5.基于近似方法的贝叶斯数据分析

戴俊毅

研究员/长聘副教授

格点近似法

- 当模型只包含少量参数(一般为1-2个)时，可以采用格点法，将连续分布离散化，得到先验分布和后验分布的近似解
- 例如，对于表达概率或者占比的参数Θ，可以将参数的可能取值范围[0,1]分成等距的有限个点，然后按照这些点的先验概率密度，将100%的总概率，成比例地分配在这些点上，作为先验分布的近似解
- 由于后验概率密度正比于先验概率密度和似然函数的乘积，以上方法也可用于后验分布的近似解

格点近似法

- 例如，假定Θ的先验分布呈单峰等腰形式，即

$$p(\theta) \propto \min(\theta, 1-\theta)$$

且将[0,1]区间分为等距的5个点，即0, 1/4, 1/2, 3/4和1，那么对应的概率密度分别正比于0, 1/4, 1/2, 1/4, 0，因此，每个格点分配到的概率分别为0, 1/4, 1/2, 1/4, 0除以(0+1/4+1/2+1/4+0), 也就是0, 1/4, 1/2, 1/4, 0

格点近似法

- 再假定数据为进行了一次试验且结果为正，那么似然函数为

$$P(D|\theta) = L(\theta; N = 1, z = 1) = \theta^z(1-\theta)^{N-z} = \theta$$

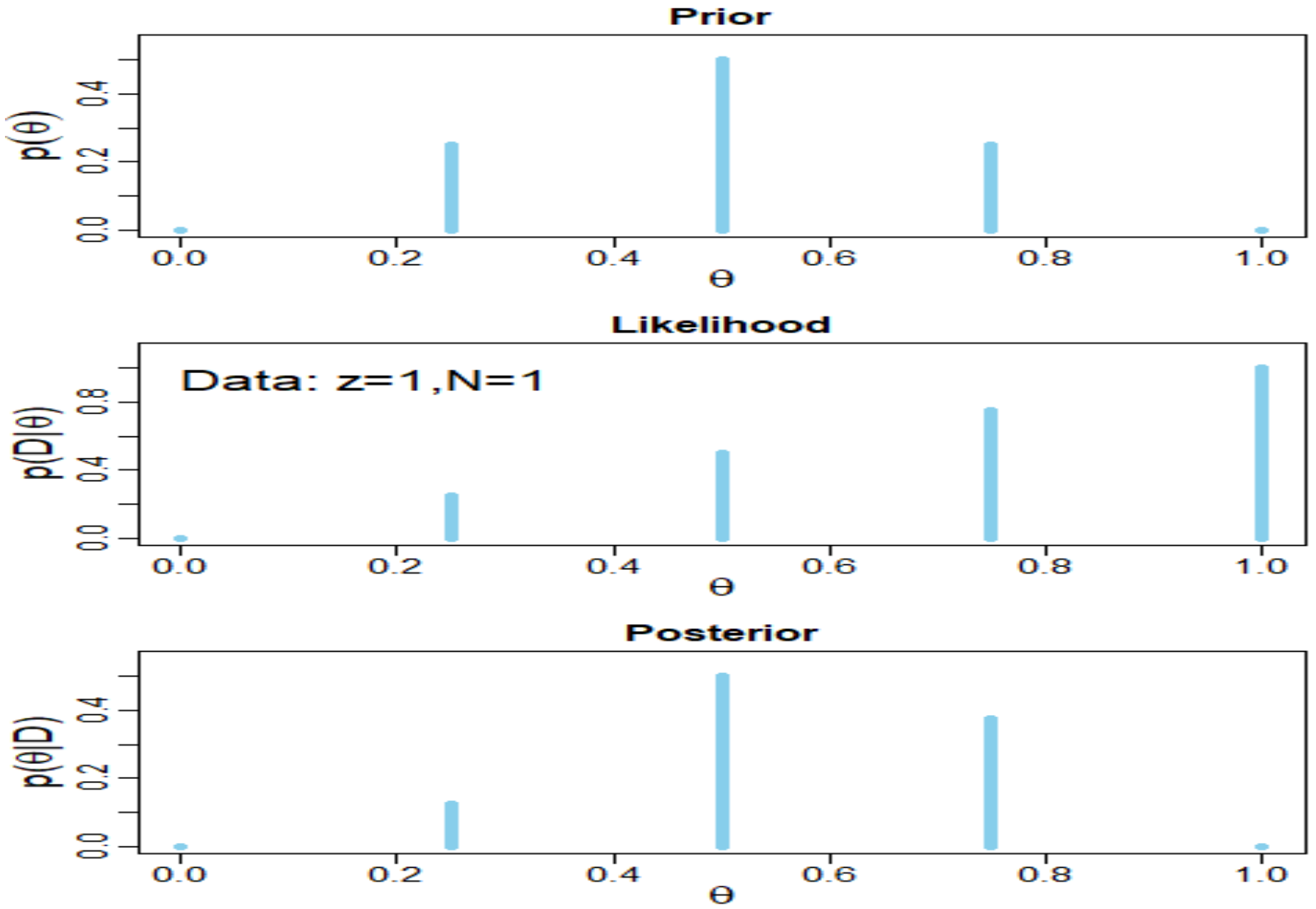
- 那么后验概率密度

$$p(\theta|D) \propto p(\theta) \times P(D|\theta) \propto \min(\theta, 1-\theta) \cdot \theta$$

- 因此，对应格点分配到的后验概率分别正比于0·0, 1/4·1/4, 1/2·1/2, 3/4·1/4, 1·0, 即0, 1/16, 1/4, 3/16, 0, 和为1/2

- 所以，后验概率近似值分别为0, 1/8, 1/2, 3/8, 0

- $P(D) \approx \sum_i^5 p(\theta)P(D|\theta) = 0 \cdot 0 + \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{2} + \frac{1}{4} \cdot \frac{3}{4} + 0 \cdot 1 = 1/2$


Prior
p(θ)
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
θ
Likelihood
Data: z=1,N=1
p(D|θ)
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
θ
Posterior
p(θ|D)
0.0
0.2
0.4
0.0
0.2
0.4
0.6
0.8
1.0
θ

