贝叶斯数据分析

贝叶斯数据分析笔记

作者：韩方成
组织：中国海洋大学水产学院
时间：Dec 5, 2021

光并非太阳的专利，你也可以发光

目录

第 1 章 概率和推理

1.1 贝叶斯数据分析的三大步骤

贝叶斯数据分析的过程可以被分为一下三步：

1. 建立全概率模型，即一个问题中所有可观测和不可观测的量的联合概率密度。
2. 计算后验概率：在给定已知数据下未知参数的分布。
3. 评估模型的拟合程度。

1.2 贝叶斯推理

1.2.1 概率符号

我们定义变异系数为 $\mathrm{sd}(\theta)/E(\theta)$，几何平均数为 $\exp(E[\log(\theta)])$，并且几何标准差为 $\exp(\mathrm{sd}[\log(\theta)])$。

1.2.2 贝叶斯法则

贝叶斯法则可以写为：

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{\int_\theta p(\theta)p(y|\theta)d\theta} \tag{1.1}$$

因为在给定 y 的条件下 $p(y)$ 为常数，与 θ 无关，所以我们常将1.1写为式1.2所示的非标准化后验分布：

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{1.2}$$

1.2.3 预测

为了对一个未知的可观测数据做预测，即预测推断。在数据 y 被考虑之前，未知但是可观测的 y 的分布为

$$p(y) = \int p(y, \theta)d\theta = \int p(\theta)p(y|\theta)d\theta \tag{1.3}$$

这被称为先验预测分布：先验是因为它没有以任何先前的观测为条件，预测是因为他是一个可观测数据的分布。

当数据 y 被观测到后，我们可以用同样的方式来预测一个未知可观测的 $\tilde{y}$。$\tilde{y}$ 的分布被称为后验预测分布，后验是因为它以观测 y 为条件，预测是因为他是可观测数据 $\tilde{y}$ 的预测：

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)d\theta \\ &= \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \end{aligned} \tag{1.4}$$

1.2.4 似然和比率

在给定模型下在点 θ_1 和 θ_2 的后验密度 $p(\theta|y)$ 比值称为 θ_1 相对于 θ_2 的后验几率：

$$\frac{p(\theta_1|y)}{p(\theta_2|y)} = \frac{p(\theta_1)p(y|\theta_1)/p(y)}{p(\theta_2)p(y|\theta_2)/p(y)} = \frac{p(\theta_1)}{p(\theta_2)}\frac{p(y|\theta_1)}{p(y|\theta_2)} \tag{1.5}$$

由此可以看出后验比率是先验比率和似然比例的乘积。

1.3 概率论中的一些有用的结果

我们经常用下列方式表示随机变量 u 的期望：

$$E(u) = E(E(u|v)) \tag{1.6}$$

公式1.6很好证明：

$$\iint up(u,v)dudv = \iint u\,p(u|v)du\,p(v)dv = \int E(u|v)p(v)dv \tag{1.7}$$

方差包括两项：

$$\mathrm{var}(u) = E(\mathrm{var}(u|v)) + \mathrm{var}(E(u|v)) \tag{1.8}$$

式1.8的证明为：

$$\begin{aligned}
\mathrm{E}(\mathrm{var}(u \mid v)) + \mathrm{var}(\mathrm{E}(u \mid v)) &= \mathrm{E}\left(\mathrm{E}\left(u^2 \mid v\right) - (\mathrm{E}(u \mid v))^2\right) + \mathrm{E}\left((\mathrm{E}(u \mid v))^2\right) - (\mathrm{E}(\mathrm{E}(u \mid v)))^2 \\
&= \mathrm{E}\left(u^2\right) - \mathrm{E}\left((\mathrm{E}(u \mid v))^2\right) + \mathrm{E}\left((\mathrm{E}(u \mid v))^2\right) - (\mathrm{E}(u))^2 \\
&= \mathrm{E}\left(u^2\right) - (\mathrm{E}(u))^2 \\
&= \mathrm{var}(u)
\end{aligned}$$

1.3.1 变量的变换

如果 p_u 为一个离散分布，并且 f 为一个一对一的函数，而 $v = f(u)$，则 v 的概率密度为

$$p_v(v) = p_u(f^{-1}(v))$$

如果 f 是一个多对一函数，那么会对 $p_v(v)$ 有一个求和。

如果 p_u 为一个连续分布且 $v = f(u)$ 是一个一对一的函数，则

$$p_v(v) = |J|p_u(f^{-1}(v))$$

其中 $|J|$ 为变换 $u = f^{-1}(v)$ 的雅可比行列式，如果 f 为多对一函数，则 $p_v(v)$ 上有积分或求和项。

当处理一维变量时，我们经常利用对数函数来将参数空间从 $(0, \infty)$ 转换到 $(-\infty, \infty)$。当处理唯一开单位区间 $(0, 1)$ 上的参数时，我们用 logistic 变换：

$$\mathrm{logit}(u) = \log\left(\frac{u}{1-u}\right) \tag{1.9}$$

它的逆变换为：

$$\mathrm{logit}^{-1}(v) = \frac{e^v}{1+e^v}$$

定理 1.1

如果 X 是一个连续随机变量，并且它的累积概率密度函数为 F_X，则 $U = F_X(X) \sim U(0,1)$ ♡

证明 在证明之前我们先定义：

$$F_X^{-1}(u) = \inf\{x : F_X(x) = u\}, 0 < u < 1$$

如果随机变量 $U \sim U(0,1)$，则对于所有 $x \in \mathbb{R}$：

$$
\begin{aligned}
P\left(F_X^{-1}(U) \leq x\right) &= P\left(\inf\left\{t : F_X(t) = U\right\} \leq x\right) \\
&= P\left(U \leq F_X(x)\right) \\
&= F_U\left(F_X(x)\right) \\
&= F_X(x)
\end{aligned}
$$

得证。

第 2 章 单参数模型

没有搞明白的章节

- 章节2.6无信息先验分布

本节我们将介绍二项分布、高斯分布、泊松分布和指数分布的一维模型。

2.1 从二项分布中估计概率

一个二项分布模型为：

$$p(y \mid \theta) = \mathrm{Bin}(y \mid n, \theta) = \begin{pmatrix} n \\ y \end{pmatrix} \theta^y (1-\theta)^{n-y} \tag{2.1}$$

