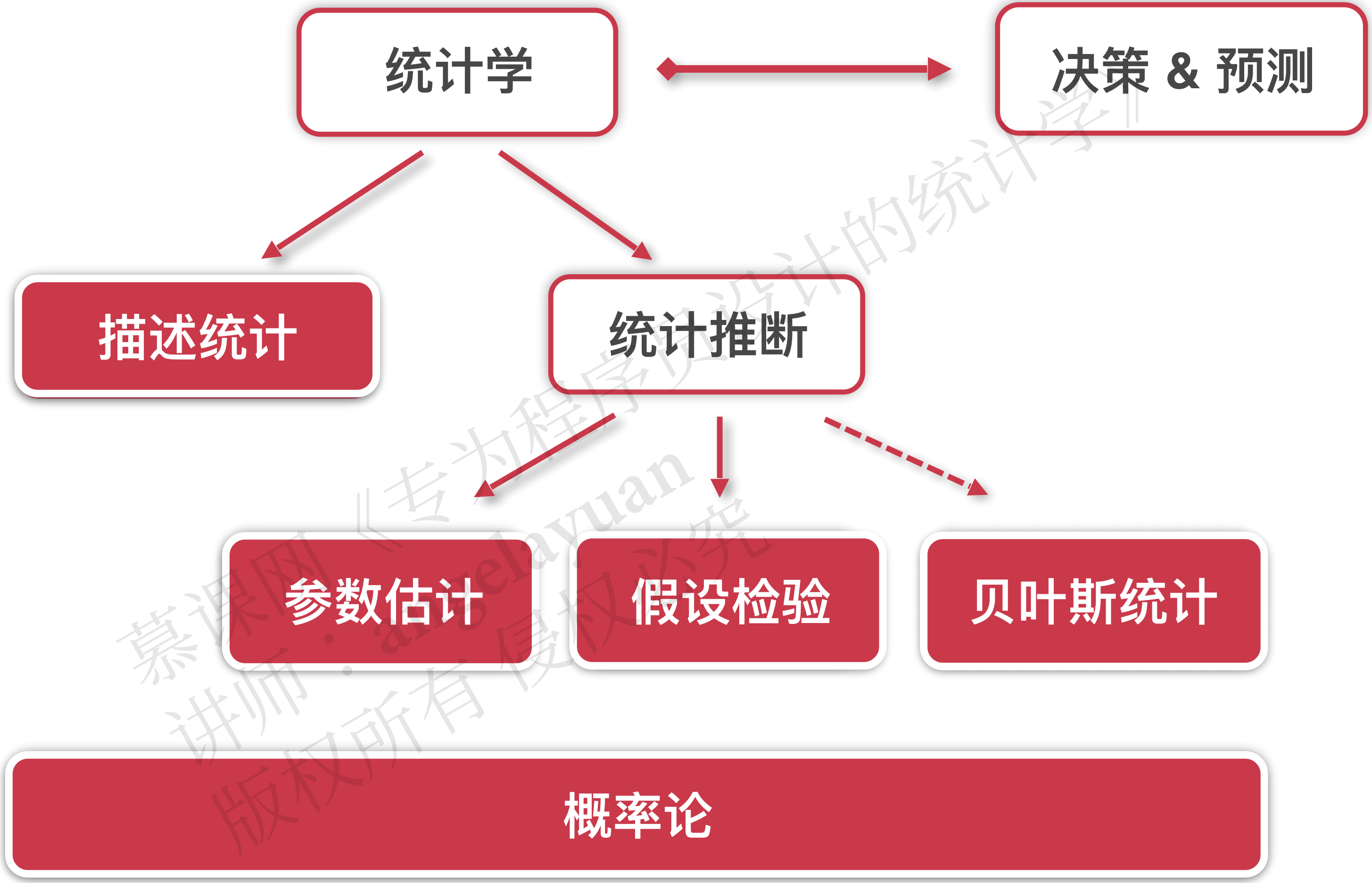
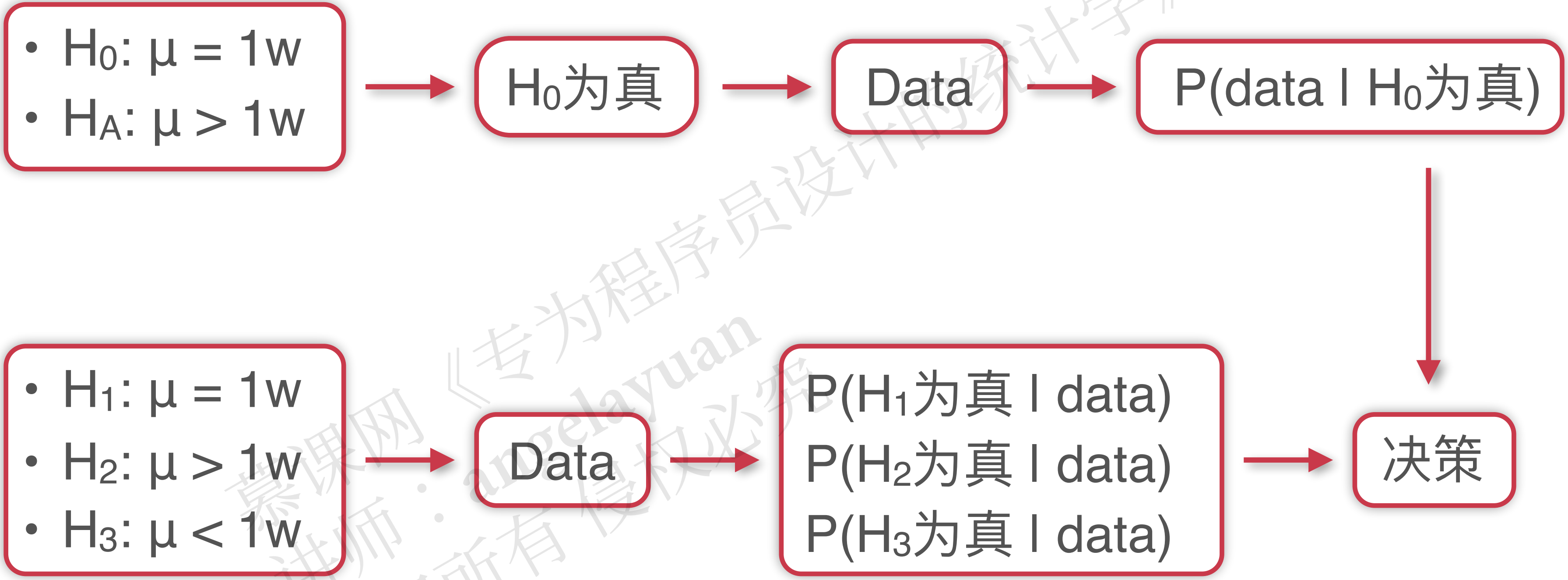

统计学
决策 & 预测
描述统计
统计推断
参数估计
假设检验
贝叶斯统计
概率论

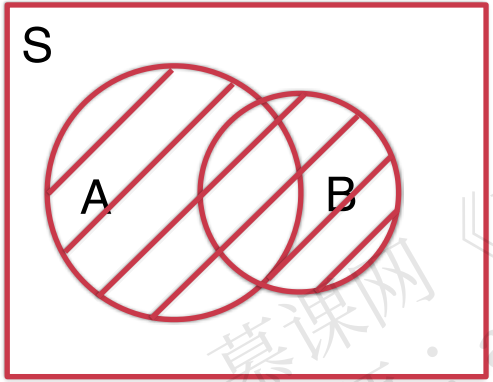
贝叶斯统计
Bayesian Statistics


• H_0: μ = 1w
• H_A: μ > 1w
H_0为真
Data
P(data | H_0为真)
• H_1: μ = 1w
• H_2: μ > 1w
• H_3: μ < 1w
Data
P(H_1为真 | data)
P(H_2为真 | data)
P(H_3为真 | data)
决策

