浙江大学 心理与行为科学系
DEPARTMENT OF PSYCHOLOGY AND BEHAVIORAL SCIENCES
ZHEJIANG UNIVERSITY

3. 贝叶斯统计学原理

戴俊毅

研究员/长聘副教授

先验信念

- 先验信念是指我们在考虑数据之前，对于客观世界的看法
- 对于客观世界的看法，可以由模型参数的概率分布来表达
- 此处的概率分布，代表的是主观信念意义上的可能性的分布

先验概率分布的设定

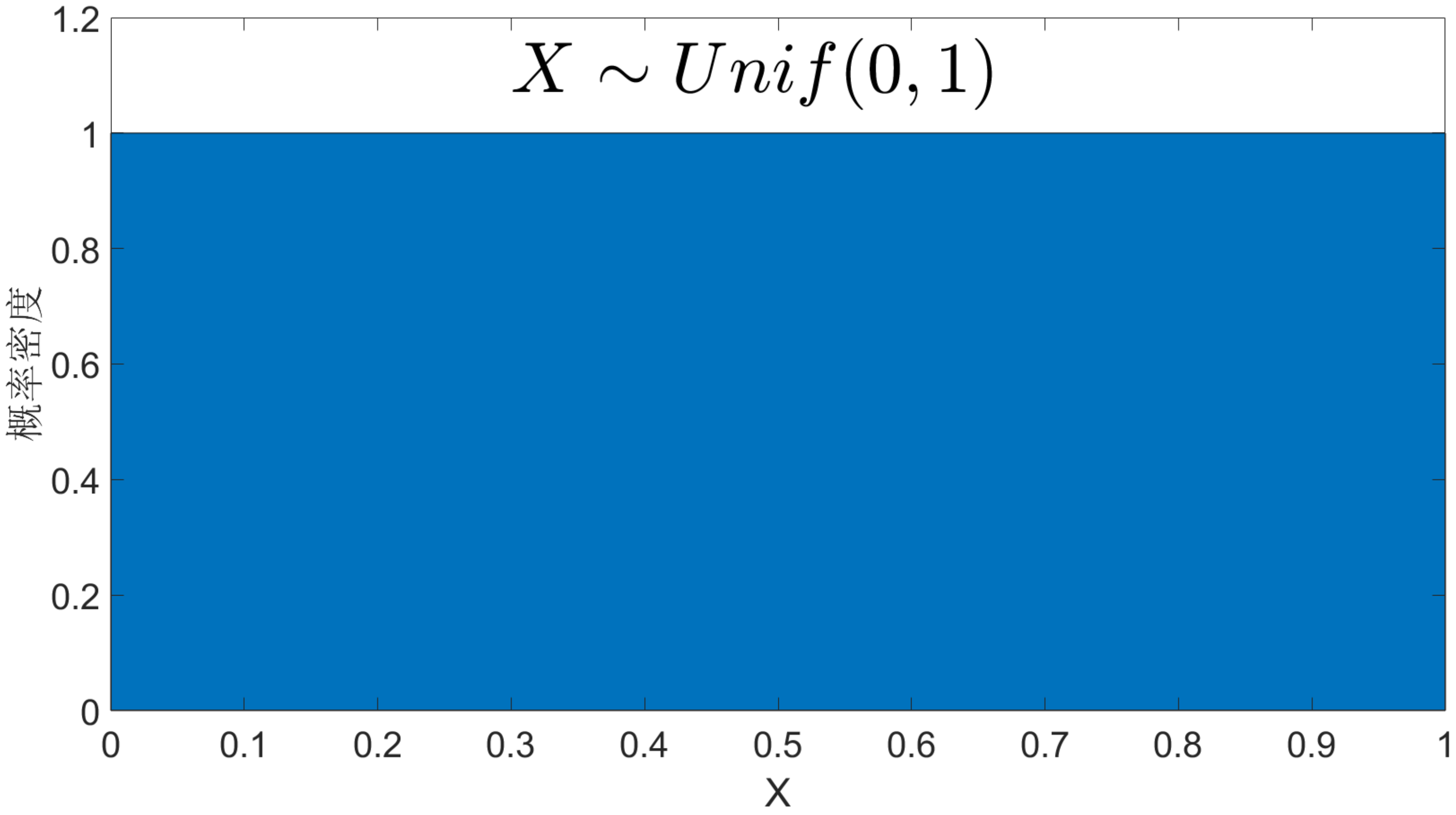
- 无信息先验分布(noninformative prior distribution）
- 有信息先验分布(informative prior distribution)

无信息先验分布

- 无信息先验分布是对后验分布影响最小的先验分布，当使用此类分布时，后验分布的形态几乎完全由数据决定
- 无信息先验分布也被称为参考先验分布(reference prior distribution)
- 无信息先验一般被描述为模糊的(vague)，平坦的(flat)或者弥散的(diffuse)

