

第二章 贝叶斯模型及推断

Wang Shujia

Contents

1	点估计	2
2	区间估计	5
3	预测	8
4	Beta-Binomial 和 Gamma-Poisson 模型	9
4.1	Beta-Binomial 模型	9
4.2	Gamma-Poisson 模型	10

记号约定 (与教材不同)

X	随机变量
\mathbf{X}	随即向量 $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$
\mathbf{x}	观察值向量 $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$
θ	未知参数
$\boldsymbol{\theta}$	参数向量 $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n)^T$
$\pi(\theta)$	先验分布密度或概率函数
$f(\mathbf{x} \theta)$	随机样本的联合分布或似然函数 (看作 θ 的函数)
$p(\theta \mathbf{x})$	后验分布密度或概率函数
\mathbf{A}	常数矩阵

贝叶斯推断

贝叶斯的一切推断均基于后验分布: $p(\theta|\mathbf{x}) \propto \pi(\theta)L(\theta|\mathbf{x})$

贝叶斯推断包括:

- 点估计
- 区间估计
- 预测
- 假设检验

1 点估计

点估计的概念

定义 1. 假设总体分布为 $f(x|\theta)$, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ 为样本观察值。把后验分布 $p(\theta|\mathbf{x})$ 归纳为一个数 $\hat{\theta}$, 用以估计未知参数 θ , 则 $\hat{\theta}$ 称为 θ 的一个点估计 (Point estimate)

常用的贝叶斯点估计:

1. 后验均值 $\hat{\theta}_E = E(\theta|\mathbf{x})$
 2. 后验中位数 $\hat{\theta}_{Me} = \text{Median}(\theta|\mathbf{x})$
 3. 后验众数 $\hat{\theta}_M = \text{Mode}(\theta|\mathbf{x})$
- 最大似然估计是后验众数估计的特例 (无信息先验)

点估计的误差

定义 2 (MSE and SE). 设参数 θ 的后验分布为 $f(\theta|\mathbf{x})$, $\hat{\theta}$ 为 θ 的一个点估计值, 则 $(\hat{\theta} - \theta)^2$ 的后验均值称为后验均方误差 (Mean Square Error), 即

$$\text{MSE}(\hat{\theta}|\mathbf{x}) = E_{\theta|\mathbf{x}}(\hat{\theta} - \theta)^2$$

$\text{SE} = \sqrt{\text{MSE}}$ 称为后验标准误差 (Standar Error)

- 一般公式: $\text{MSE}(\hat{\theta}|\mathbf{x}) = \text{Var}(\theta|\mathbf{x}) + (\hat{\theta}_E - \hat{\theta})^2$
 – 当 $\hat{\theta} = \hat{\theta}_E$ 时, MSE 最小, 等于 $\text{Var}(\theta|\mathbf{x})$
- 一般贝叶斯点估计取**后验均值**
- 后验方差 $\text{Var}(\theta|\mathbf{x})$ 反映点估计的**精度**

点估计的含义解释

- 假设总体分布为: $X|\theta \sim \text{Binomial}(n, \theta)$
- 先验分布为: $\theta \sim \text{Beta}(\alpha, \beta)$ (称为 Beta-Binomial 模型)
- 则后验分布为: $\theta|x \sim \text{Beta}(x + \alpha, n - x + \beta)$, 其数学期望为

$$\begin{aligned} E(\theta|x) &= \frac{x + \alpha}{(x + \alpha + n - x + \beta)} \\ &= \left(\frac{\alpha + \beta}{\alpha + \beta + n} \right) \underbrace{\frac{\alpha}{\alpha + \beta}}_{\text{prior mean}} + \left(\frac{n}{\alpha + \beta + n} \right) \underbrace{\frac{x}{n}}_{\text{sample mean}} \end{aligned}$$

- 含义: 后验均值等于先验均值和样本均值的加权平均

$$\text{后验均值} = w \times \text{先验均值} + (1 - w) \times \text{样本均值}$$

- 后验均值向先验均值“收缩”, 收缩多少取决于权重 w
- 如 $w = 1$, 则 $E(\theta|x) = E(\theta)$ (先验均值)
- 如果 $n \rightarrow \infty$, 则先验权重 w 趋向于 0, 后验均值 $\approx \text{MLE}$ (解释?)
- 如果样本容量 n 很小, 则先验权重 w 趋向于 1, 此时: 后验均值 \approx 先验均值

