1 The basics

1.1 加法规则和乘法规则

$$prob(X|I) + prob(\bar{X}|I) = 1 \tag{1}$$

$$prob(X, Y|I) = prob(X|Y, I) \times prob(Y|I) \tag{2}$$

1.2 贝叶斯理论和边缘化

$$prob(X|Y, I) = \frac{prob(Y|X, I) \times prob(X|I)}{prob(Y|I)} \tag{3}$$

$$prob(X|I) = \int_{-\infty}^{+\infty} prob(X, Y|I) dY \tag{4}$$

分母上的 $P(Y|I)$ 要对所有可能的 Y 进行积分，很难计算，所以这个式子只用分子，得到没有归一化的后验 pdf。

$$prob(X|Y, I) \sim prob(Y|X, I) \times prob(X|I) \tag{5}$$

边缘化在不关心的参数存在时很有用

2 Parameter estimation

2.1 研究硬币是否公平

根据经验，对于同一个研究对象也就是同一个事实真理，不论做什么假设，在 likelihood 的修正下，得到的后验概率都是事实，不同的先验概率会收敛于同一个结果，并且在每次实验数据独立情况下，后验概率的结果与数据获取顺序无关。

2.1.1 关于后验和共轭先验 (conjugate prior)

贝叶斯的一个缺点为计算比较麻烦很难保证后验分布的具有解析解，指的是后验分布的密度函数即 $prob(H|\{data\}, I)$ 具有解析解（不过实际中即使 posterior 没有解析解也可以后验分布进行采样或近似的推断）
为什么 posterior 不一定具有解析解：

针对连续性随机变量的贝叶斯公式：

$$prob(X|Y,I)=\frac{prob(Y|X,I)\times prob(X|I)}{\int_{\Omega} prob(Y|X,I)\times prob(X|I)dX} \tag{6}$$

分母为积分形式，所以 posterior 不一定具有解析解，是否有解析解取决于积分内模型和先验的选取，但并不是选择了可以使后验获得解析解的先验就可以使后验具有解析解。实际上，正如上学期讨论的新冠病毒的粒子，更新计算中的先验概率并不是一直是同一个先验，而是在获取数据之后，用得到的 posterior 当作新的先验代入公式进行更新计算。所以还要保证更新的后验概率代入公式可以使新的后验获得解析解。一个直接的方法就是让后验和先验具有相同的表达式，这样不仅保证了 posterior 具有解析解而且还可以让每轮计算都用同一个公式。

由此定义了共轭先验：在给定模型即似然函数的情况下，如果先验分布和似然函数可以使得先验分布和后验分布有相同的形式，那么就称此先验为这个模型的的共轭先验分布。
（定义：称一个分布族为模型 $Y\sim f_{Y|X}(Y|X)\&X\in\Omega$ 的共轭先验，若只要先验分布 $f_X(X|I)$ 是从该分布族中选取的，那么最终得到的后验 $f_{Y|X}(Y|X)$ 就也属于该分布族）

2.1.2 更新 likelihood 或 pirior 的例子

新冠例子: 实际有病但测得为阳性的概率为 $P(D|\bar{H})=2.3\%$, 实际没病缺测得阴性的概率为 $P(\bar{D}|H)=1.4\%$, 患病概率 $P(H)=0.1\%$ 现在求实际测得为阳性，患病的概率 $P(H|D)=?$
已知 $P(D|H)+P(\bar{D}|H)=1$，所以 $P(D|H)=98.6\%$

$$\begin{aligned} P(H|D)&=\frac{P(D|H)P(H)}{P(D)}\\ &=\frac{P(D|H)P(H)}{P(D|\bar{H})P(\bar{H})+P(D|H)P(H)} \end{aligned} \tag{7}$$

一次检测为阳性的患病概率是很小的，所以要进行多次检测，有两种方法。比如，进行三次检测，1. 一种是一次一次的检测不断地更新先验概率密度，2. 一种是直接测三次，改变似然函数。
一次一次测，$P(H)\rightarrow P(H')$，下式中用到 $P(D|H)P(H)+P(D|\bar{H})P(\bar{H})=P(D)$,$P(D)-$