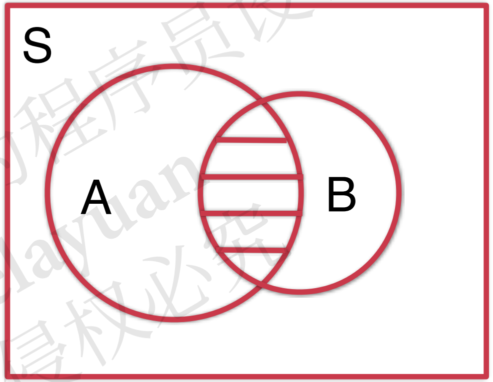
概率知识

事件间的关系

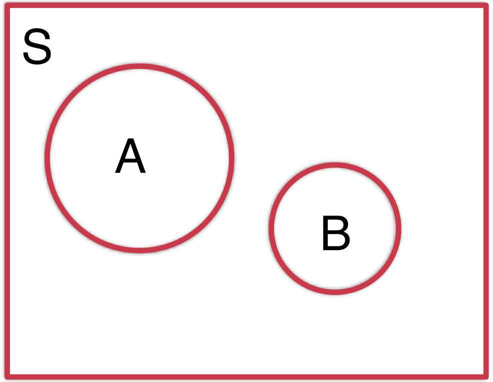
和事件 $A \cup B$



积事件 $A \cap B$



空集 $A \cap B = \emptyset$



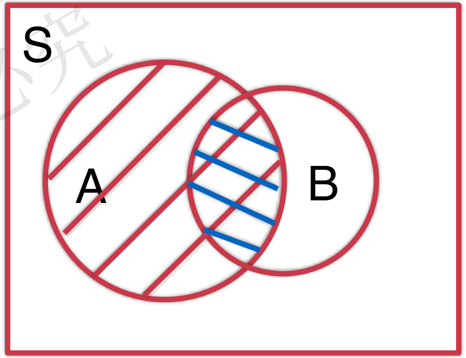
A与B是互不相容的/互斥的

条件概率和乘法定理

设A, B是两个事件， 且 $P(A) > 0$

则称 $P(B|A) = \dfrac{P(AB)}{P(A)}$ 为在事件A发生的条件下

事件B发生的条件概率



$$P(AB) = P(B|A)P(A)$$

乘法定理

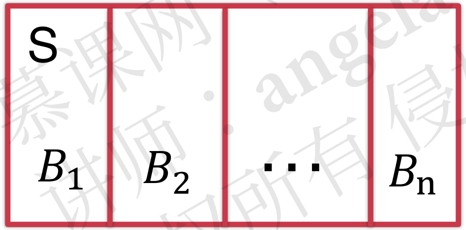
$$P(B) > 0, P(AB) = P(A|B)P(B)$$

全概率公式

S为试验E的样本空间，$B_1, B_2, \ldots, B_n$ 为E的一组事件

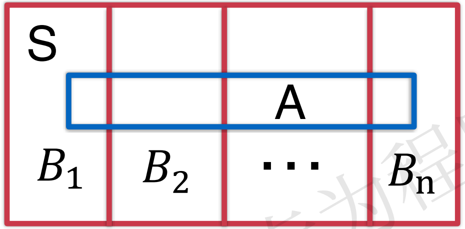
$B_iB_j = \varnothing\ (i \neq j)$ 且 $B_1 \cup B_2 \cup \cdots \cup B_n = S$

称 $B_1, B_2, \ldots, B_n$ 为样本空间S的一个划分



对每次试验，事件 $B_1, B_2, \ldots, B_n$ 必有一个且仅有一个发生

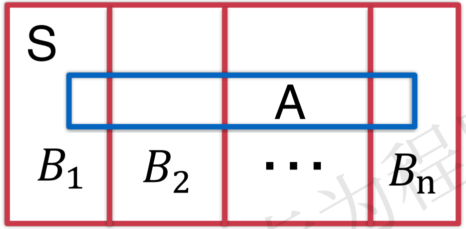
全概率公式



$$P(B_i) > 0 \ (\mathrm{i} = 1,2,\dots,\mathrm{n})$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$$

贝叶斯公式



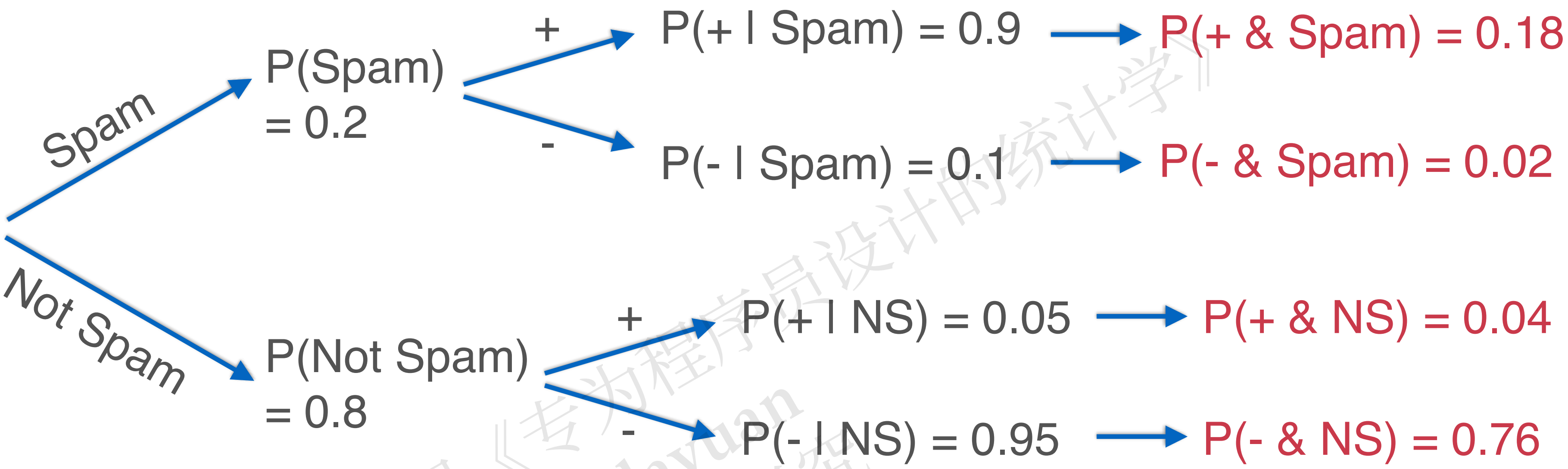
$$P(A) > 0$$

$$P(B_i) > 0 \ (\mathrm{i} = 1,2,\ldots,\mathrm{n})$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n)$$

$$P(B_i|\mathrm{A}) = \frac{P(B_iA)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$ 贝叶斯公式

概率树 (probability tree)

