第3章 极大似然估计与贝叶斯估计

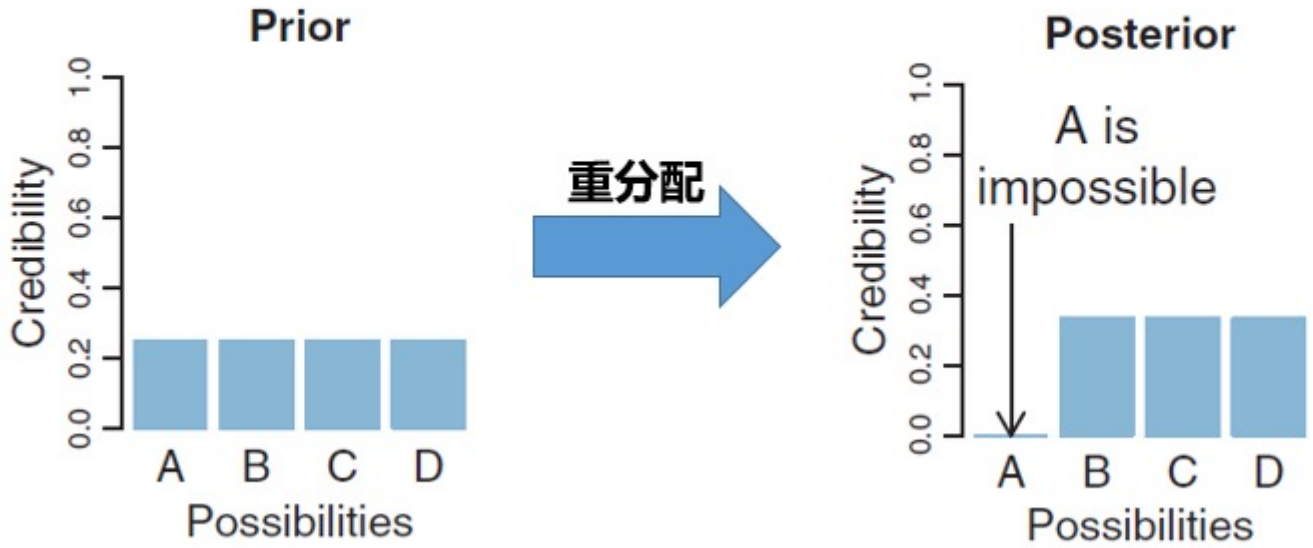
中山大学人工智能学院
毛旭东

Email：maoxd3@mail.sysu.edu.cn

回顾

- **贝叶斯推理是重新分配可信度的过程：**

福尔摩斯通过观察，收集证据，来排除一些不可能的情况。



- 贝叶斯推理是用**贝叶斯推理**来重新分配可信度。

统计学派

- 统计学主要有两个学派：

1. 贝叶斯学派（Bayesian）
2. 频率学派（Frequentist）

Thomas Bayes
托马斯·贝叶斯

Ronald Fisher
罗纳德·费希尔

历史

- 贝叶斯法则是用数学家贝叶斯（Thomas Bayes，1701 – 1761）的名字命名的。
- 1763年，发表了第一篇用贝叶斯定理来估计概率分布的参数的论文。
  - 由贝叶斯的好友Richard Price整理贝叶斯生前的工作。
- 1774年，Pierre-Simon Laplace扩展、完善了贝叶斯推理方法。
  - 可能是独立的工作，不知道贝叶斯的工作。

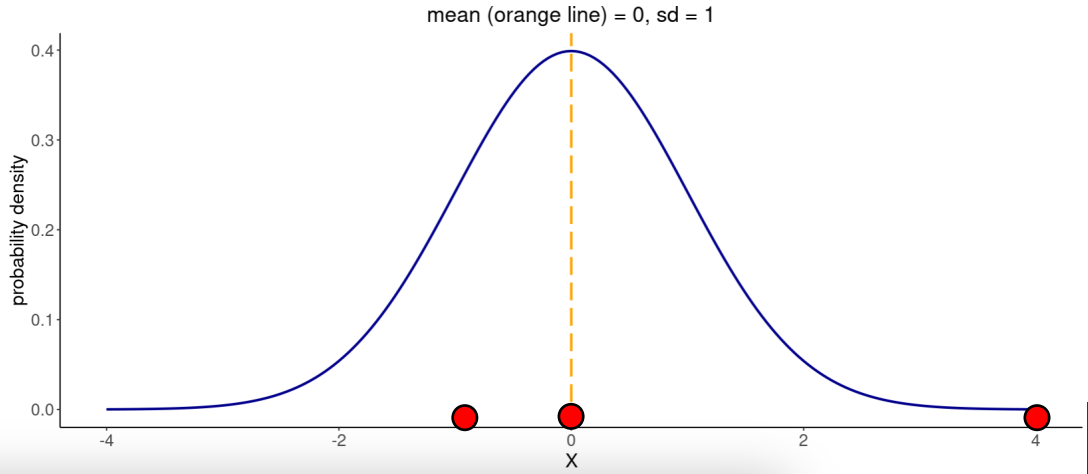
历史

- 尽管贝叶斯推理在1763年就被提出，由Ronald Fisher (1890-1962)发展及提倡的另一个统计学派，频率学派（Frequentist），在20世纪成为主流。
- 直到21世纪，贝叶斯学派才成为统计学的主流。
  - 原因包括:
  - 贝叶斯推理在计算上有困难，而且早期计算机的计算能力有限。
  - 20世纪末，MCMC近似计算方法被提出，并且计算机的计算能力提高。
  - 机器学习学者提出并发展了变分推理（variational inference）近似计算方法。

似然函数

- $L(\theta) = p(D|\theta)$，称为似然函数（likelihood function），表示在参数θ下观测到数据样本D的可能性（概率密度或概率质量）。
- 例子： 假设$p(D|\theta)$服从高斯分布$N(0,1)$，$D = \{-1,0,4\}$。
- 记$N(0,1)$的概率密度函数为$g(x)$，并且假设数据独立同分布（independent and identically distributed）则：