我们假设 θ 的先验分布为 $[0,1]$ 上的均匀分布，所以 $p(y) = 1$[1]。又因为 $\begin{pmatrix} n \\ y \end{pmatrix}$ 为常数，所以将式1.1应用大式2.1，我们有：

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y} \tag{2.2}$$

根据附录A关于 Beta 函数的讨论我们可知：

$$\theta|y \sim \mathrm{Beta}(y+1, n-y+1) \tag{2.3}$$

2.1.1 预测

假设我们现在预测抛一枚硬币为上的概率：

$$\begin{aligned} \Pr(\tilde{y}=1|y) &= \int_0^1 \Pr(\tilde{y}=1|\theta, y) p(\theta|y) d\theta \\ &= \int_0^1 \theta p(\theta|y) d\theta = E(\theta|y) = \frac{y+1}{n+2} \end{aligned} \tag{2.4}$$

2.2 后验作为数据和先验信息的折中

因为后验分布结合了数据的信息，它的方差应该比先验的要小。这个概念在下面的第二个表达式被形式化了：

$$\mathrm{E}(\theta) = \mathrm{E}(\mathrm{E}(\theta|y)) \tag{2.5}$$

和

$$\mathrm{var}(\theta) = \mathrm{E}(\mathrm{var}(\theta|y)) + \mathrm{var}(\mathrm{E}(\theta|y)) \tag{2.6}$$

这两个公式分别来自于式1.6和式1.8。从上面两个式子我们看出，θ 的先验均值是所有可能后验均值的期望，并且后验方差的期望是小于先验方差的，其大小取决于后验均值的方差。也就是说，后验均值的方差越大，θ 的不

[1]这里的 $p(y)$ 也指概率密度函数，因为连续变量的贝叶斯公式为 $f_{\theta|y}(\theta|y) = \frac{f_{y|\theta}(y|\theta) f_\theta(\theta)}{f_y(y)}$

确定性减小得就越大。

2.3 信息先验分布

在上面关于 Beta 分布的例子，我们采用的先验分布是 $[0,1]$ 上的均匀分布，那么如果我们使用其他先验分布会怎样？

2.3.1 具有不同先验分布的伯努利分布的例子

考虑似然作为 θ 的一个函数，其形式为：

$$p(y|\theta) \propto \theta^a(1-\theta)^b$$

因此如果先验分布是具有自己的 a, b 的相同形式的分布，那么后验分布也会有相同的形式。我们将这样的一个先验分布参数化为：

$$p(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

为一个参数为 α, β 的 Beta 分布：$\theta \sim \mathrm{Beta}(\alpha, \beta)$。与似然函数相比，我们可以将先验密度看作是 $\alpha-1$ 次成功和 $\beta-1$ 次失败，先验分布的参数被称为超参数。超参数的选择将在之后涉及到。

现在假设我们已经可以选择合适的 α, β，则 θ 的后验分布密度为：

$$\begin{aligned} p(\theta \mid y) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &\propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \\ &= \mathrm{Beta}(\theta \mid \alpha+y, \beta+n-y) \end{aligned}$$

先验和后验分布均为 Beta 分布，这样先验分布和后验分布具有相同的参数形式被称为共轭分布。

则后验分布的均值为：

$$\mathrm{E}(\theta|y) = \frac{\alpha+y}{\alpha+\beta+n}$$

方差为

$$\mathrm{var}(\theta|y) = \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} = \frac{\mathrm{E}(\theta|y)[1-\mathrm{E}(\theta|y)]}{\alpha+\beta+n+1}$$

实际上，在后面的章节中我们会详细介绍，根据中心极限定理我们有：

$$\left(\frac{\theta-\mathrm{E}(\theta \mid y)}{\sqrt{\mathrm{var}(\theta \mid y)}} \mid y\right) \to \mathrm{N}(0,1)$$

2.3.2 共轭先验分布

共轭定义如下。如果 $\mathcal{F}$ 是一类抽样分布 $p(y|\theta)$，并且 $\mathcal{P}$ 是 θ 的一类先验分布，则 $\mathcal{P}$ 对于 $\mathcal{F}$ 是共轭的如果：

$$p(\theta|y) \in \mathcal{P} \text{ for all } p(\cdot|\theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

2.3.3 共轭先验分布、指数族和充分统计量

族 $\mathcal{F}$ 被称为指数族如果它的所有成员都满足如下形式：

$$p(y_i\theta) = f(y_i)g(\theta)e^{\phi(\theta)^T u(y_i)}$$

向量 $\phi(\theta)$ 被称为族 $\mathcal{F}$ 的自然参数 (natural parameter)。独立同分布的观测序列 $y=(y_1,\cdots,y_n)$ 的似然函数为:

$$p(y|\theta)=\left(\prod_{i=1}^{n}f(y_i)\right)g(\theta)^n\exp\left(\phi(\theta)^T\sum_{i=1}^{n}u(y_i)\right)$$

对于所有 n 和 y，这个有一个固定的形式：

$$p(y|\theta)\propto g(\theta)^n e^{\phi(\theta)^T t(y)},\quad \text{where } t(y)=\sum_{i=1}^{n}u(y_i)$$

$t(y)$ 被称为 θ 的充分统计量，因为 θ 的似然仅通过值 $t(y)$ 依赖于数据 y。充分统计量在概率和后验分布的代数处理中是非常有用的。如果先验密度为:

$$p(\theta)\propto g(\theta)^\eta e^{\phi(\theta)^T v}$$

则后验分布为:

$$p(\theta|y)\propto g(\theta)^{\eta+n}e^{\phi(\theta)^T(v+t(y))}$$

2.4 具有已知方差的正态分布

2.4.1 一个点的似然

我们假设 σ^2 已知，则采样分布为:

$$p(y|\theta)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(y-\theta)^2}$$