格点近似法

- 我们可以增加格点的个数，使得离散近似解更接近于连续解。在此例中，存在解析解
- 先验分布的解析解为

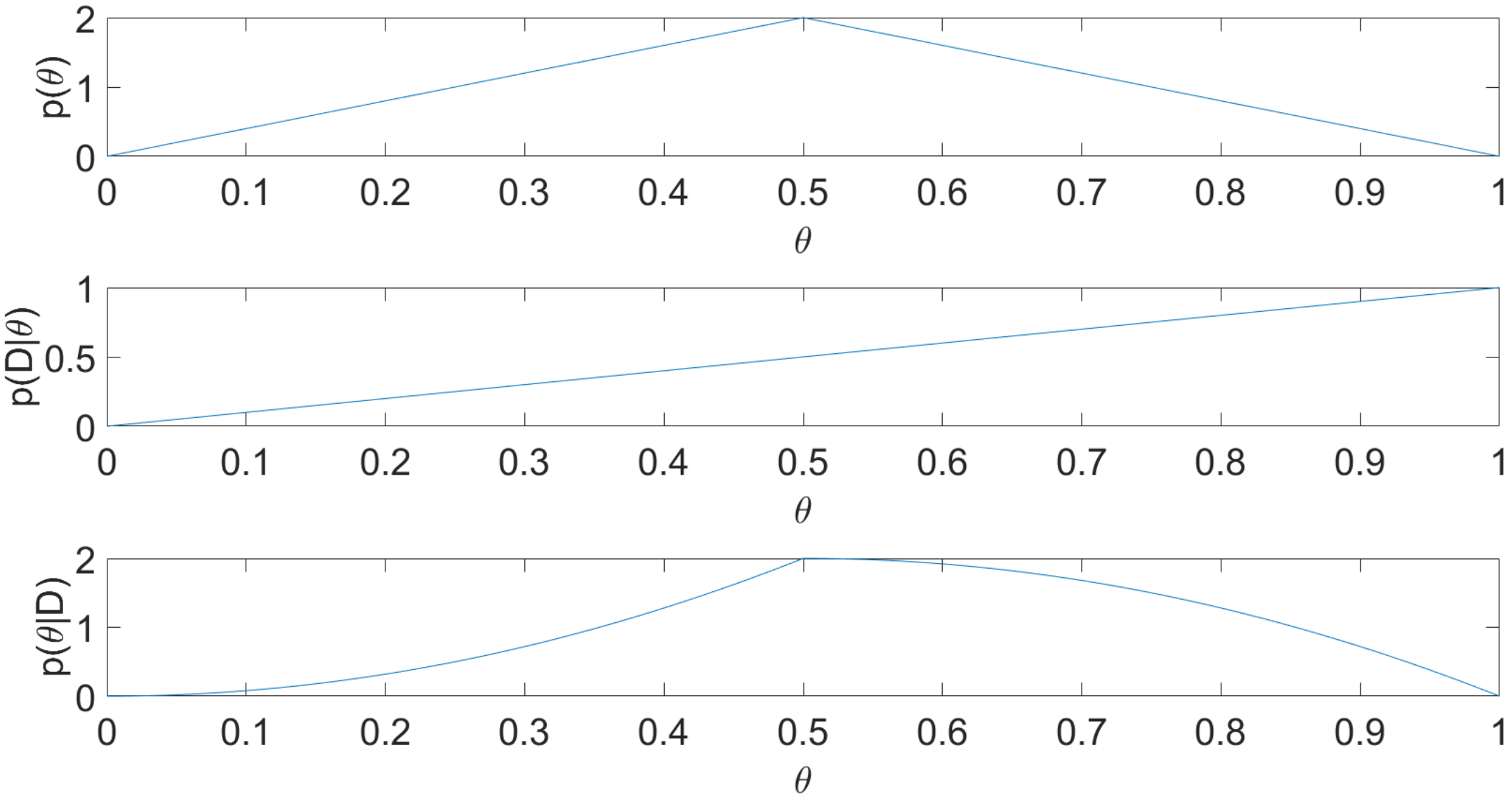
$$p(\theta) = 4 \times \min(\theta, 1-\theta)$$

