21 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

21.1 贝叶斯推断：更贴合人脑思维

一个让人“头大”的公式

本章的关键就是如何理解、应用以下公式：

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

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，也有把 x 写成 y，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 17 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)。而模型参数的分布随着外部样本数据不断输入而迭代更新。相反，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。

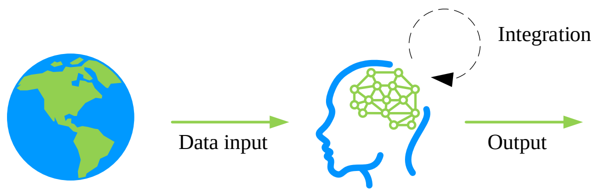

图 1. 人脑更像是一个贝叶斯推断机器

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”的结果肯定基于两方面因素：一方面，日复一日的“找手机”的经验；另一方面，“今早、昨晚在哪最后一次用过手机”的最新数据。

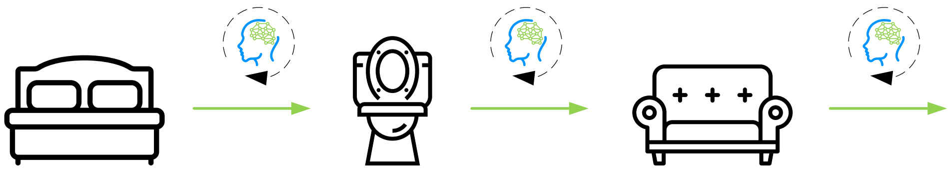

图 2. 找手机

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。

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。随后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 …

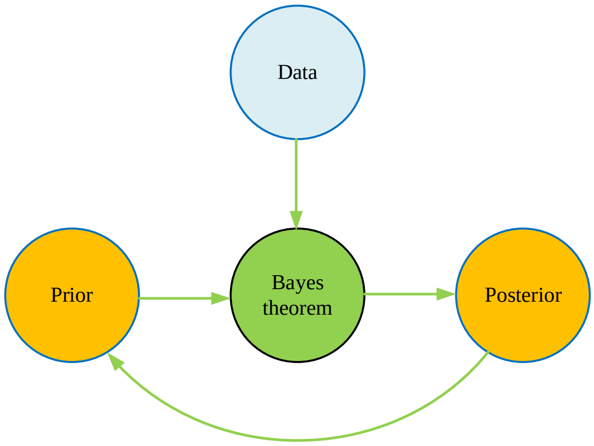

图 3. 通过贝叶斯定理迭代学习

21.2 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 $f_{X|\Theta}(x|\theta)$ 也会有相应变化。

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，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。

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 。

$f_{\Theta|X}(\theta|x)$ 叫后验概率分布 (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 $\int_\theta f_{X|\Theta}(x|\vartheta) f_\Theta(\vartheta)\mathrm{d}\vartheta$ 积分。此外，这个积分很可能没有解析解，可能需要用到数值积分。这是下一章要讲解的内容之一。

21.3 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 17 章“鸡兔同笼”的例子。一个农场散养大量“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 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，先验分布的选择独立于样本数据。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。

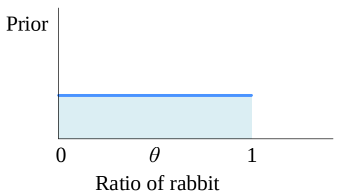

图 4. 选择连续均匀分布作为先验分布

似然

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

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

其中，$\Theta = \theta$ 代表兔子的比例，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。

实际上，上式中似然概率 $f_{X_i|\Theta}(x_i \mid \theta)$ 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。

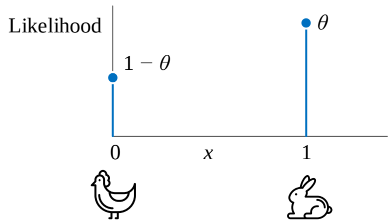

图 5. 似然分布

联合

因此，X_1、X_2 … X_n、Θ 联合分布为：

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

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：

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

证据

证据因子 $f_{X_1,X_2,\dots,X_n}\left(x_1,x_2,\dots,x_n\right)$，即 $f_{\chi}\left(\boldsymbol{x}\right)$，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：