2.4.2 共轭先验分布

作为 θ 的函数，似然是 θ 的二次指数形式，因此共轭先验分布族应该有如下形式:

$$p(\theta)=e^{A\theta^2+B\theta+C}$$

我们将这个族参数化:

$$p(\theta)\propto\exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)$$

也就是，$\theta\sim\mathrm{N}(\mu_0,\tau_0^2)$。其中 μ_0,τ_0^2 为超参数，我们假设它们已知。

则我们的后验分布为:

$$p(\theta|y)\propto\exp\left(-\frac{1}{2}\left(\frac{(y-\theta)^2}{\sigma^2}+\frac{(\theta-\mu_0)^2}{\tau_0^2}\right)\right)$$

整理后，我们有

$$p(\theta|y)\propto\exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)\tag{2.7}$$

也就是，$\theta|y\sim\mathrm{N}(\mu_1,\tau_1^2)$，其中

$$\mu_1=\frac{\frac{1}{\tau_0^2}\mu_0+\frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2}+\frac{1}{\sigma_0^2}}\quad\text{and}\quad\frac{1}{\tau_1^2}=\frac{1}{\tau_0^2}+\frac{1}{\sigma^2}\tag{2.8}$$

2.4.3 先验和后验分布的精度

我们通常称方差的倒数为精度，由式2.8可知，后验分布的均值可以看为先验分布和数据的加权平均，后验精度为先验精度和数据精度的和。我们也可以将后验均值拆成如下形式：

$$\mu_1 = \mu_0 + (y-\mu_0)\frac{\tau_0^2}{\sigma^2+\tau_0^2}$$
$$\mu_1 = y - (y-\mu_0)\frac{\sigma^2}{\sigma^2+\tau_0^2}$$

因此我们可以得到：

$$\mu_1 = \mu_0 \quad \text{if} \quad y=\mu_0 \text{ or } \tau_0^2 = 0$$
$$\mu_1 = y \quad \text{if} \quad y=\mu_0 \text{ or } \sigma^2 = 0$$

2.4.4 后验预测分布

对于未来观测 $\tilde{y}$ 的后验预测分布 $p(\tilde{y}|y)$，可以通过积分直接算出来：

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &= \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right)\exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right)d\theta \end{aligned}$$

我们知道 $\mathrm{E}(\tilde{y}|\theta)=\theta, \mathrm{var}(\tilde{y}|\theta)=\sigma^2$，于是根据式1.6和1.8我们有：

$$\mathrm{E}(\tilde{y}|y) = \mathrm{E}(\mathrm{E}(\tilde{y}|\theta,y)|y) = \mathrm{E}(\theta|y) = \mu_1$$

和

$$\begin{aligned} \mathrm{var}(\tilde{y}\mid y) &= \mathrm{E}(\mathrm{var}(\tilde{y}\mid\theta,y)\mid y) + \mathrm{var}(\mathrm{E}(\tilde{y}\mid\theta,y)\mid y) \\ &= \mathrm{E}\left(\sigma^2\mid y\right) + \mathrm{var}(\theta\mid y) \\ &= \sigma^2 + \tau_1^2 \end{aligned}$$

2.4.5 具有多个观测的正态分布模型

假设我们的数据 $y=(y_1,\cdots,y_n)$ 为独立同分布的。则后验密度为：

$$\begin{aligned} p(\theta\mid y) &\propto p(\theta)p(y\mid\theta) \\ &= p(\theta)\prod_{i=1}^n p\left(y_i\mid\theta\right) \\ &\propto \exp\left(-\frac{1}{2\tau_0^2}\left(\theta-\mu_0\right)^2\right)\prod_{i=1}^n \exp\left(-\frac{1}{2\sigma^2}\left(y_i-\theta\right)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\tau_0^2}\left(\theta-\mu_0\right)^2 + \frac{1}{\sigma^2}\sum_{i=1}^n\left(y_i-\theta\right)^2\right)\right) \end{aligned}$$

后验分布仅仅通过采样均值 $\bar{y}=\frac{1}{n}\sum_i y_i$ 与 y 有关，因此 $\bar{y}$ 为充分统计量。事实上，根据中心极限定理，我们可知，$\bar{y}|\theta\sim\mathrm{N}(\theta,\sigma^2/n)$。因此我们将 $\bar{y}$ 看作是单个数据，我们有：

$$p(\theta|y_1,\cdots,y_n) = p(\theta|\bar{y}) = \mathrm{N}(\theta|\mu_n,\tau_n^2) \tag{2.9}$$

其中

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}$$

这个结果与每次只用一个数据是一样的。

2.5 其他标准单变量模型

2.5.1 均值已知但是方差未知的正态分布

对于 $p(y|\theta,\sigma^2)=\mathrm{N}(y|\theta,\sigma^2)$，其中 θ 已知 σ^2 未知，对于 n 个独立同分布的观测序列 y 其似然函数为：

$$
\begin{aligned}
p(y|\sigma^2) &\propto \sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right)\\
&=(\sigma^2)^{-n/2}\exp\left(-\frac{n}{2\sigma^2}v\right)
\end{aligned}
$$

充分统计量为：

$$
v=\frac{1}{n}\sum_{i=1}^{n}(y_i-\theta)^2
$$

相应的共轭先验分布是逆 Gamma 分布：

$$
p(\sigma^2)\propto(\sigma^2)^{-(\alpha+1)}e^{-\beta/\sigma^2}
$$

其超参数为 α,β。一种方便的参数化方法是作为 scale 为 σ_0^2 自由度为 v_0 的 scaled 逆卡方分布[2]，也就是说，σ^2 的先验分布为分布 $\sigma_0^2 v_0/X$，其中 X 为服从 $\chi^2_{v_0}$ 的随机变量。为了方便我们使用非标准化的符号：$\sigma^2\sim \mathrm{Inv}-\chi^2(v_0,\sigma_0^2)$。

则 σ 的后验分布为：

$$
\begin{aligned}
p\left(\sigma^2\mid y\right) &\propto p\left(\sigma^2\right)p\left(y\mid\sigma^2\right)\\
&\propto\left(\frac{\sigma_0^2}{\sigma^2}\right)^{\nu_0/2+1}\exp\left(-\frac{\nu_0\sigma_0^2}{2\sigma^2}\right)\cdot\left(\sigma^2\right)^{-n/2}\exp\left(-\frac{n}{2}\frac{v}{\sigma^2}\right)\\
&\propto\left(\sigma^2\right)^{-((n+\nu_0)/2+1)}\exp\left(-\frac{1}{2\sigma^2}\left(\nu_0\sigma_0^2+nv\right)\right)
\end{aligned}
$$