$P(D|H)P(H) = P(D|\bar{H})P(\bar{H})$

$$\begin{aligned}
P'(H|D) &= \frac{P(D|H)P(H')}{P'(D)} \\
&= \frac{P(D|H)[P(D|H)P(H)]/P(D)}{P(D|\bar{H})P'(\bar{H}) + P(D|H)P'(H)} \\
&= \frac{P(D|H)P(D|H)P(H)/P(D)}{P(D|\bar{H})(1-P'(H)) + P(D|H)P(D|H)P(H)/P(D)} \\
&= \frac{[P(D|H)]^2 P(H)}{P(D|\bar{H})[P(D)-P(D|H)P(H)] + [P(D|H)]^2 P(H)} \\
&= \frac{[P(D|H)]^2 P(H)}{P(D|\bar{H})P(D|\bar{H})P(\bar{H}) + [P(D|H)]^2 P(H)} \\
&= \frac{[P(D|H)]^2 P(H)}{P(D|\bar{H})^2 P(\bar{H}) + [P(D|H)]^2 P(H)}
\end{aligned} \tag{8}$$

测量一次与测量两次的后验概率对比，

$$P(H|D) = \frac{P(D|H)P(H)}{P(D|\bar{H})P(\bar{H}) + P(D|H)P(H)} \tag{9}$$

$$P'(H|D) = \frac{[P(D|H)]^2 P(H)}{P(D|\bar{H})^2 P(\bar{H}) + [P(D|H)]^2 P(H)} \tag{10}$$

这个化简后的结果与直接测量两次的结果一样。只是一次一次测时,先验由 $P(H) \to P(H') = P(H|D)$。直接测量两次，改变的是似然函数 $P(D|H) \to [P(D|H)]^2$

2.2 关于 pdf

2.2.1 概率密度函数与分布函数

概率密度函数 f(x) 是一根山峰线，与 x 轴围成的面积为 1，(x 轴代表组距，y 轴代表频率乘以组距，取组距趋于 0 的极限，直方图就变成一根山峰线，即概率密度函数。也可以 x 轴表示频率，y 轴代表取此频率的概率，则山峰线围的面积也代表总频率等于 1）概率密度函数对 dx 进行积分，得到面积，面积的名字叫做分布函数, 分布函数代表的面积不是整个面积，分布函数 $F(x) = \int_{-\infty}^{x} f(t)dt$。
这些是对连续型随机变量的解释, 与概率密度函数相同含义，在离散型随机变量中，这个函数叫做概率质量函数。
example:Exploring pdfs

用到的函数:
mean(): 分布的均值

median(): 分布的中值
pdf(x): 概率密度函数在 x 点的值
Rvs (size=num_pts): 生成 pdf 的 num_pts 随机值
interval(alpha): 包含 alpha 百分比的分布范围的端点

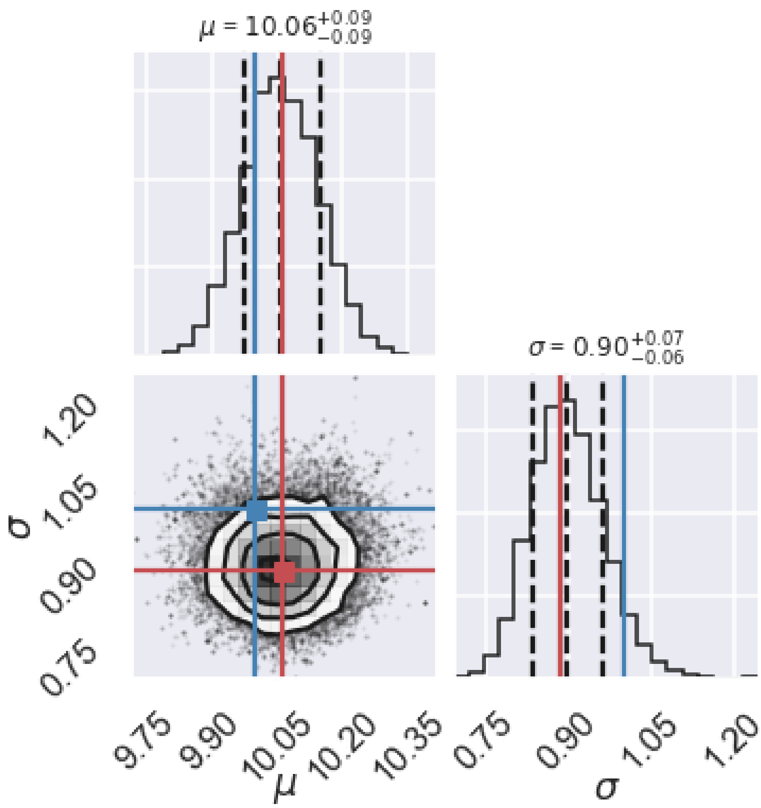

图 1:

2.2.2 常见的模型及其共轭先验

eg1:beta-伯努利共轭

伯努利分布:

取 1 的概率为 θ, 取 0 的概率为 $(1-\theta)$ 的离散型随机变量的分布。伯努利试验, 成功为 1,

失败为 0，成功的次数服从伯努利分布, 参数 θ 是试验成功的概率。其 pmf(概率质量函数) 为:

$$P(x,\theta)=\theta^{x}(1-\theta)^{1-x} \tag{11}$$

其中 $x\in\{0,1\}$

beta 分布:

beta 分布常用作先验 pirior。

beta 分布是由定义在 [0,1] 上的连续性概率分布构成的分布族，具有两个参数，$\alpha\ \beta$ ，其 pdf(概率密度函数) 为:

$$p(x;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1} \tag{12}$$

其中 B 为 beta 函数，是分布函数。先验、后验和模型都有参数，为了区分，先验和后验的参数被称为超参数，模型的参数就叫参数。

实际模型中，beta 分布常用作先验分布，当模型为二项分布或者伯努利分布时，beta 分布都是这个模型的共轭先验。

定理 (**beta-伯努利共轭**): 若 $X\rightarrow Bernouli(\Theta)$ 且 $\Theta\rightarrow Beta(\alpha,\beta)$ 则观测到 X=x 的后验分布可以选做 $Beta(x+\alpha,\beta-x+1)$。证明如下

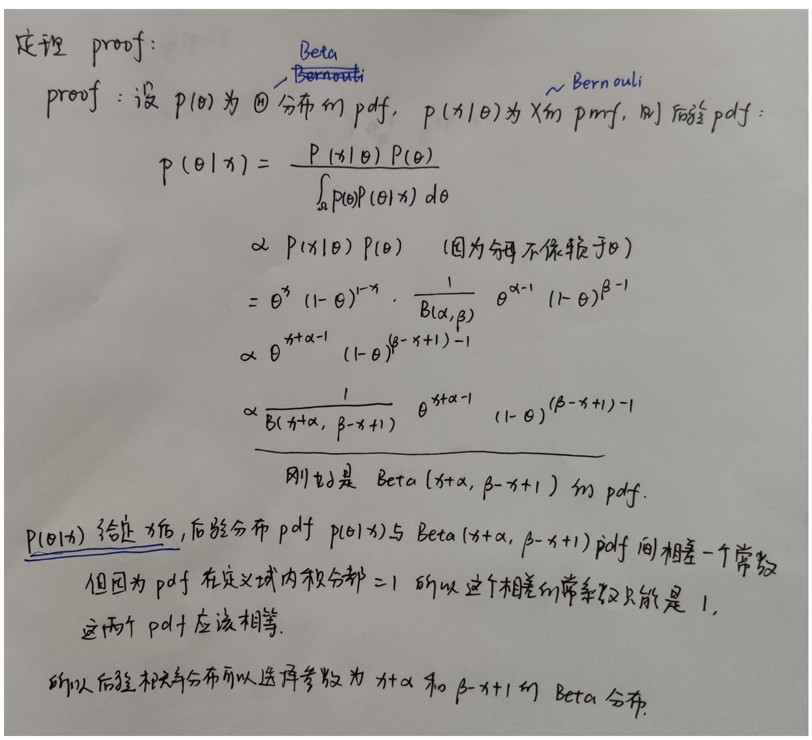

图 2:

eg2: 高斯-高斯共轭

高斯分布通常在已知均值或方差二者之一的情况下更容易找到共轭先验，因为在一个参数已知的情况下，只用关心另一个参数就行，高斯高斯共轭就是在已知模型（）likelihood 的方差的前提下的一个共轭先验。