$$L(\theta) = g(-1)\times g(0)\times g(4) = 1.3\times 10^{-5}$$

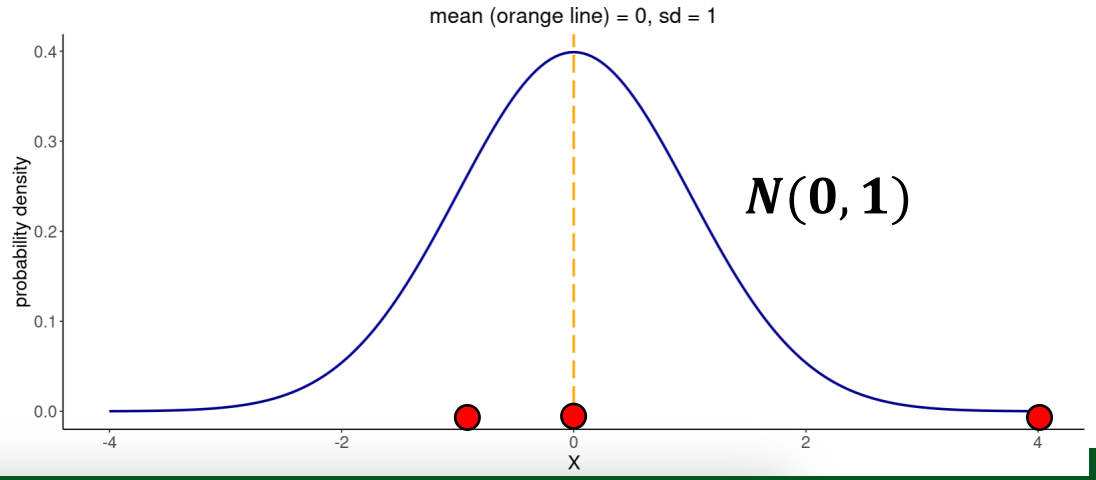


极大似然估计（Maximum Likelihood Estimation，MLE）

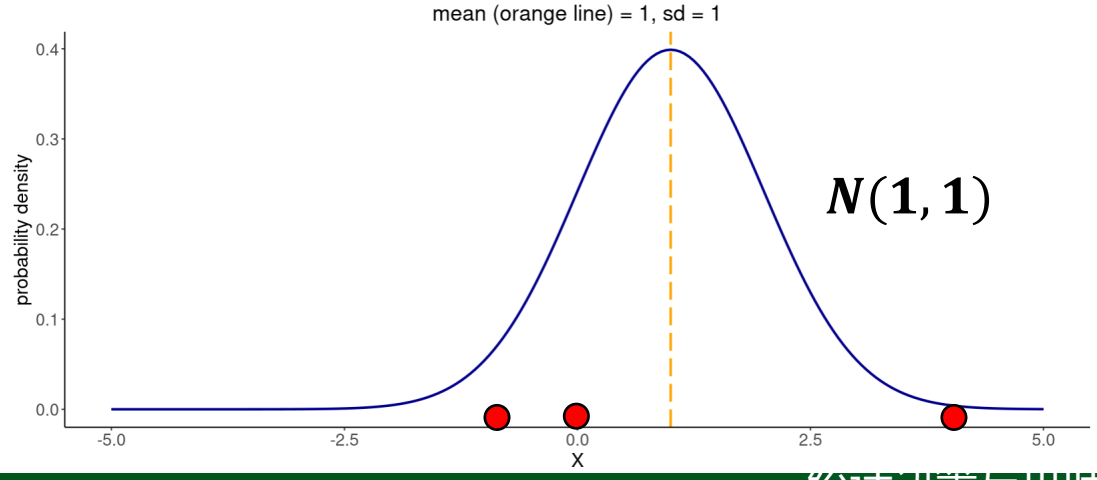
- 极大似然估计是一种参数估计量，它的核心思想是选择参数值，使得观测到的数据出现的概率（即似然）最大：

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(D|\theta)$$

$L(\mu = 0) = g_1(-1)\times g_1(0)\times g_1(4) = 1.3\times10^{-5}$



$L(\mu = 1) = g_2(-1)\times g_2(0)\times g_2(4) = 5.8\times10^{-5}$



5.3 例子：抛硬币

- 问题：给定一个硬币，估计这个硬币的正面偏向性。

- 样本空间包括{正面，反面}，我们用$y = 1$表示结果是正面，$y = 0$表示结果是反面。

- 收集一组抛硬币结果的数据，$\{y_i\}$。

确定数据的模型（似然）和相应的参数

- $y = 1$表示结果是正面，$y = 0$表示结果是反面。
- 用θ作为参数，表示硬币的正面偏向性。可得：

$$p(y = 1|\theta) = \theta$$

$$p(y = 0|\theta) = 1 - \theta$$

- 将上面2个式子合并为：

$$p(y|\theta) = \theta^{y}(1 - \theta)^{1-y}$$

- 该概率分布称为**伯努利分布（Bernoulli distribution）**，以数学家Jacob Bernoulli（1655-1705）命名。
- $p(y|\theta)$是模型（似然），θ是参数。

抛硬币：求解似然函数

- 我们考虑多次抛硬币的情况，假设抛N次的结果为$\{y_i\}$，其中正面的次数为$z = \sum_i y_i$，反面的次数为$N - z$。假设每次抛硬币之间都是相互独立的，则模型为：

$$p(\{y_i\}|\theta) = \prod_i p(y_i|\theta)$$

抛硬币：求解似然函数

- 我们考虑多次抛硬币的情况，假设抛N次的结果为$\{y_i\}$，其中正面的次数为$z = \sum_i y_i$，反面的次数为$N - z$。假设每次抛硬币之间都是相互独立的，则模型为：

$$p(\{y_i\}|\theta) = \prod_i p(y_i|\theta)$$

$$= \prod \theta^{y_i}(1-\theta)^{1-y_i}$$

抛硬币：求解似然函数

- 我们考虑多次抛硬币的情况，假设抛N次的结果为$\{y_i\}$，其中正面的次数为$z = \sum_i y_i$，反面的次数为$N - z$。假设每次抛硬币之间都是相互独立的，则模型为：

$$
\begin{aligned}
p(\{y_i\}|\theta) &= \prod_i p(y_i|\theta) \\
&= \prod_i \theta^{y_i}(1-\theta)^{1-y_i} \\
&= \theta^{\sum_i y_i}(1-\theta)^{\sum_i (1-y_i)} \\
&= \theta^{z}(1-\theta)^{N-z}
\end{aligned}
$$

似然函数

$$L(\theta) = p(\{y_i\}|\theta) = \theta^z(1-\theta)^{(N-z)}$$

- 当θ固定，y作为变量时，上式是y的一个概率分布。
- 另一个角度是：y固定，θ作为变量。此时，我们将上式称为**θ的似然函数**（likelihood function）。

注意：

- 似然函数不是概率分布，而是函数，因此似然函数的积分不是1。

极大似然估计（Maximum likelihood estimation，MLE）

- 极大似然估计经常转换成如下形式，以方便求解。

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, p(D|\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \log p(D|\theta) \\ &= \underset{\theta}{\operatorname{argmin}} -\log p(D|\theta) \\ &= \underset{\theta}{\operatorname{argmin}} -\sum_i \log p(\boldsymbol{x}^{(i)}|\theta)\end{aligned}$$

- $\log p(D|\theta)$称为Log Likelihood Function，记为$LL(\theta)$。
- 极大似然估计是一种“点估计”（point estimation）。

MLE：伯努利分布

$$p(\{y_i\}|\theta) = \theta^z(1-\theta)^{N-z}$$

其中，$\theta \in [0,1]$。

- 我们求解$LL(\theta) = z\log\theta + (N-z)\log(1-\theta)$的极值点：

$$LL'(\theta) = \frac{z}{\theta} - \frac{N-z}{1-\theta} = 0$$

可得：$\theta = \frac{z}{N}$

另一方面，$LL''(\theta) = -\frac{z}{\theta^2} - \frac{N-z}{(1-\theta)^2} < 0$

- 因此，$\theta = \frac{z}{N}$，似然取得最大值。

频率学派

- 频率学派概率（Frequentist probability）：将一个事件的“概率”定义为在多次试验中其相对频率的极限值。
- 频率学派推理（Frequentist inference）是基于重复试验。
  - 假设每一次实验的数据集采样自“真实概率分布”。

- 假设有一个估计量（estimator），$\hat{\theta} = \pi(D)$，当采样的数据集变了，参数值就会变。
  - 估计量可以是极大似然估计、极大后验估计等。

频率学派推理

- 频率学派用估计量（estimator）的**抽样分布（sampling distribution）来表示参数的不确定性。**
  - 比如，抽样100次，得到100个参数的估计值，组成一个抽样分布。

- 假设有估计量$\hat{\theta} = \pi(D)$;
- 假设有S个数据集$D^{(s)}$，采样自真实数据分布$p(x|\theta^*)$:

$$D^{(s)} = \{x_i \sim p(x|\theta^*)\}$$

- 则抽样分布为:

$$p(\hat{\theta} = \pi(D^{(s)}))$$

频率学派推理

- 抽样分布：

$$p(\hat{\theta} = \pi(D^{(s)}))$$

- 最常用的估计量是**极大似然估计**。
- 注意：很多资料中把极大似然估计等点估计方法等同于频率学派方法，这是错误的。

- 极大似然估计可以看作是频率学派方法的近似（抽样了1次）。

贝叶斯学派

- 贝叶斯概率（Bayesian probability）：将“概率”定义为对事件发生的信念程度。
- 贝叶斯推理（Bayesian inference）首先假设事件发生的先验概率，然后根据数据来更新概率。
- 贝叶斯学派将参数作为随机变量，并用这个随机变量的**概率分布来建模不确定性**。
- 利用贝叶斯法则，来求解后验概率分布$p(\theta|D)$:

$$p(\theta|D) = \frac{p(D|\theta)\, p(\theta)}{p(D)}$$

5.2 贝叶斯推理

- 模型（model）：用于对数据（data）建模。给定模型结构和参数值，输出数据的概率值。

$$p(\text{数据}|\text{参数})$$

- 我们的目标是估计参数的值，也就是：

$$p(\text{参数}|\text{数据})$$

- 可以用贝叶斯法则来转换。

贝叶斯法则用于模型参数和数据

- 用θ表示参数，用D表示数据。
- 模型（似然）：$\boldsymbol{p}$(数据|参数) $= p(D|\theta)$
- 目标：$\boldsymbol{p}$(参数|数据) $= p(\theta|D)$
- 利用贝叶斯法则:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

