Dive into Bayesian Inference

21 贝叶斯推断进阶

属于同类的后验分布与先验分布叫共轭分布

生活中没有什么是可怕的，它们只是需要被理解。现在是了解更多的时候了，这样我们就可以减少恐惧。

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

—— 玛丽·居里 (Marie Curie) | 波兰裔法国籍物理学家、化学家 | 1867 ~ 1934

- `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 `scipy.stats.beta()` Beta 分布
- `scipy.stats.beta.pdf()` Beta 分布概率密度函数
- `scipy.stats.dirichlet()` Dirichlet 分布
- `scipy.stats.dirichlet.pdf()` Dirichlet 分布概率密度函数

21.1 除了鸡兔，农场来了猪

鸡、兔、猪同笼

在确定农场走地鸡兔时，农夫发现农场还有大量的“走地”猪！

为了搞清楚农场鸡、兔、猪比例，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 $\{0, 1, 2\}$，其中 0 代表鸡，1 代表兔，2 代表猪。和上一章一样，忽略抓取动物对农场整体动物总体比例的影响。

下面我们采用和上一章完全一样，以“先验 → 似然 → 后验”的思路来进行贝叶斯推断。

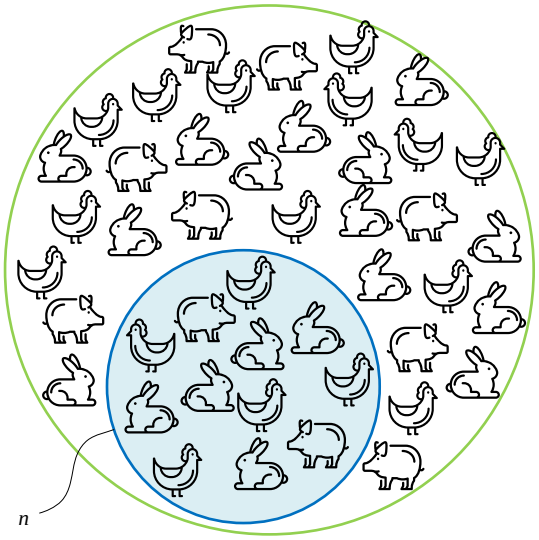


图 1. 农场有数不清的散养鸡兔猪

先验分布

在出现样本数据之前，先验分布代表我们对模型参数的既有“知识”，主观“经验”。

θ_1、θ_2、θ_3 分别为农场中鸡、兔、猪的比例，θ_1、θ_2、θ_3 的取值范围都是 $[0, 1]$。鸡兔猪比例之和为 1，即 θ_1、θ_2、θ_3 满足如下等式：

$$\theta_1 + \theta_2 + \theta_3 = 1 \tag{1}$$

我们把 θ_1、θ_2、θ_3 写成一个向量 $\boldsymbol{\theta}$。

上一章中，我们采用 Beta 分布作为先验分布。这一章，鸡兔猪问题中 $\boldsymbol{\theta} \sim \text{Dir}(\alpha_1, \alpha_2, \alpha_3)$：

$$f_{\Theta}(\boldsymbol{\theta}) = \frac{1}{\mathrm{B}(\alpha_1, \alpha_2, \alpha_3)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \theta_3^{\alpha_3 - 1} \tag{2}$$

B($\boldsymbol{\alpha}$) 起到“归一化”作用，具体定义为：

$$\mathrm{B}(\alpha_1, \alpha_2, \alpha_3) = \frac{\prod_{i=1}^{3} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{3} \alpha_i\right)} = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)} \tag{3}$$

本书前文提过，Dirichlet 分布也叫狄利克雷分布，它本质上是**多元 Beta 分布** (multivariate Beta distribution)。或者说，Beta 分布是特殊的 Dirichlet 分布。

我们也可以把 Dir(α_1, α_2, α_3) 写成 Dir($\boldsymbol{\alpha}$)。

先验分布位置

通过上一章学习我们知道，对于一个先验分布，常用众数、期望 (均值) 描述它的位置。

对于 Dir($\boldsymbol{\alpha}$)，X_i 的众数为：

$$\frac{\alpha_i - 1}{\sum_{k=1}^{K} \alpha_k - K} = \frac{\alpha_i - 1}{\alpha_0 - K}, \quad \alpha_i > 1 \tag{4}$$

这是先验初始比例所在位置，也是 MAP 的位置。其中，$\alpha_0 = \sum_{k=1}^{K} \alpha_k$ 。

特别地如果 $\alpha_1 = \alpha_2 = \ldots = \alpha_K$，$X_i$ 的众数为：

$$\frac{\alpha_i - 1}{\alpha_0 - K} = \frac{1}{K}, \quad \alpha_i > 1 \tag{5}$$

对于 Dir($\boldsymbol{\alpha}$)，X_i 的期望为：

$$\frac{\alpha_i}{\sum_{k=1}^{K} \alpha_k} = \frac{\alpha_i}{\alpha_0} \tag{6}$$

此外，大家可能会想到**中位数** (median)，也就是百分位 50-50 的位置。本章马上比较众数、期望、中位数。

似然分布

在贝叶斯推断中，我们用似然分布整合样本数据，并描述样本分布。注意，似然函数中，样本数据为给定值，而模型参数是变量。也就是说，似然分布本质上是模型参数的函数。

上一章，我们后来用二项分布作为似然分布。本章用多项分布作似然分布。二项分布可以视作是多项分布的特例。

n 为抓取动物的总数，随机变量 X_1、X_2、X_3 代表其中鸡、兔、猪数量，x_1、x_2、x_3 代表 X_1、X_2、X_3 的取值。因此，如下等式成立：

$$x_1 + x_2 + x_3 = n \tag{7}$$

在 $\Theta = \boldsymbol{\theta}$ 的条件下，(X_1, X_2, X_3) 满足如下多项分布：

$$f_{\boldsymbol{\chi}|\Theta}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right) = f_{X_1,X_2,X_3|\Theta}\left(x_1,x_2,x_K\,|\,\boldsymbol{\theta}\right) = \frac{n!}{\left(x_1!\right)\times\left(x_2!\right)\times\left(x_3!\right)} \times \theta_1^{x_1} \times \theta_2^{x_2} \times \theta_3^{x_3} \tag{8}$$

$\boldsymbol{\chi}$ 代表 X_1、X_2、X_3 构成的向量。

最大似然 MLE

似然函数 $f_{\boldsymbol{\chi}|\Theta}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right)$ 取对数，并忽略系数：

$$x_1 \ln\theta_1 + x_2 \ln\theta_2 + x_3 \ln\theta_3 \tag{9}$$

θ_1、θ_2、θ_3 存在 $\theta_1 + \theta_2 + \theta_3 = 1$ 等式约束。用拉格朗日乘子法，我们可以很容易把含约束优化问题转化为无约束问题，求得 MLE 的解为：

$$\hat{\theta}_1 = \frac{x_1}{n}, \quad \hat{\theta}_2 = \frac{x_2}{n}, \quad \hat{\theta}_3 = \frac{x_3}{n} \tag{10}$$

忘记拉格朗日乘子法读者，可以回顾《矩阵力量》第 18 章。

后验分布

后验分布代表“先验 + 数据”融合后对参数的信念。

由于后验 ∝ 先验 × 似然，后验概率 $f_{\Theta|\boldsymbol{\chi}}\left(\boldsymbol{\theta}\,|\,\boldsymbol{x}\right)$：

$$f_{\Theta|\boldsymbol{\chi}}\left(\boldsymbol{\theta}\,|\,\boldsymbol{x}\right) \propto f_{\boldsymbol{\chi}|\Theta}\left(\boldsymbol{x}\,|\,\boldsymbol{\theta}\right) f_{\Theta}\left(\boldsymbol{\theta}\right) \tag{11}$$

所以：

$$\begin{aligned} f_{\Theta|\boldsymbol{\chi}}\left(\boldsymbol{\theta}\,|\,\boldsymbol{x}\right) &\propto \theta_1^{x_1} \times \theta_2^{x_2} \times \theta_3^{x_3} \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \theta_3^{\alpha_3-1} \\ &= \theta_1^{x_1+\alpha_1-1} \times \theta_2^{x_2+\alpha_2-1} \times \theta_3^{x_3+\alpha_3-1} \end{aligned} \tag{12}$$

