第4章 贝叶斯推理方法：准确数学分析

中山大学人工智能学院
毛旭东

Email：maoxd3@mail.sysu.edu.cn

网格近似 VS. 准确数学分析

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

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

- 在网格近似中，我们用“求和”来近似计算“积分”。
- 在准确数学分析方法中，我们想办法“准确”计算出积分的表达式。

例子：抛硬币

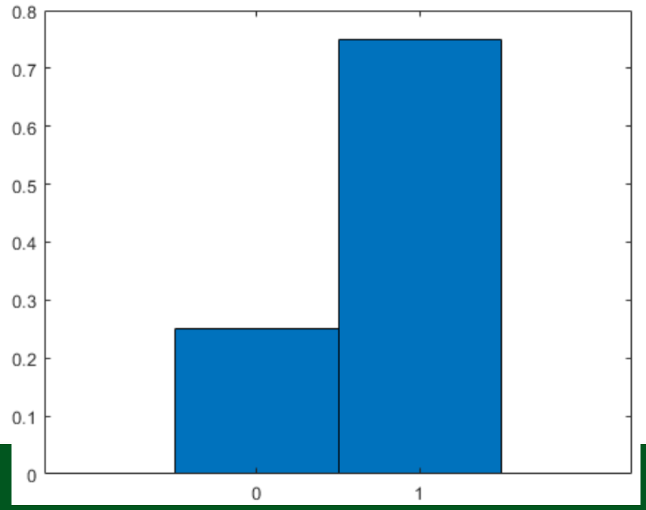
- 抛一次硬币的结果，可能是正面，也可能是反面。
- 我们用y表示一次抛硬币的结果：
  - 若为正面，我们记为$y=1$；
  - 若为反面，我们记为$y=0$。
- 结果为正面的概率，我们记为$p(y=1)$。
- 我们用参数θ来表示硬币“正面的偏向性”，从而构建一个简单的模型$p(y=1|\theta)=\theta$。
- “反面的偏向性”为$p(y=0|\theta)=1-\theta$。

模型（似然）

- $p(y=1|\theta)=\theta$
- $p(y=0|\theta)=1-\theta$
- 上式可以合并为：

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

- 上述分布称为伯努利分布（Bernoulli distribution）



模型（似然）

- 当我们抛N次硬币时，第i次的结果用y_i表示，所有结果的集合用$\{y_i\}$表示。
- 结果为正面的次数用$z=\sum_i y_i$表示，则反面的次数为 $N-z=\sum_i(1-y_i)$。
- 假设不同次抛硬币之间是独立的，我们可以得到：

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

注：和二项分布（Binomial distribution）的区别

- 伯努利分布抛N次硬币：

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

- 二项分布：

$$
p(D|\theta) = p(z|N,\theta) = \binom{N}{z}\theta^{z}(1-\theta)^{N-z}
$$

（DBDA第二版126页脚注）

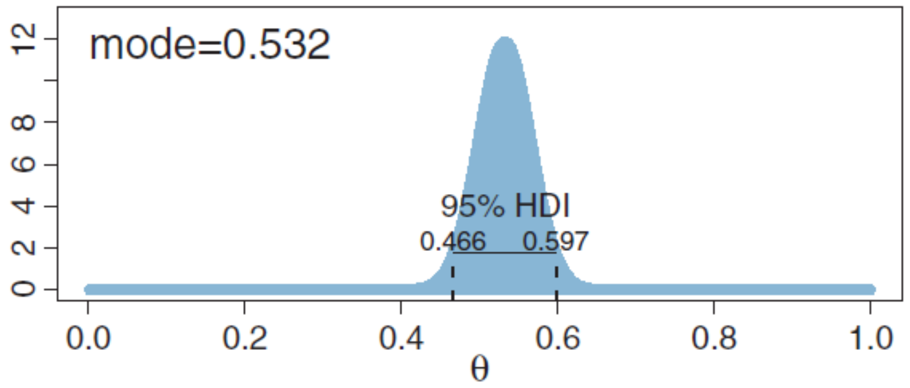
贝叶斯推理

目标：

- 给定一组抛硬币结果$\{y_i\}$，估计θ所有可能取值的概率，也就是θ的概率分布，即$p(\theta|\{y_i\})$。

注意：

- 贝叶斯推理估计的是θ的完整分布，而不是最有可能的单个θ的值。



贝叶斯推理：贝叶斯法则

目标：

- 给定一组抛硬币结果$\{y_i\}$，估计θ所有可能取值的概率，也就是θ的概率分布，即$p(\theta|\{y_i\})$。

利用贝叶斯法则：

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

- $p(\{y_i\}|\theta)$是似然函数，已假设。
- $p(\theta)$是先验（prior）概率分布，未知，由我们假设指定。
- $p(\{y_i\})$是证据（evidence），根据先验和似然函数求解，$p(\{y_i\}) = \int p(\{y_i\}|\theta)\, p(\theta)\, d\theta$。
- $p(\theta|\{y_i\})$是后验（posterior）概率分布，是求解的目标。

6.2 贝叶斯推理：先验设计

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

- 先验$p(\theta)$理论上可以是任何分布。
- 我们根据以下3个思路设计$p(\theta)$：
  - $p(\theta)$形式上最好和$p(\{y_i\}|\theta)$一致，从而使得$p(\theta|\{y_i\})$在形式上也一致，便于数学推导。
  - $p(\theta)$最好能使得$p(\{y_i\})$方便求解，$p(\{y_i\}) = \int p(\{y_i\}|\theta)\,p(\theta)\,d\theta$。
  - $p(\theta)$有足够的表达力来表达我们的先验知识（i.e., 概率密度函数曲线形状多样）。

贝叶斯推理：先验设计思路

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

- 似然函数：$p(\{y_i\}|\theta)=\theta^z(1-\theta)^{N-z}$。
- 假设先验的形式设计为$\theta^a(1-\theta)^b$， $p(\{y_i\}|\theta)$和$p(\theta)$相乘后的形式为$\theta^{(z+a)}(1-\theta)^{(N-z+b)}$。

- 当由似然函数$p(y|\theta)$和先验$p(\theta)$得到的后验$p(\theta|y)$在形式上和先验相同时，该先验被称为是似然函数$p(y|\theta)$的共轭先验（Conjugate Prior）。

先验：Beta分布

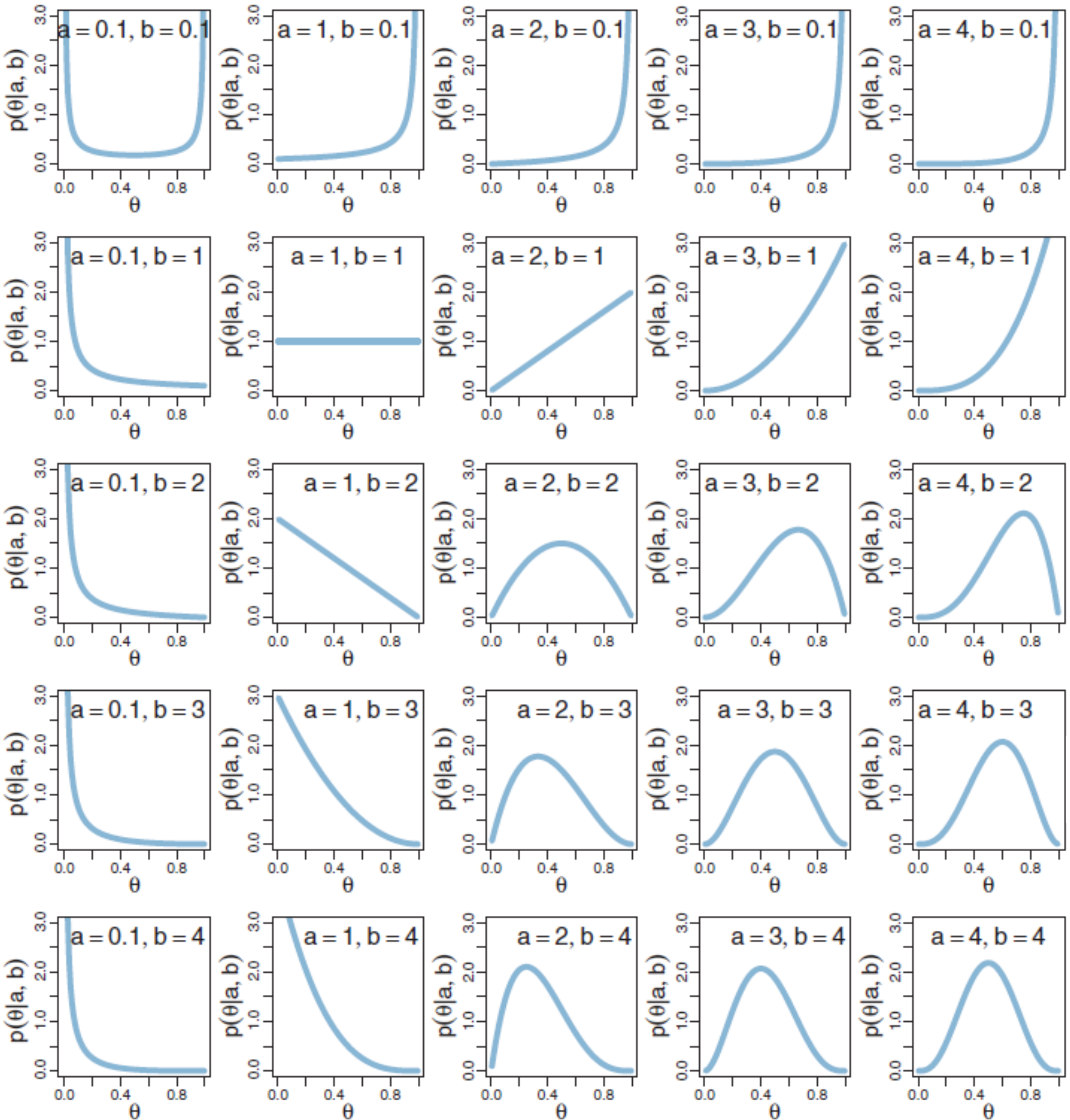
Beta分布：

$$p(\theta|a,b) = \text{beta}(\theta|a,b)$$
$$= \theta^{(a-1)}(1-\theta)^{(b-1)}/B(a,b)$$

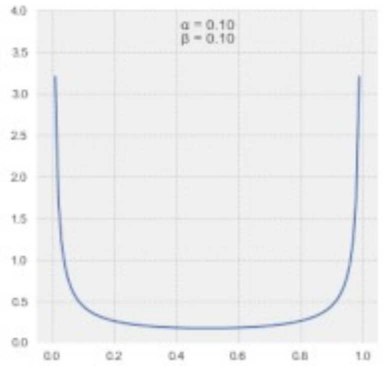
- 其中$\theta \in [0,1]$，$\mathrm{a} > 0, \mathrm{b} > 0$。
- $B(a,b)$是归一化常数，使得$p(\theta|a,b)$关于θ的积分是1，也就是：

$$B(a,b) = \int_0^1 \mathrm{d}\theta\, \theta^{(a-1)}(1-\theta)^{(b-1)}$$

- $B(a,b)$被称为beta函数，注意该函数和θ无关。



Beta分布

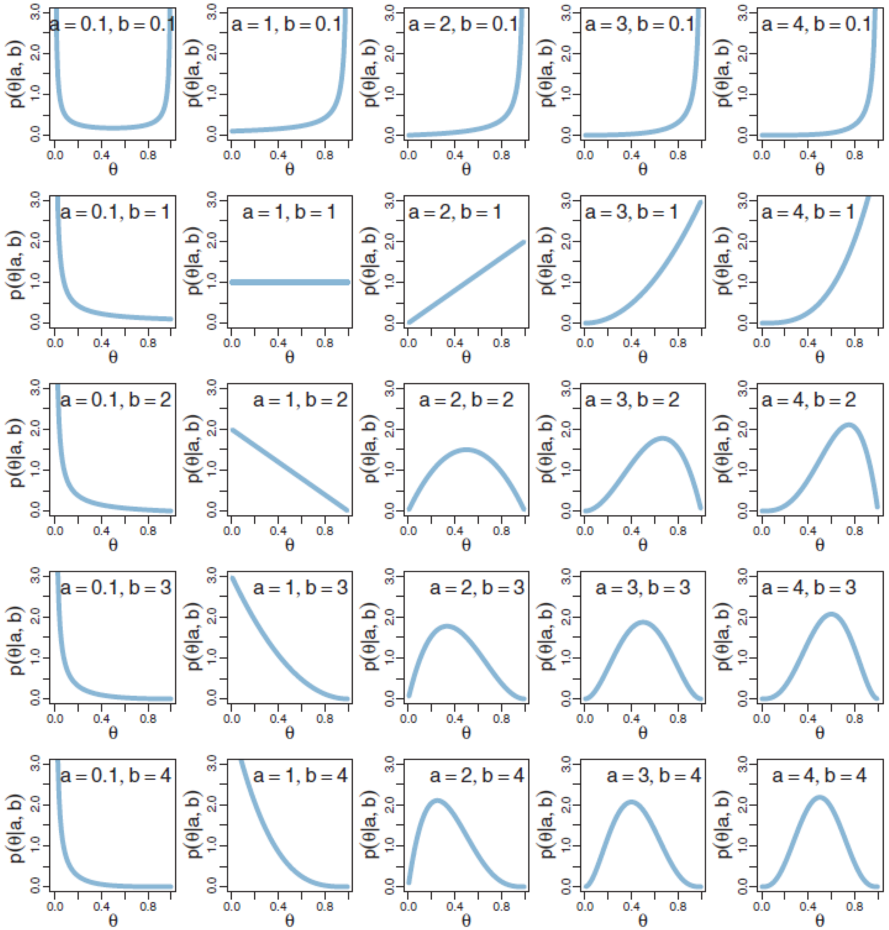


$$p(\theta|a,b) = \text{beta}(\theta|a,b) = \theta^{(a-1)}(1-\theta)^{(b-1)}/B(a,b)$$