贝叶斯法则用于模型参数和数据

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- $p(\theta)$：先验（prior）
- $p(\theta|D)$：后验（posterior）
- $p(D|\theta)$：似然（likelihood）
- $p(D)$：证据（evidence），也称为边缘似然（marginal likelihood）
  - $\sum_{\theta^*} p(D|\theta^*)\, p(\theta^*)$ 或者 $\int p(D|\theta)\, p(\theta)\, d\theta$
- 注意：我们关心的是θ，θ是变量，D是给定的。先验的“先”，后验的“后”都是对于θ而言。

5.3 例子：抛硬币

- 问题：给定一个硬币，估计这个硬币的正面偏向性。
- 收集一组抛硬币结果的数据，$\{y_i\}$。

- 似然函数：

$$L(\theta) = p(\{y_i\}|\theta) = \theta^z(1-\theta)^{(N-z)}$$

求解后验概率分布

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

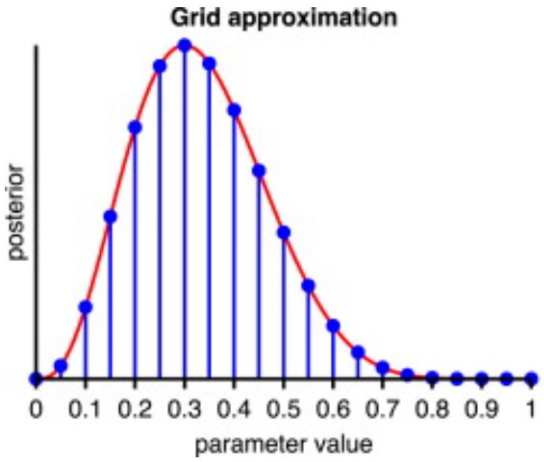
对于抛硬币问题：

- 似然：$p(D|\theta) = \theta^z(1-\theta)^{(N-z)}$
- 先验：$p(\theta)$可以假设。比如假设为均匀分布，$p(\theta) = 1$。
- 证据：$p(D) = \int p(D|\theta)p(\theta)d\theta$，不好算。
- 这里，我们先来看一种近似计算的方法，网格近似方法。

贝叶斯推理方法：网格近似（Grid Approximation）

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}$$

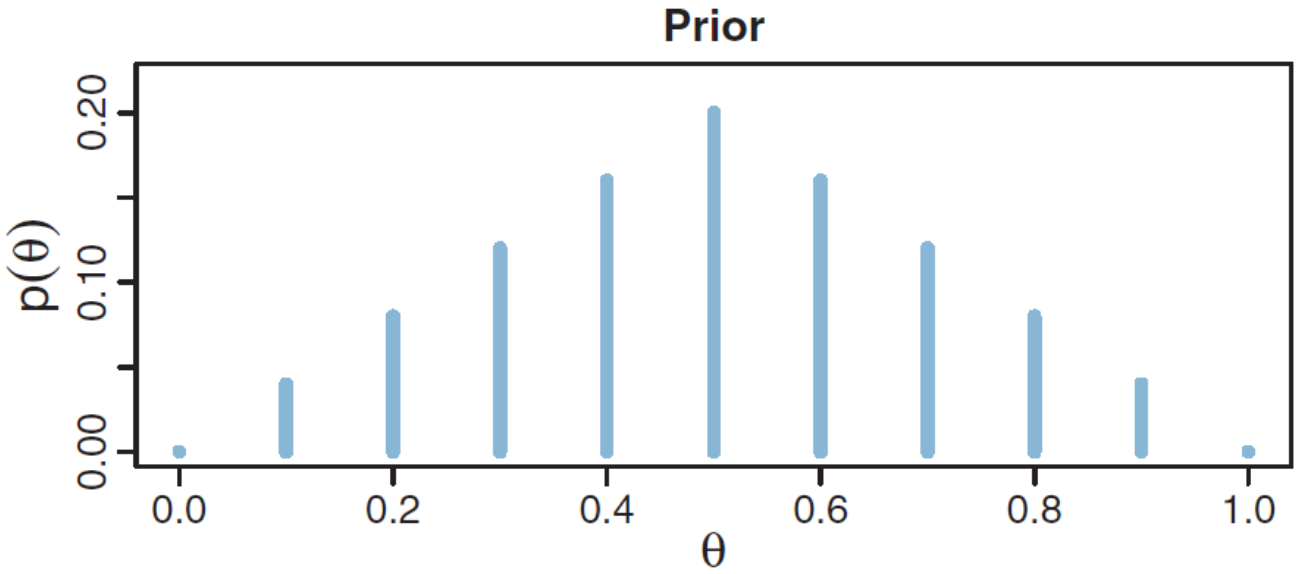
$$\approx \frac{p(D|\theta)p(\theta)}{\sum_{\theta^*} p(D|\theta^*)p(\theta^*)}$$



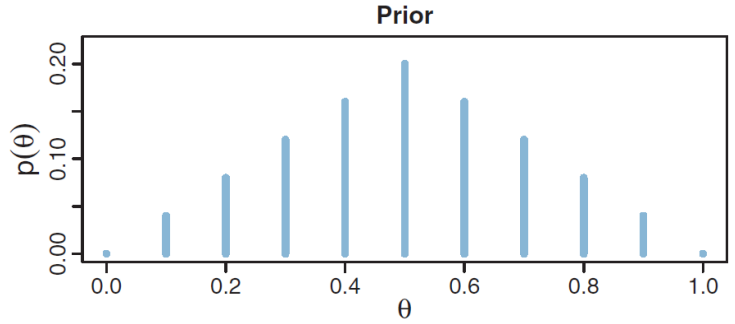
- “积分”用“求和”来近似。
- 对于连续值$\theta \in [0,1]$，**网格近似**用离散值来近似，比如$\{0.0, 0.1, 0.2, \ldots 0.9, 1.0\}$。
- 计算$p(\theta = 0.0|D)$，$p(\theta = 0.1|D)$，…，$p(\theta = 1.0|D)$，来近似后验概率分布。

假设先验概率分布

- 我们用网格近似来求解，假设θ可能的取值是$\{0.0, 0.1, 0.2, \ldots 0.9, 1.0\}$。
- 假设我们觉得$\theta = 0.5$是最有可能的，因此指定如下的先验概率分布$p(\theta)$:



假设先验概率分布



- 假设0.0至0.5的$P(\theta)$值为$(0,0.2,0.4,0.6,0.8,1)$。
- 0.6至1.0这一段是对称的，$P(\theta)$值为$(0.8,0.6,0.4,0.2,0)$。
- $P(\theta) = (0,0.2,0.4,0.6,0.8,1,0.8,0.6,0.4,0.2,0)$。
- **归一化：**

