1 The basics

1.1 加法规则和乘法规则

$$prob(X|I) + prob(\bar{X}|I) = 1 \tag{1.1}$$

$$prob(X, Y|I) = prob(X|Y, I) \times prob(Y|I) \tag{1.2}$$

成立条件?

1.2 贝叶斯理论和边缘化

$$prob(X|Y, I) = \frac{prob(Y|X, I) \times prob(X|I)}{prob(Y|I)} \tag{1.3}$$

$$prob(X|I) = \int_{-\infty}^{+\infty} prob(X, Y|I)dY \tag{1.4}$$

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。

$$prob(X|Y, I) \sim prob(Y|X, I) \times prob(X|I) \tag{1.5}$$

边缘化在不关心的参数存在时很有用

1.3 关于 pdf

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。

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 (置信区间)

几种常见 pdf 的形状:

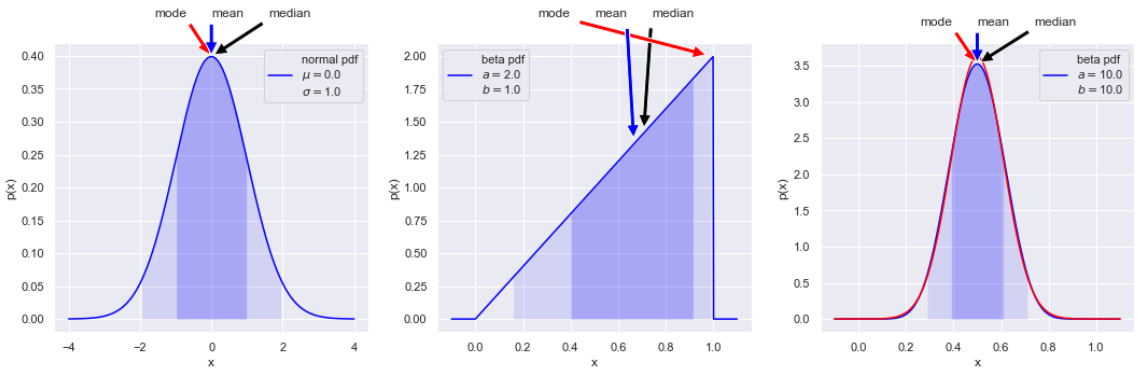

图 1.1: 从左至右分别为正态分布 ($\mu = 0.0, \sigma = 1.0$)、beta 分布 ($\alpha = 2.0, \beta = 1.0$)，第三个图中一起画了 beta 分布 ($\alpha = \beta = 10.0$) 和正态分布 ($\mu = 0.5, \sigma = 0.11$)

在图中用红色箭头指出了众数，蓝色箭头指出平均值，黑色箭头指出中位数。

intro 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

$$p(x_k|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{1.6}$$

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。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

$$p(x_i|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(x_i-\mu)^2}{2\sigma^2}\right], \tag{1.7}$$

通过计算每个数据点的概率乘积来构造似然函数:

$$\mathcal{L}(D|\mu,\sigma)=\prod_{i=1}^{M}p(x_i\ |\mu,\sigma), \tag{1.8}$$

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:

$$\mu_0=\frac{1}{M}\sum_{i=1}^{M}x_i, \tag{1.9}$$

$$\sigma_0^2=\frac{1}{M}\sum_{i=1}^{M}(x_i-\mu_0)^2, \tag{1.10}$$

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

$$p(\mu,\sigma|D,I)\propto\mathcal{L}(D|\mu,\sigma), \tag{1.11}$$

由此画出后验分布的 corner 图，并用 MCMC 进行抽样取点 (MCMC 方法是用来在概率空间，通过随机采样估算兴趣参数的后验分布。)，其中蓝线表示真实值，红线代表频率派最大似然法算出的值，对角线上的图是边缘化后的图（将另一个参数积分掉），左下角的图显示了两个模型参数的联合概率分布，线圈代表等置信区间，在圈上的点的 σ 和 μ 的取值所处的置信区间相等。图中的点就是使用 MCMC 方法进行抽样选出的点，中间最密集的地方所代表的值与频率派最大似然法得出的结果是一致的。数据虽然是按照高斯分布生成的，但是却是随机取样，所以置信区间左右不一定是对称的，数据量不够多的时候会有小的摇摆。但是这个偏差，跟下面非对称的 beta 分布对比可以看出 β 分布的上下误差的绝对值相差了 0.02, 而这里是 0 和 0.01，再下面一个图差的更多。

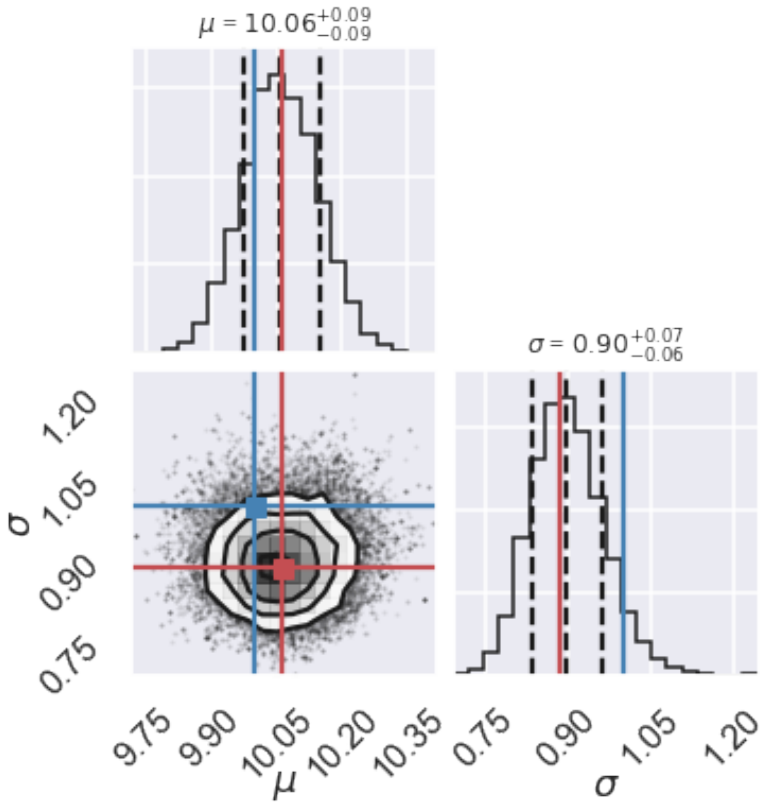

图 1.2:

除此之外，在 exploring pdf 中，还展示了其他几种 pdf 的分布 corner 图，非对称 beta 分布：

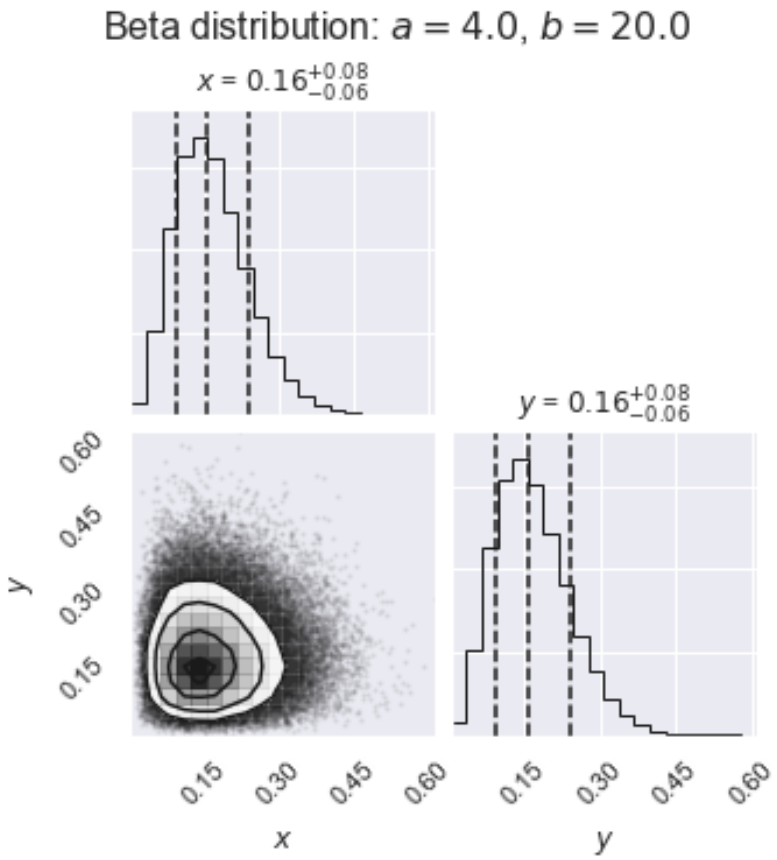

图 1.3: $\alpha = 4, \beta = 20$

多参数的 corner 图:

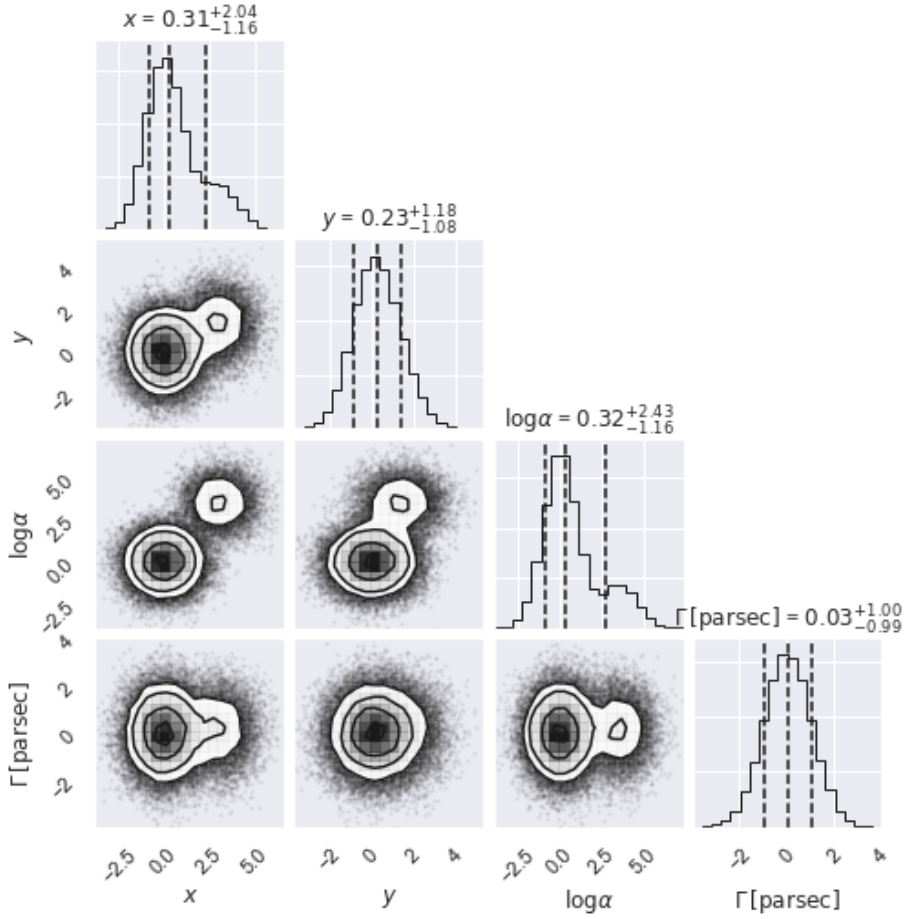

图 1.4: 显示多个参数之间关系的 corner 图

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：

$$prob(X|Y,I) = \frac{prob(Y|X,I) \times prob(X|I)}{\int_\Omega prob(Y|X,I) \times prob(X|I) dX} \tag{2.1}$$

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 $P(D|H) = 98.6\%$

$$\begin{aligned} P(H|D) &= \frac{P(D|H)P(H)}{P(D)} \\ &= \frac{P(D|H)P(H)}{P(D|\bar{H})P(\bar{H}) + P(D|H)P(H)} \end{aligned} \tag{2.2}$$

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，$P(H) \to P(H')$, 下式中用到 $P(D|H)P(H) + P(D|\bar{H})P(\bar{H}) = P(D)$,$P(D) -$

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

$$\begin{aligned} P'(H|D) &= \frac{P(D|H)P(H')}{P'(D)} \\ &= \frac{P(D|H)[P(D|H)P(H)]/P(D)}{P(D|\bar{H})P'(\bar{H}) + P(D|H)P'(H)} \\ &= \frac{P(D|H)P(D|H)P(H)/P(D)}{P(D|\bar{H})(1-P'(H)) + P(D|H)P(D|H)P(H)/P(D)} \\ &= \frac{[P(D|H)]^2P(H)}{P(D|\bar{H})[P(D) - P(D|H)P(H)] + [P(D|H)]^2P(H)} \\ &= \frac{[P(D|H)]^2P(H)}{P(D|\bar{H})P(D|\bar{H})P(\bar{H}) + [P(D|H)]^2P(H)} \\ &= \frac{[P(D|H)]^2P(H)}{P(D|\bar{H})^2P(\bar{H}) + [P(D|H)]^2P(H)} \end{aligned} \tag{2.3}$$

测量一次与测量两次的后验概率对比,

$$P(H|D) = \frac{P(D|H)P(H)}{P(D|\bar{H})P(\bar{H}) + P(D|H)P(H)} \tag{2.4}$$

$$P'(H|D) = \frac{[P(D|H)]^2P(H)}{P(D|\bar{H})^2P(\bar{H}) + [P(D|H)]^2P(H)} \tag{2.5}$$

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 $P(D|H) \to [P(D|H)]^2$

2.1.3 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:

$$P(x, \theta) = \theta^x(1-\theta)^{1-x} \tag{2.6}$$

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，α β ，其

pdf(概率密度函数) 为:

$$p(x;\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \tag{2.7}$$

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 $X \to Bernouli(\Theta)$ 且 $\Theta \to Beta(\alpha,\beta)$ 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

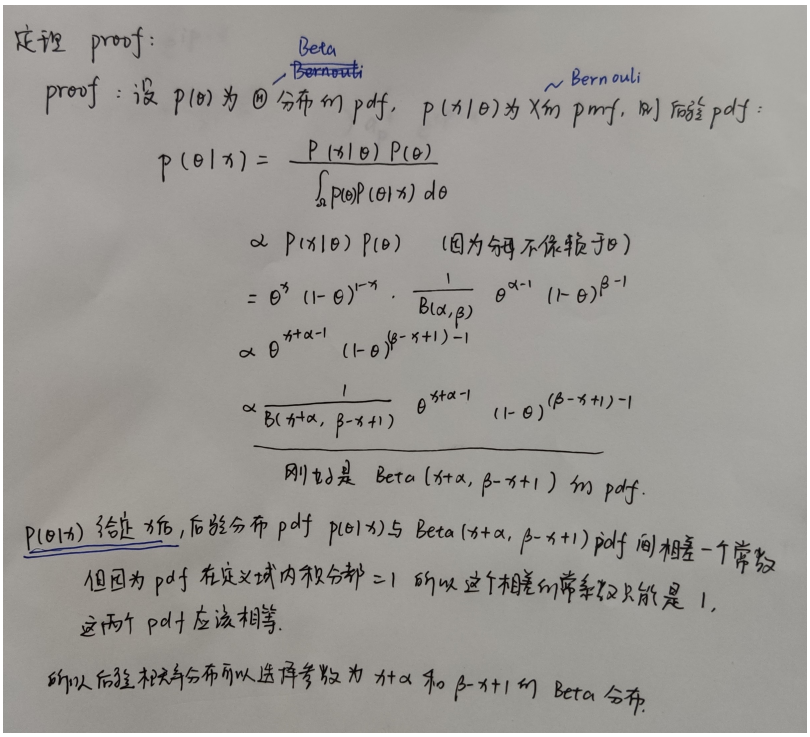

图 2.1:

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：

$$P(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2\sigma^2}(x-\mu^2)\right\} \tag{2.8}$$

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：

$$\mathcal{N}\left(\frac{\sigma_P^2}{\sigma_s^2+\sigma_P^2}x + \frac{\sigma_s^2}{\sigma_s^2+\sigma_P^2}\mu_P, (\frac{1}{\sigma_P^2}+\frac{1}{\sigma_s^2})^{-1}\right) \tag{2.9}$$

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

$$prob(\{data\}|H,I) \propto \theta^x(1-\theta)^{1-x} \tag{2.10}$$

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

$$\begin{aligned} P(data|p_h,I) &= (p_h^x(1-p_h)^{1-x})^N \\ &= p_h^{xN}(1-p_h)^{N-Nx} \end{aligned} \tag{2.11}$$

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

$$P(data|H,I) = H^{xN}(1-H)^{N-Nx} \tag{2.12}$$

先验取 beta 分布的概率密度函数：

$$P(H) = H^{\alpha-1}(1-H)^{\beta-1} \tag{2.13}$$

由

$$prob(H|\{data\},I) \propto prob(\{data\}|H,I) \times prob\ H|I \tag{2.14}$$

可以得到后验

$$P(H|data,I) = H^{\alpha+R-1}(1-H)^{\beta+N-R-1} \tag{2.15}$$

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：
蓝色线 $\alpha=\beta=1$，红色线 $\alpha=\beta=30$, 绿色线 $\alpha=\beta=0.2$