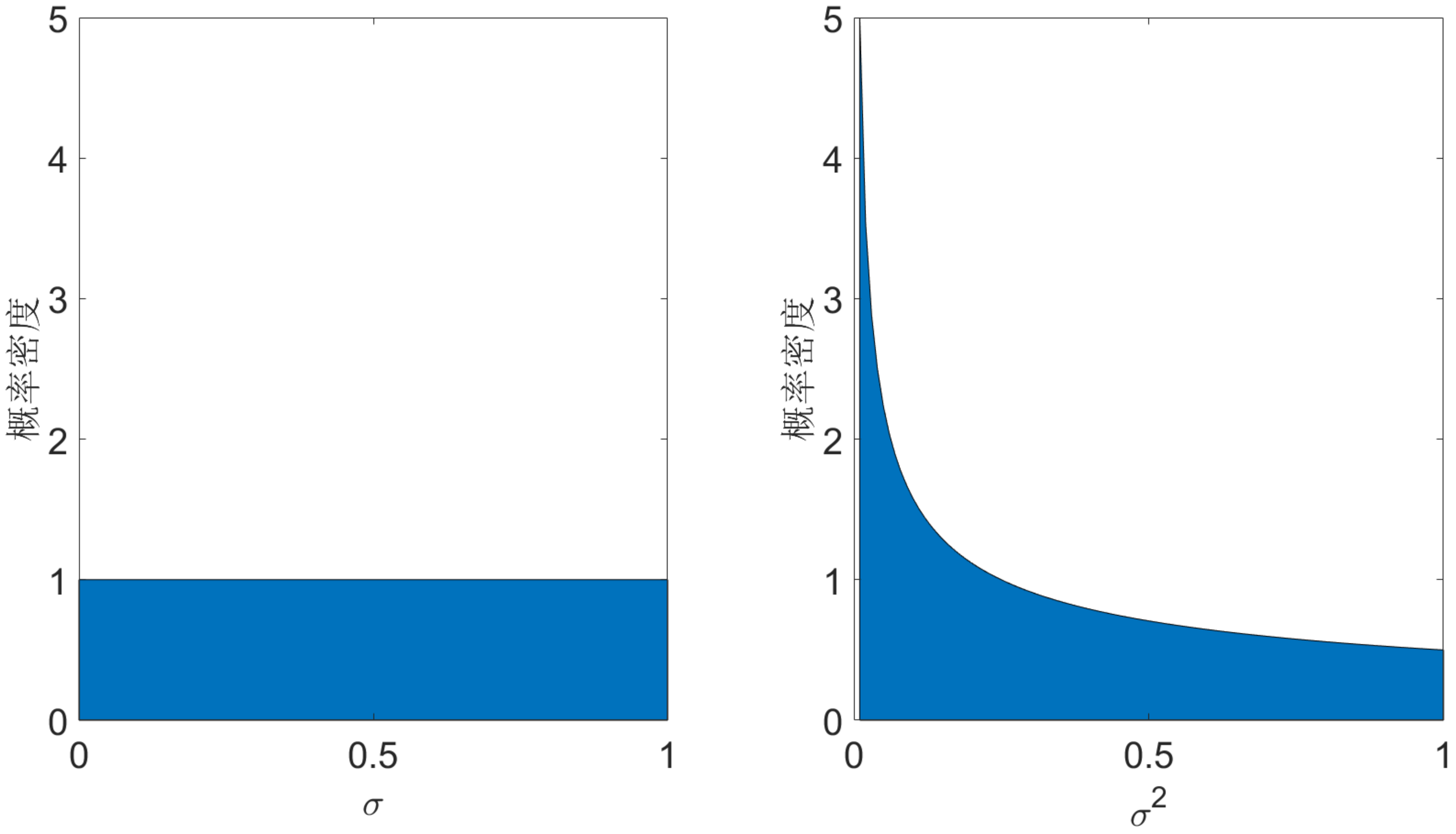
无信息先验分布实例

- 当模型参数代表占比或者概率时，常用的无信息先验是U(0,1)



