20 贝叶斯推断入门

Bayesian Inference 101

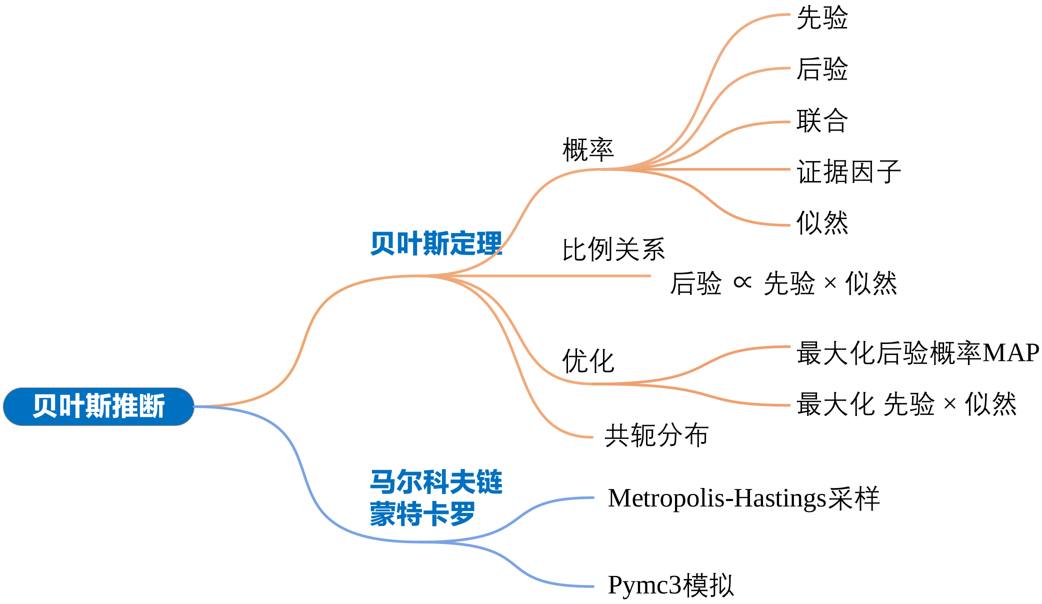
参数不确定，参数对应概率分布

没有事实，只有解释。

There are no facts, only interpretations.

—— 弗里德里希·尼采 (Friedrich Nietzsche) | 德国哲学家 | 1844 ~ 1900

- ◄ matplotlib.pyplot.axvline() 绘制竖直线
- ◄ matplotlib.pyplot.fill_between() 区域填充颜色
- ◄ numpy.cumsum() 累加
- ◄ scipy.stats.bernoulli.rvs() 满足伯努利分布的随机数
- ◄ scipy.stats.beta() Beta 分布


贝叶斯推断
贝叶斯定理
概率
先验
后验
联合
证据因子
似然
比例关系
后验 ∝ 先验 × 似然
优化
最大化后验概率MAP
最大化 先验 × 似然
共轭分布
马尔科夫链蒙特卡罗
Metropolis-Hastings采样
Pymc3模拟

20.1 贝叶斯推断：更贴合人脑思维

一个让人“头大”的公式

本章和下一章的关键就是如何理解、应用以下公式进行贝叶斯推断：

$$f_{\Theta|X}(\theta|x)=\frac{f_{X|\Theta}(x|\theta)\,f_{\Theta}(\theta)}{\int_{\vartheta} f_{X|\Theta}(x|\vartheta)\,f_{\Theta}(\vartheta)\,\mathrm{d}\vartheta} \tag{1}$$

值得注意的是这个公式还有如下常见的几种其他写法：

$$\begin{aligned}
f_{\Theta|X}(\theta|x)&=\frac{f_{X|\Theta}(x|\theta)\,f_{\Theta}(\theta)}{\int_{\theta'} f_{X|\Theta}(x|\theta')\,f_{\Theta}(\theta')\,\mathrm{d}\theta'}\\
f_{\Theta|X}(\theta|x)&=\frac{f_{X|\Theta}(x|\theta)\,g_{\Theta}(\theta)}{\int_{\vartheta} f_{X|\Theta}(x|\vartheta)\,g_{\Theta}(\vartheta)\,\mathrm{d}\vartheta}\\
p_{\Theta|X}(\theta|x)&=\frac{p_{X|\Theta}(x|\theta)\,p_{\Theta}(\theta)}{\int_{\theta'} p_{X|\Theta}(x|\theta')\,p_{\Theta}(\theta')\,\mathrm{d}\theta'}
\end{aligned} \tag{2}$$

有些书中，有把 x 写成 y 情况，也有用 $\pi()$ 代表概率密度/质量分布函数。总而言之，(1) 的表达方式很多，大家见多了，也就“见怪不怪”了。

(1) 这个公式是横在大家理解掌握贝叶斯推断之路上的一块“巨石”。本章试图用最简单的例子帮大家敲碎这块“巨石”。

在正式介绍这个公式之前，本节先用白话聊聊什么是**贝叶斯推断** (Bayesian inference)。

贝叶斯推断

本书第 16 章介绍过，在贝叶斯学派眼里，模型参数本身也是随机变量，也服从某种分布。贝叶斯推断的核心就是，在以往的经验 (先验概率) 基础上，结合新的数据，得到新的概率 (后验概率)。而模型参数分布随着外部样本数据不断输入而迭代更新。不同的是，频率派只考虑样本数据本身，不考虑先验概率。

依我看来，人脑的运作方式更贴近贝叶斯推断。

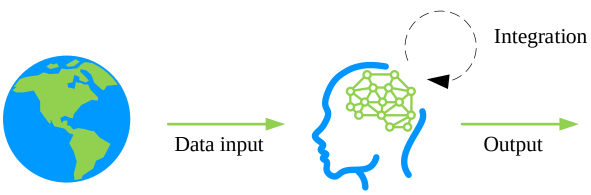


图 1. 人脑更像是一个贝叶斯推断机器

举个最简单的例子，试想你一早刚出门的时候发现忘带手机，大脑第一反应是——手机最可能在哪？

这个“贝叶斯推断”的结果一般基于两方面因素：一方面，日复一日的“找手机”的经验；另一方面，“今早、昨晚在哪用过手机”的最新数据。

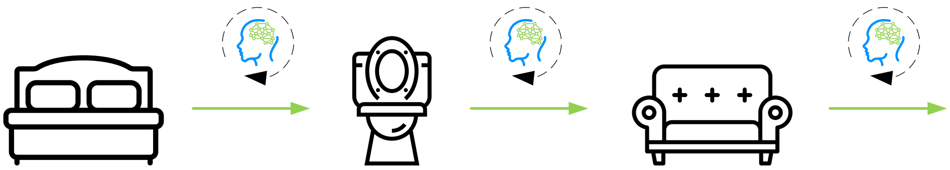


图 2. 找手机

而且在不断寻找手机的过程，大脑不断提出“下一个最有可能的地点”。

比如，昨晚睡觉前刷了一小时手机，手机肯定在床上！

跑到床头，发现手机不在床上，那很可能在马桶附近，因为早晨方便的时候一般也会刷手机！

竟然也不在马桶附近！那最可能在沙发茶几上，因为坐着看电视的时候我也爱刷手机 ...

试想，如果大脑没有以上“经验 + 最新数据”，你会怎么找手机？或者，“贝叶斯推断”找手机无果的时候，我们又会怎么办？

我们很可能会像“扫地机器人”一样，“逐点扫描”，把整个屋子从里到外歇斯底里地翻一遍。这种地毯式“采样”就类似频率派的做法。

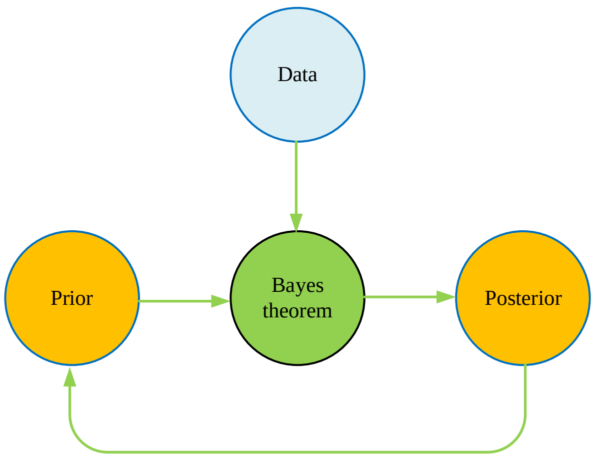


图 3. 通过贝叶斯定理迭代学习

这个找手机的过程也告诉我们，贝叶斯推断常常迭代使用。在引入新的样本数据后，先验概率产生后验概率。而这个后验概率也可以作为新的先验概率，再根据最新出现的数据，更新后验概率，如此往复。

人生来就是一个“学习机器”，“前事不忘后事之师”说的也是这个道理。通过不断学习 (数据输入)，我们不断更新自己对世界的认知 (更新模型参数)。这个过程从出生一直持续到离开这个世界为止。

往大了说，人类认识世界的机制又何尝不是贝叶斯推断。在新的数据影响下，人类一次次创造、推翻、重构知识体系。这个过程循环往复，不断推动人类认知进步。举个例子，统治西方世界思想界近千年的地心说被推翻后，日心说渐渐成了主流。在伽利略等一众巨匠的臂膀上，牛顿力学体系横空出世。在后世科学家不断努力完善下，牛顿力学体系和麦克斯韦电磁场理论为基础的物理大厦大功告成。当人们满心欢喜，以为物理学就剩下一些敲敲打打的修饰工作，结果蓝天之上又飘来了两朵乌云 …

20.2 从一元贝叶斯公式说起

先验

在任何引入任何观测数据之前，未知参数 Θ 本身是随机变量，自身对应概率分布为 $f_\Theta(\theta)$，这个分布叫做**先验分布** (prior distribution)。先验分布函数 $f_\Theta(\theta)$ 中，Θ 为随机变量，θ 是一个变量。$\Theta = \theta$ 代表随机变量 Θ 的取值为 θ。

似然

在 $\Theta = \theta$ 条件下，观察到的数据 X 的分布为**似然分布** (likelihood distribution) $f_{X|\Theta}(x|\theta)$。似然分布是一个条件概率。当 $\Theta = \theta$ 取不同值时，似然分布 $f_{X|\Theta}(x|\theta)$ 也有相应变化。

回顾本书第 17 章介绍最大似然估计 MLE，优化问题的目标函数本质上就是似然函数 $f_{X|\Theta}(x|\theta)$ 的连乘。第 17 章不涉及贝叶斯推断，因此我们没有用条件概率 $f_{X|\Theta}(x|\theta)$，用的是 $f_X(x; \theta)$。**对数似然** (log-likelihood function) 就是对似然函数取对数。

联合

根据贝叶斯定理，X 和 Θ 的**联合分布** (joint distribution) 为：

$$\underbrace{f_{X,\Theta}(x,\theta)}_{\text{Joint}} = \underbrace{f_{X|\Theta}(x|\theta)}_{\text{Likelihood}} \underbrace{f_\Theta(\theta)}_{\text{Prior}} \tag{3}$$

请大家注意，为了方便，在贝叶斯推断中，我们不再区分概率密度函数 PDF、概率质量函数 PMF，所有概率分布均用 $f()$ 记号。而且，(1) 的分母也仅仅用积分符号。

证据

如果 X 为连续随机变量，X 的边缘概率分布为：

$$\underbrace{f_X(x)}_{\text{Evidence}} = \int_\theta \underbrace{f_{X,\Theta}(x,\theta)}_{\text{Joint}} \mathrm{d}\theta = \int_\theta \underbrace{f_{X|\Theta}(x|\theta)}_{\text{Likelihood}} \underbrace{f_\Theta(\theta)}_{\text{Prior}} \mathrm{d}\theta \tag{4}$$

联合分布 $f_{X|\Theta}(x|\theta)$ 对 θ“偏积分”消去了 θ，积分结果 $f_X(x)$ 和 θ 无关。我们一般也管 $f_X(x)$ 叫做**证据因子** (evidence)，这和前两章的叫法一致。