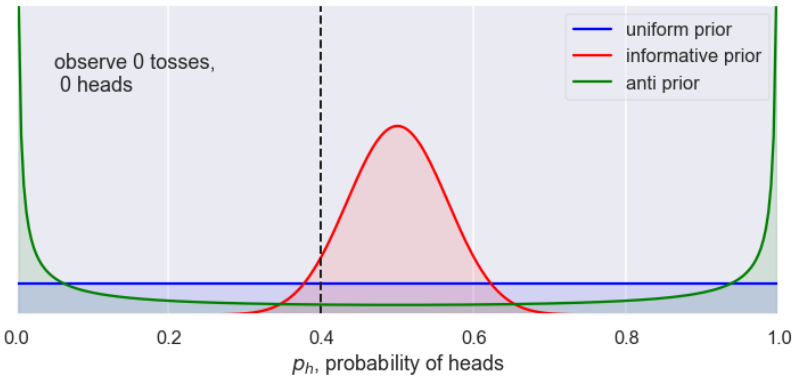

图 2.2: 0 tosse

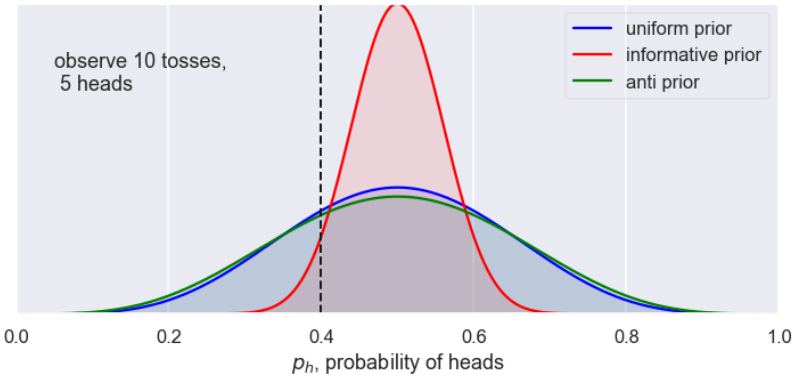

图 2.3: 10 tosses

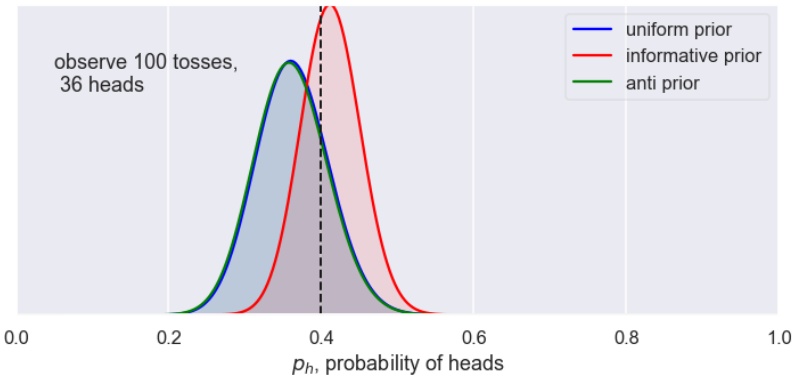

图 2.4: 100 tosses

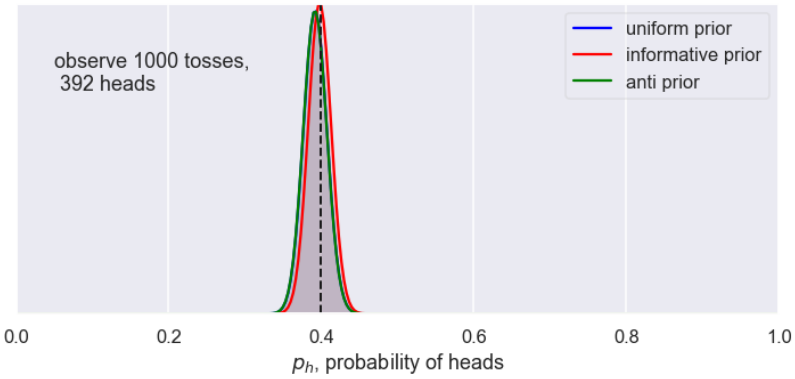

图 2.5: 1000 tosses

并且在程序中单独对平坦先验的结果进行了分析，返回的置信区间中，68% 不包含真实值，但是 98% 置信区间包含真实值。

2.2 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,

$$\left.\frac{dP}{dX}\right|_{X_0} = 0 \tag{2.16}$$

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。）

$$L = \log_e[prob(X|data, I], \tag{2.17}$$

将 L 在 X_0 处展开，

$$L = L(X_0) + \frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_0}(X - X_0)^2 + ..., \tag{2.18}$$

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，

$$prob(X|data, I) \approx A\exp[\frac{1}{2}\left.\frac{d^2L}{dX^2}\right|_{X_0}(X - X_0)^2]. \tag{2.19}$$

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，

$$prob(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}exp[-\frac{(x-\mu)^2}{2\sigma^2}], \tag{2.20}$$

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，

$$\sigma = \left(-\left.\frac{d^2L}{dX^2}\right|_{X_0}\right)^{-\frac{1}{2}}, \tag{2.21}$$

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%：

$$prob(X_0 - \sigma \le X \le X_0 + \sigma|data, I) = \int_{X_0-\sigma}^{X_0+\sigma} prob(X, data|I)dX \approx 0.67, \tag{2.22}$$

$$Bestestimate : X_0, Reliability : \sigma$$

参数 σ 也被叫做误差条 error-bar

2.2.1 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

$$\frac{dL}{dX} = \frac{R}{H}\frac{(N-R)}{1-H}, \quad \frac{d^2L}{dX^2} = -\frac{R}{H^2} - \frac{(N-R)}{1-H}^2, \tag{2.23}$$

根据

$$\left.\frac{dL}{dX}\right|_{H_0} = 0, \tag{2.24}$$

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

$$\left.\frac{d^2L}{dX^2}\right|_{H_0} = -\frac{N}{H_0(1-H_0)}, \tag{2.25}$$

因此，

$$error - bar : \sigma = \sqrt{\frac{H_0(1-H_0)}{N}}. \tag{2.26}$$

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。

2.2.2 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。

$$prob(X_1 \leq X \leq X_2|\{data\}, I) = \int_{X_1}^{X_2} prob(X|\{data\}, I)dX \propto 0.95. \tag{2.27}$$

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

$$< X >= \int prob(X|\{data\}, I)XdX, \tag{2.28}$$

(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 ($X_0 =< X >$)。

2.2.3 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

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 $X = -10 \pm 2$

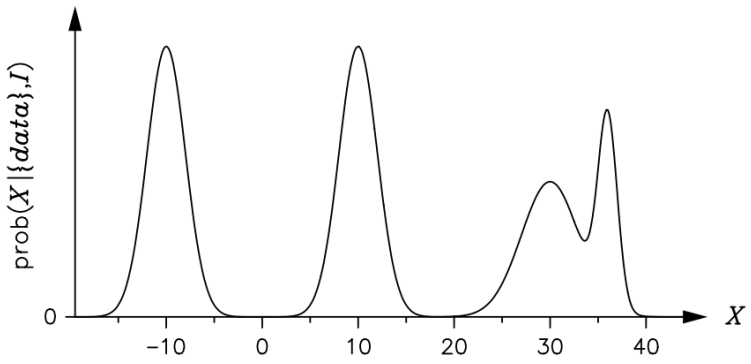

图 2.6:

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。

2.3 高斯噪声和平均值

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

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，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。

$$prob(\mu|x_k,\sigma,I) \propto prob(x_k|\mu,\sigma,I) \times prob(\mu|\sigma,I), \tag{2.29}$$

如果假设数据之间是独立的，那么

$$prob(\mu|\{x_k\},\sigma,I) = \prod_{k=1}^{N} prob(x_k|\mu,\sigma,I), \tag{2.30}$$

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

$$prob(\mu|\sigma, I) = prob(\mu|I) = \begin{cases} A & \mu_{min} \leq \mu \leq \mu_{max}, \\ 0 & otherwise, \end{cases} \tag{2.31}$$

后验 pdf 取对数后

$$L = log_e[prob(\mu|x_k, \sigma, I)] = C(Nln\frac{1}{\sigma\sqrt{2\pi}} + lnA) + \sum_{k=1}^{N} \frac{(x_k - \mu)^2}{2\sigma^2}, \tag{2.32}$$

这里的常数项与 μ 无关，后验 pdf 在 $\mu_{min} \leq \mu \leq \mu_{max}$ 范围外等于 0。为了找到最佳估计 μ_0 对 L 求一阶导等于 0，

$$\left.\frac{dL}{d\mu}\right|_{\mu_0} = \sum_{k=1}^{N} \frac{x_k - \mu_0}{\sigma^2} = 0$$

其中 σ 是与 μ 无关的常数，所以可以提到求和外，因此

$$\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} \mu_0 = N\mu_0$$

μ 的最佳估计由 xk 的数值平均 arithmetic average 决定，

$$\mu_0 = \frac{1}{N} \sum_{k=1}^{N} x_k \tag{2.33}$$

由 L 的二阶导数可以求出 σ

$$\left.\frac{d^2L}{d\mu^2}\right|_{\mu_0} = \sum_{k=1}^{N} \frac{1}{\sigma^2} = -\frac{N}{\sigma^2} \tag{2.34}$$

根据公式 (32), 误差条应该由上式的倒数的负的平方根给出，所以

$$\sigma = \sqrt{\frac{\sigma^2}{N}} = \frac{\sigma}{\sqrt{N}}$$

总结对于 μ 值的推断

$$\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}} \tag{2.35}$$

与投硬币实验一样，得到了相似的结果，估计的可靠性与实验数据的数量的平方根成正比。

在前面注意到，误差条的概念依赖于方程的二次展开的有效性，之前忽略了高阶展开，仅仅保留了泰勒展开的二次项，并把它写成了高斯的形式。而在这个高斯噪声的情况，这不是一

个近似形式而是精确的展开式因为高阶导数全为 0。因此后验 pdf 完全可以由误差条的定义来描述函数行为。

唯一的条件是 μ_{min} 和 μ_{max} 的范围。原则上可以通过让 min 和 max 趋于无穷来象征对于先验的无知，但是，如果这个范围大到了一定程度，那么他俩的值对于后验 pd 是没有影响的。如果最佳估计和误差条允许的 μ 值超出了 μ_{min} 和 μ_{max} 这个范围，那么只能去显示后验 pdf 本身及其截断？这种情况告诉我们先验知识也一样重要？

2.3.1 有不同大小 error bar 的数据

在之前的分析中，都是假设对于每个数据，误差条大小都是一样的，如果所有测量都是在一样的实验装置下进行，那么这样是合理的，但是如果实验数据来自不同精密程度的几个实验室获得的，那么应该如何结合不同实验室的数据呢？假设测量误差依然可以通过高斯 pdf 进行建模，也就是 likelihood 依然满足 guass 分布，因此第 k 个数据的值为 xk 的概率分布为

$$prob(x_k|\mu,\sigma_k,I) = \frac{1}{\sigma_k\sqrt{2\pi}} exp[-\frac{(x_k-\mu)}{2\sigma_k^2}]. \tag{2.36}$$

跟之前一样的方法

$$prob(\{x_k\}|\mu,\sigma_k,I) = \prod_{k=1}^{N} prob(x_k|\mu,\sigma_k,I), \tag{2.37}$$

$$L = Constant - \sum_{k=1}^{N} \frac{(x_k-\mu)^2}{2\sigma_k^2}, \tag{2.38}$$

同理，一阶导数等于 0，得到

$$\mu_0 = \frac{\sum_{k=1}^{N}\omega_k x_k}{\sum_{k=1}^{N}\omega_k}, \tag{2.39}$$

其中 $\omega_k = \frac{1}{\sigma_k^2}$ **这里计算 best estimate 使用加权平均 weighted average 而不是算术平均 arithmetic mean，这样不可靠的数据会对应的更大的 error bar，和相应更低的权值**。L 的二阶导数产生最佳估计的误差条，关于 μ，

$$\mu = \mu_0 \pm (\sum_{k=1}^{N}\omega_k)^{-1/2}. \tag{2.40}$$

Note that if all the data were of comparable quality,so that sigmak=sigma,….?