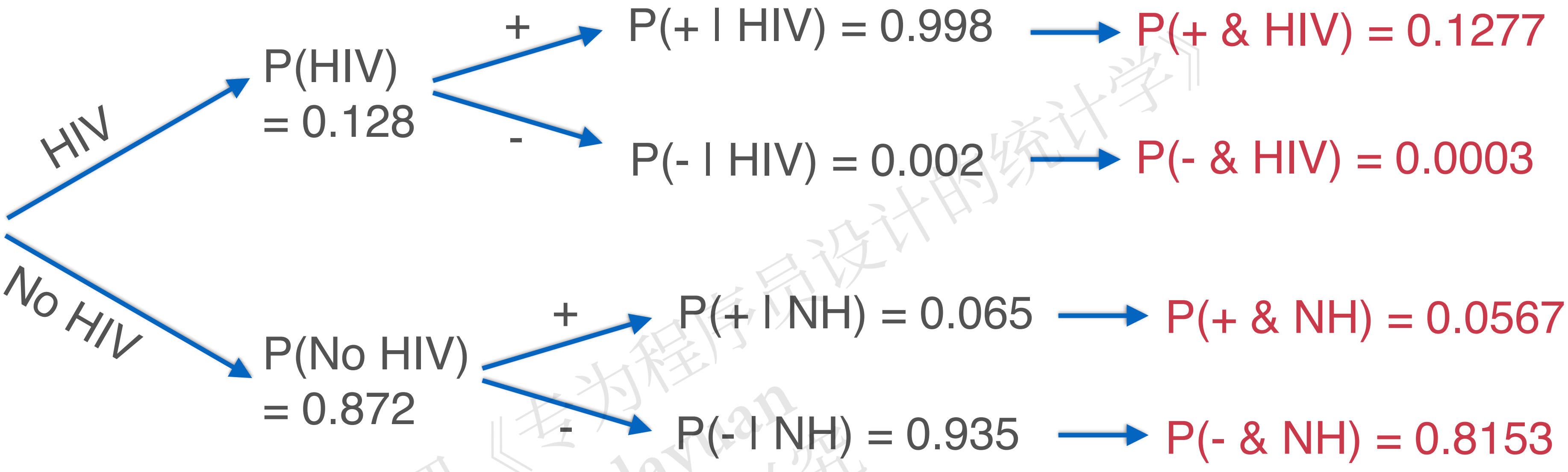


P(Spam | +)= 0.18/(0.18+0.04) = 0.818

P(Spam | -) = 0.02/(0.02+0.76) = 0.026

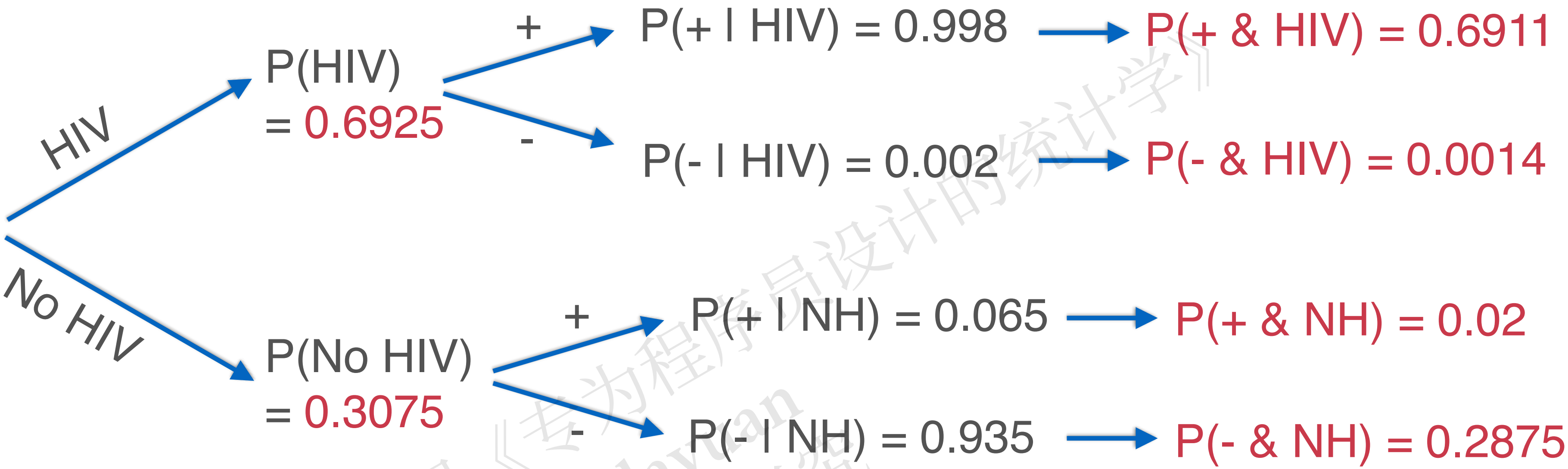
P(NS | +) = ?

P(NS | -) = ?



P(HIV | +) = 0.1277/(0.1277+0.0567) = 0.6925

P(NH | -) = 0.8153/(0.8153+0.0003) = 0.9996



P(HIV | +) = 0.6911/(0.6911+0.02) = 0.972

P(NH | -) = 0.2875/(0.2875+0.0014) = 0.995

贝叶斯推断

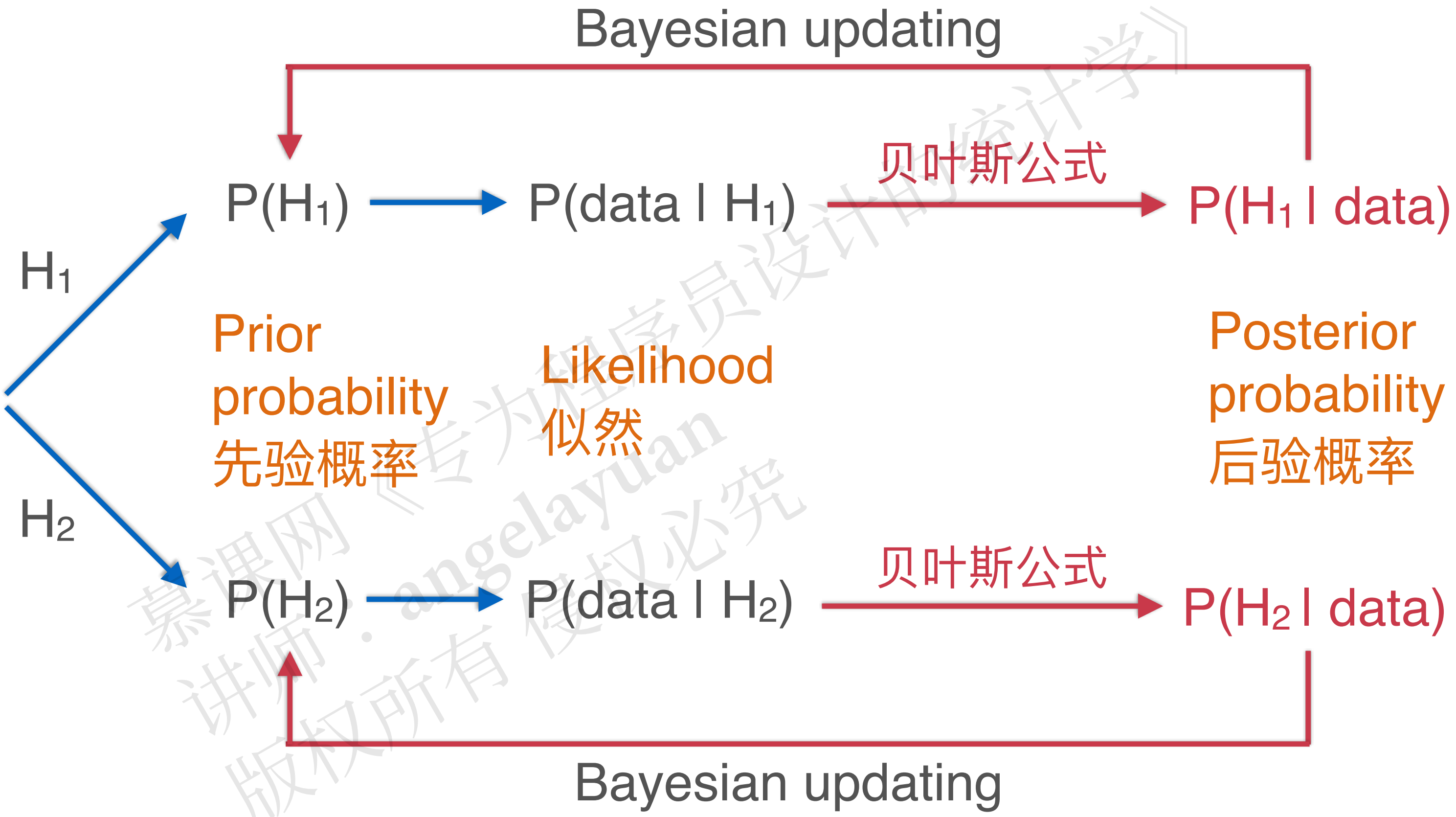
$$P(B_i|\mathrm{A}) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_j)P(B_j)}$$