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

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，上述积分很有可能没有解析解。

想到本书第 7 章介绍的 Beta 函数，(12) 可以写成：

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

图 6 所示为 B(s + 1, n – s + 1) 函数随着 s、n 变化的平面等高线。

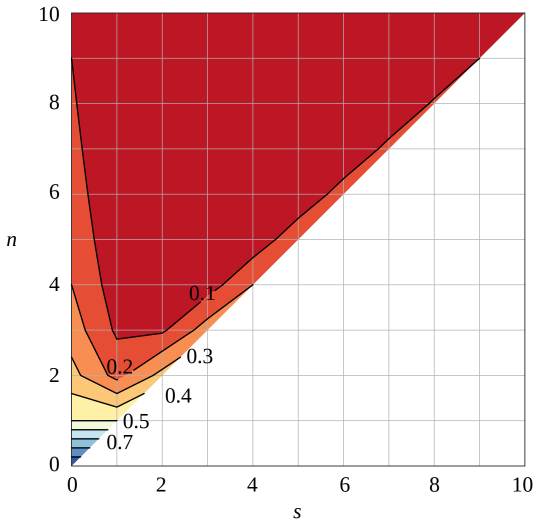

图 6. B(s + 1, n – s + 1) 函数图像平面等高线

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：

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

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，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。

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：

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

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) 给出的确切的概率值。

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 7 (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 (x = 0)，蓝点 ● 代表兔 (x = 1)。

通过图 7 (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)。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 7 (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 7 (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 7 (c)，看看能不能发现有趣的规律。

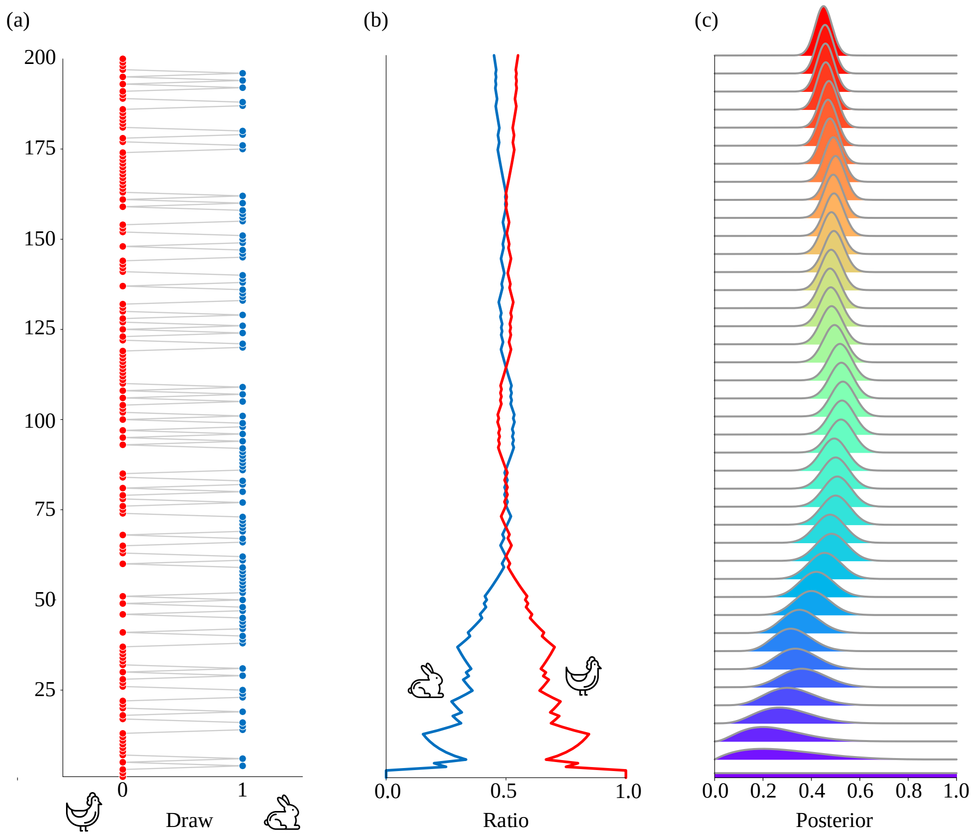

图 7. 某次试验的模拟结果，先验分布为 Beta(1, 1)