$f_X(x)$ 和 θ 无关，这意味着观测到的数据对先验的选择没有影响。

后验

给定 $X = x$ 条件下，Θ 的条件概率为：

$$f_{\Theta|X}(\theta|x) = \frac{\overbrace{f_{X,\Theta}(x,\theta)}^{\text{Joint}}}{\underbrace{f_X(x)}_{\text{Evidence}}} = \frac{f_{X|\Theta}(x|\theta) f_\Theta(\theta)}{\int_\vartheta \underbrace{f_{X|\Theta}(x|\vartheta)}_{\text{Likelihood}} \underbrace{f_\Theta(\vartheta)}_{\text{Prior}} \mathrm{d}\vartheta} \tag{5}$$

为了避免混淆，上式分母中用了花写 ϑ。

$f_{\Theta|X}(\theta|x)$ 叫**后验分布** (posterior distribution)，它代表在整合“先验 + 样本数据”之后，我们对参数 Θ 的新的“认识”。在连续迭代贝叶斯学习中，这个后验概率分布是下一个迭代的先验概率分布。

正比关系

通过前两章的学习，我们知道后验与先验和似然乘积成正比：

$$\underbrace{f_{\Theta|X}(\theta|x)}_{\text{Posterior}} \propto \underbrace{f_{X|\Theta}(x|\theta)}_{\text{Likelihood}} \underbrace{f_\Theta(\theta)}_{\text{Prior}} \tag{6}$$

即，后验 ∝ 似然 × 先验。

但是为了得出真正的后验概率密度，本章的例子中我们还是要完成 $\int_\vartheta f_{X|\Theta}(x|\vartheta) f_\Theta(\vartheta)\mathrm{d}\vartheta$ 积分。此外，这个积分很可能没有解析解 (闭式解)，可能需要用到数值积分或蒙特卡洛模拟。这是本书第 22 章要讲解的内容之一。

注意，先验分布、后验分布是关于模型参数的分布。此外，通过一定的转化，我们可以把似然函数也变成有关模型参数的“分布”。

下面，我们便结合实例讲解贝叶斯推断。

20.3 走地鸡兔：比例完全不确定

本节举一个例子，展开讲解贝叶斯推断。

回到本书第 16 章“鸡兔同笼”的例子。一个巨大无比农场散养大量“走地”鸡、兔。但是，农夫自己也说不清楚鸡兔的比例。

用 Θ 代表兔子的比例随机变量，这意味着 Θ 的取值范围为 [0, 1]。即，$\Theta = 0.5$ 意味着农场有 50%兔、50%鸡，$\Theta = 0.3$ 意味着有 30%兔、70%鸡。

为了搞清楚农场鸡兔比例，农夫决定随机抓 n 只动物。X_1、X_2 … X_n 为每次抓取动物的结果。X_i ($i = 1, 2, \ldots, n$) 的样本空间为 {0, 1}，其中 0 代表鸡，1 代表兔。

注意，抓取动物过程，我们同样忽略这对农场整体动物总体比例的影响。

先验

由于农夫完全不确定鸡兔的比例，我们选择连续均匀分布 uniform(0, 1) 为先验分布，所以 $f_\Theta(\theta)$ 为：

$$f_\Theta(\theta) = 1, \quad \theta \in [0,1] \tag{7}$$

再次强调，先验分布代表我们对模型参数的“主观经验”，先验分布的选择独立于“客观”样本数据。

图 4 所示为 [0, 1] 区间上的均匀分布，也就是说兔子比例 Θ 可以是 [0, 1] 区间内的任意一个数，而且可能性相同。

这个例子告诉我们，没有先验信息，或者先验分布不清楚，也不要紧！我们可以用常数或均匀分布作为先验分布。这种情况也叫**无信息先验** (uninformative prior)。

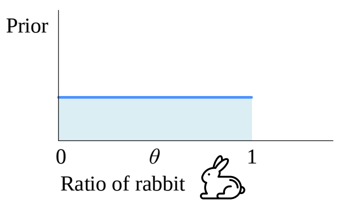


图 4. 选择连续均匀分布作为先验分布

似然

给定 $\Theta = \theta$ 条件下，X_1、X_2 … X_n 服从 IID 的伯努利分布 Bernoulli(θ)，即：

$$\underbrace{f_{X_i|\Theta}\left(x_i \mid \theta\right)}_{\text{Likelihood}} = \theta^{x_i}\left(1-\theta\right)^{1-x_i} \tag{8}$$

其中，$\Theta = \theta$ 代表农场中兔子的比例，取值范围为 [0, 1] 区间任意数值；$1 - \theta$ 代表鸡的比例。$X_i = x_i$ 代表某一次抓到的动物，0 代表鸡，1 代表兔。

也就是说，(8) 中，θ 是未知量。实际上，上式中似然概率 $f_{X_i|\Theta}\left(x_i \mid \theta\right)$ 代表概率质量函数。

本书前文提过，IID 的含义是**独立同分布** (Independent Identically Distribution)。在随机过程中，任何时刻的取值都为随机变量，如果这些随机变量服从同一分布，并且互相独立，那么这些随机变量是独立同分布。

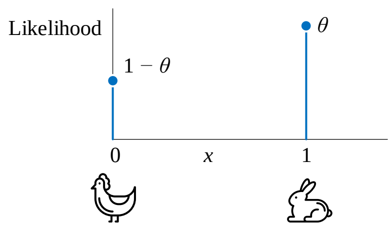


图 5. 似然分布

联合

因此，X_1、X_2 … X_n、Θ 联合分布为：

$$\begin{aligned}\underbrace{f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right)}_{\text{Joint}} &= \underbrace{f_{X_1,X_2,\dots,X_n|\Theta}\left(x_1,x_2,\dots,x_n \mid \theta\right)}_{\text{Likelihood}}\underbrace{f_{\Theta}\left(\theta\right)}_{\text{Prior}} \\ &= f_{X_1|\Theta}\left(x_1 \mid \theta\right)\cdot f_{X_2|\Theta}\left(x_2 \mid \theta\right)\cdots f_{X_n|\Theta}\left(x_n \mid \theta\right)\cdot \underbrace{f_{\Theta}\left(\theta\right)}_{1} \\ &= \prod_{i=1}^{n}\theta^{x_i}\left(1-\theta\right)^{1-x_i} = \theta^{\overbrace{\sum_{i=1}^{n}x_i}^{s}}\left(1-\theta\right)^{n-\overbrace{\sum_{i=1}^{n}x_i}^{s}}\end{aligned} \tag{9}$$

令：

$$s = \sum_{i=1}^{n} x_i \tag{10}$$

s 的含义是 n 次抽取中兔子的总数。

这样 (9) 可以写成：

$$f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right) = \theta^{s}\left(1-\theta\right)^{n-s} \tag{11}$$

上式中，$n - s$ 代表 n 次抽取中鸡的总数。

证据

证据因子 $f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)$，即 $f_{\chi}(\boldsymbol{x})$，可以通过 $f_{X_1,X_2,\dots,X_n,\Theta}(x_1,x_2,\dots,x_n,\theta)$ 对 θ“偏积分”得到：

$$
\begin{aligned}
f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n) &= \int_{\theta} f_{X_1,X_2,\dots,X_n,\Theta}(x_1,x_2,\dots,x_n,\theta)\,\mathrm{d}\theta \\
&= \int_{\theta} \theta^{s}(1-\theta)^{n-s}\,\mathrm{d}\theta
\end{aligned}
\tag{12}
$$

以上积分相当于在 θ 维度上压缩，结果 $f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)$ 和 θ 无关。再次强调，在贝叶斯推断中，上述积分很可能没有解析解。

想到本书第 7 章介绍的 Beta 函数，(12) 可以写成：

$$
\begin{aligned}
f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n) &= \int_{\theta} \theta^{s+1-1}(1-\theta)^{n-s+1-1}\,\mathrm{d}\theta \\
&= \mathrm{B}(s+1,n-s+1) = \frac{s!(n-s)!}{(n+1)!}
\end{aligned}
\tag{13}
$$

利用 Beta 函数的性质，我们“逃过”积分运算。

图 6 所示为 $\mathrm{B}(s + 1, n - s + 1)$ 函数随着 s、n 变化的平面等高线。

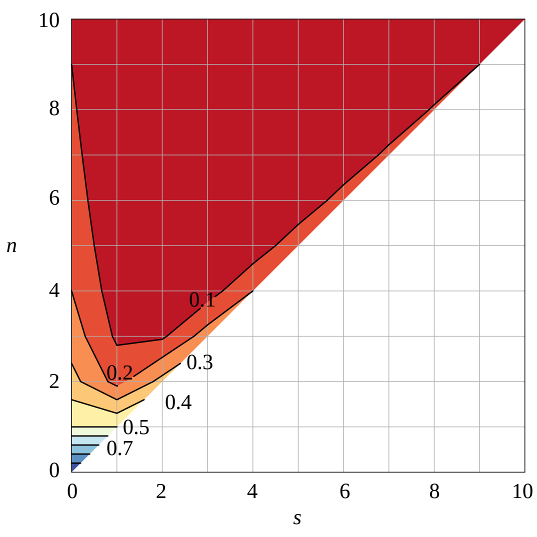


图 6. $\mathrm{B}(s + 1, n - s + 1)$ 函数图像平面等高线

后验

由此，在 $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ 条件下，Θ 的后验分布为：

$$
\begin{aligned}
f_{\Theta|X_1,X_2,\dots,X_n}(\theta|x_1,x_2,\dots,x_n) &= \frac{\overbrace{f_{X_1,X_2,\dots,X_n,\Theta}(x_1,x_2,\dots,x_n,\theta)}^{\text{Joint}}}{\underbrace{f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)}_{\text{Evidence}}} \\
&= \frac{\theta^s(1-\theta)^{n-s}}{\mathrm{B}(s+1,n-s+1)} = \frac{\theta^{(s+1)-1}(1-\theta)^{(n-s+1)-1}}{\mathrm{B}(s+1,n-s+1)}
\end{aligned} \tag{14}
$$

我们惊奇地发现，上式对应 Beta(s + 1, n – s + 1) 分布。

总结来说，农夫完全不清楚鸡兔的比例，因此选择先验概率为 uniform(0, 1)。抓取 n 只动物，知道其中有 s 只兔子，n – s 只鸡，利用贝叶斯定理整合“先验概率 + 样本数据”得到后验概率为 Beta(s + 1, n – s + 1) 分布。马上，我们就通过蒙特卡罗模拟结果将具体数值代入后验概率 Beta(s + 1, n – s + 1)，这样就可以看到后验分布的形状。

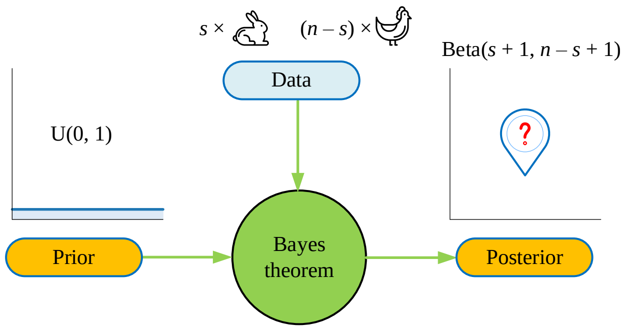


图 7. 先验 U(0, 1) + 样本 (s, n – s) → 后验 Beta(s + 1, n – s + 1)

正比关系

(14) 中分母 $\mathrm{B}(s+1,n-s+1)$ 的作用是条件概率归一化。实际上，根据 (6)，我们只需要知道：

$$
f_{\Theta|X_1,X_2,\dots,X_n}(\theta|x_1,x_2,\dots,x_n) \propto f_{X_1,X_2,\dots,X_n|\Theta}(x_1,x_2,\dots,x_n|\theta)\, f_\Theta(\theta) = \theta^s(1-\theta)^{n-s} \tag{15}
$$