想要把 (12) 变成概率密度函数，我们需要一个归一化系数，使得 PDF 在整个定义域上积分为 1。

很明显，我们需要的就是如下 Beta 函数：

$$\mathrm{B}\left(\alpha_1 + x_1, \alpha_2 + x_2, \alpha_3 + x_3\right) = \mathrm{B}\left(\boldsymbol{x} + \boldsymbol{\alpha}\right) = \frac{\prod_{i=1}^{K} \Gamma\left(\alpha_i + x_i\right)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i + x_i\right)} \tag{13}$$

由此可知后验分布 $f_{\Theta|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$ 服从 Dir($x_1 + \alpha_1, x_2 + \alpha_2, x_3 + \alpha_3$)，可以写成 Dir($\boldsymbol{x} + \boldsymbol{\alpha}$)。

也就是说，在这个鸡兔猪贝叶斯推断问题中，如果先验概率为 Dir($\boldsymbol{\alpha}$)，则后验概率为 Dir($\boldsymbol{x} + \boldsymbol{\alpha}$)。

最大后验 MAP

对于 Dir($\boldsymbol{x} + \boldsymbol{\alpha}$)，$X_i$ 的众数为：

$$\frac{x_i+\alpha_i-1}{\sum_{k=1}^{K}(x_k+\alpha_k)-K}=\frac{x_i+\alpha_i-1}{n+\alpha_0-K},\quad x_i+\alpha_i>1 \tag{14}$$

这就是最大后验估计 MAP 的解析解位置所在。

当 $K = 3$ 时，最大后验 MAP 的位置为：

$$\frac{x_i+\alpha_i-1}{n+\alpha_0-3} \tag{15}$$

特别地，当 $\alpha_1 = \alpha_2 = \alpha_3 = 1$ 时，最大后验 MAP 的位置为：

$$\frac{x_i}{n} \tag{16}$$

此时，MAP 的解和 MLE 的解相同。

边缘分布

根据本书第 7 章，先验分布 Dir($\boldsymbol{\alpha}$) 的三个边缘分布分别为：

$$\begin{aligned}&\text{Beta}(\alpha_1,\alpha_0-\alpha_1)\\&\text{Beta}(\alpha_2,\alpha_0-\alpha_2)\\&\text{Beta}(\alpha_3,\alpha_0-\alpha_3)\end{aligned} \tag{17}$$

后验分布 Dir($\boldsymbol{x} + \boldsymbol{\alpha}$) 的三个边缘分布分别为：

$$\begin{aligned}&\text{Beta}(x_1+\alpha_1,\alpha_0+n-(x_1+\alpha_1))\\&\text{Beta}(x_2+\alpha_2,\alpha_0+n-(x_2+\alpha_2))\\&\text{Beta}(x_3+\alpha_3,\alpha_0+n-(x_3+\alpha_3))\end{aligned} \tag{18}$$

后验分布的位置

Dir($\boldsymbol{x} + \boldsymbol{\alpha}$) 的三个边缘分布各自的众数分别为：

$$\frac{x_i+\alpha_i-1}{n+\alpha_0-2} \tag{19}$$

它们的期望值位置为：

$$\frac{x_i+\alpha_i}{n+\alpha_0} \tag{20}$$

可见当 n 足够大时，(20) 可以用来近似 (19)。而 (19) 则可以用来近似 (14)，后验分布 MAP 优化解。也就是说，我们可以用三个边缘 Beta 分布的期望 (均值) 来近似后验分布 Dir($\boldsymbol{x}+\boldsymbol{\alpha}$) 的 MAP 优化解。特别是在下一章中，大家会看到我们直接用后验边缘 Beta 分布的均值作为 MAP 的优化解。

表 1 比较先验、后验分布的众数和期望。

表 1. 比较先验、后验分布的众数和期望

分布	类型	统计量	位置
Dir($\boldsymbol{\alpha}$)	先验	众数 (联合 PDF 曲面最大值)	$\frac{\alpha_i-1}{\alpha_0-K}$，$\alpha_i>1$
		期望 (联合 PDF 质心)	$\frac{\alpha_i}{\alpha_0}$
Beta(α_i, $\alpha_0-\alpha_i$)	先验边缘	众数 (先验边缘分布 PDF 曲线最大值)	$\frac{\alpha_i-1}{\alpha_0-2}$，$\alpha_i>1$
		期望 (先验边缘分布均值)	$\frac{\alpha_i}{\alpha_0}$
Dir($\boldsymbol{x}+\boldsymbol{\alpha}$)	后验	众数 (联合 PDF 曲面最大值) * MAP 优化解	$\frac{x_i+\alpha_i-1}{n+\alpha_0-K}$，$x_i+\alpha_i>1$
		期望 (联合 PDF 质心) * 最大化期望值	$\frac{x_i+\alpha_i}{n+\alpha_0}$
Beta(α_i+x_i, $\alpha_0-(\alpha_i+x_i)$)	后验边缘	众数 (边缘 PDF 曲线最大值)	$\frac{x_i+\alpha_i-1}{n+\alpha_0-2}$，$x_i+\alpha_i>1$
		期望 (边缘 PDF 均值) * 常用来近似 MAP 优化解	$\frac{x_i+\alpha_i}{n+\alpha_0}$

比较 Beta 分布的众数、中位数、均值

本节最后比较 Beta(α, β) 众数、中位数、均值。众数、中位数、均值都可以用来表征 Beta(α, β) 分布的具体位置。实际上，在贝叶斯推断中，对模型参数有三种不同的**点估计** (point estimate)：1) 后验众数，2) 后验中位数，3) 后验均值。

图 2 所示为不同 Beta(α, β) 分布众数 (蓝色划线)、中位数 (黑色划线)、均值 (红色划线)。这幅图中，请大家回顾本书第 2 章介绍的有关左偏、右偏的内容。

Beta(α, β) 分布的众数有明显的缺点。我们在本书第 7 章介绍过，当 α 或 β 小于等于 1 时，Beta(α, β) 的众数可能位于分布的某一端，0 或 1。比如图 2 中，Beta(2, 1) 的众数位于 1，而 Beta(1, 2) 的众数位于 0。这两个众数显然不能合理地表征分布的具体位置。