例: Florida 总统选举

- 美国 Florida 州在 2000 年 3 月对将于 11 月举行的总统选举进行一项民意调查, 结果: $n = 621$, Bush 45% ($n_1 = 279$), Gore 37% (230), Buchanan 3% (19) and undecided 15% (93).
- 简单起见, 仅考虑 Bush 和 Gore 两个候选人, 结果:
- $n = 509$, Bush(55%, $n_1 = 279$), Gore(45%, $n_2 = 230$).
- 以 θ 表示 Bush 的支持率, 并假设该调查是简单随机抽样。
- 判断布什是否能获胜。

Florida 总统选举的贝叶斯推断

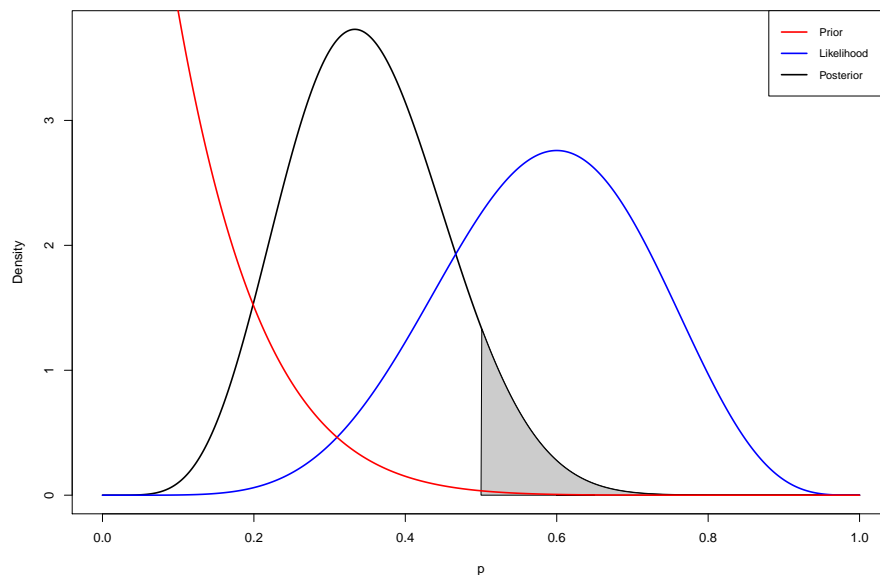
- 最大似然估计: $X = \text{支持布什的人数}$, 观察值 $x = 279$, 则二项分布 $X|\theta \sim \text{Bin}(509, \theta)$
 - 似然函数: $L(\theta) = f(x|\theta) \propto \theta^{279}(1-\theta)^{509-279}$
 - 最大似然估计: $\hat{\theta} = 279/509 = 0.548$
- 这个点估计的精度 (误差多少)? 可能区间?
- 贝叶斯点估计
 - 无信息先验: $\theta \sim \text{Beta}(1, 1) = U[0, 1]$
 $X|\theta \sim \text{Bin}(509, \theta)$
 - 后验分布: $\theta|X \sim \text{Beta}(280, 231)$
 - $E(\theta|x) = 280/(280 + 231) = 0.548$
 - 标准差: $\text{sd}(\text{rbeta}(10000, 280, 231)) = 0.022$
 - 区间估计: $> \text{qbeta}(c(0.025, 0.975), 280, 231) = [0.5046756, 0.5908593]$

女士品茶: 贝叶斯估计

假设女士、音乐家和醉汉都随机测试 10 次, 结果说对 6 次, 即 $X|\theta \sim \text{Bin}(10, \theta)$, $x = 6$ 。

1. 女士品茶: 先验分布 $\theta \sim \text{Beta}(1, 1) = U[0, 1]$
 - 后验分布: $\theta|X \sim \text{Beta}(7, 5)$
 - 后验概率: $P(\theta > 0.5|x = 6) = 0.73$
 - 后验机会比: $\text{odds} = 0.73/0.27 = 2.7$
2. 音乐家识谱: 先验分布 $\theta \sim \text{Beta}(2, 1)$
 - 后验分布: $\theta|X \sim \text{Beta}(8, 5)$
 - 后验概率: $P(\theta > 0.5|x = 6) = 0.81$
 - 后验机会比: $\text{odds} = 0.81/0.19 = 4.3$
3. 醉汉猜硬币: 先验分布 $\theta \sim \text{Beta}(1, 9)$
 - 后验分布: $\theta|X \sim \text{Beta}(7, 13)$
 - 后验概率: $P(\theta > 0.5|x = 6) = 0.08$
 - 后验机会比: $\text{odds} = 0.08/0.992 = 0.087$

