大学数学实验

Mathematical Experiments

第10讲

统计方法I 统计量和MC算法

本实验基本内容

一. 统计量的概念

二. Monte Carlo方法计算积分

三. Monte Carlo方法计算矩阵积和式

一. 数据的整理和描述

- **数据的收集和样本的概念**
- **数据的整理、频数表和直方图**
- **统计量**
- **MATLAB命令**

数据的收集

- 银行随机选了50名顾客进行调查
- 测量每个顾客感觉舒适时的柜台高度(单位：厘米)

100	110	136	97	104	100	95	120	119	99
126	113	115	108	93	116	102	122	121	122
118	117	114	106	110	119	127	119	125	119
105	95	117	109	140	121	122	131	108	120
115	112	130	116	119	134	124	128	115	110

- 银行怎样依据它确定柜台高度呢？

样本：统计研究的主要对象

- 总体--研究对象的全体。如所有顾客感觉舒适的高度
- 个体--总体中一个基本单位。如一位顾客的舒适高度
- 样本--若干个体的集合。如50位顾客的舒适高度
- 样本容量--样本中个体数。如50

■ 顾客群体的舒适高度~随机变量X，概率分布$F(x)$

■ n位顾客的舒适高度$\{x_i, i=1,\ldots n\}$ (样本)~相互独立的、分布均为F(x)的一组随机变量。

样本：随机取值的一组数据；

一组相互独立的、同分布的随机变量。

频数表

将数据的取值范围划分为若干个区间，统计这组数据在每个区间中出现的次数，称为频数，得到一个频数表。

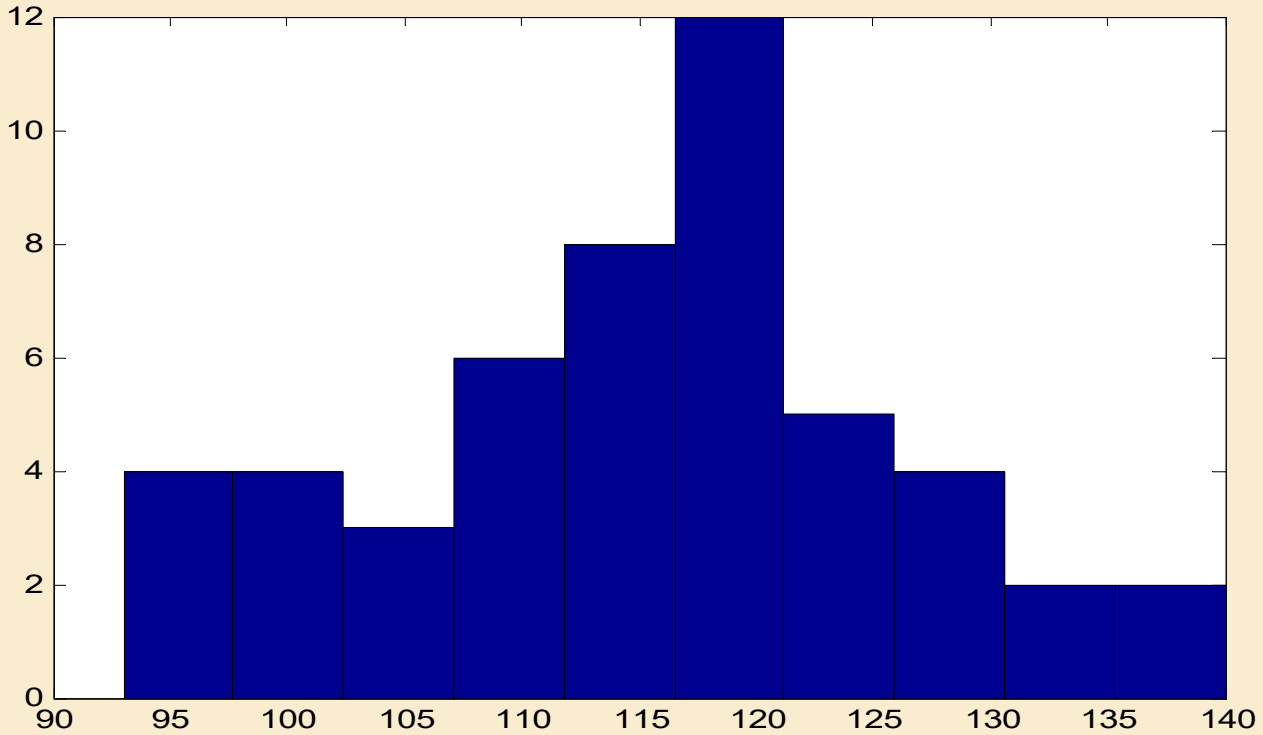
柜台高度频数表

中点	95.35	100.05	104.75	109.45	114.15	118.85	123.55	128.25	132.95	137.65
频数	4	4	3	6	8	12	5	4	2	2

作用：推测出总体的某些简单性质。

如上表表明选择柜台高度在107.10至125.90的有31人，占总人数的62%，柜台高度设计在这个范围内，会得到大多数顾客的满意。

直方图(histogram)：频数分布图



柜台高度直方图

平均值

频数表和直方图给出某个范围的状况，

无法直接给出具体值，如确定柜台具体高度

平均值 (mean，简称样本均值)定义为

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = 115.26$$

可作为设计柜台高度的参考值

例：两个班的一次考试成绩

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
甲班	92	88	85	92	95	79	84	87	88	65	93	73	88	87	94	80
乙班	84	83	82	85	82	81	82	90	84	78	75	83	78	85	84	79
序号	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
甲班	69	86	88	78	79	68	88	87	55	93	79	85	90	53	99	81
乙班	85	73	90	77	81	82	82	80	86	83	77	78				