2.4 example3:lighthouse

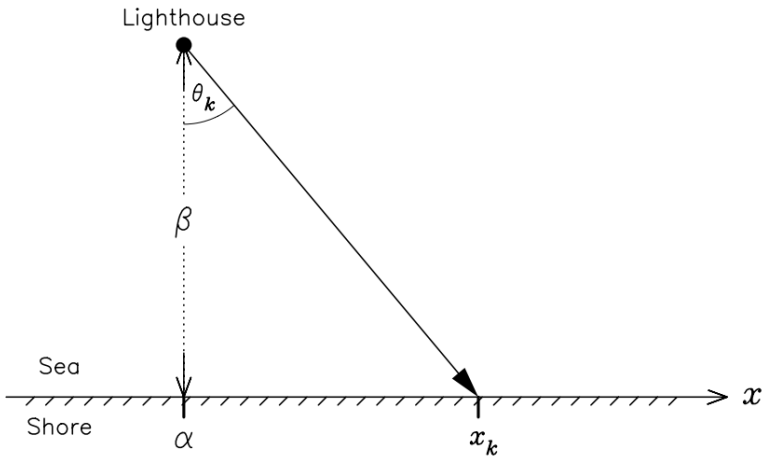

图 2.7:

灯塔的位置对应海岸线上的 α，离海岸的距离为 β，灯塔随机间隔、随即方位角发射一系列闪光照到岸上，岸上记录了 N 次数据x_k, 求灯塔在哪里。考虑灯塔发射的性质，为第 k 个数据的方位角 θ_k 分配一个均匀的 pdf 是合理的，

$$prob(\theta_k|\alpha,\beta,I) = \frac{1}{\pi}, \tag{2.41}$$

θ 的取值范围在 $\pm\pi/2$ 之间。将 θ_k 和 x_k 联系起来

$$\beta tan\theta_k = x_k - \alpha, \tag{2.42}$$

3.6 节可以看到，在处理 changing variables 时，可以用下面的式子重写 (54) 式。

$$prob(x_k|\alpha,\beta,I) = \frac{\beta}{\pi[\beta^2 + (x_k-\alpha)^2]}. \tag{2.43}$$

由此，已知灯塔 (α,β) 的坐标，第 k 次闪光记录在位置 x_k 的概率由柯西分布给出。

这种 pdf 的函数形式在物理学中经常遇到，通常称为洛伦兹函数。它关于最大值 $x_k = \alpha$ 对称，其 FWHM 为 2β(FWHM 是半峰全宽，峰一半高处的峰宽度); 分布如图 2.8 所示。

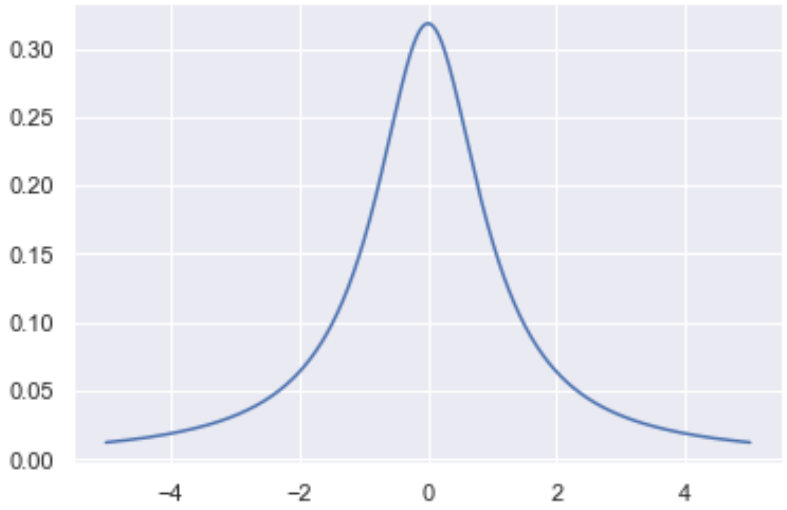

图 2.8:

柯西分布是一个数学期望不存在的连续型概率分布。当随机变量 X 满足它的概率密度函数时，称 X 服从柯西分布。

从数据中推断灯塔位置，需要同时对 α 和 β 进行估计; 这章讨论单参数问题，所以假定灯塔到海面的距离 β 已知，并将其简化为一个单一参数的问题。对于灯塔位置的 inference 由后验 pdf$prob(\alpha|x_k,\beta,I)$ 表示。作为已知数据，β 没有提供任何关于 α 的信息，所以可以将先验取为平坦的:

$$prob(\alpha|\beta,I) = prob(\alpha|I) = \begin{cases} A & \alpha_{min} \le \alpha \le \alpha_{max}, \\ 0 & otherwise, \end{cases} \tag{2.44}$$

α_{min} 和 α_{max} 可以代表海岸线的边界，这些独立数据的概率函数是获得 N 个单独探测的概率的乘积:

$$prob(\{x_k\}|\alpha,\beta,I) = \prod_{k=1}^{N} prob(x_k|\alpha,\beta,I), \tag{2.45}$$

跟之前一样，将由 (55) 得到的 prior，和由 (54)(56) 得到的 likelihood 代入贝叶斯公式，得到 posterior 并对它取对数，得到

$$L = log_e[prob(\alpha|x_k,\beta,I)] = constant - \sum_{k=1}^{N} log_e[\beta^2 + (x_k - \alpha)^2], \tag{2.46}$$

其中常数不包含 α。假设先验的范围无限大，也就是海岸线无限长，这样就不用担心后验函数

的截断。位置的最佳估计由后验函数最大值 α_0 给出，由此

$$\left.\frac{dL}{d\alpha}\right|_{\alpha_0} = 2\sum_{k=1}^{N}\frac{x_k-\alpha_0}{\beta^2+(x_k-\alpha_0)^2} = 0. \tag{2.47}$$

这个方程很难重新排列，所以用 β 和 x_k 来表示 α_0。虽然解析解不好搞，但是可以**用数值法解**这个式子: 从 (57) 式，为 α 的一系列不同的可能值计算 L。得到最大 L 的 α 就是最佳估计。如果我们在纵轴上画出 L 的指数 exp(L)，横轴为 α，就得到了灯塔位置的后验 pdf，提供了推断的图像，优点是不需要担心后验 pdf 是否是不对称的或多模态的。

对于 (57)，倾向于从总和中忽略常数项，因为它既不影响最佳估计，也不影响误差条。它的值对于后验 pdf 来说只是乘了一个常数，只是将函数整体放大。

生成随机数：根据 (52) 生成均匀分布的方位角样本 θ_k，用 (53) 将其转换为位置数据 x_k，假定已知灯塔距离海岸 1 公里 ($\beta = 1$)。如图 12：在前几个图中，闪光的位置由图形顶部的小圆圈标记，数据的数量显示在右上角。

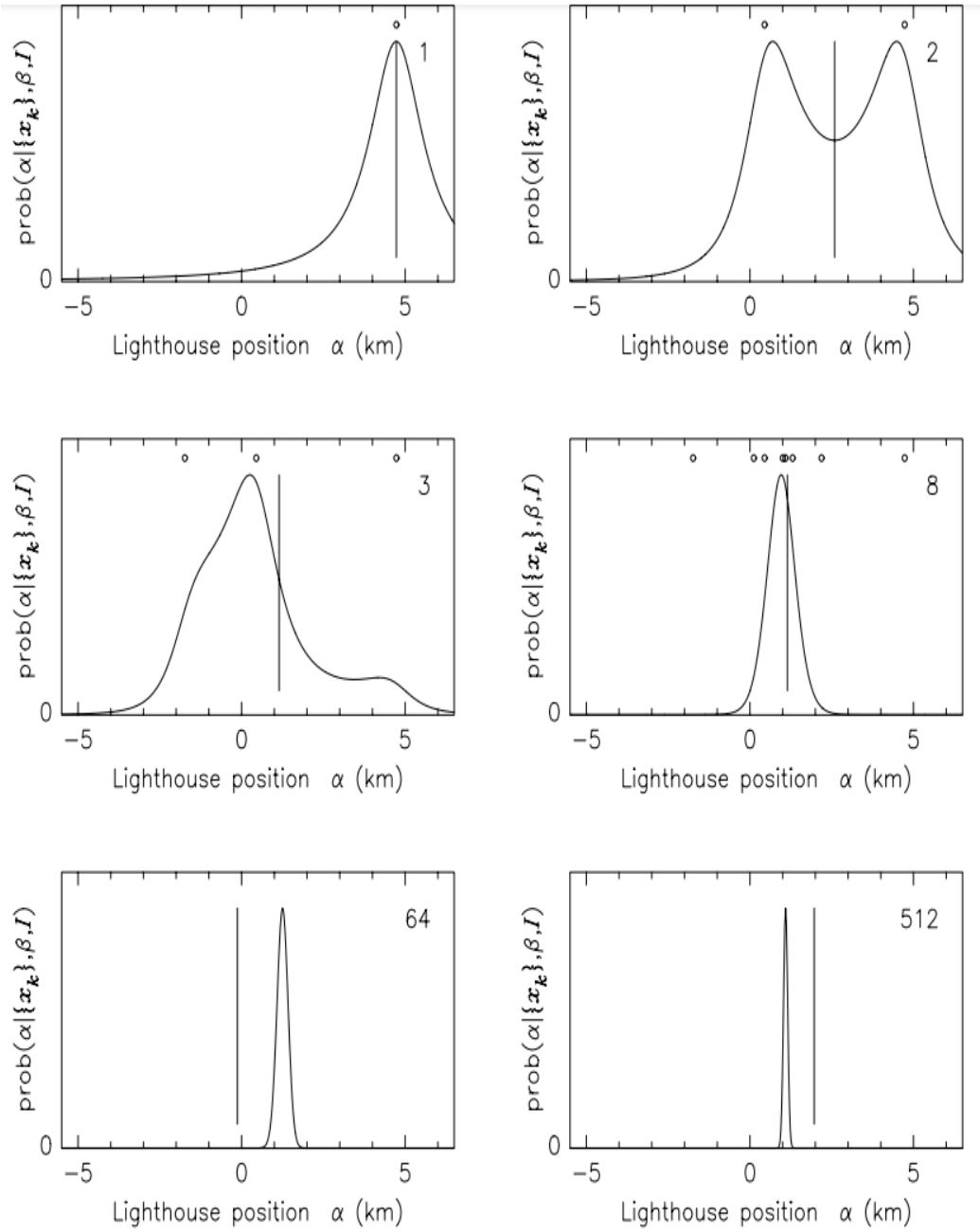

图 2.9: 随着海岸上探测到的闪光数目的增加，灯塔位置的后验 pdf 如何演变，直线表示平均值

数据量很少时很容易成为多模态，经过大约十几次的测量，后验接近高斯 pdf 形式。随着实验证据的增加，后 pdf 变窄，峰值收敛到 α= 1。与预期一致，因为数据是在沿海岸 1 公里