醉汉猜硬币的贝叶斯模型图示



醉汉猜硬币模型作图代码

```
p=seq(0,1,length=500)
a=1; b=9
y=6; n=10
prior=dbeta(p,a,b)
like=dbeta(p,y+1,n-y+1)
post=dbeta(p,y+a,n-y+b)
plot(p,post,type='l',ylab="Density",lwd=2,col='black')
x1<-p[p>=0.5]
y1<-dbeta(x1,y+a,n-y+b)
polygon(c(0.5,x1,0.6,0.65),c(0,y1,0,0),col='grey80')
lines(p,like,lwd=2,col='blue')
lines(p,prior,lwd=2,col='red')
legend("topright",c("Prior","Likelihood","Posterior"),
      col=c('red','blue','black'),lwd=c(2,2,2),cex=0.8)
```

2 区间估计

可信区间

定义 3 (Credible region). 对给定样本观察值 \mathbf{x} , 参数 θ 的后验分布为 $p(\theta|\mathbf{x})$ 。如果一个区间 $C = (L, U)$ 使得

$$P(\theta \in C|\mathbf{x}) = 1 - \alpha$$

则称区间 C 为参数 θ 的一个 $100(1-\alpha)\%$ 可信区间 (Credible Interval)。
一般取等尾可信区间 (Equal Tail) :

$$P(\theta \leq L|x) = \int_{-\infty}^L p(\theta|\mathbf{x})d\theta = \frac{\alpha}{2}$$

$$P(\theta \geq U|x) = \int_U^{\infty} p(\theta|\mathbf{x})d\theta = \frac{\alpha}{2}$$

频率学派的置信区间

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)^T$ be a random sample from a population $X \sim f(x|\theta)$, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ be the observed values, and θ be an unknown parameter.

Suppose that we can find $L(\mathbf{X})$ and $U(\mathbf{X})$ such that

$$P(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})) = 1 - \alpha$$

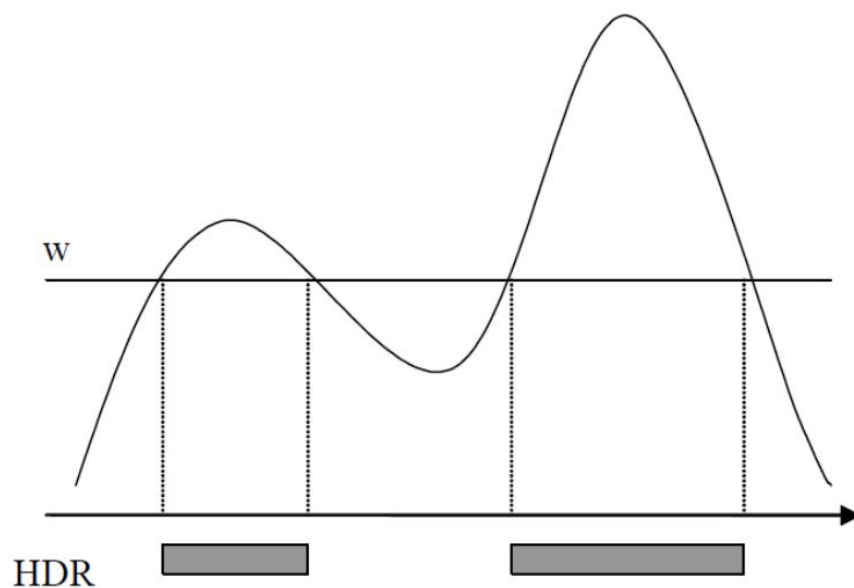
Then $[L(\mathbf{x}), U(\mathbf{x})]$ is called a **confidence interval** for θ , $(1-\alpha) \times 100\%$ is called the **confidence level**.

- $\alpha = 0.05$ is a standard 95% confidence interval.
- The random variable is \mathbf{X} , not the θ .
- Interpret: the *random interval* will overlap the parameter θ 95% of the time.
- “The probability that a confidence interval $[L(\mathbf{x}), U(\mathbf{x})]$ contains the true population parameter is $(1 - \alpha)$ ” (not true).