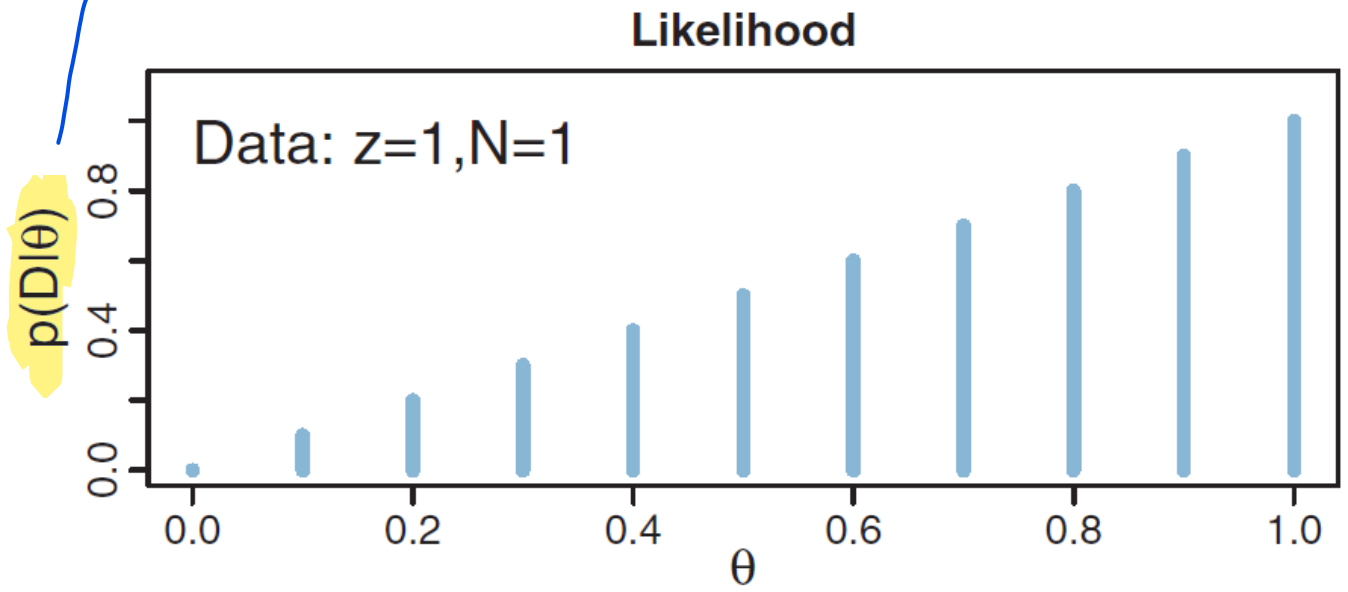
$$p(\theta_i) = \frac{P(\theta_i)}{\sum_i P(\theta_i)}$$

- $p(\theta) = (0,0.04,0.06,0.08,0.12,0.14,0.16,0.2,\ldots,0.04,0)$

计算后验概率分布

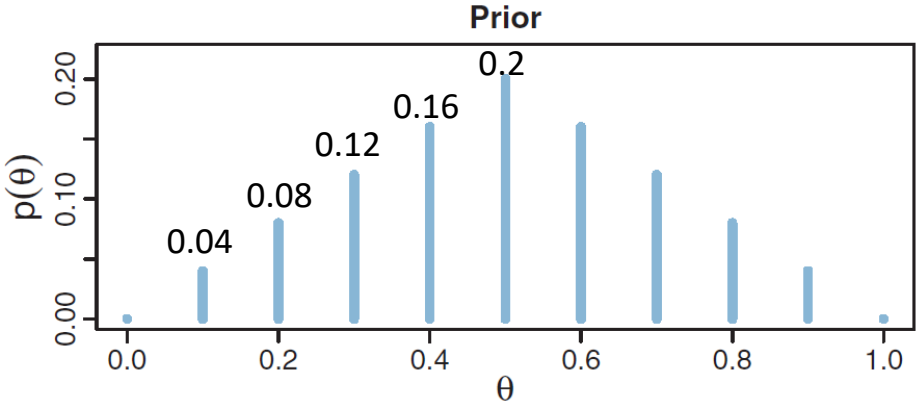
- 假设抛了1次硬币，结果是正面，即$N = 1, z = 1$。
- 根据$p(\{y_i\}|\theta) = \theta^z(1-\theta)^{N-z}$，得到似然函数为：

$$L(\theta) = \theta$$

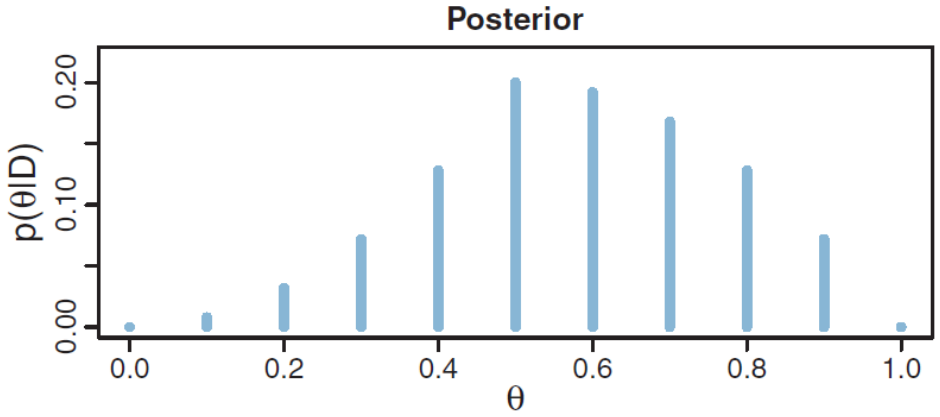


计算后验概率分布

- 根据先验$p(\theta)$和似然$p(D|\theta)$，我们计算所有θ的后验概率$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{p(D)}$。
- 先计算$p(D)=\sum_{\theta^*}p(D|\theta^*)p(\theta^*)=0.5$
- $p(\theta=0.2|D)=\frac{0.2*0.08}{0.5}=0.032$
- $p(\theta=0.5|D)=\frac{0.5*0.2}{0.5}=0.2$
- ....

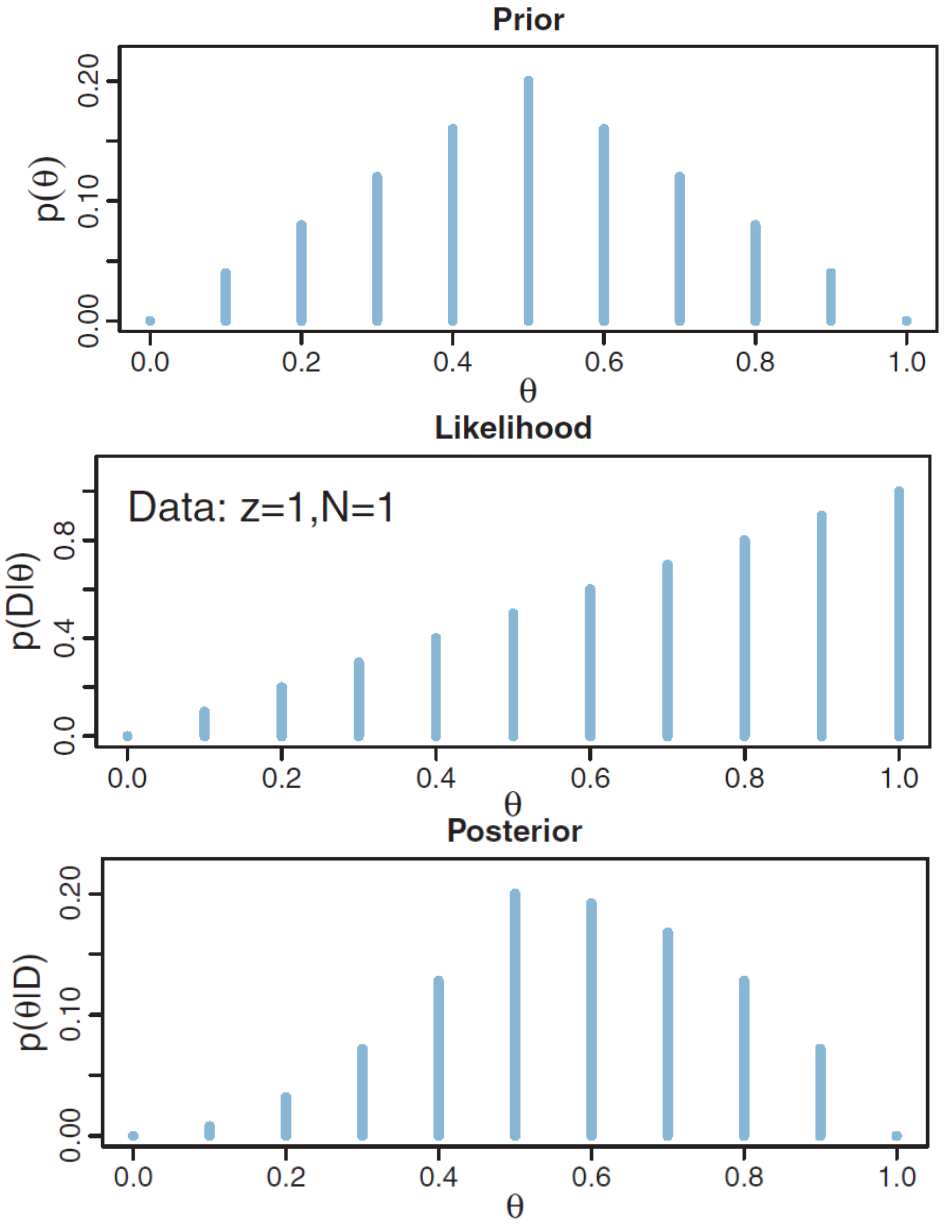


- 根据先验$p(\theta)$和似然$p(D|\theta)$，我们计算所有θ的后验概率$p(\theta|D)=\frac{p(D|\theta)p(\theta)}{p(D)}$。
- 先计算$p(D)=\sum_{\theta^*}p(D|\theta^*)p(\theta^*)=0.5$
- $p(\theta=0.2|D)=\frac{0.2*0.08}{0.5}=0.032$
- $p(\theta=0.5|D)=\frac{0.5*0.2}{0.5}=0.2$
- ....