图 7 (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，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 8 中看得更清楚。

图 8 (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 8 (b) 所示为，抓到一只动物发现是鸡。利用贝叶斯定理，通过图 8 (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 8 (b) 所示的后验概率分布 Beta(1,2)，这一过程如图 9 所示。

图 8 (b) 这个分布显然认为鸡的比例可能更高，但是不排除其他可能。“不排除其他可能”对应图 8 (b) 的三角形，θ 在 (0, 1) 区间取值时，$f_{\Theta|X}(\theta|x)$ 都不为 0。

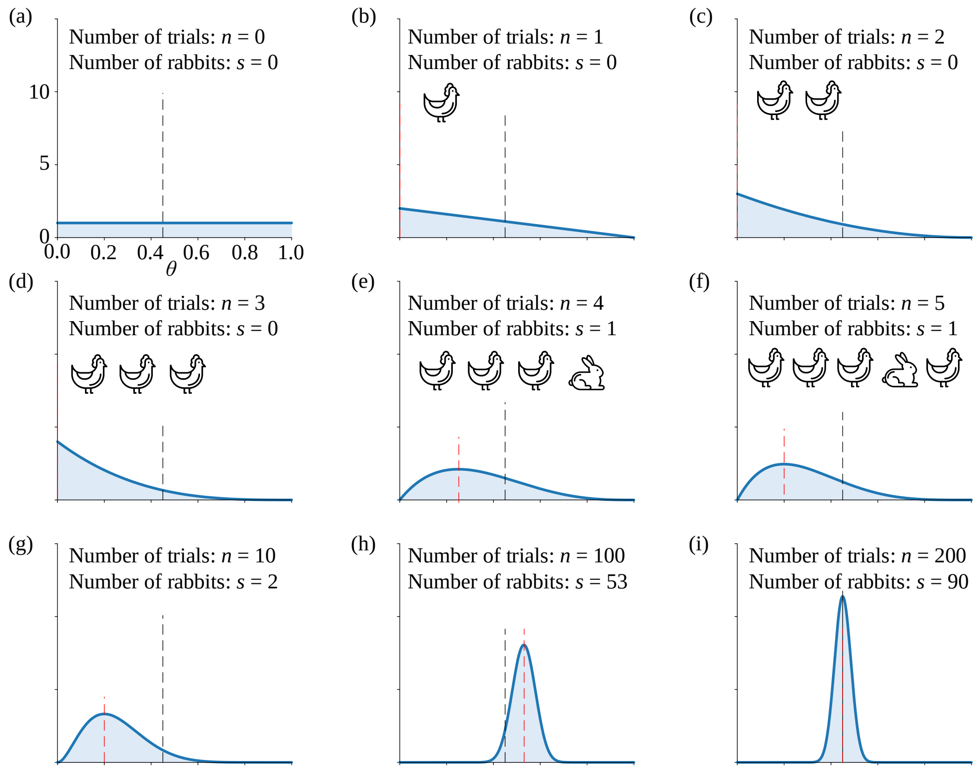

图 8. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(1, 1)

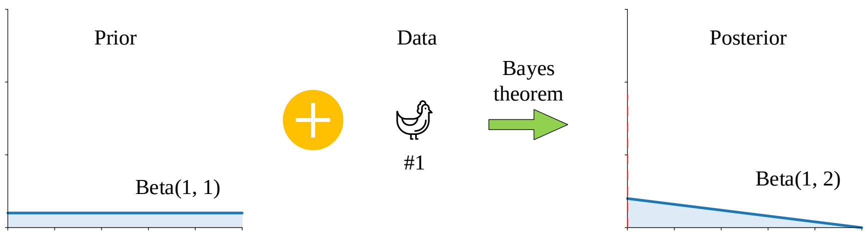

图 9. 不确定鸡兔比例，先验概率 Beta(1, 1) + 一只鸡 (数据) 推导得到后验概率 Beta(1, 2)