无信息先验分布设定问题

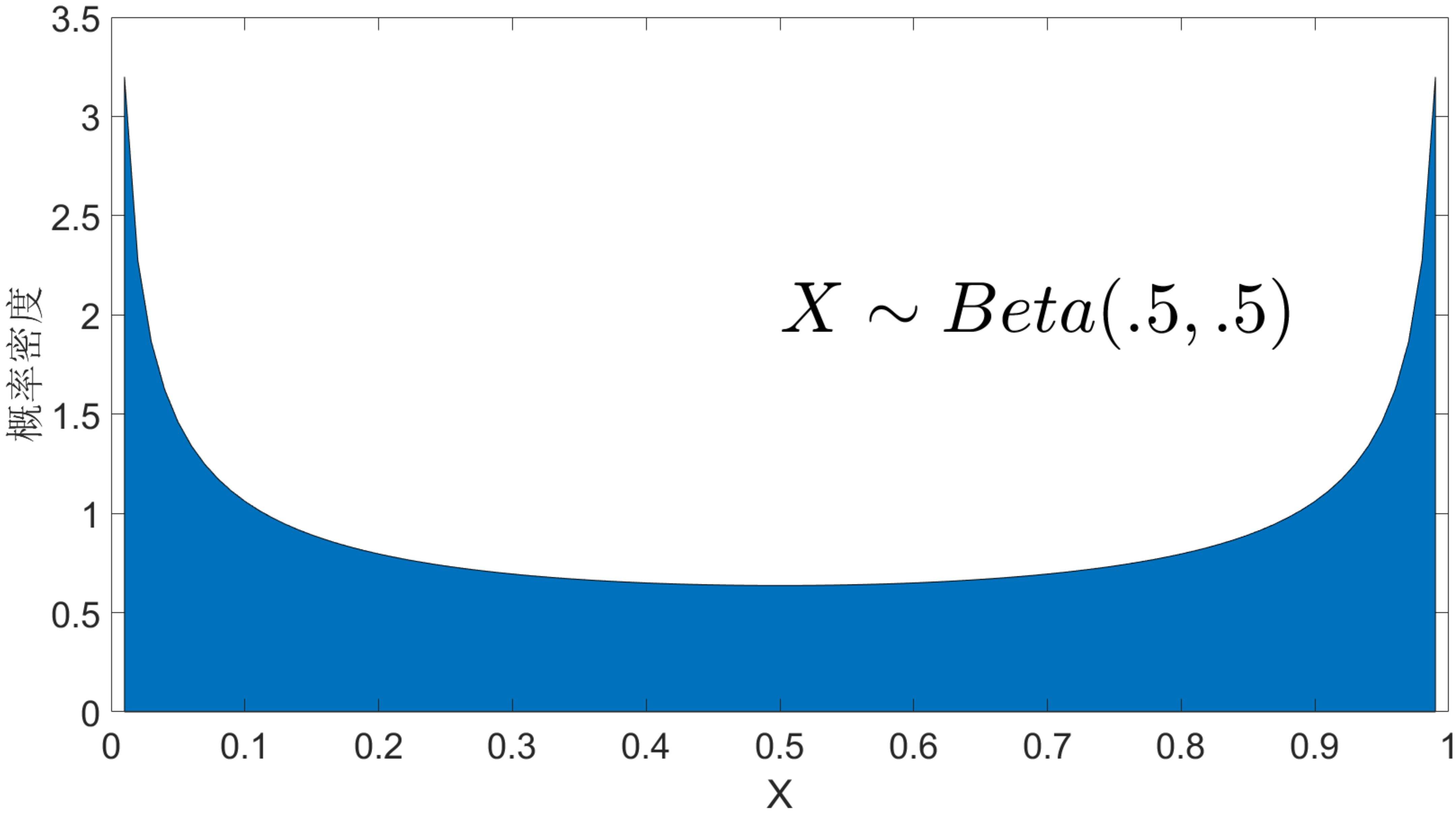
- “平坦”的分布并非一定是无信息的，或者无偏向的
- 例如，当假定正态分布的标准差这一参数满足均匀分布时，方差这一参数的分布就会变成不均匀的



5
4
3
2
1
0
概率密度
0
0.5
1
σ
5
4
3
2
1
0
概率密度
0
0.5
1
σ2

无信息先验分布设定问题

- 另外，“平坦”的分布往往是“不当的”，例如方差参数可以取任意非负实数，此时的“平坦”分布积分不等于1
- 因此，无信息先验分布的设定，需要考虑众多因素，不应将“平坦”分布作为无信息先验分布的默认设置
- 另一种常用的无信息先验分布是Jeffreys先验分布，它的出发点是保证参数变换后的先验分布的等价性