A = 数据(data)

B_i = 第i个假设H_i

$$P(H_i|\mathrm{data}) = \frac{P(data|H_i)P(H_i)}{\sum_{j=1}^{n} P(data|H_j)P(H_j)}$$

$P(\mathrm{data} \mid H_i) \longrightarrow P(H_i \mid \mathrm{data})$


Bayesian updating
H1
H2
P(H1)
P(data | H1)
贝叶斯公式
P(H1 | data)
Prior
probability
先验概率
Likelihood
似然
Posterior
probability
后验概率
P(H2)
P(data | H2)
贝叶斯公式
P(H2 | data)
Bayesian updating

Likelihood

Posterior probability

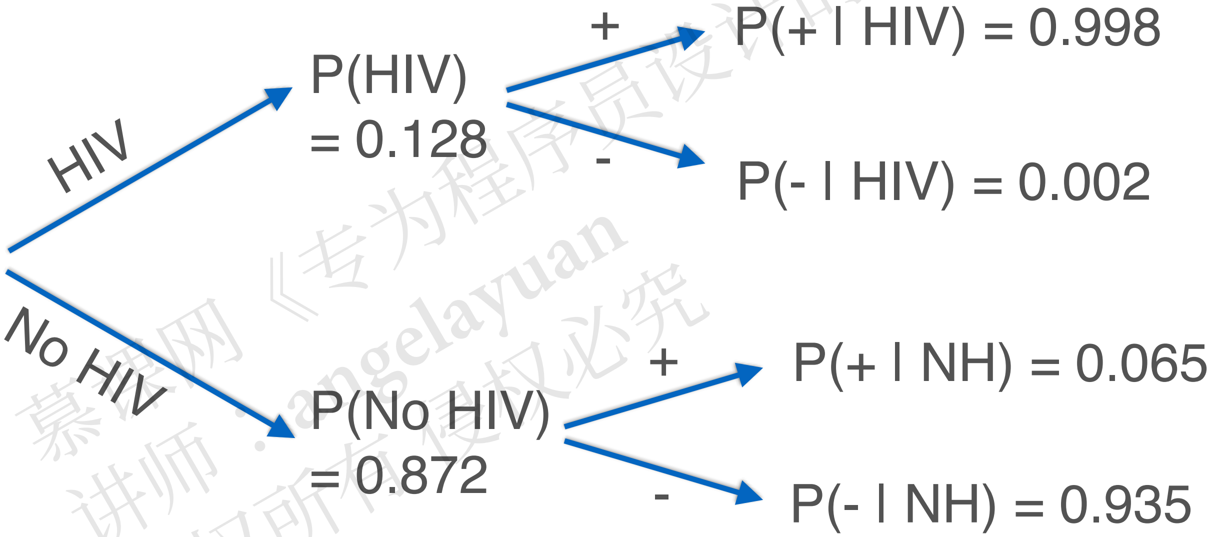
Prior probability

$$P(H_i|\text{data}) = \frac{P(data|H_i)P(H_i)}{\sum_{j=1}^{n} P(data|H_j)P(H_j)}$$

Marginal likelihood

贝叶斯推断 - 例1

贝叶斯推断 - 例1


HIV
No HIV
P(HIV)
= 0.128
P(No HIV)
= 0.872
+
-
+
-
P(+ | HIV) = 0.998
P(- | HIV) = 0.002
P(+ | NH) = 0.065
P(- | NH) = 0.935

贝叶斯推断 - 例1

Hypothesis/Model	HIV	No HIV	Total
Prior: P(model)	0.128	0.872	1
Data	+		
Likelihood P(data I model)	0.998	0.065	
P(data I mode)xP(model)	0.1277	0.0567	0.1844
Posterior: P(model I data)	0.6925	0.3075	1

贝叶斯推断 - 例1

Hypothesis/Model	HIV	No HIV	Total
Prior: P(model)	0.6925	0.3075	1
Data	+		
Likelihood P(data I model)	0.998	0.065	
P(data I mode)xP(model)	0.6911	0.02	0.7111
Posterior: P(model I data)	0.972	0.028	1

贝叶斯因子(Bayes Factor)

$$\text{Bayes Factor} = \frac{P(\text{data} \mid H_1)}{P(\text{data} \mid H_2)}$$ 数据更支持哪一个假设/模型

$$BF = \frac{P(+ \mid HIV)}{P(+ \mid NH)} = \frac{0.998}{0.065} = 15.35$$

$$BF = \frac{P(- \mid HIV)}{P(- \mid NH)} = \frac{0.002}{0.935} = 0.002$$

贝叶斯因子(Bayes Factor)

$$BF = \frac{P(+ \mid HIV)}{P(+ \mid NH)} = 15.35$$

$$BF = \frac{P(- \mid HIV)}{P(- \mid NH)} = 0.002$$

BF	解释
< 1	没有证据支持H_1 (支持H_2)
1 ~ 3	较弱的证据支持H_1
3 ~ 10	中等程度的证据支持H_1
10 ~ 30	较强的证据支持H_1
30 ~ 100	非常强的证据支持H_1
> 100	极强的证据支持H_1

Odds和Odds ratio

Model	HIV	No HIV	odds of HIV = $\frac{P(HIV)}{P(No\ HIV)}$
Prior: P(model)	0.128	0.872	0.147
Posterior P(model I data)	0.6925	0.3075	2.252

$$\text{odds ratio} = \frac{\text{Posterior odds}}{\text{Prior odds}} = 2.252/0.147 = 15.4$$

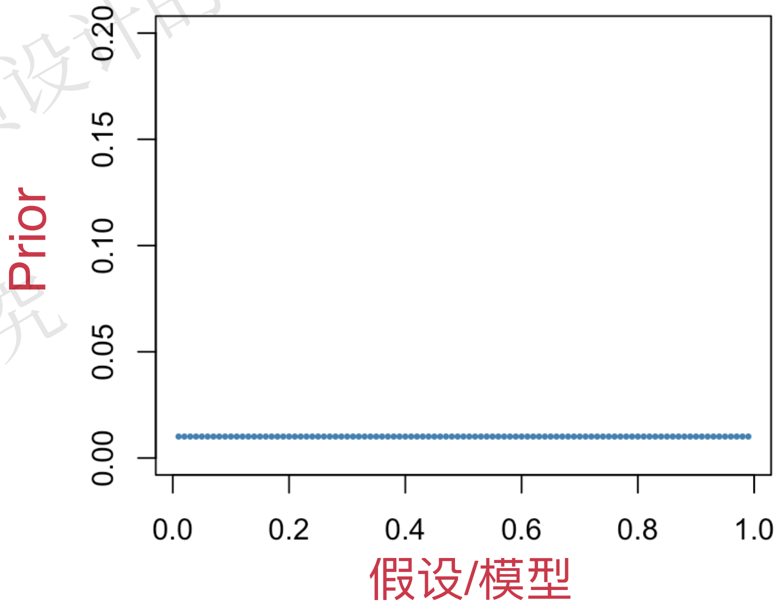
贝叶斯推断 - 例2

贝叶斯推断 - 例2

- 研究问题: 治疗高血压的新药是否有效
- 方法: 让高血压患者服用新药一段时间, 然后检查高血压症状是否有所改善
- 使用症状得到改善的患者占总患者数的比例来近似代表新药的有效程度
- 数据: 100名高血压患者服用新药, 其中78名患者症状有所改善