抓第二只动物，发现还是鸡。如图 8 (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 8 (c) 的后验概率分布有两条路径。

第一条如图 10 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 11 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 9 中的后验概率。

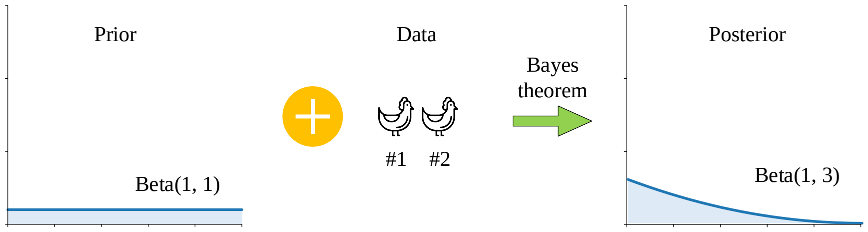

图 10. 第一条路径：先验概率 Beta(1, 1) + 两只鸡 (数据) 推导得到后验概率 Beta(1, 3)

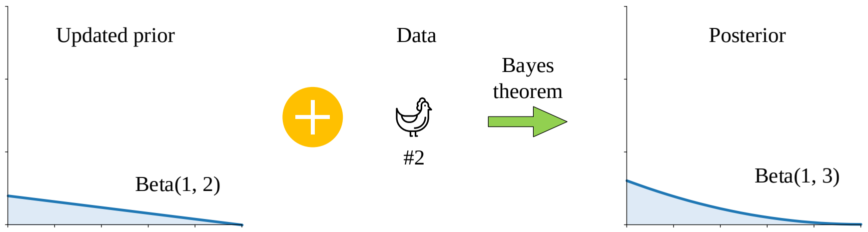

图 11. 第二条路径：更新先验概率 Beta(1, 2) + 第二只鸡 (数据) 推导得到后验概率 Beta(1, 3)

抓第三只动物，竟然还是鸡！如图 8 (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 12 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。

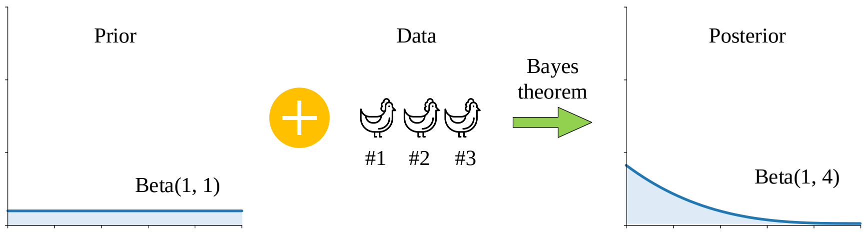

图 12. 先验概率 Beta(1, 1) + 三只鸡 (数据) 推导得到后验概率 Beta(1, 4)

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 8 (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 8 (i) 的后验概率曲线，发现 $\theta = 45\%$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观点。

最大化后验概率 MAP

图 8 子图中红色划线对应的就是混合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：

$$\hat{\theta}_{\text{MAP}} = \frac{s}{n} \tag{20}$$

兜兜转转，结果这个贝叶斯派 MAP 优化解和频率派 MLE 一致？

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 8 (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。

更准确地说，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)，图像为图 8 (i)。

21.4 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：

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

图 13 所示为 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/(8α + 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) 假设条件。