处生成的。

2.4.1 对应的代码

首先 $num\_pts = 512$ 用 $x\_pts = np.arange(num\_pts)$ 生成 512 个坐标点数组，然后用 $dist = cauchy(x0\_true, y0\_true), dist\_pts = dist.rvs(num\_pts)$ 生成了满足柯西分布的数据数组 () 柯西分布的两个参数选用了真实值 $x0\_true = 1.y0\_true = 1.$ ，并从中随机抽取 512 个随机样本，并与坐标点一一对应，绘图，以下三个图：

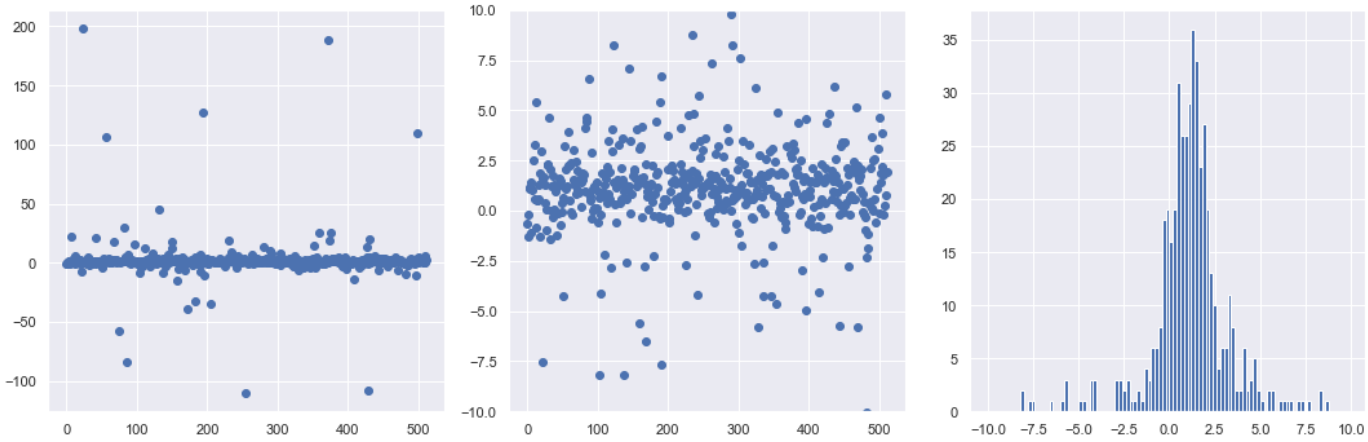

图 2.10:

第一个图：以 x_pts512 个坐标点为横轴，以 $dist\_pts$ 从柯西分布中抽出的 512 个样本值为纵轴，可以看出样本值在 0 附近的点显著的多。

第二个图：是把第一个图放大了，显示样本值在-10 到 10 附近的数据点。第三个图：横坐标是样本取值，纵坐标是取该值的样本的个数。

上面的例子展示了柯西分布，接下来对于灯塔的例子。

同样首先 $num\_pts = 512$ 用 $x\_pts = np.arange(num\_pts)$ 生成 512 个坐标点数组，已知 $prob(\theta_k|\alpha, \beta, I) = 1/\pi$，所以用 $theta\_dist = uniform(-np.pi/2., np.pi/2.)$，均匀的在 $-\pi/2$ 到 $\pi/2$ 之间取点，再用 rvs 对其随机取样取了 512 个；也可以用 np.random.uniform($-\pi/2, \pi/2$,512) 直接随机取出 512 个样本，画出来的图是一样的，只是两个 uniform 来自于不同的库。

对于取出的 θ_k 的值，需要转化为坐标的样本，利用公式

$$\beta tan\theta_k = x_k - \alpha.$$

并且，似然函数的表达式为

$$prob(x_k|\alpha,\beta,I)=\frac{\beta}{\pi[\beta^2+(x_k-\alpha)^2]},$$

将 $x_k-\alpha$ 整体带入表达式右侧，并且已经假设 β 已知，所以将 $\beta=1$ 代入似然函数，得到似然函数的表示式

$$prob(x_k|\alpha,\beta,I)=\frac{1}{\pi[1+(tan\theta_k)^2]},$$

$prob(x_k|\alpha,\beta,I)=dist\_pts\_alt=1/(np.pi*(1+(np.tan(theta\_dist\_pts).T)*(np.tan(theta\_dist\_pts))))$

直接对似然函数绘图

画出来三个图：

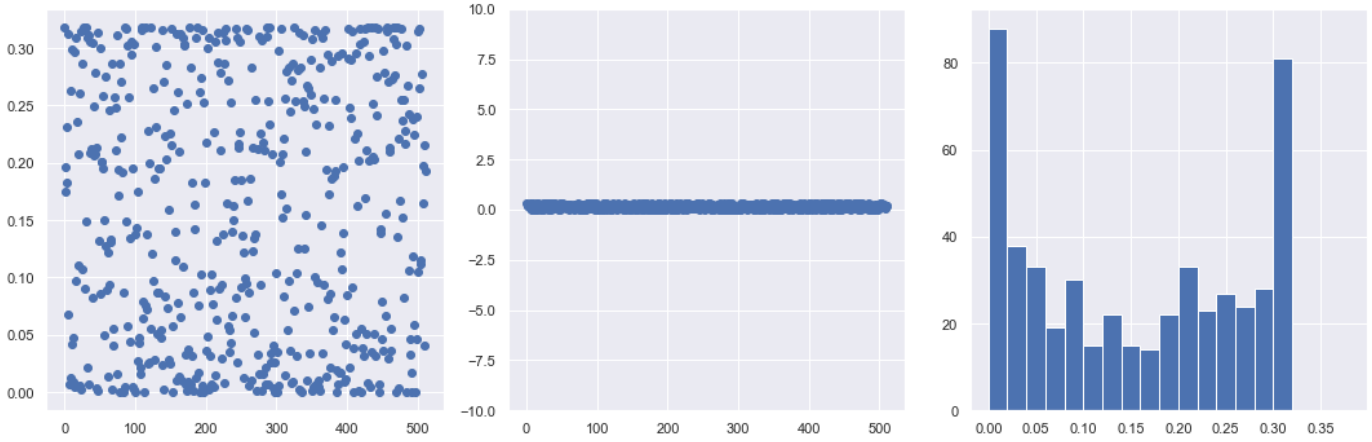

图 2.11:

这三个图可以看出似然函数的值分布在，0 到 0.3 多，并且从第三个图中看出，似然函数取值再 0 附近和右边界附近的数量明显比中间区域更多。

这样分布是合理的。$\tan\theta_k$ 的图像如下，

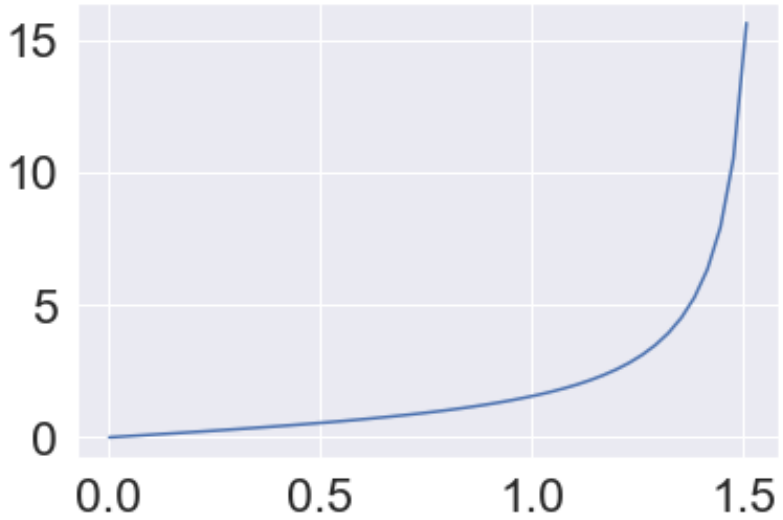

图 2.12:

根据计算，tan 值处在 1.128 到 1.524 的时候，似然函数的值才会洒在 0.05 到 0.25 之间，下面的 tan 函数形状中可以看出，tan 处在 1.128 到 1.524 的区间范围很小，对应的上面第三个图中，生成的数据点大多在 0 和 0.3 附近。

根据 $x_k = \beta tan\theta_k = x_k + \alpha$，画出 $prob(x_k|\alpha, \beta, I) - x_k$ 图

