则

$$r=\frac{l}{m}=\frac{0.9}{3\times0.4}=0.75$$

$$P_0=[1+\frac{0.9}{0.4}+\frac{1}{2!}(\frac{0.9}{0.4})^2+\frac{1}{3!}(\frac{0.9}{0.4})^3\frac{1}{1-0.75}]^{-1}=0.0743$$

即整个售票处空闲的概率为 0.0743。

平均队长 $L_q=\frac{(0.9/0.4)^3\times3/4}{3!(1/4)^2}\times0.0743=1.7$

平均等待时间 $W_q=1.7/0.9=1.89\,\text{min}$

平均逗留时间 $W_s=1.89+1/0.4=4.39\,\text{min}$

5.4 排队系统的最优化问题

5.4.1 一般排队系统的最优化问题

排队系统的最优化问题可分为系统设计最优化和系统控制最优化，系统设计最优化又称静态最优化，是指在服务系统设置以前根据一定的质量指标，找出参数的最优值，从而使系统设计最经济.例如：服务机构的规模大小、服务台的个数、系统容量大小等.系统控制最优化又称动态最优化，是指对已有的排队系统寻求使其某一目标函数达到最优的运行指标。

在排队系统中还有一个费用问题，它是指服务机构的服务成本和顾客等待的费用，一般来说，提高服务机构的服务水平（即增加了服务机构的成本），自然会降低顾客的等待费用（损失），最优化的目标之一是使二者费用之和为最小，另一个目标是使服务机构的纯收入（利润）为最大，如图 5-1.

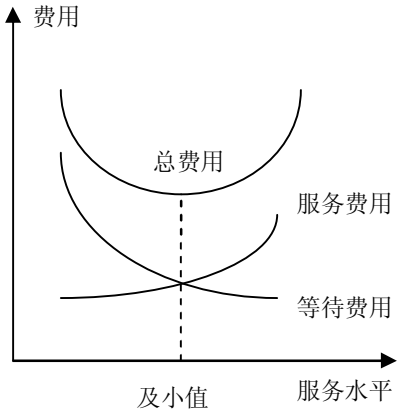

图 5-1 系统费用

5.4.2 模型 M/M/1 中的最优服务率 μ

1.标准型：*M*/*M*/1

设目标函数为 $z=C_s\mu+C_wL_s$,即单位时间服务成本与顾客等待费用之和的期望值，其中 C_s 表示当 $\mu=1$（单位时间内服务完 1 个顾客）时服务机构的服务费用，C_{ww} 为每个顾客在系统中停留单位时间的费用，由 $L_s=\dfrac{l}{m-l}$,则 $z=C_s\mu+C_w\dfrac{l}{m-l}$，求其极小值，即令 $\dfrac{dz}{dm}=0$，则 $C_s-\dfrac{C_w l}{(m-l)^2}$，解出最优解 $m^*=l+\sqrt{\dfrac{C_w}{C_s}}\,l$，即为最优服务率.

2.系统容量有限的：*M*/*M*/1/*N*/∞

如果系统中已有 N 个顾客，则后来的顾客将被拒绝，于是可设 P_N 为被拒绝的概率，1-P_N 即为接受服务的概率.l（1-P_N）表示单位时间内实际进入服务机构的顾客数，在稳定状态下，即为单位时间内实际服务完成的顾客数.

设系统服务完 1 个顾客能收入 G 元，于是单位时间收入的期望值为l (1-P_N)G，则系统的纯利润为

$$z=l(1-P_N)G-C_sm=lG\frac{1-r^N}{1-r^{N+1}}-C_sm=lG\frac{m^N-l^N}{m^{N+1}-l^{N+1}}-C_sm\ .$$

令 $\dfrac{dz}{dm}=0$，可解得

$$r^{N+1}\frac{N-(N+1)r+r^{N+1}}{(1-r^{N+1})^2}=\frac{C_s}{G}\ ,$$

其中 $P_N=\dfrac{r^N-r^{N+1}}{(1-r^{N+1})}, r=\dfrac{l}{m}$，而 C_s,G,l,N 均为已知的，用数值方法求解出 μ*的数值解.