此外，下一章中大家会看到通过数值方法得到后验分布的曲线可能有若干局部极大值，这给 MAP 求解增加了麻烦。

因此，实践中当样本足够大时，我们常用后验边缘分布均值代替后验众数作为 MAP 的结果。

此外，后验中位数也是一个不错的选择。对于厚尾的后验分布，后验中位数要好过后验均值。因为后验均值的位置会受到厚尾影响。但是，后验中位数计算更困难。

特别地，如果后验分布对称，众数、均值、中位数重合。

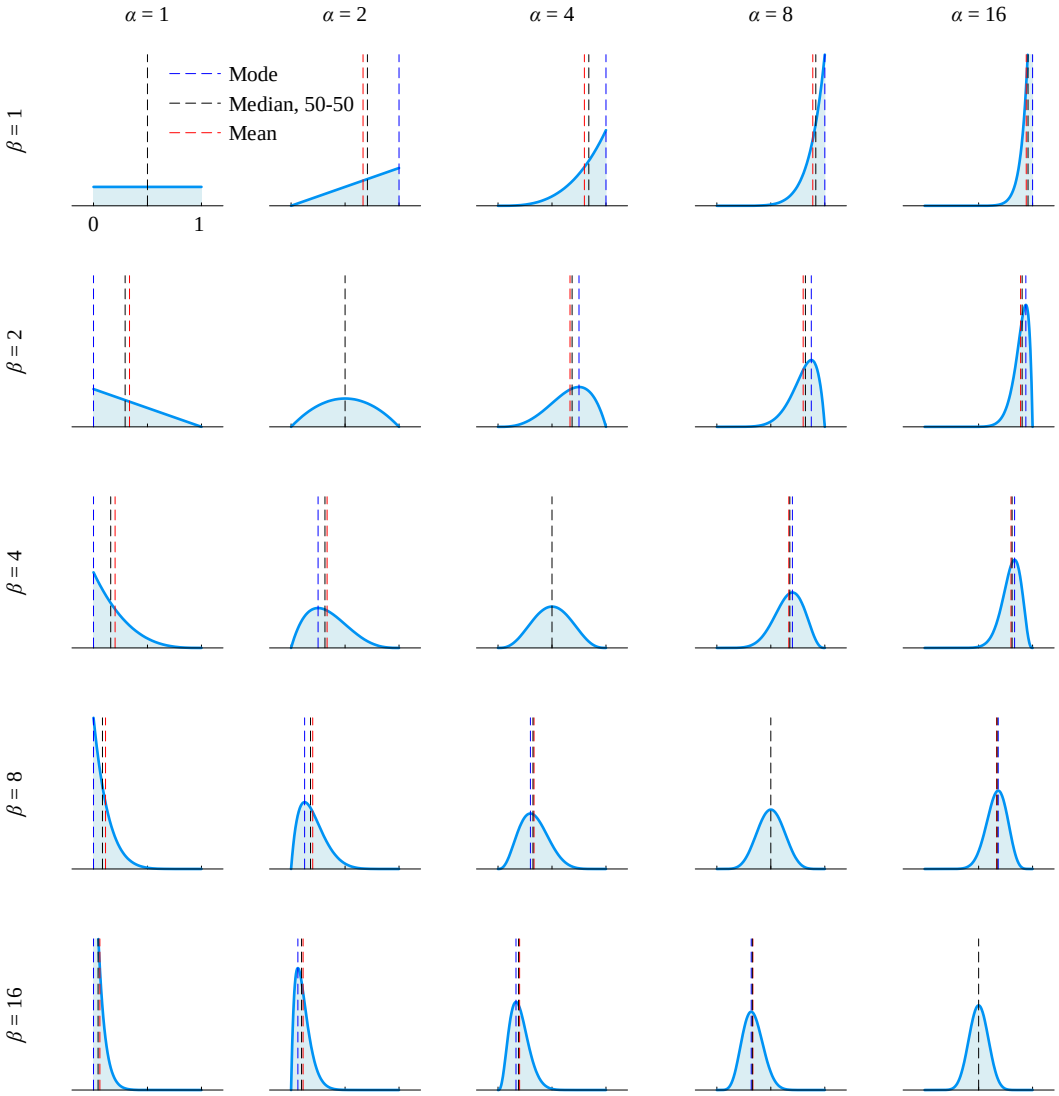


图 2. 比较不同 Beta(α, β) 分布众数、中位数、均值

有了本节理论铺垫，下面我们结合具体实例展开讲解。本章后续三节和上一章最后三节结构相似，请大家对比阅读。

21.2 走地鸡兔猪：比例完全不确定

上一章提过，如果我们事先对动物比例值一无所知的话，我们就可以采用一个“不偏不倚”的先验分布。Dir(1, 1, 1) 显然就满足本节这个要求。这种 Dirichlet 分布又叫 flat (uniform) Dirichlet distribution。

Dir(1, 1, 1) 分布概率密度值为定值，它代表我们试图保持“客观”，而不是将“主观”先验经验代入贝叶斯推断中去。图 3 所示为四种三元 Dirichlet 分布的可视化方案，本章将采用第一种，$\theta_1\theta_2$ 平面直角坐标系投影。

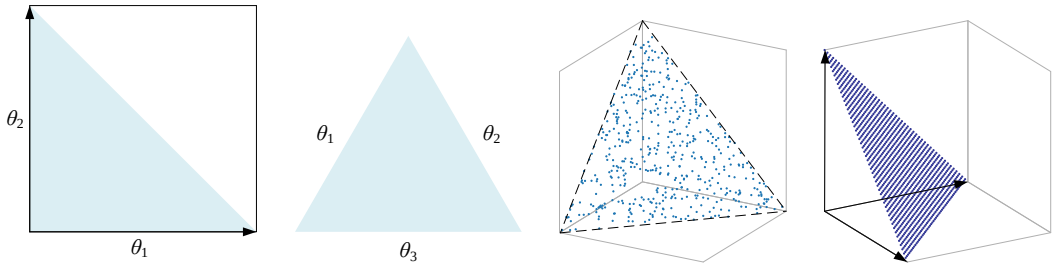


图 3. Dirichlet 分布的几种可视化方案，$\alpha_1 = 1, \alpha_2 = 1, \alpha_3 = 1$

图 4 所示为某次采样的结果。注意，采样结果和先验分布无关。图 4 (a) 中，0 代表鸡，1 代表兔，2 代表猪。图 4 (b) 中，随着样本数量不断增加，三种动物的比例逐渐稳定。仅仅依赖样本数据进行推断，特别是样本数量足够大时，我们已经可以得知所谓“客观”概率结果。

利用贝叶斯定理，整合“先验分布 + 样本”，我么可以得到后验分布。图 5 (a) 所示为 Dir(1, 1, 1) 对应的图像。图 5 剩余 8 个不同子图展示随着样本数据 (x_1, x_2, x_3) 不断增加后验分布 $\text{Dir}(\boldsymbol{x} + \boldsymbol{\alpha})$ 的变化。