贝叶斯推断 - 例2

- 明确假设(模型)和先验概率
  - 假设/模型: 新药的有效性为 0.01, 0.02, …, 0.99 (以 0.01为步长)
  - 先验概率: 各模型的概率相等, 均为1/99



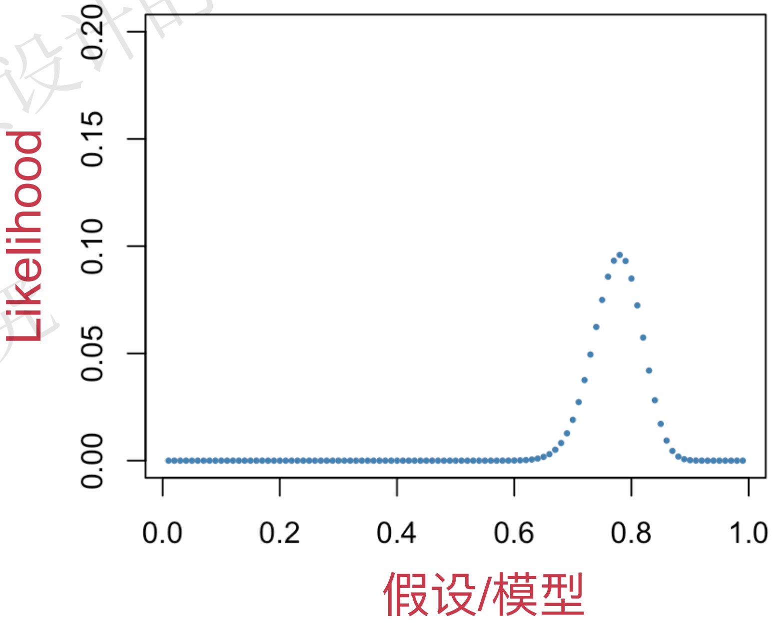
贝叶斯推断 - 例2

- 计算各假设下得到观测数据的可能性(likelihood)

$$P\{X=k\}=\binom{n}{k}p^k(1-p)^{n-k}$$

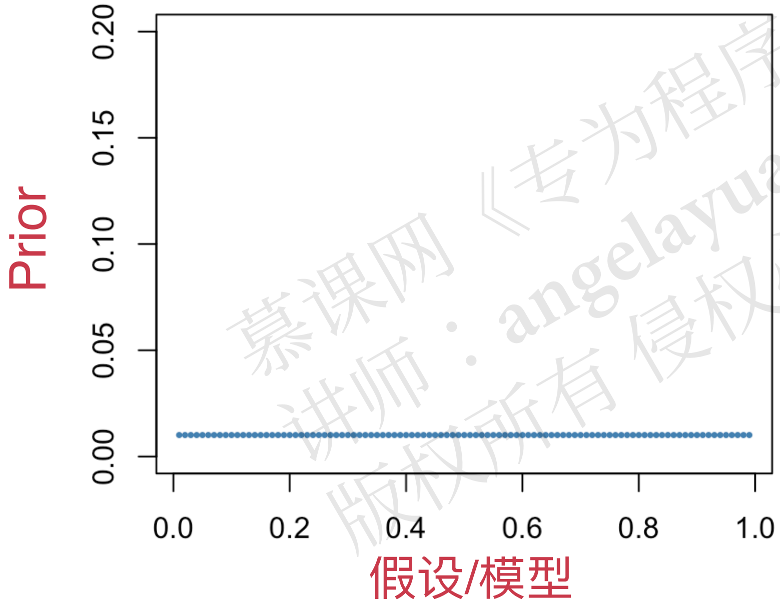
$$(k=0,1,2,\ldots,n; 0<p<1)$$

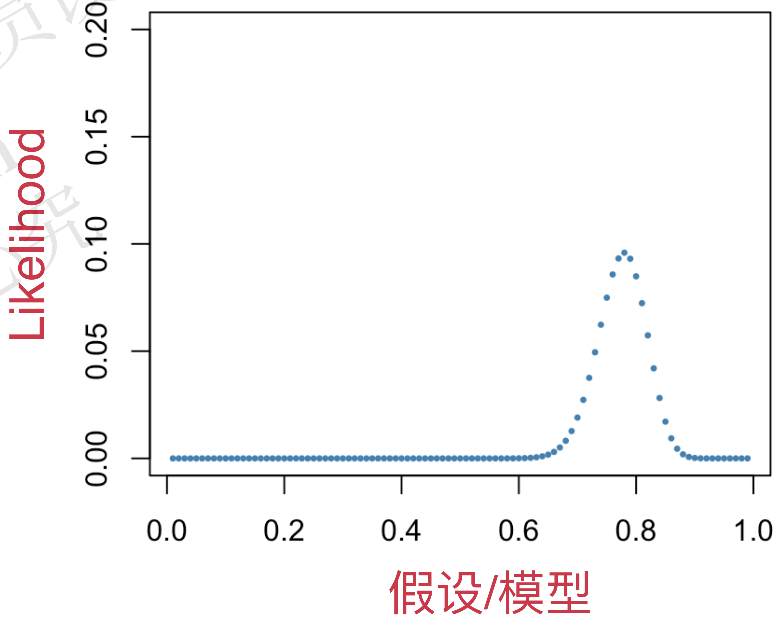
k = 78, n = 100,
p = 0.01, 0.02, …, 0.99



贝叶斯推断 - 例2

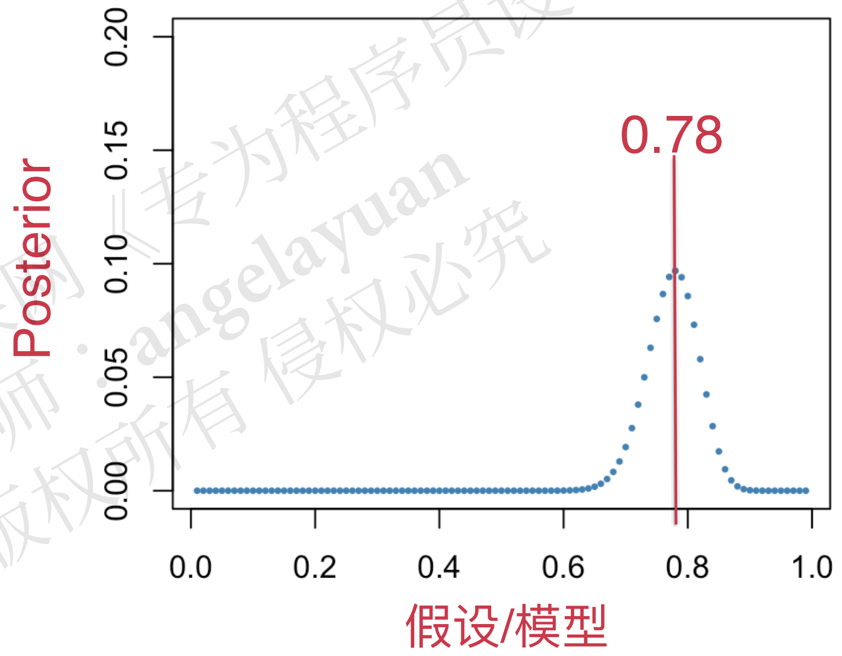
- 计算marginal likelihood = sum(prior x likelihood)