后验、先验、似然的关系

- 后验和先验更接近。
  - 因为只有1次抛硬币。
- 似然对后验有影响。
  - 后验的右半边概率值更大。
- 后验是先验和似然之间的妥协。



伪代码

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

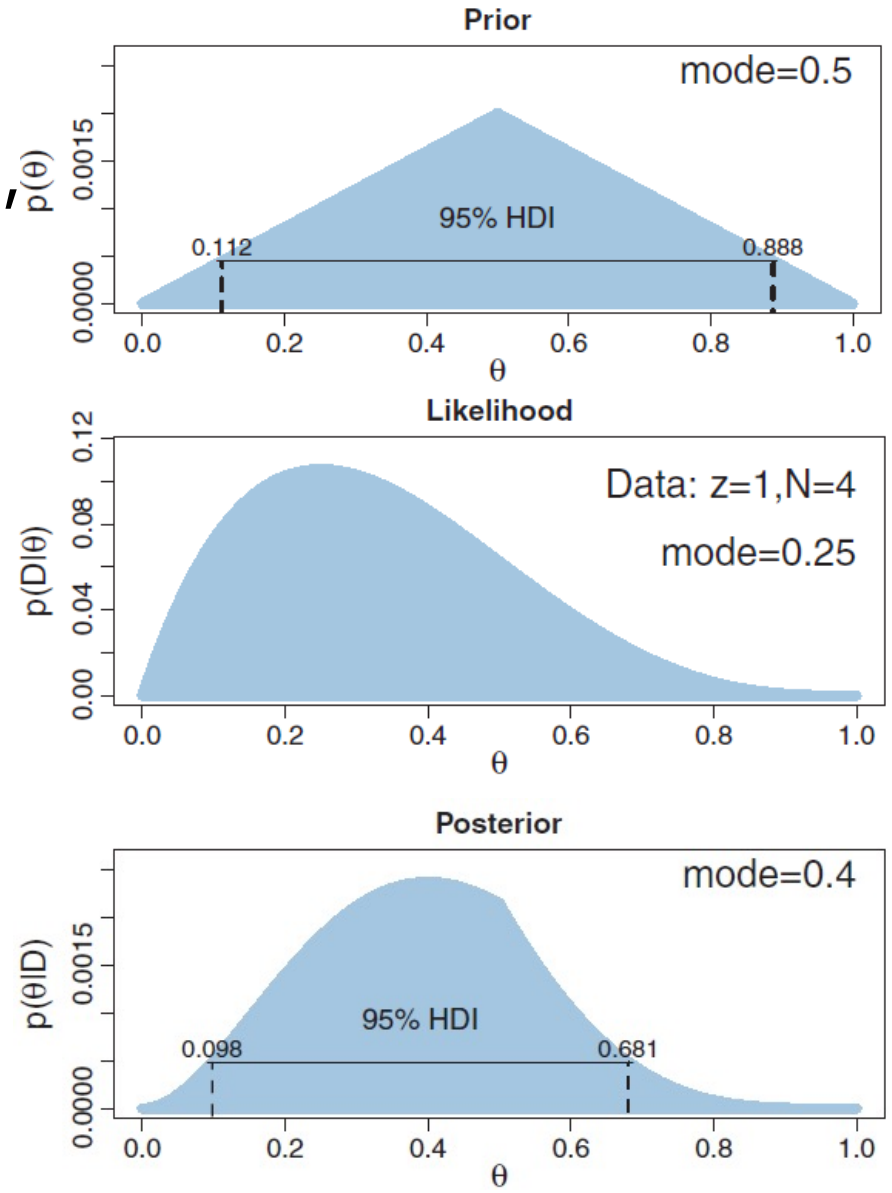
```
def bern_grid(theta, p_theta, z, N):
    #遍历所有θ,计算θ^z(1 − θ^(1−z))
    p_D_given_theta = likelihood(theta, z, N)
    #遍历所有θ,计算p(D|θ)p(θ)，然后求和
    p_D = evidence(p_theta, p_D_given_theta)
    #遍历所有θ,计算p(D|θ)p(θ)/p(D)
    p_theta_given_D =
      posterior(p_D_given_theta, p_theta, p_D)
```

python实现

```python
def bern_grid(theta, p_theta, z, N):

    p_D_given_theta = theta**z * (1 - theta)**(N - z)

    p_D = sum(p_D_given_theta * p_theta)

    p_theta_given_D = p_D_given_theta * p_theta / p_D

    return p_theta_given_D
```

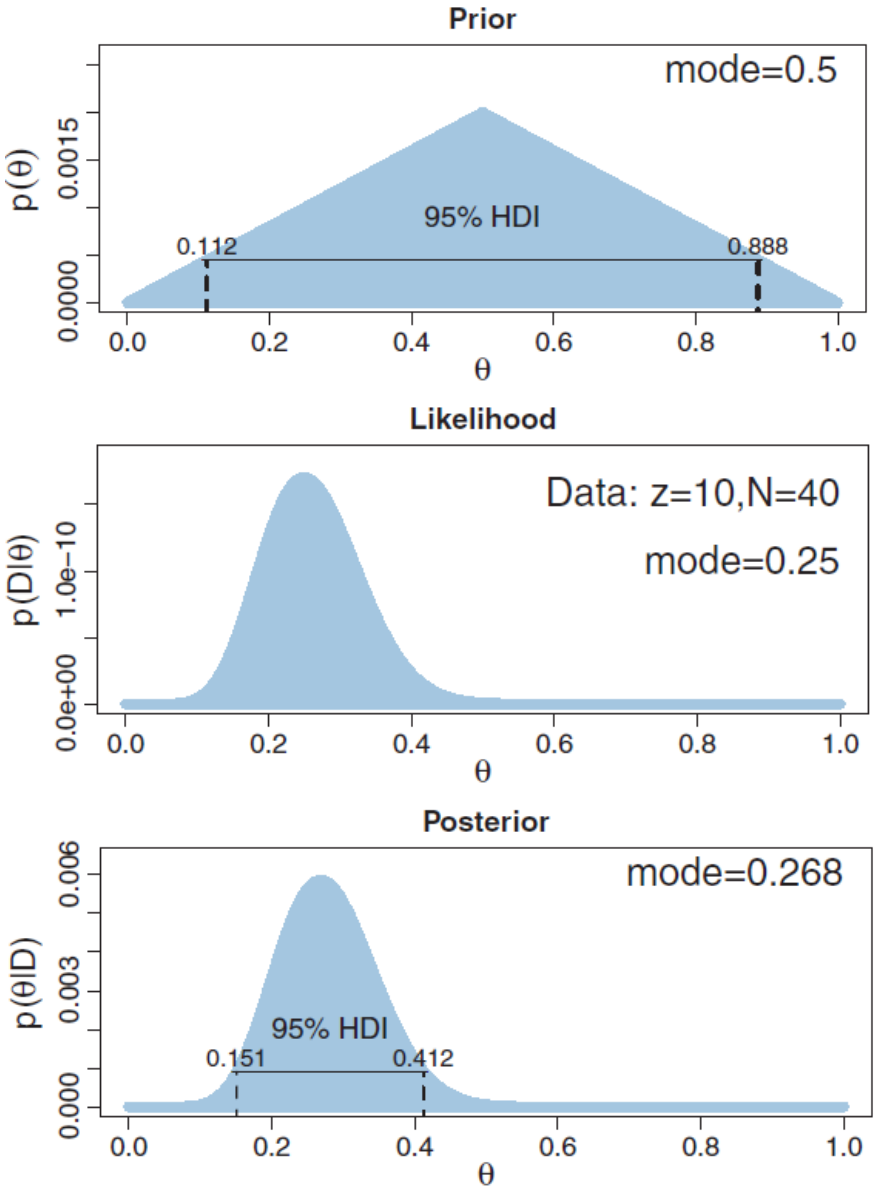
5.3.1 样本数量对后验的影响

- 我们将 θ 的离散值从10个提高到1000个，也就是$\{0.0, 0.001, 0.002, \ldots 0.999, 1.0\}$
- 抛硬币次数$N = 4$，正面次数$z = 1$。
  - $p(\{y_i\}|\theta) = \theta^z(1-\theta)^{N-z}$