Beta分布

Beta分布：

$$p(\theta|a,b) = \text{beta}(\theta|a,b) = \theta^{(a-1)}(1-\theta)^{(b-1)}/B(a,b)$$



观察上图，我们可以得到：

- 当a变大时，概率值的主体向右偏移，也就是集中在θ大的区域。
- 当b变大时，概率值的主体向左偏移。
- 当a和b同时变大时，概率密度函数变窄，也就是值更集中。
- a和b被称为Beta分布的形状参数（Shape Parameters）。

Beta分布的均值、众数、集中度

对于$\text{beta}(\theta|a,b)$：

- 均值（mean）为：

$$\mu = \frac{a}{a+b}$$

- 峰值（mode）为：

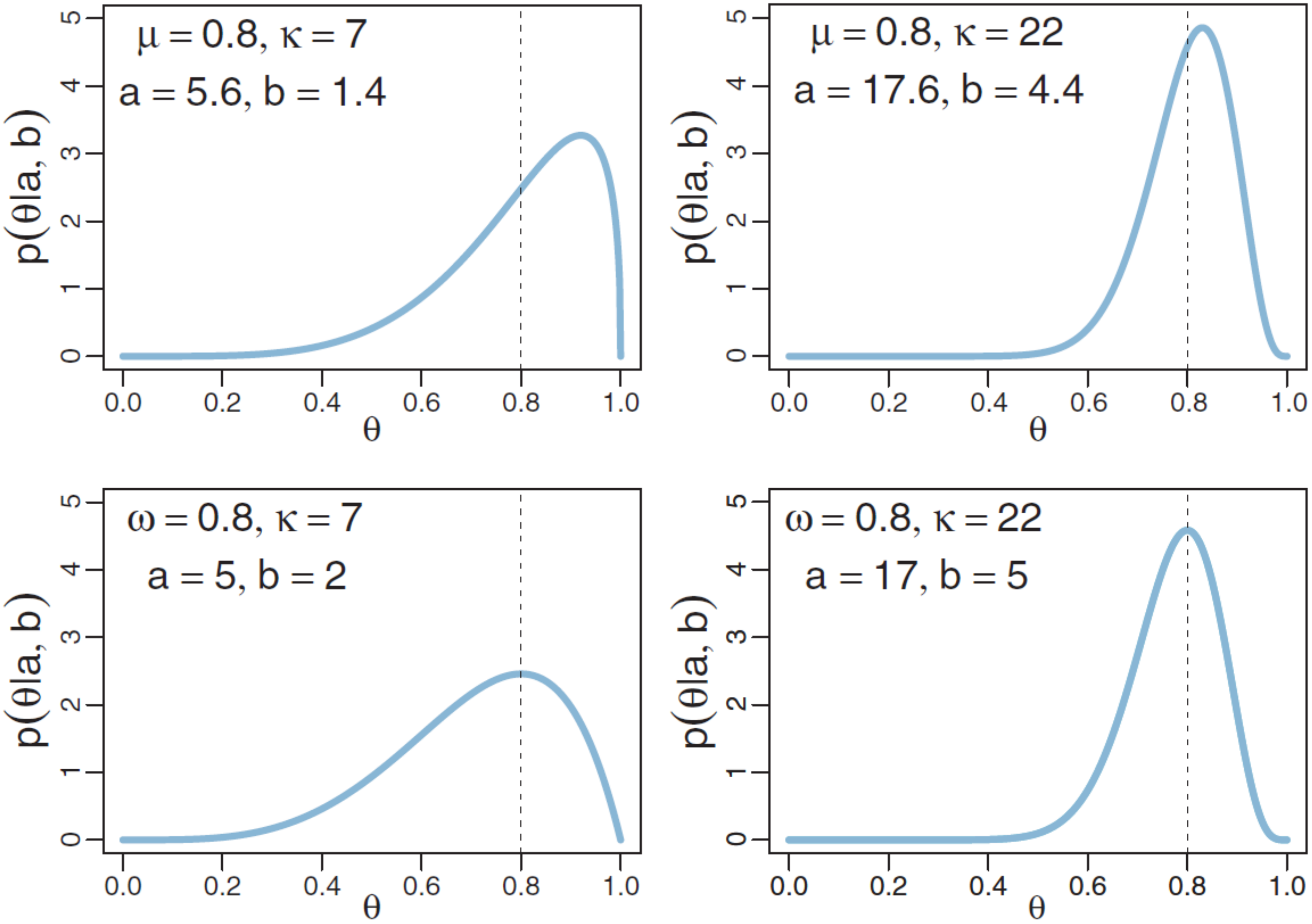
$$\omega = \frac{a-1}{a+b-2}，\text{其中}a > 1，\ b > 1$$

- 集中度（concentration）为：

$$\kappa = a + b$$

  - κ越大，越集中。

- $a = \omega(\kappa - 2) + 1，b = (1-\omega)(\kappa - 2) + 1$



μ = 0.8, κ = 7
a = 5.6, b = 1.4
μ = 0.8, κ = 22
a = 17.6, b = 4.4
ω = 0.8, κ = 7
a = 5, b = 2
ω = 0.8, κ = 22
a = 17, b = 5
p(θ|a, b)
θ
0.0 0.2 0.4 0.6 0.8 1.0
0 1 2 3 4 5

6.3 求解后验

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

- 后验：$p(\theta|\{y_i\})$。
- 似然函数：$p(\{y_i\}|\theta) = \theta^{z}(1-\theta)^{N-z}$
- 假设先验是Beta分布：$p(\theta) = \theta^{a-1}(1-\theta)^{b-1}/B(a,b)$
- 假设一组抛硬币的结果$\{y_i\}$为：N次中有z次是正面。

求解后验

$$p(\theta|\{y_i\}) = p(\{y_i\}|\theta)\,p(\theta)/p(\{y_i\})$$

$$= \frac{\theta^{z}(1-\theta)^{N-z}\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)p(\{y_i\})}$$

$$= \frac{\theta^{(z+a-1)}(1-\theta)^{(N-z+b-1)}}{B(a,b)p(\{y_i\})}$$

- 对比beta分布的表达式：

$$p(\theta|a,b) = \text{beta}(\theta|a,b) = \theta^{(a-1)}(1-\theta)^{(b-1)}/B(a,b)$$

- 可以看出$p(\theta|\{y_i\})$服从beta分布。

$$p(\theta|a,b) = \text{beta}(\theta|a,b) = \theta^{(a-1)}(1-\theta)^{(b-1)}/B(a,b)$$

Beta分布

中山大學 SUN YAT-SEN UNIVERSITY

求解后验

$$p(\theta|\{y_i\}) = \frac{\theta^{(z+a-1)}(1-\theta)^{(N-z+b-1)}}{B(a,b)p(\{y_i\})}$$

- 根据分子$\theta^{((z+a)-1)}(1-\theta)^{((N-z+b)-1)}$，可以得出后验$p(\theta|\{y_i\})$服从$\text{beta}(\theta|z+a, N-z+b)$。
- 分母是归一化常数，我们可以“倒推”出分母为：

$$B(a,b)p(z,N) = B(z+a, N-z+b)$$

- 因此：

$$p(\theta|\{y_i\}) = \text{beta}(\theta|z+a, N-z+b)$$

简化推导

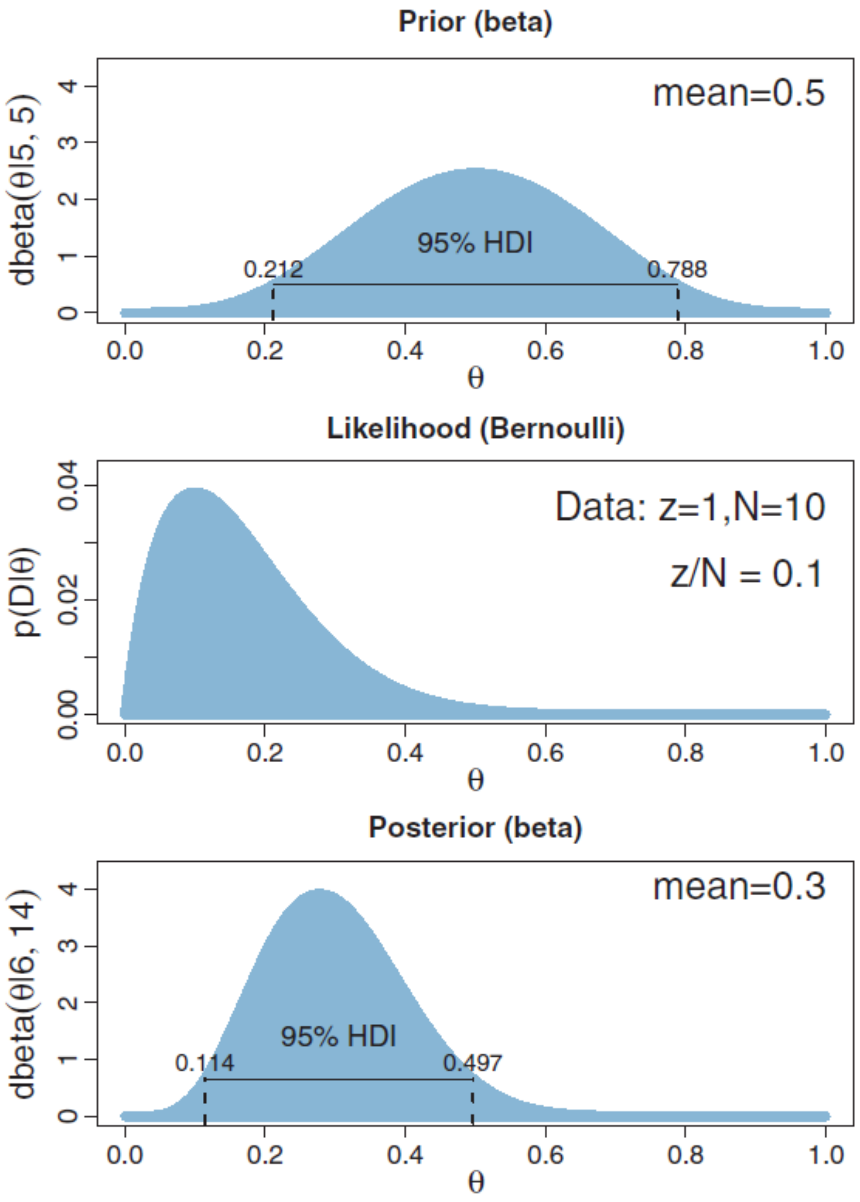
- 省略常数项，只保留和θ有关的项，用“正比于”来推导：

$$
\begin{aligned}
p(\theta|\{y_i\}) &\propto p(\{y_i\}|\theta)\, p(\theta) \\
&\propto \theta^{z}(1-\theta)^{N-z}\theta^{a-1}(1-\theta)^{b-1} \\
&\propto \theta^{(z+a-1)}(1-\theta)^{(N-z+b-1)}
\end{aligned}
$$

- 可得，$p(\theta|\{y_i\}) = \mathrm{beta}(\theta|z+a, N-z+b)$

6.3.1 后验是先验和似然之间的妥协

- 先验：
  - $\text{beta}(\theta|a=5,b=5)$
  - 均值 ＝ 峰值 ＝ 0.5
- 数据：$z=1, N=10$
  - 似然：$\theta^z(1-\theta)^{N-z}$
  - 峰值 = 0.1
- 后验：
  - $\text{beta}(\theta|z+a, N-z+b)$
  - $\text{beta}(\theta|6,14)$
  - 均值 = 0.3，峰值=0.28



后验是先验和似然之间的妥协

- 先验： $\mathrm{beta}(\theta|a,b)$
  - 均值 = $\frac{a}{a+b}$
- 数据： 峰值= $\frac{z}{N}$
- 后验： $\mathrm{beta}(\theta|z+a,N-z+b)$
  - 均值 = $\frac{z+a}{z+a+N-z+b} = \frac{z+a}{N+a+b}$
- 可得：

$$\underbrace{\frac{z+a}{N+a+b}}_{\text{posterior}} = \underbrace{\frac{z}{N}}_{\text{data}} \underbrace{\frac{N}{N+a+b}}_{\text{weight}} + \underbrace{\frac{a}{a+b}}_{\text{prior}} \underbrace{\frac{a+b}{N+a+b}}_{\text{weight}}$$

后验是先验和似然之间的妥协

$$\underbrace{\frac{z+a}{N+a+b}}_{\text{posterior}} = \underbrace{\frac{z}{N}}_{\text{data}} \underbrace{\frac{N}{N+a+b}}_{\text{weight}} + \underbrace{\frac{a}{a+b}}_{\text{prior}} \underbrace{\frac{a+b}{N+a+b}}_{\text{weight}}$$

- 当$N > (a+b)$时，后验的均值主要由数据决定。
- 当$N < (\mathrm{a}+\mathrm{b})$时，后验的均值主要由先验决定。

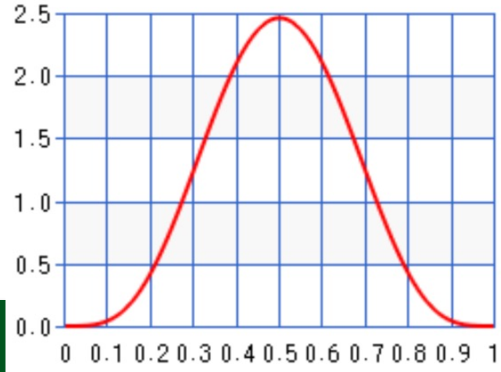
上式可以理解为：

- 先验：抛了$a+b$次硬币，其中a次是正面。
- 似然：抛了N次硬币，其中z次是正面。
- 后验：将先验和似然按一定权重组合。
  - 权重（weight）：一共$N+a+b$，其中先验占$a+b$，似然占N。