HPD 区域

一个区域 C 称为 θ 的一个 $100(1-\alpha)\%$ 最高后验密度区域 (Highest Probability Density Region, HPD), 如果 $C = \{\theta : p(\theta|\mathbf{x}) > w\}$, 其中 w 满足

$$\int_C p(\theta|\mathbf{x})d\theta = 1 - \alpha$$



计算 HPD

1. 调用软件包 TeachingDemos:

```
hpd(posterior.icdf, conf=0.95, tol=1e-8,...)
```

2. 自定义 R 函数:

```
HPD = function(ICDFname,credMass=0.95,tol=1e-8,...){
  incredMass=1.0-credMass
  intervalWidth=function(lowTailPr,ICDFname,credMass,...){
    ICDFname(credMass+lowTailPr,...)-ICDFname(lowTailPr,...)}
  optInfo=optimize(intervalWidth,
                    c(0,incredMass),ICDFname=ICDFname,
                    credMass=credMass,tol=tol,...)
  HDIlowTailPr=optInfo$minimum
  return(c(ICDFname(HDIlowTailPr,...),
           ICDFname(credMass+HDIlowTailPr,...)))
  #ICDF-分布函数的反函数
```

可信区间和 HPD

醉汉猜硬币：先验分布 $\theta \sim \text{Beta}(1,9)$ ，后验分布 $\theta|X \sim \text{Beta}(7,13)$

- 可信区间: `> qbeta(c(0.025,0.975),7,13)`
`= [0.1628859,0.5655016]`
- 最高后验密度区间 (HPD):
 - 用 hpd 函数:

```
>library(TeachingDemos)
>hpd(qbeta, shape1 = 7, shape2= 13, conf=0.95)
[1] 0.1537483 0.5543178
```

– 用 R 自定义函数

```
(此处省略函数定义部分)
>HPD(qbeta,shape1=7,shape2=13)
[1] 0.1537483 0.5543178
```

- 可信区间与 HPD 不一致，HPD 区间长度稍短

3 预测

预测分布

总体分布为: $Y \sim f(y|\theta)$, 先验分布为: $\theta \sim \pi(\theta)$, 后验分布 $p(\theta|\mathbf{y})$ 。

已有观察值 $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, 设 \tilde{Y} 是一个未来可能的观察值, 则其分布称为预测分布。

显然, 未来观察值来自相同总体, 因此: $\tilde{Y}|\theta \sim f(\tilde{y}|\theta)$

给定观察值 \mathbf{y} , \tilde{Y} 的后验预测分布 (Posterior predictive distribution) 为:

$$p(\tilde{y}|\mathbf{y}) = \int f(\tilde{y}|\theta)p(\theta|\mathbf{y})d\theta$$

应用:

- 预测: 点预测, 预测区间
- 模型检验: 数据分为两部分: 训练样本 + 检验样本

预测分布的计算

- 直接计算积分: 经常积不出来
- 随机模拟法: 用 MCMC 抽样 (抽取预测分布的样本 $\tilde{y}^{(1)}, \tilde{y}^{(2)}, \dots, \tilde{y}^{(m)}$)
 1. 抽取样本: $\theta^{(k)}|\mathbf{y} \sim p(\theta|\mathbf{y})$
 2. 抽取预测样本: $\tilde{y}^{(k)}|\theta^{(k)} \sim f(\tilde{y}|\theta^{(k)})$

条件分布的期望和方差 (重要公式)

定理 1 (Double Expectation). 设 u, v 是两个随机变量, 如果 $u|v$ 的分布已知, 则

$$\begin{aligned} E(u) &= E_v[E(u|v)] \\ \text{Var}(u) &= \text{Var}_v[E(u|v)] + E_v[\text{Var}(u|v)] \end{aligned}$$

例: 假设一只母虫能孵化出 X 个下一代小虫, 试求 X 的均值和方差 (该母虫的产卵数 $U \sim \text{Poisson}(\lambda)$, 每个卵能孵化成小虫子的概率是 p , 且相互独立)。