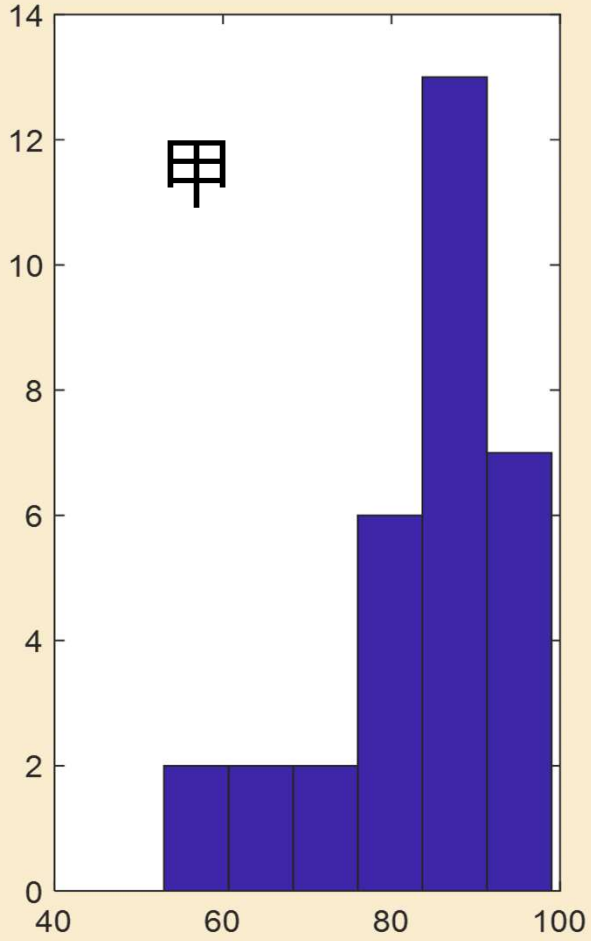
现象1：甲班平均值：82.75分，乙班平均值：81.75分

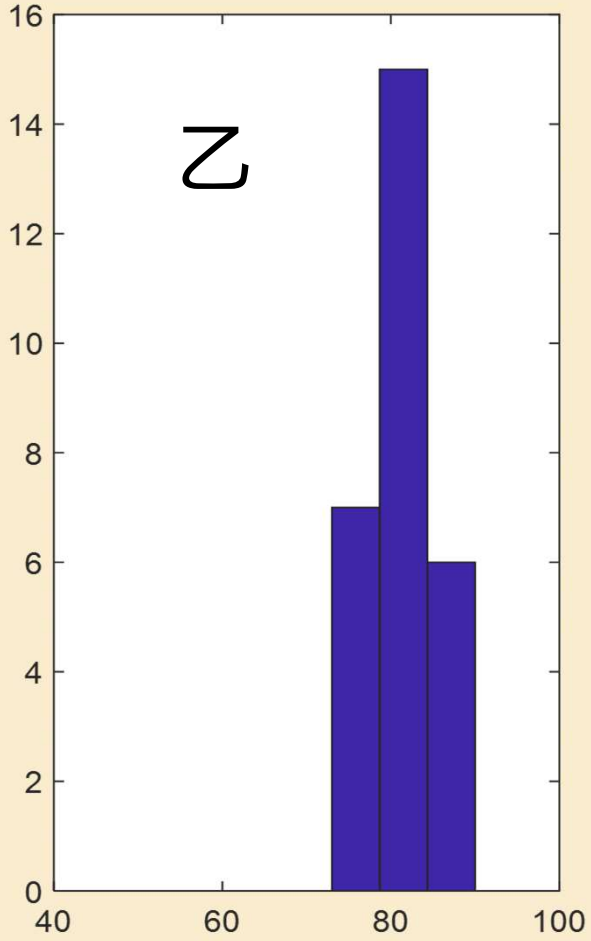
结　论：大致表明甲班的平均成绩稍高于乙班

现象2：甲班90分以上7人，但有2人不及格，分数分散

乙班全在73分到90分之间，分数相对集中

考试成绩直方图





标准差

描述数据的分散程度（统计上称为变异）

样本$x=(x_1, x_2, \ldots, x_n)$的标准差(Standard deviation)为：

$$s=\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{1/2}$$

甲班的标准差为10.98分，乙班的标准差为3.98分，表明甲班成绩的分散程度远大于乙班。

统计量：由样本加工出来的、集中反映样本特征的函数，$T(x_1, x_2, \ldots, x_n)$。

常用统计量

中位数(median)：将数据由小到大排序后处于中间位置的那个数值。

*n*为奇数时，中位数唯一确定；
*n*为偶数时，定义为中间两数的平均值

表示变异程度的还有：

极差(range)：$x_1, x_2, \ldots, x_n$的最大值与最小值之差。

表示分布形状的：

偏度(skewness)：分布对称性

峰度(kurtosis)：分布形状

MATLAB数据描述的常用命令

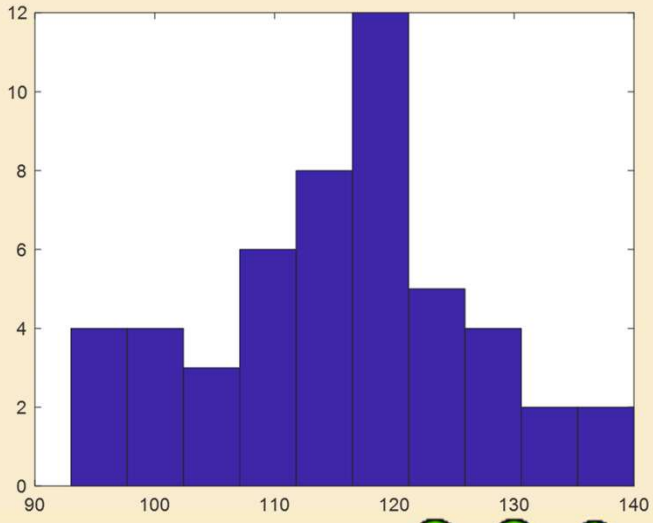
命令	名称	输入	输出
[n,y]=hist(x,k)	频数表	x: 原始数据行向量 k：等分区间数	n: 频数行向量 y: 区间中点行向量
hist(x,k)	直方图	同上	直方图
mean(x)	均值	x: 原始数据行向量	样本均值
median(x)	中位数	同上	中位数
range(x)	极差	同上	极差
std(x)	标准差	同上	样本标准差s
var(x)	方差	同上	样本方差s^2
skewness(x)	偏度	同上	偏度
kurtosis(x)	峰度	同上	峰度