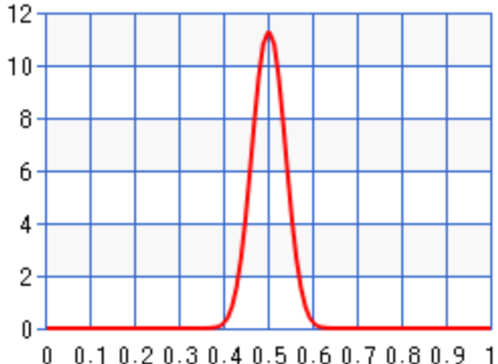
如何指定beta先验的参数

- 指定的先验一般需要考虑的是**集中趋势**（Central Tendency）和**对该集中趋势的确信度。**
  - 集中趋势包括平均数（mean）、众数（mode）等。
- 比如我们感觉正面的概率应该是50%，但不是很确定。
  - 假设先验：10次抛硬币，有5次是正面，$\text{beta}(\theta|5,5)$。
- 比如我们很确定正面的概率是50%。
  - 假设先验：200次抛硬币，有100次是正面，$\text{beta}(\theta|100,100)$

$\mathbf{beta}(\boldsymbol{\theta}|\mathbf{5},\mathbf{5})$:



$\mathbf{beta}(\boldsymbol{\theta}|\mathbf{100},\mathbf{100})$:



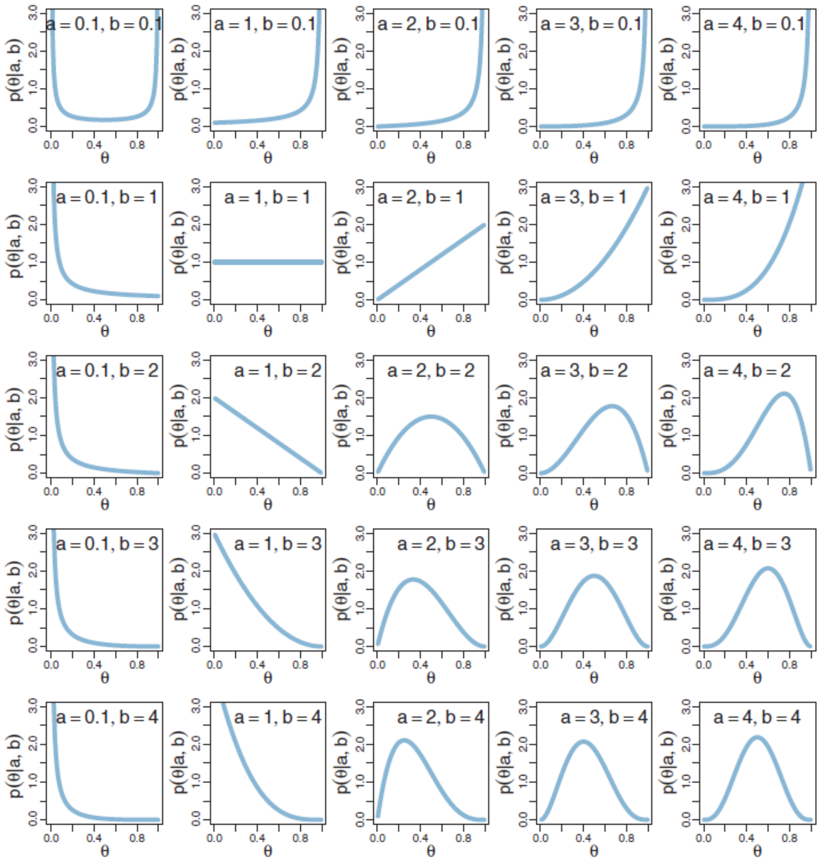
指定Beta先验的参数

- 对于Beta分布，峰值比均值更直观。

- 一种设定参数的方式是：

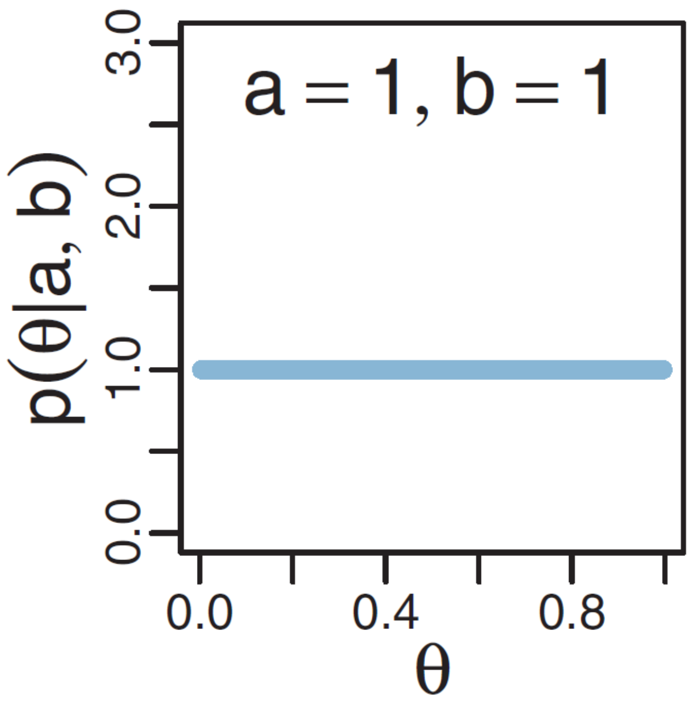
1. 根据我们相信的概率值，设定峰值$\omega = \frac{a-1}{a+b-2}$；
2. 根据我们的确信度，设定集中度$\kappa = a + b$；
3. 计算$a = \omega(\kappa - 2) + 1$，$b = (1 - \omega)(\kappa - 2) + 1$。

- 大部分情况下，我们使用$a \geq 1$并且$b \geq 1$。



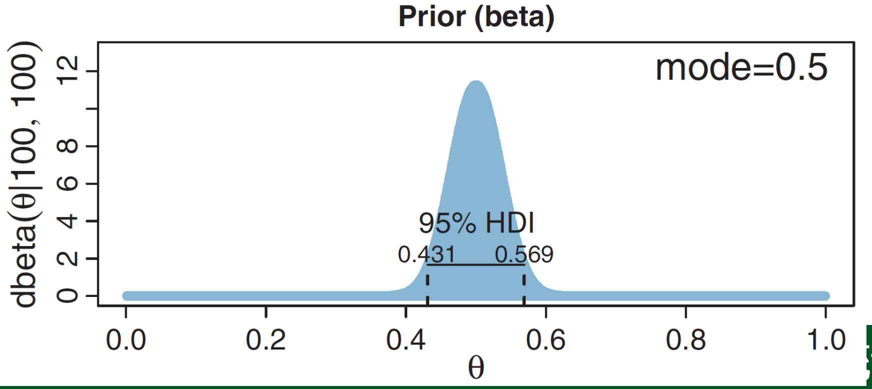
一种常用的先验：$\mathbf{beta}(\boldsymbol{\theta}|\mathbf{1},\mathbf{1})$

- $\mathrm{beta}(\theta|1,1)$是[0,1]上的均匀分布，如下图所示。
- 当我们对θ的值完全不确定时，可以采用这种不确定的先验（a vague and noncommittal prior）。



6.4.1 可以用Beta分布表达的先验知识

- **例子1：抛硬币**
- 先验知识：大部分的硬币都是均匀的，也就是**正面的概率是**0.5，并且我们对这个先验知识**很确定。**
- 假设先验为：$\text{beta}(\theta|100{,}100)$
  - $\omega = \frac{100-1}{200-2} = 0.5$ 表达了：正面概率是0.5。
  - $\kappa = 200$ 表达了：对“正面概率是0.5”很确定。
    - $\kappa = 200$，相当于200次抛硬币。

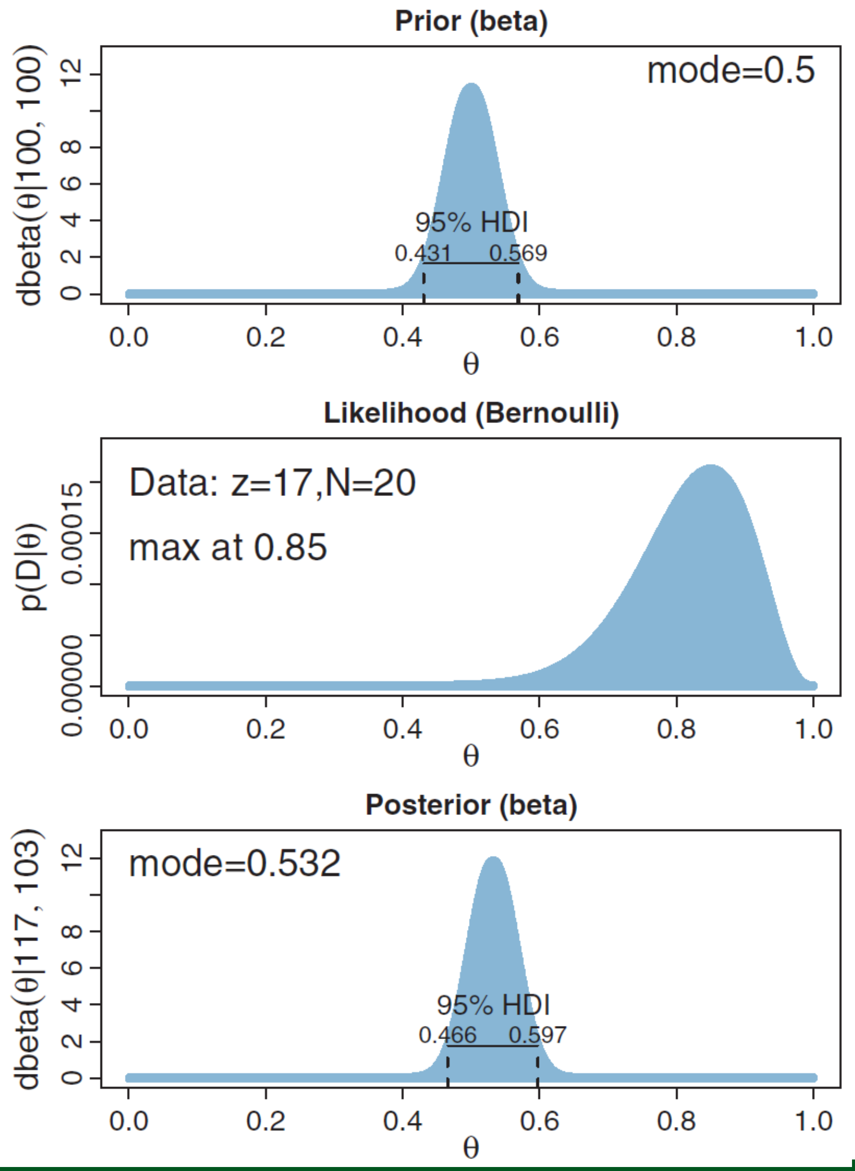


6.4.1 可以用Beta分布表达的先验知识

- **例子1：抛硬币**
- 先验：$\mathrm{beta}(\theta|100{,}100)$
- 似然：$N = 20, z = 17$
- 后验：

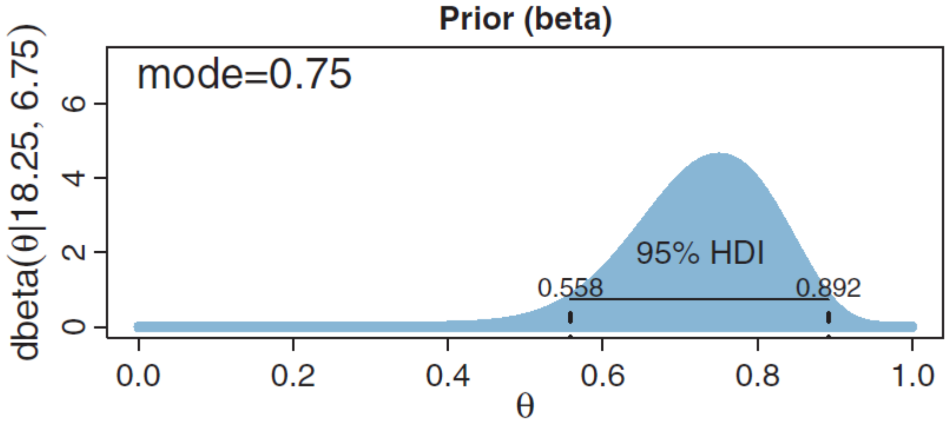
$$\mathrm{beta}(\theta|z + a, N - z + b)$$

- 即便20次里面有17次是正面，但是后验概率分布的峰值（0.532）仍是接近0.5。
- 20次抛硬币，相比于先验的“200次”，很少。



6.4.1 可以用Beta分布表达的先验知识

- **例子2：篮球罚球命中率**
- 先验知识：某个专业联赛的运动员，罚球的**平均命中率是75%，大部分运动员的命中率在50%至90%**之间。
- 假设先验为：$\text{beta}(\theta|18.25, 6.75)$
  - $\omega = \frac{18.25-1}{18.25+6.75-2} = 0.75$　　表达了：平均命中率是75%。
  - 95% HDI是[0.558, 0.892]　表达了：大部分运动员的命中率在50%至90%之间。