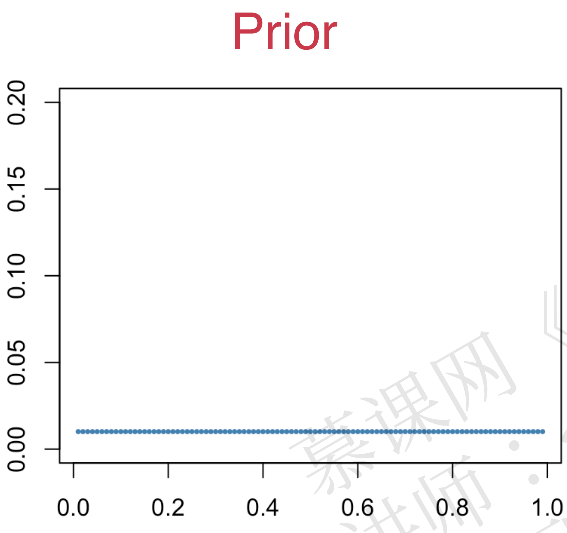


贝叶斯推断 - 例2

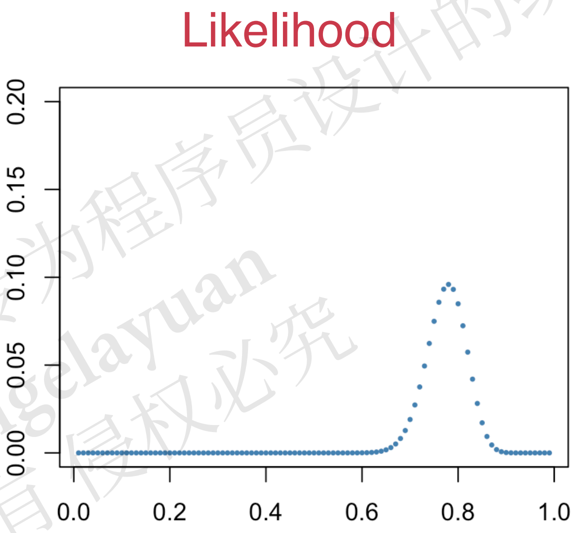
- 计算后验概率 = (likelihood x prior) / marginal likelihood