因此，

$$
\sigma^2|y\sim \mathrm{Inv}-\chi^2\left(v_0+n,\frac{v_0\sigma_0^2+nv}{v_0+n}\right)
$$

2.5.2 泊松模型

泊松分布在以计数形式研究数据时自然产生；例如，一个主要的应用领域是流行病学，研究疾病的发病率。

如果数据点 y 服从参数为 θ 的泊松分布，则单个观察 y 的概率分布为：

$$
p(y|\theta)=\frac{\theta^y e^{-\theta}}{y!},\quad \text{for } y=0,1,2,\cdots
$$

对于独立同分布的观测序列 $y=(y_1,\cdots,y_n)$，似然函数为：

$$
\begin{aligned}
p(y|\theta)&=\prod_{i=1}^{n}\frac{1}{y_i!}\theta^{y_i}e^{-\theta}\\
&\propto\theta^{t(y)}e^{-n\theta}
\end{aligned}
$$

其中 $t(y)=\sum_{i=1}^{n}y_i$ 为独立统计量。我们可以重现将其写为指数族的形式：

$$
p(y|\theta)\propto e^{-n\theta}e^{t(y)\log\theta}
$$

[2]不会翻译哈哈

这说明自然参数 $\phi(\theta)=\log\theta$，并且自然共轭先验为：

$$p(\theta)\propto(e^{-\theta})^{\eta}e^{v\log\theta}$$

超参数为 (η,v)。换句话说，似然函数的形式为 $\theta^{a}e^{-b\theta}$，因此共轭先验也必须有 $p(\theta)\propto\theta^{A}e^{-B\theta}$ 的形式。我们用更方便的一种参数化形式：

$$p(\theta)\propto e^{\beta\theta}\theta^{\alpha-1}$$

为参数为 α,β 的 Gamma 分布[3]。有了这样一个共轭先验分布，后验分布为：

$$\theta|y\sim\text{Gamma}(\alpha+n\bar{y},\beta+n)$$

2.5.3 负二项分布

利用共轭族，已知后验分布和后验分布可以被用来找到边缘分布 $p(y)$，利用如下形式，

$$p(y)=\frac{p(y|\theta)p(\theta)}{p(\theta|y)}$$

例如，对于单个观测变量 y 的泊松模型，y 具有如下的先验预测分布：

$$\begin{aligned}p(y)&=\frac{\text{Poisson}(y\mid\theta)\,\text{Gamma}(\theta\mid\alpha,\beta)}{\text{Gamma}(\theta\mid\alpha+y,1+\beta)}\\&=\frac{\Gamma(\alpha+y)\beta^{\alpha}}{\Gamma(\alpha)y!(1+\beta)^{\alpha+y}}\end{aligned}$$

化简为：

$$p(y)=\begin{pmatrix}\alpha+y-1\\y\end{pmatrix}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y}$$

这就是负二项分布：

$$y\sim\text{Neg-bin}(\alpha,\beta)$$

上述推导说明负二项分布为泊松分布的混合，其中泊松分布的参数 θ 符合下列 Gamma 分布：

$$\text{Neg-bin}(y|\alpha,\beta)=\int\text{Poisson}(y|\theta)\,\text{Gamma}(\theta|\alpha,\beta)d\theta$$

2.5.4 根据 rate 和 exposure 参数化的泊松模型

在很多应用中，我们将泊松分布写为如下形式：

$$y_i\sim\text{Poisson}(x_i\theta)\tag{2.10}$$

其中 x_i 为已知的正解释变量，θ 是未知的参数。在流行病学中，θ 被称为 rate，x_i 被称为第 i 个单元的 exposure。则该模型的似然表示为：

$$p(y|\theta)\propto\theta^{\sum_{i=1}^{n}y_i}e^{-(\sum_{i=1}^{n}x_i)\theta}$$

因此 Gamma 分布的先验是共轭的。则后验分布为

$$\theta|y\sim\text{Gamma}\left(\alpha+\sum_{i=1}^{n}y_i,\beta+\sum_{i=1}^{n}x_i\right)$$

[3]Gamma 分布的期望为 $\frac{\alpha}{\beta}$，其方差为 $\frac{\alpha}{\beta^2}$

2.5.5 指数模型

指数分布通常用于模拟“等待时间”和其他连续的，正的，实值的随机变量，通常在时间尺度。在给定参数 θ 下的 y 的采样概率为：

$$p(y|\theta) = \theta\exp(-y\theta), \text{ for } y > 0$$

并且 $\theta = 1/E(y|\theta)$ 被称为 rate。

指数分布具有无记忆性，对象存活额外时间长度的概率与到目前为止所经过的时间无关，即 $\Pr(y > t+s|y > s, \theta) = \Pr(y > t|\theta)$。具有观测序列 $y = (y_1, \cdots, y_n)$ 的似然函数为：

$$p(y|\theta) = \theta^n\exp(-n\bar{y}\theta), \text{ for } \bar{y} \geq 0$$

。

2.6 无信息 (Noninformative) 的先验分布

2.6.1 合适和不合适的先验分布

我们回到估计已知方差为 σ^2 的正态模型的均值 θ 的例子，它的先验分布为 $\mathrm{N}(\mu_0, \tau_0^2)$。如果先验精度 $1/\tau_0^2$ 相对于数据精度 n/σ^2 来说很小，则后验分布近似于

$$p(\theta|y) \approx \mathrm{N}(\theta|\sigma^2/n)$$

但是这样的先验分布的积分不为 1[4]。一般来说，我们认为先验分布 $p(\theta)$ 是合理的 (proper) 如果它不依赖于数据并且积分和为 1。本例中的先验分布是不合理的，但是后验分布式合理的。

2.6.2 Jeffreys's invariance principle

通过函数 $\phi = h(\theta)$ 对参数进行一对一的变换，变换之后得到的先验分布 $p(\phi)$ 与之前的先验分布 $p(\theta)$ 表达同样的信息 (beliefs)。变换之后的先验分布为：

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = p(\theta)|h'(\theta)|^{-1} \tag{2.11}$$