高斯分布：

高斯分布（或正态分布）是一个具有两个参数，由连续性分布所构成的参数化分布族，其参数为均值 μ 和方差 σ^2。高斯分布的 pdf 为：

$$P(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2\sigma^2}(x-\mu^2)\right\} \tag{13}$$

高斯-高斯共轭：若 $X \sim \mathcal{N}(\mu_s, \sigma_s^2)$,$\mu_s \sim \mathcal{N}(\mu_P, \sigma_P^2)$，则观测到 X=x 的后验分布可以选做：

$$\mathcal{N}\left(\frac{\sigma_P^2}{\sigma_s^2+\sigma_P^2}x + \frac{\sigma_s^2}{\sigma_s^2+\sigma_P^2}\mu_P, (\frac{1}{\sigma_P^2}+\frac{1}{\sigma_s^2})^{-1}\right) \tag{14}$$

其中 σ_s, μ_s 是模型的参数，σ_P, μ_P 是先验的超参数。脚标 s 是信号 signal 的缩写，所以模型参数也叫信号均值和信号方差，同时 P 是 pirior 的缩写，所以超参数也被称为先验均值和先验方差。在**高斯共轭中，信号方差是已知的**。

与高斯分布对应的共轭先验有很多，高斯-高斯共轭只适用于信号方差已知，信号均值未知的情况。

对于抛硬币，由于每次数据都是独立的，likelihood（即模型）由二项分布给出，二项分布是 N 次伯努利试验。
一次伯努利试验

$$prob(\{data\}|H,I) \propto \theta^x(1-\theta)^{1-x} \tag{15}$$

因为生成的实际的实验数据，所以 x=p_h 真实值 0.4。与新冠例子中证明的一样，进行 N 次测量，一次次更新实验数据与一次更新完数据得到的测量结果是一样的。所以可以将 N 次实验后得到的 likelihood 变为

$$\begin{aligned} P(data|p_h,I) &= (p_h^x(1-p_h)^{1-x})^(N) \\ &= p_h^{xN}(1-p_h)^{N-Nx} \end{aligned} \tag{16}$$

其中 Nx=R: 头朝上次数，N-Nx=N-R：头朝下次数。这样计算与二项分布也是相同的。即

$$P(data|H,I) = H^{xN}(1-H)^{N-Nx} \tag{17}$$

先验取 beta 分布的概率密度函数：

$$P(H) = H^{\alpha-1}(1-H)^{\beta-1} \tag{18}$$

由

$$prob(H|\{data\},I) \propto prob(\{data\}|H,I) \times prob\ H|I \tag{19}$$

可以得到后验

$$P(H|data,I) = H^{\alpha+R-1}(1-H)^{\beta+N-R-1} \tag{20}$$

即参数为 $\alpha + R$ 和 $\beta + N - R$, 按照三行代码：
$y_1 = stats.beta.pdf(x, alpha_1 + heads, beta_1 + N - heads)$.... 画出三条不同 pirior 下的后验概率函数图。

实验数据 $\{data\}$ 由 $generate_data$ 定义的伯努利分布函数生成，超参数为 R，实验数据按照伯努利分布参数 $H = p_h$ 真实值 0.4 生成，产生一组 0 和 1 的数列。

在程序中对先验进行计算时，选用了共轭先验, 采用共轭先验的原因是可以使得先验分布和后验分布的形式相同，这样计算起来就较为方便。前面的证明可以看出后验分布可以选择参数为 $\alpha + N$ 和 $\beta - N + 1$ 的 Beta 分布。
随着实验数据增加，后验 pdf 的形状如下所示：