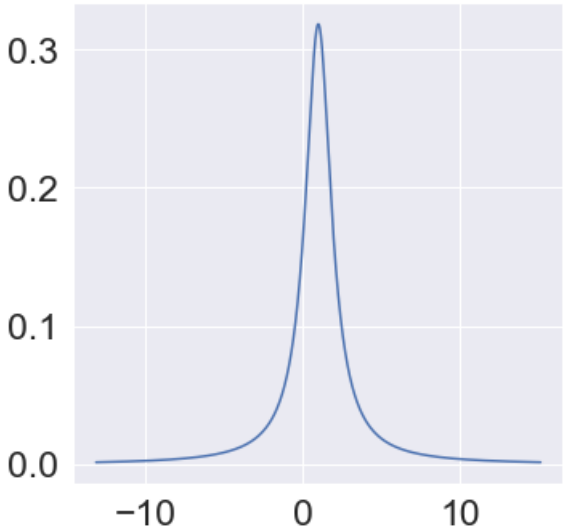

图 2.13:

在 α 和 β 已知的情况下，生成的样本 x_k 落在 $x_k = 1$ 的概率最大。

接下来绘制了计算并绘制了不同数据量下的 xo 后验 $prob(x_k|\alpha, \beta, I)$

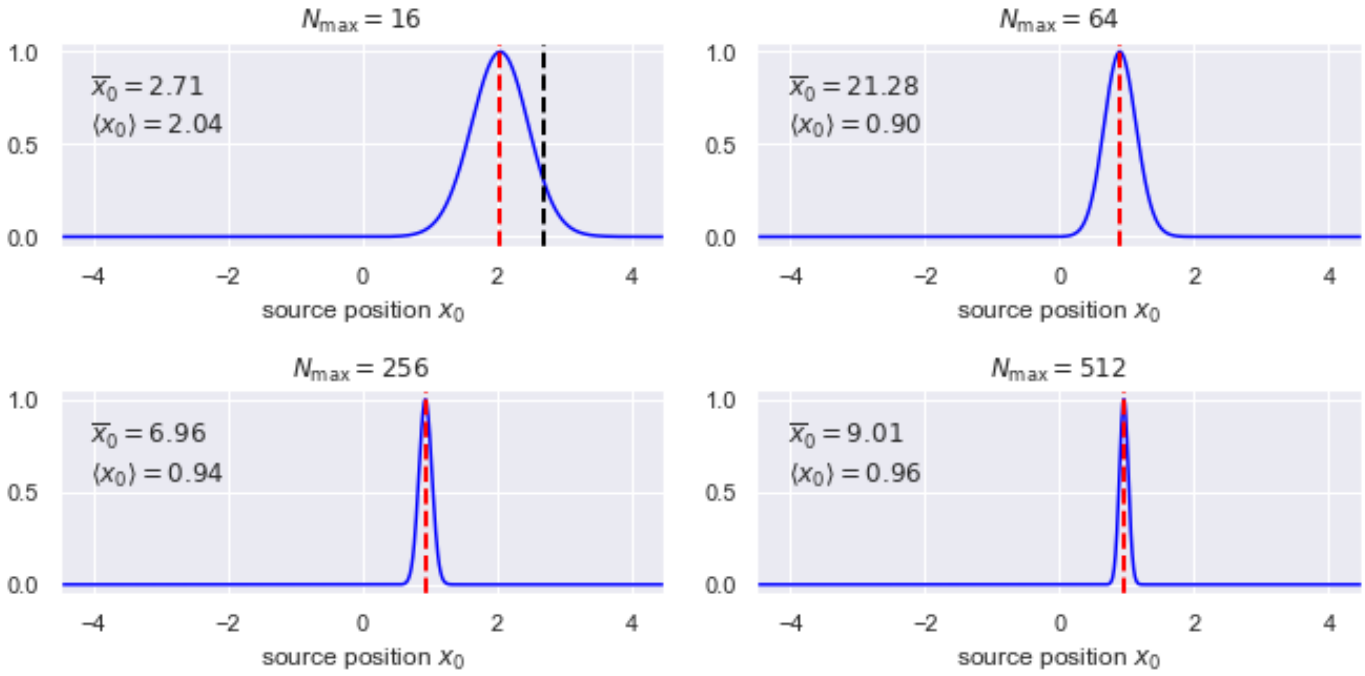

图 2.14:

红线是 $< X_0 >$

$$< X_0 >= \frac{\sum X_0 * posterior}{\sum posterior},$$

黑线是 $\bar{X}_0$

$$\bar{X}_0 = mean(X_0[0 : N_max]).$$

数据的均值并不能很好的描述参数的最佳估计，图上也可以看出黑线与最佳估计之间有一定的差距。与之前不同，在高斯分布中，参数 μ 的最佳估计 $\mu_0 = \frac{1}{N}\sum_{k=1}^{N} x_k$，由样本简单的算数平均给出。在灯塔例子中，数据来自柯西 pdf，由于这个分布也是关于我们感兴趣的参数 α 对称的，所可能会认为，数据的平均值也可以提供对灯塔位置的良好估计。实际上样本均值并不是 X 的最佳估计。图中也可以看到，数据的平均值由黑色长竖线表示。可以看到对于这个问题，样本均值不是一个很好的估计。

中心极限定理: 如果从 (几乎) 任何 pdf 中随机抽取样本，pdf 的均值为 μ，那么在数据充足的情况下，数据的平均值将趋向于这个值 μ;μ 和样本均值之间的差值的误差条会以 $\sqrt{N}$ 下降，随着数据量增加，在除以 N 来计算平均值时，平均值上的误差条会以 $\sqrt{N}$ 的量级减小。但

对于柯西分布，由于在 $-\infty$ 到 ∞ 是不可积的，所以柯西分布的期望值是没有定义的，σ^2 无限大，并且 μ 不确定，所以数据平均值的可变性并不会随着测量次数的增加而减少，而且在测量了一千个或一百万个数据后，其“错误”可能和测量了一个数据后一样。

反而后验均值对于最佳估计有一个很好的描述，也就是红线。

$$< X_0 > \frac{\sum X_0 * posterior}{\sum posterior}.$$

3 参数估计

之前的内容只涉及一个未知变量，接下来考虑有好几个参数的情况，对其中一些感兴趣。对error-bar、边缘化进行推广。还将看到某些近似如何自然地产生一些最常用的分析过程，并讨论所谓的误差传播 (propagation of errors)。

3.1 exampli 4: 存在信号本底时的信号振幅

最简单的情况，如图

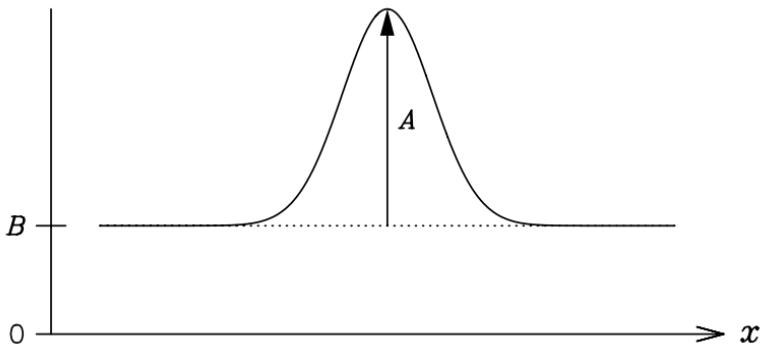

图 3.1:

横轴 c 为测量的变量。背景 (比如系统误差) 可以被认为是平坦的，幅度是 B 是未知的，而感兴趣的信号是已知形状和位置峰，幅值为 A。问题的数据通常都是整数值。给定一组实验装置测量的计数 N_k ，实验探测装置位于 X_k，**对信号峰值和背景的振幅的最佳估计是什么?**

先对数据的性质进行判断，应该期望第 k 个数据通道中的计数与 x_k 处的信号和背景之和成正比；取峰值的形状为 gauss 分布，宽度为 ω，中心值在 x_0 处**理想数据**由 D_k 给出：

$$D_k = n_0[Ae^{-(x_k-x_0)^2/2\omega^2} + B], \tag{3.1}$$

n_0 是与测量时间有关的常数。然而，与计数 N_k 的数量不同，上式中的 D_k 通常不是一个整数。因此，实际的数据将是这种理想数据附近的一个大于等于 0 的整数。**泊松分布**是一个满足这种属性的 pdf，通常在**这样的计数实验**中被调用。

泊松分布：

日常生活中，大量事件是有固定频率的。比如某医院平均每小时出生 3 个婴儿，这种事件的特点是可以预估这些事件的总数，但是没法知道具体的发生时间。已知平均每小时出生 3 个婴儿，那么两个小时中的第二个小时，会出生几个? 有可能一下子出生 6 个，也有可能一个都不出生。

泊松分布就是描述某段时间内，事件具体的发生概率，

$$P(N(t)=n)=\frac{(\lambda t)^n e^{-\lambda}}{n!}, \tag{3.2}$$

等号的左边，P 表示概率，N 表示某种函数关系，t 表示时间，n 表示数量，λ1 小时内出生 3 个婴儿的概率，就表示为 P(N(1) = 3) 。等号的右边，λ **表示事件的频次**。接下来两个小时，一个婴儿都不出生的概率是 0.25%，基本不可能发生。与这个例子相对应，单位时间 t=1 内，已知理想数据数据为 D_k 周围的整数，所以 D 对应频率 λ，$N_k = N$ 的概率分布函数为

$$prob(N|D)=\frac{D^N e^{-D}}{N!} \tag{3.3}$$

与之前的定义 <X> 相似，将上式代入离散形式定义，期望值 =D ：

$$< N >= \sum_{N=0}^{\infty} N prob(N|D) = D, \tag{3.4}$$

根据式子 (60)，数据为 N_k 的似然函数:

$$prob(N_k|A,B,I)=\frac{D_k^{N_k} e^{-D_k}}{N_k!}, \tag{3.5}$$

其中，背景信息 I 包括：计数的期望数量 D_k 与感兴趣的参数 A 和 B 之间的关系的知识; 对于 (59) 的高斯峰形模型，这意味着 x_0、ω 和 n_0 取给定 (以及 x_k)。如果数据是独立的，那么，当 A，B 给定时，在一个通道内观察到的 N_k 不影响在另一个通道内发现的粒子数，所以 likelihood 是单个测量的 prob 的乘积：

$$prob(\{N_k\}|A,B,I)=\prod_{k=1}^{M} prob(N_k|A,B,I), \tag{3.6}$$

对信号振幅和信号本底的推断体现在 posterior 中，

$$prob(A,B|\{N_k\},I) \propto prob(\{N_k\}|A,B,I) \times prob(A,B|I) \tag{3.7}$$

最简单的是均匀的 pdf，由于振幅和信号本底都不能是负的，所以：

$$prob(A,B|I)\begin{cases} Constant \quad for A \geq 0 and B \geq 0, \\ 0 \quad otherwise, \end{cases} \tag{3.8}$$