Jeffreys 的一般原则是，任何用于确定先验分布 $p(\theta)$ 的规则应该与应用于变换参数产生等效的结果：也就是说，通过确定 $p(\theta)$ 和应用变换2.11得到的 $p(\phi)$ 应该与直接通过变换模型得到的结果相一致。

根据 Jeffreys 定理我们可以定义无信息先验密度为 $p(\theta) \propto [J(\theta)]^{1/2}$，其中 $J(\theta)$ 为 Fisher 信息：

$$J(\theta) = \mathrm{E}\left(\left(\frac{d\log p(y\mid\theta)}{d\theta}\right)^2 \mid \theta\right) = -\mathrm{E}\left(\frac{d^2\log p(y\mid\theta)}{d\theta^2} \mid \theta\right) \tag{2.12}$$

证明 Jeffreys 先验模型对于参数化是不变的，在 $\theta = h^{-1}(\phi)$ 处计算 $J(\theta)$：

$$\begin{aligned} J(\phi) &= -\mathrm{E}\left(\frac{d^2\log p(y\mid\phi)}{d\phi^2}\right) \\ &= -\mathrm{E}\left(\frac{d^2\log p\left(y\mid\theta = h^{-1}(\phi)\right)}{d\theta^2}\left|\frac{d\theta}{d\phi}\right|^2\right) \\ &= J(\theta)\left|\frac{d\theta}{d\phi}\right|^2 \end{aligned} \tag{2.13}$$

[4]这样就相当于 $p(\theta)$ 为常数，所以无穷积分不为 0

因此 $J(\phi)^{1/2} = J(\theta)^{1/2}|\frac{d\theta}{d\phi}|$。

第 3 章 无信息先验分布

以下基于西安交大梅长林老师的讲解。

假设我们没有任何先验信息可用，那么我们应该如何确定先验分布呢？

3.1 广义先验分布

贝叶斯假设：我们对 θ 的任何取值都同等无知，无偏爱。即

$$p(\theta)=\begin{cases} c & \theta\in\Theta \\ 0 & \theta\notin\Theta \end{cases}$$

如果按照这种方法处理的话，我们很容易得到：

- 当 θ 有可数个取值 $(\theta_1,\cdots,\theta_m)$ 时，则
$$p(\theta_i)=\frac{1}{m}$$
- 当 θ 为 (a,b) 上的连续分布时，我们有
$$p(\theta)=\frac{1}{b-a}$$

但是这样就存在问题了，如果 θ 的取值为可列集或者说 $(-\infty,\infty)$，那么我们就无法按照上述方式来定义其先验分布。那么我们需要对概念进行推广。

下面我们举一个例子，这个例子的先验分布不是一个正常的先验分布 (积分不为 1)，但是他能得到一个正常的后验分布。

例题 3.1 设总体 $\mathrm{X}\sim\mathrm{N}(\theta,1),-\infty<\theta<+\infty$，我们设先验分布为 $p(\theta)=c$。则

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{\int_{-\infty}^{\infty}p(x|\theta)p(\theta)d\theta} \\ &= \frac{\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\theta)^2}{2}}c}{\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\theta)^2}{2}}cd\theta} \\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{(\theta-x)^2}{2}}=\mathrm{N}(x,1) \end{aligned}$$

由此可以看出，虽然先验分布不是正常的分布，但是得到的后验分布确是正常的分布。

我们要做的就是寻找和上述例子相似的先验分布，即广义先验分布。

定义 3.1 (广义先验分布)

设总体 $\mathrm{X}\sim p(x|\theta)$。若 θ 的先验分布满足：

1. $p(\theta)\geq 0,\int_{\Theta}p(\theta)d\theta=+\infty$
2. 有 $p(\theta)$ 确定的后验分布 $p(\theta|x)$ 是正常的概率分布。

我们就称 $p(\theta)$ 为广义先验分布。

♣

但是贝叶斯假设还存在一个缺陷，就是它对变量变换不具有不变性。

例题 3.2 假设 $\mathrm{X}\sim\mathrm{N}(0,\sigma^2)$，我们假设 $\theta=\sigma,\theta\in(0,+\infty)$，而 $\eta=\sigma^2,\eta\in(0,+\infty)$。我们可以看出 θ 和 η 是一一对应的，也就说这样的变换并不损失信息。根据贝叶斯假设 $p(\theta)=c,p(\eta)=c'$。但是根据概率密度变量的变换

的公式，我们有

$$p(\eta) = p(\sqrt{\eta})\left|\frac{d\theta}{d\eta}\right| = \frac{1}{2\sqrt{\eta}}p(\sqrt{\eta}) = \frac{1}{2\sqrt{\eta}}c$$

。显然不是常数。于是违背了贝叶斯假设。

3.2 位置参数的无信息先验

定义 3.2 (位置参数)

若概率密度族有如下形式：

$$\{p(x-\theta), -\infty < \theta < +\infty\}$$

则称它为位置参数族，参数 θ 被称为位置参数。♣

例题 3.3 设 $\mathrm{X} \sim p(x-\theta)$。令 $\mathrm{Y} = \mathrm{X} + c, \eta = \theta + c$。则对于 $\mathrm{Y}: p(y-c-\theta) = p(y-\eta)$，所以 Y 的分布仍为位置参数族的成员，所以 (x,θ) 和 (y,η) 的统计问题结构是相同的，所以它们先验分布的结构也应该是相同的。那么 $p(\theta)$ 也应该具有位置不变性。也就是说 $p(\theta) = p(\theta + c)$，对于任意 c，所以其先验分布应该为常数，我们常把常数设为 1，即 $p(\theta) = 1$。

下面证明为什么先验分布为常数：

证明 我们令 $\eta = \theta + c$，则：

$$p(\eta) = p(\theta)\cdot\left|\frac{d\theta}{d\eta}\right| = p(\eta - c)$$

又因为 c 为任意常数，所以 $p(\theta)$ 为常数。

第 A 章 Gamma,Beta 分布

以下内容基于 LDA 的数学八卦。

A.1 Gamma 函数

定义 A.1 (Gamma 函数)