图 6 所示为，n 不断增加，三个后验边缘分布位置逐渐稳定。而后验边缘分布本身变得越发“细高”，标准差不断减小，这意味着鸡兔猪的比例变得更值得信任。

图 7 比较三个不同后验边缘分布曲线形状。请大家写出每幅子图中不同后验边缘分布对应的 Beta 分布。

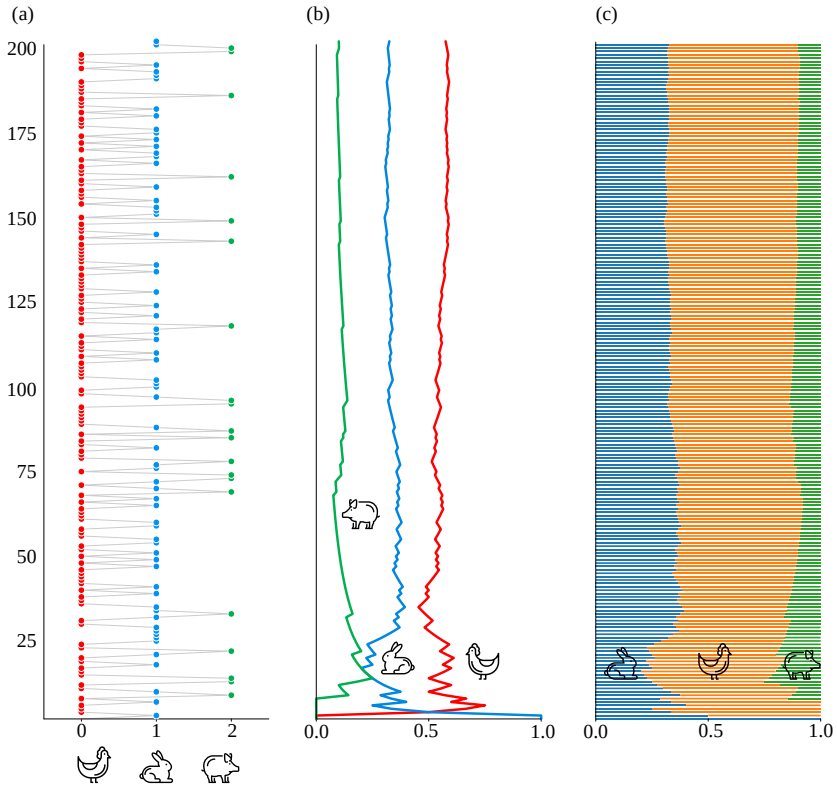


图 4. 某次试验的蒙特卡罗模拟结果

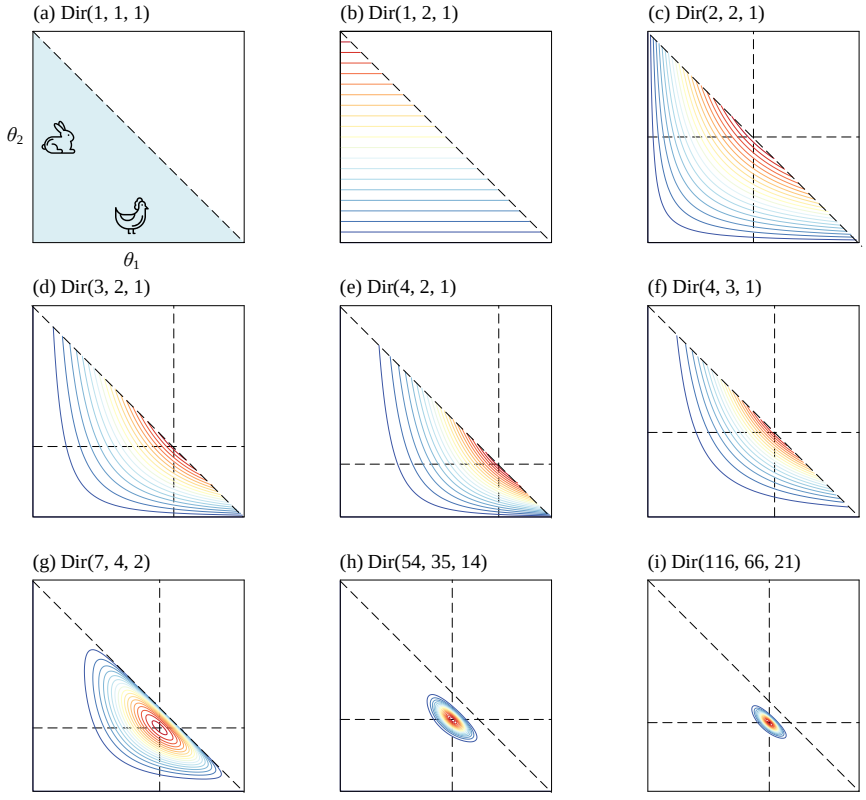


图 5. 九张 Dirichlet 分布，$\theta_1\theta_2$平面直角坐标系，先验分布为 Dir(1, 1, 1)

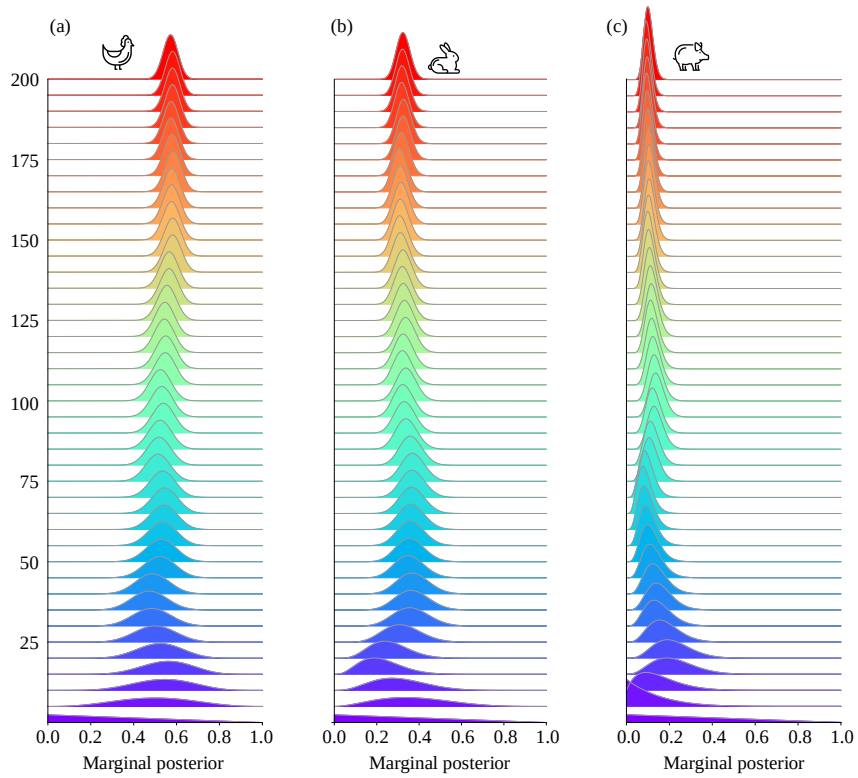


图 6. 某次试验的后验边缘分布山脊图，先验分布为 Dir(1, 1, 1)

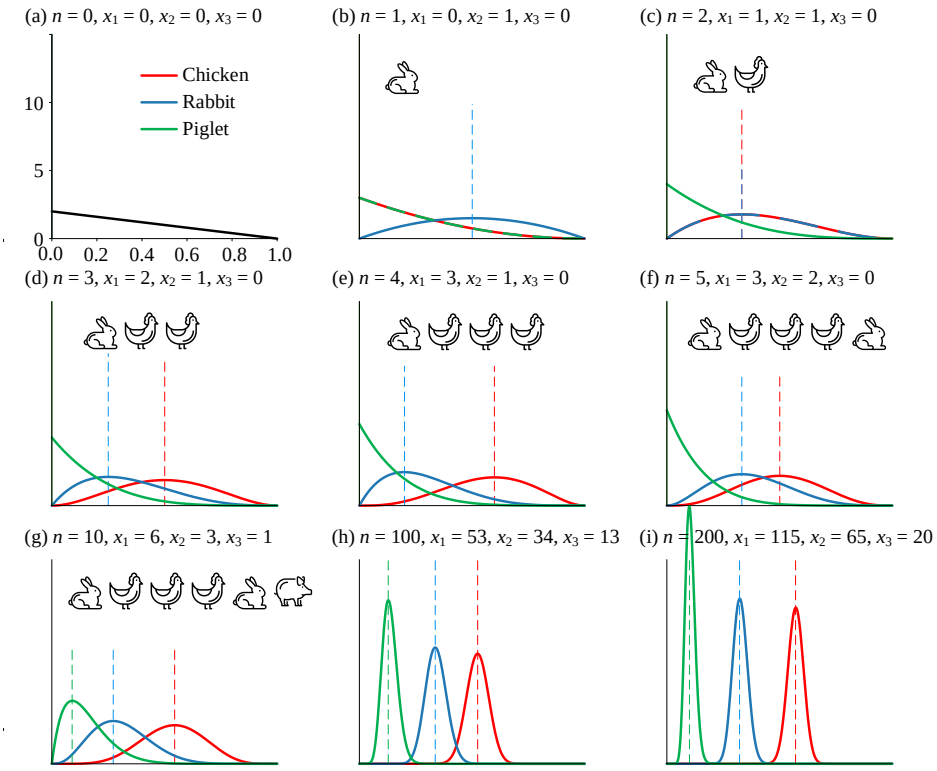


图 7. 九张不同节点的后验边缘 PDF 曲线快照，先验分布为 Dir(1, 1, 1)

21.3 走地鸡兔猪：很可能各 1/3

如果农夫认为农场的鸡兔猪的比例都是 1/3，我们就需要选用不同于前文的先验分布。这种情况，先验 Dirichlet 分布三个参数相同。如图 8 所示为 $\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 2$ 时，Dirichlet 分布的四种可视化方案。请大家分别计算 Dir(2, 2, 2) 的众数、均值，并计算其边缘分布的众数、均值。

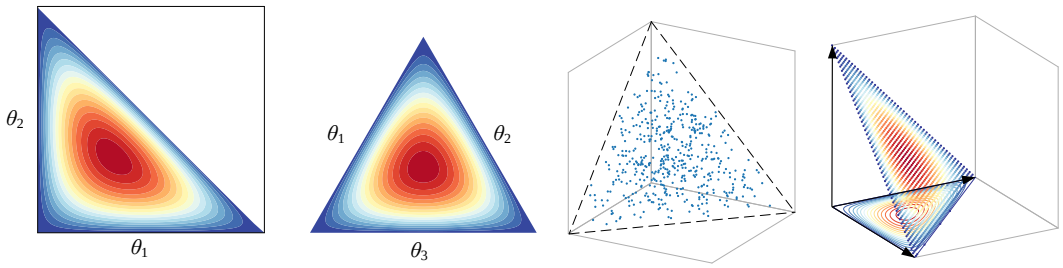


图 8. Dirichlet 分布的几种可视化方案，$\alpha_1 = 2, \alpha_2 = 2, \alpha_3 = 2$

图 9 所示为 4 种不同确信度的先验分布参数设定条件下，Dirichlet 分布等高线和边缘分布曲线。图中黑色划线代表 Dirichlet 分布众数 (MAP 优化解) 所在位置。蓝色划线为边缘 Beta 分布众数位置。