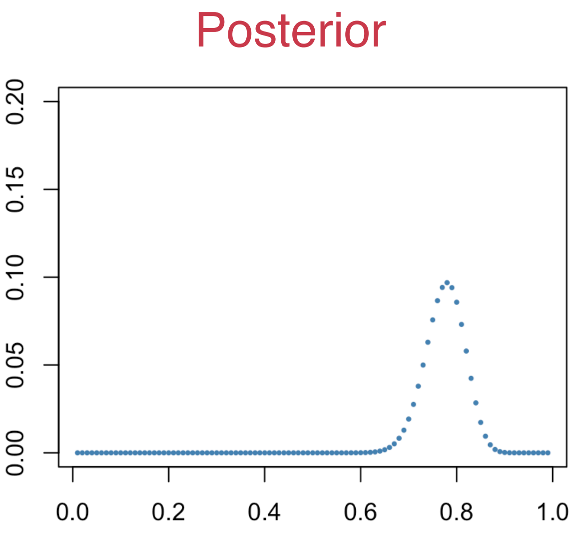
先验概率的选择对后验概率影响



假设/模型



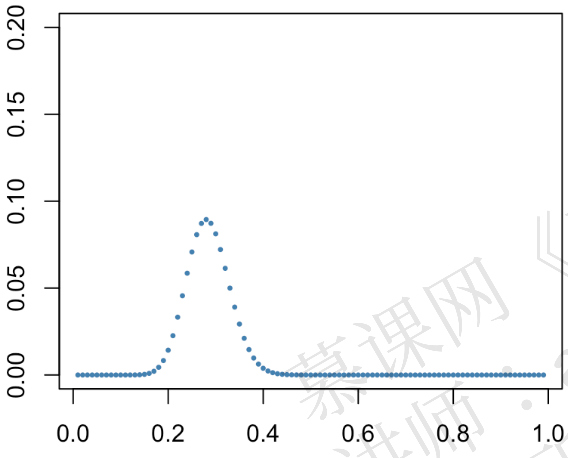
假设/模型



假设/模型

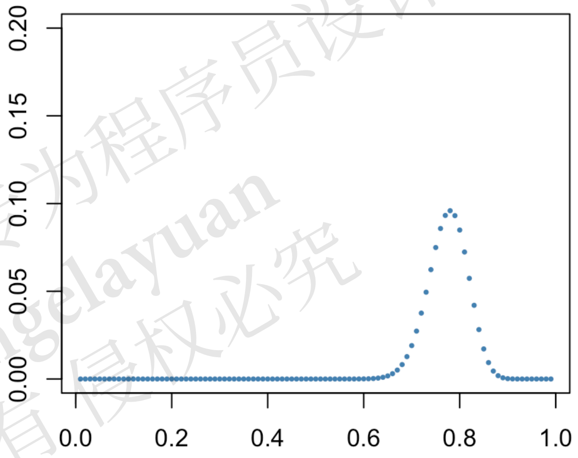
先验概率的选择对后验概率影响

Prior



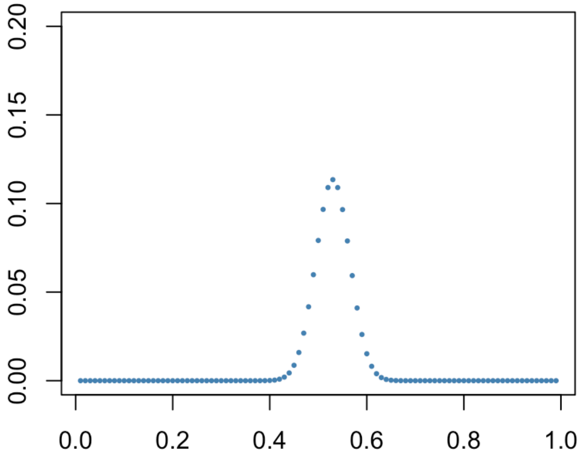
假设/模型

Likelihood



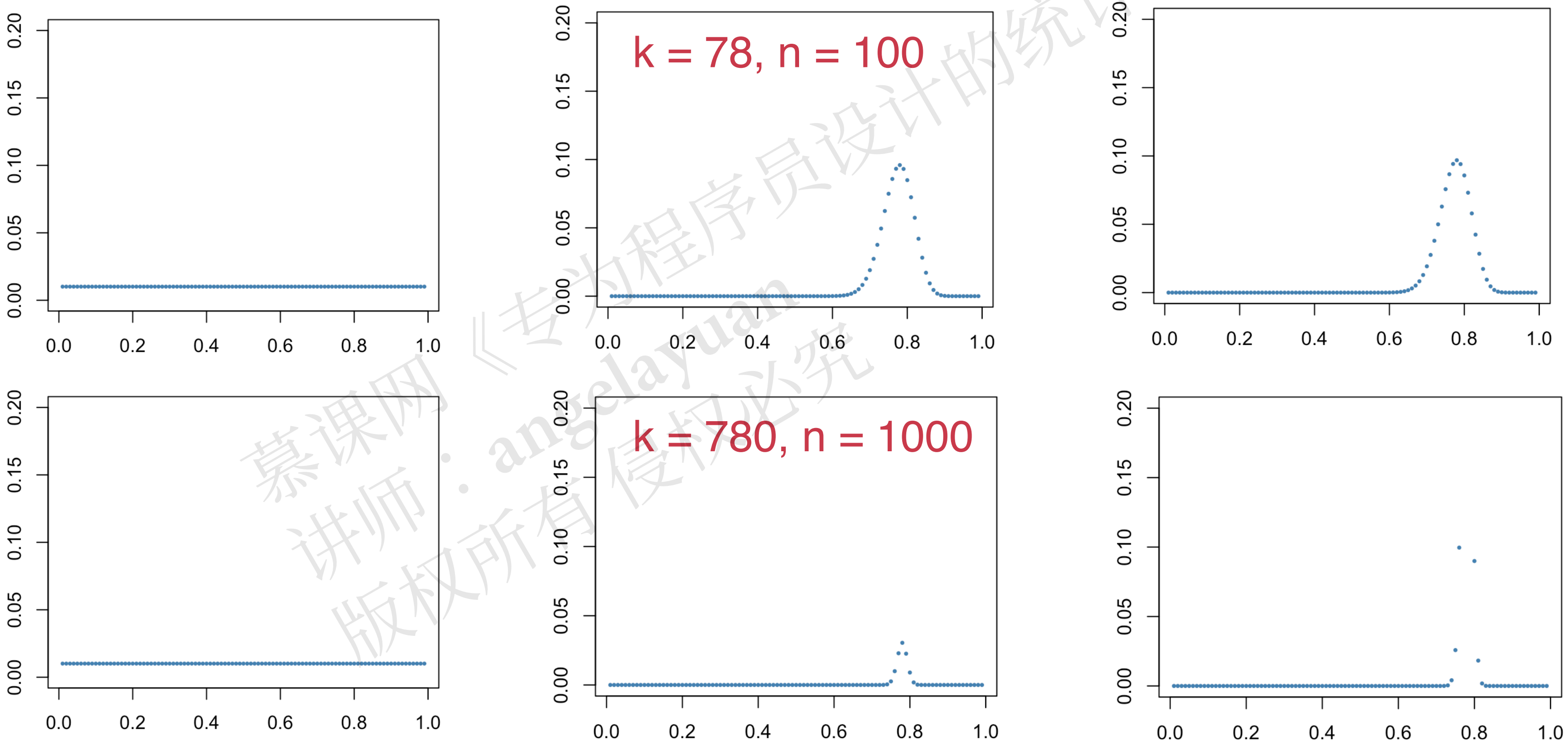
假设/模型

Posterior



假设/模型

样本容量对后验概率的影响



置信区间(credible interval)

置信区间 (credible interval)

- 使用拒绝采样(rejection sampling)方法从后验概率分布中抽样，然后计算分位数以得到置信区间
  - 抽样次数: n = 100,000
  - x代表$p_{respond}$
  - y代表$p_{respond}$的后验概率

置信区间 (credible interval)

- 重复下列步骤n次
  - 从均匀分布U(0,1)中抽取1个数, 赋值给x
  - 从均匀分布U(0,1)中抽取1个数, 赋值给y
  - 从真实的后验概率分布中找到x对应的后验概率f(x)
  - 如果y < f(x), 接受x, 记录下该x的值

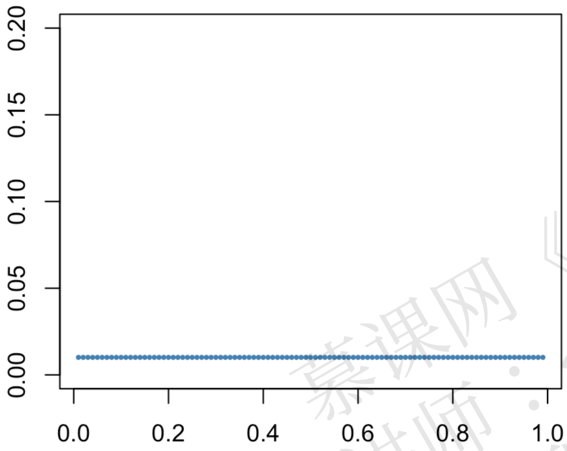
置信区间 (credible interval)

- 从所有被接受的x值中, 找到两个数值L和H, 使落在这两个数值之间的x的值的个数占x值总数的95%, 则这两个数值构成的区间(L,H)就是一个95%的置信区间

- 置信区间的含义: $p_{respond}$落在区间(L,H)内的概率是95%

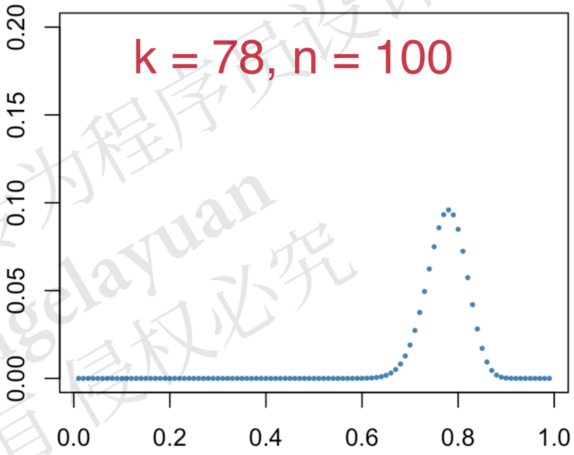
置信区间 (credible interval)

Prior



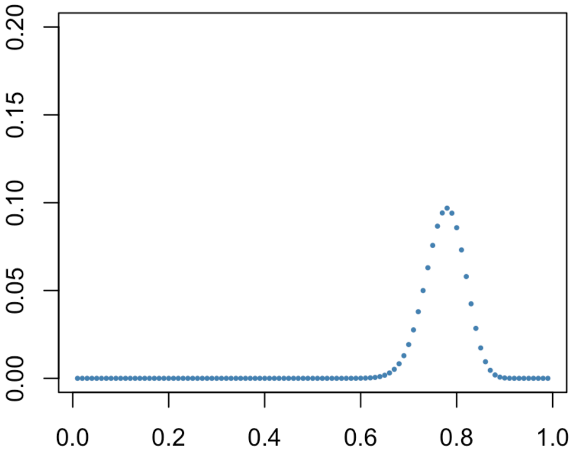
假设/模型

Likelihood



假设/模型

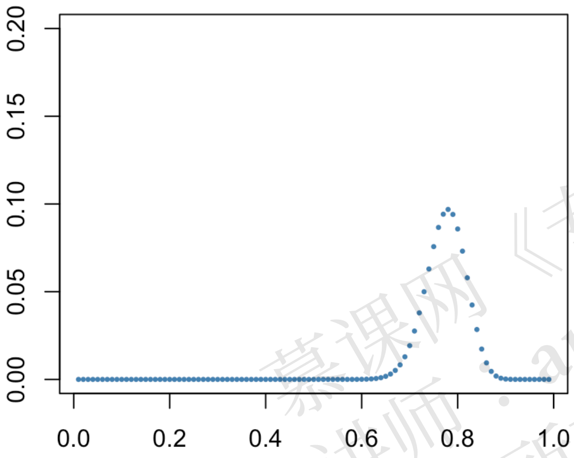
Posterior



假设/模型

置信区间 (credible interval)