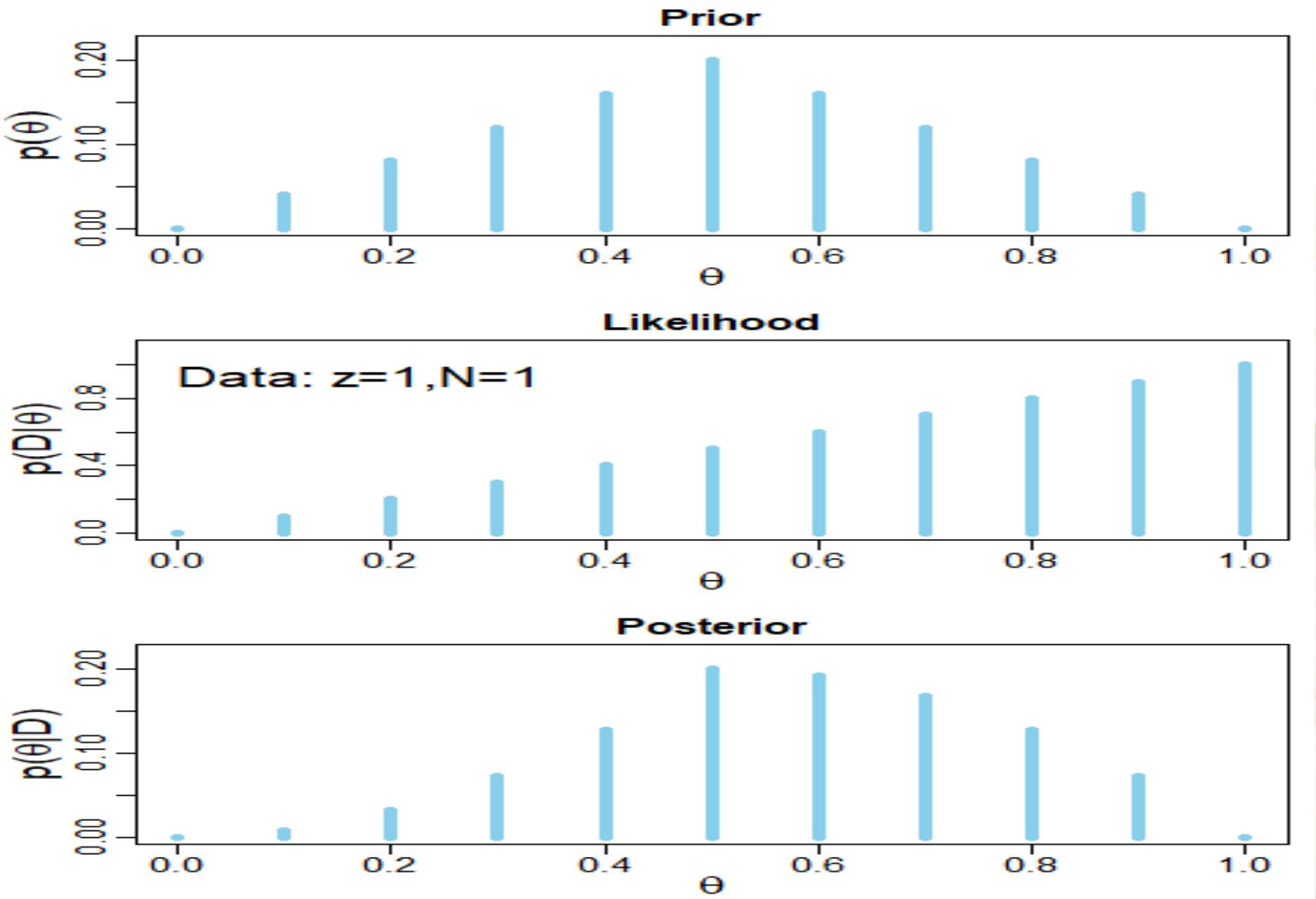
- P(D)的解析解为

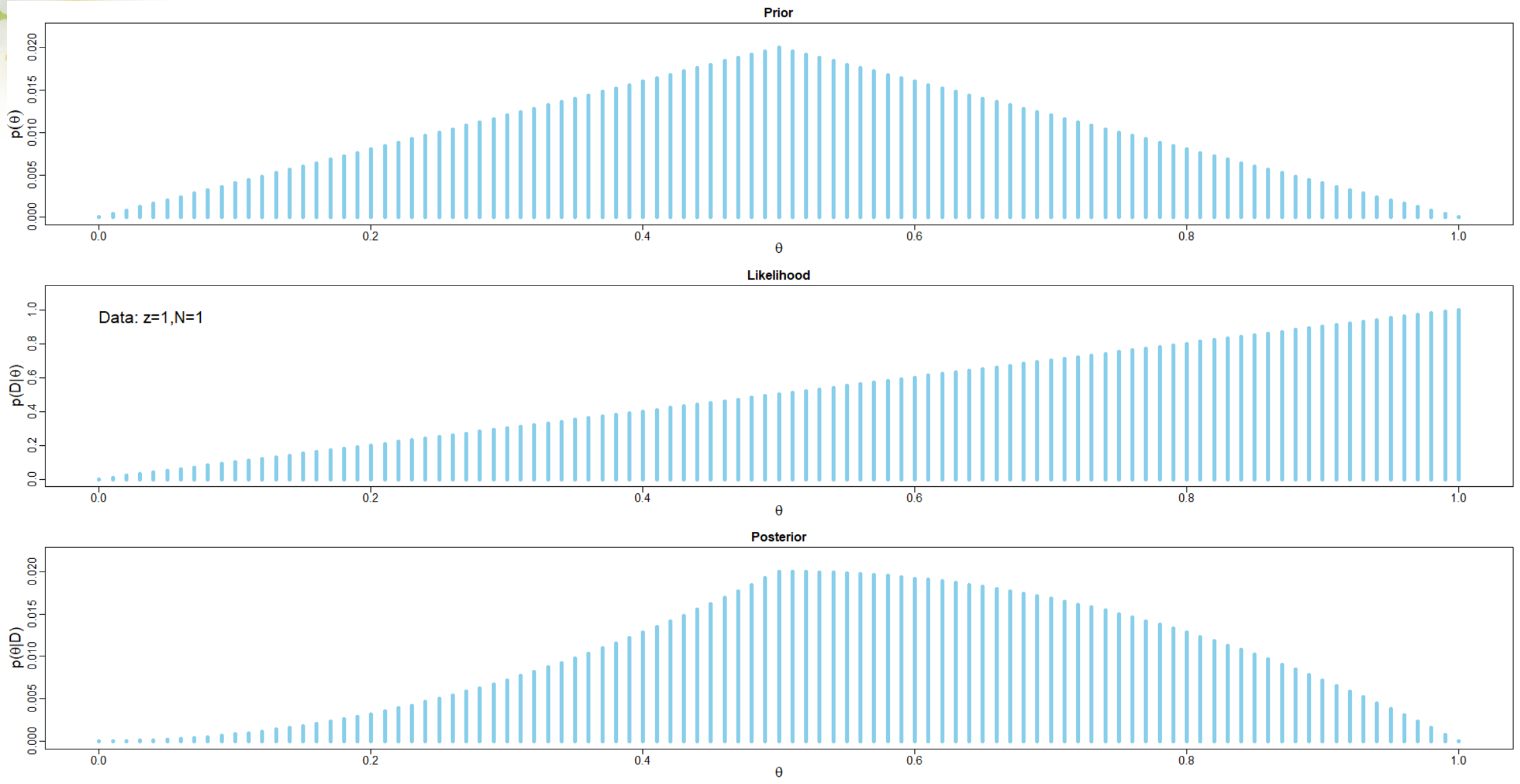
$$\int_0^{1/2} 4\theta \cdot \theta d\theta + \int_{\frac{1}{2}}^{1} 4\theta \cdot (1-\theta) d\theta = 1/2$$