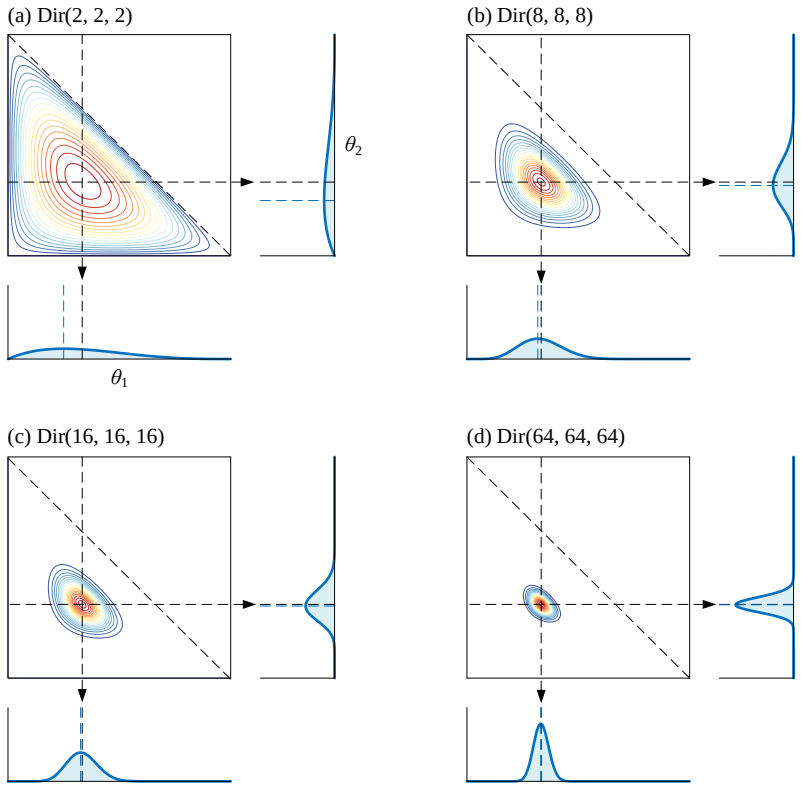


图 9. 四个不同置信度

下面，我们分两种情况完成本节蒙特卡罗模拟。随机数发生器的结果和上一节图 4 完全一致。

确信度不高

确信度不高的情况下，选择 Dir(2, 2, 2) 为先验分布，如图 10 (a) 所示。

随着样本数据不断整合，图 10 剩余八幅子图所示为后验分布变化。比较图 5 (i)、图 10 (i)，可以发现样本数量较大时，后验分布受先验分布影响较小。

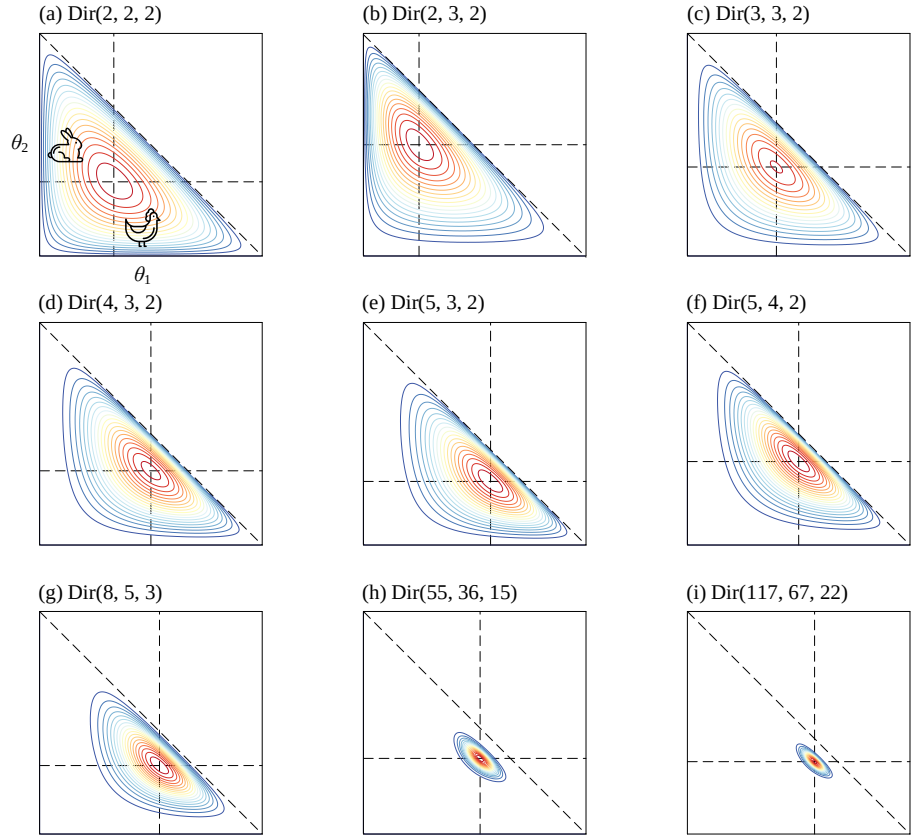


图 10. 九张 Dirichlet 分布，$\theta_1\theta_2$平面直角坐标系，先验分布为 Dir(2, 2, 2)

图 10 (g) 代表“先验 Dir(2, 2, 2) + 样本 (x_1 = 6, x_1 = 3, x_1 = 1) → 后验 Dir(8, 5, 3)”。具体过程如图 11 所示。

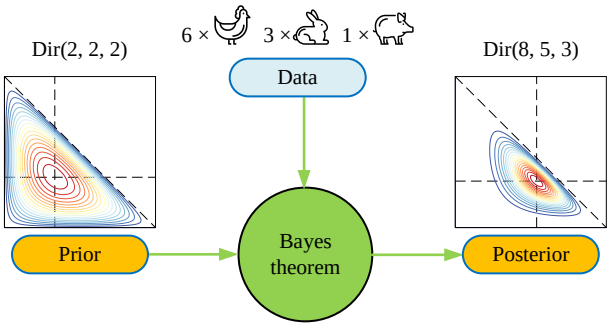


图 11. 先验 Dir(2, 2, 2) + 样本 → 后验 Dir(8, 5, 3)

图 12 所示为后验边缘分布的山脊图。比较图 6、图 12，容易发现当 n 比较小时，后验边缘分布曲线差异较大；n 增大后，后验边缘分布趋同。

图 13 比较三个不同的后验边缘分布。

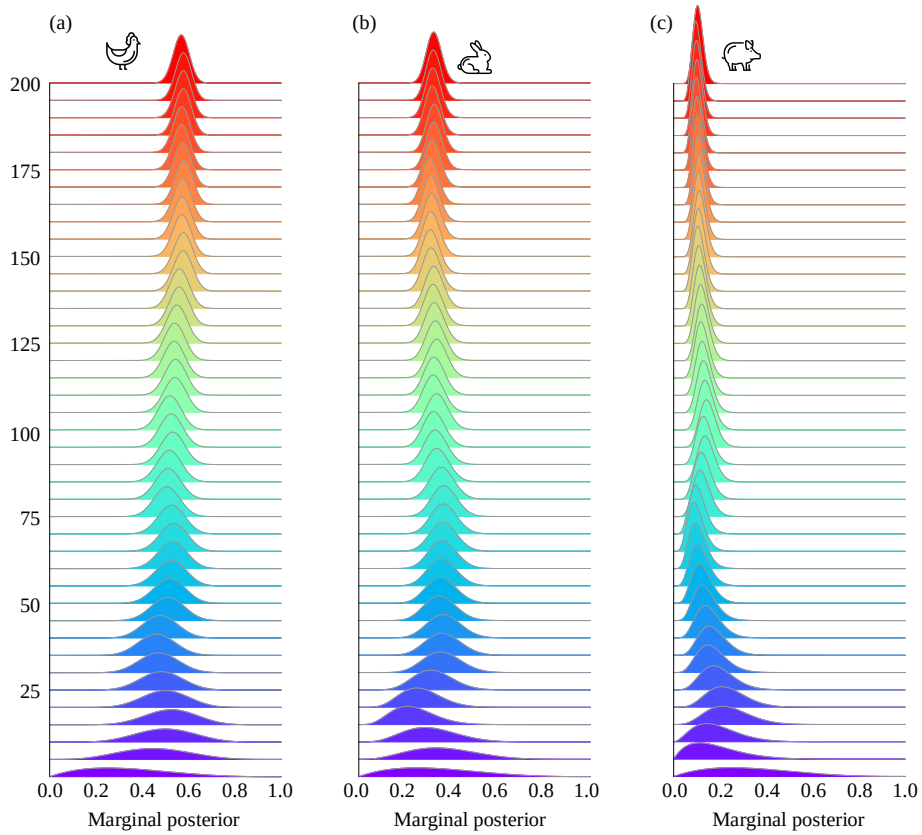


图 12. 某次试验的后验边缘分布山脊图，先验分布为 Dir(2, 2, 2)