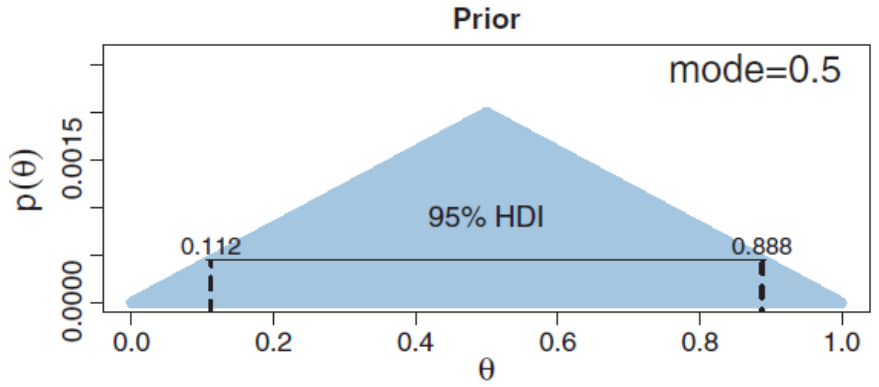
样本数量对后验的影响

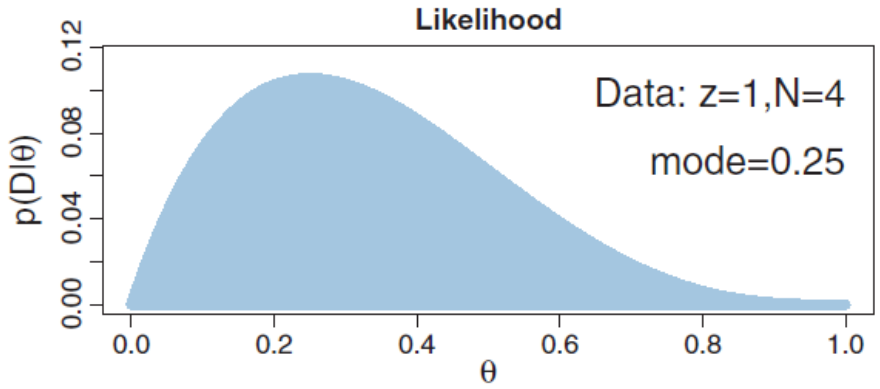
- 抛硬币次数$N = 40$，正面次数$z = 10$。
  - $p(\{y_i\}|\theta) = \theta^z(1-\theta)^{N-z}$