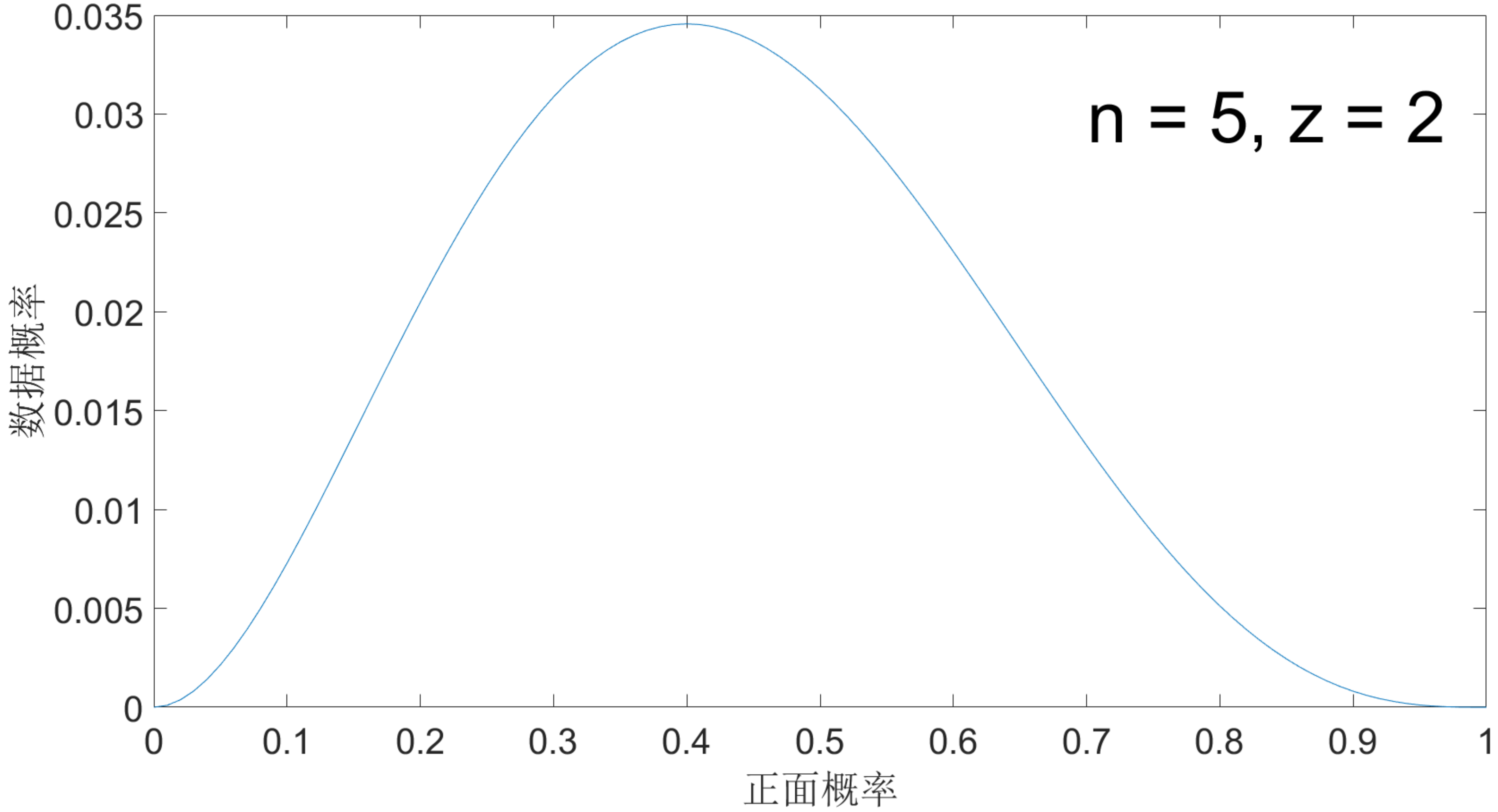

3.5
3
2.5
2
1.5
1
0.5
0
概率密度
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
X
$X \sim Beta(.5, .5)$

有信息先验分布的设定

- 在有些情况下，我们在考虑数据前，已经对模型参数有了一定的知识或者看法，此时可以使用有信息先验分布来反映这种知识和看法
- 例如，对于在所有人口中男性所占比例这一参数，我们可能会认为，最有可能的值是0.5，它处于0.45到0.55之间的概率（主观信念意义上的可能性）为99%。如果我们进一步采用贝塔分布来描述这种先验信念，那么对应的有信息先验分布为beta(330,330)，因为它是唯一满足以上两个条件的贝塔分布。
- 不同的先验信念条件及分布类型假定，会产生不同的有信息先验分布

似然函数

- 似然函数反映的是数据提供的关于模型参数的信息
- 似然值代表的是在给定模型参数取值的条件下数据出现的可能性，以概率或概率密度形式出现
- 似然值越大，说明数据对于对应参数取值的支持力度越大


n = 5, z = 2
数据概率
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
正面概率

后验概率分布

由贝叶斯公式可知，

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

其中，D代表观测数据,θ代表参数可能取值

$P(\theta|\mathrm{D})$代表在出现数据D的前提下参数取值为θ的概率(如果Θ离散)或概率密度(如果Θ连续)，即后验概率（或概率密度）

$P(\theta)$代表参数取值为θ的概率（或概率密度），即先验概率（或概率密度）

$P(\mathrm{D}|\theta)$代表在参数取值为θ的条件下出现数据D的概率(如果D离散)或概率密度(如果D连续)，即θ对应的似然值

$P(\mathrm{D})$代表出现数据D的边缘概率(如果D离散)或概率密度(如果D连续)

当Θ连续时，$P(\mathrm{D}) = \int P(D|\theta)P(\theta)d\theta$，当Θ离散时，$P(\mathrm{D}) = \sum P(D|\theta)P(\theta)$

后验分布计算问题

- 在大多数情况下，无法得到P(D)的解析解，因此也无法得到后验分布概率密度（或者概率）的解析解
- P(D)其实就是给定某一先验分布P(θ)(也就是对应的假设)条件下的P(D|H)
- 所以无法得到P(D)的解析解，也就难以计算贝叶斯因子