3.顾客源为有限的：M/M/1/∞/***m***

设顾客数为 m，单个服务台、服务时间服从负指数分布，当服务率为 μ=1 时，服务机构的成本费为 C_s，单位时间内服务完 1 个顾客的收入为 G 元，单位时间内服务完的顾客数为

m-L_s，则单位时间内的纯利润为

$$z=(m-L_s)G-C_sm=\frac{mG}{r}\cdot\frac{E_{m-1}\left(\frac{m}{r}\right)}{E_m\left(\frac{m}{r}\right)}-C_sm\ ,$$

其中 $E_m\left(\frac{m}{r}\right)=\sum_{k=1}^{m}\frac{\left(\frac{m}{r}\right)^k}{k!}e^{-\frac{m}{r}}$ 为泊松和，$r=\frac{ml}{m}$，令 $\frac{dz}{dm}=0$，则得

$$\frac{E_{m-1}\left(\frac{m}{r}\right)E_m\left(\frac{m}{r}\right)+\frac{m}{r}\left[E_m\left(\frac{m}{r}\right)E_{m-1}\left(\frac{m}{r}\right)-E_{m-1}^2\left(\frac{m}{r}\right)\right]}{E_m^2\left(\frac{m}{r}\right)}=\frac{C_sl}{G}.$$

当给定 C_s,G,l, m 后，利用泊松分布表和数值方法计算求得最优服务率 μ*.

5.4.3 模型 M/M/c 中的最优服务台数

这里仅讨论标准模型.在稳态假设下，单位时间内每个服务台的成本费为 C_s，每个顾客在系统中停留单位时间的费用为 C_w,则单位时间内的费用（服务成本和等待的费用）的期望值：z=C_sc+C_w L_s,其中 L_s=L_s(c),即与服务台数 c 有关，因此总费用为 z=z(c),记 c 的最优值为 c^*,则 $z(c^*)$是最小费用.由于 c 只能取整数，即 z(c)是离散函数，所以只能用边际分析方法求解.事实上，根据 $z(c^*)$为最小值，可有

$$\begin{cases}z(c^*)\le z(c^*-1)\\ z(c^*)\le z(c^*+1)\end{cases}$$

由 z=C_sc+C_w L_s,则有

$$\begin{cases}C_sc^*+C_wL_s(c^*)\le C_s(c^*-1)+C_wL_s(c^*-1)\\ C_sc^*+C_wL_s(c^*)\le C_s(c^*+1)+C_wL_s(c^*+1)\end{cases}$$

化简整理得

$$L_s(c^*)-L_s(c^*+1)\le\frac{C_s}{C_w}\le L_s(c^*-1)-L_s(c^*).$$

由此可求得 c^*.

5.5 案例分析:校园网的设计和调节收费问题

5.5.1 问题的提出

随着计算机技术的飞速发展，校园信息网已在全国高校中普及。某高校拟建一个校园信息网，并与 Internet 连接，用户可以通过网络通信端口拨号上网.因此，需要根据用户的数量研究通信端口的设计规模.通常的通信端口分为 16 口、32 口、64 口、128 口等，实际中，随着随着通信端口数量的增加，其成本费将成倍增加。如何根据实际情况在保证基本满足用户

需求的条件下，确定合适的通信端口数，以减少费用的开支和资源的浪费。

当网络建成以后。为了保证用户有效的使用信息网，必须要通过适当的收取线路调节费来控制上网时间，一般认为，采用分段计时收费较为合理，例如按上网时间长短分为“免费→半费→全费→2 倍→3 倍→4 倍……”等时段.

现在的问题是：

（1）假设有 m 个用户，每个用户每天（按 16h 计算）平均上网 1.5h，试确定通信端口数 n 与 m 之比 n/m;

（2）假设 m＝150，按所设定的通信端口数 n，试讨论平均每天每个用户上网 1h 、2h、3h、4h、5h 的可能性，出现因线路忙，导致用户想上网而上不去产生抱怨的可能性，以及通信端口的平均使用率；

（3）为了控制上网时间，学校要求适当收取线路调节费，试给出一种合理的分段计时收费方案.