我们在前两章也看到了这个正比关系的应用。但是为了方便蒙特卡罗模拟，本节还是会使用 (14) 给出的后验分布解析式。

蒙特卡罗模拟

下面，我们编写 Python 代码来进行上述贝叶斯推断的蒙特卡洛模拟。先验分布为 Uniform(0, 1)，这意味着各种鸡兔比例可能性相同。

大家查看代码会发现，代码中实际用的分布是 Beta(1, 1)。Uniform(0, 1) 和 Beta(1, 1) 形状相同，而且方便本章后续模拟。

本章代码用到伯努利分布随机数发生器。假设兔子占整体的真实比例为 0.45 (45%)。图 8 (a) 所示为用伯努利随机数发生器产生的随机数，红点 ● 代表鸡 (0)，蓝点 ● 代表兔 (1)。

通过图 8 (a) 样本数据做推断便是频率学派的思路。频率学派依靠样本数据，而不引入先验概率 (已有知识或主观经验)。当样本数量较大时，频率学派可以做出合理判断；但是，当样本数量很小时，频率学派做出的推断往往不可信。

图 8 (b) 中，从下到上所示为不断抓取动物中鸡、兔各自的比例变化。当动物的数量 n 不断增多时，我们发现比例趋于稳定，并逼近真实值 (0.45)。

图 8 (c) 所示为随着样本数据不断导入，后验概率分布曲线的渐变过程。请大家仔细观察图 8 (c)，看看能不能发现有趣的规律。

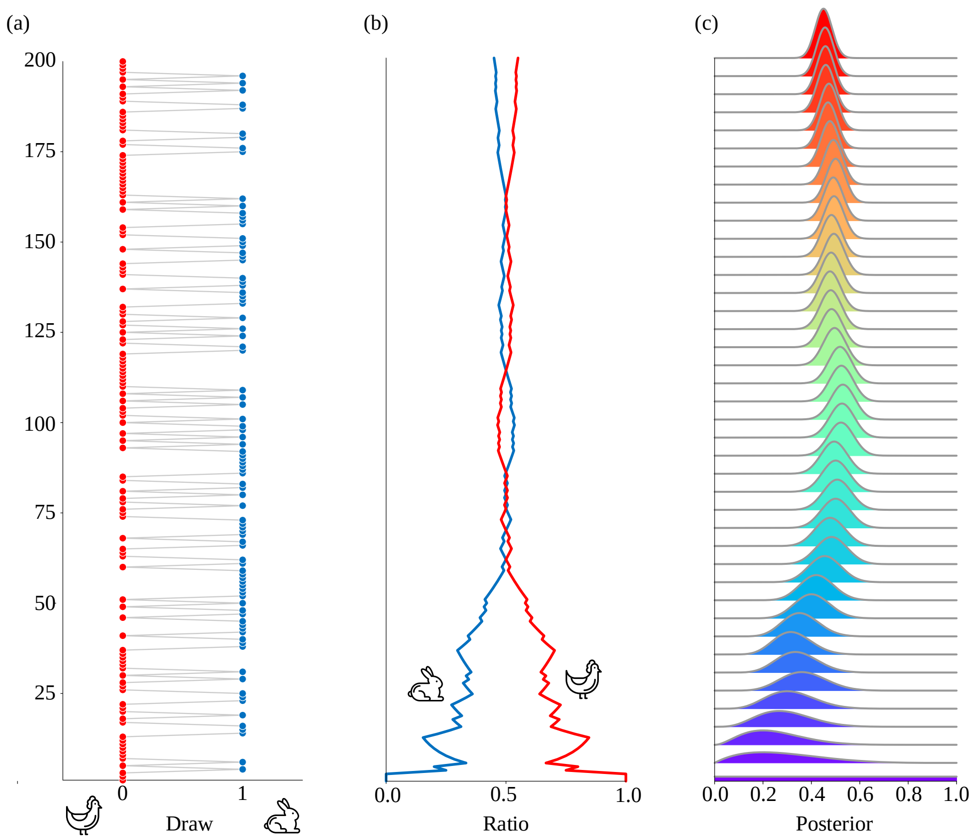


图 8. 某次试验的模拟结果，先验分布为 Beta(1, 1)

图 8 (c) 给出的这个过程中，请大家注意两个细节。

第一，后验概率分布 $f_{\Theta|X}(\theta|x)$ 曲线不断变的细高，也就是后验标准差不断变小。这是因为样本数据不断增多，大家对鸡兔比例变得越发“确信”。

第二，后验概率分布 $f_{\Theta|X}(\theta|x)$ 的最大值，也就是峰值，所在位置逐渐逼近鸡兔的真实比例 0.45。第二点在图 9 中看得更清楚。

图 9 (a) 中，先验概率分布为均匀分布，这代表老农对鸡兔比例一无所知。兔子的比例在 0 和 1 之间，任何值皆有可能，而且可能性均等。

图 9 (b) 所示为，抓到第一只动物发现是鸡。利用贝叶斯定理，通过图 9 (a) 的先验概率 (连续均匀分布 Beta(1,1)) 和样本数据 (一只鸡)，计算得到图 9 (b) 所示的后验概率分布 Beta(1,2)，这一过程如图 10 所示。

图 9 (b) 这个分布显然认为“农场全是鸡”的可能性更高，但是不排除其他可能。“不排除其他可能”对应图 9 (b) 的三角形，θ 在 [0, 1) 区间取值时，后验概率 $f_{\Theta|X}(\theta|x)$ 都不为 0。

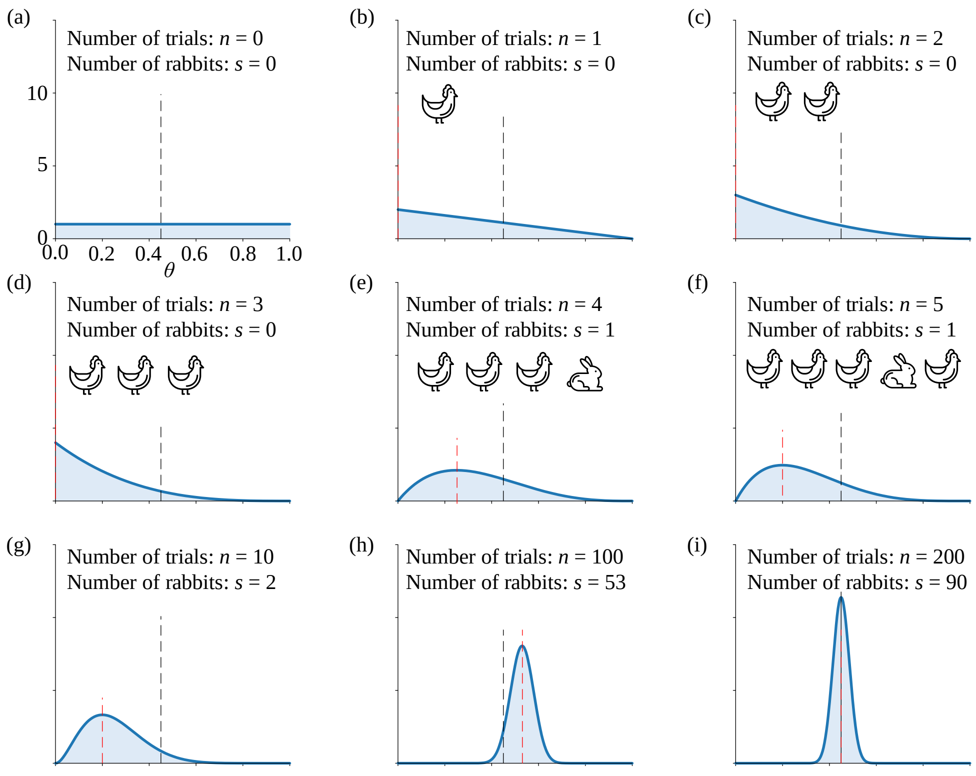


图 9. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(1, 1)

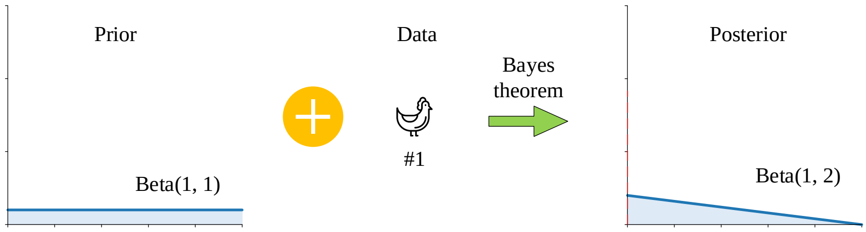


图 10. 不确定鸡兔比例，先验概率 Beta(1, 1) + 一只鸡 (数据) 推导得到后验概率 Beta(1, 2)

抓第二只动物，发现还是鸡。如图 9 (c) 后验概率分布所示，显然农夫心中的天平发生倾斜，认为农场的鸡的比例肯定很高。

获得图 9 (c) 的后验概率分布有两条路径。

第一条如图 11 所示，先验概率 Beta(1, 1) + 两只鸡 (数据) 推导得到后验概率 Beta(1, 3)。

第二条如图 12 所示，更新先验概率 Beta(1, 2) + 第二只鸡 (数据) 推导得到后验概率 Beta(1, 3)。而更新先验概率 Beta(1, 2) 就是图 10 中的后验概率。

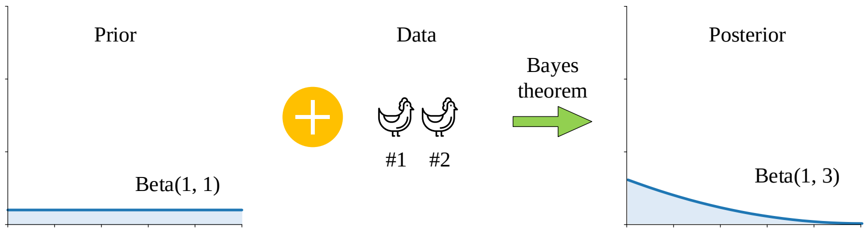


图 11. 第一条路径：先验概率 Beta(1, 1) + 两只鸡 (数据) 推导得到后验概率 Beta(1, 3)

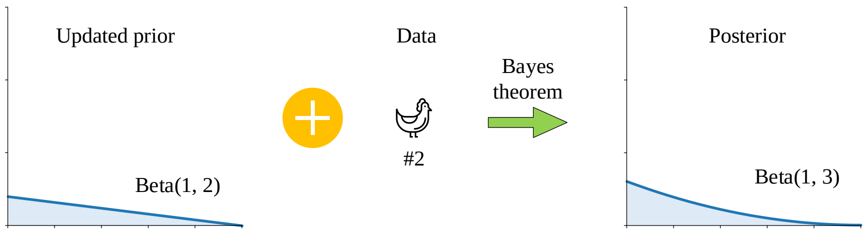


图 12. 第二条路径：更新先验概率 Beta(1, 2) + 第二只鸡 (数据) 推导得到后验概率 Beta(1, 3)

抓第三只动物，竟然还是鸡！如图 9 (d) 所示，农夫心中比例进一步向“鸡”倾斜，但是仍然不能排除其他可能。

理解这步运算则有三条路径！图 13 所示为三条路径中的第一条，请大家自己绘制另外两条。

如果采样此时停止，依照频率派的观点，农场 100%都是鸡。

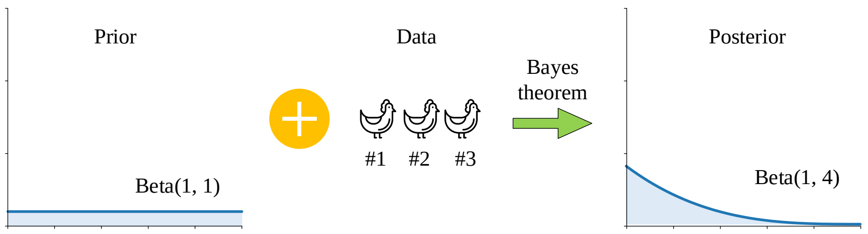


图 13. 先验概率 Beta(1, 1) + 三只鸡 (数据) 推导得到后验概率 Beta(1, 4)