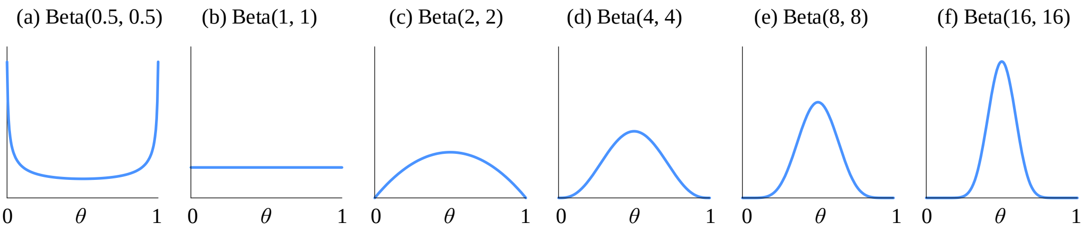

图 13. 五个不同参数 α 取不同值时 Beta(α, α) 分布 PDF 图像

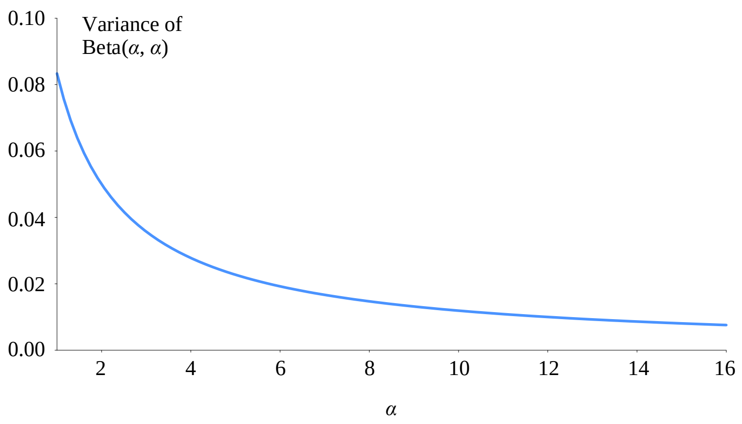

图 14. Beta(α, α) 方差随参数 α 变化

而 α 取不同大于 1 的值时，代表农夫的对鸡兔比例的确信程度。如图 14 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。$\alpha = 2$ 代表农夫认为兔子的比例大致 50%，但是确信度不高。$\alpha = 16$ 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。

联合

因此，联合分布为：

$$
\begin{aligned}
f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right) &= \frac{1}{\mathrm{B}(\alpha,\alpha)}\theta^{\alpha-1}(1-\theta)^{\alpha-1}\theta^{s}(1-\theta)^{n-s} \\
&= \frac{1}{\mathrm{B}(\alpha,\alpha)}\theta^{s+\alpha-1}(1-\theta)^{n-s+\alpha-1}
\end{aligned}
\tag{23}
$$

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：

$$
\begin{aligned}
f_{X_1,X_2,\dots,X_n}\left(x_1,x_2,\dots,x_n\right) &= \int_{\theta} f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right)\mathrm{d}\theta \\
&= \frac{1}{\mathrm{B}(\alpha,\alpha)}\int_{\theta}\theta^{s+\alpha-1}(1-\theta)^{n-s+\alpha-1}\,\mathrm{d}\theta \\
&= \frac{\mathrm{B}(s+\alpha,n-s+\alpha)}{\mathrm{B}(\alpha,\alpha)}
\end{aligned}
\tag{24}
$$

后验

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

$$
\begin{aligned}
f_{\Theta|X_1,X_2,\dots,X_n}\left(\theta \mid x_1,x_2,\dots,x_n\right) &= \frac{f_{X_1,X_2,\dots,X_n,\Theta}\left(x_1,x_2,\dots,x_n,\theta\right)}{f_{X_1,X_2,\dots,X_n}\left(x_1,x_2,\dots,x_n\right)} \\
&= \frac{\dfrac{1}{\mathrm{B}(\alpha,\alpha)}\theta^{s+\alpha-1}(1-\theta)^{n-s+\alpha-1}}{\dfrac{\mathrm{B}(s+\alpha,n-s+\alpha)}{\mathrm{B}(\alpha,\alpha)}} = \frac{\theta^{s+\alpha-1}(1-\theta)^{n-s+\alpha-1}}{\mathrm{B}(s+\alpha,n-s+\alpha)}
\end{aligned}
\tag{25}
$$

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

正比关系

类似 (15)，后验概率存在如下正比关系：

$$f_{\Theta|X_1,X_2,\dots,X_n}\left(\theta \mid x_1,x_2,\dots,x_n\right) \propto f_{X_1,X_2,\dots,X_n|\Theta}\left(x_1,x_2,\dots,x_n \mid \theta\right) f_{\Theta}\left(\theta\right) = \theta^{s+\alpha-1}\left(1-\theta\right)^{n-s+\alpha-1} \tag{26}$$

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 15 (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 15 (b) 所示，随着样本数 n 增大，鸡兔的比例趋于稳定。