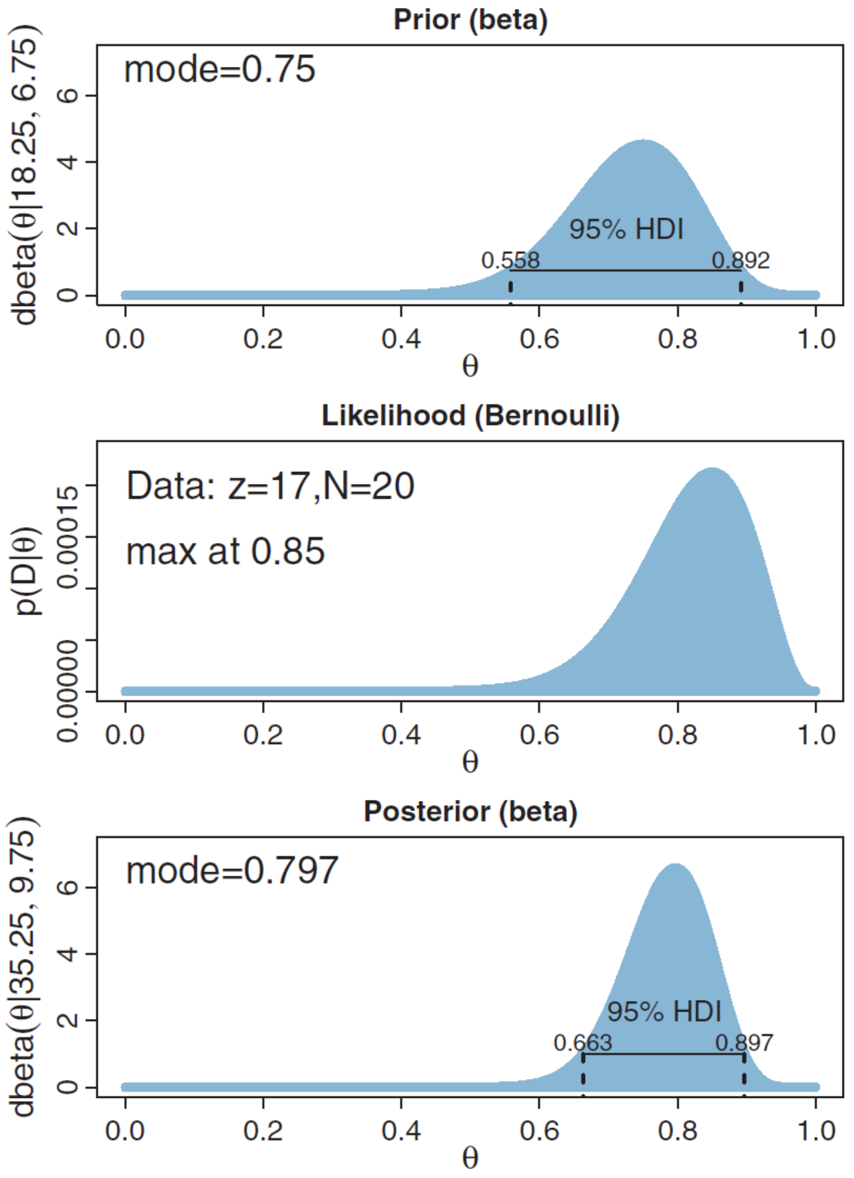


6.4.1 可以用Beta分布表达的先验知识

- 先验： $\text{beta}(\theta|18.25,6.75)$
- 似然： $N=20, z=17$
  - 投20次，中17次。
- 后验：

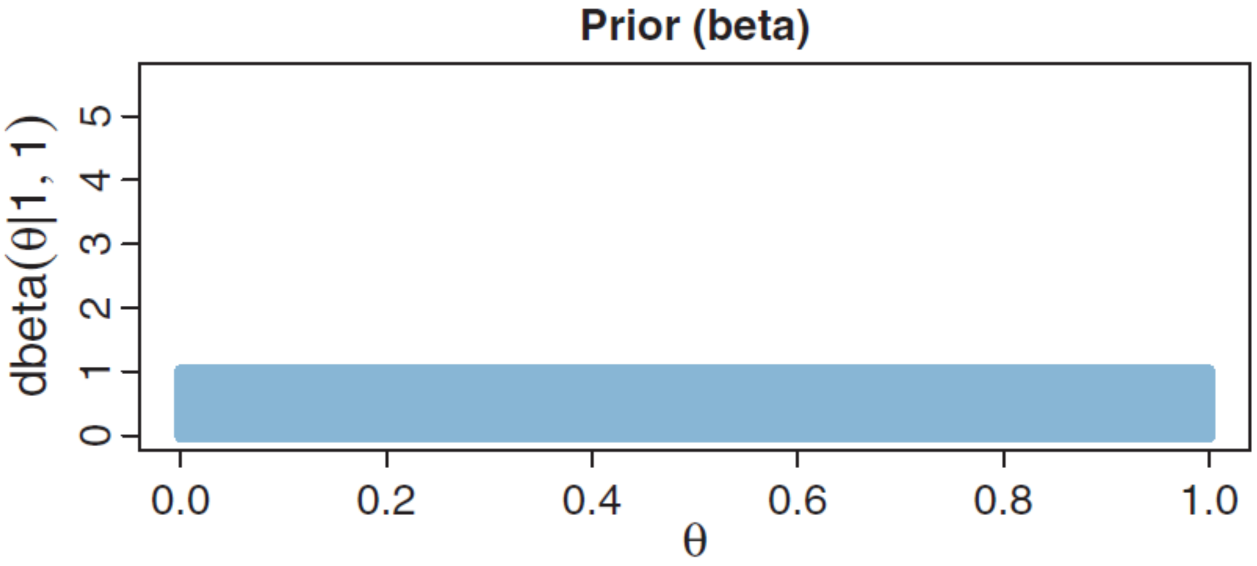
$$\text{beta}(\theta|z+a, N-z+b)$$

- 后验概率的峰值是0.797。是先验0.75和似然0.85的中间值。



6.4.1 可以用Beta分布表达的先验知识

- **例子3**：我们用一个外形探测器，探测某个外星球的物质的颜色，可能是蓝色或者绿色，估计该星球物质是蓝色的概率。
- 先验知识：我们对该星球的物质毫无所知。
- 因此假设“不确定的”先验：$\mathrm{beta}(\theta|1,1)$

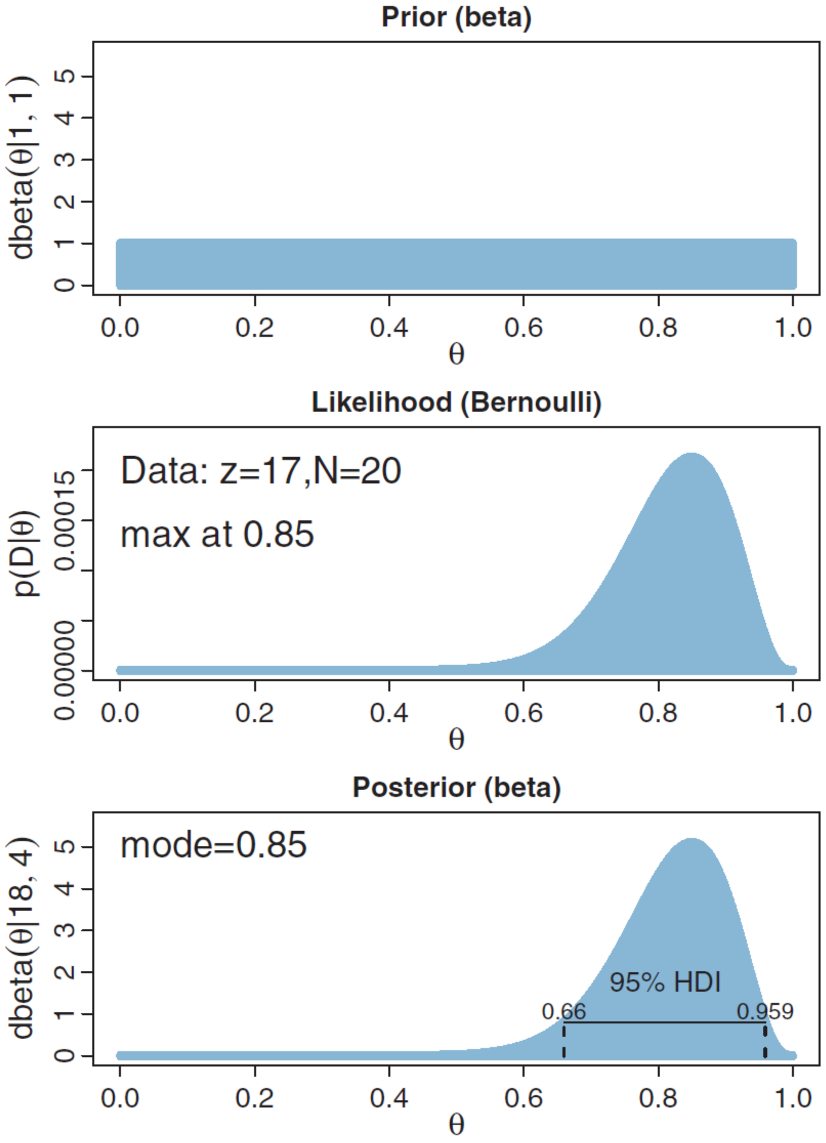


6.4.1 可以用Beta分布表达的先验知识

- 先验：$\text{beta}(\theta|1,1)$
- 似然：$N = 20, z = 17$
  - 采集20个样本，17个是蓝色。
- 后验：

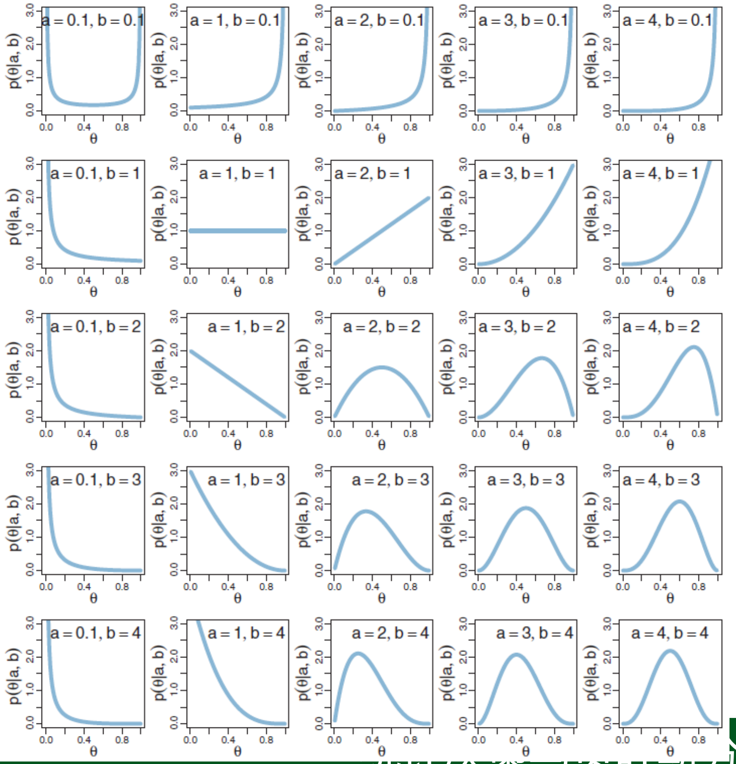
$$\text{beta}(\theta|z + a, N - z + b)$$

  - 峰值=$\frac{z+a-1}{N+a+b-2}$
- 探测器随机似然的峰值是0.85。
- 后验概率的峰值也是0.85。



6.4.2 不可以用Beta分布表达的先验知识

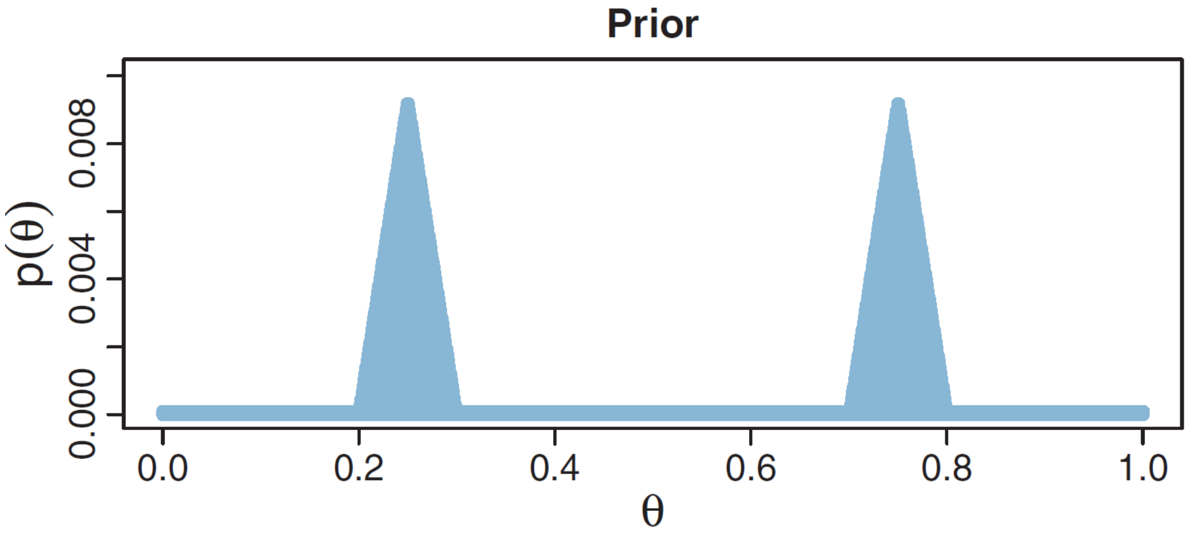
- 并不是所有的先验知识都能用Beta分布来表达，因为Beta分布受限制于下图的“表达能力”。
- 当先验知识无法用Beta分布表达时，我们需要用其他分布作为先验概率分布。
- 此时无法准确推导出后验概率分布的表达式。
- 需要用近似方法，比如网格近似、MCMC等。



