# 随机模拟

- 马尔可夫链
- 仿真随机服务系统
- 案例分析

随机模拟是运用计算机对随机系统进行的一种仿真研究方法，其另一种叫法是蒙特卡罗方法。随机模拟方法的历史可以追溯到1777年法国人蒲丰的投针试验，但在没有计算机时代的这种人工试验方法显然无法被推广发展。随机模拟方法的基本思想是由计算机生成随机数序列，从而可以由此模拟出各种随机事件。

# 1 马尔可夫链

如果我们将处于全空间$\Omega$中的单个概率试验扩展到按时间$t = 1, 2, \cdots$依次进行试验的实验序列，这样就获得了一连串前后相关的试验结果。

令$X_t$表示第$t$次试验结果，我们将它看成是从状态$X_{t-1}$转移而来的，并称这种从状态$X_{t-1}$转移到状态$X_t$的过程为第$t$次迭代。俄国数学家马尔可夫对此作了一个重要假设：这种转移仅依赖于当前状态，而与以前的状态无关，由此所产生的试验序列$\{X_t\}$就被称为马尔可夫链。决定马尔可夫链的关键就是其转移概率，即从一个状态转移到其他状态（可能包括自身）的概率。如果这些转移概率不随时间而变化，那么我们称这种马尔可夫链是齐次的，否则它是非齐次的。通常，从任意一个给定状态$X$出发，一步迭代只能转移到$\Omega$的部分状态，这个部分状态称为$X$的领域。

马尔可夫链的精髓是简化了系统从一个时刻到下一个时刻的随机演化方式，即假设该系统的下一步状态只取决于当前时刻的状态，而无关于整个历史。

事实证明，马尔可夫链模型有着广泛的适用性。例如，一个随时间变化的热力学系统，一天又一天的股价波动，一个赌徒的赌博资金等。马尔可夫链后来又被推广到时间连续和状态连续的情况，统称为马尔可夫过程。因为随机模拟仅涉及时间离散的马尔可夫过程，因此这里我们只介绍这方面的知识。

## 1.1. 马尔可夫链的表示方式

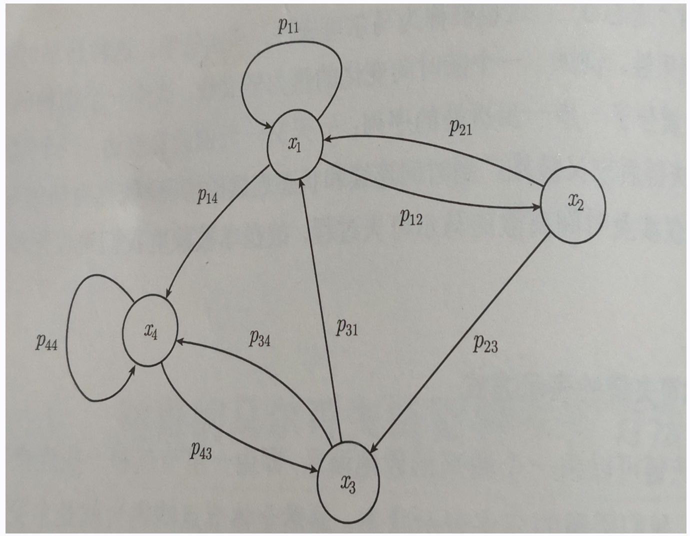
马尔可夫链可以由一个图来形象地演示，即由一组节点和一些连接节点的边所构成的几何图形。我们用图的节点表示状态，显然全体节点的集合就是全空间$\Omega$。这里我们考虑状态数是有限的情形。我们不妨将状态予以编号，即$\Omega = \{x_1, x_2, \cdots, x_n\}$。另外，我们用连接节点$x_i$到另一个节点$x_j$的有向边来表示从$x_i$到$x_j$的转移，并在此条边上标注该转移概率$p_{ij}$。

在一次迭代中，从一个节点$x_i$（状态）可能转移到的所有节点就是它的领域，将这些节点的编号集合记为$N_i$。对于在一次迭代中那些不可能转移到的节点（状态），我们赋予其转移概率为零。这样，我们可以方便地定义一个$n$阶的转移矩阵：

$$P=\begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n} \\ \vdots & \vdots & & \vdots \\ p_{n1} & p_{n2} & \cdots & p_{nn} \end{pmatrix}$$

它必定满足归一性条件：$\sum_{j=1}^{n} p_{ij}=1$。马尔可夫链的转移矩阵是一种非负矩阵，并且每行都是归一的。具有这样性质的矩阵也被称为随机矩阵。

**例1：** 考察一个拥有4个状态的马尔可夫链，其转移情况由下图所示，




相应的转移矩阵为

$$P=\begin{pmatrix} p_{11} & p_{12} & 0 & p_{14} \\ p_{21} & 0 & p_{23} & 0 \\ p_{31} & 0 & 0 & p_{34} \\ 0 & 0 & p_{43} & p_{44} \end{pmatrix}$$

我们设初始的状态是$x_1$，那么下一步的状态将是哪个呢？

我们只需要生成服从均匀分布的一个随机数$r \sim U(0,1)$，看$r$落在$[0,p_{11}],(p_{11},p_{11}+p_{12}]$或$(p_{11}+p_{12},1]$的哪个区间里，则下一步状态就转移到它们分别对应于的状态$x_1,x_2$或$x_4$。依次类推，不断地进行转移，则产生了该马尔可夫链的一条状态演化路径。

显然，状态的演化是随机的，我们以向量$p_t = (p_1,p_2,p_3,p_4)$表示经过$t$步转移后状态的概率分布，这个向量被称为概率向量。根据概率公式，我们可得概率分布的转移关系式

$$p_{t+1} = p_t P = p_{t-1}P^2 = \cdots = p_0 P^{t+1}$$

因此，马尔可夫链的状态演化完全由其转移概率所决定。

## 1.2. 一个股市的马尔可夫链模型

考虑一个股票市场的模型。根据金融市场有效性假设，股价的下一次变化只取决于当前状态，具有马尔可夫性，所以股价的变化可以用马尔可夫链来建模描述。

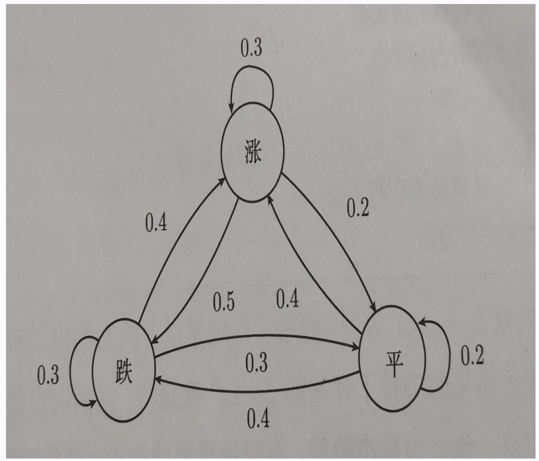
例2：众所周知，明天的股价相对于今天来说有“涨”、“持平”或者“跌”三种可能的状态，分别以数字1,2,3表示。我们再假设其状态转移的变化规律是：

（1）如果股价今天是涨，那么明天再涨的概率是0.3，持平的概率是0.2，而跌的概率是0.5；

（2）如果股价今天是持平，那么明天上涨的概率是0.4，再持平的概率是0.2，而跌的概率是0.4；

（3）如果股价今天是跌，那么明天会涨的概率是0.4，持平的概率是0.3，而再跌的概率是0.3。

我们用一个图来描述这个马尔可夫链，其中节点表示三个状态：涨、平和跌；有向边和标出的数值分别表示状态转移方向及其概率。




其转移概率矩阵为

$$P=\begin{pmatrix}0.3 & 0.2 & 0.5\\ 0.4 & 0.2 & 0.4\\ 0.4 & 0.3 & 0.3\end{pmatrix}$$

我们可以来模拟股市的变化状况。首先需要给定一个起始状态，我们可以随机选择或确定一个状态。如果我们假定今天状态是“涨”，接下来我们需要生成服从于转移概率分布$\{p_{11}=0.3, p_{12}=0.2, p_{13}=0.5\}$的随机变量$x$。为此，我们使用连续均匀分布的随机数$r\sim U(0,1)$来产生下面的随机数：

$$x=\begin{cases}1 & 0\le r\le p_{11}\\ 2 & p_{11}<r\le p_{11}+p_{12}\\ 3 & p_{11}+p_{12}<r\le 1\end{cases}$$