求银行柜台高度的频数表、直方图及均值等统计量：

Exp10.m

```
% 柜台数据
X =[100  110  136  97  104  100  95  120  119  99 ...
       126  113  115  108  93  116  102  122  121  122 ...
       118  117  114  106  110  119  127  119  125  119 ...
       105  95  117  109  140  121  122  131  108  120 ...
       115  112  130  116  119  134  124  128  115  110];
[N,Y]=hist(X);                                  % 频数表
hist(X)                                         % 直方图
x(1)=mean(X);        x(2)=median(X);            % 各统计量
x(3)=range(X);       x(4)=std(X);
x(5)=skewness(X);    x(6)=kurtosis(X)
```

```
115.2600   116.5000    47.0000
10.9690    -0.0971     2.6216
```

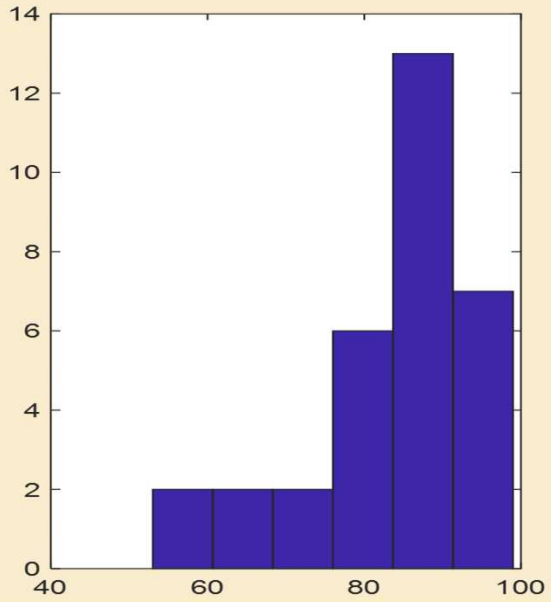


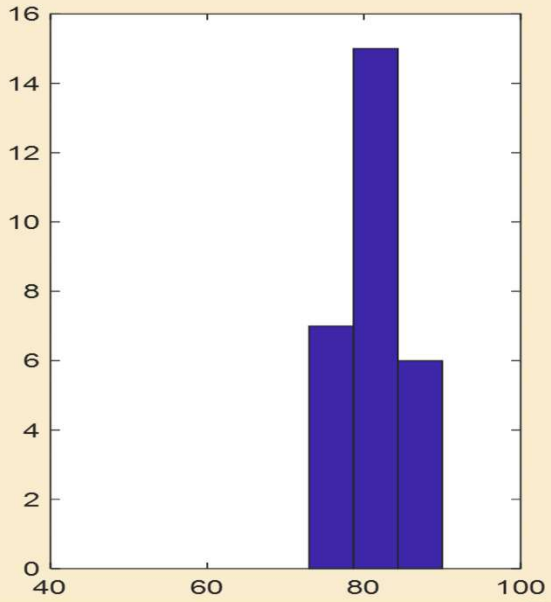
例：两个班的一次考试成绩

```
clear; close;
a1=[92,69,88,86,85,88,92,78,95,79,79,68,84,88,87,87,88,...
55,65,93,93,79,73,85,88,90,87,53,94,99,80,81];
a2=[84,83,82,85,82,81,82,90,84,78,75,83,78,85,84,79,85,...
73,90,77,81,82,82,80,86,83,77,78];
subplot(1,2,1); hist(a1,6);axis([40 100 0 14])
subplot(1,2,2); hist(a2,3);axis([40 100 0 16])
x(1,1)=skewness(a1);  x(1,2)=kurtosis(a1);
x(2,1)=skewness(a2);  x(2,2)=kurtosis(a2)
```

```
-1.1800    3.9469
-0.0204    3.0105
```

Exp10.m





MATLAB命令

分布	均匀分布	指数分布	正态分布	卡方分布	t分布	F分布	二项分布	泊松分布
字符	unif	exp	norm	chi2	t	f	bino	poiss

功能	概率密度	分布函数	逆概率分布	均值与方差	随机数生成
字符	pdf	cdf	inv	stat	rand

y=normpdf(1.5,1,2) 正态分布(μ=1, σ=2) 在x=1.5处的概率密度（标准正态分布的μ, σ可省略）

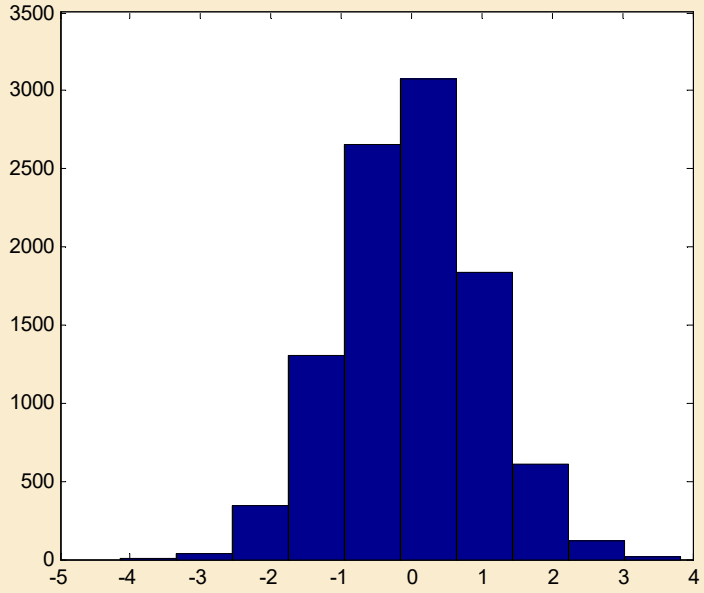
y=normcdf([-1 0 1.5],0,2) $N(0,2^2)$ 在 x= -1, 0, 1.5处分布函数值