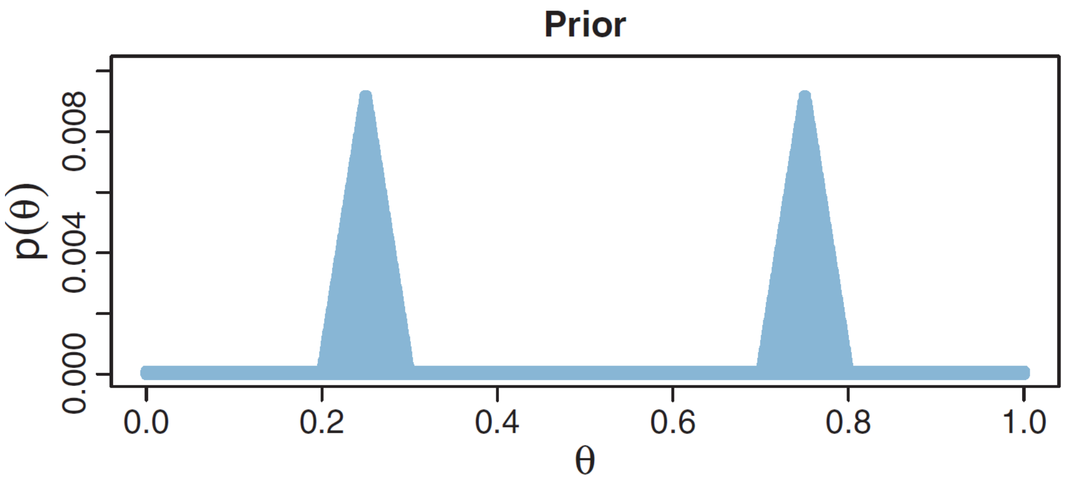
6.4.2 不可以用Beta分布表达的先验知识

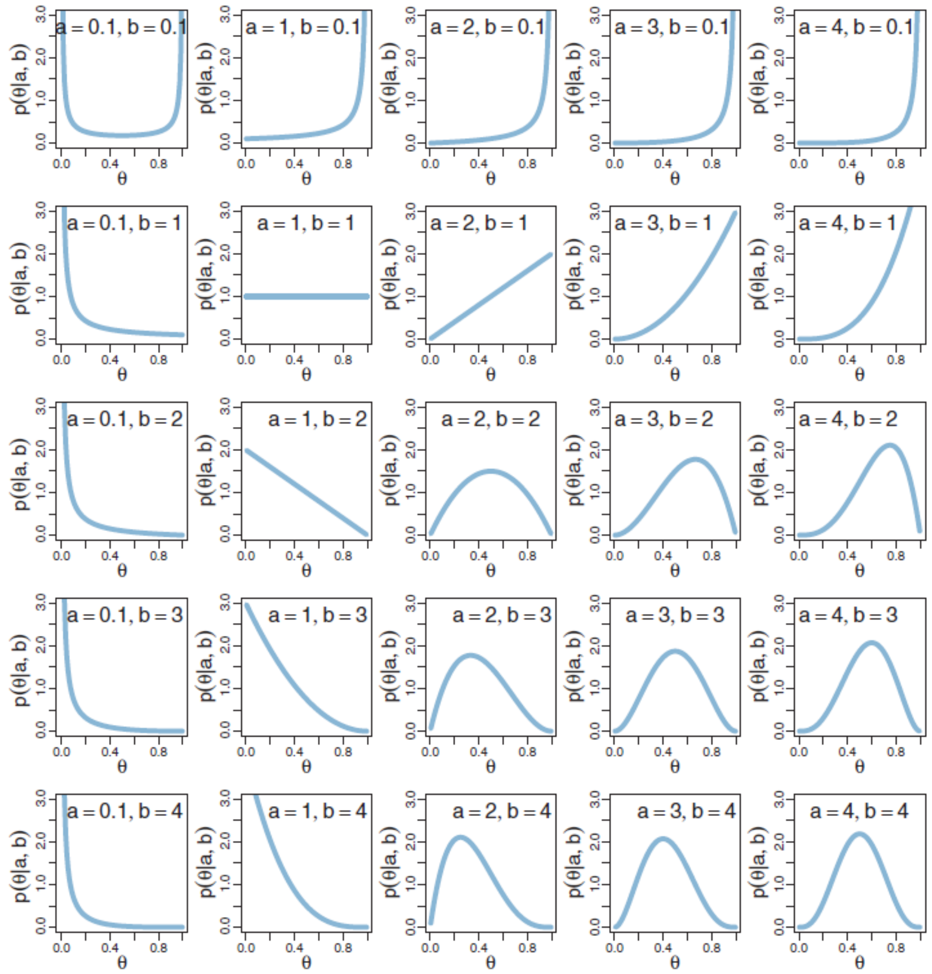
中山大學 SUN YAT-SEN UNIVERSITY

- **例子：**假设我们知道硬币是由A硬币厂或者B硬币厂制造，A厂制造的硬币，“正面偏向性”的均值是25%，B厂制造的硬币，“正面偏向性”的均值是75%。
- 我们需要用如下图所示的先验概率分布来表达先验知识：



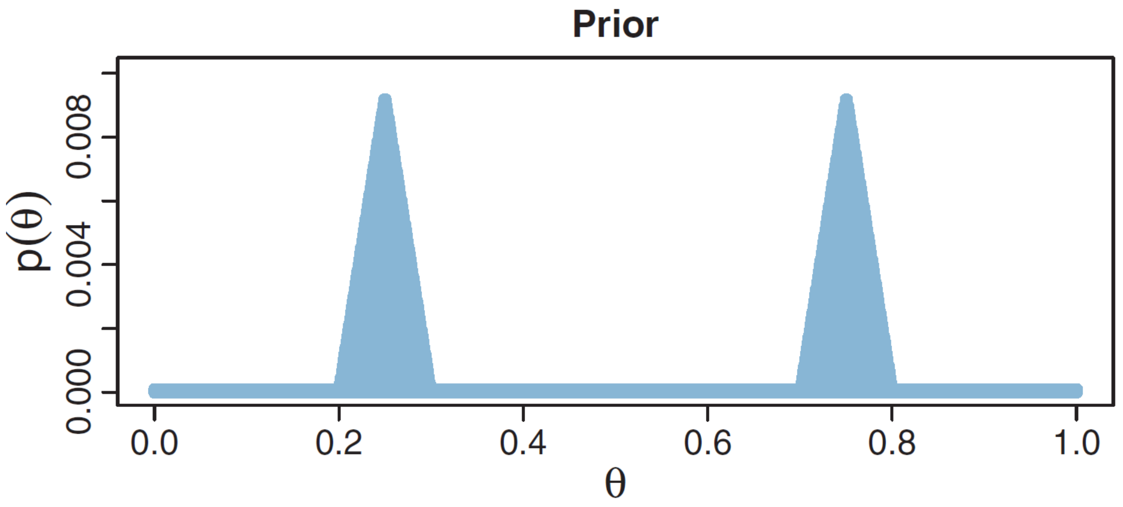
6.4.2 不可以用Beta分布表达的先验知识





- Beta分布没有双峰值的形式，无法表达该先验知识。

6.4.2 不可以用Beta分布表达的先验知识



- 网格近似是一种解决方法。

1. 先设置好θ和$p(\theta)$的离散值，比如：
   - $\theta = \{0.0, 0.001, 0.002, \ldots 0.999, 1.0\}$
   - $$P(\theta) = \begin{cases} \mathrm{seq}(1,100,50), & 0.2 \le \theta \le 0.25 \\ \mathrm{seq}(100,1,50), & 0.25 \le \theta \le 0.3 \\ \mathrm{seq}(1,100,50), & 0.7 \le \theta \le 0.75 \\ \mathrm{seq}(100,1,50), & 0.75 \le \theta \le 0.8 \\ 1, & \text{其他} \end{cases}$$
   - 对$P(\theta)$的值做归一化： $p(\theta) = P(\theta)/\mathrm{sum}\big(P(\theta)\big)$
2. 用第3章的bern_grid()函数，即可得到近似的后验概率分布。

伪代码

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

#遍历所有θ,计算$\theta^z(1-\theta^{1-z})$

```
p_D_given_theta = likelihood(theta, z, N)
```

#遍历所有θ,计算$p(D|\theta)p(\theta)$，然后求和

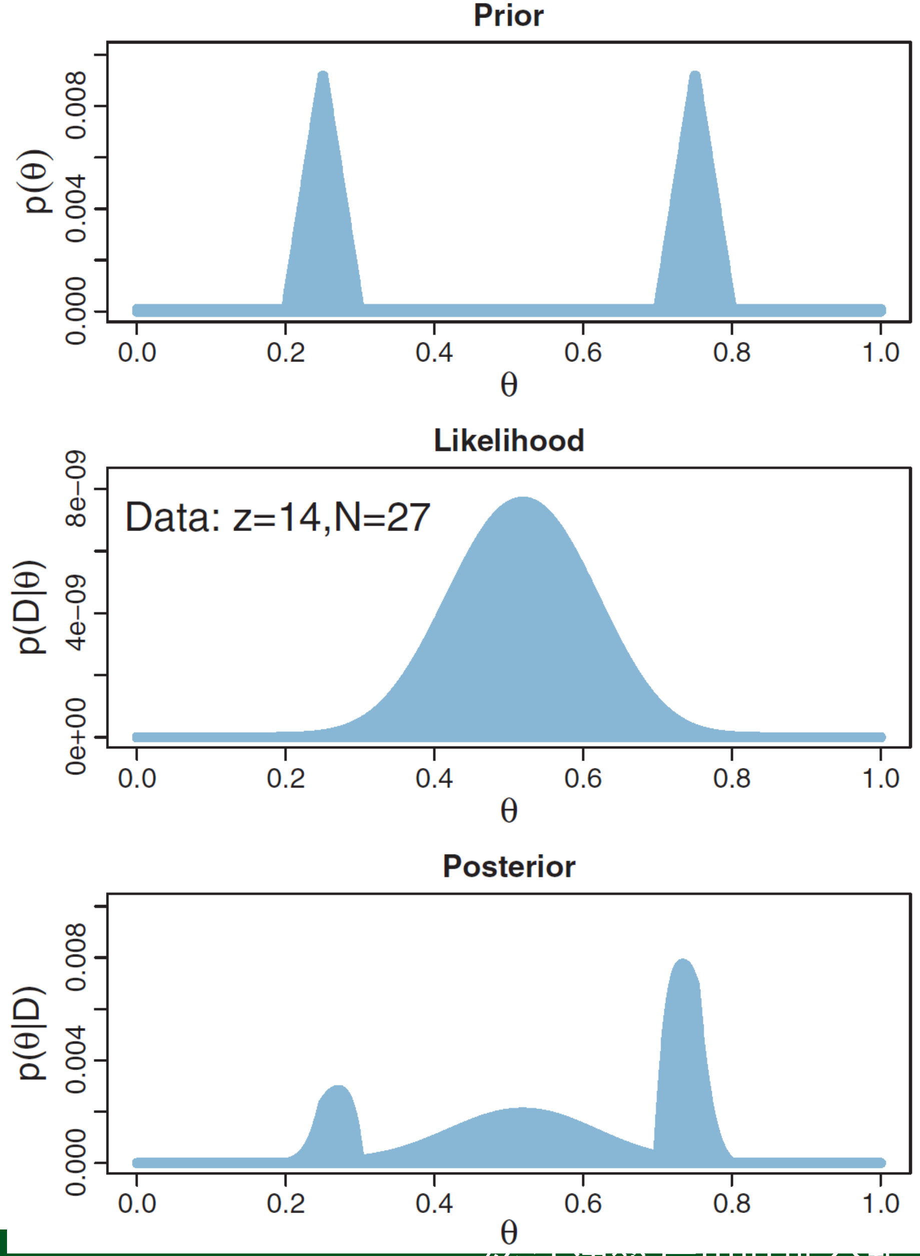
```
p_D = evidence(p_theta, p_D_given_theta)
```

#遍历所有θ,计算$p(D|\theta)p(\theta)/p(D)$

```
p_theta_given_D =
  posterior(p_D_given_theta, p_theta, p_D)
```

结果

- 先验：0.0至0.2等区域，概率密度值并不是0。
- 似然：$p(D|\theta = 0.75)$比$p(D|\theta = 0.25)$大很多。
- 后验：有3个峰值，$\theta = 0.75$是最高峰值。
  - 也是不能用Beta分布表达。



极大后验估计（Maximum a posteriori estimation，MAP）

- 极大似然估计：

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

- 极大后验估计：

$$\begin{aligned}\hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, p(\theta|D) \\ &= \underset{\theta}{\operatorname{argmax}}\, p(D|\theta)\, p(\theta) \\ &= \underset{\theta}{\operatorname{argmax}}\, \log p(D|\theta) + \log p(\theta)\end{aligned}$$

极大后验估计：伯努利分布

- 似然：

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

- 假设先验为Beta分布：

$$p(\theta) = \mathrm{Beta}(\theta|a,b)$$

- 极大后验估计：

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

极大后验估计：伯努利分布

$$f(\theta) = $$
$$z\log\theta + (N-z)\log(1-\theta) + (a-1)\log\theta + (b-1)\log(1-\theta)$$