抓第四只动物时，终于抓住一只兔子！农夫才确定农场不都是鸡，确信还是有兔子！观察图 9 (e) 会发现，$\theta = 0$，即兔子比例为 0，对应的概率密度骤降为 0。

随着抓到的动物不断送来验明正身，农夫的“后验概率”、“先验概率”依次更新。最终，在抓获的 200 只动物中，有 90 只兔子，也就是说兔子比例 45%。但是观察图 9 (i) 的后验概率曲线，发现 θ = 45%左右的其他 θ 值也不小。从农夫的视角，农场的鸡兔比例很可能是 45%，但是不排除其他比例的可能性，也就是贝叶斯推断的结论观点。

此外，图 9 (i) 的后验概率的“高矮胖瘦”，也决定了对结论观点的“确信度”。本章后文将展开讲解。

最大化后验概率 MAP

图 9 中黑色划线为农场兔子的真实比例。

而图 9 各个子图中红色划线对应的就是后验概率分布的最大值。这便对应贝叶斯推断的优化问题，**最大化后验概率** (Maximum A Posteriori estimation, MAP)：

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} f_{\Theta|X}\left(\theta \mid x\right) \tag{16}$$

将 (1) 代入上式：

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} \frac{f_{X|\Theta}\left(x \mid \theta\right) f_{\Theta}\left(\theta\right)}{\int_{\vartheta} f_{X|\Theta}\left(x \mid \vartheta\right) f_{\Theta}\left(\vartheta\right) \mathrm{d}\vartheta} \tag{17}$$

进一步根据 (6)，这个优化问题可以简化为：

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} f_{X|\Theta}\left(x \mid \theta\right) f_{\Theta}\left(\theta\right) \tag{18}$$

本书第 7 章介绍过 Beta(α, β) 分布的众数为：

$$\frac{\alpha - 1}{\alpha + \beta - 2}, \quad \alpha, \beta > 1 \tag{19}$$

对于本节例子，MAP 的优化解为 Beta(s + 1, n – s + 1) 的众数，即概率密度最大值：

$$\hat{\theta}_{\text{MAP}} = \frac{s + 1 - 1}{s + 1 + n - s + 1 - 2} = \frac{s}{n} \tag{20}$$

兜兜转转，结果这个贝叶斯派 MAP 优化解和频率派 MLE 一致？

MAP 和 MLE 当然不同！

首先，MAP 和 MLE 的优化问题完全不一样，两者分析问题的视角完全不同。回顾 MLE 优化问题：

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} \prod_{i=1}^{n} f_{X_i}\left(x_i; \theta\right) \tag{21}$$

请大家自行对比 (16) 和 (21)。

此外，(20) 中这个比例是在先验概率为 uniform(0, 1) 条件下得到的，下一节大家会看到不同的 MAP 优化结果。

更重要的是，贝叶斯派得到的结论是图 9 (i) 中这个分布。也就是说，最优解虽然在 Θ = 0.45，但是不排除其他可能。

以图 9 (i) 为例，本例中贝叶斯派得到的参数 Θ 为 Beta(s + 1, n – s + 1) 这个分布。代入具体数据 (n = 200, s = 90)，贝叶斯推断的结果为 Beta(91, 111)，整个过程如图 14 所示。

图 14 中，先验分布为 Beta(1, 1)，括号内的样本数据为 (兔，鸡)，即 (90, 110)，获得的后验概率为 Beta(1 + 90, 1 + 110)。Beta(1 + 90, 1 + 110) 的标准差可以度量我们对贝叶斯推断结论的确信程度，这是本章最后要讨论的话题之一。

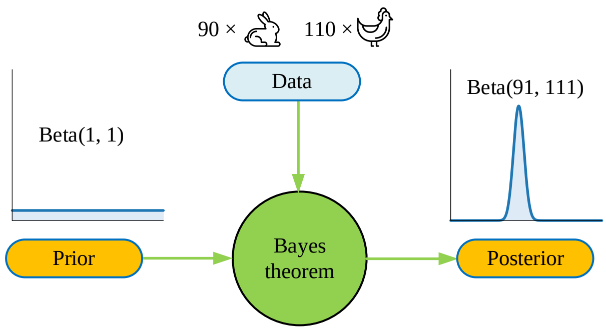


图 14. 先验 Beta(1, 1) + 样本 (90, 110) → 后验 Beta(91, 111)

先验分布的选择和参数的确定代表“经验”，也代表某种“信念”。先验分布的选择和样本数据无关，不需要通过样本数据构造。反过来，观测到的样本数据对先验的选择没有任何影响。

此外，讲解图 12 时，我们看到贝叶斯推断可以采用迭代方式，即后验概率可以成为新样本数据的先验概率。

20.4 走地鸡兔：很可能一半一半

本节我们更换场景，假设农夫认为鸡兔的比例接近 1:1，也就是说，兔子的比例为 50%。但是，农夫对这个比例的确信程度不同。

先验

由于农夫认为鸡兔的比例为 1:1，我们选用 Beta(α, α) 作为先验分布。Beta(α, α) 具体的概率密度函数为：

$$f_{\Theta}(\theta)=\frac{1}{\mathrm{B}(\alpha,\alpha)}\theta^{\alpha-1}(1-\theta)^{\alpha-1} \tag{22}$$

其中，Beta(α, α) 为：

$$\mathrm{B}(\alpha,\alpha)=\frac{\Gamma(\alpha)\Gamma(\alpha)}{\Gamma(\alpha+\alpha)} \tag{23}$$

再次强调，选取 Beta(α, α) 和样本无关，Beta(α, α) 代表事前主观经验。

不同确信程度

图 15 所示为 α 取不同值时 Beta(α, α) 分布 PDF 图像。

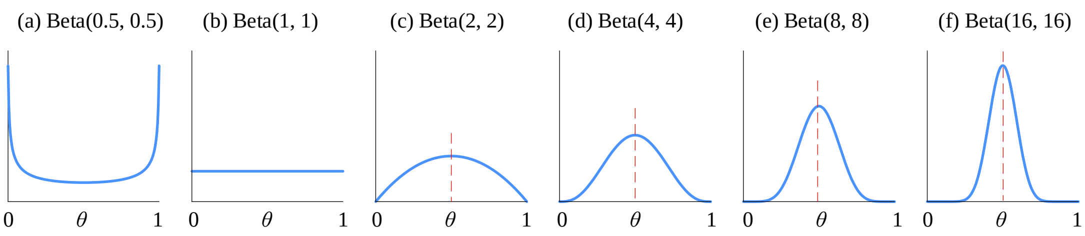


图 15. 五个不同参数 α 取不同值时 Beta(α, α) 分布 PDF 图像

容易发现发现 Beta(α, α) 图像为对称，Beta(α, α) 的均值和众数为 1/2，方差为 $1/(8\alpha + 4)$。显然，参数 α 小于 1 不合适。

α 等于 1 就是本章前文的先验分布为 uniform(0, 1)，即 Beta(1, 1)，假设条件。也就是说，当我们事先对比例不持立场，对 [0, 1] 范围内任何一个 θ 值不偏不倚，Beta(1, 1) 就是最佳的先验分布。

而 α 取不同大于 1 的值时，代表农夫的对鸡兔比例 1:1 的确信程度。如图 16 所示，α 越大 Beta(α, α) 的方差越小，这意味着先验分布的图像越窄、越细高，这代表农夫对兔子比例为 50%这个观点的确信度越高。本章后文会用 Beta 分布的标准差作为“确信程度”的度量，原因是标准差和众数、均值量纲一致。

本节后续的蒙特卡洛模拟中参数 α 的取值分为 2、16 两种情况。α = 2 代表农夫认为兔子的比例大致 50%，但是确信度不高。α = 16 则对应农夫认为兔子的比例很可能 50%，但是绝不排除其他比例的可能性，确信度相对高很多。

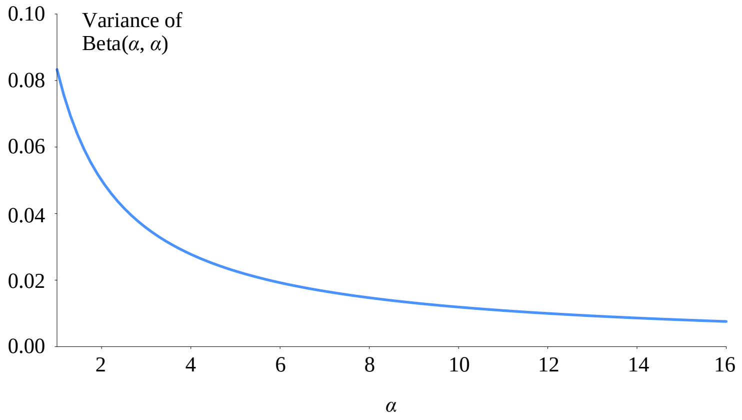


图 16. Beta(α, α) 方差随参数 α 变化

似然

和前文一致，给定 $\Theta = \theta$ 条件下，X_1、X_2 … X_n服从 IID 的伯努利分布 Bernoulli(θ)，即：

$$\underbrace{f_{X_i|\Theta}\left(x_i \mid \theta\right)}_{\text{Likelihood}} = \theta^{x_i}\left(1-\theta\right)^{1-x_i} \tag{24}$$

似然函数为：

$$f_{X_1,X_2,\dots,X_n|\Theta}\left(x_1,x_2,\dots,x_n \mid \theta\right) = \theta^{s}\left(1-\theta\right)^{n-s} \tag{25}$$

大家可能已经发现，(25) 本质上就是二项分布。二项分布是若干独立的伯努利分布。我们把似然分布记做 $f_{X|\Theta}\left(x \mid \theta\right)$：

$$f_{X|\Theta}\left(x \mid \theta\right) = C_n^s \cdot \theta^{s}\left(1-\theta\right)^{n-s} \tag{26}$$

C_n^s 和 θ 无关，(46) 和 (26) 成正比关系。也就是说，C_n^s 仅仅提供缩放。

本书第 5 章中，我们这样解读二项分布。给定任意一次试验成功的概率为 θ，(26) 计算 n 次试验中 s 次成功的概率。对于本例，(26) 的含义是，给定兔子的占比为 θ，n 只动物中正好有 s 只兔子的概率。

本章中，我们需要换一个视角理解 (26)。它是给定 n 次试验中 s 次成功，而 θ 变化导致概率的变化。而 θ 是在 (0, 1) 区间上连续变化。

图 17 (a) 所示为一组似然分布，其中 n = 20，这些曲线 s 的取值为 1 ~ 19 整数。θ 是在 (0, 1) 区间上连续变化。

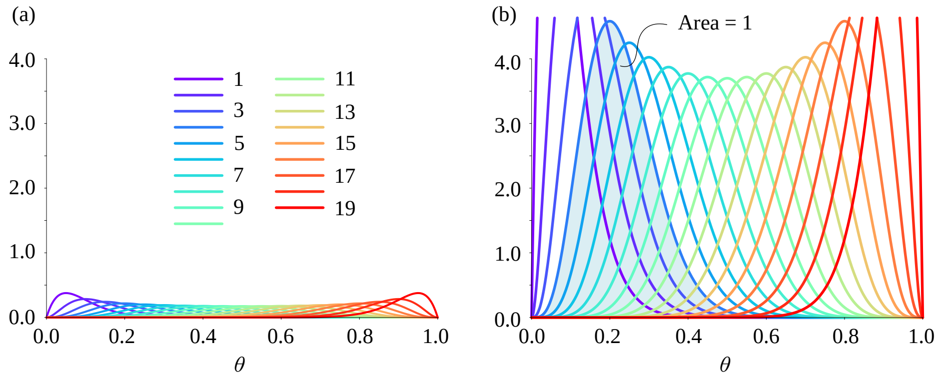


图 17. 似然分布，$n = 20$

注意，似然函数本身是关于 θ 的函数，和先验分布 Beta(α, α) 中的 α 无关。似然函数值通常是很小的数，所以我们一般会取对数 ln() 获得对数似然函数。