将公式代入 (65), 并取对数，可以得到

$$L = log_e[prob(A, B|N_k, I)] = constant + \sum_{k=1}^{M}[N_k log_e(D_k) - D_k], \tag{3.9}$$

常数项不包含 AB 系数并且求和项大于 0，对信号峰和背景幅度的最佳估计是由 L 的最大值给出的，可靠性由 posterior 在最佳点附近的宽度给出。

数据是根据方程 (63)，用泊松分布随机数发生器从方程 (59) “平坦背景上的高斯模型” 生成的，以下是四组数据以及得到的后验，绘制为直方图因为观测通道是离散的，并且统一观测装置的宽度。基础信号以原点为中心，所以 $x_0 = 0$，并且半高宽为 5 个单位，假设这些对于分析是已知的

后验现在是二维的因为它同时是 A 和 B 的函数，可以用等高线表示，也就是等置信区间线。最外圈到最内圈，分别是 $10\%, 30\%, 50\%, 70\%, 90\%$ 等置信区间线。

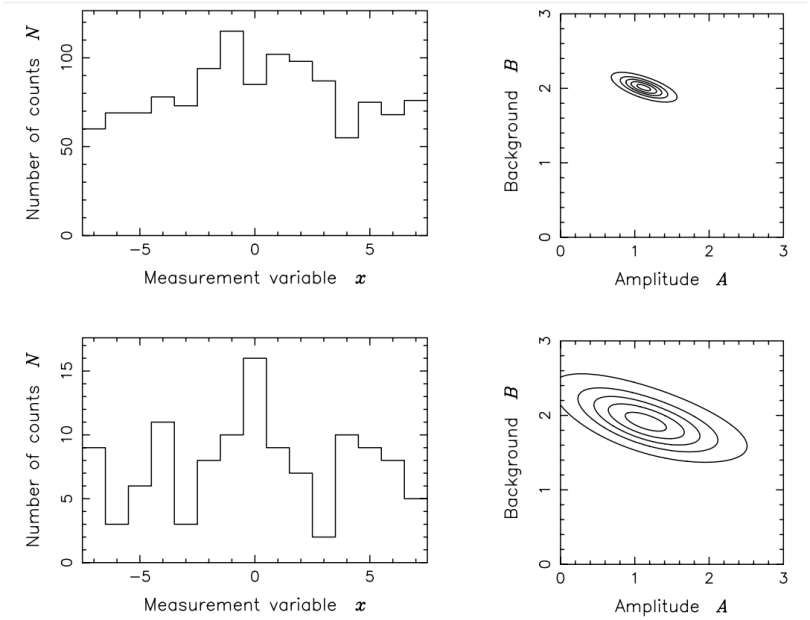

图 3.2:

第一张图显示在 15 个数据箱中检测到的计数数，参数 n_0 是给定的，n_0 **是与时间有关的常数**，理想数据 D_k 的最大值给出 n_0，$n_0 = D_{kmax}/A + B$，最大 expect data 的值为 100，由此而得到 n_0，由方程 (67) 的指数得到对应的后验 pdf，画在第一行的第二个图中；表明信号幅度的最佳估计大约 1，大约是背景幅度一半。

第二张图是相同的实验设置下，但是实验只进行了十分之一的时间，理想计数数量是之前的 1/10 倍即 $D'_{kmax} = 0.1D_{kmax}$，从而得到新的 n_0，实验数据减少，数据看起来更嘈杂。右

边的图在两个方向上的后验 pdf 大约比之前多了三倍宽，与之前 $\mu = \mu_0 \pm \frac{\sigma}{\sqrt{N}}$ 类似，数据量减少，置信区间宽度以 $\sqrt{N}$ 的比例增大，大约是三倍多宽。并且可以看到第二行右边的图中 A 小于 0 的后验被截断了体现了当数据不够好时，先验的重要性。

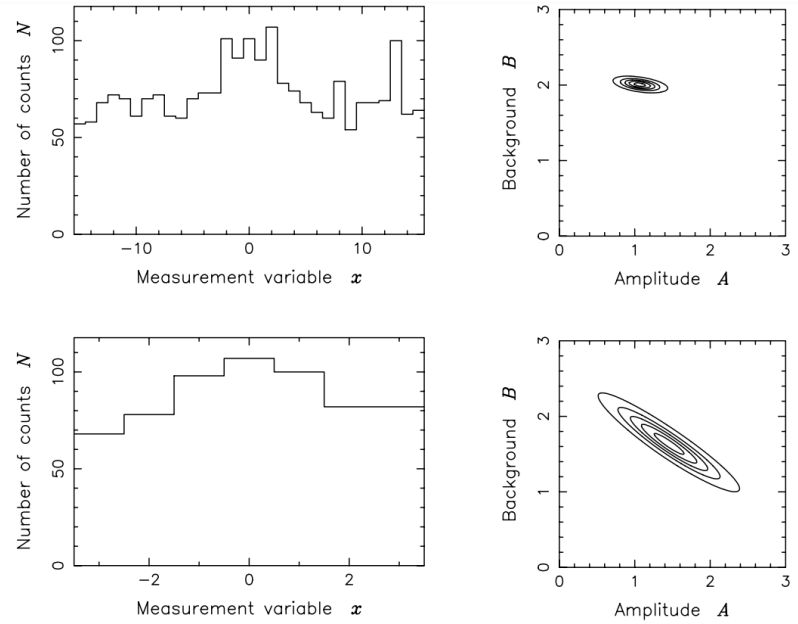

图 3.3:

这张图的第一行与上一张图第一行的计数率相同，有着相同的 D_kmax 和 n_0 但是粒子探测通道增加，总探测范围变为原来两倍，探测器间距不变，可以看到 A 和 B 的置信区间都缩小了，等置信区间图变得更加平稳。

第二行探测器数量减少，探测范围减少，数据量是是之前的图第一行的数据的一半，但是置信区间的宽度明显大于 $\sqrt{2}$ 倍，这是因为只有**在** $x_0 = 0$ **附近的数据**才能体现信号峰的信息，而**远离** x_0 **的数据**会提供背景幅度的信息，所以虽然计数率提高，但是得到的结果依然不好，并且图像明显倾斜。这些特征表明了对 A 和 B 的估计之间有很强的**相关性**。由于收集数据的 x 的范围受到了严重的限制，因此很**难将信号与背景区分**开。

3.1.1 Marginal distributions

二维后验 pdf 很好的描述了对 A 和 B 的值的联合推断。但是实际上通常对背景信息不感兴趣，只想计算对 A 的估计，也就是需要后验 $prob(A|\{N_k\}, I)$。根据边缘化规则可以通过积分得到：

$$prob(A|\{N_k\}, I) = \int_0^\infty prob(A, B|\{N_k\}, I)dB. \tag{3.10}$$

也可以对 A 积分得到关于背景振幅 B 的后验 pdf:

$$prob(B|\{N_k\},I) = \int_0^\infty prob(A,B|\{N_k\},I)dA. \tag{3.11}$$

四组边缘化后的数据如图所示，实验设置与是一致的，图 3.2 和图 3.3 中数据集对应的四组边缘分布和后验 pdf 画在图 3.4 和图 3.5 中，图 3.4、3.5 中更容易看到不同的实验设置对推断 A 和 B 的值的可靠性的影响。

这里应该注意到 $prob(A|\{N_k\},I)$ 与 $prob(A|\{N_k\},B,I)$ 是不同的，第一个 prob 表示了对于 B 的值的无知，但是第二个 prob 表示已知 B，在图 3.4、3.5 中用虚线表示已知 B=2 的后验。