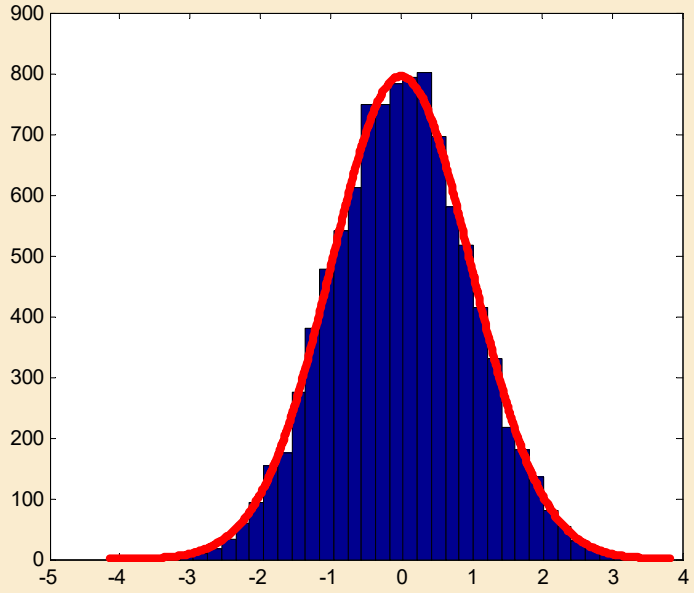
[m,v]=fstat(3,5) 计算$F(3,5)$的期望和方差

x=tinv(0.3,10) 计算$t(10)$的0.3-分位数

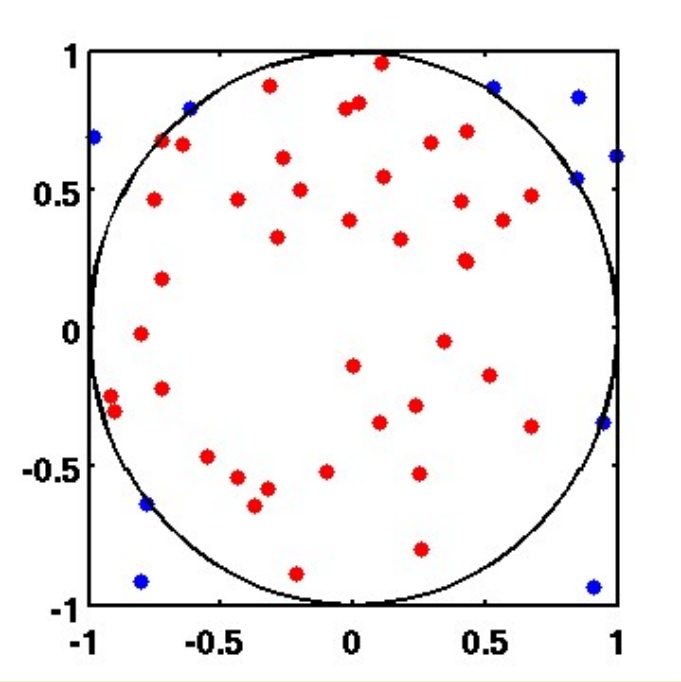
a=randn(10000,1); hist(a)

a=randn(10000,1); hist(a,40)

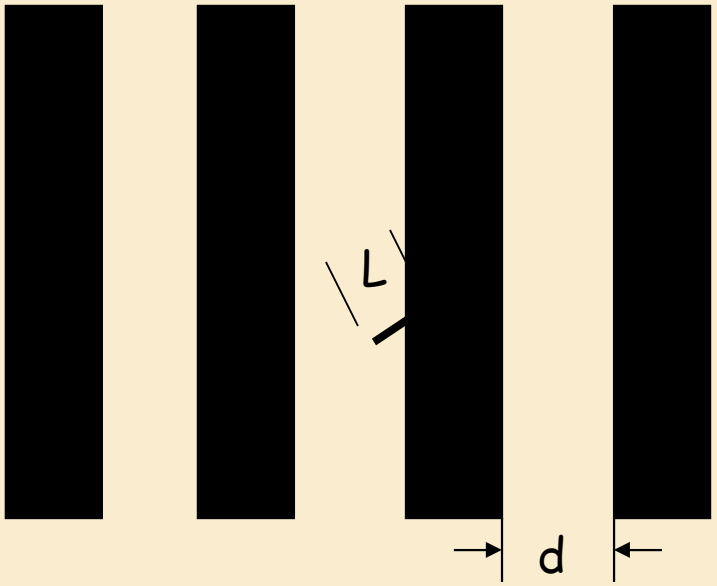




二. Monte Carla 方法



布丰投针实验(1777)



$$p=\frac{2L}{\pi d},\qquad L<d$$

利用布丰投针试验估计圆周率的一些历史资料

Experimenter	Length of needle	Number of casts	Number of crossings	Estimate for π
Wolf, 1850	.8	5000	2532	3.1596
Smith, 1855	.6	3204	1218.5	3.1553
De Morgan, c.1860	1.0	600	382.5	3.137
Fox, 1864	.75	1030	489	3.1595
Lazzerini, 1901	.83	3408	1808	3.1415929
Reina, 1925	.5419	2520	869	3.1795

Stanislaw Ulam (1909-1984)

S. Ulam is credited as the inventor of Monte Carlo method in 1940s, which solves mathematical problems using statistical sampling.

Nicholas Metropolis (1915-1999)

The algorithm by Metropolis (and A Rosenbluth, M Rosenbluth, A Teller and E Teller, 1953) has been cited as among the top 10 algorithms having the "greatest influence on the development and practice of science and engineering in the 20th century."