为了和先验分布、后验分布直接比较，需要归一化 (26)：

$$f_{X|\Theta}(x|\theta) = \frac{\overbrace{C_n^s \theta^s (1-\theta)^{n-s}}^{\text{Binomial distribution}}}{C_n^s \int_\theta \theta^s (1-\theta)^{n-s} \,\mathrm{d}\theta} \tag{27}$$

这样似然函数曲线和横轴围成的面积也是 1。

前文提过，(27) 的分子可以视作二项分布。利用 Beta 函数，(27) 的分母可以进一步化简：

$$C_n^s \int_\theta \theta^s (1-\theta)^{n-s} \,\mathrm{d}\theta = C_n^s \cdot \mathrm{B}(s+1, n-s+1) = \frac{n!}{s!(n-s)!} \frac{s!(n-s)!}{(n+1)!} = \frac{1}{n+1} \tag{28}$$

上式就是似然函数的归一化因子。图 17 (b) 所示为归一化后的似然分布。当然我们也可以用数值积分归一化似然函数。

因此，(27) 可以写成：

$$f_{X|\Theta}(x|\theta) = (n+1) \cdot \overbrace{C_n^s \theta^s (1-\theta)^{n-s}}^{\text{Binomial distribution}} \tag{29}$$

在本书第 17 章中，我们知道似然函数的最大值位置为 s/n，也就是最大似然估计 MLE 的解，具体位置如图 18 所示。注意图 18 中，s 为 0 ~ 20 的整数。

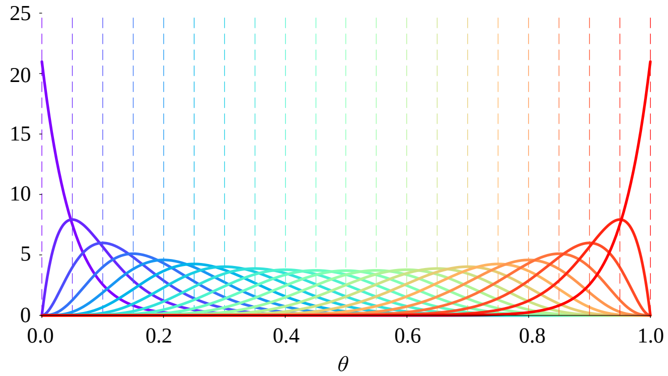


图 18. 似然分布和 MLE 优化解的位置，$n = 20$

再换个视角，看到 (25) 这种形式，大家是否立刻想到，这不正是一个 Beta 分布！缺的就是归一化系数！补齐这个归一化系数，我们便得到 Beta(s + 1, n – s + 1) 分布：

$$\frac{\Gamma(s+1+n-s+1)}{\Gamma(s+1)\Gamma(n-s+1)}\theta^{s+1-1}(1-\theta)^{n-s+1-1} = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)}\theta^{s+1-1}(1-\theta)^{n-s+1-1} \tag{30}$$

而 Beta(s + 1, n – s + 1) 分布的众数位置为：

$$\frac{s+1-1}{s+1+n-s+1-2} = \frac{s}{n} \tag{31}$$

这和之前的结论一致。请大家自己绘制 n = 20、s 为 0 ~ 20 整数时，Beta(s + 1, n – s + 1) 的 PDF 曲线，并和图 18 比较。

回看 (14)，本节的似然分布 Beta(s + 1, n – s + 1) 相当于对鸡兔比例“不持立场”，一切均以客观样本数据为准。

再换个角度来看，上述讨论似乎说明，贝叶斯推断“包含了”频率推断。MLE 是 MAP 的特例 (无信息先验)。

先验 vs 似然

图 19 中灰色曲线对应“归一化”的似然分布 $f_{X|\Theta}(x|\theta)$，它相当于 Beta(s + 1, n – s + 1)。灰色划线对应 MLE 的解，$f_{X|\Theta}(x|\theta)$ 的最大值。

图 19 中粉色曲线对应 $f_{\Theta}(\theta)$，即 Beta(α, α)。如 (22) 所示，$f_{\Theta}(\theta)$ 和 α 有关；α 越大，$f_{\Theta}(\theta)$ 曲线越细高。$f_{\Theta}(\theta)$ 曲线的最大值是 Beta(α, α) 的众数，θ = 1/2。

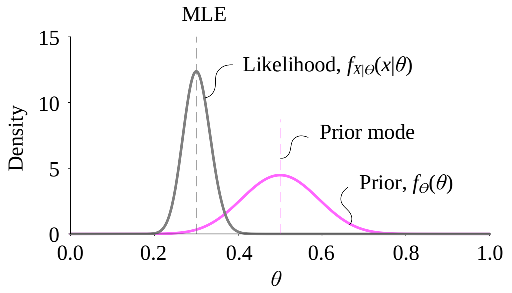


图 19. 对比先验分布、似然分布，$\alpha = 16$

联合

联合分布为：

$$
\begin{aligned}
f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right) &= \underbrace{f_{X_1,X_2,\dots,X_n|\Theta}\left(x_1,x_2,\dots,x_n \mid \theta\right)}_{\text{Likelihood}} \underbrace{f_{\Theta}\left(\theta\right)}_{\text{Prior}} \\
&= \theta^{s}\left(1-\theta\right)^{n-s} \frac{1}{\mathrm{B}\left(\alpha,\alpha\right)} \theta^{\alpha-1}\left(1-\theta\right)^{\alpha-1} \\
&= \frac{1}{\mathrm{B}\left(\alpha,\alpha\right)} \theta^{s+\alpha-1}\left(1-\theta\right)^{n-s+\alpha-1}
\end{aligned} \tag{32}
$$

证据

证据因子 $f_{X_1,X_2,\dots,X_n}\left(x_1,x_2,\dots,x_n\right)$ 可以通过 $f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right)$ 对 θ“偏积分”得到：

$$
\begin{aligned}
f_{X_1,X_2,\dots,X_n}\left(x_1,x_2,\dots,x_n\right) &= \int_{\theta} f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right)\mathrm{d}\theta \\
&= \frac{1}{\mathrm{B}\left(\alpha,\alpha\right)} \int_{\theta} \theta^{s+\alpha-1}\left(1-\theta\right)^{n-s+\alpha-1}\mathrm{d}\theta \\
&= \frac{\mathrm{B}\left(s+\alpha,n-s+\alpha\right)}{\mathrm{B}\left(\alpha,\alpha\right)}
\end{aligned} \tag{33}
$$

后验

在 $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ 条件下，Θ 的后验分布为：

$$
\begin{aligned}
f_{\Theta|X_1,X_2,\dots,X_n}\left(\theta \mid x_1,x_2,\dots,x_n\right) &= \frac{f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right)}{f_{X_1,X_2,\dots,X_n}\left(x_1,x_2,\dots,x_n\right)} \\
&= \frac{\frac{1}{\mathrm{B}\left(\alpha,\alpha\right)}\theta^{s+\alpha-1}\left(1-\theta\right)^{n-s+\alpha-1}}{\frac{\mathrm{B}\left(s+\alpha,n-s+\alpha\right)}{\mathrm{B}\left(\alpha,\alpha\right)}} = \frac{\theta^{s+\alpha-1}\left(1-\theta\right)^{n-s+\alpha-1}}{\mathrm{B}\left(s+\alpha,n-s+\alpha\right)}
\end{aligned} \tag{34}
$$

上式对应 Beta($s + \alpha$, $n - s + \alpha$) 分布。

幸运的是，我们实际上“避开”(33) 这个复杂积分。但是，并不是所有情况都存在积分的**闭式解** (closed form solution)，也叫**解析解** (analytical solution)。本书第 22 章将介绍蒙特卡洛模拟方式近似获得后验分布。

先验 vs 似然 vs 后验

图 20 对比对比先验分布 Beta(α, α)、似然分布 Beta($s + 1$, $n - s + 1$)、后验分布 Beta($s + \alpha$, $n - s + \alpha$)。

比较这三个分布，直觉告诉我们后验分布 Beta($s + \alpha$, $n - s + \alpha$) 好像是先验分布 Beta(α, α)、似然分布 Beta($s + 1$, $n - s + 1$) 的某种“糅合”！本章最后会继续这个思路探讨贝叶斯推断。

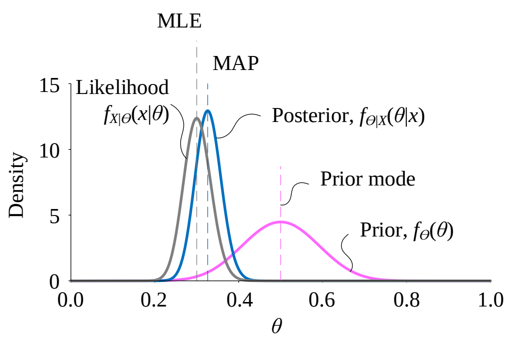


图 20. 对比先验分布、似然分布、后验分布，$\alpha = 16$

正比关系

类似 (15)，后验概率存在如下正比关系：

$$f_{\Theta|X_1,X_2,\ldots,X_n}\left(\theta \mid x_1,x_2,\ldots,x_n\right) \propto f_{X_1,X_2,\ldots,X_n|\Theta}\left(x_1,x_2,\ldots,x_n \mid \theta\right) f_{\Theta}\left(\theta\right) = \theta^{s+\alpha-1}\left(1-\theta\right)^{n-s+\alpha-1} \tag{35}$$

蒙特卡罗模拟：确信度不高

前文提到，农夫认为农场兔子的比例大致为 50%，因此我们选择 Beta(α, α) 作为先验概率分布。下面的蒙特卡罗模拟中，我们设定 $\alpha = 2$。

图 21 (a) 所示为伯努利随机数发生器产生的随机数。和前文一样，0 代表鸡，1 代表兔。不同的是，我们设定兔子的真实比例为 0.3。

如图 21 (b) 所示，随着样本数 n 增大，鸡兔的比例趋于稳定。

图 21 (c) 所示为后验概率分布随着 n 的变化。自下而上，后验概率曲线从平缓逐渐过渡到细高，这代表确信度不断升高。

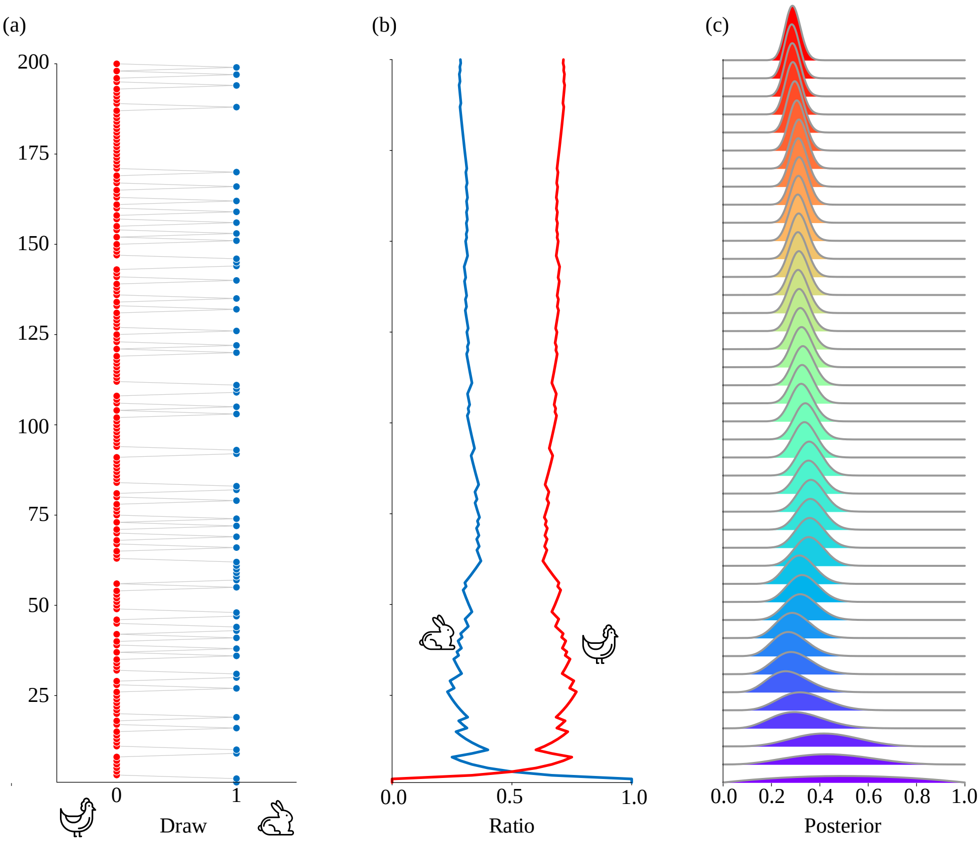


图 21. 某次试验的模拟结果，先验分布为 Beta(2, 2)