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt \tag{A.1}$$

通过分布积分的方法，可以推导出这个函数有如下的递归性质：

$$\Gamma(x+1) = x\Gamma(x)$$

于是很容易证明，$\Gamma(x)$ 函数可以当成是阶乘在实数集上的延拓，具有如下性质：

$$\Gamma(n) = (n-1)!$$

♣

那么 Gamma 函数是怎么来的呢？这个来源于哥德巴赫在将阶乘推广到实数域的问题上，欧拉解决了这个问题，他发现 $n!$ 可以用如下的一个无穷乘积表达：

$$\left[\left(\frac{2}{1}\right)^n \frac{1}{n+1}\right]\left[\left(\frac{3}{2}\right)^n \frac{2}{n+2}\right]\left[\left(\frac{4}{3}\right)^n \frac{3}{n+3}\right]\cdots = n! \tag{A.2}$$

用极限形式，这个式子整理后可以写为

$$\lim_{m\to\infty} \frac{1\cdot 2\cdot 3\cdots m}{(1+n)(2+n)\cdots(m+n)}(m+1)^n = n! \tag{A.3}$$

左边可以整理为：

$$\begin{aligned}
&\frac{1\cdot 2\cdot 3\cdots m}{(1+n)(2+n)\cdots(m+n)}(m+1)^n \\
=&1\cdot 2\cdot 3\cdots n\cdot \frac{(n+1)(n+2)m}{(1+n)(2+n)\cdots m}\cdot \frac{(m+1)^n}{(m+1)(m+2)\cdots(m+n)} \\
=&n!\frac{(m+1)^n}{(m+1)(m+2)\cdots(m+n)} \\
=&n!\prod_{k=1}^{n}\frac{m+1}{m+k} \to n! \quad (m\to\infty)
\end{aligned}$$

所以上式成立。

之后欧拉对得到的式子进行了处理，得到了积分的形式：

$$n! = \int_0^1 (-\log t)^n dt$$

如果我们令 $t = e^{-u}$，就可得我们常见的 Gamma 函数形式：

$$n! = \int_0^{\infty} u^n e^{-u} du$$

于是，利用上式把阶乘延拓到实数集上，我们就得到了 Gamma 函数的一般形式：

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

Gamma 函数不仅可以用来计算 $(1/2)!$，还可以拓展到许多其他的数学概念。比如导数，我们可以把导数的定义

拓展到实数集，从而可以计算 1/2 阶导数。我们考虑 x^n 的导数，我们知道，x^n 的 k 阶导数为：

$$n(n-1)(n-2)\cdots(n-k+1)x^{n-k}=\frac{n!}{(n-k)!}x^{n-k}$$

由于 k 阶导数可以用阶乘表达，于是我们用 Gamma 函数表达式为：

$$\frac{\Gamma(n+1)}{\Gamma(n-k+1)}x^{n-k}$$

于是基于上式，我们可以把导数的阶从整数延拓到实数集。例如，取 $n=1,k=\frac{1}{2}$，我们可以计算 x 的 $\frac{1}{2}$ 阶导数为：

$$\frac{\Gamma(1+1)}{\Gamma(1-1/2+1)}x^{1-1/2}=\frac{2\sqrt{x}}{\sqrt{\pi}}$$

Gamma 函数的图像如图A.1所示，由此可以看出 Gamma 函数为一个凸函数。不仅如此，$\log(\Gamma(x))$ 也是一个凸函

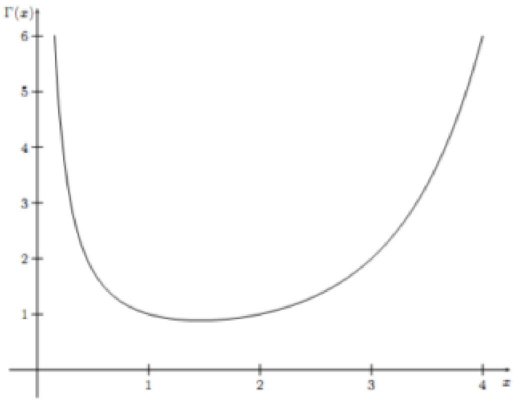

图 A.1: $\Gamma(x)$

数。我们可以用以下定理证明：

定理 A.1 (Bohr-Mullerup)

如果 $f:(0,\infty)\to(0,\infty)$，且满足

1. $f(1)=1$
2. $f(x+1)=xf(x)$
3. $\log f(x)$ 是凸函数

那么 $f(x)=\Gamma(x)$，也就是 $\Gamma(x)$ 是唯一满足以上条件的函数。

如下函数被称为 Digamma 函数，

$$\Psi(x)=\frac{d\log\Gamma(x)}{dx}$$

这也是一个很重要的函数，在涉及求 Dirichlet 分布相关的参数的极大似然估计时，往往需要用到这个函数。Digamma 函数具有如下一个漂亮的性质：

$$\Psi(x+1)=\Psi(x)+\frac{1}{x}$$

A.2 从二项分布到 Gamma 分布

对 Gamma 函数的定义做一个变形，就可以得到如下式子：

$$\int_0^\infty \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}dx = 1$$

于是，取积分中的函数作为概率密度，就得到一个形式最简单的 Gamma 分布的密度函数：

$$\text{Gamma}(x|\alpha) = \frac{x^{\alpha-1}e^{-x}}{\Gamma(\alpha)}$$

如果做一个变换 $x = \beta t$，就得到 Gamma 分布的更一般形式

$$\text{Gamma}(t|\alpha,\beta) = \frac{\beta^\alpha t^{\alpha-1}e^{-\beta t}}{\Gamma(\alpha)}$$

其中 α 为 shape parameter，主要决定分布曲线的形状，而 β 称为 rate parameter 或者 inverse scale parameter($\frac{1}{\beta}$ 被称为 scale parameter)，主要决定曲线有多陡。如图A.2所示。