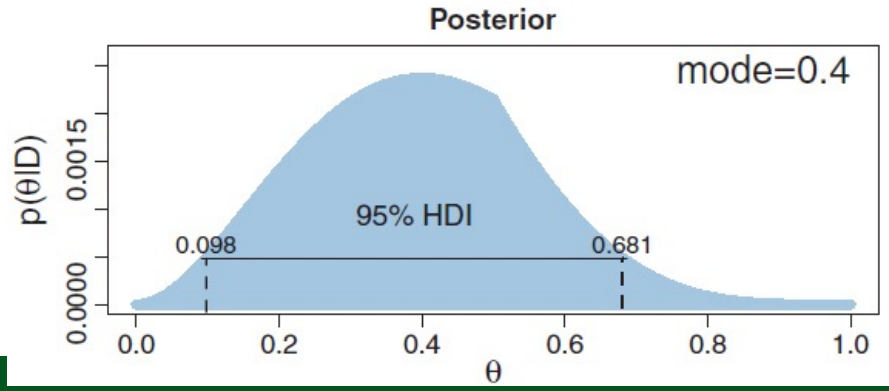


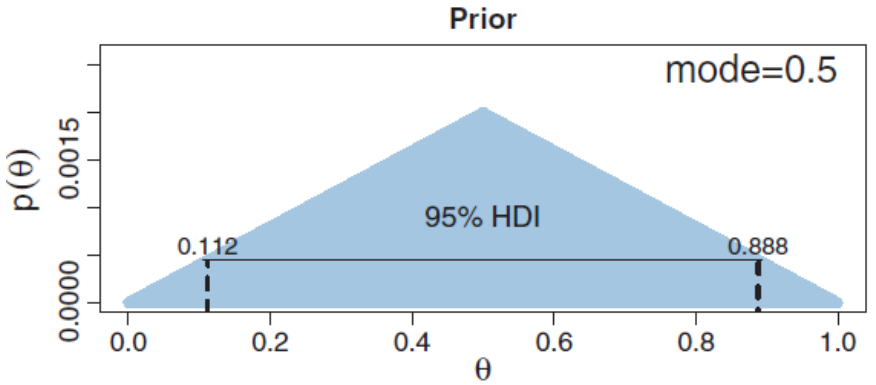
样本数量对后验的影响

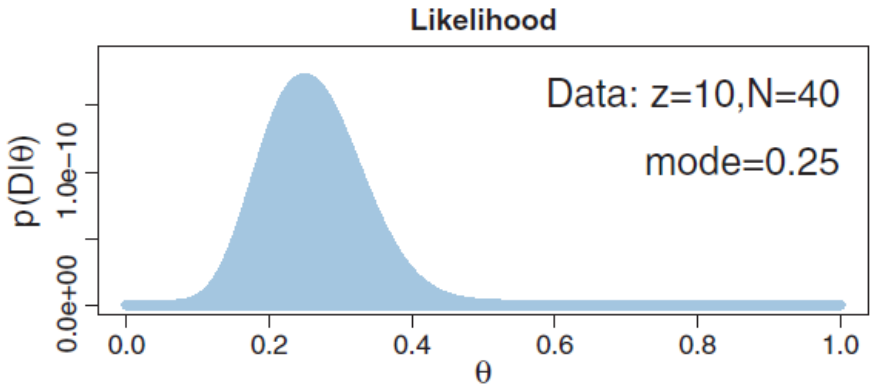
$N = 4$
$z = 1$

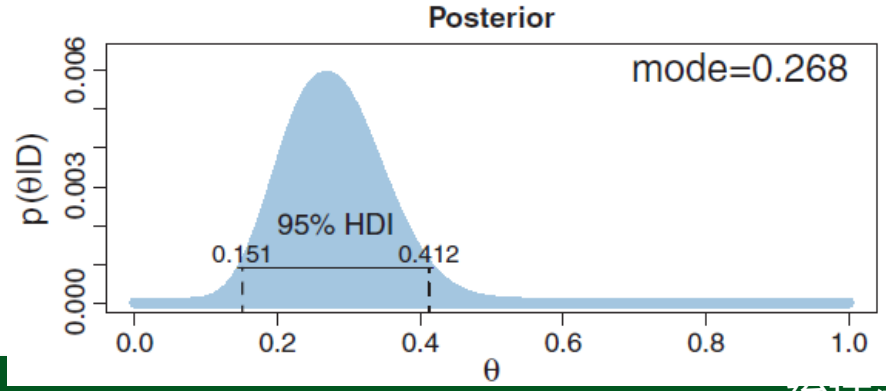








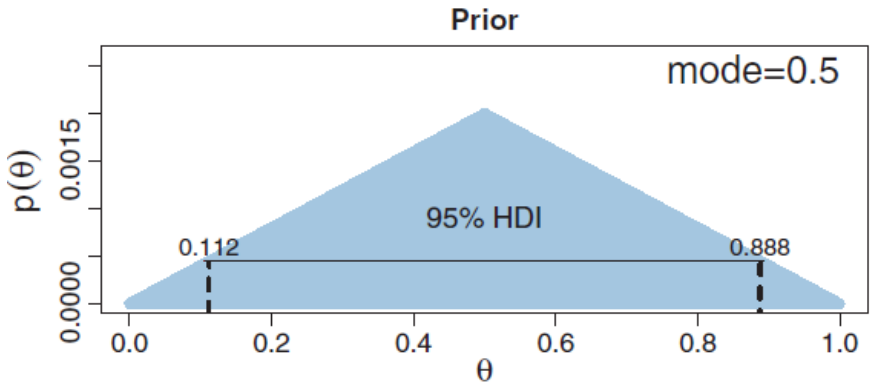


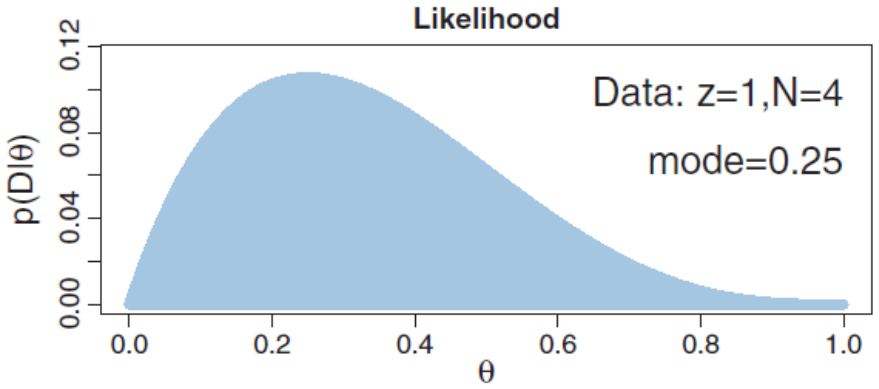


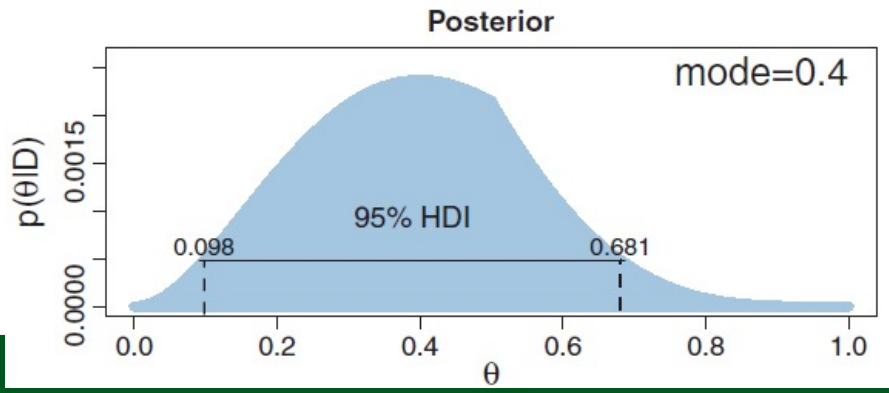
$N = 40$
$z = 10$

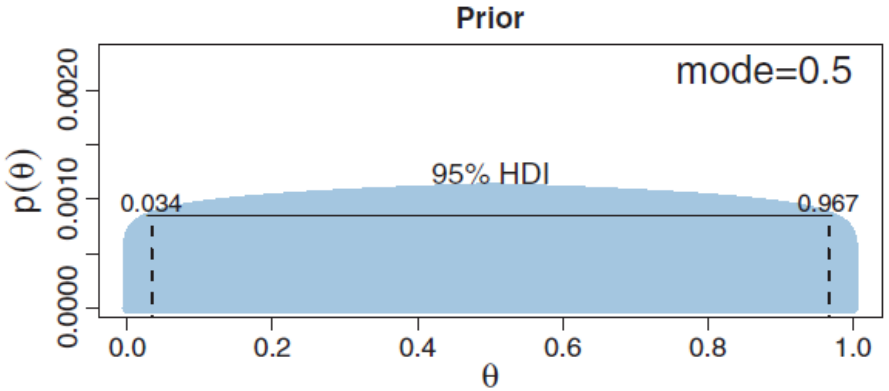
5.3.2 先验对后验的影响

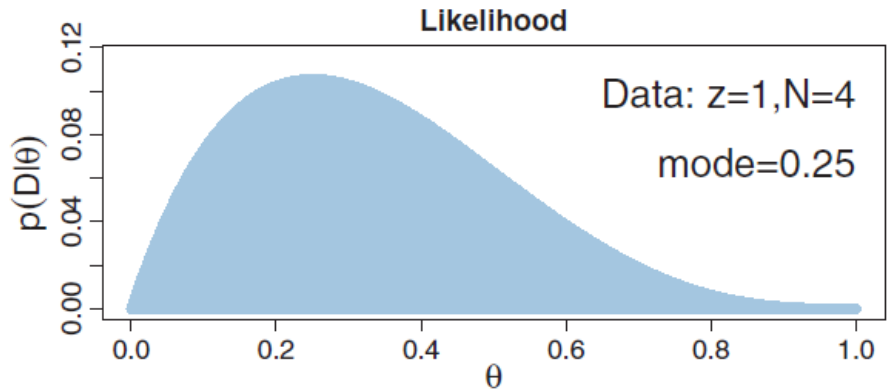
$N = 4$
$z = 1$

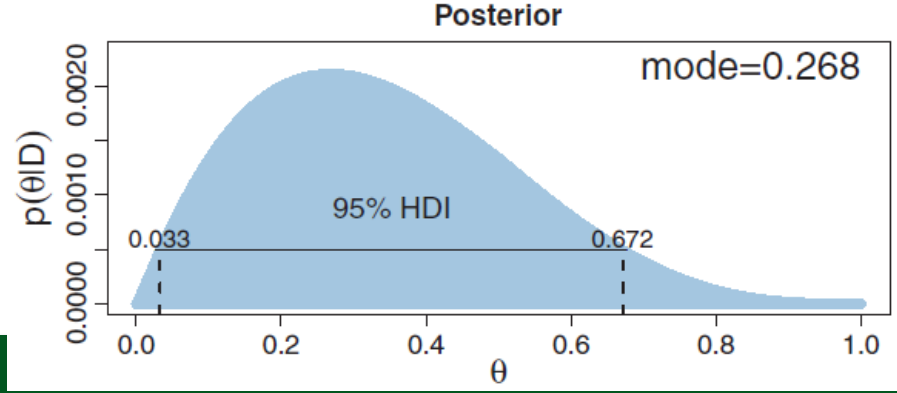










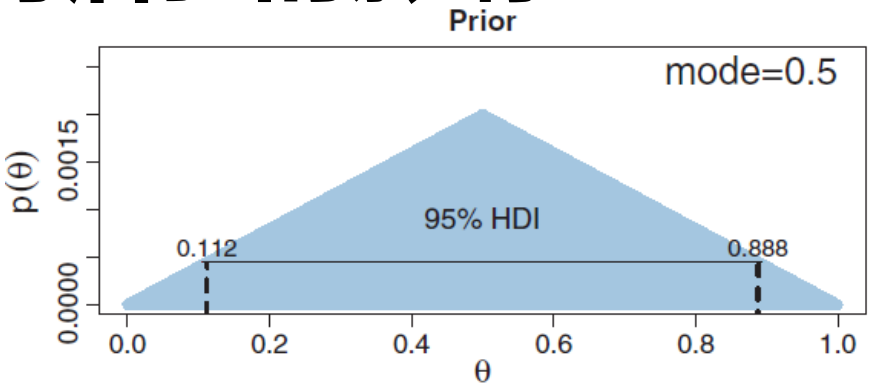


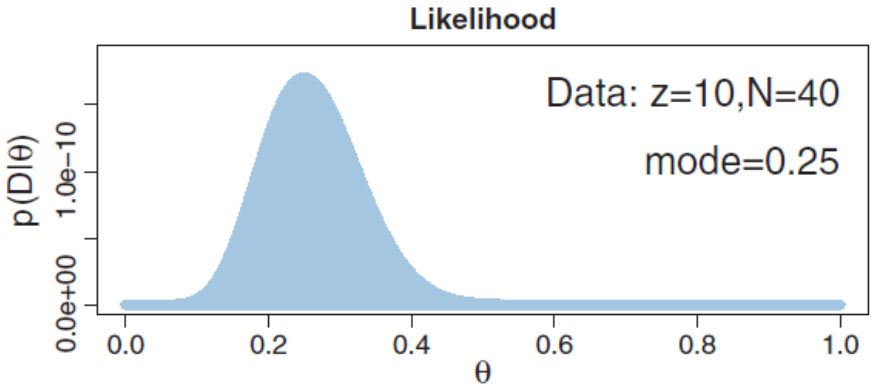
$N = 4$
$z = 1$

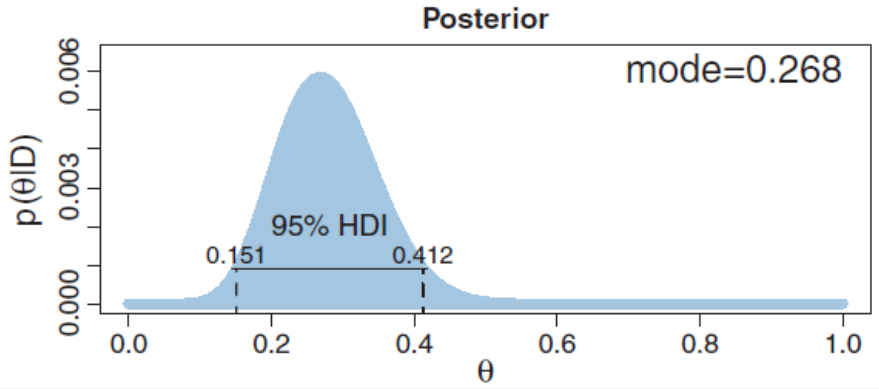
先验对后验的影响

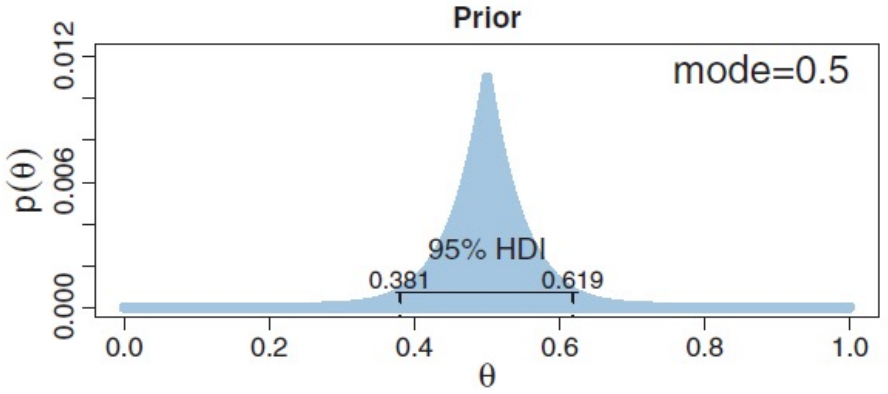
$N = 40$

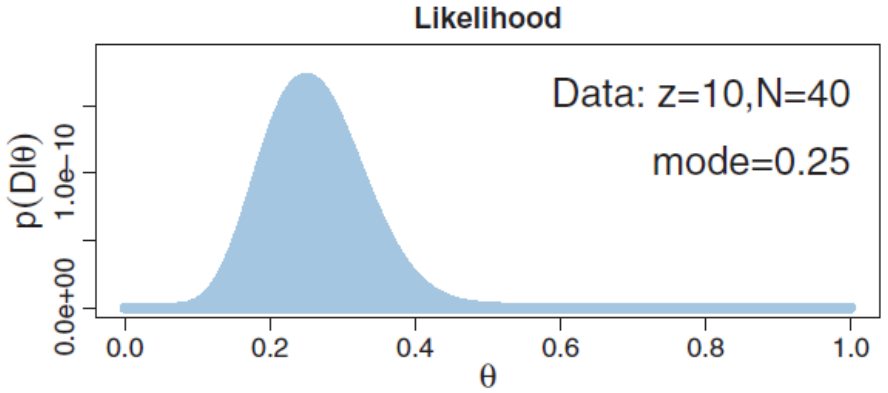
$z = 10$

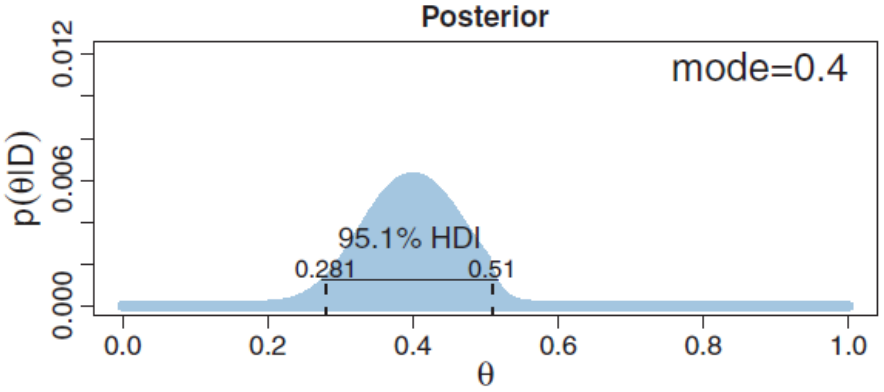












$N = 40$

$z = 10$

极大似然估计 vs. 贝叶斯估计

- 极大似然估计得到的θ的值只有一个，因此是一种“点估计”（point estimate）。

- 和“点估计”不同的是，贝叶斯估计能够估计θ所有可能取值的概率，也就是θ的概率分布。
  - 能够看出对θ估计的不确定性。

- 极大似然估计是一种估计量（estimator）,可以作为“频率学派推理”的估计量。

240

Discussion [D]What are some "important" problems in machine learning/AI? (self.MachineLearning)

Netero1999 於 17 天前 發表

I am not talking about "hot stuff" like self driving cars or anything, but topics important to the field(like maybe interpretability of machine learning?) which is fundamental to the advancement of the field.

146 留言 分享 儲存 隱藏 give award 檢舉 crosspost