后验分布计算问题

- 计算后验分布的四种方法

1. 使用共轭(conjugate)先验分布，此时后验分布有解析解，P(D)也有解析解
2. 使用变分(variational)方法给出后验分布的近似解，以及P(D)的近似解
3. 使用格点(grid)法给出后验分布的近似解，以及P(D)的近似解
4. 使用MCMC（Markov Chain Monte Carlo）方法给出后验分布的近似解，以及P(D)的近似解

贝叶斯因子回顾

$$BF_{10}=\frac{P(D|H_1)}{P(D|H_0)}$$

根据数据对两个假设的概率比进行调整的幅度

$$\frac{P(H_1|D)}{P(H_0|D)}=\frac{P(D|H_1)}{P(D|H_0)}\times\frac{P(H_1)}{P(H_0)}=BF_{10}\times\frac{P(H_1)}{P(H_0)}$$

后验概率之比

先验概率之比

贝叶斯因子回顾

- 从根本上说，应该以两个假设的后验概率比为依据，来进行模型比较和选择
- 所以，基于贝叶斯因子的模型比较和选择是有前提的，前提是先验地认为两个假设的概率是相等的，即$P(H_1) = P(H_0)$
- 此时，贝叶斯因子等于后验概率比

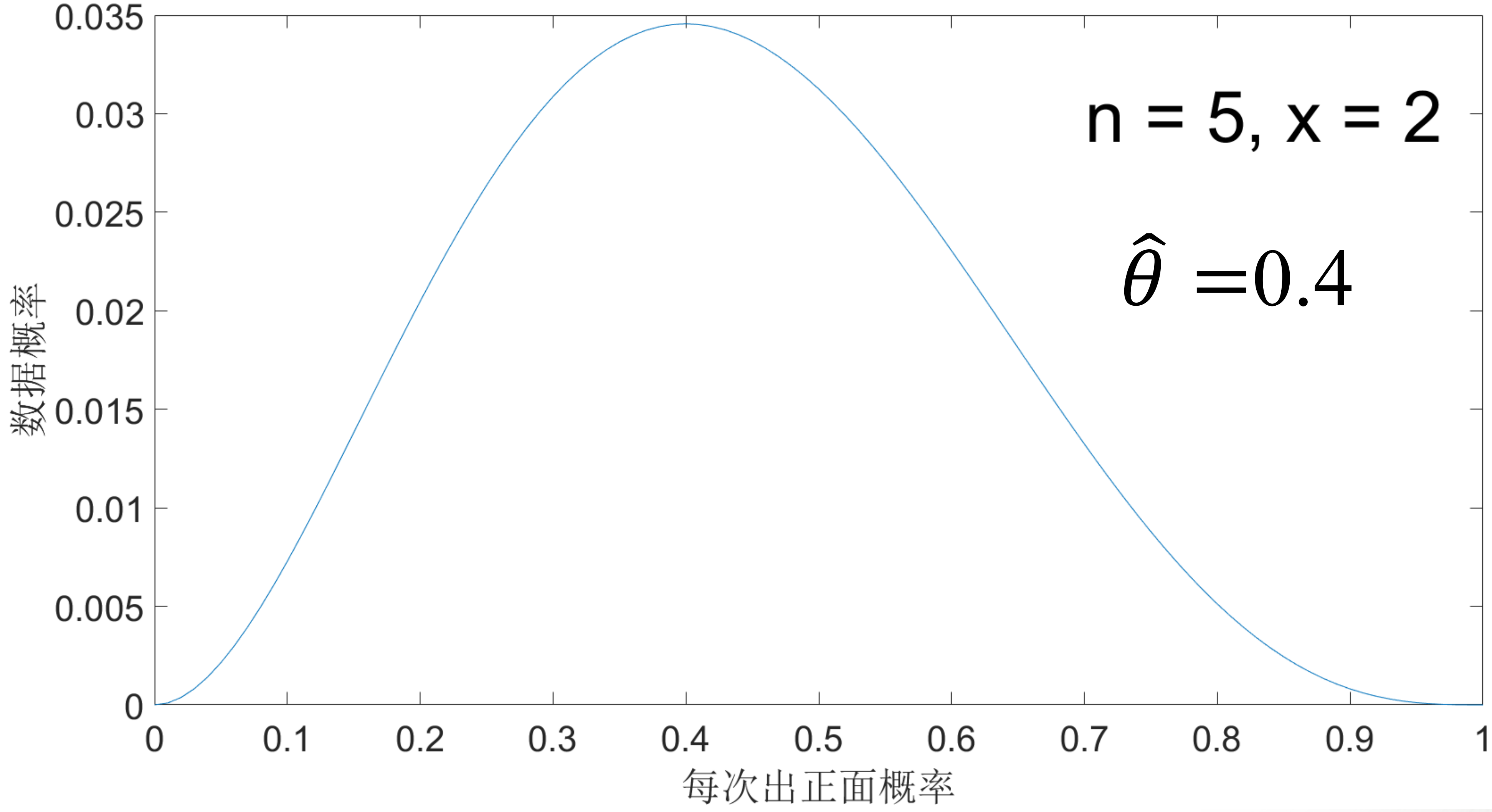
统计推断的三大任务

- 估计参数
- 检验假设
- 预测数据

参数估计

- 频率学派统计学因为关注参数的唯一真值，所以对参数的估计一般为点估计，比如用极大似然估计值作为参数估计值
- 贝叶斯统计学因为关注对于参数真值的主观认识，所以对于参数的估计，既可以是点估计，也可以是区间估计
- 频率学派统计学由于把数据看成随机的，而把参数看成固定的，所以它在设定参数估计方法时，会考虑未出现的数据，例如根据所有可能数据，找出无偏估计量作为参数估计值
- 贝叶斯统计学认为未出现的数据不应影响我们的统计推断，因此完全根据先验信念和观察到的数据为依据进行参数估计

频率学派参数估计

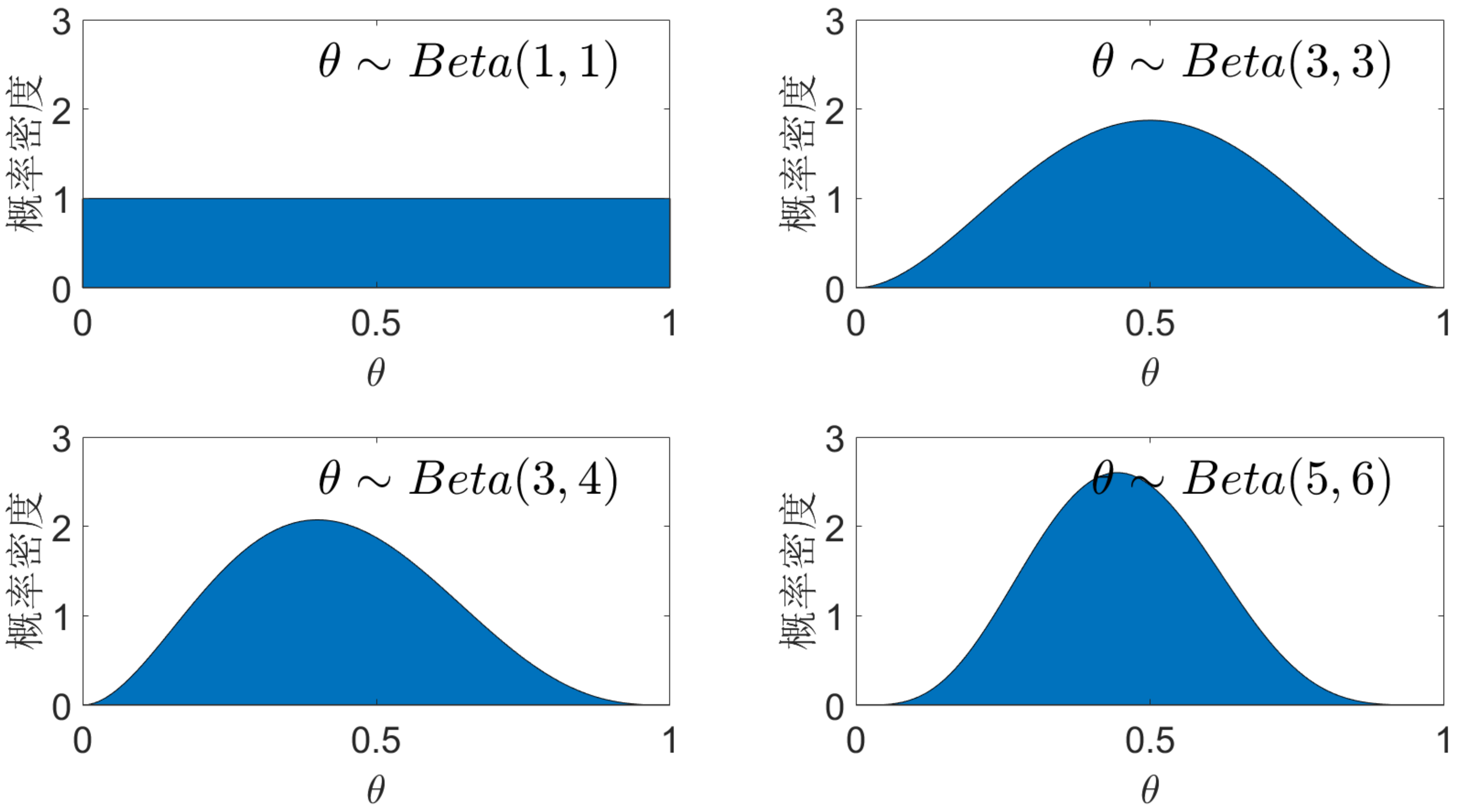


0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
数据概率
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
每次出正面概率
n = 5, x = 2
$\hat{\theta} = 0.4$