图 22 所示为九张不同节点的后验概率分布曲线快照。

图 22 (a) 代表农夫最初的先验概率 Beta(2, 2)。Beta(2, 2) 曲线关于 $\theta = 0.5$ 对称，并在 $\theta = 0.5$ 取得最大值。Beta(2, 2) 很平缓，这代表农夫对 50%的比例不够确信。

抓到第一只动物是兔子，这个样本导致图 22 (b) 中后验概率最大值向右移动。请大家自己写出后验 Beta 分布的参数。

抓到的第二只动物还是兔子，后验概率最大值进一步向右移动，具体如图 22 (c) 所示。

第三只动物是鸡，后验概率最大值所在位置向左移动了一点。

请大家自行分析图 22 剩下几幅子图，注意后验概率形状、最大值位置变化。

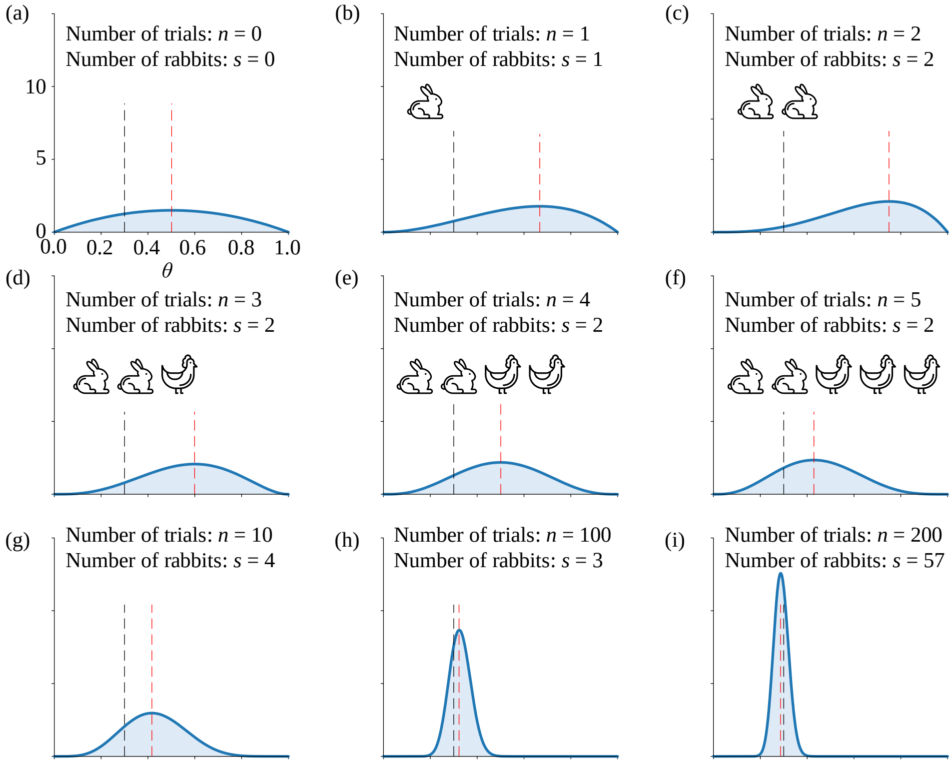


图 22. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(2, 2)

蒙特卡罗模拟：确信度很高

α = 16 则对应农夫认为兔子的比例很可能 50%，但是绝不排除其他比例的可能性，确信度相对高很多。请大家对比前文蒙特卡洛模拟结果，自行分析图 23 和图 24。强烈建议大家把图 24 每幅子图的 Beta 分布的参数写出来。

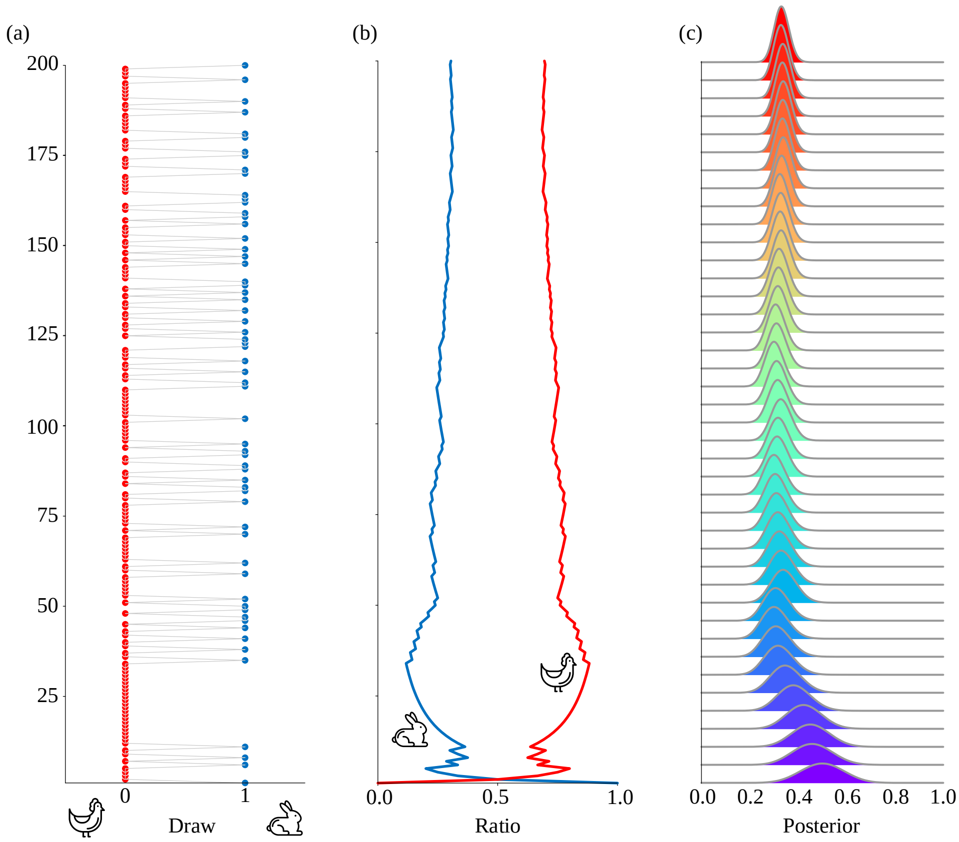


图 23. 某次试验的模拟结果，先验分布为 Beta(16, 16)

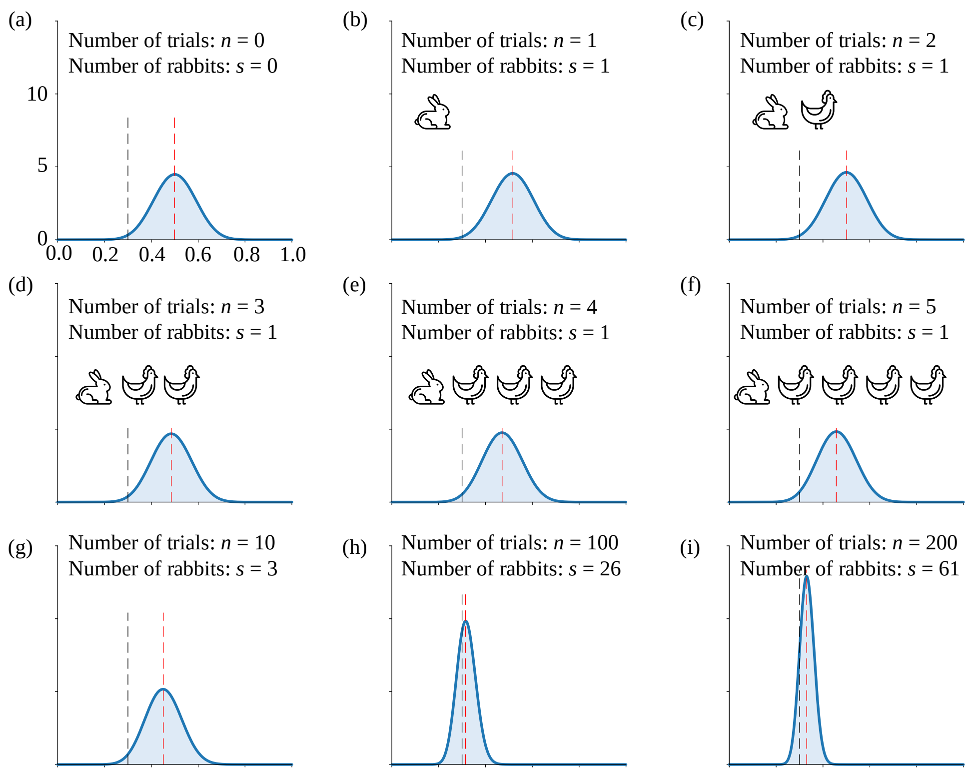


图 24. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(16, 16)

代码 Bk5_Ch20_01.py 完成本章前文蒙特卡洛模拟和可视化。

最大后验 MAP

Beta($s + \alpha$, $n - s + \alpha$) 的众数，即 MAP 的优化解，为：

$$\hat{\theta}_{\text{MAP}} = \frac{s+\alpha-1}{n+2\alpha-2} \tag{36}$$

特别地，当 $\alpha = 1$ 时，MAP 和 MLE 的解相同，即：

$$\hat{\theta}_{\text{MAP}} = \hat{\theta}_{\text{MLE}} = \frac{s}{n} \tag{37}$$

图 25 对比 α 取不同值时先验分布、似然分布、后验分布。先验分布 Beta(α, α) 中 α 越大，代表主观经验越发“先入为主”，对贝叶斯推断最终结果越强。表现在图 25 中就是，随着 α 增大，似然分布和后验分布差异越大，MAP 优化解越发偏离 MLE 优化解。