- 我们求解上式的极值点：

$$f'(\theta) = \frac{z+a-1}{\theta} - \frac{N-z+b-1}{1-\theta} = 0$$

可得：$\theta = \frac{z+a-1}{N+b+a-2}$

另一方面，$LL''(\theta) < 0$

- 因此，$\hat{\theta}_{\mathrm{MAP}} = \frac{z+a-1}{N+b+a-2}$。

比较

	MLE	MAP	Bayesian Inference
$\hat{\theta}$	$\hat{\theta}_{\text{MLE}} = \frac{z}{N}$	$\hat{\theta}_{\text{MAP}} = \frac{z+a-1}{N+b+a-2}$	$\text{beta}(\theta \mid z+a, N-z+b)$
$p(\theta \mid D)$	$\delta(\mu - \hat{\mu}_{\text{MLE}})$	$\delta(\mu - \hat{\mu}_{\text{MAP}})$	$\text{beta}(\theta \mid z+a, N-z+b)$

此处，δ表示狄拉克函数（dirac delta function）。

后验预测（Posterior prediction）：

- 后验预测：

$$p(y'|D) = \int p(y', \theta|D)d\theta$$

$$= \int p(y'|\theta, D)p(\theta|D)d\theta$$

$$= \int p(y'|\theta)p(\theta|D)d\theta$$

- 对于点估计：

$$p(y'|D) = p(y'|\hat{\theta})$$

后验预测：伯努利分布

$$p(y' = 1|D) = \int p(y'|\theta)p(\theta|D)d\theta$$

$$= \int \theta \mathrm{beta}(\theta|a_D, b_D)d\theta$$

$$= E[\theta] = \frac{a_D}{a_D + b_D}$$

- 其中，$a_D = z + a$，$b_D = N - z + b$。

后验预测：伯努利分布

$$p(y' = 1|D) = \frac{a_D}{a_D + b_D}$$

- 可得：

$$p(y' = 0|D) = \frac{b_D}{a_D + b_D}$$

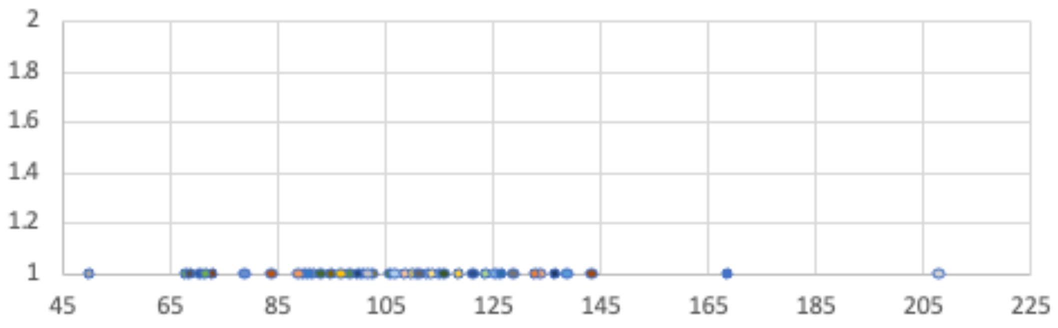
合并上面两式子，可得：

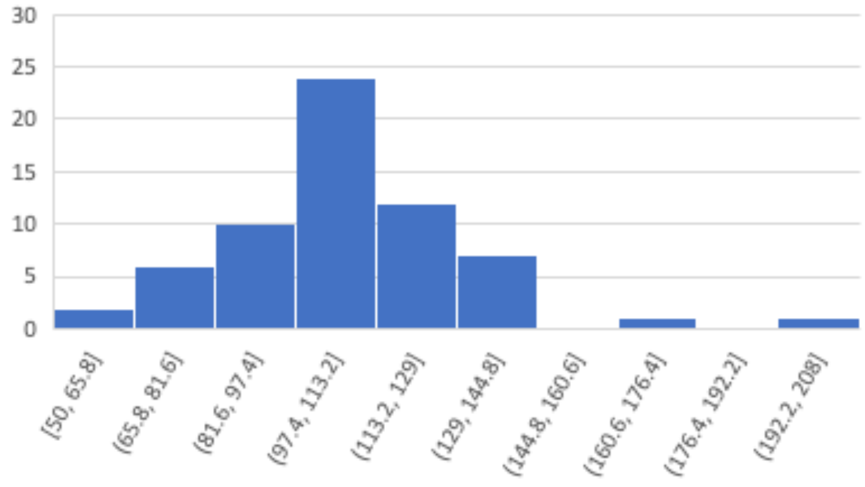
$$p(y'|D) = \theta'^{y'}(1 - \theta')^{1-y'}$$

- 其中，$\theta' = \frac{a_D}{a_D+b_D}$。

贝叶斯推理：高斯分布
第1步：确定和问题相关的数据

- 测试“聪明药”是否能使人聪明。
- 下图是63个人吃了“聪明药”以后的IQ数据。
- 人类的IQ均值是100，标准差是15。

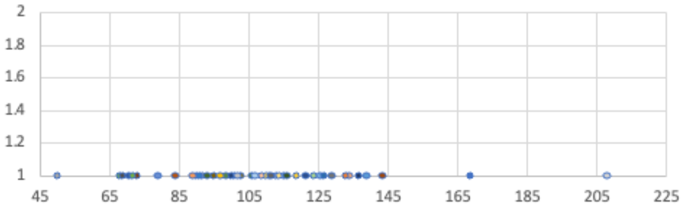


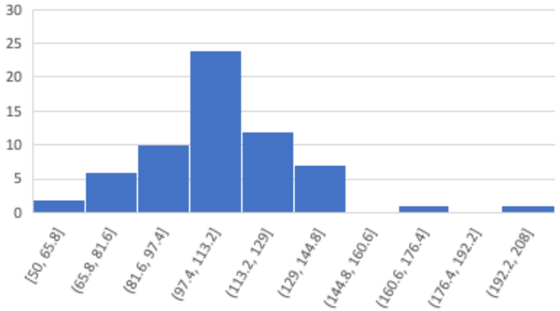


(数据来源：DBDA)

第2步：确定适合数据的模型和相应的参数

- 下图的数据看起来像高斯分布，我们可以假设似然服从高斯分布：$p(D|\theta) = N(IQ|\mu,\sigma^2)$。
- 参数是均值μ和方差σ^2。
- 通过估计参数μ和σ^2，和人类的正常值（均值100，标准差15）来比较和分析。
    - 比如估计出来的μ大于100，而且95% HDI区间的下界大于100，我们就有理由相信“聪明药”有效果。





居来源：DBDA)

第2步：确定适合数据的模型和相应的参数

- 高斯分布的概率密度函数：

$$p(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

- 假设N个人之间的IQ是相互独立的，则似然为：

$$p(D|\mu,\sigma^2) = \prod_i p(x_i|\mu,\sigma^2)$$

第2步：确定适合数据的模型和相应的参数

$$p(D|\mu,\sigma^2)=\prod_i p(x_i|\mu,\sigma^2)$$

$$=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N\prod_i \exp\left[-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right]$$

$$=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}\exp\left[-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right]$$

$$p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

多参数估计

- 对于多参数的估计：

1. 可以单独估计一个参数，假设其他参数已知。
   - 比如对于高斯分布，假设σ^2已知，估计μ的后验$p(\mu|D)$。
     - 实践中，可以指定σ^2 =数据的方差。
2. 可以同时估计所有参数的联合概率分布。
   - 比如对于高斯分布，估计$p(\mu, \sigma^2|D)$。

- 在这里，我们用第一种方法，第二种方法作为阅读材料。

第3步：给要估计的参数指定一个先验

- 假设σ^2已知，估计μ。
- 观察似然的形式：

$$p(D|\mu,\sigma^2)=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}\exp\left[-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right]$$

- 形式上是$\frac{1}{a}\exp[-\frac{1}{b}(\mu-c)^2]$，
- 和高斯分布的概率密度函数$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$，形式上一致。
- μ的共轭先验是高斯分布。

第3步：给要估计的参数指定一个先验

- 指定μ的共轭先验：

$$\mu \sim N(\mu | \mu_0, \sigma_0^2)$$

$$p(\mu) = N(\mu | \mu_0, \sigma_0^2) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[-\frac{1}{2}\frac{(\mu - \mu_0)^2}{\sigma_0^2}\right]$$

第4步：求解后验概率分布

$$p(D|\mu,\sigma^2)=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}\exp\left[-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right]$$

$$p(\mu)=N(\mu|\mu_0,\sigma_0^2)=\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left[-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right]$$

- 我们使用“简化推导”，将常数省去，最后做归一化：

$$p(\mu|D)\propto p(D|\mu)p(\mu)$$

$$\propto\exp\left[-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right]\exp\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$

第4步：求解后验概率分布

- 我们使用“简化推导”，将常数省去，最后做归一化：

$$p(\mu|D) \propto p(D|\mu)p(\mu)$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right]\exp\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_i (\mu^2-2\mu x_i)-\frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0)\right]$$

第4步：求解后验概率分布

- 我们使用“简化推导”，将常数省去，最后做归一化：

$$p(\mu|D) \propto p(D|\mu)p(\mu)$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right]\exp\left[-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right]$$

$$\propto \exp\left[-\frac{1}{2\sigma^2}\sum_i (\mu^2-2\mu x_i)-\frac{1}{2\sigma_0^2}(\mu^2-2\mu\mu_0)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\left(N\mu^2-2\mu\sum_i x_i\right)+\frac{1}{\sigma_0^2}(\mu^2-2\mu\mu_0)\right)\right]$$

第4步：求解后验概率分布

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\left(N\mu^2 - 2\mu\sum_i x_i\right) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0)\right)\right]$$

第4步：求解后验概率分布

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\left(N\mu^2 - 2\mu\sum_i x_i\right) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0)\right)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(N\mu^2 - 2\mu N\bar{x}) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0)\right)\right]$$

第4步：求解后验概率分布

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\left(N\mu^2 - 2\mu\sum_i x_i\right) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0)\right)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(N\mu^2 - 2\mu N\bar{x}) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0)\right)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{\mu_0}{\sigma_0^2} + \frac{N\bar{x}}{\sigma^2}\right)\mu\right)\right]$$

第4步：求解后验概率分布

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\left(N\mu^2 - 2\mu\sum_i x_i\right) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0)\right)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}(N\mu^2 - 2\mu N\bar{x}) + \frac{1}{\sigma_0^2}(\mu^2 - 2\mu\mu_0)\right)\right]$$

$$\propto \exp\left[-\frac{1}{2}\left(\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\mu^2 - 2\left(\frac{\mu_0}{\sigma_0^2} + \frac{N\bar{x}}{\sigma^2}\right)\mu\right)\right]$$

$$\propto \exp\left[-\frac{1}{2\hat{\sigma}^2}(\mu - \hat{\mu})^2\right]$$

- 其中，$\bar{x} = \frac{\sum_i x_i}{N}$。

第4步：求解后验概率分布

$$\propto \exp\left[-\frac{1}{2}\left(\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\frac{\mu_0}{\sigma_0^2}+\frac{N\bar{x}}{\sigma^2}\right)\mu\right)\right]$$

$$\propto \exp\left[-\frac{1}{2\hat{\sigma}^2}(\mu-\hat{\mu})^2\right]$$

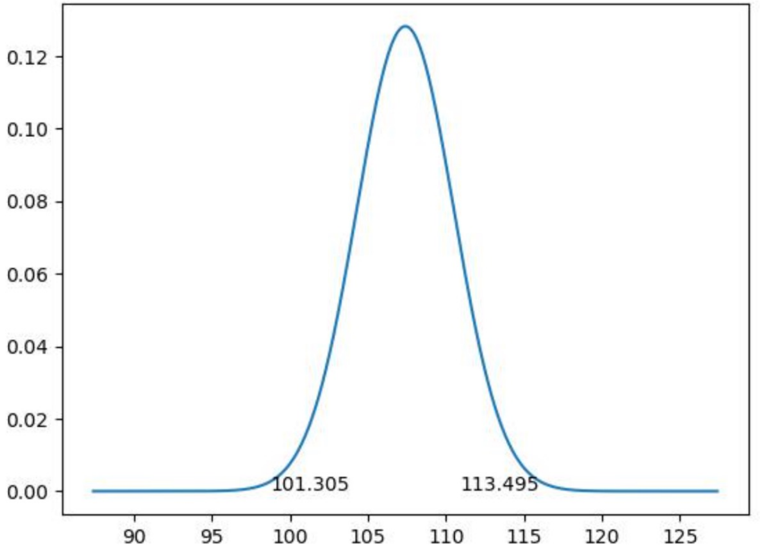
- $p(\mu|D)=N(\mu|\hat{\mu},\hat{\sigma}^2)$，其中，
- $\hat{\sigma}^2=\frac{\sigma^2\sigma_0^2}{\sigma^2+N\sigma_0^2}$
- $\hat{\mu}=\hat{\sigma}^2\left(\frac{\mu_0}{\sigma_0^2}+\frac{N\bar{x}}{\sigma^2}\right)=\frac{\sigma^2}{\sigma^2+N\sigma_0^2}\mu_0+\frac{N\sigma_0^2}{\sigma^2+N\sigma_0^2}\bar{x}$

回到“聪明药”

- $p(\mu|D) = N(\mu|\hat{\mu}, \hat{\sigma}^2)$
- $\hat{\sigma}^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + N\sigma_0^2}$
- $\hat{\mu} = \frac{\sigma^2}{\sigma^2 + N\sigma_0^2}\mu_0 + \frac{N\sigma_0^2}{\sigma^2 + N\sigma_0^2}\bar{x}$
- 根据全人类的IQ，假设$\mu_0 = 100$， $\sigma_0^2 = 225$。
- 根据吃“聪明药”人的IQ方差是637，指定$\sigma^2 = 637$。
- $\hat{\sigma}^2 = \frac{637*225}{637+63*225} = 9.68$
- $\hat{\mu} = \frac{637}{637+63*225} * 100 + \frac{63*225}{637+63*225} * 107.8 = 107.4$