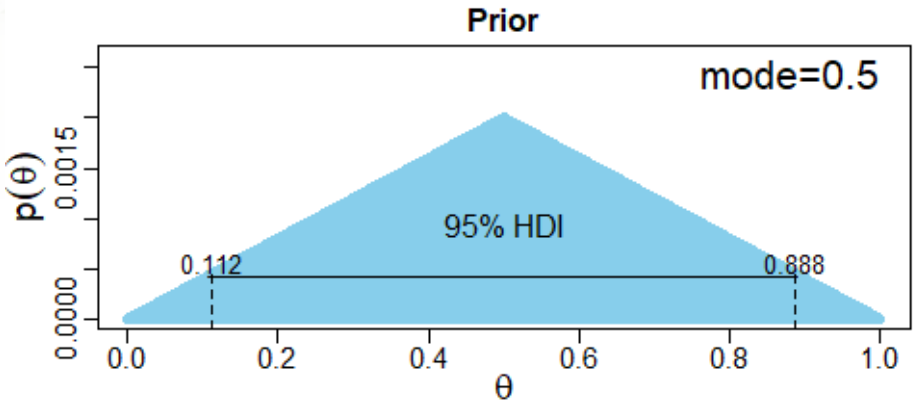
- 后验分布的解析解为

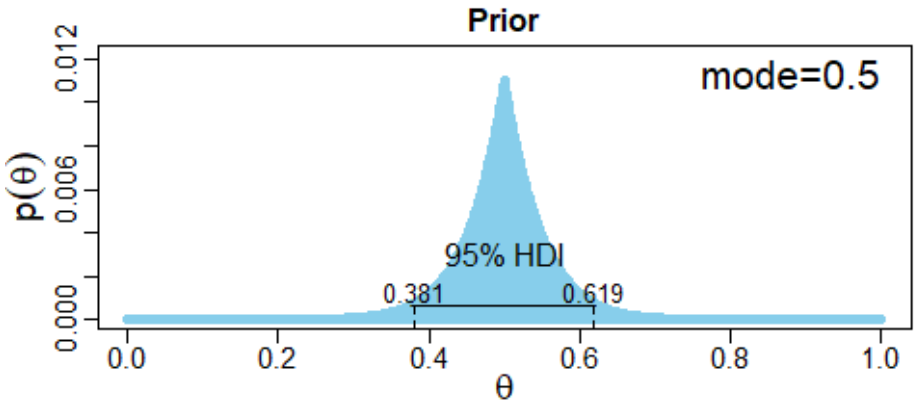
$$p(\theta|D) = 8\theta \times \min(\theta, 1-\theta)$$

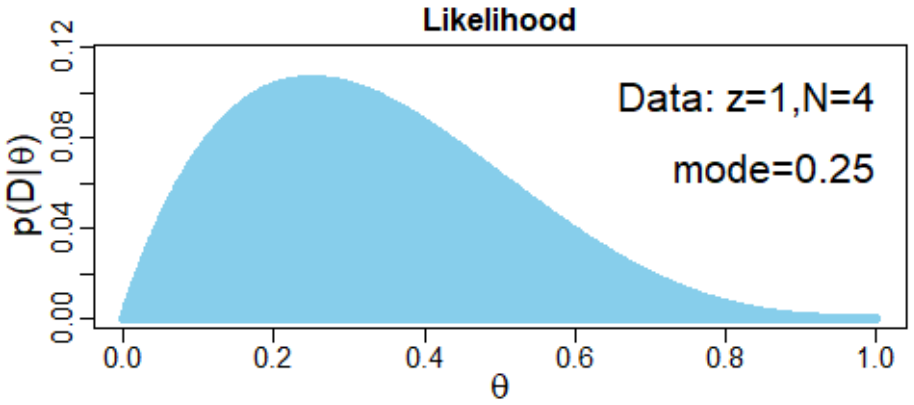

p(θ)
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
θ
p(D|θ)
1
0.5
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
θ
p(θ|D)
2
1
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
θ

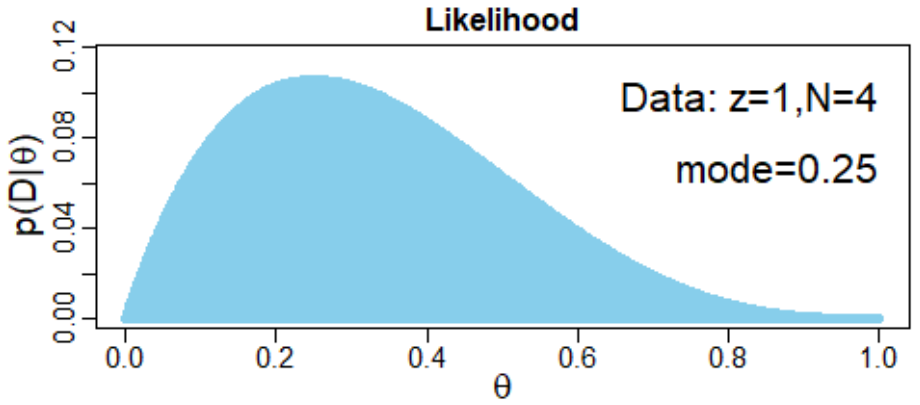

Prior
p(θ)
0.00
0.10
0.20
0.0
0.2
0.4
0.6
0.8
1.0
θ
Likelihood
Data: z=1,N=1
p(D|θ)
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
θ
Posterior
p(θ|D)
0.00
0.10
0.20
0.0
0.2
0.4
0.6
0.8
1.0
θ

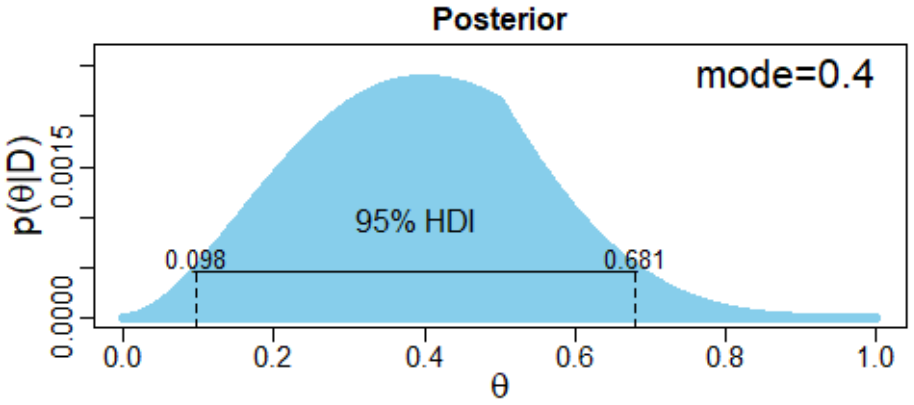

Prior
p(θ)
0.000
0.005
0.010
0.015
0.020
0.0
0.2
0.4
0.6
0.8
1.0
θ
Likelihood
Data: z=1,N=1
p(D|θ)
0.0
0.2
0.4
0.6
0.8
1.0
θ
Posterior
p(θ|D)
0.000
0.005
0.010
0.015
0.020
0.0
0.2
0.4
0.6
0.8
1.0
θ