类似地，对于其他初始状态情形，我们可以生成所要的随机数。事实上，我们可以定义关于转移概率的累积分布矩阵

$$C=\begin{pmatrix}0.3 & 0.5 & 1\\ 0.4 & 0.6 & 1\\ 0.4 & 0.7 & 1\end{pmatrix}$$

马尔可夫链
仿真随机服务系统
案例分析
访问主页
标题页
返回
全屏显示
关闭
退出

我们使用Matlab编制模拟该马尔可夫链的程序。

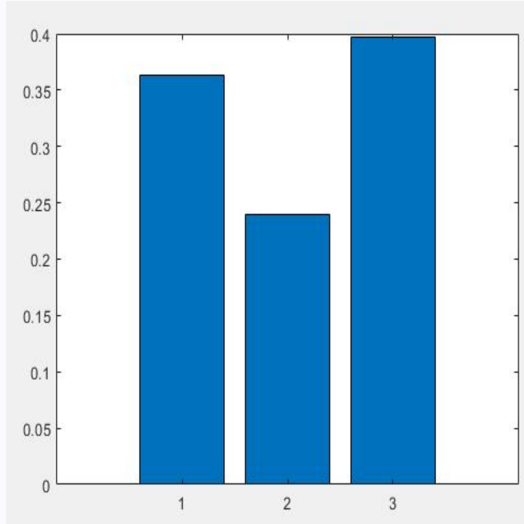
```
n=2500000;t0=2000000; %n迭代次数，t0暂态的长度
P=[0.3 0.2 0.5;0.4 0.2 0.4;0.4 0.3 0.3]; %转移概率矩阵
C=cumsum(P,2); %产生累积转移概率矩阵
state=ones(1,n); %将存放迭代产生的状态序列
r=unifrnd(0,1,[1,n]); %在[0,1]中的均匀随机数
i0=input('初始状态');
i=i0;
for t=1:n
    j=state(t);
    while r(t)>C(i,j)  %此段循环生成随机状态
        j=j+1;
    end
state(t)=j; %转移到的状态
i=j; %更新下次的出发状态
end
h=hist(state(t0+1:n),[1,2,3]); %统计次数
p=h/(n-t0); %计算频率获得不变分布
bar(p)  %画频率的直方图
```

我们选择初始状态$x_0 = 1$为例，程序运行结果以各状态的出现频率的直方图形式呈现，其中在统计结果时去掉了开始的$2 \times 10^6$次迭代，因为它们是系统运行还未进入稳定阶段之前的状态，称之为暂态阶段。

暂态的结果是依赖于所产生的随机数的特定序列，虽然充分长的链序列能够忽略暂态的影响，但去掉它们能提高结果的精确度。所以，我们应该仅统计系统进入稳态后的结果。

这个例子所显示的分布（频率）是十分稳定的，也就是说，即时再增加更多的迭代次数，其分布仍然不变，这正是马尔可夫链的一个重要性质：存在不变分布，也被称为稳定分布。从其他初始状态出发，程序运行的结果也将呈现同一个不变分布，即不变分布作为极限的分布是和初始状态无关的。

股市的马尔可夫链的频率直方图：




程序给出的不变概率分布：

```
初始状态1
>> p

p =

    0.3634    0.2398    0.3968
```

因此，在这个股市模型中，如果我们知道转移概率，就能根据不变分布来预测今后股价所呈现的涨跌趋势。结果告诉我们：今后股价出现上涨的比例约为36%，而出现下跌的概率约为40%，高于上涨的比例。

这个例子还有两个问题值得注意：

（1）如何获得转移概率？

（2）为什么要假设该马尔可夫链是齐次的，即转移概率不随时间变化？

首先回答第二个问题。因为只有当股价变化有稳定的转移规律存在，转移概率才有意义。应该知道，假设马尔可夫链是齐次的往往是一种对问题的简化，这使得我们能够方便地获得结果。

如果认为齐次性是可接受的，那么我们就可以收集股市上那只股票在较长一段时期内的价格变化数据，然后简单地统计在此段时期内所有连续两天股价的如下9种变化情况的出现频率：涨-涨、涨-平、涨-跌、平-涨、平-平、平-跌、跌-涨、跌-平、跌-跌。这样，我们就可以获得所要的转移概率。

# 2 仿真随机服务系统

随机服务系统是人们在日常生活中经常要与之打交道的对象，如顾客到商店购物有时需要排队，患者到医院求诊常需要排队等待。原因是到达的顾客数超过了服务机构的容量（如服务台、库容、舱位等），到达的顾客不能立即得到服务，因而出现了排队现象。随机服务系统还包括了像公路收费站、机场等交通枢纽，水库的库容调节等，这些都可以归结为一种随机服务系统。

服务系统的这种不确定现象的随机源来自：

（1）顾客的到达时间和数量的不确定；

（2）每一顾客的服务时间通常也不确定。

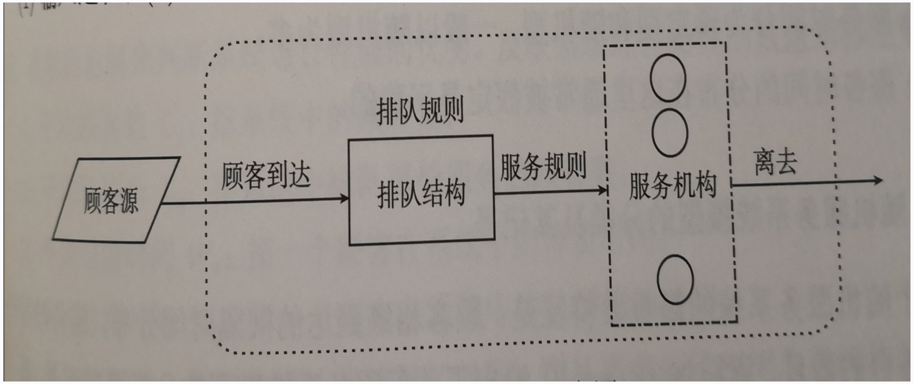
研究随机服务系统的数学理论起源于1990年丹麦电话工程师爱尔朗的工作，他的研究成果为运筹学的分支–排队论奠定了基础，该理论研究的目标就是既要保证系统服务质量指标较好，又要使服务系统的成本费用经济合理。

排队论的应用涉及广泛的领域，但其中许多实际问题所写出的数学模型并不是排队论中的标准可解模型，除非对模型做出不切实际的简化。然而，采用随机模拟的方法来分析这些系统就不存在技术障碍，即使对那些排队论可解模型，随机模拟方法也是简便直观的。

## 2.1. 随机服务系统的组成与特征

随机服务系统一般有如下三个基本组成部分：

（1）输入过程；（2）排队规则；（3）服务机构。



首先，输入过程是指顾客的到达情况，可分为下列几种情形：

（1）顾客源可能是有限的，也可能是无限的。

（2）顾客是成批到达，或是单个到达。

（3）顾客到达时间间隔可能是随机的，或确定的。

（4）顾客到达可能是相互独立，也可能是不独立的。所谓独立就是先前顾客的到达对以后顾客的到达无影响。

（5）输入过程可能是平稳的，也可能是非平稳的。所谓平稳过程是指顾客相继到达的时间间隔所服从的概率分布与时间无关（通常只要求均值和方差与时间无关），而非平稳的则依赖于时间。

其次，排队规则是指顾客进入服务系统的等待方式，可分为损失制、等待制、混合制三大类。

（1）损失制：指如果顾客到达排队系统时，所有服务台都已被先来的顾客占用，那么他们就自动离开系统。例如，停车场，顾客来时如果遇到位满，则自动离去。

（2）等待制：当顾客来到系统时，所有服务台都不空，顾客加入排队行列等待服务。例如，排队等待售票，故障设备等待维修等。等待制中，服务台在选择顾客进行服务时，还有四种规则：先到先服务（FCFS）；后到先服务（LCFS）；随机服务；优先权服务。

（3）混合制：这是等待制与损失制相结合的一种服务规则，一般是指允许排队，但又不允许队列无限长下去。大致分为三种，其一是限制队长，如水库库容、展览馆等；另两种指等待时间和逗留时间有限制的情形。