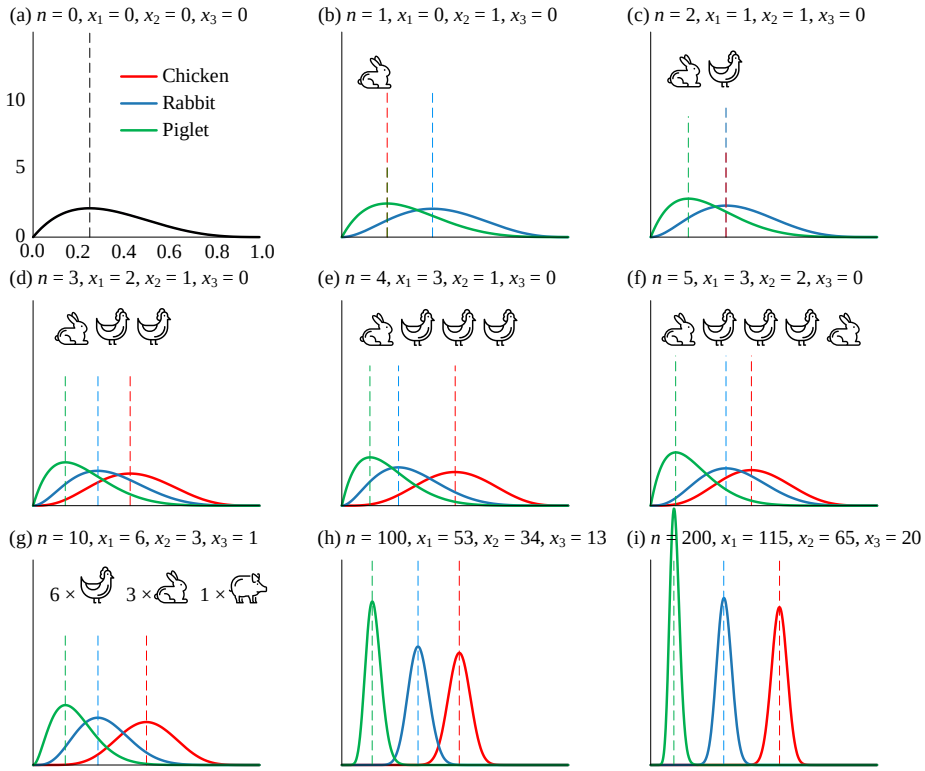


图 13. 九张不同节点的后验边缘 PDF 曲线快照，先验分布为 Dir(2, 2, 2)

确信度很高

当农夫对 1/3 的比例确信度比较高时，我们可以选择 Dir(8, 8, 8) 作为先验分布。比较图 10 (a)、图 14 (a)，我们可以发现先验分布变得更加细高，这意味着边缘分布的均方差减小，确信度提高。

请大家自行分析图 14 剩余子图，并对比图 10。

图 15 先验分布为 Dir(8, 8, 8) 条件下，后验边缘分布的山脊图。图 16 比较不同后验边缘分布。请大家自行分析这两图图像。

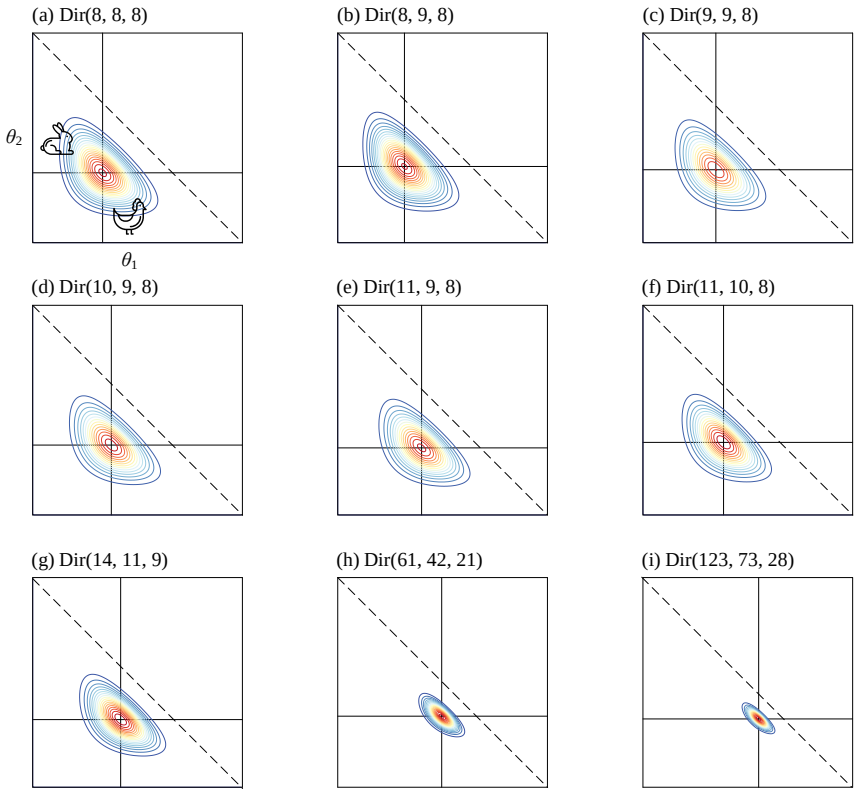


图 14. 九张 Dirichlet 分布，$\theta_1\theta_2$平面直角坐标系，先验分布为 Dir(8, 8, 8)

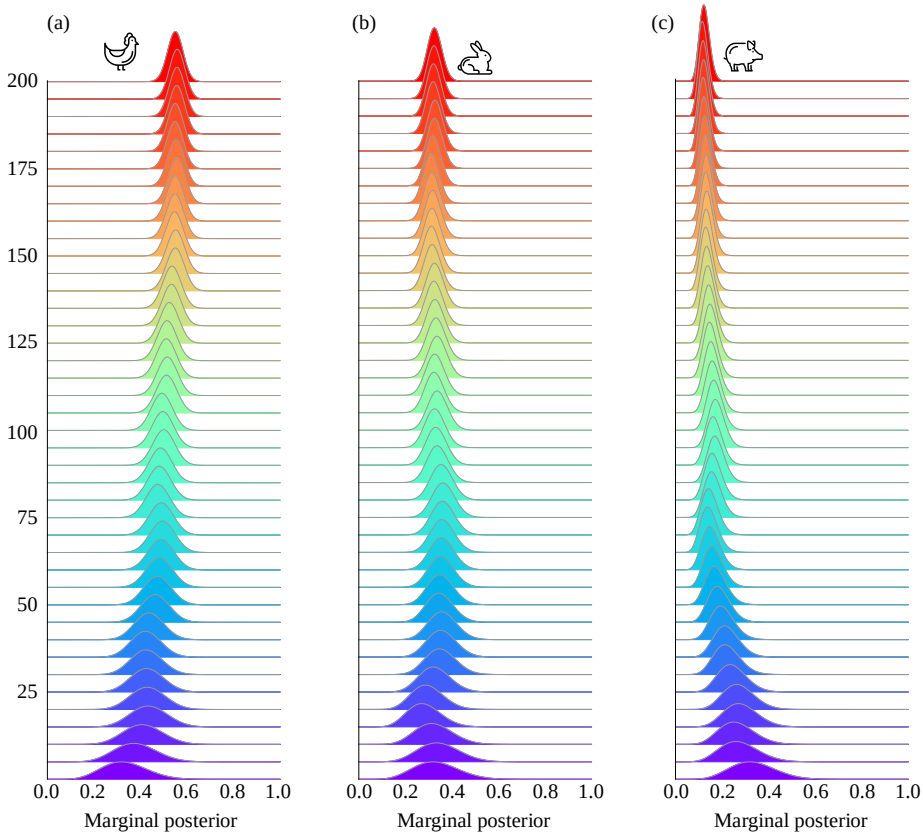


图 15. 某次试验的后验边缘分布山脊图，先验分布为 Dir(8, 8, 8)