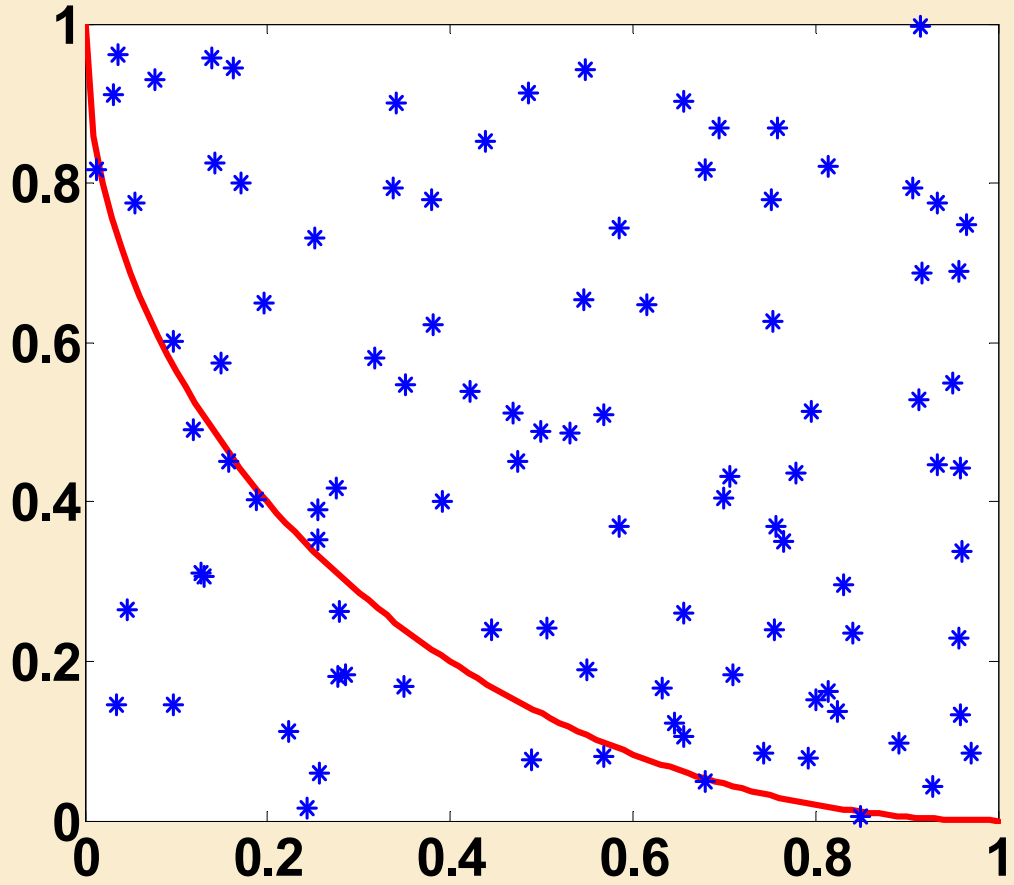
Monte Carlo方法名称的由来

Metropolis coined the name "Monte Carlo", from its gambling Casino.

Monte-Carlo, Monaco

计算积分的Monte Carlo方法

$$p=\int_0^1\left(1-\sqrt{x(2-x)}\right)dx\approx\frac{m}{n}$$



$$X\sim\begin{pmatrix}0 & 1\\ 1-p & p\end{pmatrix}$$

$$E(X)=p$$

$$\frac{X_1+X_2+\cdots+X_n}{n}\xrightarrow{P}p$$

$$\frac{m}{n}\approx E(X)=p$$

$$p=\int_0^1\left(1-\sqrt{x(2-x)}\right)dx\approx\frac{m}{n}$$

Exp10.m

```
n=10000;
x=rand(n,2);
y=1-sqrt(x(:,1).*(2-x(:,1)));
m=sum(x(:,2)<y);

f=inline('1-sqrt(x.*(2-x))');  % inline object

[m/n, quad(f,0,1), 1-pi/4]
```

0.2135 0.2146 0.2146

计算积分的Monte Carlo方法

$\iint_D f(x)\,dx$ 生成 D 内的均匀随机数 $x_1, x_2, \cdots x_n$,

$$\iint_D f(x)\,dx \approx S_D \cdot \frac{f(x_1)+f(x_2)+\cdots+f(x_n)}{n}$$

$$X \sim U(D), \quad E(f(X)) = \iint_D f(x)\cdot\frac{1}{S_D}dx, \quad \text{其中 } S_D = \iint_D dx$$

$$E(f(X)) \approx \frac{1}{n}\sum_{k=1}^{n} f(x_k) \quad \Rightarrow \quad \iint_D f(x)\,dx \approx S_D \cdot \frac{1}{n}\sum_{k=1}^{n} f(x_k)$$

一般区间重积分的计算 $\iint_{\Omega} f(x,y)dxdy$

分别为[a, b]和[c, d] 区间上的均匀分布随机数，$x_i, y_i \ (i = 1, \cdots n)$

判断每个点是否落在Ω域内，将落在Ω域内的m个点记作

$(x_k, y_k), \ k = 1, \cdots m$ 则

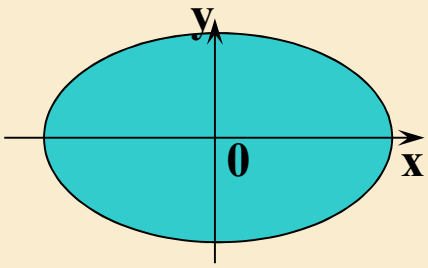
$$\iint_{\Omega} f(x,y)dxdy = S_{\Omega}\iint_{\Omega} f(x,y)\frac{1}{S_{\Omega}}dxdy = S_{\Omega} \bullet E\left(f\left(X,Y\right)\right)$$

$$\approx \frac{m}{n}(b-a)(d-c)\frac{1}{m}\sum_{k=1}^{m} f(x_k, y_k) = \frac{(b-a)(d-c)}{n}\sum_{k=1}^{m} f(x_k, y_k)$$