最后，服务机构是指服务系统的功能情况，通常关心的是服务时间。它可分为：

（1）服务机构可以是单个服务台或者并列的多服务台组成；

（2）服务方式分单个顾客服务和成批顾客服务；

（3）服务时间分为确定型和随机型，一般以随机型为多；

（4）服务时间的分布在这里通常被假定是平稳的。

## 2.2. 随机服务系统模型的分类

一个随机服务系统的最重要特征是：顾客相继到达的间隔时间分布，服务时间的分布和服务台的数目。肯特尔给出了关于服务系统模型的分类及其记号，即

$$X/Y/Z/A/B/C$$

- $X$ – 顾客相继到达间隔时间分布，常见的是，M-指数分布（泊松过程），这种输入过程属于马尔可夫过程；D-确定型情形。
- $Y$ – 服务时间分布，常见记号同上。
- $Z$ – 并列的服务台数。
- $A$ – 排队系统的最大容量。
- $B$ – 顾客源数量。
- $C$ – 排队规则，如先到先服务等。

同时约定，如略去记号的后三项，即指：$X/Y/Z/\infty/\infty/FCFS$。另外，$M/M/1/\infty/\infty/FCFS$可简写为$M/M/1$，它指顾客到达为泊松过程、服务时间为指数分布和单服务台的随机服务系统模型（注：到达流为泊松过程的时间间隔是服从指数分布的）。

## 2.3. 随机服务系统的基本模型

对于一个实际的随机服务系统问题，一般需要做如下两项工作：

（1）确定或拟合系统中顾客到达的时间间隔分布（或者到达流的分布）和服务时间分布；

（2）研究系统状态的概率特性，评估服务系统的性能指标。系统状态是指系统中顾客数$n$，状态概率用$P_n(t)$表示，即在$t$时刻，系统中有$n$个顾客的概率，也称暂态概率。

求解状态概率的方法是建立含$P_n(t)$的微分差分方程，通过求解微分差分方程得到系统的暂态解。由于暂态解一般来说难以被用来评估系统的性能状况，因此常常使用它的极限（如果存在的话）情况，即稳态解：

$$\lim_{t\to\infty} P_n(t) = p_n$$

然后，根据稳态解来判断系统运行性能的优劣。

反映系统运行特性的数量指标主要有：

（1）平均顾客数$L_s$：系统中的顾客数；

（2）平均队列长$L_q$：系统中排队等待服务的顾客数；

（3）平均逗留时间$W_s$：一个顾客在系统中的停留时间；

（4）平均等待时间$W_q$：一个顾客在系统中排队等待的时间；

（5）平均忙期$T_b$：服务机构连续繁忙时间，即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。忙期和一个忙期中平均完成服务顾客数都是衡量服务机构效率的指标，忙期与工作强度有关。

注意：以上这些指标均为期望值，所以它们非常适合采用随机模拟方法来统计确定。

- **泊松过程**

在随机服务系统中，顾客的到达过程常常被假设为一个泊松过程。设$N(t)$表示在时间区间$[0,t)$内到达的顾客数，以$P_n(t_1,t_2)$表示在时间区间$[t_1,t_2)$内有$n$个顾客到达的概率。所谓的泊松过程是指$P_n(t_1,t_2)$满足如下的泊松概率分布：

$$P_n(t_1,t_2)=P(N(t_2)-N(t_1)=n)=\frac{(\lambda(t_2-t_1))^n}{n!}e^{-\lambda(t_2-t_1)} \tag{1}$$

并且对于足够小的$\Delta t$，显然有

$$P_1(t,t+\Delta t)=\lambda\Delta t+o(\Delta t) \tag{2}$$

其中，$\lambda>0$是常数，它表示单位时间顾客的平均到达数。上式表明，在$[t,t+\Delta t)$内有一个顾客到达的概率与$t$无关，而与$\Delta t$成正比。

泊松过程是一种马尔可夫过程，它具有如下性质：

（1）马尔可夫性：各时间区间的顾客到达相互独立，即无后效性。也就是说，过程在$t+\Delta t$所处的状态与$t$以前所处的状态无关。

（2）平稳性：式（1）表明$P_n(t_1,t_2)$在$[t_1,t_2)$内有顾客到达的概率与$t_1$无关，而仅与时间的间隔$(t_2-t_1)$有关。

（3）排斥性：对充分小的$\Delta t$，在时间区间$[t,t+\Delta t)$内有2个或2个以上顾客到达的概率是一高阶无穷小，即

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

因此，我们可以忽略它们的发生。

因为$P_0+P_1+P_{\geq 2}=1$，于是可知在$[t,t+\Delta t)$区间内没有顾客到达的概率为

$$P_0(t,t+\Delta t) = 1-\lambda\Delta t + o(\Delta t) \approx 1-\lambda\Delta t \qquad (3)$$

- **指数分布**

当输入过程是泊松过程时，两顾客相继到达的时间间隔的概率分布是什么？

设$T$为时间间隔，分布函数为$F_T(t)$，则$F_T(t)=P(T\leq t)$。此概率等价于在$[0,t)$区间内至少有1个顾客到达的概率。

由于没有顾客到达的概率为：$P_0(t)=e^{-\lambda t}$，则概率分布函数为

$$F_T(t)=1-P_0(t)=1-e^{-\lambda t},t>0$$

其概率密度函数为

$$f_T(t)=\frac{dF_t}{dt}=\lambda e^{-\lambda t},t>0$$

即$T$服从指数分布，它的期望及方差分别是$1/\lambda,1/\lambda^2$。这里$\lambda$就是前面的顾客在单位时间内的平均到达数。这样，$1/\lambda$就表示顾客到达的平均间隔时间。

这就证明了：顾客到达的间隔时间$T$服从指数分布，当然前后不同时间的顾客到达间隔时间是相互独立的。同时，我们看到，顾客到达的间隔时间服从指数分布等价于顾客的到达流服从泊松过程，两者的参数相同。

- **服务时间的分布**

对顾客的服务时间$H$是指系统处于忙期时两顾客相继离开系统的时间间隔。一般地，我们也假设它服从指数分布，即设它的概率密度函数为

$$f_H(t) = \mu e^{-\mu t}$$

其中，$\mu$表示单位时间内被服务的顾客数，即平均服务率，则$1/\mu$表示一个顾客的平均服务时间。

服务时间的分布与上面泊松过程有完全类似的性质。若要计算在$\Delta t$时间内有一个顾客服务完毕和没有顾客服务完毕的概率，只要将上面公式（2）和（3）中的参数$\lambda$换成$\mu$即可。

令

$$\rho = \lambda/\mu$$

则$\rho$称为服务强度，一般要求它满足条件$\rho < 1$，即顾客的平均到达率小于系统的平均服务率，否则队列会无限，永远达不到稳态。

## 2.4. M/M/1 模型

在给定输入和服务条件下，这个模型在数学上是一个可解的系统。我们可以先求出系统状态的暂态概率，即在时刻$t$，系统状态为$n$的概率$P_n(t)$，然后对$P_n(t)$取极限来求得稳态解。

我们可以求得系统的稳态概率为

$$P_n = (1-\rho)\rho^n$$

这就是系统达到平衡状态后队长的概率分布。由此可以看到，服务强度$\rho$是系统中至少有一个顾客的概率，也就是服务系统处于忙得状态的概率。

## 几个主要数量指标

系统中的平均顾客数

$$L_s = \frac{\lambda}{\mu - \lambda}$$

队列中等待的平均顾客数

$$L_q = L_s - \rho = \frac{\rho^2}{1-\rho} = \frac{\rho\lambda}{\mu - \lambda}$$

顾客在系统中的平均逗留时间

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

在$L_s$与$W_s$之间存在关系式：$L_s = \lambda W_s$.

平均等待时间可由平均逗留时间减去平均服务时间得到，

$$W_q = W_s - \frac{1}{\mu} = \frac{\rho}{\mu - \lambda}$$

同样，在$L_q$与$W_q$之间存在关系式：$L_q = \lambda W_q$.