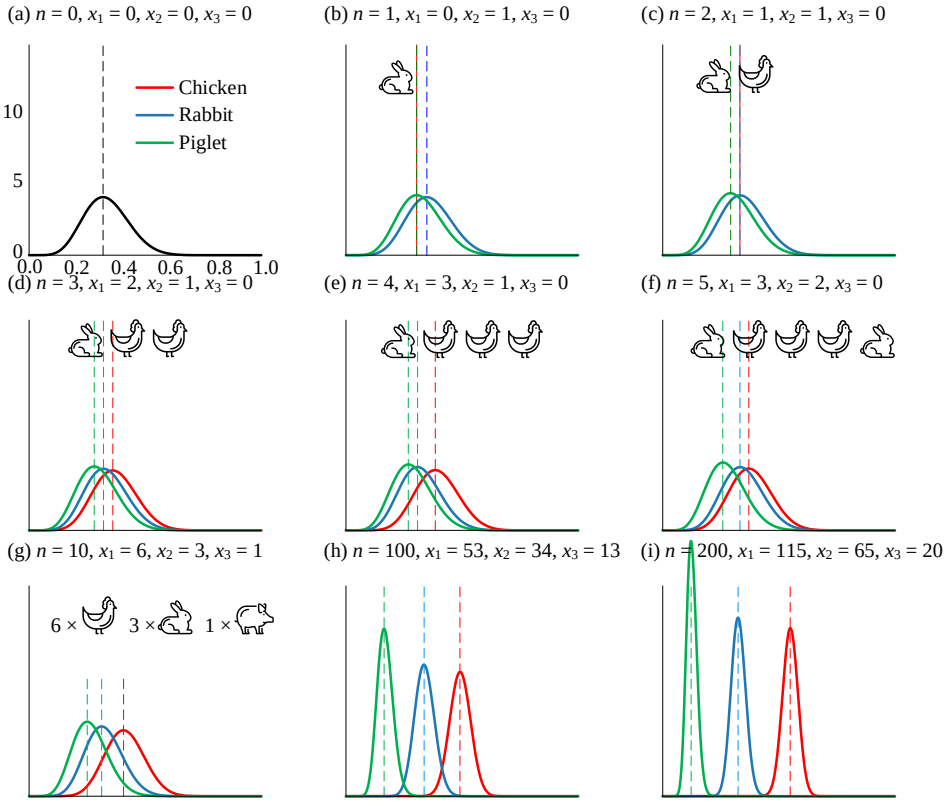


图 16. 九张不同节点的后验边缘 PDF 曲线快照，先验分布为 Dir(8, 8, 8)

代码 Bk5_Ch21_01.py 完成本章前文蒙特卡洛模拟和可视化。

21.4 走地鸡兔猪：更一般的情况

不同先验

上一章提过，如果样本数据足够大，先验对后验的影响微乎其微。如图 17 所示，从完全不同的先验出发得到的后验结果非常相似。

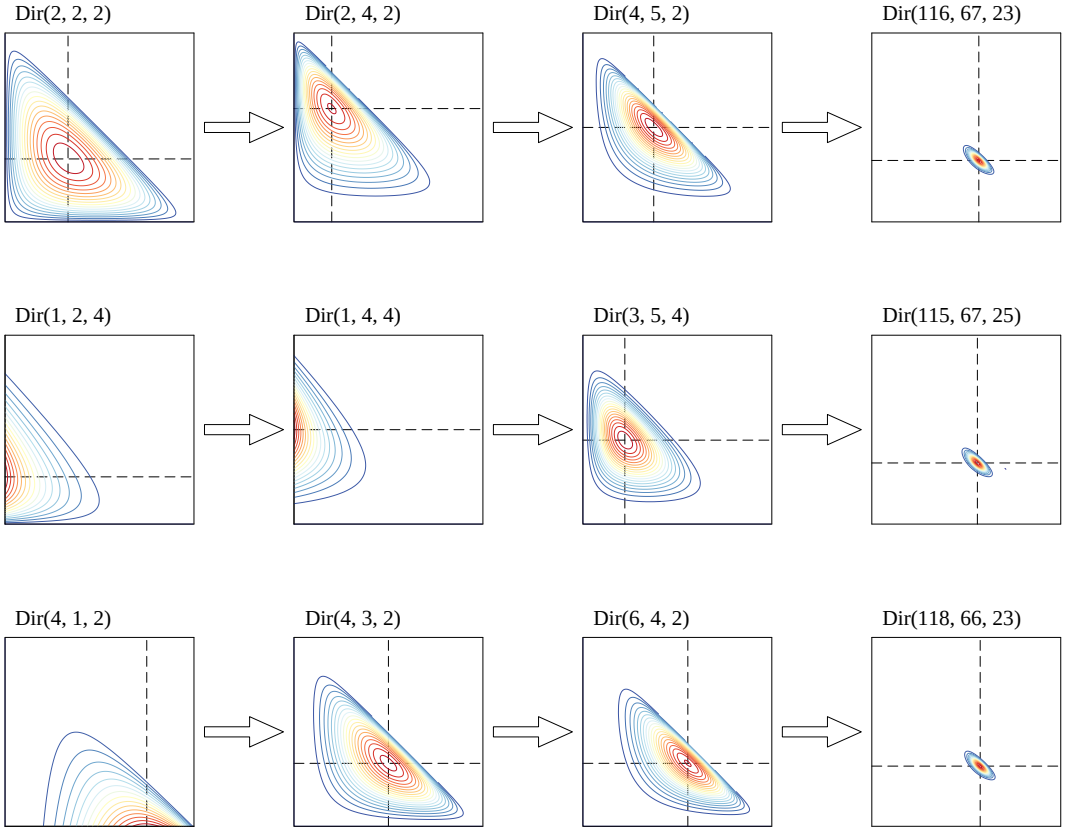


图 17. 如果样本数据足够大，先验对后验的影响微乎其微

贝叶斯收缩

上一章介绍了贝叶斯收缩，本章贝叶斯推断的结果也可以用这个视角来理解。

Dir($\boldsymbol{x}$ + $\boldsymbol{\alpha}$) 的后验边缘分布的期望也可以写成两部分：

$$\begin{aligned}\frac{x_i+\alpha_i}{n+\alpha_0} &= \frac{\alpha_i}{n+\alpha_0}+\frac{x_i}{n+\alpha_0} \\ &= \underbrace{\frac{\alpha_0}{n+\alpha_0}\times\frac{\alpha_i}{\alpha_0}}_{\text{Prior mean}}+\underbrace{\frac{n}{n+\alpha_0}\times\frac{x_i}{n}}_{\text{Sample mean}}\end{aligned} \tag{21}$$

其中，$\alpha_0=\sum_{i=1}^{K}\alpha_i$，$n=\sum_{i=1}^{K}x_i$。

以本章“鸡兔猪”为例，先验分布为 Dir(α_1, α_2, α_3)，α_1/α_0 代表动物中鸡的比例，α_2/α_0 为兔子比例，α_3/α_0 为猪的比例。

抽取 n 只动物，其中 x_1 只鸡、x_2 只兔、x_3 只猪，比例分别对应 x_1/n、x_2/n、x_3/n。

如图 18 所示，后验分布 Dir($\alpha_1 + x_1$, $\alpha_2 + x_2$, $\alpha_3 + x_3$) 代表“先验 + 数据”融合得到“后验”。