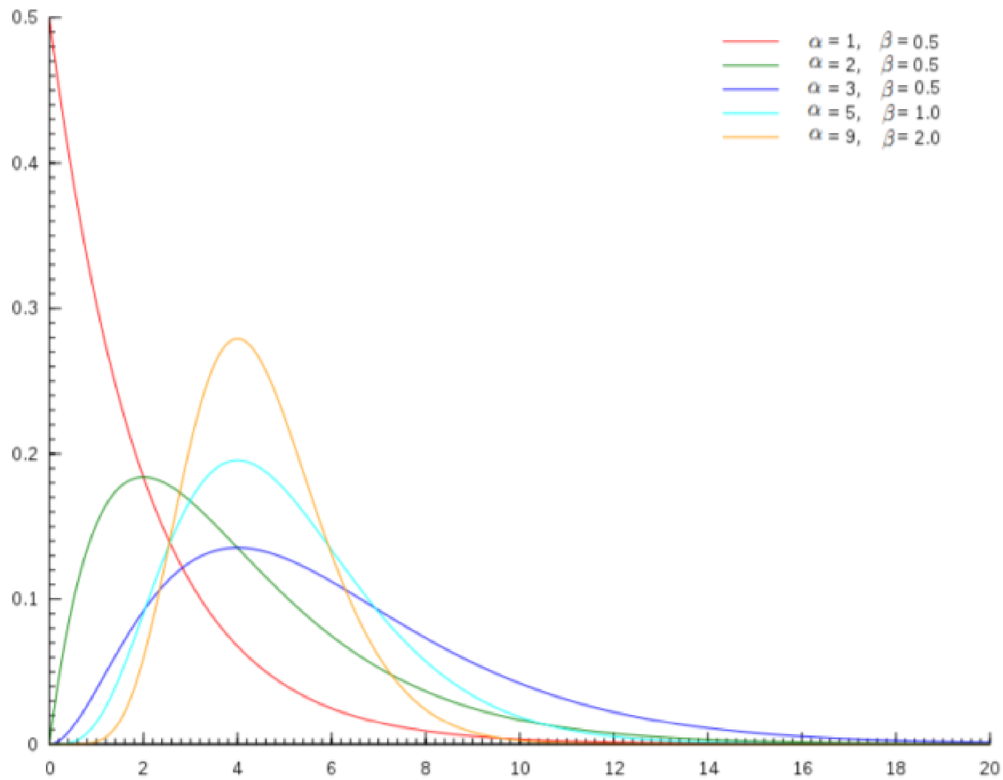

图 A.2: Γ 分布

下面我们先讨论 $\beta = 1$ 情况。Gamma 分布首先与 Poisson 分布、Poisson 过程发生密切的联系。参数为 λ 的 Poisson 分布，概率写为：

$$\text{Poisson}(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

在 Gamma 分布的密度函数取 $\alpha = k + 1$ 得到

$$\text{Gamma}(x|\alpha = k+1) = \frac{x^k e^{-x}}{k!}$$

所以这两个分布数学形式上是一样的。这个并不是偶然的。我们在数理统计中学过，Poisson(λ) 可以看作是二项分布 $B(n,p)$ 在 $np=\lambda, n\to\infty$ 条件下的极限分布。其实二项分布的随机变量 $X\sim B(n,p)$ 满足如下一个很奇妙的恒等式：

$$P(X\le k)=\frac{n!}{k!(n-k-1)!}\int_p^1 t^k(1-t)^{n-k-1}dt \tag{A.4}$$

我们之后再证明它。我们先暂且承认A.4是成立的。我们在等式右边做一个变换 $t=\frac{x}{n}$ 得到：

$$\begin{aligned}P(X\le k)&=\frac{n!}{k!(n-k-1)!}\int_p^1 t^k(1-t)^{n-k-1}dt\\&=\frac{n!}{k!(n-k-1)!}\int_{np}^n\left(\frac{x}{n}\right)^k\left(1-\frac{x}{n}\right)^{n-k-1}d\frac{x}{n}\\&=\frac{(n-1)!}{k!(n-k-1)!}\int_{np}^n\left(\frac{x}{n}\right)^k\left(1-\frac{x}{n}\right)^{n-k-1}dx\\&=\int_{np}^n\left(\begin{array}{c}n-1\\k\end{array}\right)\left(\frac{x}{n}\right)^k\left(1-\frac{x}{n}\right)^{n-k-1}dx\\&=\int_{np}^n\text{Binomial}\left(Y=k\mid n-1,\frac{x}{n}\right)dx\end{aligned} \tag{A.5}$$

于是我们有：

$$\text{Binomial}(X\le k|n,p)=\int_{np}^n\text{Binomial}(Y=k|n-1,\frac{x}{n})dx \tag{A.6}$$

我们对式A.6在条件 $np=\lambda, n\to\infty$ 下取极限，则左边有 $B(n,p)\to$ Poisson(λ)，而右边有 $B(n-1,\frac{x}{n})\to$ Poisson(x)，所以得到：

$$\text{Poisson}(X\le k|\lambda)=\int_\lambda^\infty\text{Poisson}(Y=k|x)dx \tag{A.7}$$

把上式右边的 Possion 分布展开，得到：

$$\text{Poisson}(X\le k|\lambda)=\int_\lambda^\infty\frac{x^ke^{-x}}{k!}dx \tag{A.8}$$

在 $\lambda\to 0$ 情况下取极限，得：

$$1=\int_0^\infty\frac{x^ke^{-x}}{k!}dx$$

这就得到了 Gamma 函数，另外

$$k!=\int_0^\infty x^ke^{-x}dx$$

即 $k!$ 的积分表示方法。

最后我们将式A.8变形一下，得到：

$$\text{Poisson}(X\le k|\lambda)+\int_0^\lambda\frac{x^ke^{-x}}{k!}dx=1$$

由此可以看出 Poisson 分布的累积概率和 Gamma 分布的累积概率具有互补的性质。

A.3 认识 Beta/Dirichlet 分布

假设我们现在有这样一个问题：

1. $X_1,X_2,\cdots,X_n\overset{\text{iid}}{\sim}\text{Uniform}(0,1)$
2. 把这 n 各随机变量排序后得到顺序统计量 $X_{(1)},X_{(2)},\cdots,X_{(n)}$
3. 问 $X_{(k)}$ 的分布是什么