例3: 某修理店只有一个修理工，要求提供服务的顾客到达过程为Poisson流，平均4人/h；修理时间服从指数分布，平均需要6min。试求：

（1）修理店空闲的概率；

（2）店内恰有3个顾客的概率；

（3）店内至少有1个顾客的概率；

（4）在店内的平均顾客数；

（5）每位顾客在店内的平均逗留时间；

（6）等待服务的平均顾客数；

（7）每位顾客平均等待服务时间。

**解:** 本例可看成一个M/M/1系统，其中$\lambda = 4, \mu = 1/0.1 = 10, \rho = 0.4 < 1$。

（1）修理店空闲的概率

$$P_0 = 1 - \rho = 0.6$$

（2）店内恰有3个顾客的概率

$$P_3 = \rho^3(1-\rho) = 0.4^3 \times 0.6 = 0.038$$

（3）店内至少有1个顾客的概率

$$P\{N \geq 1\} = 1 - P_0 = \rho = 0.4$$

（4）在店内的平均顾客数

$$L_q = \frac{\rho}{1-\rho} = 0.67$$

（5）每位顾客在店内的平均逗留时间

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

（6）等待服务的平均顾客数

$$L_q = \frac{\rho^2}{1-\rho} = 0.267$$

（7）每位顾客平均等待服务时间

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

例4：某医院某科室有一位医生值班，每小时平均有4个病人，医生每小时平均可诊5个病人。如要满足99%以上的病人有座位，至少应设多少座位？如果每小时可诊6个病人，可减少多少个座位？病人平均等待时间是多少？

解：设病人到来服从Poisson分布，医生诊断时间服从指数分布（M/M/1系统），则$\lambda = 4, \mu = 5, \rho = 0.8 < 1$。

$$L_s = \frac{\rho}{1-\rho} = 4$$

$$W_q = \frac{\rho}{\mu - \lambda} = 0.8$$

为满足99%以上病人有座，设应设$m$个座位，则

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

$$\rho^{m+1} \leq 0.01 \Rightarrow m \geq \frac{\ln 0.01}{\ln \rho} - 1 \simeq 20$$

若$\lambda = 4, \mu = 6, \rho = 2/3 < 1$。

$$m \geq \frac{\ln 0.01}{\ln \rho} - 1 = 11$$

所以可减少9个座位。

$$W_q = \frac{\rho}{\mu - \lambda} = 1/3$$

## 2.5. 随机服务系统的模拟

我们采用Matlab软件来模拟随机服务系统。系统的状态（顾客数）是随着各个事件的发生而变化的，其中影响系统状态的事件有：顾客的到达、服务完毕的顾客的离去。为此，我们要记录每个顾客的全部有关信息。我们可以建立一个5行多列的矩阵变量guests来记录这些信息。这个矩阵的各列对应于依时间先后顺序到达的各个顾客，矩阵的各行分别表示顾客的各个主要信息变量：

第1行：顾客的到达时刻；第2行：顾客的服务时间；第3行：顾客的逗留时间；第4行：顾客的离开时刻；第5行：顾客的附加信息，此行信息为可选的。

关于第5行信息的具体内容可视系统的类型而定。例如，在容量有限的系统，它可用来表示该顾客是被接纳还是被拒绝的标志；如果有多个服务台的话，它可用来表示该顾客选择的是哪个服务台的编号；如果顾客享有某种优先权的话，那它可用来表示顾客的优先级。

## 模拟算法：

（1）生成指定分布的随机序列

（a）根据到达率$\lambda$和服务率$\mu$来分别确定每个顾客的到达时间间隔和服务时间间隔。因为M/M/1/N型的服务系统，其服务间隔时间可以用指数分布函数$exprnd()$来生成。由于泊松过程的时间间隔也服从指数分布，故也可用此函数生成顾客的到达时间间隔。需要注意的是，$exprnd()$的输入参数不是到达率$\lambda$和服务率$\mu$，而是平均到达时间间隔$1/\lambda$和平均服务时间$1/\mu$。

（b）根据到达时间间隔，确定每个顾客的到达时刻。在Matlab中提供了一个进行累加计算的函数$cumsum()$。

（c）对开始顾客的变量进行初始化。第1个到达系统的顾客不需要等待就可以直接接受服务，其离开时刻等于到达时刻与服务时间之和。

（2）模拟顾客的到达与离开

按照时间顺序来模拟顾客流和服务过程。在当前顾客到达时刻，根据系统内已有的顾客数和系统的容量来确定当前顾客是否进入该系统。如进入系统，则根据前面顾客的离开时刻来确定该顾客的等待时间、服务时间和离开时间。如不进入，则只需再附加信息中做出标示即可。

如果服务系统有营业时间的限制，每一轮模拟的时间将要根据采用的时间单位来确定。据此，我们可以设置每一轮模拟的迭代终止条件。

# 3 案例分析

## 3.1. 案例1(医院预检处)

一般一个中等规模的医院都设有预检处，预检员的工作主要是帮助患者或访客对病类的甄别，使他们能正确地挂号，到医院的相关诊疗科室就诊。随着计算机的普及，医院会在接待大厅设置计算机自助挂号的触摸屏来取代人工接待，患者或访客要使访客用计算机提供的导航信息为自己挂号。但有些患者或访客可能不会操作或因不熟练的操作而引起耽误，所以往往会因此出现排队现象。这种现象就需要医院的信息管理工程师来评估计算机挂号的效率及可能引起的意外耽误程度。

我们首先对下面的情况进行初步分析：信息工程师收集计算机的使用记录来估计患者或访客到达流的分布及其平均到达率、患者或访客使用计算机的时间分布和平均使用时间（服务时间）。当然，关于顾客到达流的估计也可以根据原先人工预检时的信息数据来做出。假如预检系统的到达流的分布特性不随时间变化而改变，则可以近似地假设这个系统是一种M/M/1型的随机服务系统。于是，我们假设客户到达率服从均值为$\lambda$的泊松分布，而其服务时间也是相互独立同分布的随机变量，服从其均值为$\tau$的指数分布。

假设$A_1, A_2, \cdots$是顾客到达的间隔时间均值为$1/\lambda$的独立同分布的随机变量序列（其中$A_1$是每天第一个客户到达的实际时间）；类似地，令$X_1, X_2, \cdots$是顾客的服务时间序列，它们是均值为$\tau$且标准差为$\sigma$的独立同分布的随机变量；令$Y_1, Y_2, \cdots$是顾客在排队预检时的等待时间序列（即从顾客到达直到服务开始之间的时间）。这样，显然有

$$Y_i = \max\{0, Y_{i-1} + X_{i-1} - A_i\}, i = 1, 2, \cdots$$

这里我们规定$Y_0 = X_0 = 0$。这个递归式被称为林德利方程，我们可以用它来方便地模拟这种M/M/1型系统的运行情况，并且可以通过统计重复多次的模拟结果来为改进预检系统提供一些相应地建议。

需要指出的是：在上面的模型中，等待时间这个随机变量是相关的，因为显然$Y_i$取决于$Y_{i-1}$（如果前面的顾客等了很久，则后面的顾客大概也会一样等很久）；并且$Y_i$不会是同分布的（显然，第一个顾客不需要等待，即$Y_1 = 0$），但其他的$Y_i$就不一定是0。还有，无论我们模拟多长时间，等待时间不是相互独立的且分布也不相同，即使做统计平均处理，它们也不会收敛于某个常数。然而，数学上还是证明了该系统在一定条件下具有某些概率意义上的极限特征，因此这也表明模拟方法确实可以给出可靠的结果。

## 模拟程序

我们假设医院门诊每天营业10个小时，选择小时作为单位时间。从历史数据中已估计得出该医院患者或访客的到达率为65（人/小时），服务率为60（人/小时）。虽然没有容量的限制，但由于每天有营业时间，所以每天患者或访客的总数量是有限的。我们可以用下面公式来估算每天可能的最大顾客数：

$$最大顾客数=总时间\times 到达率\times 2$$

这样，我们可以预先定义程序中数组变量的大小，而不需要采用动态数组。否则的话，采用可变大小的动态数组将会降低计算速度。另外，我们将顾客数组中附加信息的那行用来记录当前顾客进入系统时，在他之前系统已有的顾客数。