例:炮弹命中概率

目标: $a=1.2, b=0.8$(椭圆)

炮弹: $\sigma_x=0.6, \sigma_y=0.4$(独立)



$$P=\iint_{\Omega} p(x,y)dxdy, \quad \Omega: \frac{x^2}{a^2}+\frac{y^2}{b^2}\le 1$$

Ω_1是椭圆Ω在第1象限的部分

$$p(x,y)=\frac{1}{2\pi\sigma_x\sigma_y}e^{-\frac{1}{2}(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2})}$$

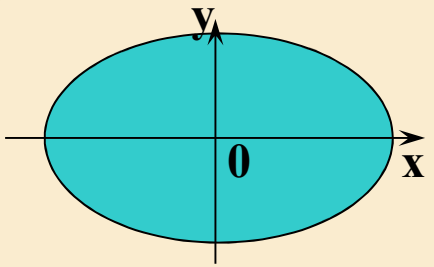
积分域和被积函数的对称性

$$P=4\iint_{\Omega_1} f(x,y)dxdy \cong \frac{4ab}{n}\sum_{k=1}^{m} p(x_k,y_k)$$

蒙特卡罗方法:x取$(0,a)$随机数, y取$(0,b)$随机数

```
a=1.2; b=0.8; sx=0.6;sy=0.4;
n=100000;m=0;z=0;
x=unifrnd(0,1.2,1,n);  y=unifrnd(0,0.8,1,n);
for i=1:n
   u=0;
   if x(i)^2/a^2+y(i)^2/b^2<=1
      u=exp(-0.5*(x(i)^2/sx^2+y(i)^2/sy^2));
      z=z+u;   m=m+1;
   end
end
p1=4*a*b*z/2/pi/sx/sy/n;

z=exp(-0.5*(x.^2/sx^2+y.^2/sy^2));
u=(x.^2/a^2+y.^2/b^2)<1;
p2=4*a*b*sum(z.*u)/2/pi/sx/sy/n;
```



Exp10.m

$$(X,Y)\sim N(0,0,1,1,0)$$

$$(0.6X,0.4Y)\sim N(0,0,0.6^2,0.4^2,0)$$

$$P\left(\frac{(0.6X)^2}{a^2}+\frac{(0.4Y)^2}{b^2}\le 1\right)$$

```
x1=(randn(n,2).^2)*[(sx/a)^2;(sy/b)^2]; sum(x1<1)/n
```

无偏估计

估计某参数 θ, $\hat{\theta} = g(X_1, X_2, \cdots, X_n)$

如果 $E(\hat{\theta}) = \theta$, 称 $\hat{\theta}$ 为参数 θ 的无偏估计,

$\hat{\theta}_1$ 与 $\hat{\theta}_2$ 均为参数 θ 的无偏估计, 若 $Var(\hat{\theta}_1) \le Var(\hat{\theta}_2)$,
则 $\hat{\theta}_1$ 优于 $\hat{\theta}_2$

MC方法应用实例II

矩阵积和式 permanent

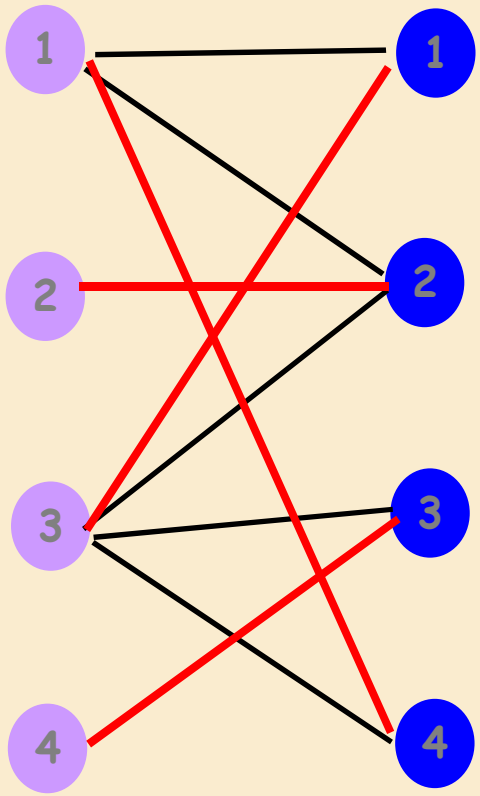
$$\operatorname{per} A = \sum_{\pi} a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)}$$

$$\det A = \sum_{\pi} \operatorname{sign} \pi \cdot a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)}$$

Theorem: Computing the permanent of a (0,1) matrix is #P-complete.