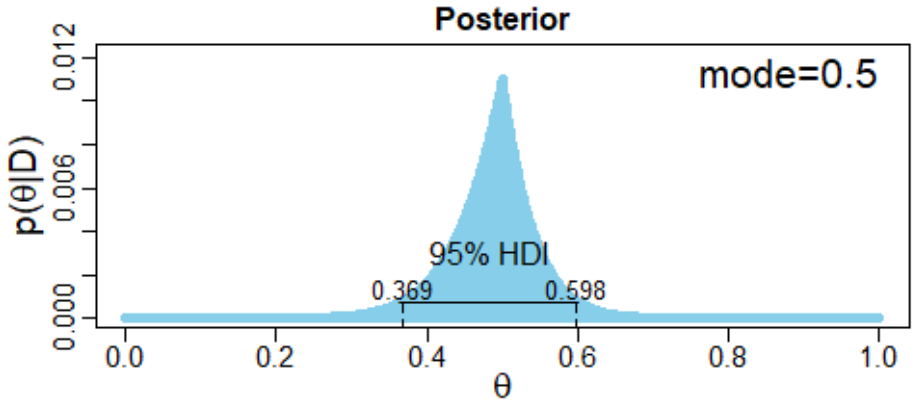

Prior
mode=0.5
95% HDI
0.112
0.888
p(θ)
θ

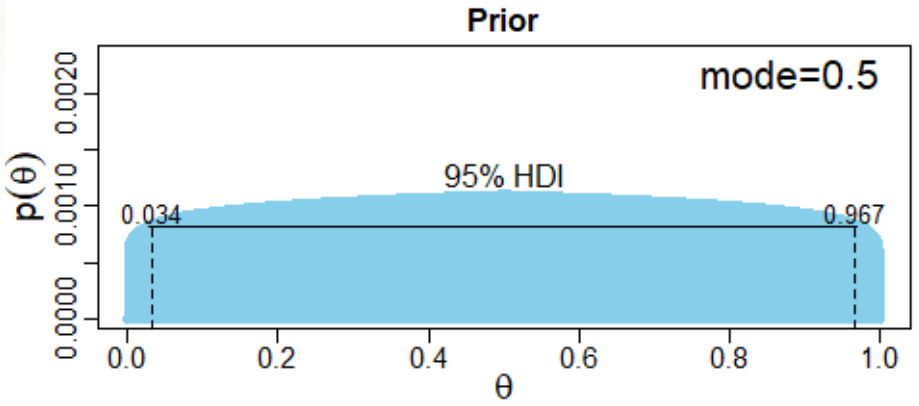

Prior
mode=0.5
95% HDI
0.381
0.619
p(θ)
θ

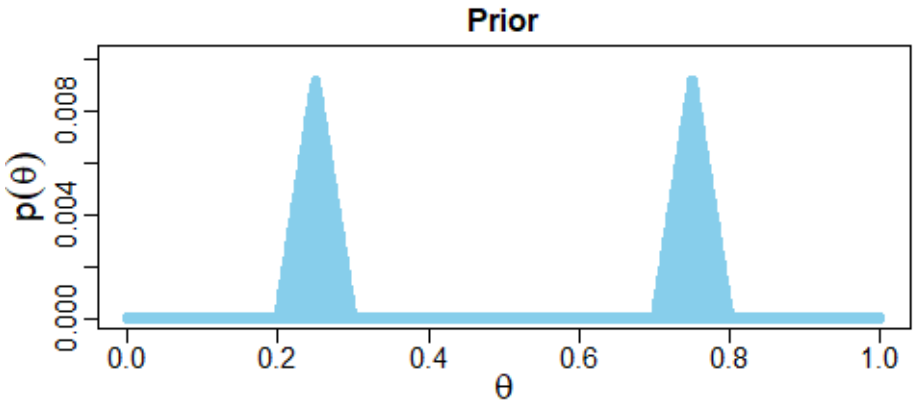

Likelihood
Data: z=1,N=4
mode=0.25
p(D|θ)
θ

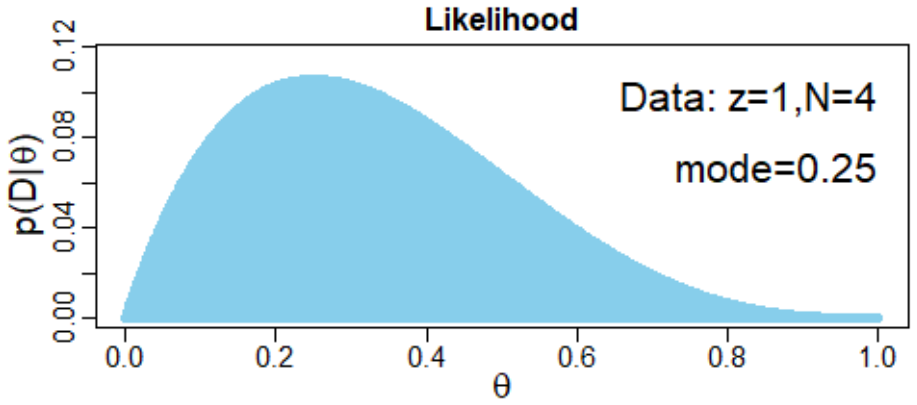

Likelihood
Data: z=1,N=4
mode=0.25
p(D|θ)
θ

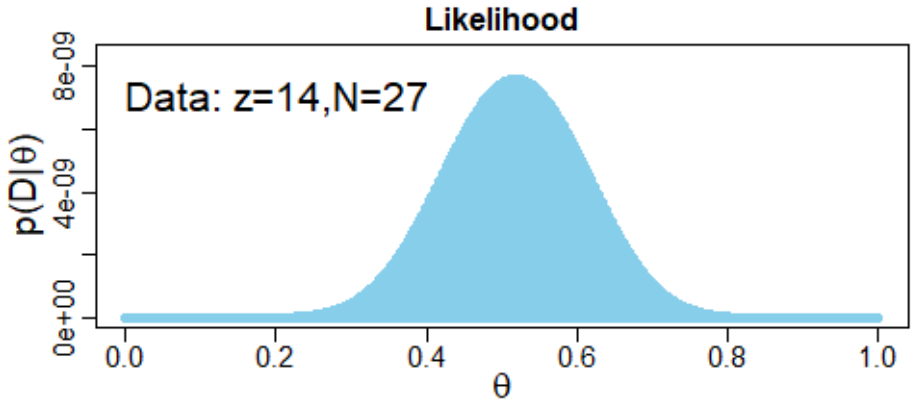

Posterior
mode=0.4
95% HDI
0.098
0.681
p(θ|D)
θ

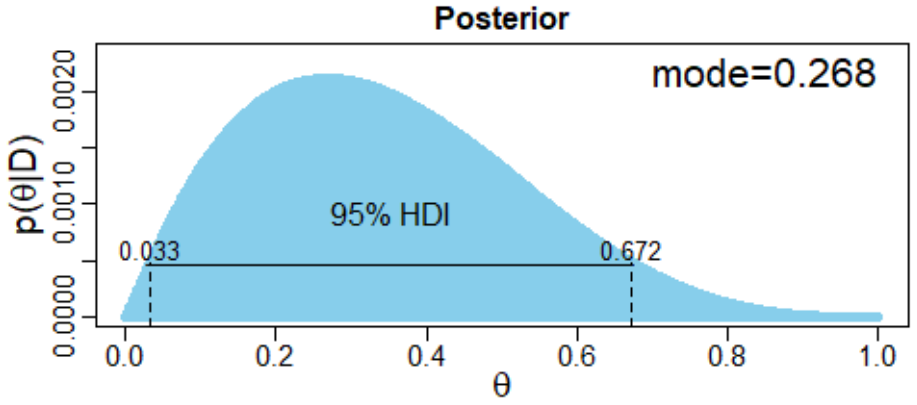

Posterior
mode=0.5
95% HDI
0.369
0.598
p(θ|D)
θ

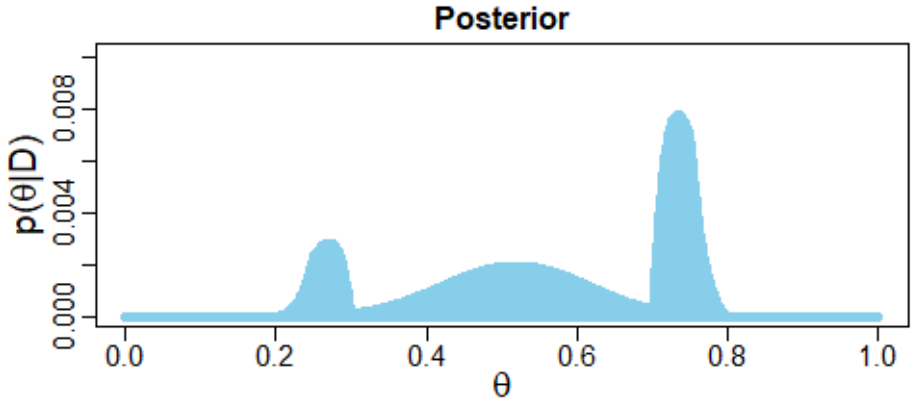

Prior
mode=0.5
95% HDI
0.034
0.967
p(θ)
θ


Prior
p(θ)
θ


Likelihood
Data: z=1,N=4
mode=0.25
p(D|θ)
θ