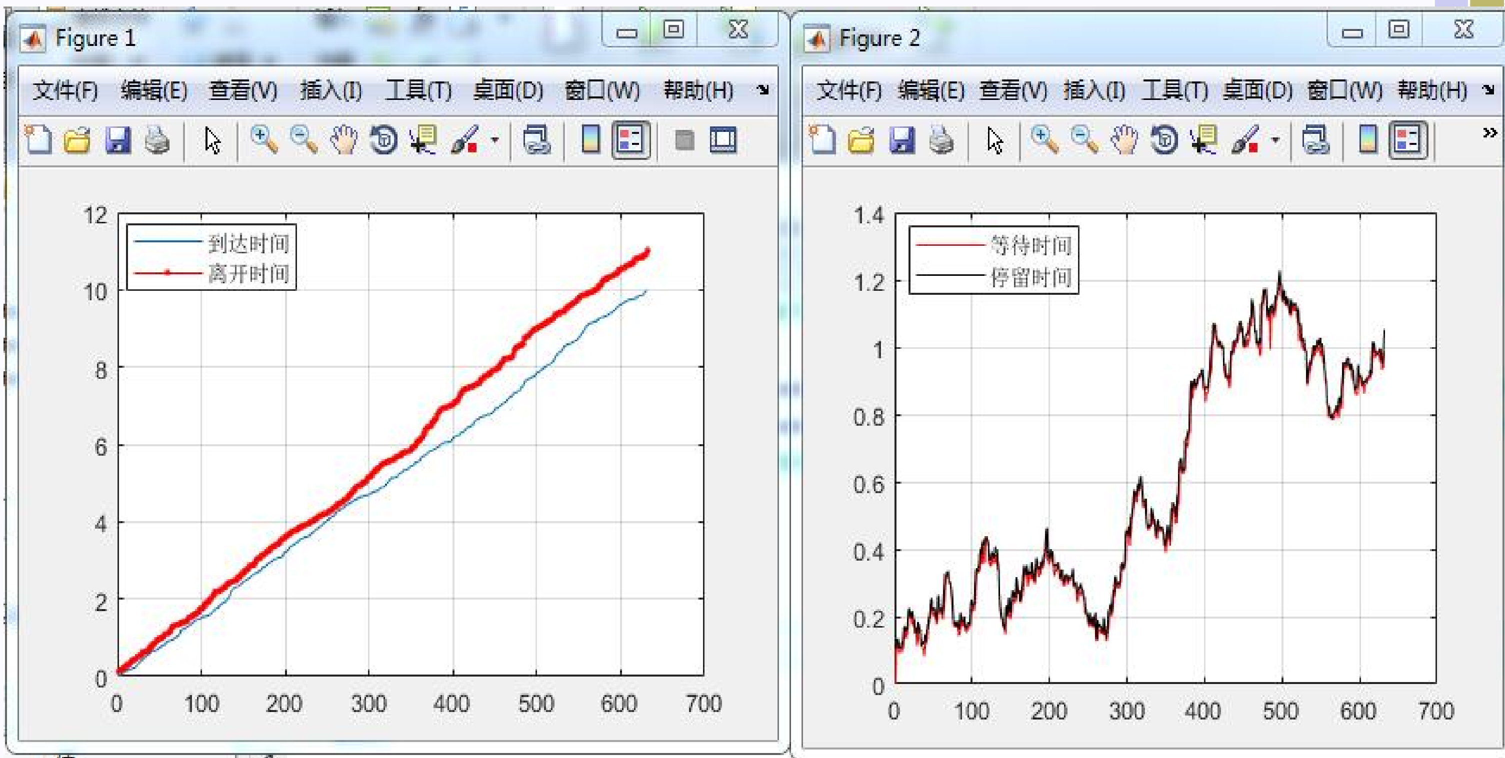
```
Total_time=10;
lambda=65;mu=60;
arr_mean=1/lambda;ser_mean=1/mu;  %平均到达时间与平均服务时间
arr_num=round(Total_time*lambda*2); %可能到达的最大顾客数（round四舍五入求整数）
guests=zeros(5,arr_num); %定义顾客信息的数组
guests(1,:)=exprnd(arr_mean,1,arr_num); %按指数分布产生各顾客到达的时间间隔
guests(1,:)=cumsum(guests(1,:)); %各顾客的到达时刻等于时间间隔的累积和
guests(2,:)=exprnd(ser_mean,1,arr_num); %按指数分布产生各顾客服务时间
len_sim=sum(guests(1,:)<=Total_time); %计算模拟的顾客个数，即到达时刻在模拟时间内的顾客数

%初始化第1个顾客的信息
guests(3,1)=0; %第1个顾客进入系统后直接接受服务，无需等待
guests(4,1)=guests(1,1)+guests(2,1); %其离开时刻等于其到达时刻与服务时间之和
guests(5,1)=0; %此时系统内没有其他顾客，故附加信息为0
member=[1]; %其进入系统后，系统内已有成员序号为1
```

```
%计算第i个顾客的信息
for i=2:arr_num
    if guests(1,i)>Total_time  %如果第i个顾客的到达时间超过了迭代时间，则跳出循环
        break;
    else
        %如果第i个顾客的到达时间在迭代时间内，则计算在其到达时刻系统中已有的顾客数
        number=sum(guests(4,member)>guests(1,i));
        if number==0
            guests(3,i)=0;  %如果系统为空，则第i个顾客直接接受服务，其等待时间为0
            guests(4,i)=guests(1,i)+guests(2,i);%其离开时刻等于到达时刻与服务时间之和
            guests(5,i)=0; %其附加信息为0
            member=[member,i];
        else
            %如果系统有顾客正在接受服务，且系统等待队列未满，则第i个顾客进入系统
            len_mem=length(member);
            guests(3,i)=guests(4,member(len_mem))-guests(1,i); %等待时间等于前一个顾客的离开时刻减去到达时刻
            guests(4,i)=guests(4,member(len_mem))+guests(2,i); %离开时刻等于前一个顾客的离开时刻加上服务时间
            guests(5,i)=number;%附加信息表示其进入系统时在他之前系统已有的顾客数
            member=[member,i];
        end
    end
end
len_mem=length(member); %模拟结束时，进入系统的总顾客数
```

```
%输出结果，绘制在模拟时间内，进入系统的所有顾客的到达时刻和离开时刻的曲线图
figure(1)
plot(1:len_mem,guests(1,member));hold on
plot(1:len_mem,guests(4,member),'.r-')
legend('到达时间','离开时间')
hold off,grid on

%绘制在模拟时间内，进入系统的所有顾客的停留时间和等待时间的曲线图
figure(2)
plot(1:len_mem,guests(3,member),'r-',1:len_mem,guests(2,member)+guests(3,member),'k-')
legend('等待时间','停留时间')
grid on
```




程序绘制除了各顾客到达时间与离开时间的阶梯图，同时也给出了各顾客等待时间与停留时间的曲线。

## 3.2. 案例2(停车库问题)

在建设一个大型商业中心之前，商业中心的设计者必须决定建设多大容量的地下停车库。无疑，每天车的到达率将是随时间而变化的，每天在变化，甚至每小时都在变化。一些顾客光顾商场可能很短的时间，而有些顾客可能光顾一整天。一旦车库建成后，商业中心可以根据负载量开放或关闭一些停车层（或区域），但首要问题是，停车库的最大容量应该为多大?

分析：由于商业中心的开发商希望几乎每个光临的顾客都能够停车，所以，在建模上我们就应该考察无穷大容量的停车库情形。也就是说，在事先不设置限制的条件下，拟合出实际停车数的概率分布。然后，停车库的容量应该被设置为这样一个值，它使得发生超出此库容量的情况为很小的概率。当然这个小概率的具体确定还要权衡考虑商业中心的可用空间和成本因素。

很明显，这是一个随机服务系统的问题。由于车辆到达流呈现明显的时间周期性，合适的假设是到达流服从非平稳的泊松过程，认为平均到达率$\lambda(t)$是随时间呈现某种周期性变化的泊松过程。我们将这种非平稳的泊松过程记为$M(t)$，它是对平稳泊松过程的一种推广。

停车时间仍然可以假设为服从平均服务时间为$\tau$的指数分布。这样，停车库问题就成为$M(t)/M/\infty$型的随机服务系统模型，其中以小时作为单位时间。

确定停车库容量的一个简便方法是，假设在时刻$t \geq 0$，车辆的到达数$N(t)$服从均值为$m(t)$的泊松分布，其中$m(t)$满足如下的微分方程：

$$\frac{d}{dt}m(t) = \lambda(t) - \frac{m(t)}{\tau} \tag{4}$$

其初始条件是$m(0) = 0$，即初始车库是空的。

通过求解上面方程（或者计算方程的数值解），找到$m(t)$的最大值，记为$m^*$。然后，在车库最拥堵状况下，考虑以$m^*$为均值的车辆数的泊松分布，就是说

$$P(N(t) \leq c) = \sum_{n=0}^{c} \frac{(m^*)^n}{n!} e^{-m^*}$$

最后，将$c$调整直到上面的概率足够接近于1，此时的$c$就是我们要确定的车库容量。

这种简化方法除了一些数值计算方法外几乎不需要模拟来求解。但对原模型做随机模拟将更有效。

## 模拟程序

我们假设停车库是每天连续24小时营业，时间单位是小时，即$T=24$。从已收集到的类似地方的停车库的数据中，估计出该停车库车辆的最大到达率为65（车/小时），其波动幅度约为20（车/小时），而服务率为60（车/小时）。我们以每天的营业时间作为一个周期来模拟每天的停车情况，所以同上例一样，每天的车辆总数也是有限的。我们可以用同样的公式来估算每天可能的最大车辆数。在这个问题中，并列的服务台个数是无穷多的。与上面例子的不同之处还在于，车辆到达流是非平稳的泊松过程。考虑到每天车辆活动的周期性，我们简单地假设每天车辆的到达率呈周期性变化，即

$$\lambda(t)=(\lambda_{max}-a)+a\sin(2\pi t/T)$$

其中，$\lambda_{max}=\max\lambda(t)$是可能的最大到达率，$a$是到达率的波动幅度（$0\le a\le\lambda_{max}$）。为了能生成服从这种非平稳泊松过程的随机序列，我们采用拒绝法的技术来采样这些随机数。