5.5.2 问题的分析与假设

根据题目中给出的信息，我们可以用排队理论来研究这个问题。假设校园信息网络和用户构成一个排队系统，网络的通信端口为服务台，个数为 n，用户为顾客，顾客源数为 m，平均忙期（即一天连续工作时间）为 16h.但是要注意到：实际中不限制用户的上网次数，虽然实际用户数为 m，但我们可以认为顾客总体是无限的。

另一方面，在同一时间，当 n 个用户端全部被占用（即系统满员）时，再有用户拨号上网，系统将会拒绝。此时，这些用户将产生抱怨，只有当网上的用户下网后才能有新的用户上网，可以这样周而复始的进行下去。这表明，只要时间允许，系统不限制上网人数，但不允许顾客在系统内排队等候，即系统的服务是即时制的。为此，给出如下几个假设：

(1) 每个用户的上网是随机、且相互独立的，单位时间的平均到达（上网）率为 l；

(2)n 个通信端口的使用是随机独立的，即任一用户可以使用空闲的任一端口，单位时间的平均服务（上网人数）率为 μ；

(3)不限制用户每天的上网次数，即顾客接受一次服务后仍回到顾客总体；

(4)学校对用户一般要收取一定数量的线路基本费，在模型中不考虑此费用；

(5)学校的目的不是营利，完全是为了调节线路，控制上网时间。因此，不需要追求经济利益。

5.5.3 模型的建立与求解

由上面的分析，假设用户平均上网的人数（即顾客平均到达率）服从参数为 l 的泊松分布，平均服务（上网）时间服从参数为 μ 的负指数分布，故问题的排队模型为 $M/M/n/n/\infty$.

问题(1):已知每个用户每天平均的上网 1.5h，则每天的总上网时间为 T=1.5m(h)，一天按 16h 计算，根据题意，要求在基本满足需要的条件下节省费用，通讯端口数尽量少为好。为此，设想让所有的端口满负荷运转，则每天每个通信端口占用的时间为 $\frac{T}{n}=\frac{1.5m}{n}=16(\mathrm{h})$，故 $\frac{n}{m}=\frac{1.5}{16}=\frac{1}{10.7}$，即通信端口数 n 与用户数 m 的比为 1:10.7， 这与实际中通常采用 1：10 的比例是相符的.

问题(2)：由问题 1 的结果，当 m＝150 时,通信端口数 n＝14.由假设（1），用户的平均上网率为 $l=\frac{150}{16}=\frac{75}{8}$(人/h).由假设(2)，各端口的平均服务率（单位时间上网人数）为 μ，

即每个用户平均上网时间为 $t=1/\mu$.

假设系统的状态为 k（即有 k 个用户在上网）的概率为 $P_k(k=0,1,2,\ldots,n)$，状态转移概率为 $k\mu P_k$，故得状态的平衡方程为

$$\begin{cases} mP_1 = lP_0, \\ lP_{k-1} = kmP_k\,(1<k<n), \\ lP_{n-1} = nmP_n, \end{cases}$$

其中 $\sum_{k=1}^{n} P_k = 1$.求解此差分方程，得到各状态的转移概率为

$$P_k = \frac{1}{k!}\left(\frac{l}{m}\right)^k P_0, 0<k\le n,\ P_0 = \left[\sum_{j=0}^{n}\frac{1}{j!}\left(\frac{l}{m}\right)^j\right]^{-1}.$$

而且通信端口的平均使用数为 $L_s = n\left(\frac{l}{m}\right)(1-P_n)$.

系统满员的概率为 $P_n = \frac{1}{n!}\left(\frac{l}{m}\right)^n P_0$.

于是，通信端口有空闲用户能上网的概率为 $\overline{P} = \sum_{k=0}^{n-1} P_k = 1-P_n$.

当 l=75/8 (人/h)，n＝16，平均每天单个用户上网 1h、1.5h、2h、3h、4h、5h，即 μ=1，2/3,1/2,1/3,1/4,1/5 时，用户能上网的概率 $\overline{P}$ 、因线路忙用户上不了因而产生抱怨的概率 P_{16}、和单位时间端口的平均使用率 L_s 的计算结果如表 5-1

表 5-1 问题（2）的计算结果

μ	P_{16}	$\overline{P}$	L_s
1	0.01466671	0.985333	9.2375
2/3	0.116352	0.883648	12.4263
1/2	0.257403	0.742597	13.9237
1/3	0.467174	0.532826	14.9857
1/4	0.590668	0.409332	15.35
1/5	0.668793	0.331207	15.5253