Likelihood
Data: z=14,N=27
p(D|θ)
θ


Posterior
mode=0.268
95% HDI
0.033
0.672
p(θ|D)
θ


Posterior
p(θ|D)
θ

马尔科夫链蒙特卡洛(MCMC)

- 对于高维问题，上述格点近似法一般难以操作，这主要是因为为了达到一定的拟合精度所需要的格点数随着维度的上升指数上升。
- 例如，如果给每个维度（参数）设置100个格点，那么当维数（参数个数）为3时，需要100万个格点，维数为4时，需要1亿个格点，维数为5时，需要100亿个格点。
- 此时，需要使用更为有效的马尔科夫链蒙特卡洛(Markov chain Monte Carlo)方法得到后验分布的近似解，以及相关指标的近似解

马尔可夫链简介

- 马尔可夫链是一种随机过程
- 它的基本特征是过程中下一刻的状态，由当前所处的状态依特定的概率分布决定，与之前所处的状态无关。
- 例如，一维随机游走(random walk)就是一种马尔科夫链

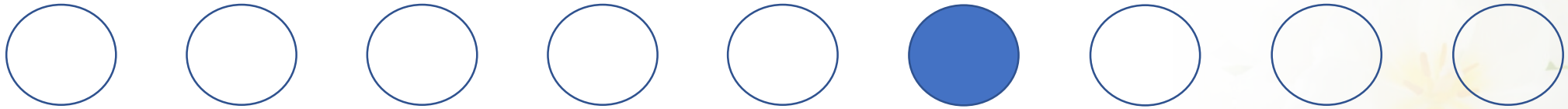
随机游走示例

- 一维随机游走是一种离散状态离散时间的随机过程
- 状态为0, -1, 1, -2, 2, -3, 3……
- 每一次移动所需时间是固定的，且状态变化为1
- Pr(L) = p, Pr(R) = q = 1-p
- 所以，如果当前状态为s, 那么下一步的状态只能是s-1或者s+1,且前者的概率为p, 后者的概率为q，无论当前时刻之前的状态如何