采样随机数的伪代码如下：

初始化$t_0 = 0, \tilde{t}_0 = 0$

令$k = 1$ for $n = 1, 2, \cdots$ do

    生成$\tilde{d}_n \sim E(\lambda_{max})$

    $\tilde{t}_n = \tilde{t}_{n-1} + \tilde{d}_n$

    生成随机数$rand \sim U(0, 1)$

    if $rand \le \lambda(\tilde{t}_n)/\lambda_{max}$ then

        $t_k = \tilde{t}_n$

        $d_k = t_k - t_{k-1}$

        $k = k + 1$

    end if

end for

在这样所生成的时间间隔的随机数序列$\{d_k\}$过程中，有一定比例的随机数被拒绝掉了。下面给出模拟程序，程序的基本构造与上例基本相同，但顾客信息中不需要等待时间和附加信息。

```
%初始化车辆流信息
%模拟一天的总时间
Total_time=24;
nTrials=1000; %试验次数
lambda=120;mu=0.4; %车辆的平均到达率与服务率
arr_mean=1/lambda;ser_mean=1/mu; %车辆的平均到达时间与平均服务时间
arr_num=round(Total_time*lambda*2); %可能到达的最大车辆数
%定义车辆信息数组和输出结果数组
guests=zeros(4,arr_num);
results=zeros(2,nTrials);
for iter=1:nTrials
    rng('shuffle')
    for i=1:arr_num
        arr_time=guests(1,i)+exprnd(arr_mean);
        while ~(rand<((lambda-40)+40*sin(2*pi*arr_time/Total_time))/lambda)
            arr_time=arr_time+exprnd(arr_mean);
        end
    guests(1,i)=arr_time-guests(1,i);
    end
    guests(1,:)=cumsum(guests(1,:)); % 各车辆的到达时间等于时间间隔的累积和
    guests(2,:)=exprnd(ser_mean,1,arr_num); %按指数分布产生各车辆的服务时间
    len_sim=sum(guests(1,:)<=Total_time); %计算模拟的车辆个数，即到达时刻在模拟时间内的车辆数
```

```
    %计算第i辆车的信息
    for i=2:arr_num
        if guests(1,i)>Total_time  %如果第i辆车的到达时间超过了模拟时间，则跳出循环
            break;
        else
            %如果第i辆车的到达时间未超过模拟时间，则计算在其到达时刻停车场中已有的车辆数
            number=sum(guests(3,member)>guests(1,i));
            guests(3,i)=guests(1,i)+guests(2,i); %车辆的离开时刻等于其到达时刻加上其服务时间
            guests(4,i)=number+1; %计算其进入停车库后，停车库内全部的车辆数
            member=[member,i];
        end
    end
    len_mem=length(member);%模拟结束时进入停车库的车辆总数
    %计算在停车库中当天车辆的平均数和最大数
    results(1,iter)=mean(guests(4,:));
    results(2,iter)=max(guests(4,:));
end
```

```
%输出结果，绘制在模拟时间内，进入停车库的所有车辆的到达时刻和离开时刻曲线图
figure(1)
plot(1:len_mem,[guests(1,member)],'k-');
axis([1 len_sim 0 Total_time+1])
title('车辆的到达时间')
xlabel('车辆的序号'),ylabel('时间/小时'),grid on
figure(2)
plot(1:len_mem,[guests(3,member)],'k-')
Ymax=fix(max(guests(3,member)))+1;
axis([1 len_sim 0 Ymax]);
title('车辆的离开时间');
xlabel('车辆的序号'),ylabel('时间/小时'),grid on
%绘制在模拟时间内，在停车库中各时刻车辆总数的曲线图
figure(3)
plot([0 guests(1,member)],[0 guests(4,member)],'k-')
Ymax=fix(max(guests(4,member)))+10;
axis([0 Total_time 0 Ymax])
title('一天内车库的停车数')
xlabel('时间/小时'),ylabel('车辆数/量'),grid on
```

```
%重复模拟试验结束后，绘制在停车库中每天车辆的最大数的经验分布图
%重复模拟试验结束后，绘制在停车库中每天车辆的平均数的直方图
figure(4)
hist(results(1,:),15)
figure(5)
[y,x]=hist(results(2,:),50);y=y/sum(y);y=cumsum(y);
plot(x,y,'k-')
title('停车库最大车辆数的经验分布函数')
xlabel('车辆数'),ylabel('概率'),grid on
```

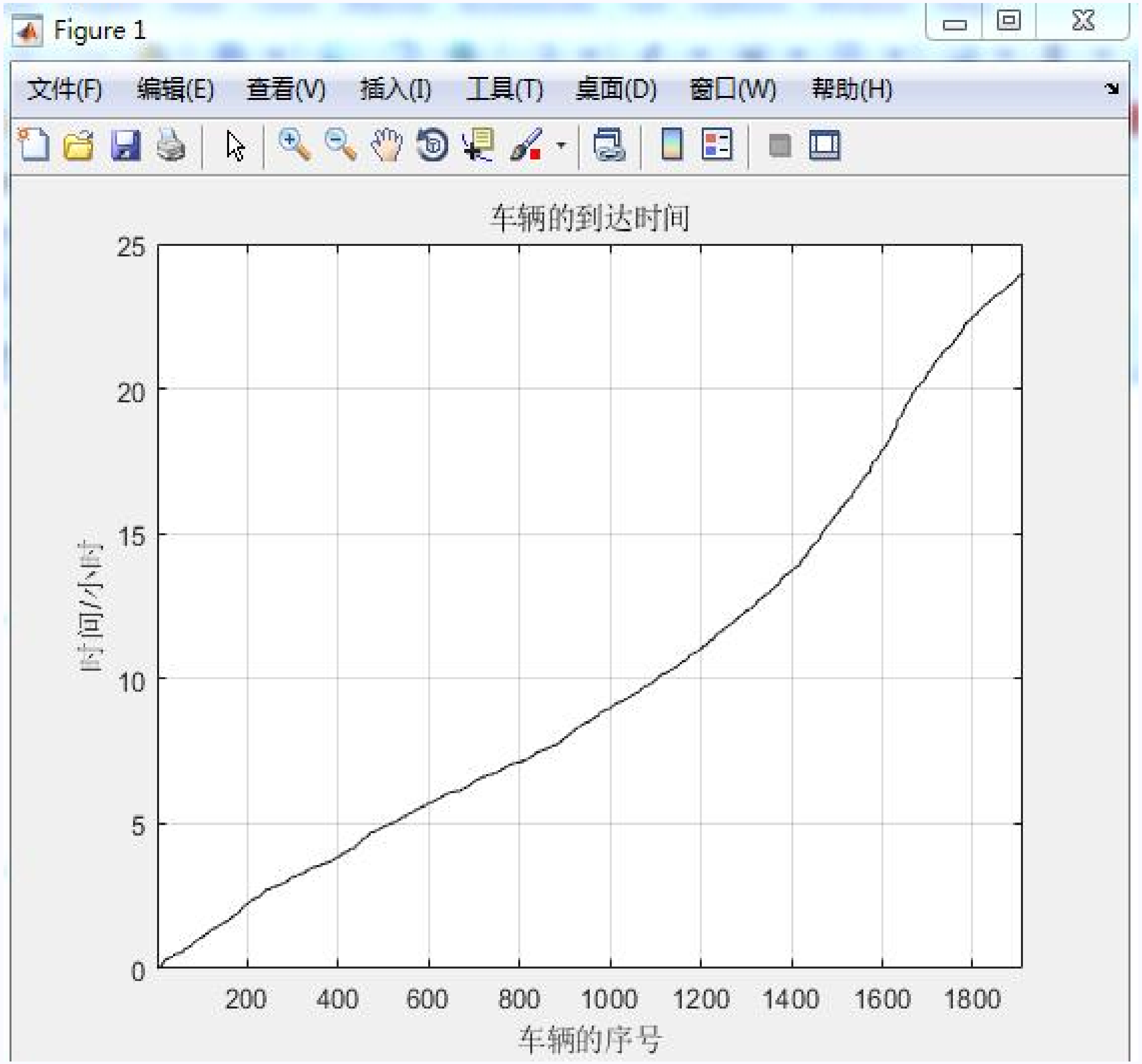


Figure 1
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
车辆的到达时间
25
20
15
10
5
0
时间/小时
200
400
600
800
1000
1200
1400
1600
1800
车辆的序号