问题（3）：根据问题的要求，采用分段计时收费方案。分为“免费→半费→全费→2 倍→3 倍……”等时段，首先应确定免费时段。按 m＝150 人，n＝14 口，要保证平均每天每端口为 150/14＝10.7 人次，平均上网时间大约为 1.5h. 同时考虑到问题（1）中的端口设计按照 1.5h 设计。于是,不妨确定免费上网时间为 1.5h，则相应的抱怨概率为 $P_{16}(1.5)=0.116352$. 满员的概率为 $P_{16}(1.5)=0.116352$.随着时间 $t=1/\mu$ 的增大，$P_{16}(t)$也增大，因此就按照 $P_{16}(t)$

随时间 t 对 $P_{16}(1.5)$增加的倍数来确定对应的上网时间段，即为分段加倍收费的时间.

事实上，由 $P_{16}(t)=\frac{P_0}{16!}\left(\frac{l}{m}\right)^{16}$ 和 $P_0=\left[\sum_{j=0}^{16}\frac{1}{j!}\left(\frac{l}{m}\right)^j\right]^{-1}$ 可以求出 $t=\frac{1}{m}$.为此根据问题（2）中的计算结果作拟合数据得到 $P_{16}(t)$的近似表达式，则得到分段计时收费的时间段为：

$2P_{16}(1.7)=0.346774, t\approx 2.383\approx 2.4$;

$3P_{16}(1.7)=0.520161, t\approx 3.35039\approx 3.4$;

$4P_{16}(1.7)=0.693548, t\approx 5.23768\approx 5.2$;

$5P_{16}(1.7)=0.866935, t\approx 6.10164\approx 6.1$.

于是可以得到分段收费方案：

当 t<1.5 时，免费；1.5 ≤t<1.7 时，收半费 $\frac{d}{2}$；当 1.7 ≤t<2.4 时，收全费 *d*; 当 2.4 ≤t<2.4 时，2 倍收费 2*d*; 当 3.4≤t<5.2 时，3 倍收费 3*d*; 当 5.2≤ t<6.1 时，4 倍收费 4*d*；当 *t*≥6.1 时，依此类推，其中 d 根据学校的实际情况确定.

5.5.4 两点说明

（1）实际中，用户上网数量的多少一定与时间有关系，早中晚的人数一定是不均衡的，因此，因在收费方案中考虑时间因素。对早上、上午、中午、下午、晚上分时间段赋予不同权重。得到，“分段加权计时”的收费方案，这种方案可以更好、更有效地起到对通信端口的调节作用；

（2）本问题的模型为网络的设计提供了一定的理论依据，实践证明该模型是符合实际的，具有一定的应用和推广价值。

【修理工录用问题】解答

解　用 N 表示每天发生故障机器的平均数，包括正在修理和等待修理的机器数，即等于队长 L_s，C_1 和 C_2 分别表示修理一台机器的费用和工人的工资，则工厂每天平均损失费用为：

$$R=NC_1+NC_2$$

(1) 若录用 A 种修理工，据题意知 $\lambda=1$，$\mu_A=0.2$，$\rho_A=\lambda/\mu_A=5/6$，$L_s^A=\rho_A/(1-\rho_A)=5$ 台，则若录用 A 种修理工，工厂每天平均损失费用为：

$$R_A=5\times 20+3=103 \text{ 元}$$

(2) 若录用 B 种修理工，据题意知 $\lambda=1$，$\mu_B=1.5$，$\rho_b=2/3$，$L_s^B=2$ 台，则若录用 B 种修理工，工厂每天平均损失费用为：

$$R_B=2\times 20+3=43 \text{ 元}$$

比较可知，工厂录用 B 种修理工较为合算。如果计入机器停工损失费用，这一选择更为上算。

习题

1. 机器发生故障后排队等待修理，队伍越长因停产造成的损失越大。提高维修工人和设备的服务速度或增加其数量可以减少队长，但将使修理费用上升，选择怎样的服务速度，或者确定几个维修工人和设备使损失和修理的总费用最小。

参考文献

[1] 韩中庚，数学建模及其应用
[2]姜启源，谢金星，叶俊，数学模型高等教育出版社.2003.