我们尝试先计算一下 $X_{(k)}$ 落在一个区间 $[x, x+\Delta x]$ 的概率，也就是以下概率取值：

$$P(x \le X_{(k)} \le x + \Delta x) = ?$$

我们把区间 $[0, 1]$ 分为三段 $[0, x), [x, x+\Delta x], (x+\Delta, 1]$，我们先考虑简单的情形，假设 n 各数只有一个落在区间 $[x, x+\Delta x]$ 内，则因为这个区间内的数 $X_{(k)}$ 是第 k 大的，则 $[0, x)$ 中应有 $k-1$ 个数，$(x, 1]$ 区间应该有 $n-k$ 个数。不失一般性，我们先考虑如下一个符合上述要求的事件 E：

$$\begin{aligned} & X_1 \in [x, x+\Delta x], \\ E = \; & X_i \in [0, x) \quad (i = 2, \cdots, k) \\ & X_j \in (x+\Delta x, 1] \quad (j = k+1, \cdots, n) \end{aligned}$$

如图A.3所示。则有：

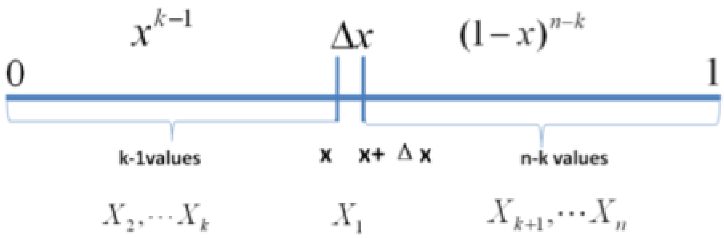

图 **A.3:** 事件 E

$$\begin{aligned} P(E) &= \prod_{i=1}^{n} P(X_i) \\ &= x^{k-1}(1-x-\Delta x)^{n-k}\Delta x \\ &= x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x) \end{aligned}$$

显然，由于不同的排列组合，即 n 个数中有一个落在 $[x, x+\Delta x]$ 区间的有 n 种取法，余下 $n-1$ 数中有 $k-1$ 个落在 $[0, x)$ 的有 $\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}$，所以与 E 等价的事件一共有 $n \begin{pmatrix} n-1 \\ k-1 \end{pmatrix}$ 个。

下面我们考虑一个更为复杂的情况，假设 n 个数中有两个数落在了区间 $[x, x+\Delta x]$，

$$\begin{aligned} & X_1, X_2 \in [x, x+\Delta], \\ E' = \; & X_i \in [0, x) \quad (i = 3, \cdots, k) \\ & X_j \in (x+\Delta x, 1] \quad (k = k+1, \cdots, n) \end{aligned}$$

则有：

$$P(E') = x^{k-2}(1-x-\Delta x)^{n-k}(\Delta x)^2 = o(\Delta x)$$

所以只有落入 $[x, x+\Delta x]$ 内的数字超过一个，对应的概率就是 $o(\Delta x)$。于是：

$$\begin{aligned} & P\left(x \le X_{(k)} \le x+\Delta x\right) \\ &= n \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} P(E) + o(\Delta x) \\ &= n \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} x^{k-1}(1-x)^{n-k}\Delta x + o(\Delta x) \end{aligned}$$

所以，可以得到 $X_{(k)}$ 的概率密度函数为：

$$
\begin{aligned}
f(x) &= \lim_{\Delta x \to 0} \frac{P\left(x \le X_{(k)} \le x + \Delta x\right)}{\Delta x} \\
&= n \begin{pmatrix} n-1 \\ k-1 \end{pmatrix} x^{k-1}(1-x)^{n-k} \\
&= \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k} \quad x \in [0,1]
\end{aligned}
$$

，利用 Gamma 我们可以将 $f(x)$ 的表达式写为：

$$
f(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} x^{k-1}(1-x)^{n-k}
$$

我们在上式中取 $\alpha = k, \beta = n-k+1$，于是我们得到：

$$
f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{A.9}
$$

这就是一般意义上的 Beta 分布。

A.4 Beta-Binomial 共轭

假设我们的问题变为：

1. $X_1, X_2, \cdots, X_n \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$，排序后对应的顺序统计量 $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$，我们要猜测 $p = X_{(k)}$
2. $Y_1, Y_2, \cdots, Y_m \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$，$Y_i$ 中有 m_1 个比 p 小，m_2 个比 p 大
3. 问 $P(p|Y_1, Y_2, \cdots, Y_m)$ 的分布是什么。

由于 $p = X_{(k)}$ 在 $X_1, X_2, \cdots, X_n$ 中是第 k 大的，利用 Y_i 的信息，我们容易推理得到 $p = X_{(k)}$ 在 $X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m \overset{\text{iid}}{\sim} \text{Uniform}(0,1)$ 这 $(m+n)$ 个独立随机变量中是第 $k+m_1$ 大的，于是按照上一个小节的推理，此时 $p = X_{(k)}$ 的概率密度函数是 $\text{Beta}(p|k+m_1, n-k+1+m_2)$。按照贝叶斯推理的过程，我们可以把以上过程表述为：

1. $p = X_{(k)}$ 是我们要猜测的参数，我们推导出 p 的分布为 $f(p) = \text{Beta}(p|k, n-k+1)$，称为 p 的先验分布
2. 数据 Y_i 中有 m_1 个比 p 小，m_2 个比 p 大，Y_i 相当于是做了 m 次贝努力实验，所以 m_1 服从二项分布 $\text{B}(m,p)$
3. 在给定了来自数据提供的 (m_1, m_2) 的知识后，p 的后验分布变为 $f(p|m_1, m_2) = \text{Beta}(p|k+m_1, n-k+1+m_2)$

上述可以表示为：

$$
\text{Beta}(p|\alpha,\beta) + \text{BinomCount}(m_1, m_2) = \text{Beta}(p|\alpha+m_1, \beta+m_2) \tag{A.10}
$$

这个式子描述的就是 Beta-Binomial 共轭。从上面过程我们可以看到，Beta 分布中的 α, β 参数都可以理解为物理计数，这两个参数经常被称为伪计数。基于以上逻辑，我们也可以把 $\text{Beta}(p|\alpha,\beta)$ 写成下式来理解：

$$
\text{Beta}(p|1,1) + \text{BinomCount}(\alpha-1, \beta-1) = \text{Beta}(p|\alpha,\beta) \tag{A.11}
$$

其中 $\text{Beta}(0,1)$ 恰好就是 $[0,1]$ 上的均匀分布。