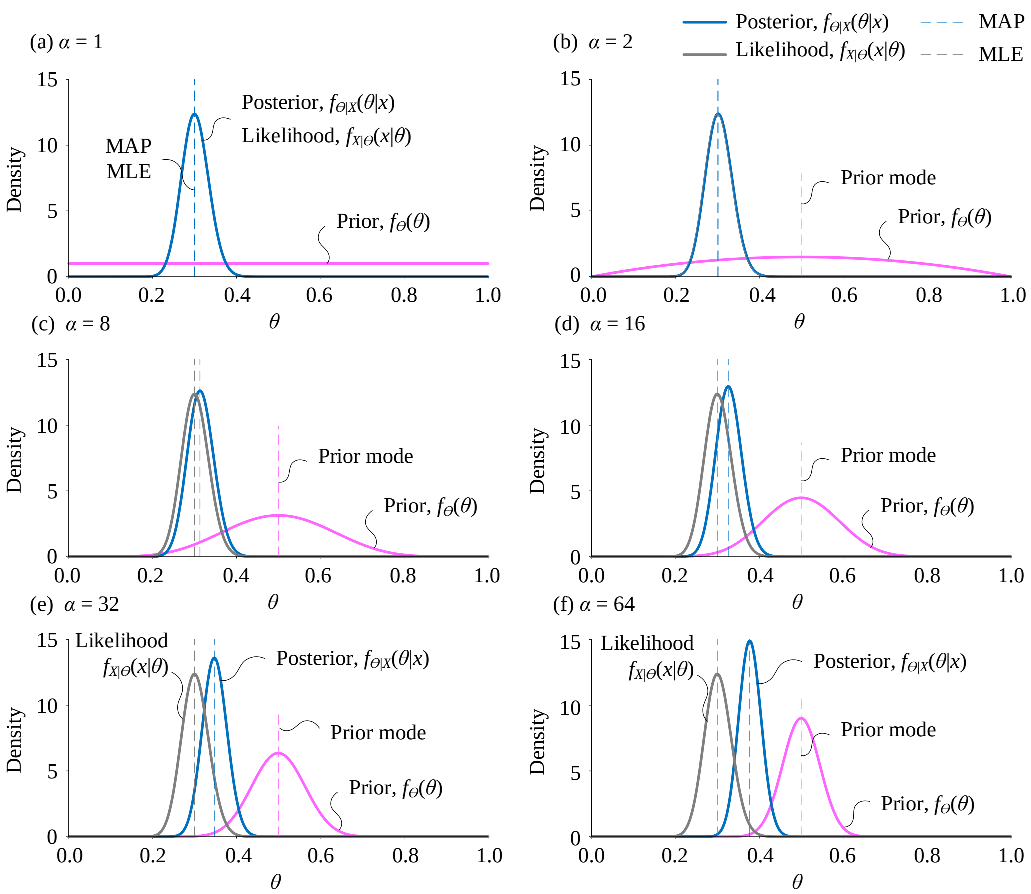


图 25. 对比先验分布、似然分布、后验分布，α 取不同值时

图 26 和图 27 以另外一种可视化方案对比 α 取不同值时先验分布对后验分布的影响。

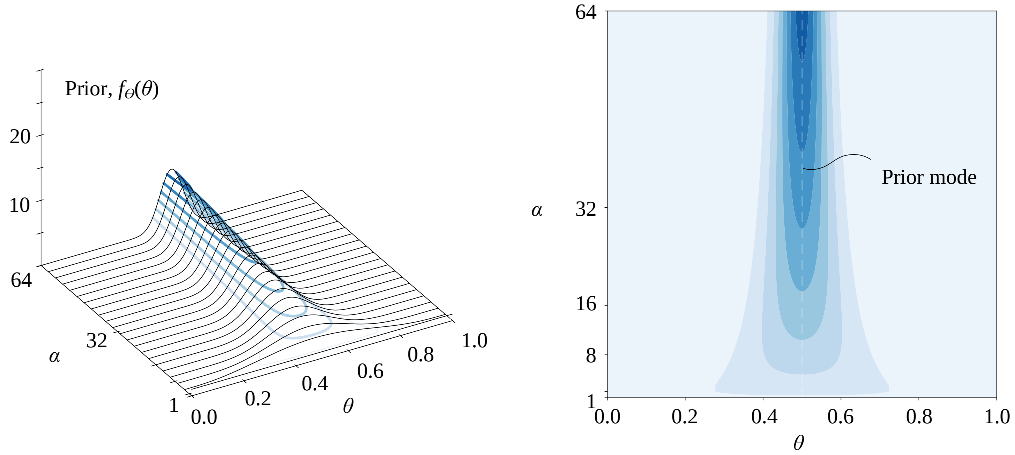


图 26. 先验分布，α 取不同值时

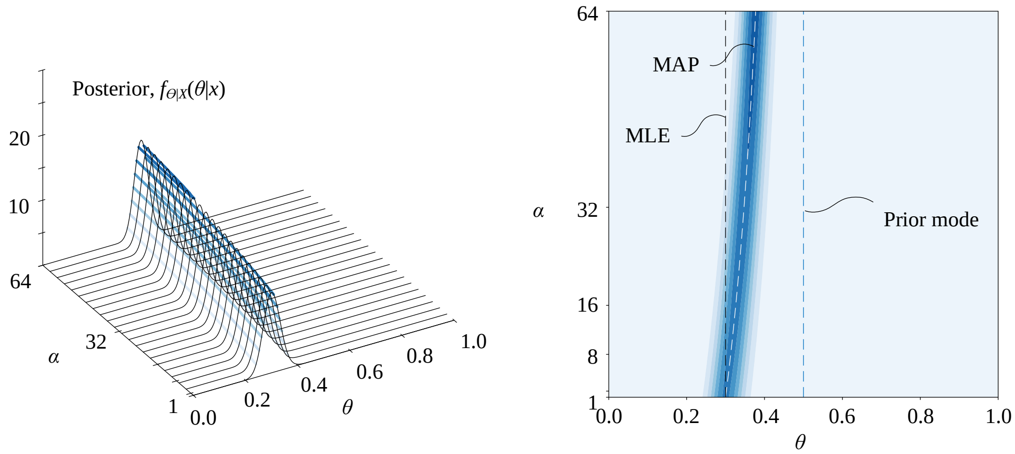


图 27. 后验分布，α 取不同值时

代码 Bk5_Ch021_02.py 绘制图 25、图 26、图 27。

20.5 走地鸡兔：更一般的情况

有了前文的两个例子，下面我们看一下更为一般的情况。

先验

选用 Beta(α, β) 作为先验分布。Beta(α, β) 具体的概率密度函数为：

$$f_\Theta(\theta) = \frac{1}{\mathrm{B}(\alpha,\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \tag{38}$$

先验分布 Beta(α, β) 的众数为：

$$\frac{\alpha-1}{\alpha+\beta-2},\quad \alpha,\beta>1 \tag{39}$$

其他比例

举个例子，假设农夫认为兔子比例为 1/3，则：

$$\frac{\alpha-1}{\alpha+\beta-2}=\frac{1}{3} \tag{40}$$

即 α 和 β 关系为：

$$\beta=2\alpha-1 \tag{41}$$

图 28 所示为 α 和 β 取不同值时 Beta(α, β) 分布 PDF 图像。这些图像有一个共同特点，众数都是 1/3。

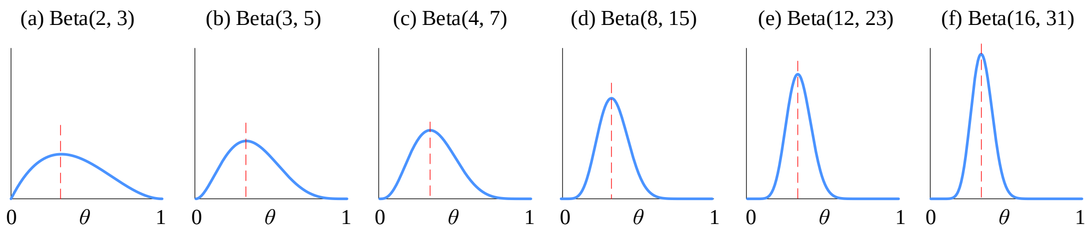


图 28. 五个不同 Beta(α, β) 分布 PDF 图像，众数都是 1/3

如果农夫认为兔子比例为 1/4，则：

$$\frac{\alpha-1}{\alpha+\beta-2}=\frac{1}{4} \tag{42}$$

即 α 和 β 关系为：

$$\beta=3\alpha-2 \tag{43}$$

满足上式条件下，当 α 不断增大，兔子的比例虽然还是 1/4，但是如图 29 所示，先验分布变得越发细高，这代表着确信程度提高，“信念”增强。

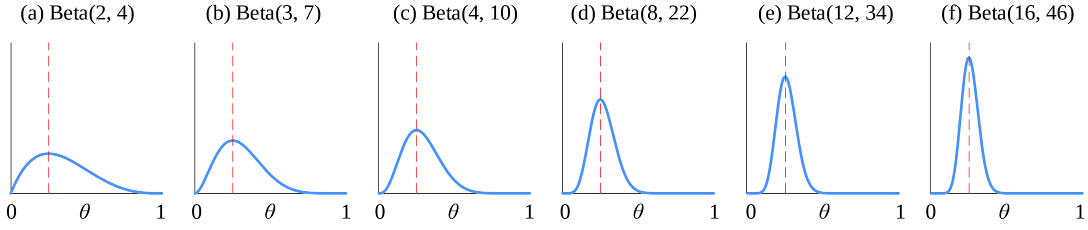


图 29. 五个不同 Beta(α, β) 先验分布 PDF 图像，众数都是 1/4

确信程度

我们可以用 Beta(α, β) 分布的标准差量化所谓“确信程度”。

Beta(α, β) 的标准差为：

$$\operatorname{std}(X)=\sqrt{\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}} \tag{44}$$

如果 α、β 满足 (43) 等式，Beta(α, β) 的标准差随 α 变化如图 30 所示。更准确地说，随着标准差减小，对比例的“怀疑程度”不断减小。

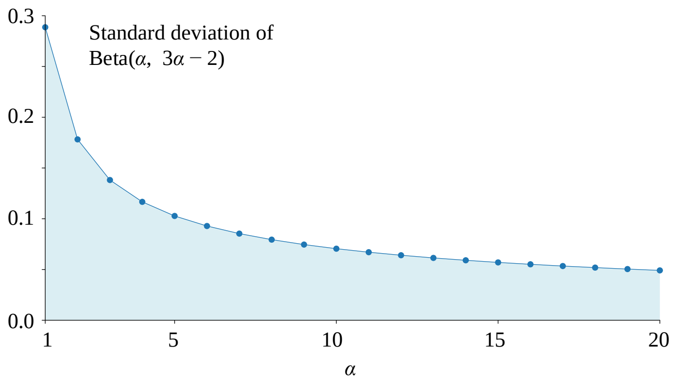


图 30. 随着 α 增大，“怀疑程度”不断减小

换一个方式，为了方便和下一章的 Dirichlet 分布对照，令 $\alpha_0=\alpha+\beta$，Beta(α, β) 的均方差可以进一步写成：

$$\operatorname{std}(X)=\sqrt{\frac{\alpha/\alpha_0\left(1-\alpha/\alpha_0\right)}{\alpha_0+1}} \tag{45}$$