“聪明药”

- 后验：$p(\mu|D) = N(\mu|107.4, 9.68)$，如下图所示。
- 95% HDI：[101.305, 113.495]。
- 均值是107.4，“聪明药”看起来是提高了IQ。
- HDI区间的下界101.3和100比较接近，效果不是那么明显。



假设μ已知，估计σ^2

- 观察似然关于σ^2的形式：

$$p(D|\mu,\sigma^2)=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}\exp\left[-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right]$$

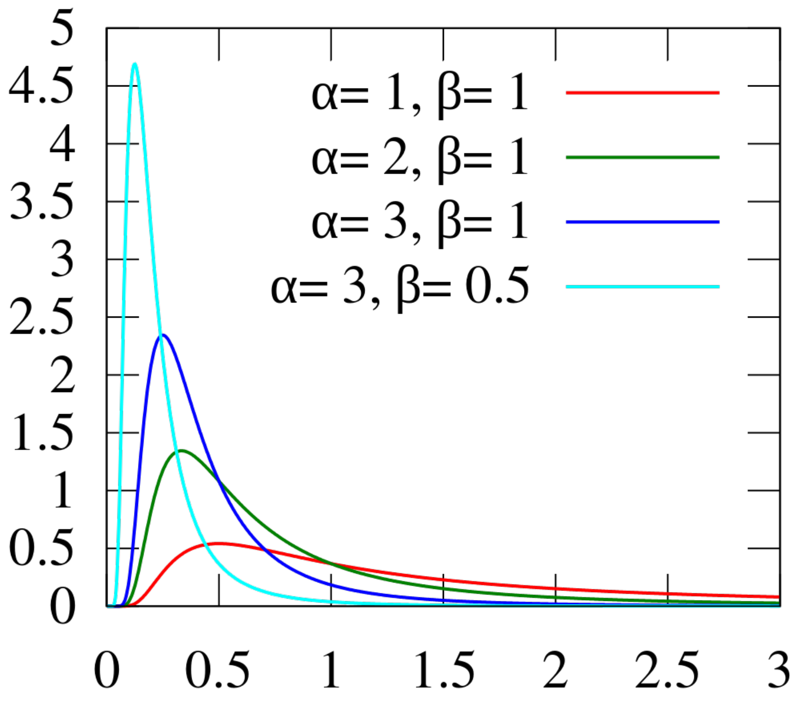
- 形式上是$(\sigma^2)^{-a}\exp[-b\,\frac{1}{\sigma^2}]$。
- 和高斯分布$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$，形式上不一致。

反向Gamma分布

- 反向Gamma分布：

$$\mathrm{IG}(x|a,b)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$$

- 其中，$x>0, a>0, b>0$。
- 类似Beta分布，$\Gamma(a)$也是归一化常数，称为Gamma函数。
- 峰值：$\frac{b}{a+1}$



形式对比

- σ^2的形式：

$$(\sigma^2)^{-a}\exp[-b\frac{1}{\sigma^2}]$$

- 反向Gamma分布：

$$\mathrm{IG}(x|a,b)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$$

- 反向Gamma分布在形式上和σ^2的形式一致。
- 反向Gamma分布是σ^2的共轭先验。

求解σ^2的后验

$$\mathrm{IG}(x|a,b)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$$

中山大學
SUN YAT-SEN UNIVERSITY

- 假设先验$p(\sigma^2)=\mathrm{IG}(\sigma^2|a_0,b_0)$。
- 仍然省去常数，最后做归一化：

$$p(\sigma^2|D)\propto p(D|\sigma^2)p(\sigma^2)$$

求解σ^2的后验

$$\mathrm{IG}(x|a,b)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$$

中山大學 SUN YAT-SEN UNIVERSITY

- 假设先验$p(\sigma^2)=\mathrm{IG}(\sigma^2|a_0,b_0)$。
- 仍然省去常数，最后做归一化：

$$p(\sigma^2|D)\propto p(D|\sigma^2)p(\sigma^2)$$

$$=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)}e^{-\frac{b_0}{\sigma^2}}|$$

求解σ^2的后验

$$\mathrm{IG}(x|a,b)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$$

中山大學 SUN YAT-SEN UNIVERSITY

- 假设先验$p(\sigma^2)=\mathrm{IG}(\sigma^2|a_0,b_0)$。
- 仍然省去常数，最后做归一化：

$$p(\sigma^2|D)\propto p(D|\sigma^2)p(\sigma^2)$$

$$=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)}e^{-\frac{b_0}{\sigma^2}}$$

$$\propto(\sigma^2)^{-\frac{N}{2}}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}(\sigma^2)^{-(a_0+1)}e^{-\frac{b_0}{\sigma^2}}$$

求解σ^2的后验

$$\mathrm{IG}(x|a,b)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$$

- 假设先验$p(\sigma^2)=\mathrm{IG}(\sigma^2|a_0,b_0)$。
- 仍然省去常数，最后做归一化：

$$p(\sigma^2|D)\propto p(D|\sigma^2)p(\sigma^2)$$

$$=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)}e^{-\frac{b_0}{\sigma^2}}$$

$$\propto(\sigma^2)^{-\frac{N}{2}}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}(\sigma^2)^{-(a_0+1)}e^{-\frac{b_0}{\sigma^2}}$$

$$\propto(\sigma^2)^{-\left(\frac{N}{2}+a_0+1\right)}e^{-\left(\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2+\frac{b_0}{\sigma^2}\right)}$$

求解σ^2的后验

$$\mathrm{IG}(x|a,b)=\frac{b^a}{\Gamma(a)}x^{-(a+1)}e^{-\frac{b}{x}}$$

中山大學 SUN YAT-SEN UNIVERSITY

- 假设先验$p(\sigma^2)=\mathrm{IG}(\sigma^2|a_0,b_0)$。
- 仍然省去常数，最后做归一化：

$$p(\sigma^2|D)\propto p(D|\sigma^2)p(\sigma^2)$$

$$=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma^2)^{-(a_0+1)}e^{-\frac{b_0}{\sigma^2}}|$$

$$\propto(\sigma^2)^{-\frac{N}{2}}e^{-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2}(\sigma^2)^{-(a_0+1)}e^{-\frac{b_0}{\sigma^2}}$$

$$\propto(\sigma^2)^{-\left(\frac{N}{2}+a_0+1\right)}e^{-\left(\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2+\frac{b_0}{\sigma^2}\right)}$$

$$\propto(\sigma^2)^{-\left(\frac{N}{2}+a_0+1\right)}e^{-\left(\frac{\frac{1}{2}\sum_i(x_i-\mu)^2+b_0}{\sigma^2}\right)}$$

求解σ^2的后验

$$p(\sigma^2|D) \propto (\sigma^2)^{-\left(\frac{N}{2}+a_0+1\right)} e^{-\left(\frac{\frac{1}{2}\sum_i (x_i-\mu)^2+b_0}{\sigma^2}\right)}$$

- 对比反向Gamma分布：

$$\mathrm{IG}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{-(a+1)} e^{-\frac{b}{x}}$$

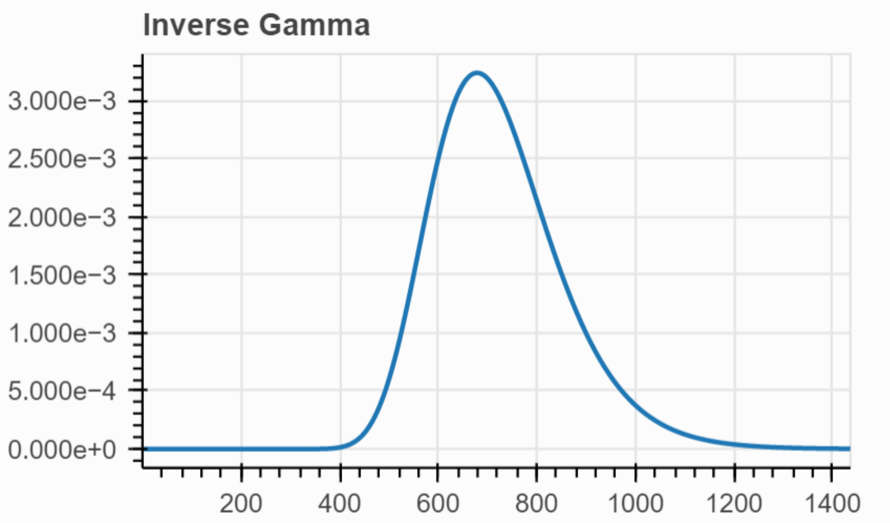
- 可得， $p(\sigma^2|D)$服从反向Gamma分布，$\mathrm{IG}(\sigma^2|\hat{a},\hat{b})$。
- $\hat{a} = a_0 + \frac{N}{2}$
- $\hat{b} = b_0 + \frac{1}{2}\sum_i (x_i - \mu)^2$

回到“聪明药”

- $p(\sigma^2|D) = \mathrm{IG}(\sigma^2|\hat{a}, \hat{b})$
- $\hat{a} = a_0 + \frac{N}{2}$
- $\hat{b} = b_0 + \frac{1}{2}\sum_i (x_i - \mu)^2$
- 根据样本IQ均值是107.84，指定$\mu = 107.84$。
- 我们设定一个对后验影响小的先验：$a_0 = 0, b_0 = 0$。
- 根据上式计算得到：$\hat{a} = 31.5$，$\hat{b} = 22071.2$
- 后验：$p(\sigma^2|D) = \mathrm{IG}(\sigma^2|31.5{,}22071.2)$

回到“聪明药”

- 后验：$p(\sigma^2|D) = \mathrm{IG}(\sigma^2|31.5,22071.2)$
- 峰值：$\frac{b}{a+1} = 679.1$
- 吃了“聪明药”的IQ标准差为26.1，大于人类的25。
  - “聪明药”可能有副作用，让一些人提高IQ，让一些人降低IQ。



极大似然估计（MLE）：高斯分布

$$p(D|\mu,\sigma^2) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N} \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right]$$

- 求解$\underset{\mu}{\operatorname{argmin}} -\log p(D|\mu)$，可得：

$$\hat{\mu}_{\mathrm{MLE}} = \bar{x}$$

极大后验估计 (Maximum A Posteriori Estimation, MAP)

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

极大后验估计（MAP）：高斯分布

- 似然：$p(D|\mu,\sigma^2)=\frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^N}\exp\left[-\frac{1}{2\sigma^2}\sum_i(x_i-\mu)^2\right]$
- 先验：　$\mu\sim N(\mu|\mu_0,\sigma_0^2)$
- 求解$\underset{\mu}{\operatorname{argmin}}-\log p(D|\mu)-\log p(\mu)$，可得：

$$\hat{\mu}_{\mathrm{MAP}}=\frac{\sigma^2}{\sigma^2+N\sigma_0^2}\mu_0+\frac{N\sigma_0^2}{\sigma^2+N\sigma_0^2}\bar{x}$$

比较

	MLE	MAP	Bayesian Inference
$\hat{\theta}$	$\hat{\mu}_{\text{MLE}} = \bar{x}$	$\hat{\mu}_{\text{MAP}} = \alpha\mu_0 + (1-\alpha)\bar{x}$	$N(\mu\|\hat{\mu},\hat{\sigma}^2)$ $\hat{\mu} = \hat{\mu}_{\text{MAP}}$ $\hat{\sigma}^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2+N\sigma_0^2}$
$p(\theta\|D)$	$\delta(\mu - \hat{\mu}_{\text{MLE}})$	$\delta(\mu - \hat{\mu}_{\text{MAP}})$	$N(\mu\|\hat{\mu},\hat{\sigma}^2)$

此处，δ表示狄拉克函数（dirac delta function）。

The distribution zoo

https://ben18785.shinyapps.io/distribution-zoo/

*同时估计所有参数（以下均为阅读材料）

中山大學 SUN YAT-SEN UNIVERSITY

- 对于多参数的估计：

1. 可以单独估计一个参数，假设其他参数已知。
   - 比如对于高斯分布，假设σ^2已知，估计μ的后验$p(\mu|D)$。
     - 实践中，可以指定σ^2 =数据的方差。
2. 可以同时估计所有参数的联合概率分布。
   - 比如对于高斯分布，估计$p(\mu,\sigma^2|D)$。

正态分布---精度（precision）

- 高斯分布的概率密度函数：

$$p(x)=\sqrt{\frac{\lambda}{2\pi}}\exp\left[-\frac{\lambda}{2}(x-\mu)^2\right]$$

- 其中，$\lambda=(\sigma^2)^{-1}$，被称为精度（precision），表示的是“值有多集中在均值”， λ越大越集中。
  - σ表示的是“值有多分散”。