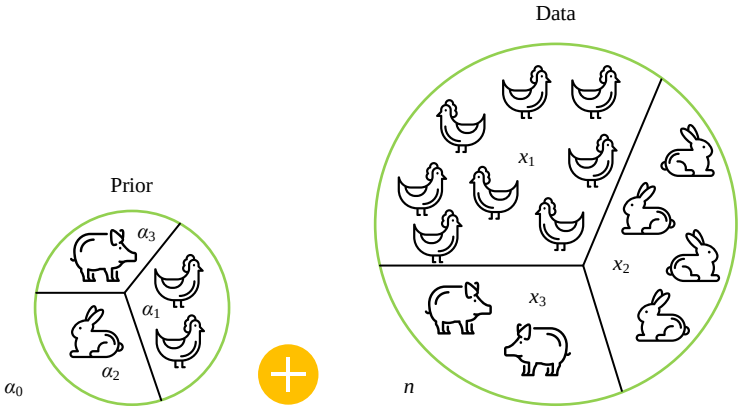


图 18. “混合”先验、样本数据

贝叶斯可信区间

实际上，贝叶斯推断中，我们直接采用后验分布得到模型参数的各种推断，比如点估计、区间估计等等。最大化后验 MAP 就是点估计的一种。贝叶斯推断中，我们还会遇到**可信区间** (credible interval)。

贝叶斯推断的可信区间不同于本书第 16 章介绍的置信区间。在频率学派中，模型参数是固定值，而样本是随机的。因此，样本的**置信区间** (confidence interval) 代表参数的真实值落在该区间内的概率为 $1-\alpha$。

由于贝叶斯学派认为模型参数是一个随机变量，可信区间本身就是随机变量的一个取值范围。随着样本增多，对参数信心增强，可信区间缩窄。

下一章中，大家会发现贝叶斯推断中常用 94%双尾可信区间。图 19 所示为不同 Beta 分布的 94%双尾可信区间，左、右尾分别对应 3%。当概率密度曲线非对称时，我们可以发现区间左右端点对应的概率密度值一般不同。

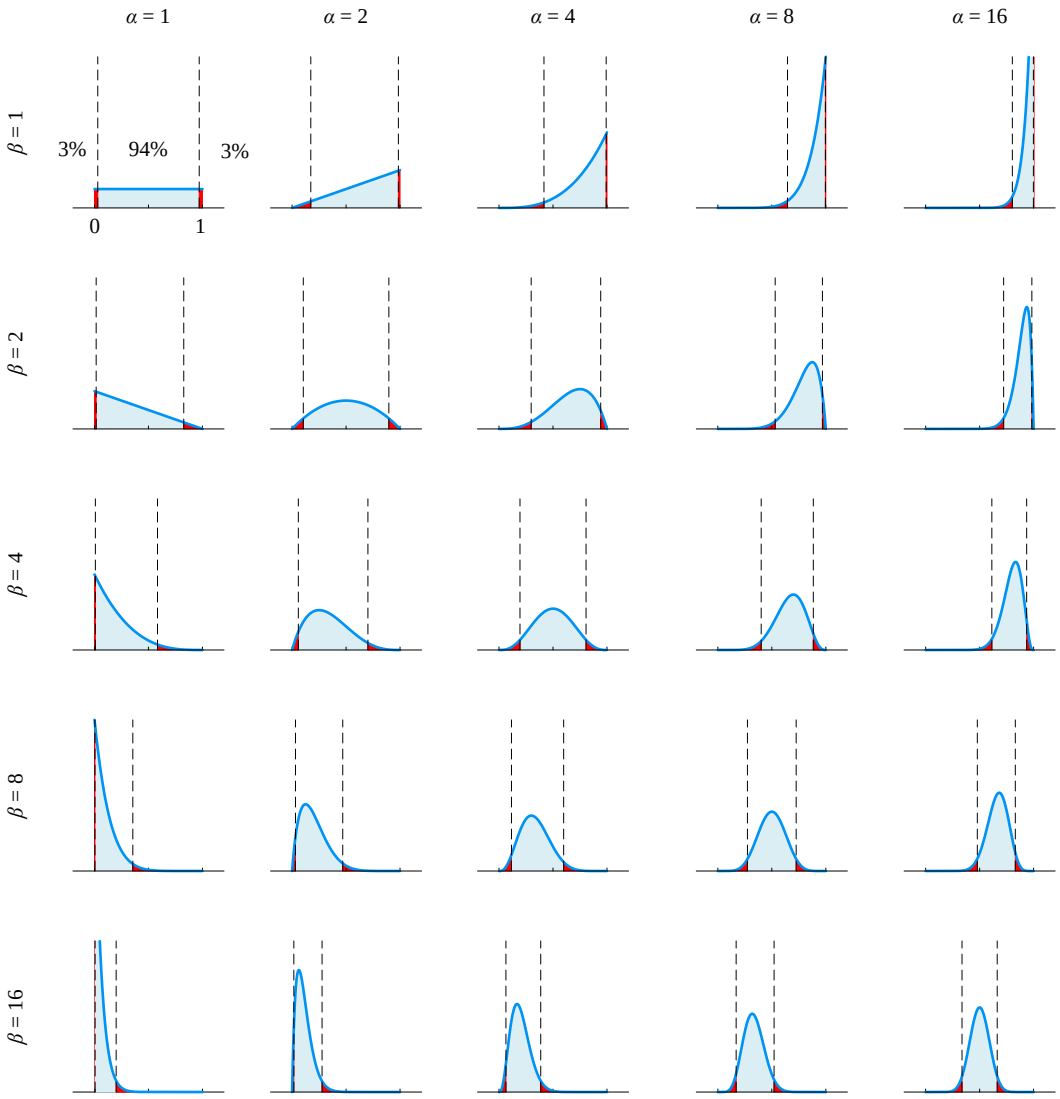


图 19. 比较 Beta(α, β) 分布 94%双尾可信区间

共轭先验

选择先验是有技巧的！

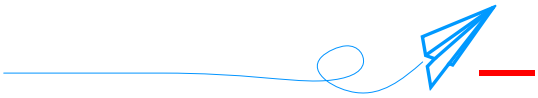
为了方便运算，在 $f_{\Theta|X}(\theta|x)=\dfrac{f_{X|\Theta}(x|\theta)\,f_{\Theta}(\theta)}{\int_{\vartheta} f_{X|\Theta}(x|\vartheta)\,f_{\Theta}(\vartheta)\mathrm{d}\vartheta}$ 中，选取合适的先验分布 $f_{\Theta}(\theta)$ 能让后验分布 $f_{\Theta|X}(\theta|x)$ 和先验分布 $f_{\Theta}(\theta)$ 具有相同的数学形式。

这就是上一章提到的，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)。

简单来说，在贝叶斯统计学中，如果我们选择先验分布和似然函数为特定的概率分布，那么我们可以计算得到一个具有相同函数形式的后验分布，这种性质被称为共轭性，对应的先验分布和后验分布就被称为共轭先验分布和共轭后验分布。

使用共轭先验，无需计算积分就可以得到后验的闭式解。我们仅仅需要跟新观察到的样本数据即可。

上一章的二项分布、Beta 分布，这一章的多项分布、Dirichlet 分布都是成对共轭分布。其他常用的成对共轭分布有：泊松分布-Gamma 分布，正态分布-正态分布，几何分布-Gamma 分布。



本章把贝叶斯推断的维度从二元提高到了三元。先验分布采用了 Dirichlet 分布，似然分布采用多项分布，而后验分布还是 Dirichlet 分布。Beta 分布可以视作 Dirichlet 分布的特例。同理，二项分布可以视作多项分布的特例。

贝叶斯推断中，后验 $\propto$ 似然 $\times$ 先验，无疑是最重要的关系。这个比例关系足可以确定后验概率的形状，我们只需要找到一个归一化常数让后验分布在整个域上积分为 1。

本章还比较了不同 Beta 分布的众数、中位数、均值，以及它们在贝叶斯统计中的适用场合。

上一章和本章中，我们很“幸运地”避免了复杂积分运算，这是因为我们选用了共轭分布。下一章将介绍如何用马尔科夫链蒙特卡罗模拟获得后验分布。