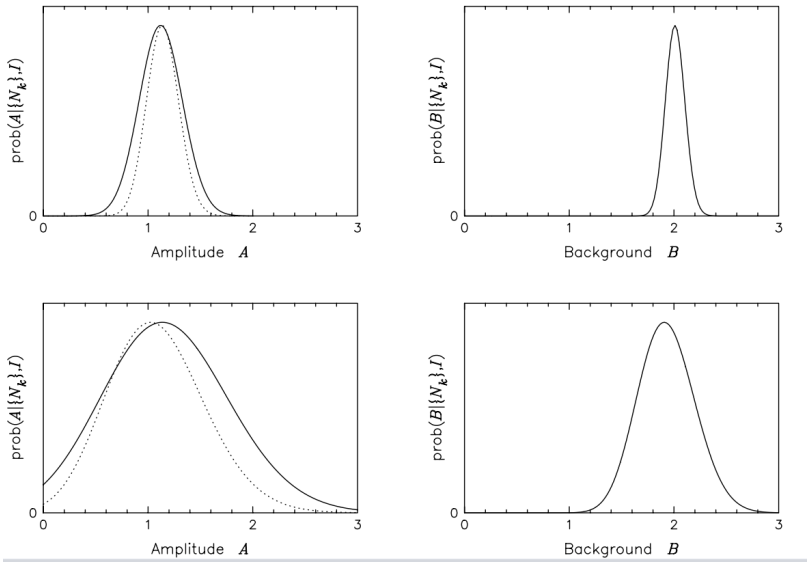

图 3.4: 一、二组数据

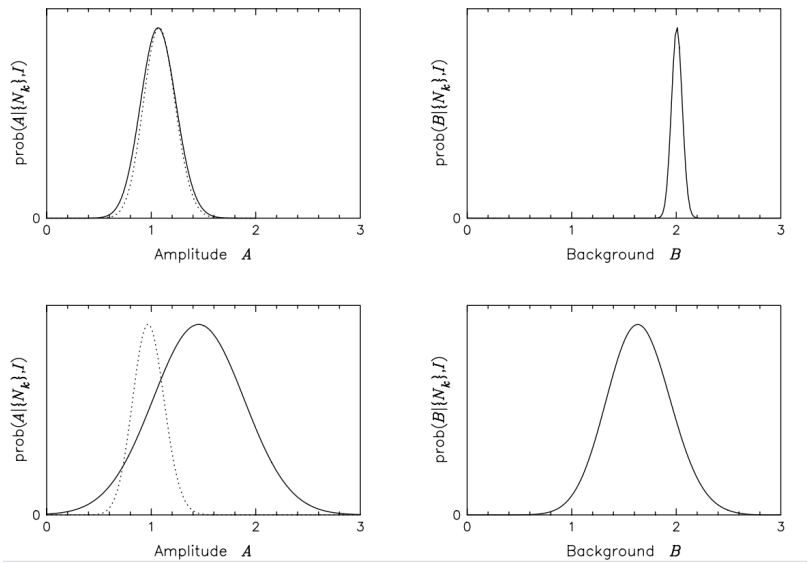

图 3.5: 三、四组数据

与边缘化后验对比：

最后一个数据集，相比边缘化 pdf，虚线的后验概率显著缩小，因为第四组实验并不能很好的区分信号振幅和背景，所以已知 B 可以更好地对 A 进行分析。

第三组数据集，测量范围远超过信号峰范围的情况下，实线和虚线差异最小。
当能够很好地区分背景和信号峰时，对 B 进行单独的校准实验所得到的东西很少, 如果数据集被严重截断 (比如第四组数据)，或者背景是高度结构化的，这些额外的信息 (比如已知 B 的值) 非常有益。

假设信号峰的形状和位置已知，在高斯模型方程中

$$D_k = n_0[Ae^{-(x_k-x_0)^2/2\omega^2} + B],$$

隐含了背景信息 I 中的 ω 和 x_0 的值。如果不知道这些值，从边缘化的讨论中可以看出，面对妨害参数应该整合相关变量：

$$prob(A, B|\{N_k\}, I) = \int\int prob(A, B, \omega, x_0|\{N_k\}, I)d\omega dx_0, \tag{3.12}$$

假设 I 包含用高斯模型模型提供理想数据，但不一定包含对高斯峰的宽度和位置的知识。二重积分下的四参数后验 pdf 本身可以展开：

$$prob(A, B, \omega, x_0|\{N_k\}, I) \propto prob(\{N_k\}|A, B, \omega, x_0, I) \times prob(A, B, \omega, x_0|I). \tag{3.13}$$

右边第二项是 A, B, ω, x_0 的 prior pdf，可以展开为

$$prob(A, B, \omega, x_0|I) = prob(A, B|I) \times prob(\omega, x_0|I). \tag{3.14}$$

如果已知高斯峰的宽度和位置，那么 ω, x_0 的 prior 是非常尖锐的。在完全确定这两个参数的极限下，有

$$prob(\omega, x_0|I) = \delta(\omega - 2.12)\delta(x_0), \tag{3.15}$$

非零点在 $\omega = 2.12$(即 FWHM=5) 和 $x_0 = 0$，在这种情况下，积分方程 (3.12) 非常容易计算

$$prob(A, B|\{N_k\}, I) = \int\int prob(A, B, \omega, x_0|\{N_k\}, I)d\omega dx_0,$$

代入 delta 函数

$$prob(A, B|\{N_k\}, I) \propto prob(\{N_k\}|A, B, \omega = 2.12, x_0 = 0) \times prob(A, B|I) \tag{3.16}$$

这个表达式回到了方程 (3.7)

$$prob(A, B|\{N_k\}, I) \propto prob(\{N_k\}|A, B, I) \times prob(A, B|I)$$

如果不知道 ω 和 x_0 的值，就必须为这些参数 (包括 A 和 B) 分配一个较宽的先验。边缘化积分相比于已知参数时的 delta 函数，需要做更多的计算，既可以数值计算，也可以解析近似。

3.1.2 绑定数据 Binning the data

用直方图将数据绘制在图 3.2、3.3 中时，提到用直方图是因为实验测量通常在有限宽度的通道中检测计数。这意味着，对于理想数据 D_k，eqn(3.1)，实际上应该被写成第 k 个数据箱上的一个积分：

$$D_k = \int_{x_k-\Delta/2}^{x_k+\Delta/2} n_0[Ae^{-(x_k-x_0)^2/2\omega^2} + B]dx, \tag{3.17}$$

这里假设所有的测量通道都有相同的宽度 。只要箱的宽度不太大，方程式（3.12）的积分可以近似为长矩形的面积：

$$D_k = n_0[Ae^{-(x_k-x_0)^2/2\omega^2} + B]\Delta. \tag{3.18}$$

因此，方程 (3.1) 是合理的，因为固定大小的 Δ 可以被吸收进 n_0。新的 n_0 反映了进行实验测量的时间和 “收集区域” 的大小。然而，容器宽度 Δ 并不总是由检测器的物理大小决定的，**但通常被选择为足够大，以便在由此产生的复合数据通道中有合理数量的计数。???**

对图 3.2 中第一个面板的实验设置对应的数据进行分析，但箱子变窄了 4 倍。

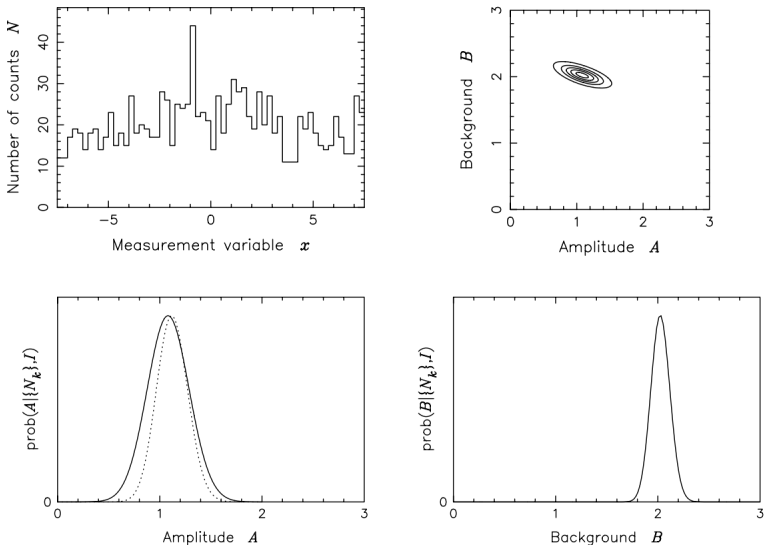

图 3.6:

这个数据集看起来更嘈杂，因为平均看，每个通道只有之前计数的四分之一。这张图还显示了 A 和 B 的后验 pdf，及其边缘分布，假设 $\omega = 2.12$ 和 $x_0 = 0$。. 与图 3.2 和图 3.4 中相应的 pdf 相比，所推断参数的可靠性几乎是相同的。虽然这种视觉增益并非没有价值，而且在处理较少的 “测量” 方面有计算优势，但我们必须始终意识到，过于粗糙的装箱 (Δ 过大) 会破坏数据中的有用信息。比如用一个很宽的大箱子计数，这样会把所有的计数加成一个数字，这样会完全没办法从区分背景和信号的信息，同时 (3.18) 的公式近似也会出现问题。?? p43)

3.2 Reliabiilties:best estimates,correlations and error-bars

对于多参数问题, 如果用 $\{X_j\}$ 表示感兴趣的参数集, 那么对于这些参数的最佳估计 $\{X_{Oj}\}$ 由后验 pdf:$prob(\{X_j\}|\{data\}, I)$ 的一阶导数给出

$$\left.\frac{\partial P}{\partial X_i}\right|_{\{X_{Oj}\}} = 0, \tag{3.19}$$

严格来说，还需要二阶导数小于 0，写出这个微分时，就隐含了假设后验关于这些参数是连续的。使用 P 的对数还是更加方便，所以

$$L = log_e[prob(\{X_j\}|\{data\}, I)]. \tag{3.20}$$

3.19 式也由 L 代替 P。首先考虑**两个变量**的具体情况：用 X 和 Y 表示。求解联立方程得到最佳估计 X_0, Y_0：

$$\left.\frac{\partial P}{\partial X}\right|_{X_0,Y_0} = 0 and \left.\frac{\partial P}{\partial Y}\right|_{X_0,Y_0} = 0 \tag{3.21}$$

这里

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

为了对该最佳估计的可靠性进行度量，需要要看二维后验 pdf 关于点 (X_0, Y_0) 附近的函数行为。如第 2.2 节所述，可以**通过泰勒展开**分析 pdf 的局部行为：

$$\begin{aligned} L = L\left(X_{\mathrm{O}}, Y_{\mathrm{O}}\right) + \frac{1}{2}\Bigg[& \left.\frac{\partial^2 L}{\partial X^2}\right|_{X_{\mathrm{o}},Y_{\mathrm{O}}} \left(X - X_{\mathrm{O}}\right)^2 + \left.\frac{\partial^2 L}{\partial Y^2}\right|_{X_{\mathrm{o}},Y_{\mathrm{O}}} \left(Y - Y_{\mathrm{O}}\right)^2 \\ & + 2\left.\frac{\partial^2 L}{\partial X \partial Y}\right|_{X_{\mathrm{o}},Y_{\mathrm{o}}} \left(X - X_{\mathrm{O}}\right)\left(Y - Y_{\mathrm{O}}\right)\Bigg] + \cdots, \end{aligned} \tag{3.22}$$