- 应用: 计算预测分布的期望、方差和概率。

预测下一个黑天鹅出现的概率

一个人看到 n 只天鹅都是白天鹅，请预测他看到下一只天鹅仍是白天鹅的概率

- 传统方法：假设 Ta 每次观察中看到白天鹅的概率为 θ ， $Y_i = 1 (i = 1, \dots, n)$ 表示第 i 次观察到白天鹅，否则为 0。设 Y 为 n 次观察中白天鹅的个数，则
 - $Y \sim \text{Binomial}(n, \theta)$ ，现在观察到 $y = n$
 - θ 的最大似然估计为： $\hat{\theta} = \bar{y} = n/n = 1$ ，即预测下一只天鹅仍是白天鹅的概率为 100%。
 - 传统方法无法预测到黑天鹅
- 贝叶斯预测：
 - 无信息先验 $\theta \sim \text{Beta}(1, 1)$ ，则后验分布为 $\text{Beta}(y + 1, n - y + 1)$
 - 设 \tilde{Y} 为下一只天鹅的颜色（白：1，黑：0），则 $P(\tilde{Y} = 1|\theta) = \theta$ ，
 - \tilde{Y} 的预测分布为：

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

- 观察值 $y = n$ ，下一只天鹅是白天鹅的预测概率为： $(n+1)/(n+2)$

4 Beta-Binomial 和 Gamma-Poisson 模型

4.1 Beta-Binomial 模型

Beta-Binomial Model

在 n 次独立重复试验中，每次试验事件 A 发生的概率为 θ ，设 X 为事件 A 发生的次数，则 $X|\theta \sim \text{Bin}(n, \theta)$ 。现在某实际试验中观察到 $X = x$ ，试对概率 θ 进行贝叶斯估计。

- 贝叶斯模型：
 - 先验分布： $\theta \sim \text{Beta}(\alpha, \beta)$
 - 总体模型：二项分布 $X|\theta \sim \text{Bin}(n, \theta)$ ： $f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$
 - 后验分布： $\theta|X \sim p(\theta|x) = \pi(\theta)f(x|\theta) \propto \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$
 - 即 $\theta|X \sim \text{Beta}(x + \alpha, n - x + \beta)$
- 先验分布与后验分布属于同一个分布族 (Beta 分布)，称为共轭先验 (Conjugate prior)
- 称为 Beta-Binomial 模型

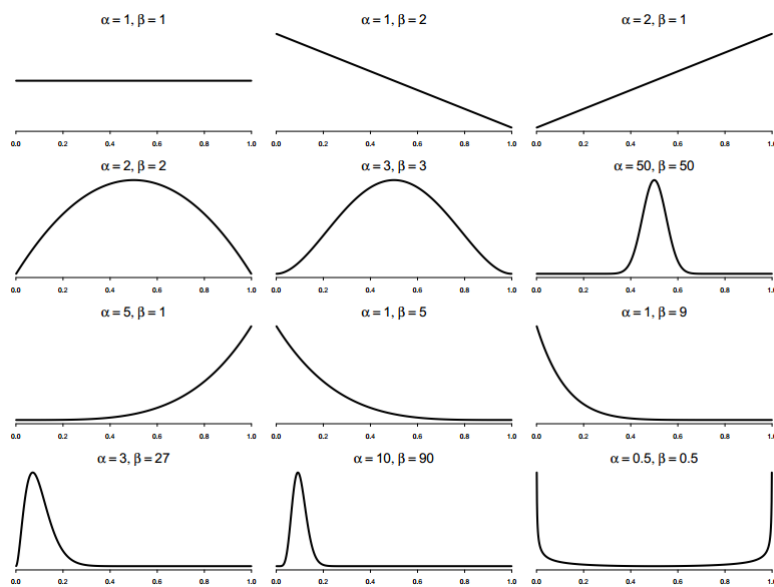
Beta 分布

定义 4. 称随机变量 X 服从 $\text{Beta}(\alpha, \beta)$ 分布, 如果其 pdf 为

$$f(\theta|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1 - \theta)^{\beta-1} (0 < \theta < 1; \alpha > 0, \beta > 0)$$

- 当 $\alpha = 1, \beta = 1$, $\text{Beta}(1, 1) = U[0, 1]$, 均匀分布
- 均值 $E(\theta) = \frac{\alpha}{\alpha + \beta}$
- 方差 $D(\theta) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$
- 众数 $\text{Mode} = \frac{\alpha - 1}{\alpha + \beta - 2}$ ($\alpha > 1, \beta > 1$)

Beta 分布密度函数



4.2 Gamma-Poisson 模型

泊松分布定义

定义 5. $X =$ 计数数据 (*Count data*, 如单位时间内事件发生次数), 如果

$$P(X = x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}, \quad x = 0, 1, \dots, \lambda > 0$$

则称 $X|\lambda \sim \text{Poisson}(\lambda)$

- 某地一年发生恐怖袭击的次数

- 某大学每位教师发表论文数
- $E(X|\lambda) = D(X|\lambda) = \lambda$
- 参数 λ 取什么先验分布?