假定p = q = 0.5, S(0) = 0

S(0) = 0

假定p = q = 0.5, S(0) = 0

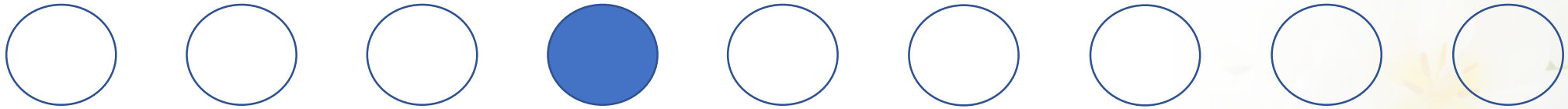


S(0) = 0
S(1) = 1

假定p = q = 0.5, S(0) = 0

S(0) = 0
S(1) = 1
S(2) = 0

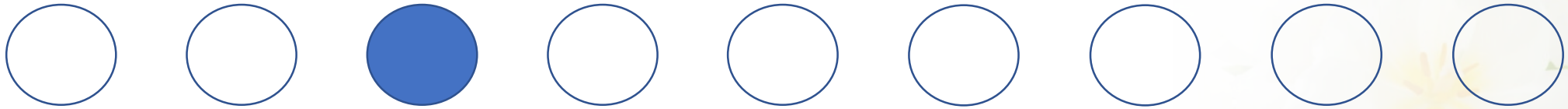
假定p = q = 0.5, S(0) = 0



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

S(0) = 0
S(1) = 1
S(2) = 0
S(3) = -1

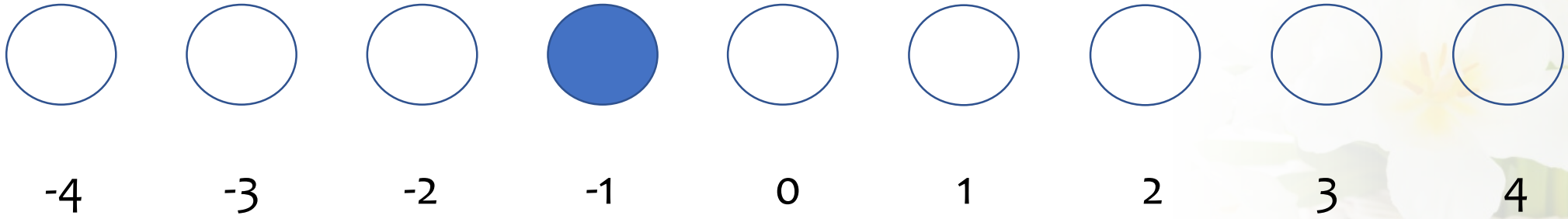
假定p = q = 0.5, S(0) = 0



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

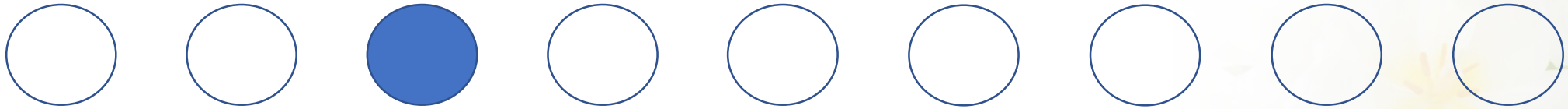
S(0) = 0
S(1) = 1
S(2) = 0
S(3) = -1
S(4) = -2

假定p = q = 0.5, S(0) = 0



S(0) = 0
S(1) = 1
S(2) = 0
S(3) = -1
S(4) = -2
S(5) = -1

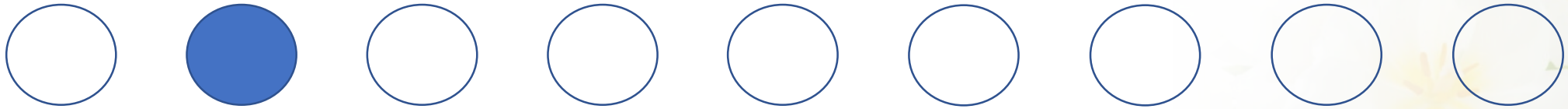
假定p = q = 0.5, S(0) = 0



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

S(0) = 0
S(1) = 1
S(2) = 0
S(3) = -1
S(4) = -2
S(5) = -1
S(6) = -2

假定p = q = 0.5, S(0) = 0



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

S(0) = 0
S(1) = 1
S(2) = 0
S(3) = -1
S(4) = -2
S(5) = -1

S(6) = -2
S(7) = -3
……

马尔可夫链的性质

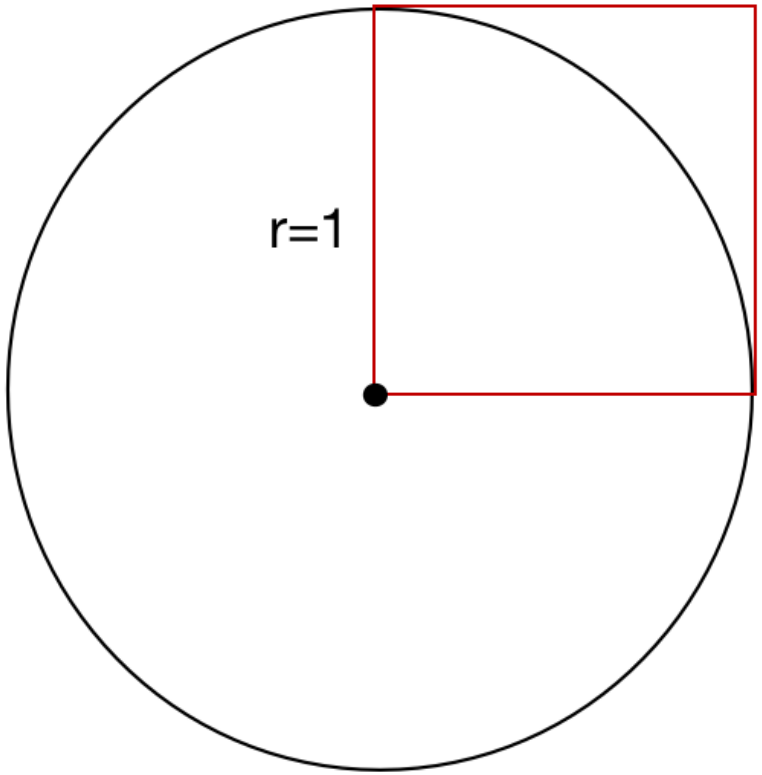
- 对于某些特定的马尔可夫链，数学上可以证明，随着移动次数的增加，随机过程处于各个状态的概率或者概率密度将趋于稳定。这样的稳定分布，称为马尔可夫链的稳态（或者平稳）分布。
- 这样的马尔科夫链，需要满足平稳性、正常返性和非周期性