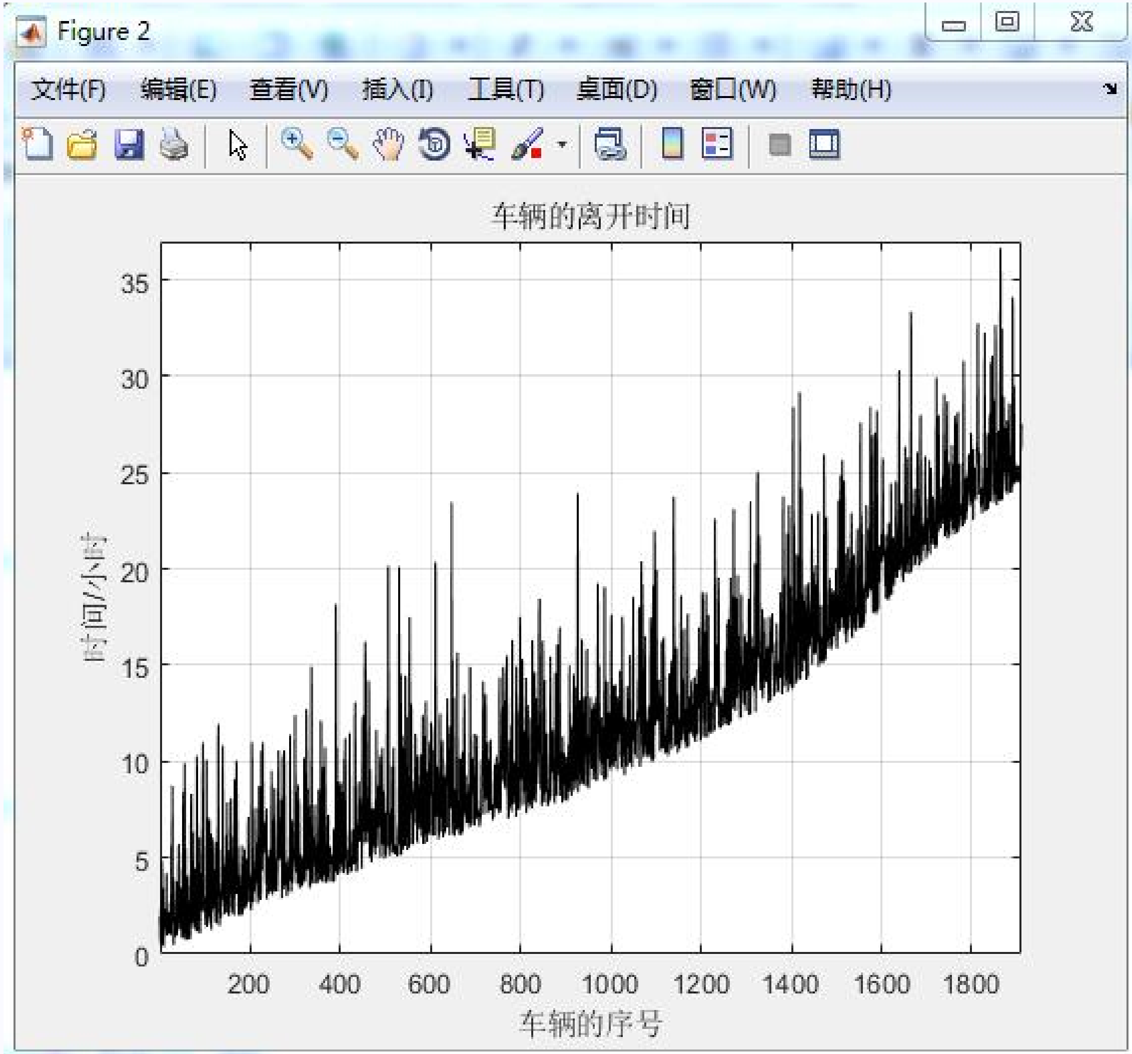


Figure 2
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
车辆的离开时间
时间/小时
0
5
10
15
20
25
30
35
200
400
600
800
1000
1200
1400
1600
1800
车辆的序号

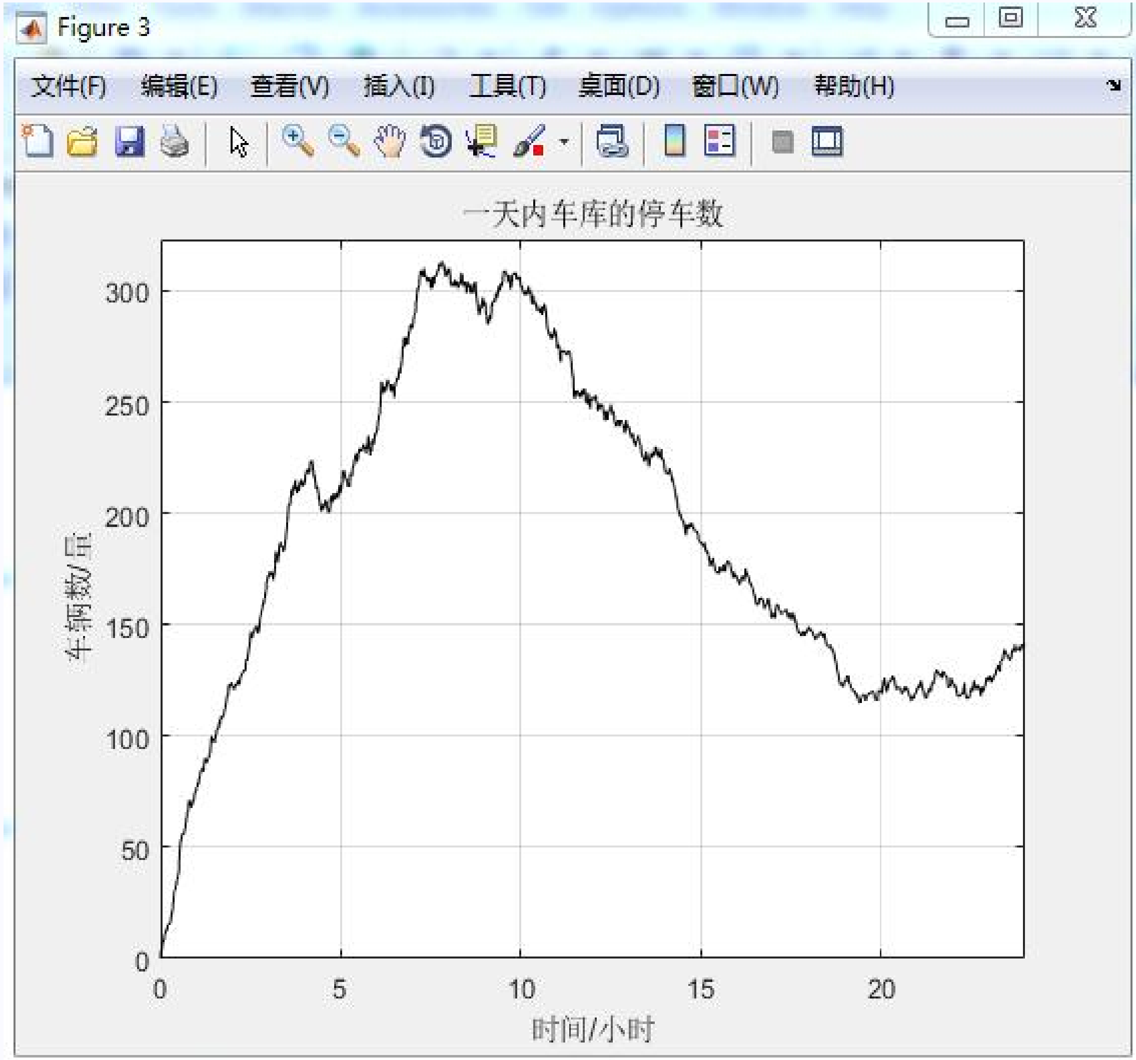
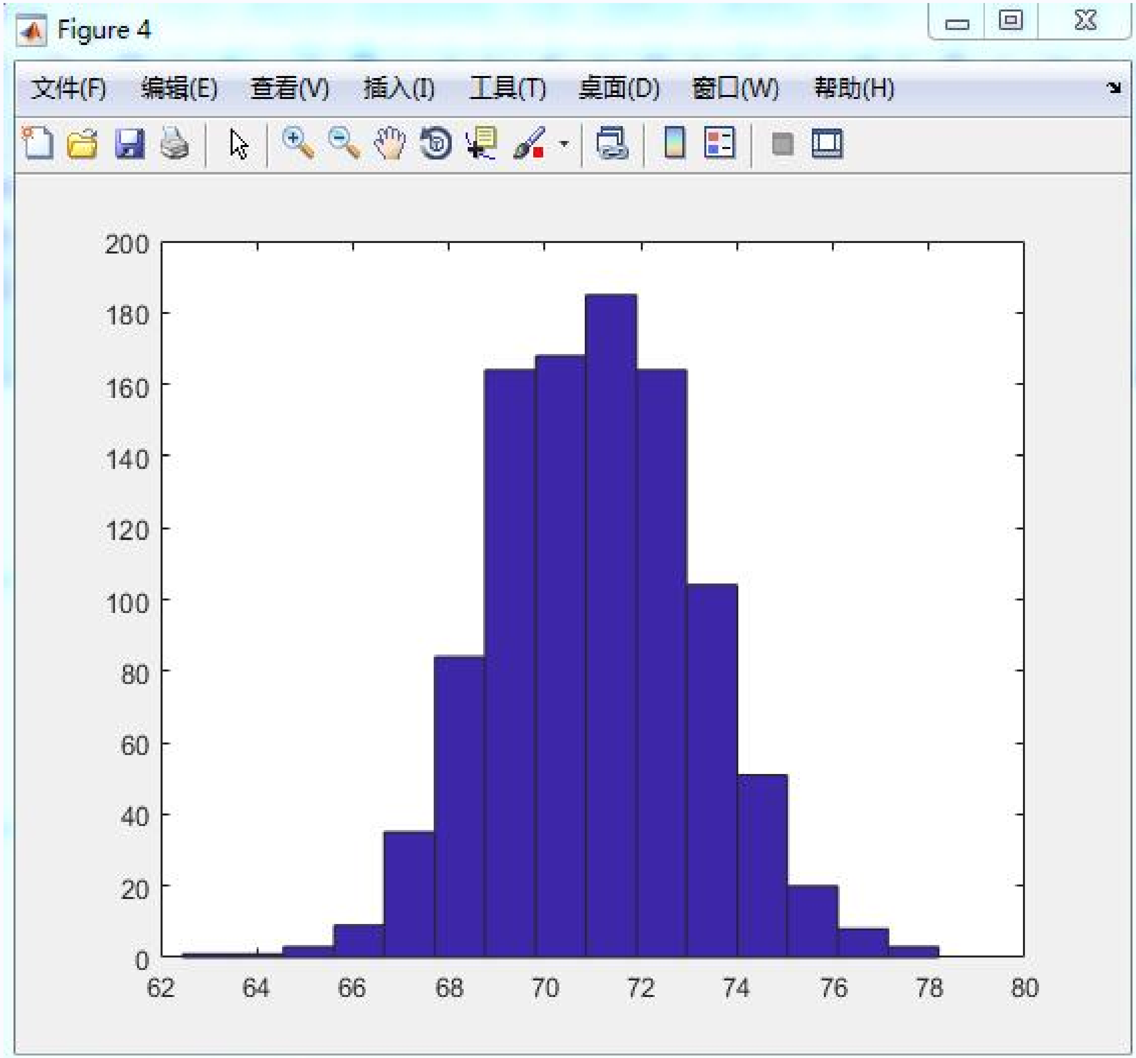