Gamma 分布定义

定义 6. 随机变量 X 的密度函数

$$f(x|a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \quad (x > 0, a > 0, b > 0)$$

记为 $X \sim \text{Gamma}(a, b)$

- Shape: a ; rate: b 或 scale = $1/b$
- $E(X) = ab^{-1}, D(X) = ab^{-2}, \text{Mod}(X) = (a-1)/b \quad (a > 1)$
- 如果 $X \sim \text{Gamma}(n/2, 1/2)$, 则 $X \sim \chi^2(n)$
- 如果 $X \sim \text{Gamma}(1, b)$, 则 $X \sim \exp(b)$ (指数分布)

Gamma-Poisson Bayesian Model

假设 X_1, X_2, \dots, X_n 是泊松分布 $\text{Poisson}(\lambda)$ 的独立同分布样本, $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$

- 先验分布: $\lambda \sim \text{Gamma}(a, b)$
- 似然函数: $L(\lambda|\mathbf{x}) = \prod \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \propto \lambda^{\sum x_i} e^{-n\lambda}$
- 后验分布为:

$$p(\lambda|\mathbf{x}) \propto \lambda^{a-1} e^{-b\lambda} \cdot \lambda^{\sum x_i} e^{-n\lambda} \sim \text{Gamma}(a + \sum x_i, b + n)$$

– 也是共轭先验

- 后验均值:

$$E(\lambda|\mathbf{x}) = \frac{b}{b+n} \frac{a}{b} + \frac{n}{b+n} \frac{\sum x_i}{n} = wE(\lambda) + (1-w)\bar{x}$$

– 当 $n \rightarrow \infty$ 和 $n \rightarrow 0$ 时, 后验分布结果如何?

Gamma-Poisson Model: 美国大规模枪击案

2012 年 12 月, 美国康涅狄格州发生校园枪击案, 造成 28 人死亡。

资料显示, 1982 年至 2012 年, 美国共发生 62 起 (大规模) 枪击案。其中, 2012 年发生了 7 起, 是次数最多的一年。

2012 年有这么多枪击案, 正常吗? 这是巧合, 还是美国治安恶化?



1982-2012 年美国枪击案数据

一年中发生枪击案次数	年数
0	3
1	13
2	5
3	5
4	3
5	1
6	0
7	1

数据来源:<http://www.motherjones.com/politics/2012/12/mass-shootings-mother-jones-full-data>
 参考:Aatish Bhatia,2012:Are mass shootings really random events? A look at the US numbers, <http://www.wired.com/2012/12/are-mass-shootings-really-random-events-a-look-at-the-us-numbers/>

美国枪击案：MLE

- 目的：利用过去 30 年数据（不包含 2012 年），判断 2012 年是否属于正常的泊松分布
- 总体分布： $X \sim \text{Poisson}(\theta)$, θ 为平均每年枪击案发生率, 观察值 X_1, \dots, X_{30}
- 最大似然估计： $\hat{\theta} = \bar{x} = 1.83$

美国枪击案：贝叶斯模型

- 先验分布：选共轭先验 $\theta \sim \text{Gamma}(a, b)$ ，如何确定参数 a, b ? 观察过去数据，先验分布均值为 1.83，出现次数最多的年数为 1 年，因此

$$E(\theta) = a/b = 1.8; \text{Mod}(\theta) = (a - 1)/b = 1$$

得到超参数估计值: $a = 2.25, b = 1.25$

- 后验分布: $\theta|\mathbf{x} \sim \text{Gamma}(57.25, 31.25)$, 其中 $\sum x_i = 55, n = 30$
- 95%CI: $> \text{qgamma}(c(0.025, 0.975), 57.25, 31.25) = (1.388, 2.336)$
- 95%CI for noninformative prior: (1.410, 2.386)
- 无论那种先验, $X_{31} = 7$ 都远离该可信区间, 属于异常。