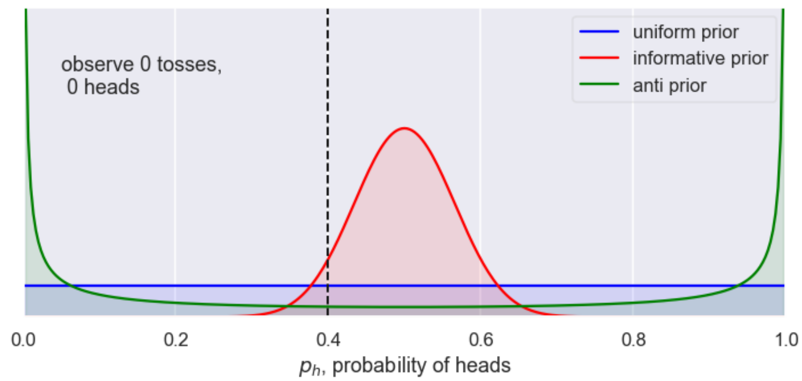

图 3: 0 tosse

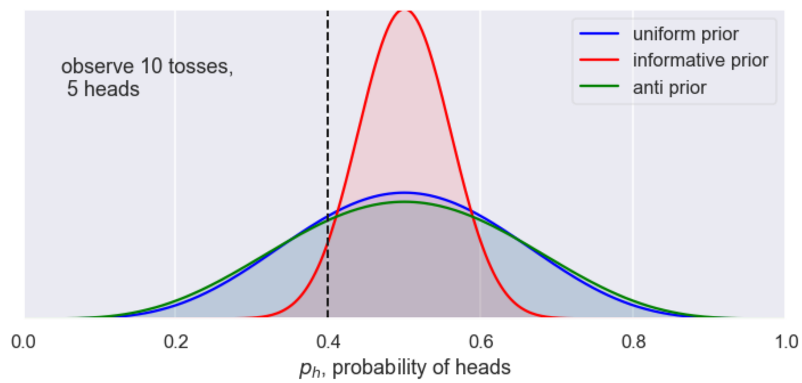

图 4: 10 tosses

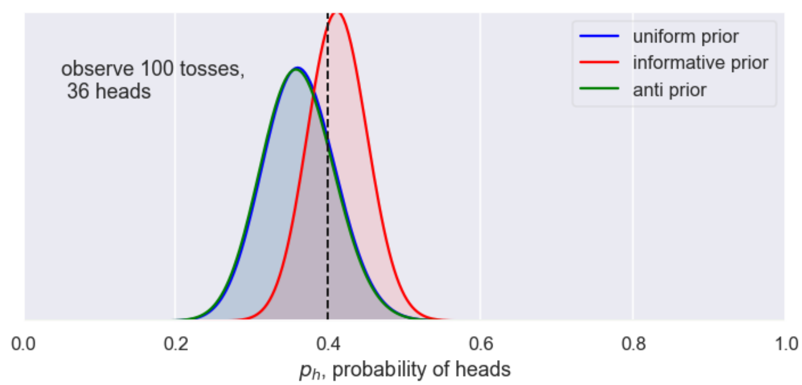

图 5: 100 tosses

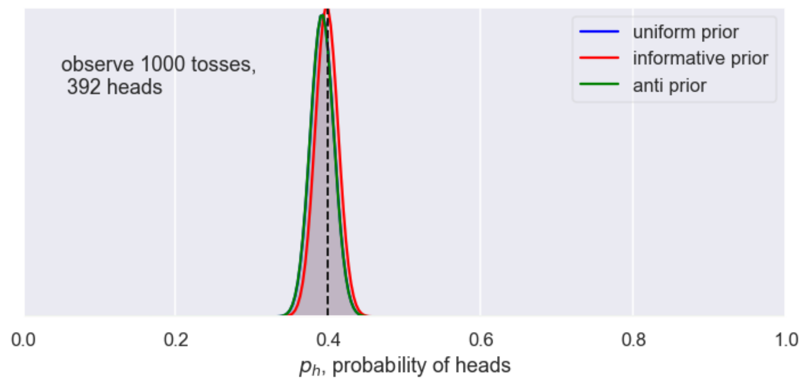

图 6: 1000 tosses

2.3 最佳估计，误差条，可信度

现在已经知道后验 pdf 如何在给定数据和相关背景信息的情况下，对关于参数值的推断进行概率计算。接下来还需要用两个参数总结这些，即 best estimate 最佳估计和 confidence 对结果可靠性的衡量。

最佳估计由后验 pdf 的最大值给出，如果用 X 表示感兴趣的参数，后验表示为 $P = prob(X|data, I)$, 最佳估计 X_0,

$$\left.\frac{dP}{dX}\right|_{X_0} = 0 \tag{21}$$

严格来说，应该检查二阶导数小于零以确保 X_0 代表最大值。

由于进行了微分操作，所以假设了 X 是一个连续参数。如果参数 X 只能取离散值，best estimate 依然是最大后验概率的估计，但是不能用上面的微分表达式。? 那么应该怎么计算呢?