Figure 3
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
一天内车库的停车数
车辆数/台
时间/小时
300
250
200
150
100
50
0
0
5
10
15
20




马尔可夫链
仿真随机服务系统
案例分析
访 问 主 页
标 题 页
第 60 页 共 62 页
返 回
全 屏 显 示
关 闭
退 出

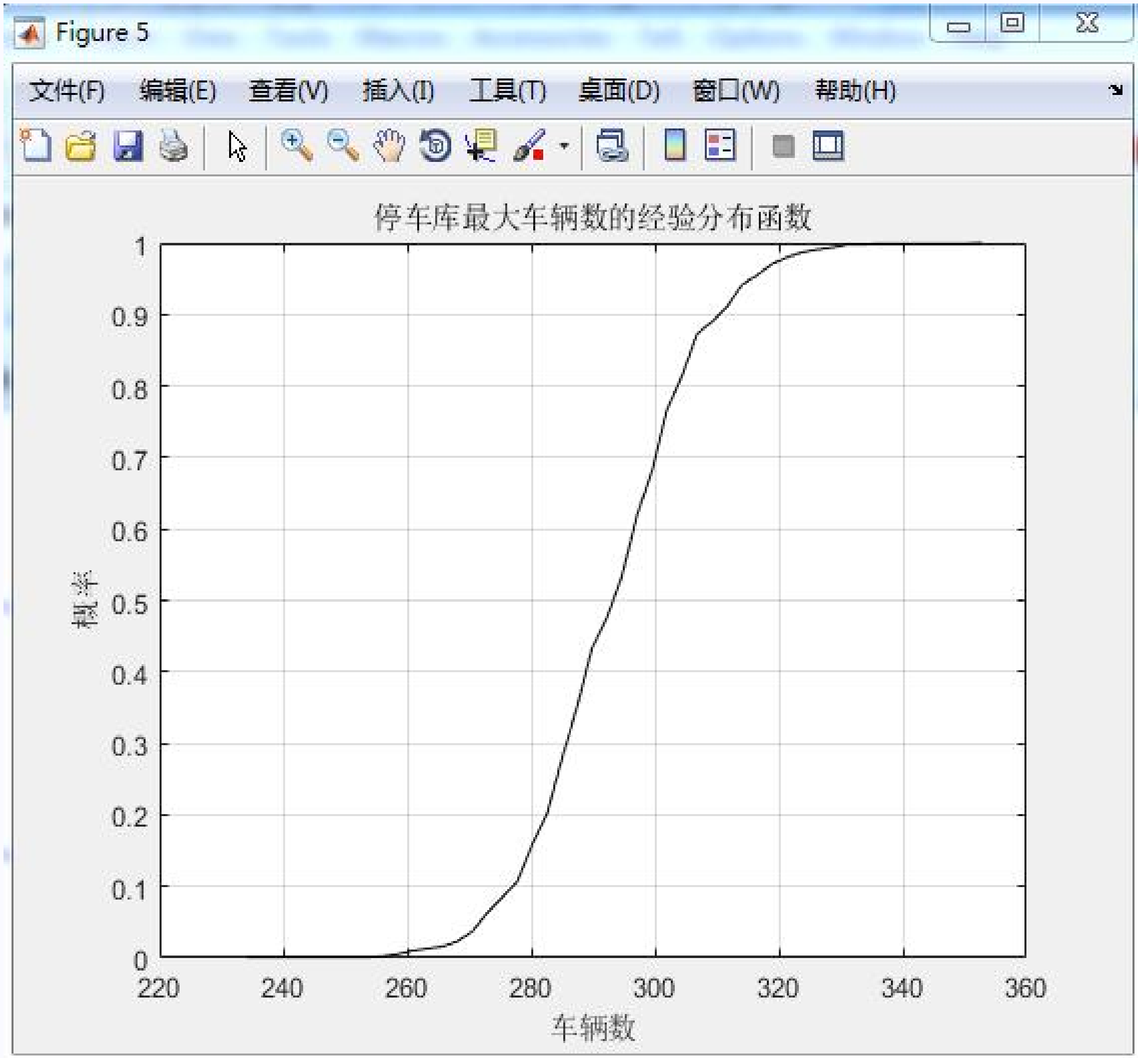


Figure 5
文件(F) 编辑(E) 查看(V) 插入(I) 工具(T) 桌面(D) 窗口(W) 帮助(H)
停车库最大车辆数的经验分布函数
概率
车辆数
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
220 240 260 280 300 320 340 360

从每天最大车辆数的经验分布可以看到，最大车辆数不超过328的概率为0.99，而最大车辆数不超过317的概率为0.95。停车库设计的目的是尽可能不失去每一位光临的顾客，所以模拟结果是建议该停车库的设计容量应该不低于328辆车。

从每天平均停车数来看，平均数不超过100辆，而且从一天的停车量看，大约有一半时间车辆数超过200辆，这说明可以考虑建二层停车库。