图 15 (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。

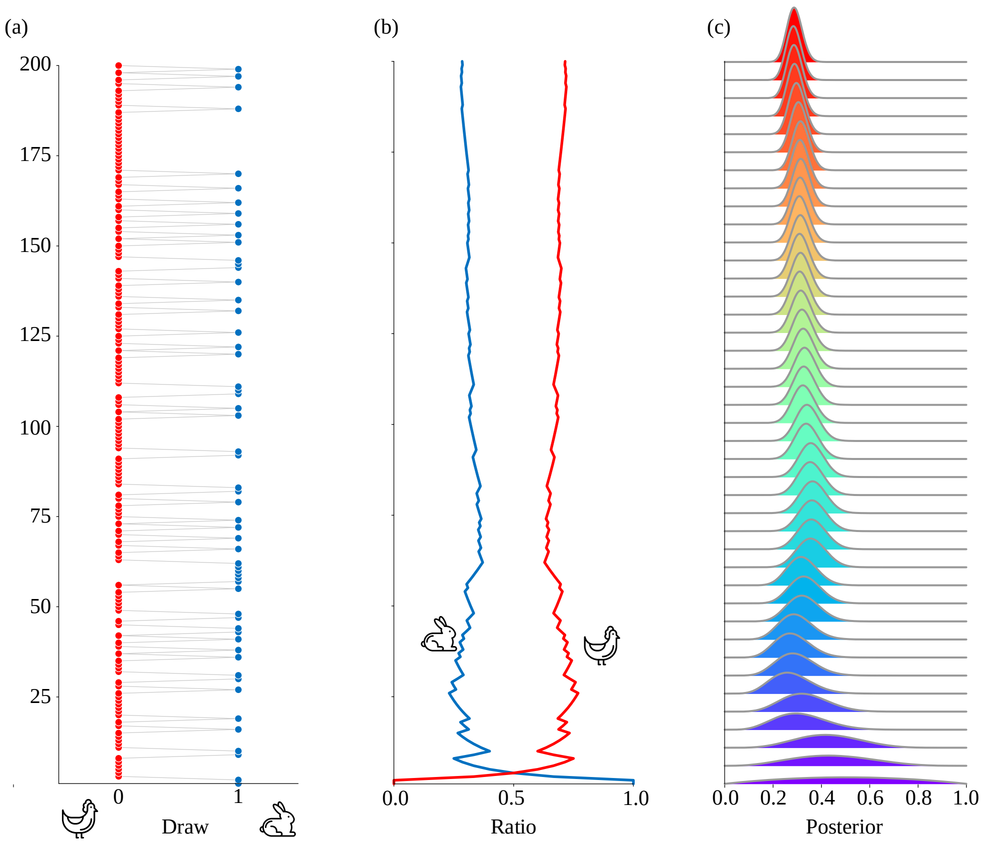

图 15. 某次试验的模拟结果，先验分布为 Beta(2, 2)

图 16 所示为九张不同节点的后验概率分布曲线快照。

图 16 (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 16 (b) 中后验概率最大值向右移动。请大家自己写出这个 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 16 (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 16 剩下几幅子图，注意后验概率形状、最大值位置变化。