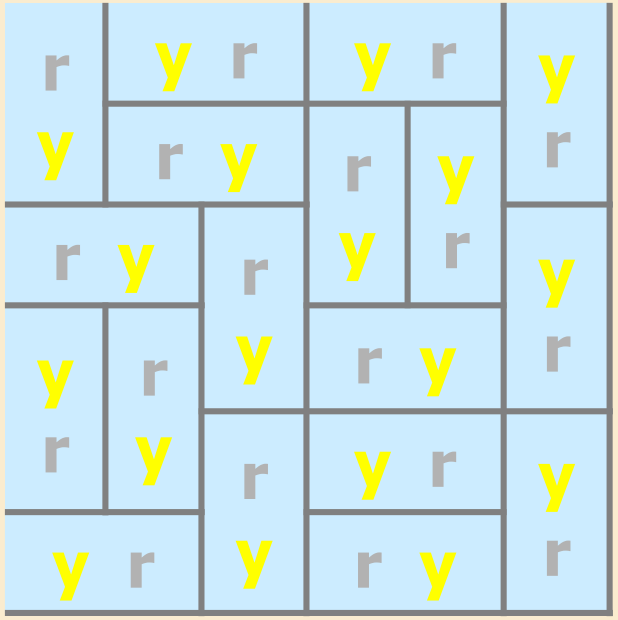
（Valiant *SIAM COMP 1979* *TCS 1979*）

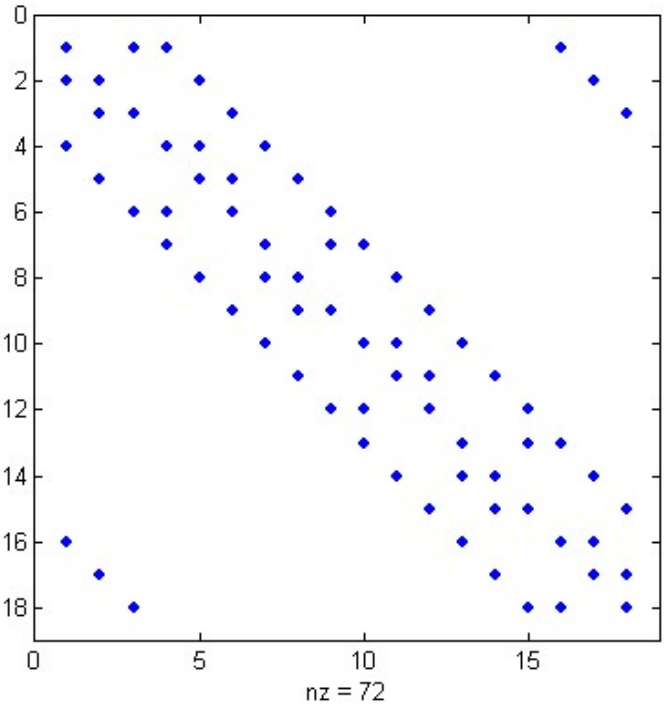
Perfect Matching and Permanent



	1	2	3	4
1	1	1	0	1
2	0	1	0	0
3	1	1	1	1
4	0	0	1	0

双分子覆盖模型





$$\lambda_d = \lim_{m\to\infty} \frac{\ln[Per(A_m)]}{m^d}$$

P.W. Kastelyn, The statistics of dimers on a lattice, Physica 27 (1961) 1209-1225

关于矩阵积和式的著名问题

- **Polya problem(1913)**

  For 0-1 matrix A, under what conditions does there exist a matrix B obtained from A by changing some of the 1's to -1's such a way that perm(A)=det(B)

  solved by Seymour et al. (Ann. Math. 1999)

- **van der Waerden conjecture(1926)**

  For all double stochastic matrix A, $perm(A) \geqslant n!/n^n$

  proved in 1980 (Ann. Math. 1962,1979)

- **M.Agrawal proposed new conjecture on expressing permanent of matrices as determinant of (possibly larger) matrices**

  (2006 ICM 45 minutes invited talk)

计算积和式的MC方法

0-1矩阵积和式计算的Godsil / Gutman估计子（1980）

将0-1矩阵A随机化，0的位置保持不变，1的位置以$\frac{1}{2}$概率取1和-1

则：$\mathrm{E}\left(\det\left(\mathrm{B}\right)^2\right)=\mathrm{per}(\mathrm{A})$

$$\begin{pmatrix}1 & 0\\ 1 & 1\end{pmatrix}\Rightarrow\begin{pmatrix}b_{11} & 0\\ b_{21} & b_{22}\end{pmatrix}$$

A　　B

$\det(\mathrm{B})$是一个随机变量

计算积和式的MC方法的精度分析

$$Y=\left(\det(B)\right)^2$$

$$E(Y)=E(\bar{Y})=per(A)$$

$$\bar{Y}=\frac{\sum_{k=1}^{n}Y_k}{n}=\frac{\sum_{k=1}^{n}\left(\det(B_k)\right)^2}{n}$$

$$P\left[(1-\varepsilon)\text{per}\,A\le\bar{Y}\le(1+\varepsilon)\text{per}\,A\right]\ge 1-\delta$$

$$P\left(\left|\bar{Y}-E(\bar{Y})\right|\le\varepsilon Per(A)\right)\ge 1-\frac{Var(\bar{Y})}{\varepsilon^2 Per(A)^2}=1-\frac{Var(Y)}{n\varepsilon^2 E(Y)^2}\ge 1-\delta$$

$$\Rightarrow\quad n\ge\frac{1}{\varepsilon^2\delta}\cdot\frac{Var(Y)}{E(Y)^2}$$

SIAM J. COMPUT.
Vol. 22, No. 2, pp. 284–293, April 1993

© 1993 Society for Industrial and Applied Mathematics
005

A MONTE-CARLO ALGORITHM FOR ESTIMATING THE PERMANENT*

N. KARMARKAR†, R. KARP‡, R. LIPTON§, L. LOVÁSZ¶, AND M. LUBY[1]

Abstract. Let A be an $n \times n$ matrix with 0-1 valued entries, and let $\mathrm{per}(A)$ be the permanent of A. This paper describes a Monte-Carlo algorithm that produces a "good in the relative sense" estimate of $\mathrm{per}(A)$ and has running time $\mathrm{poly}(n)2^{n/2}$, where $\mathrm{poly}(n)$ denotes a function that grows polynomially with n.

Key words. permanent, matching, Monte-Carlo algorithm, algorithm, bipartite graph, determinant

AMS(MOS) subject classifications. 05C50, 05C70, 68Q25, 68R05, 68R10

*Received by the editors November 12, 1990; accepted for publication (in revised form) December 19, 1991.

†AT&T Bell Laboratories, Incorporated, Murray Hill, New Jersey 07974-2010.

‡Computer Science Division, University of California, Berkeley, California 94720, and International Computer Science Institute, Berkeley, California 94704-1105. The research of this author was supported by National Science Foundation grant CCR-9005448.

§Computer Science Department, Princeton University, Princeton, New Jersey 08544-0001.