[–] **lorepieri** 242 指標 17 天前

Estimating the uncertainty and confidence interval in AI/ML predictions.

永久連結 embed 儲存 檢舉 give award 回覆

[–] **quadprog** 111 指標 17 天前

Commenting because this answer is short so people might overlook it. This is a huge, important problem. Simpler ML models were good at providing uncertainty estimates. Now deep networks do not have analytical solutions and the obvious technique (MCMC) is extremely computationally expensive. Anyone who figures out a reliable and inexpensive way to get uncertainty estimates from deep NNs will be a hero of ML.

永久連結 embed 儲存 上層留言 檢舉 give award 回覆

5.4 贝叶斯推理的难点

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- 实际问题中，θ一般是连续值，上式的难点在于计算$p(D) = \int p(D|\theta)\, p(\theta)\, d\theta$。
  4种解决方法：

1. 网格近似（grid approximation），将θ用离散值近似。
2. 准确数学分析，需要通过构造合适的先验$p(\theta)$形式。
3. 马尔科夫蒙特卡洛近似（MCMC approximation），通过采样很多θ的值。
4. 变分近似（variational approximation），用另一个$q(\theta)$来近似$p(\theta|D)$，转换为一个优化问题。

5.4 贝叶斯推理的难点

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}$$

- 实际问题中，θ一般是连续值，上式的难点在于计算$p(D) = \int p(D|\theta)\, p(\theta)\, d\theta$。4种解决方法：

1. 网格近似（grid approximation），将θ用离散的值近似。
2. 准确数学分析，需要通过构造合适的先验$p(\theta)$形式。
3. 马尔科夫蒙特卡洛近似（MCMC approximation），通过采样很多θ的值。
4. 变分近似（variational approximation），用另一个$q(\theta)$来近似$p(\theta|D)$，转换为一个优化问题。

总结

- 似然函数
- 极大似然估计
- 频率学派推理
- 贝叶斯学派推理
- 网格近似