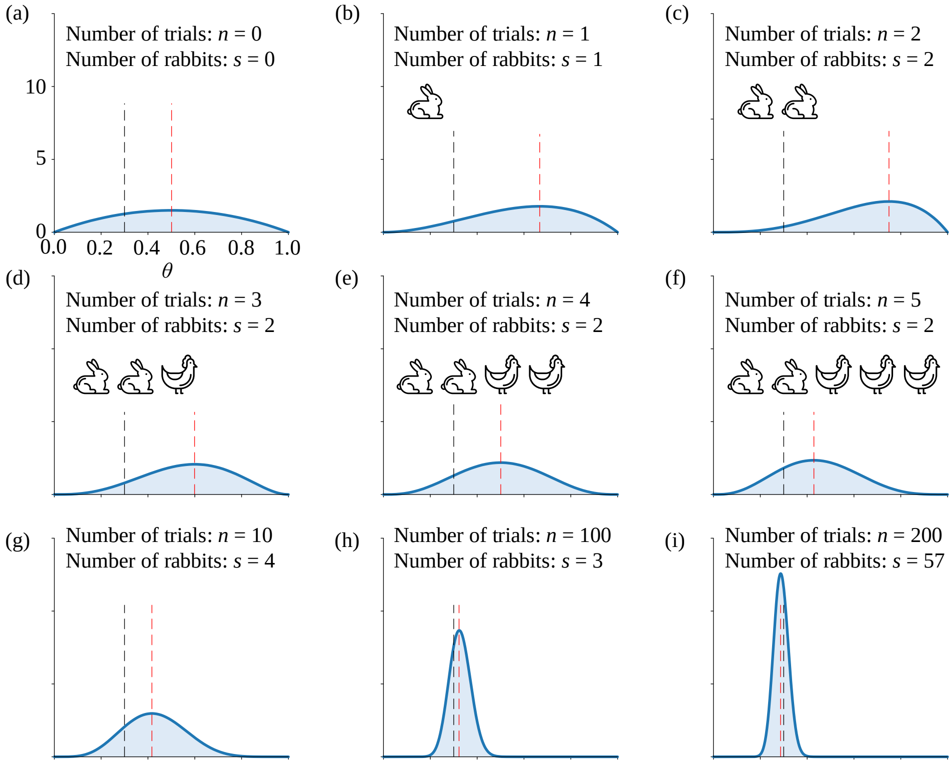

图 16. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(2, 2)

蒙特卡罗模拟：确信度很高

$\alpha = 16$ 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 17 和图 18。强烈建议大家把图 18 每幅子图的 Beta 分布的参数写出来。

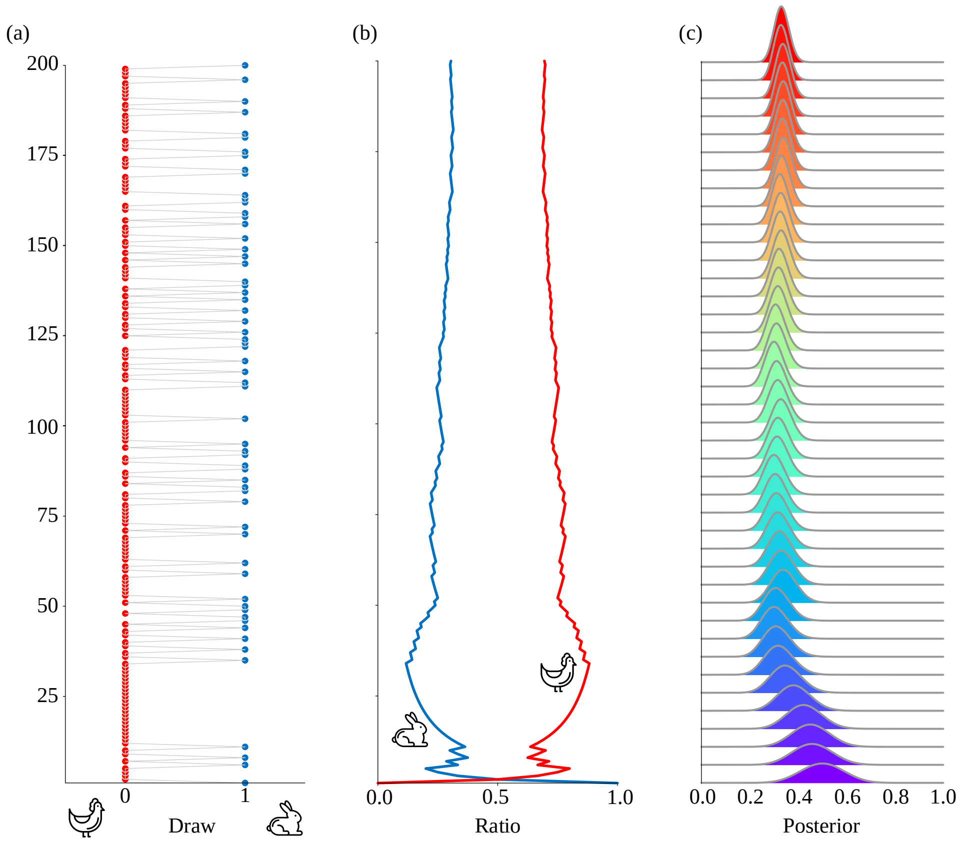

图 17. 某次试验的模拟结果，先验分布为 Beta(16, 16)