¶Hungarian Academy of Sciences, 1051 Budapest, Roosevelt-Ter9, Hungary and Computer Science Department, Princeton University, Princeton, New Jersey 08544-0001.

[1]International Computer Science Institute, Berkeley, California 94720. The research of this author was partially supported by National Science Foundation operating grant CCR-9016468 and by grant number 89-00312 from the United States–Israel Binational Science Foundation (BSF), Jerusalem, Israel.

2021/4/14

The Godsil/Gutman estimator is defined as follows:

(1) An $n \times n$ matrix B is formed from A as follows:

For all i, j, $1 \le i, j \le n$,

If $A_{ij} = 0$ then $B_{ij} \leftarrow 0$

Elseif $A_{ij} = 1$ then randomly and independently choose $B_{i,j} \in \{-1, 1\}$, each choice with probability $\frac{1}{2}$.

(2) $Y \leftarrow (\det(B))^2$.

$\mathrm{per}(A)$. Let

$$w_0 = 1, \quad w_1 = -\frac{1}{2} + \frac{\sqrt{3}}{2}i, \quad w_2 = -\frac{1}{2} - \frac{\sqrt{3}}{2}i$$

be the three cube roots of unity. If $y = a + bi$ is a complex number, then $\overline{y} = a - bi$ is the complex conjugate of y.

The estimator is computed as follows.

(1) An $n \times n$ matrix B is formed from A as follows:

For all i, j, $1 \le i, j \le n$,

If $A_{i,j} = 0$ then $B_{i,j} \leftarrow 0$

Elseif $A_{i,j} = 1$ then randomly and independently choose $B_{i,j} \in \{w_0, w_1, w_2\}$, each choice with probability $\frac{1}{3}$.

(2) $Z \leftarrow \det(B)\overline{\det(B)}$.

Exp10.m

```
% MC for permanent
nn=10;  %  the order of the matrix
A=rand(nn)>0.4;  p(1)=Nperm(A);  n=1000;

% GG estimator
for k=1:n
   C=rand(nn)>0.5;   C=2*C-1;    B=A.*C;    gg(k)=det(B)^2;
end
p(2)=mean(gg);

% KKLLL estimator
clear i;
for k=1:n
   C=rand(nn);    C1=(-1/2+i*sqrt(3)/2)*(C>2/3);    C2=(-1/2-i*sqrt(3)/2)*(C<1/3);
   CC=C1+C2+((C>1/3).*(C<2/3));    B=A.*CC;
   kk(k)=abs(det(B))^2;
end
p(3)=mean(kk);      p
```

无偏性的证明

$$E\left[(\det \boldsymbol{B})^2\right]$$

$$= E\left[\sum_{j_1 j_2 \cdots j_n} (-1)^{\tau(j_1 j_2 \cdots j_n)} b_{1 j_1} b_{2 j_2} \cdots b_{n j_n} \cdot \sum_{i_1 i_2 \cdots i_n} (-1)^{\tau(i_1 i_2 \cdots i_n)} b_{1 i_1} b_{2 i_2} \cdots b_{n i_n}\right]$$

$$= E\left[\sum_{j_1 j_2 \cdots j_n} \left(b_{1 j_1} b_{2 j_2} \cdots b_{n j_n}\right)^2 + \sum_{\sigma_s \neq \sigma_t} (-1)^{\tau(\sigma_s)+\tau(\sigma_t)} b_{1\sigma_s(1)} \cdots b_{n\sigma_s(n)} b_{1\sigma_t(1)} \cdots b_{n\sigma_t(n)}\right]$$

$$= E\left[\sum_{j_1 j_2 \cdots j_n} \left(b_{1 j_1} b_{2 j_2} \cdots b_{n j_n}\right)^2\right] = \sum_{j_1 j_2 \cdots j_n} E\left(b_{1 j_1}{}^2\right) \cdots E\left(b_{n j_n}{}^2\right) = per(A)$$

COROLLARY 5.3. *The Godsil/Gutman estimator yields an* (ϵ,δ)*- approximation algorithm for estimating* $\mathrm{per}(A)$ *which runs in time* $\mathrm{poly}(n)3^{n/2}\frac{1}{\epsilon^2}\log(\frac{1}{\delta})$.

Proof. Each evaluation of the estimator can be performed in time $\mathrm{poly}(n)$. Also,

$$\frac{\mathrm{E}[Y^2]}{\mathrm{E}[Y]^2} = \frac{\sum_{G\in D} 6^{c(G)}}{\sum_{G\in D} 2^{c(G)}} \leq \max_{G\in D} 3^{c(G)} \leq 3^{n/2},$$

COROLLARY 6.3. *The estimator Z yields an* (ϵ, δ)*-approximation algorithm for estimating* $\mathrm{per}(A)$ *which runs in time* $\mathrm{poly}(n)2^{n/2}\frac{1}{\epsilon^2}\log(\frac{1}{\delta})$.

Proof. Each evaluation of the estimator can be performed in time $\mathrm{poly}(n)$. Also,

$$\frac{\mathrm{E}[Z^2]}{\mathrm{E}[Z]^2} = \frac{\sum_{G\in D} 4^{c(G)}}{\sum_{G\in D} 2^{c(G)}} \leq \max_{G\in D} 2^{c(G)} \leq 2^{n/2},$$

布置实验

目的

1) 掌握统计量的基本概念；

2) 掌握用随机的方法(蒙特卡罗法)计算积分；

3) 运用蒙特卡洛方法解决一般问题

内容　见网络学堂

作业

1. 炮弹射击的目标为一半径100m的圆形区域，弹着点服从以目标为中心的二元正态分布,X和Y方向上的标准差分别为80m和50m，相关系数为0.4。用Monte Carlo方法求炮弹命中圆形区域的概率，并与利用数值积分方法得到的结果进行比较。

2. 考察不同规模与不同稀疏度的0-1矩阵，对GG与KKLLL两种计算矩阵积和式的Monte Carlo方法进行比较。