α/α_0 也可以看做兔子的比例。不同的是，α/α_0 代表 Beta(α, β) 的期望 (均值)，不是众数。下一章会比较 Beta 分布的期望和均值。

图 31 所示一组图像代表比例和确信度同时变化。

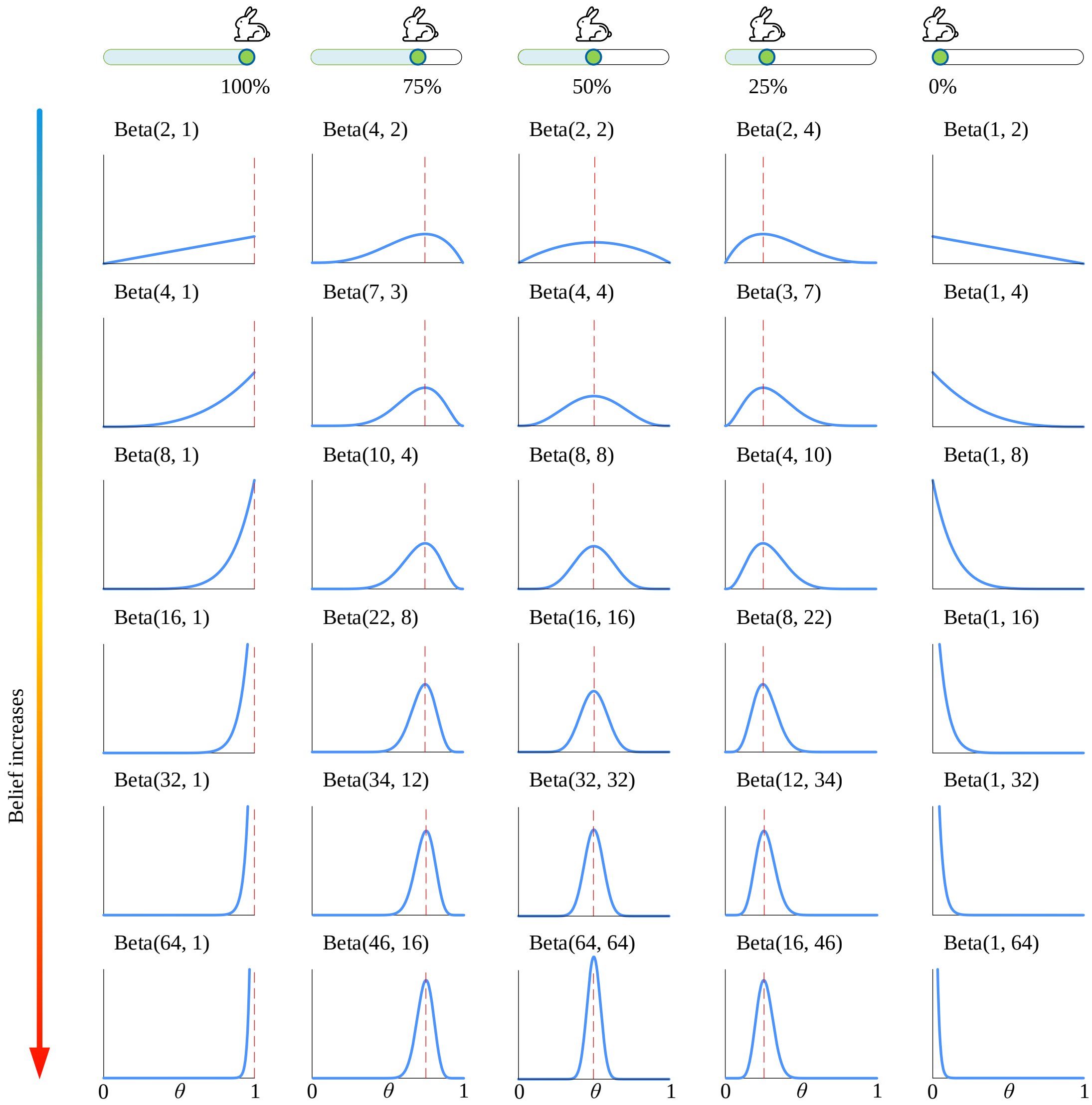


图 31. 比例和确信程度同时变化

似然

和前文一致，似然函数为：

$$f_{X_1,X_2,\ldots,X_n|\Theta}\left(x_1,x_2,\ldots,x_n \mid \theta\right)=\theta^s\left(1-\theta\right)^{n-s} \tag{46}$$

本章前文介绍过，似然函数可以看成 IID 伯努利分布、二项分布，甚至用 Beta 分布代替。

联合

因此，联合分布为：

$$
\begin{aligned}
f_{X_1,X_2,\dots,X_n,\Theta}(x_1,x_2,\dots,x_n,\theta) &= \underbrace{f_{X_1,X_2,\dots,X_n|\Theta}(x_1,x_2,\dots,x_n\,|\,\theta)}_{\text{Likelihood}}\underbrace{f_{\Theta}(\theta)}_{\text{Prior}} \\
&= \theta^{s}(1-\theta)^{n-s}\frac{1}{\mathrm{B}(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}\theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}
\end{aligned} \tag{47}
$$

证据

证据因子 $f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)$ 可以通过 $f_{X_1,X_2,\dots,X_n,\Theta}(x_1,x_2,\dots,x_n,\theta)$ 对 θ"偏积分"得到：

$$
\begin{aligned}
f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n) &= \int_{\theta} f_{X_1,X_2,\dots,X_n,\Theta}(x_1,x_2,\dots,x_n,\theta)\,\mathrm{d}\theta \\
&= \frac{1}{\mathrm{B}(\alpha,\beta)}\int_{\theta}\theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}\,\mathrm{d}\theta \\
&= \frac{\mathrm{B}(s+\alpha,n-s+\beta)}{\mathrm{B}(\alpha,\beta)}
\end{aligned} \tag{48}
$$

后验

在 $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ 条件下，Θ 的后验分布为：

$$
\begin{aligned}
f_{\Theta|X_1,X_2,\dots,X_n}(\theta\,|\,x_1,x_2,\dots,x_n) &= \frac{f_{X_1,X_2,\dots,X_n,\Theta}(x_1,x_2,\dots,x_n,\theta)}{f_{X_1,X_2,\dots,X_n}(x_1,x_2,\dots,x_n)} \\
&= \frac{\frac{1}{\mathrm{B}(\alpha,\beta)}\theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}}{\frac{\mathrm{B}(s+\alpha,n-s+\beta)}{\mathrm{B}(\alpha,\beta)}} = \frac{\theta^{s+\alpha-1}(1-\theta)^{n-s+\beta-1}}{\mathrm{B}(s+\alpha,n-s+\beta)}
\end{aligned} \tag{49}
$$

上式对应 $\mathrm{Beta}(s + \alpha, n - s + \beta)$ 分布。

看到这里，大家肯定会想我们是幸运的，因为我们再次成功地避开了 (48) 这个复杂的积分。而这绝不是巧合！在贝叶斯统计中，如果后验分布 $\mathrm{Beta}(s + \alpha, n - s + \beta)$ 与先验分布 $\mathrm{Beta}(\alpha, \beta)$ 属于同类，则先验分布与后验分布被称为**共轭分布** (conjugate distribution 或 conjugate pair)，而先验分布被称为似然函数的**共轭先验** (conjugate prior)。下一章还会探讨这一话题。

贝叶斯收缩

$\mathrm{Beta}(s + \alpha, n - s + \beta)$ 的众数为：

$$
\frac{s+\alpha-1}{n+\alpha+\beta-2} \tag{50}
$$

我们可以把上式写成两个部分：

$$\begin{aligned}\frac{s+\alpha-1}{n+\alpha+\beta-2}&=\frac{\alpha-1}{n+\alpha+\beta-2}+\frac{s}{n+\alpha+\beta-2}\\&=\frac{\alpha+\beta-2}{n+\alpha+\beta-2}\times\underbrace{\frac{\alpha-1}{\alpha+\beta-2}}_{\text{Prior mode}}+\frac{n}{n+\alpha+\beta-2}\times\underbrace{\frac{s}{n}}_{\text{Sample mean}}\end{aligned}\tag{51}$$

定义权重：

$$\begin{aligned}w&=\frac{\alpha+\beta-2}{n+\alpha+\beta-2}\\1-w&=\frac{n}{n+\alpha+\beta-2}\end{aligned}\tag{52}$$

(51) 可以写成：

$$\frac{s+\alpha-1}{n+\alpha+\beta-2}=w\times\underbrace{\frac{\alpha-1}{\alpha+\beta-2}}_{\text{Prior mode}}+(1-w)\times\underbrace{\frac{s}{n}}_{\text{Sample mean}}\tag{53}$$

随着 n 不断增大，w 趋向于 0，而 $1-w$ 趋向于 1。也就是说，随着样本数据量不断增多，先验的影响力不断减小。$n\to\infty$时，MAP 和 MLE 的结果趋同。

相反，当 n 较小的时候，特别是当 α 和 β 比较大，则先验的影响力很大，MAP 的结果向先验均值“收缩”。这种效果常被称作**贝叶斯收缩** (Bayes shrinkage)。

贝叶斯收缩也可以从期望角度理解。Beta($s+\alpha$, $n-s+\beta$) 的期望也可以写成两部分：

$$\begin{aligned}\frac{s+\alpha}{n+\alpha+\beta}&=\frac{\alpha}{n+\alpha+\beta}+\frac{s}{n+\alpha+\beta}\\&=\frac{\alpha+\beta}{n+\alpha+\beta}\times\underbrace{\frac{\alpha}{\alpha+\beta}}_{\text{Prior mean}}+\frac{n}{n+\alpha+\beta}\times\underbrace{\frac{s}{n}}_{\text{Sample mean}}\end{aligned}\tag{54}$$

从贝叶斯收缩角度，让我们再回过头来看本节上述结果。

首先，换个视角理解先验分布 Beta(α, β) 中的 α 和 β。

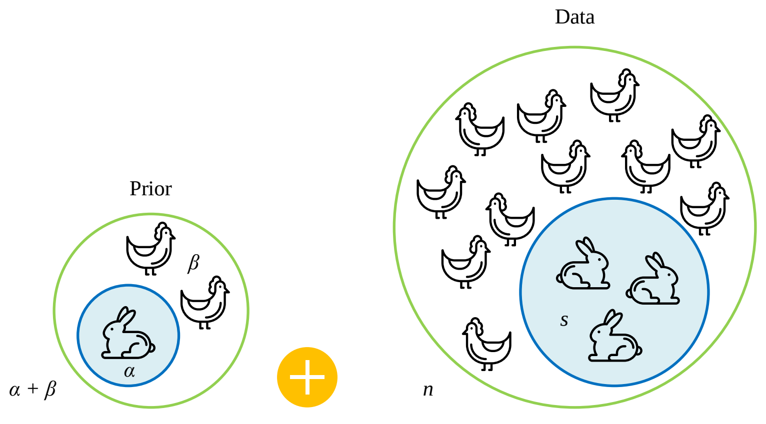


图 32. “混合”先验、样本数据

先验分布中的 α 和 β 之和可以看做“先验”动物总数。即没有数据时，根据先验经验，农夫认为农场动物总数为 $\alpha + \beta$，其中兔子的比例为 $\alpha/(\alpha + \beta)$。

样本数据中，s 代表 n 只动物中兔子的数量，$n - s$ 代表鸡的数量，兔子比例为 s/n。

而 (54) 就可以简单理解成“先验 + 数据”融合得到“后验”。

后验分布 $\text{Beta}(s + \alpha, n - s + \beta)$ 则代表“先验 $\text{Beta}(\alpha, \beta)$ + 数据 $(s, n - s)$”。兔子 α 从增加到 $s + \alpha$，鸡从 β 增加到 $n - s + \beta$。

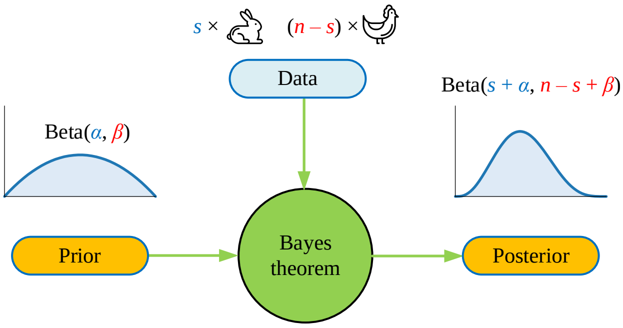


图 33. 先验 $\text{Beta}(\alpha, \beta)$ + 样本 $(s, n - s) \to$ 后验 $\text{Beta}(s + \alpha, n - s + \beta)$

当然，α 和 β 越大，先验的“主观”影响力越大。但是随着样本数量不断增大，先验的影响力逐步下降。当样本数量趋近无穷时，先验不再有任何影响力，MAP 优化解趋向于 MLE 优化解。

换个角度，当我们对参数先验知识模糊不清时，Beta(1, 1) 并非唯一选择。任何 α 和 β 较小的 Beta 分布都可以。因为随着样本数量不断增大，先验分布的较小参数对后验影响微乎其微。

有趣的是，贝叶斯推断所体现出来的“学习过程”和人类认知过程极为相似。

总结来说，贝叶斯推断的过程包括以下几个步骤：1) 确定模型和参数空间，建立参数的先验分布；2) 收集数据；3) 根据样本数据，计算似然函数；4) 利用贝叶斯定理，将似然函数与先验概率相结合，计算后验概率；5) 根据后验概率，更新先验概率，得到更准确的参数估计。

贝叶斯推断的优点在于其能够利用先验信息和后验概率，通过不断更新来获得更准确的估计结果。本章透过二项比例的贝叶斯推断，以 Beta 分布为先验，以伯努利分布或二项分布作为似然分布，讨论不同参数对贝叶斯推断结果的影响。

请大家格外注意，这仅仅是众多贝叶斯推断中较为简单的一种。虽然以管窥豹，希望大家能通过本章例子理解贝叶斯推断背后的思想，以及整条技术路线。此外，本章和下两章共用一幅思维导图。

本章农场仅仅有鸡、兔，即二元。下一章中，农场又来了猪，贝叶斯推断变成了三元，进一步“升维”。先验分布则变成了 Dirichlet 分布，似然分布为多项分布。