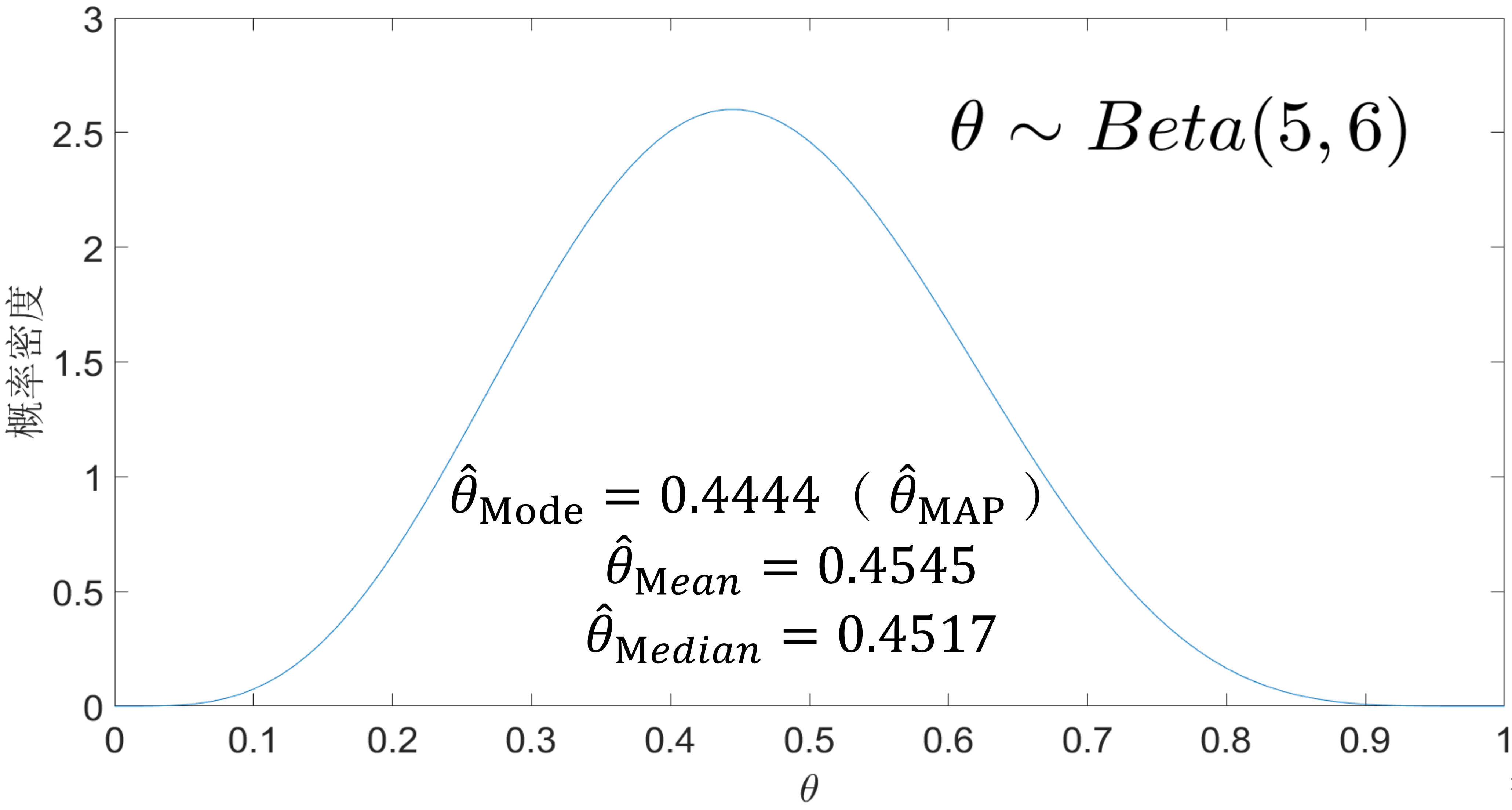
贝叶斯参数估计

先验分布

后验分布

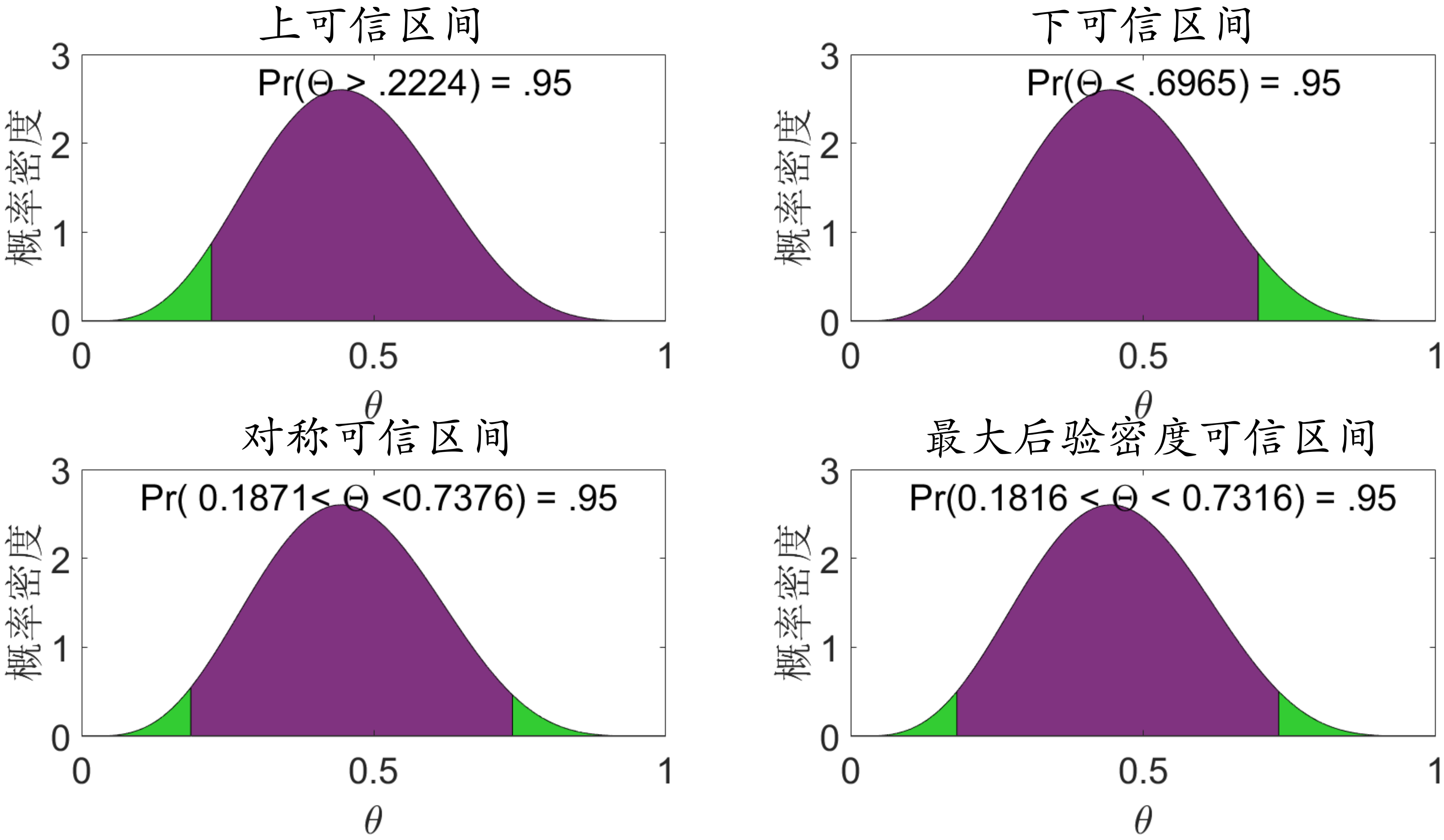


贝叶斯参数点估计



$\theta \sim Beta(5,6)$
$\hat{\theta}_{\text{Mode}} = 0.4444$ （$\hat{\theta}_{\text{MAP}}$）
$\hat{\theta}_{Mean} = 0.4545$
$\hat{\theta}_{Median} = 0.4517$
概率密度
θ
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
1.5
2
2.5
3

贝叶斯参数区间估计


上可信区间
Pr(Θ > .2224) = .95
下可信区间
Pr(Θ < .6965) = .95
对称可信区间
Pr(0.1871< Θ <0.7376) = .95
最大后验密度可信区间
Pr(0.1816 < Θ < 0.7316) = .95
概率密度
θ

贝叶斯参数区间估计