Gamma分布

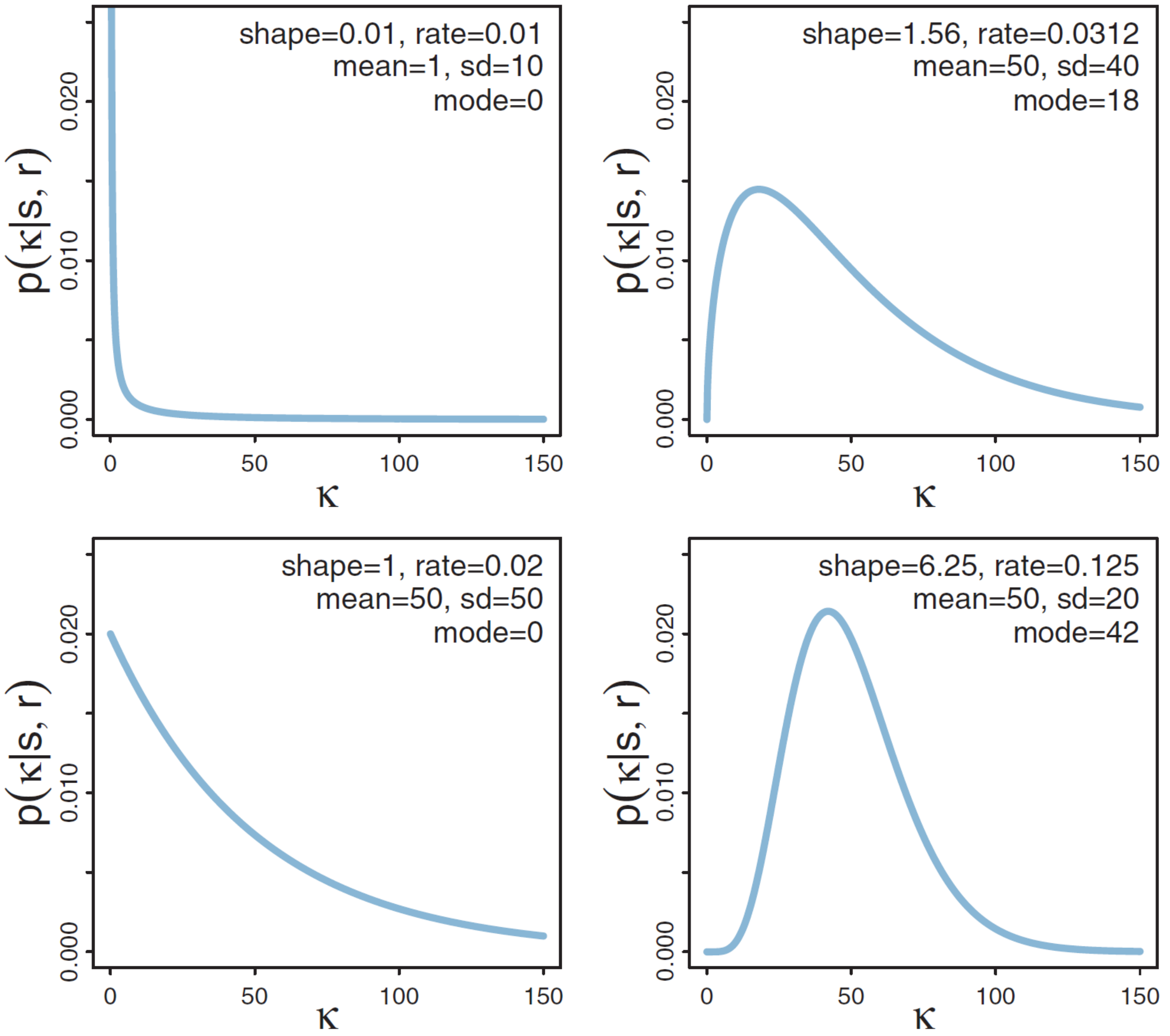
$$\mathrm{Gamma}(x|s,r)=\frac{r^s}{\Gamma(s)}x^{s-1}e^{-rx}$$

- 其中，$x\geq 0, s>0, r>0$。
- $\Gamma(s)=\int_0^\infty t^{s-1}e^{-t}dt$，被称为Gamma函数。
- s是形状参数（shape），r是率参数（rate），r又被称为反向尺度参数（inverse scale parameter）。

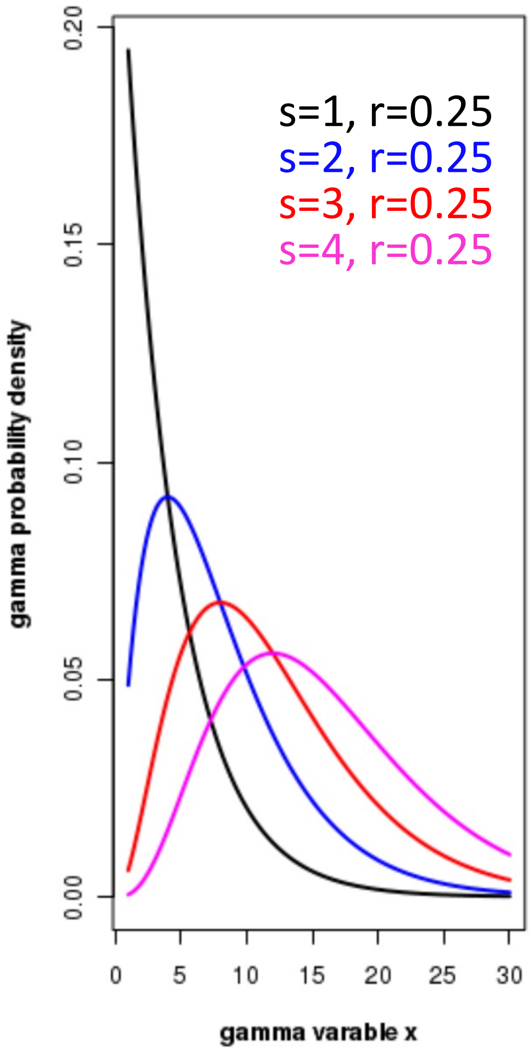
- Gamma分布可以被用作很多似然的共轭先验，比如正态分布的精度，指数分布等。

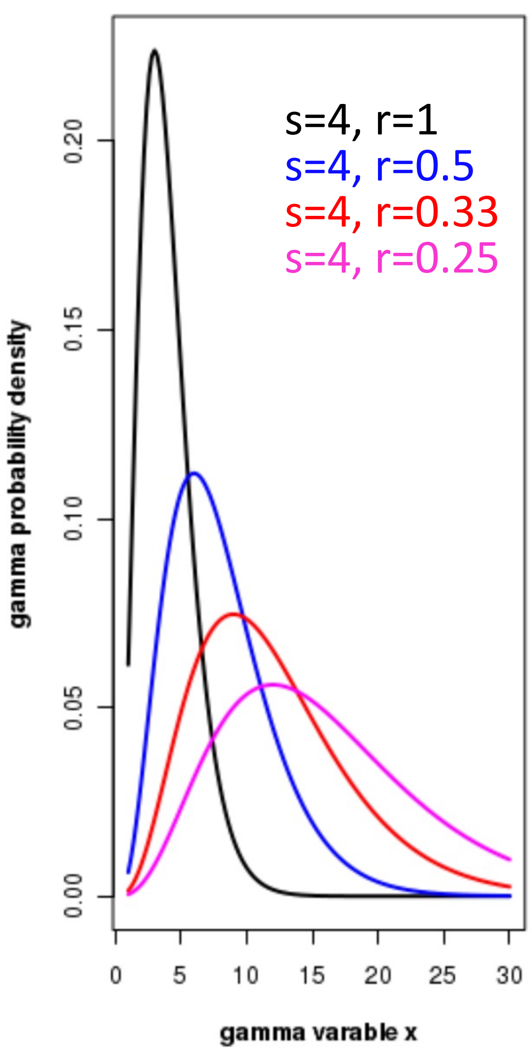
Gamma分布

$$\mathrm{Gamma}(x|s,r) = \frac{r^s}{\Gamma(s)} x^{s-1} e^{-rx}$$



Gamma分布





（图片来源：http://www.countbio.com/web_pages/left_object/R_for_biology/R_biostatistics_part-1/gamma_distribution.html）

Normal-Gamma先验

$$p(D|\mu,\lambda)=\left(\sqrt{\frac{\lambda}{2\pi}}\right)^N \exp\left[-\frac{\lambda}{2}\sum_i (x_i-\mu)^2\right]$$

- 其中，$\lambda=(\sigma^2)^{-1}$。
- Normal-Gamma prior:

$$p(\mu,\lambda)=p(\mu|\lambda)p(\lambda)$$

假设$p(\mu|\lambda)\sim \mathrm{N}(\mu|\mu_0,(\kappa_0\lambda)^{-1}), p(\lambda)\sim \mathrm{Gamma}(\lambda|\alpha_0,\beta_0)$：

$$\mathrm{NG}(\mu,\sigma^2|\mu_0,\kappa_0,\alpha_0,\beta_0)\triangleq N(\mu|\mu_0,(\kappa_0\lambda)^{-1})\mathrm{G}a(\sigma^2|a_0,b_0)$$

Normal-Gamma先验

$$
\begin{aligned}
NG(\mu, \lambda | \mu_0, \kappa_0, \alpha_0, \beta_0) &\overset{\text{def}}{=} \mathcal{N}(\mu | \mu_0, (\kappa_0 \lambda)^{-1}) Ga(\lambda | \alpha_0, \text{rate} = \beta_0) \\
&= \frac{1}{Z_{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0)} \lambda^{\frac{1}{2}} \exp(-\frac{\kappa_0 \lambda}{2}(\mu - \mu_0)^2) \lambda^{\alpha_0 - 1} e^{-\lambda \beta_0} \\
&= \frac{1}{Z_{NG}} \lambda^{\alpha_0 - \frac{1}{2}} \exp\left(-\frac{\lambda}{2}\left[\kappa_0 (\mu - \mu_0)^2 + 2\beta_0\right]\right) \\
Z_{NG}(\mu_0, \kappa_0, \alpha_0, \beta_0) &= \frac{\Gamma(\alpha_0)}{\beta_0^{\alpha_0}} \left(\frac{2\pi}{\kappa_0}\right)^{\frac{1}{2}}
\end{aligned}
$$

后验

$$
\begin{aligned}
p(\mu, \lambda | D) &\propto NG(\mu, \lambda | \mu_0, \kappa_0, \alpha_0, \beta_0) p(D | \mu, \lambda) \\
&\propto \lambda^{\frac{1}{2}} e^{-(\kappa_0 \lambda (\mu - \mu_0)^2)/2} \lambda^{\alpha_0 - 1} e^{-\beta_0 \lambda} \times \lambda^{n/2} e^{-\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2} \\
&\propto \lambda^{\frac{1}{2}} \lambda^{\alpha_0 + n/2 - 1} e^{-\beta_0 \lambda} e^{-(\lambda/2)[\kappa_0 (\mu - \mu_0)^2 + \sum_i (x_i - \mu)^2]}
\end{aligned}
$$

- 把μ相关的整理出来，并配平方：

$$
\begin{aligned}
\kappa_0 (\mu - \mu_0)^2 + \sum_i (x_i - \mu)^2 &= \kappa_0 (\mu - \mu_0)^2 + n(\mu - \overline{x})^2 + \sum_i (x_i - \overline{x})^2 \\
&= (\kappa_0 + n)(\mu - \mu_n)^2 + \frac{\kappa_0 n (\overline{x} - \mu_0)^2}{\kappa_0 + n} + \sum_i (x_i - \overline{x})^2
\end{aligned}
$$

- 其中，$\mu_n = \dfrac{\kappa_0 \mu_0 + n\overline{x}}{\kappa_0 + n}$

后验

$$
\begin{aligned}
p(\mu,\lambda|D) &\propto NG(\mu,\lambda|\mu_0,\kappa_0,\alpha_0,\beta_0)p(D|\mu,\lambda) \\
&\propto \lambda^{\frac{1}{2}} e^{-(\kappa_0\lambda(\mu-\mu_0)^2)/2}\lambda^{\alpha_0-1}e^{-\beta_0\lambda} \times \lambda^{n/2}e^{-\frac{\lambda}{2}\sum_{i=1}^n (x_i-\mu)^2} \\
&\propto \lambda^{\frac{1}{2}}\lambda^{\alpha_0+n/2-1}e^{-\beta_0\lambda}e^{-(\lambda/2)[\kappa_0(\mu-\mu_0)^2+\sum_i (x_i-\mu)^2]}
\end{aligned}
$$

$$
\begin{aligned}
\kappa_0(\mu-\mu_0)^2+\sum_i (x_i-\mu)^2 &= \kappa_0(\mu-\mu_0)^2+n(\mu-\overline{x})^2+\sum_i (x_i-\overline{x})^2 \\
&= (\kappa_0+n)(\mu-\mu_n)^2+\frac{\kappa_0 n(\overline{x}-\mu_0)^2}{\kappa_0+n}+\sum_i (x_i-\overline{x})^2
\end{aligned}
$$

- 对比 $p(\mu) \propto \lambda^{\frac{1}{2}} \exp\left[-\frac{\lambda}{2}(\mu-\mu_n)^2\right]$ $\mathrm{Gamma}(\lambda|s,r) \propto \lambda^{s-1}e^{-r\lambda}$ ，可得：

$$
\begin{aligned}
p(\mu,\lambda|D) &\propto \lambda^{\frac{1}{2}} e^{-(\lambda/2)(\kappa_0+n)(\mu-\mu_n)^2} \\
&\quad \times \lambda^{\alpha_0+n/2-1}e^{-\beta_0\lambda}e^{-(\lambda/2)\sum_i (x_i-\overline{x})^2}e^{-(\lambda/2)\frac{\kappa_0 n(\overline{x}-\mu_0)^2}{\kappa_0+n}} \\
&\propto \mathcal{N}(\mu|\mu_n,((\kappa_0+n)\lambda)^{-1}) \times Ga(\lambda|\alpha_0+n/2,\beta_n)
\end{aligned}
$$

- 其中，$\beta_n = \beta_0+\frac{1}{2}\sum_{i=1}^n (x_i-\overline{x})^2+\frac{\kappa_0 n(\overline{x}-\mu_0)^2}{2(\kappa_0+n)}$