为了获得这一最佳估计的可靠性的度量，需要观察的 pdf 在 X_0 附近的宽度或分布。要考虑函数在特定点附近的行为时，泰勒展开会比较有帮助，泰勒展开是一个简单且标准的工具，用于用低阶多项式逼近一个复杂的函数。

计算 best estimate 时，很难找到解析解，所以可以对后验 pdf 取对数（1. 避免计算机精度引起的误差。2. 可以把乘法化为加减法简化计算减少计算周期。）

$$L = \log_e[prob(X|data, I], \tag{22}$$

将 L 在 X_0 处展开，

$$L = L(X_0) + \frac{1}{2}\left.\frac{d^2P}{dX^2}\right|_{X_0}(X - X_0)^2 + ..., \tag{23}$$

这里的 $L(X_0)$ 是一个常数，并且一阶导等于 0 了，所以二次项是决定后验 pdf 宽度的主导因

素，在可靠性分析中担任中心角色。忽略所有的高阶项，

$$prob(X|data,I) \approx A\exp[\frac{1}{2}\left.\frac{d^2P}{dX^2}\right|_{X_0}(X-X_0)^2] \tag{24}$$

A 是归一化常数。这么做的目的是用简单的高斯分布（正态分布）来近似后验 pdf，

$$prob(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}exp[-\frac{(x-\mu)^2}{2\sigma^2}], \tag{25}$$

这个式子与后验 pdf 近似式相比可以发现，后验 pdf 是最大值在 $\mu = X_0$ 处，由参数 σ 描述的高斯分布，

$$\sigma = \frac{d^2P}{dX^2}|_{X_0}(X-X_0)^2)^{-\frac{1}{2}}, \tag{26}$$

由高斯积分性质可以看出，X 的真实值落在 $X_0-\sigma$ 到 $X_0+\sigma$ 范围内的概率为 67%：

$$prob(X_0-\sigma \le X \le X_0+\sigma|data,I) = \int_{X_0-\sigma}^{X_0+\sigma} prob(X,data|I)dX \approx 0.67 \tag{27}$$

$$Bestestimate : X_0, Reliability : \sigma \tag{28}$$

参数 σ 也被叫做误差条 error-bar

2.3.1 硬币例子

在硬币例子中 $prob(H|\{data\},I) \propto H^R(1-H)^{N-R}$，对其取对数，并计算

$$\frac{R}{H}\frac{(N-R)}{1-H}, \frac{d^2P}{dX^2} = -\frac{R}{H^2} - \frac{(N-R)}{1-H}^2, \tag{29}$$

根据

$$\left.\frac{dP}{dX}\right|_{H_0} = 0, \tag{30}$$

推出最佳估计 $bestesitimate : H_0 = \frac{R}{N}$, 所以

$$\frac{d^2P}{dX^2}|_{H_0} = -\frac{N}{H_0(1-H_0)}, \tag{31}$$

因此，

$$error - bar : \sigma = \sqrt{\frac{H_0(1-H_0)}{N}}. \tag{32}$$

H_0 在经过一定数量的数据分析后变化不大，分子趋于一个定值，因此后验的宽度与数据

量的平方根成反比，并且可以看出证明一个硬币不公平比证明他是公平的更简单。

2.3.2 非对称后验 pdfs

在之前的图中可以注意到峰并不是一直对称的，随着数据量的增加，后验 pdf 的形状会越来越像高斯分布，在数据量不够时，error-barσ 的使用就有一些不足，因为误差条隐含了对称的信息，pdf 不对称时，后验最大值 X_0 依然表示最佳估计, 但真实值可能在这个峰的旁边。

一个很好的办法是通过置信区间来推断参数的可靠性，若 pdf 已经归一化，考虑 95% 置信区间。

$$prob(X_1 \leq X \leq X_2|\{data\}, I) = \int_{X_1}^{X_2} prob(X|\{data\}, I)dX \propto 0.95. \tag{33}$$