- 相比于点估计，后验分布以及基于后验分布的区间估计提供了更加丰富的有关参数的信息，例如

1. 在日常统计问题中，我们往往想要知道参数的各种可能值，以及参数位于某一区域内的可能性大小，这只有在使用后验分布的情况下才能实现
2. 参数的各种可能值，代表了不同的效应量大小，对应的后验分布和区间估计，就为我们提供了丰富的信息。相反，在频率学派统计学下，同样的一批数据只能产生一个效应量的估计值，因此提供的信息非常有限

假设检验

- 在假设检验方面，贝叶斯统计学相比于频率学派统计学也要灵活得多

1. 既可以处理点假设，也可以处理区间假设
2. 既可以比较两个假设，也可以同时比较多个假设
3. 既可能做出接受备择假设的推断，也可能做出接受零假设的推断

- 无论是基于贝叶斯因子的推断，还是基于后验分布的推断，本质上都是基于后验分布的

数据预测

- 由于贝叶斯统计学提供了模型参数的后验分布而不仅仅是点估计，它在预测新的可观测数据方面，也要比频率学派统计学更为灵活，信息更为丰富。
- 从根本上来说，贝叶斯统计学在统计推断上的优势，主要源于它考虑了所有可能的参数值，以及各种可能值的可能性大小。所以，它是一种建立在对于模型参数更加系统，更加全面的认识上的统计推断方法。