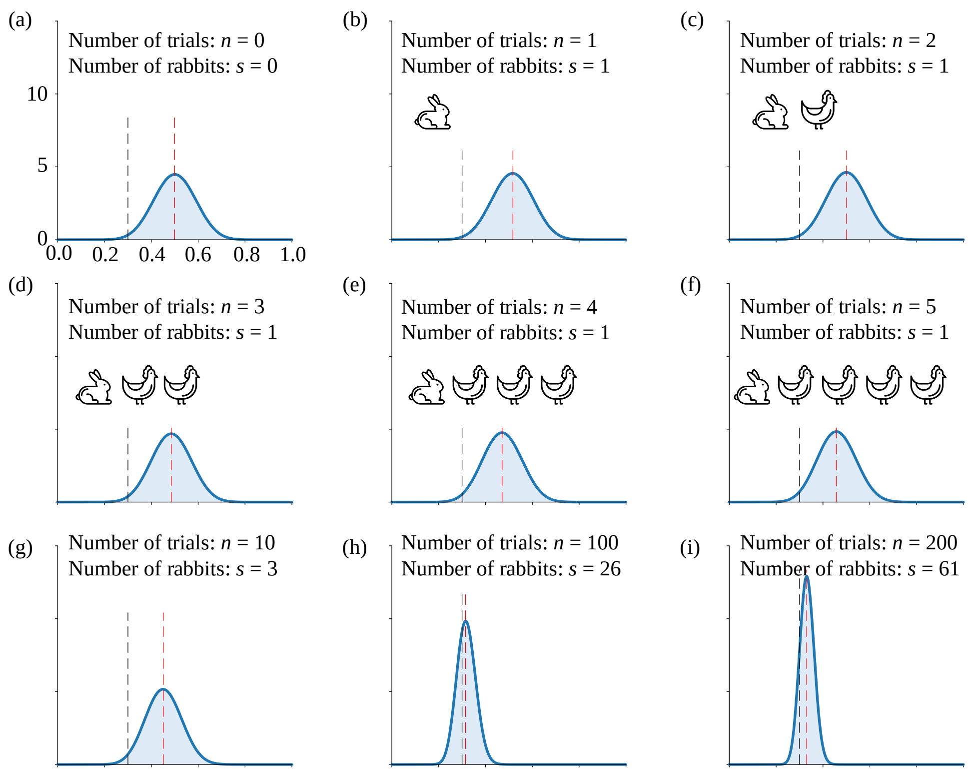

图 18. 九张不同节点的后验概率分布曲线快照，先验分布为 Beta(16, 16)

共轭先验

选择先验是有技巧的！

为了方便运算，在 $f_{\Theta|X}(\theta|x)=\dfrac{f_{X|\Theta}(x|\theta)f_{\Theta}(\theta)}{\int_{\vartheta} f_{X|\Theta}(x|\vartheta)f_{\Theta}(\vartheta)\mathrm{d}\vartheta}$ 中，选取合适的先验分布 $f_{\Theta}(\theta)$ 能让后验分布 $f_{\Theta|X}(\theta|x)$ 和先验分布 $f_{\Theta}(\theta)$ 具有相同的数学形式。

在贝叶斯推断中，如果后验分布与先验分布属于同类，则先验分布与后验分布被称为共轭分布 (conjugate distribution)，而先验分布被称为似然函数的共轭先验 (conjugate prior)。

比如，本章中的 Beta 分布就是二项分布的共轭先验。

代码 Bk5_Ch021_01.py 完成本章蒙特卡洛模拟和可视化。

想深入学习贝叶斯推断的读者可以参考如下两本开源图书：

https://bayesiancomputationbook.com/welcome.html

https://github.com/CamDavidsonPilon/Probabilistic-Programming-and-Bayesian-Methods-for-Hackers