其中 X1，X2 的差值越小越好。同时可以考虑均值 mean 和期望值 expection，他们考虑到了 pdf 的不对称性、偏度。归一化的 pdf 的加权平均

$$< X >= \int prob(X|\{data\}, I)XdX, \tag{34}$$

(X 只能取离散值时积分用求和代替), 如果后验 pdf 没有归一化, 那么右边必须除以一个归一化系数 $\int prob(X|\{data\}, I)dX$, 如果是高斯分布, 则均值与最佳估计 X_0 刚好相等 ($X_0 =< X >$)。

2.3.3 多模态后验

如果后验 pdf 一个极大值比其他的都大时，可以简单的忽略附属解，但是如果是几个规模相当的极大值那么 best estimate 是不能做出解释的。因为后验 pdf 给出了完整地描述，可以根据数据和相关的先验知识推断出参数的值。但是企图用最佳估计、误差条、置信区间两三个数字总结后验 pdf 有的时候是没办法做到的。不过后验 pdf 是存在的，可以从中得到适当的结论。

如图

图 7:

忽略 $X = 20$ 右边的结构，则这个后验传递的 $X = -10$ 或 $+10$，可以写成 $X = -10 \pm 2$ 或 $X = 10 \pm 2$，然而 pdf 的均值仍然是唯一的，所以有时考虑用均值表示最佳估计，但是在图中的 pdf 上，忽略右边的结构后，期望值 $< X >= 0$，在后验 pdf 图画中显示这个取值是很不可能的，即使这样也依然使用均值表示最佳估计的话，需要对其分配一个很大的 error bar 来让置信区间内包含这个值，因此也不能很好的反应后验 pdf 中固有的信息。对于双峰 pdf 可以用几个数据来描述后验 pdf：两个最佳估计，及两个最佳估计分别相关的 error bar，或者不相

交的置信区间。一般对于多模态，我们能做的只是诚实的显示后验 pdf 本身。

2.4 高斯噪声和平均值

2.4.1 bayesTALENT into

程序中比较了频率派和贝叶斯方法程序是 Sivia 的书中第 2.3 节 Example2 的扩展。考虑正态分布的均值和方差的估计问题。正态分布

$$p(x_k|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{35}$$

通常用于作为理论模型来描述与实验数据相关的噪声。假设有一系列 M 测量值 $D \equiv X_k = (x_1 \ldots x_M)$, 样本服从一个正态分布 $N(\mu,\sigma^2)$, 现在想知道的是参数 μ 和 σ。频率论的方法：最大似然方法, 贝叶斯方法：计算模型参数的后验分布 u 和 o。这里，bayesTALENT into 中使用 Python 生成一些数据，演示了解决该问题的两种方法。

首先使用按照真实值生成满足高斯分布的数据，然后画了数据的散点图和条形图。高斯参数估计的频率派：

从经典的频率最大似然法开始，一次测量的概率 D_i 的值为 x_i 的概率是由 $p(x_i \mid \mu,\sigma)$ 给出的

$$p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(x_i-\mu)^2}{2\sigma^2}\right], \tag{36}$$

这里的 μ 和 σ 都是模型 (likelihood) 参数的真实值，通过计算每个数据点的概率乘积来构造似然函数:

$$\mathcal{L}(D|\mu,\sigma) = \prod_{i=1}^{M} p(x_i\ |\mu,\sigma), \tag{37}$$

目的是找到 μ_0,σ_0，使 likelihood(或对数 likelihood) 最大化。对于这个简单的问题，最大化可以用解析的方法计算, 即通过 $\left(\frac{\partial\log\mathcal{L}}{\partial\mu}\right)|_{\mu_0\sigma_0} = \left(\frac{\partial\log\mathcal{L}}{\partial\sigma}\right)|_{\mu_0\sigma_0} = 0$ 这将导致以下真实参数的最大似然估计:

$$\mu_0 = \frac{1}{M}\sum_{i=1}^{M} x_i, \tag{38}$$

$$\sigma_0^2 = \frac{1}{M}\sum_{i=1}^{M}(x_i-\mu_0)^2, \tag{39}$$

程序用这两个表达式计算出了 μ_0 和 σ_0，与真实值非常接近。10，10.06；1，0.89 (基于 100 次实验)。

Bayes 方法中，首先定义了先验后验和似然函数，取先验为平坦的 1，则后验

$$p(\mu,\sigma|D,I) \propto \mathcal{L}(D|\mu,\sigma), \tag{40}$$

使用蒙特卡洛抽样计算，还在学，还没学明白。根据计算出的后验绘图:

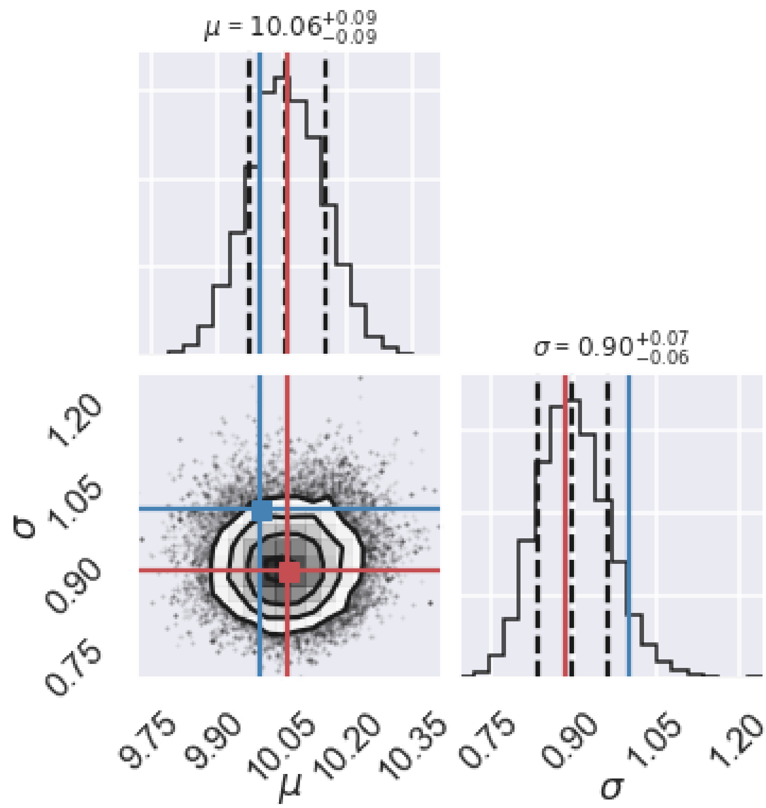

图 8:

左下角的图显示了两个模型参数的联合概率分布。对角线上的两个图显示了两个模型参数的边缘分布 (其中一个参数被积掉)。蓝线表示模型参数的真实值，标题和虚线表示最可能的值和 68% 的置信区间。频率派最大可能性估计由红线表示。对于这个特定的设置，最大似然估计与来自贝叶斯后验分布的 MCMC 抽样的最大似然值一致。这个问题中给出了最佳估计和 error bar 的值，但是贝叶斯的结果将是报告后验分布本身。

2.4.2 课本上的内容

$$p(x_k|\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

第 k 个数据的值为 x_k 的概率由该式给出,μ 是感兴趣的参数的真实值，σ 是实验中对测量误差的测量值。给出一组数据 x_k，μ 的最佳估计是什么？对于这一预测的 confident 是多大？

在这个例子中，σ 的值是已知的，因此对 μ 值的推断通过后验 pdf$prob(\mu|x_k,\sigma,I)$ 来计算。

$$prob(\mu|x_k,\sigma,I)\propto prob(x_k|\mu,\sigma,I)\times prob(\mu|\sigma,I), \tag{41}$$

如果假设数据之间是独立的，那么

$$prob(\mu|x_k,\sigma,I)=\prod_{k=1}^{N}prob(x_k|\mu,\sigma,I), \tag{42}$$

由于高斯峰的宽度没有关于中心值的信息，并选取一个平坦的先验

$$prob(\mu|\sigma,I)=prob(\mu|I)=\begin{cases}A & \mu_{min}\le\mu\le\mu_{max},\\ 0 & otherwise,\end{cases} \tag{43}$$

后验 pdf 取对数后

$$L=log_e[prob(\mu|x_k,\sigma,I)]=C(Nln\frac{1}{\sigma\sqrt{2\pi}}+lnA)+\sum_{k=1}^{N}\frac{(x_k-\mu)^2}{2\sigma^2}, \tag{44}$$

这里的常数项与 μ 无关，后验 pdf 在 $\mu_{min}\le\mu\le\mu_{max}$ 范围外等于 0。