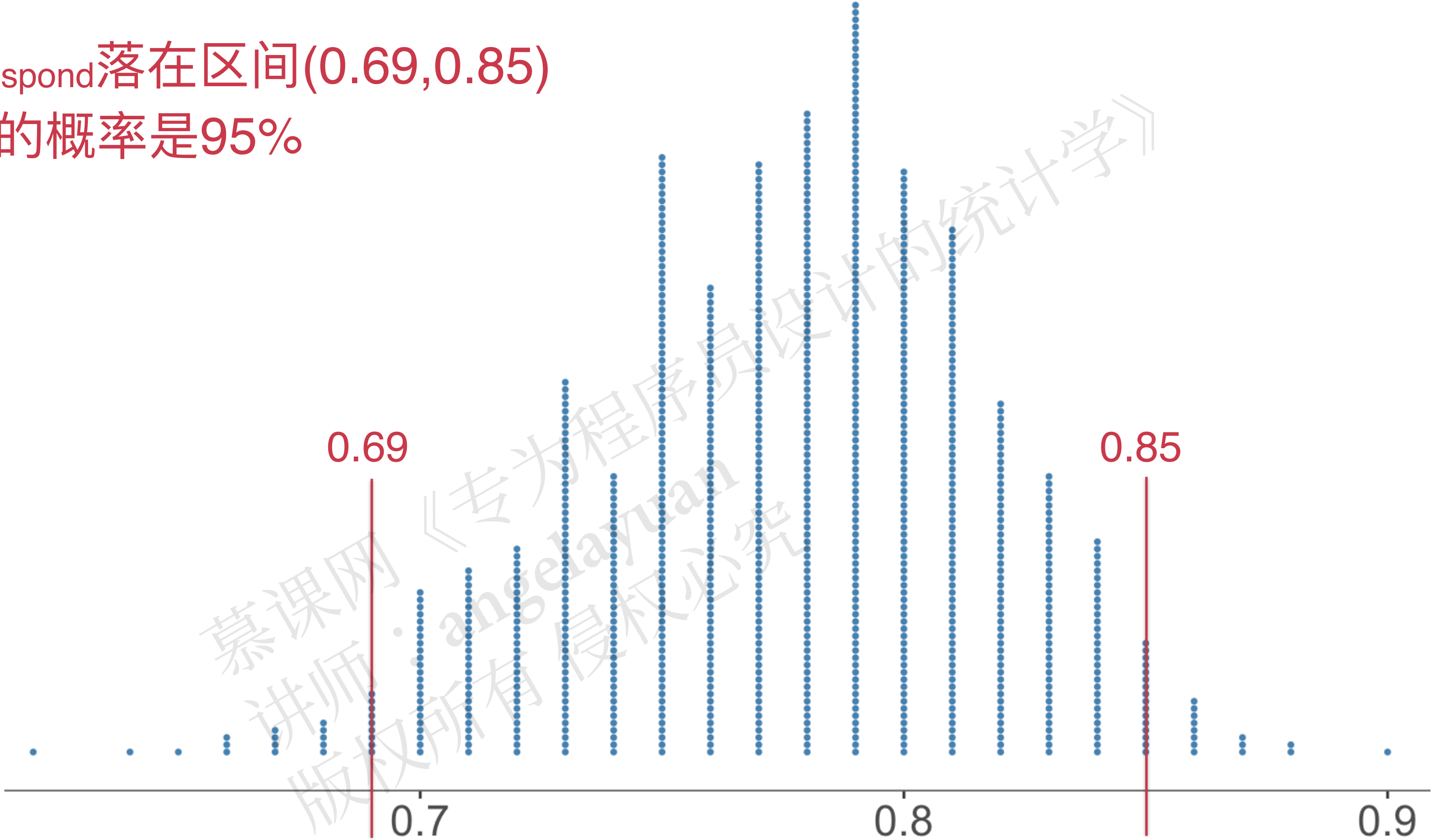
Posterior



假设/模型

- x = 0.8, y = 0.3
- Posterior(x = 0.8) = 0.086
- y > f(x), 拒绝x

$p_{respond}$落在区间(0.69,0.85)内的概率是95%



本章小结

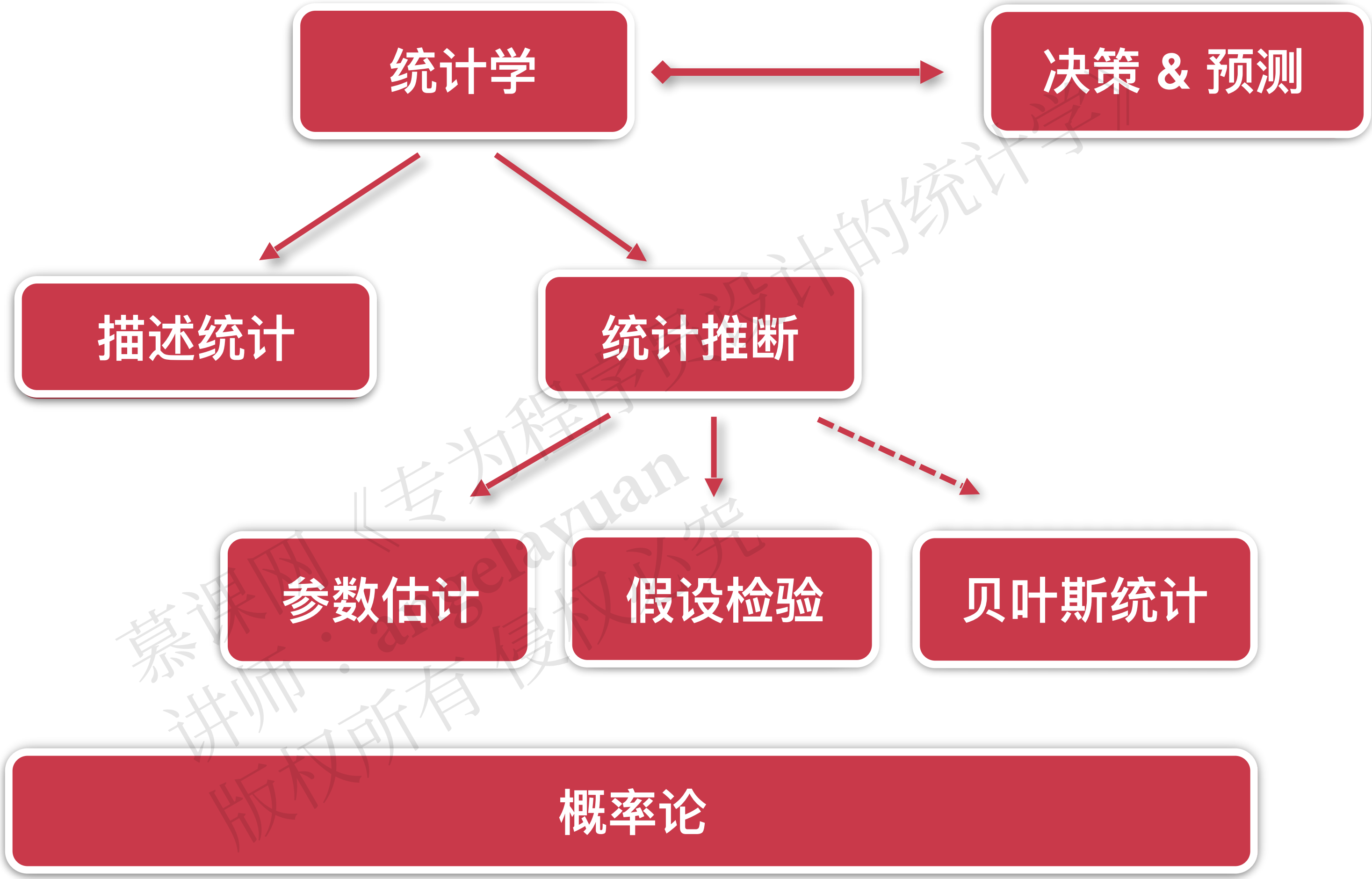
贝叶斯统计

条件概率, 乘法定理, 全概率公式, 贝叶斯公式

概率树

贝叶斯推断
- 疾病诊断; 参数估计
- 先验概率, 样本容量对后验概率的影响

置信区间 (credible interval)



统计学
决策 & 预测
描述统计
统计推断
参数估计
假设检验
贝叶斯统计
概率论