蒙特卡洛方法简介

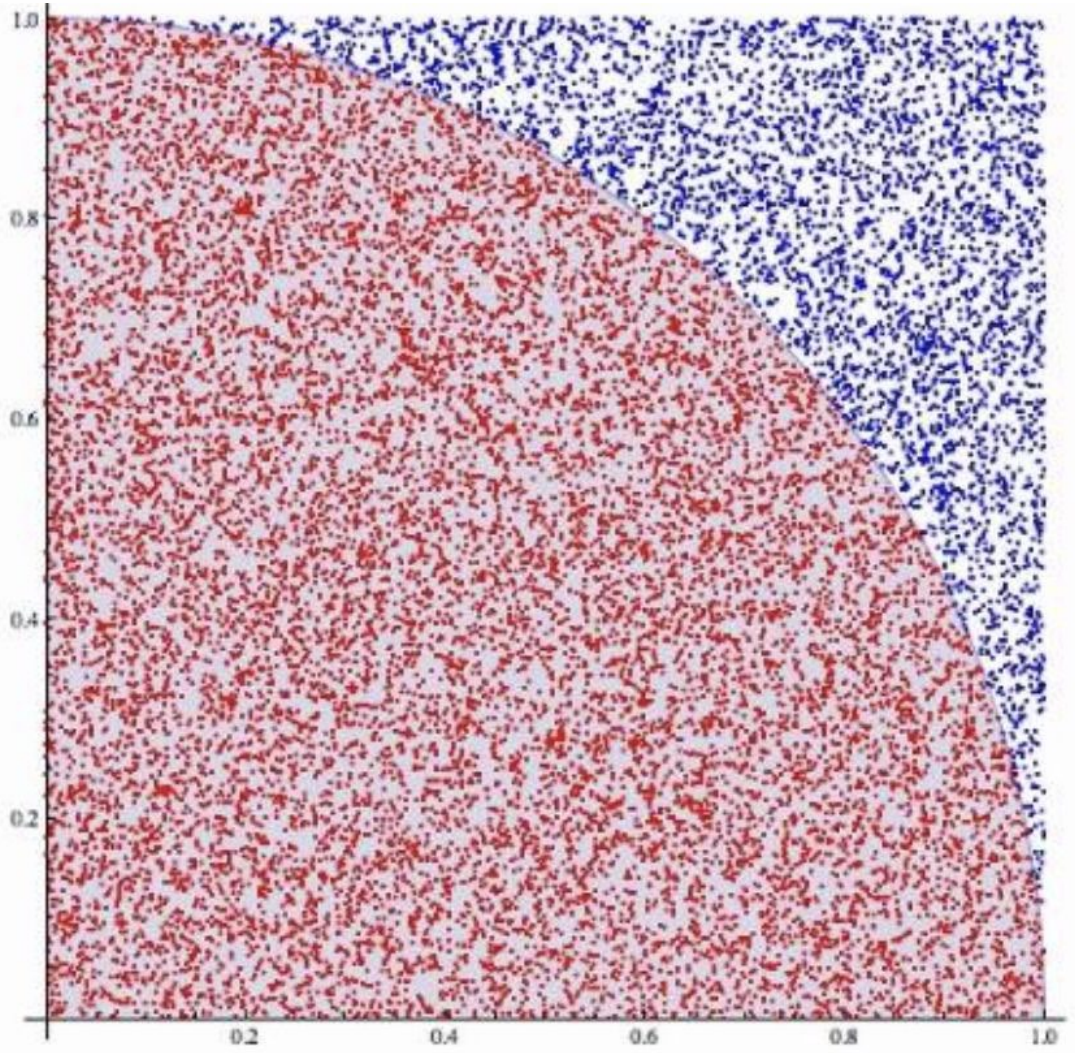
- 定义：通过使用随机数来解决计算问题的一种统计模拟方法
- 工作原理：通过不断抽样逐渐逼近目标

示例

- 圆周率的求解



示例




```
import random
total = [10, 100, 1000, 10000, 100000, 1000000, 5000000]  #随机点数
for t in total:
    in_count = 0
    for i in range(t):
        x = random.random()
        y = random.random()
        dis = (x**2 + y**2)**0.5
        if dis<=1:
            in_count += 1
    print(t,'个随机点时，π是：', 4*in_count/t)
```

```
10 个随机点时，π是： 3.2
100 个随机点时，π是： 3.04
1000 个随机点时，π是： 3.108
10000 个随机点时，π是： 3.136
100000 个随机点时，π是： 3.13916
1000000 个随机点时，π是： 3.140868
5000000 个随机点时，π是： 3.1423208
```

MCMC基本思路

- 在无法得出后验分布解析解的情况下，首先建构特定的马尔可夫链(Markov chain)，使得它的稳态分布和后验分布相一致
- 然后，使用蒙特卡洛(MC)方法和符合上述要求的马尔可夫链，产生后验分布的随机近似样本，并求得贝叶斯数据分析所需要的各种指标

MCMC涉及的基本概念

- 由于马尔可夫链进入稳态分布需要时间，所以最初生成的样本不能作为对应后验分布的近似解的一部分，需要舍弃，这在贝叶斯数据分析中一般称为burn-in(预烧)或者warmup(预热)阶段。
- 另外，由于马尔可夫链相邻时间点所处状态的相关性，即使进入了稳态分布阶段，所生成的样本点也不是独立的，而是存在时间序列上的自相关关系。可以通过设置thin-in(稀释)操作，取时间间隔较大的样本点来构建后验分布的近似解。

MCMC基本操作

- 实际计算过程中，需要合理设置burn-in阶段的长度、thin-in的程度，从而生成有代表性的后验分布近似样本，并且需要对代表性进行检验。
- 检验的指标通常为Gelman-Rubin统计量，也称为Brooks-Gelman-Rubin统计量、潜在尺度缩减因子(potential scale reduction factor)或者收缩因子(shrink factor), 一般记作R-hat。

R-hat简介

- R-hat的基本思路是通过使用不同的随机过程起始点，考察由此产生的随机样本（随机链）的同质性。
- 如果设置合理，每条随机链都进入了稳态分布，那么它们相互间同质性较高，R-hat接近于1。
- R-hat的计算方式类似于方差分析中的F值，每条随机链相当于一组数据，当组间差异相对于组内差异较小时，说明不同组之间同质性较高，随机过程已经进入了稳态分布。

常用MCMC方法

- Metropolis算法 (BUGS, JAGS)
- Gibbs算法（BUGS, JAGS）
- Hamiltonian MC方法（Stan）
- 技术细节参见DBDA,Ch 7, 8, 14