# 统计学笔记

Ruijun Shi[1]

2021 年 12 月 12 日

[1]GitHub：https://github.com/RuijunShi

## 摘要

简单的统计学笔记，主要是在天文学，特别是引力波和pulsar timing中遇到的统计学，现在没写多少内容。目前图片也没有绘制，等有空再说吧！同时也在缓慢更新中。内容相对比较基础。这也是我第一次用latex编写书籍，学习过程艰难啊。有错误请大家指出！基本上都是在自己不想读文献的时候就写写笔记，不知不觉的积累。不过目前的编排非常混乱，表述不够精炼，逻辑感不强，也有重复造轮子的嫌疑。希望慢慢改进吧。

# Contents

# Chapter 1

# 数理统计基础

## 1.1 概率

1. 概率的定义（略）概率满足：非负性，规范性，可列可加性

2. 概率的性质：

   重点：逆事件概率；加法公式；有限可加性

3. 条件概率：

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{1.1}$$

4. 乘法定理：

$$P(AB) = P(A|B)P(A) \tag{1.2}$$

5. 全概率公式：

$$P(A) = P\left(A \mid B_1\right) P\left(B_1\right) + P\left(A \mid B_2\right) P\left(B_2\right) + \ldots + P\left(A \mid B_n\right) P\left(B_n\right) \tag{1.3}$$

6. 独立性：满足

$$P(AB) = P(A)P(B) \tag{1.4a}$$

$$P(B|A) = P(B) \tag{1.4b}$$

## 1.2 单变量分布和多变量分布

### 1.2.1 单变量分布

1. 随机变量的概念（略）

2. 分布函数的概念（略）和性质：不减函数；$0 \leq F(x) \leq 1$; $F(x+0) = F(x)$

3. 概率密度函数

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{dt} \tag{1.5}$$

**性质 1.1** 概率分布的性质

$$f(x) \leq 0 \tag{1.6a}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{1.6b}$$

$$P\{x_1 < X < x_2\} = \int_{x_1}^{x_2} f(x)dx \tag{1.6c}$$

$$F'(x) = f(x) \tag{1.6d}$$

### 1.2.2 多变量分布

**定义 1.1** 二维随机变量
设$(X,Y)$是二维随机变量，对于任意实数$(x,y)$，二元函数

$$F(x,y) = P\{X \leq x, Y \leq y\}$$

为二维随机变量$(X,Y)$的**分布函数**。在$(X,Y)$为离散型，离散状态下，二维随机变量$(x,y)$取值$(x_i,y_j)$，$i,j = 1,2,\cdots$，则有：

$$P\{x = x_i, Y = y_j\} = p_{ij}, \qquad i,j = 1,2,\cdots$$

其中$p_{ij} \leq 0$且$\sum_i \sum_j p_{ij} = 1$，则称为离散型随机随机变量的**分布律**或**联合分布律**

**定义 1.2** 联合概率密度
对于非负可积的函数$f(x,y)$，对于任意实数$x,y$有

$$F(X,Y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(u,v)\mathrm{d}x\mathrm{d}y \tag{1.7}$$

则称为函数$f(x,y)$为$(X,Y)$的**概率密度**或**联合概率密度**

### 1.2.3 边缘分布

**定义 1.3 边缘分布** $F(x,y)$是随机变量$(X,Y)$的分布函数，$F_X(x)$和$F_Y(y)$分别是$X$和$Y$的分布函数：

$$F_X(x) = p\{X \leq x\} = p\{X \leq x, Y \leq \infty\} = F(x, +\infty) \tag{1.8}$$

对于随机变量$Y$同理。分布$F_X(x)$和$F_Y(y)$称为$(X,Y)$关于$X$和$Y$的**边缘分布函数**。

对于离散型随机变量，边缘分布为：

$$\begin{aligned} P\{X = x_i\} &= \sum_j p_{ij} = p_{\cdot i} \\ P\{Y = y_i\} &= \sum_i p_{ij} = p_{\cdot j} \end{aligned} \tag{1.9}$$

对于连续性随机变量，边缘概率密度为：

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy f_Y(y) = \int_{-\infty}^{+\infty} f(x,y)dx \tag{1.10}$$

同理，我们可以扩展到n维分布中，假设要求维度a的边缘分布，则对其他维度$^-a$即可

### 1.2.4 条件分布

**定义 1.4** 条件分布
对于二维随机变量$(X,Y)$，其分布律为：

$$P\{X = x_i, Y = y_j\} = p_{ij}, \qquad i,j = 1,2,\cdots$$

对于固定的$j$，若$P\{Y = y_j\} > 0$，则称为：

$$P\{X = x_i | Y = y_j\} = \frac{P\{X = x_i, Y = y_j\}}{PY = y_j} = \frac{p_{ij}}{p_{i\cdot}} \tag{1.11}$$

称为在$Y = y_i$条件下$X$的**条件分布率**

## 1.3 随机变量的数学特征

### 1.3.1 数学期望与方差

**定义 1.5** 数学期望
积分：

$$E(X) = \int_{-\infty}^{\infty} x f(x) \mathrm{d}x \tag{1.12}$$

为连续性随机变量的数学期望，离散状态下为:

$$E(X)=\sum_{k=1}^{\infty}x_k p_k \tag{1.13}$$

**定义 1.6** *方差*
设$X$是一个随机变量，若$E\{[X-E(X)]^2\}$存在，则称为$E\{[X-E(X)]^2\}$为随机变量$X$的方差，记为$D(X)$或者$\mathrm{Var}(X)$

我们可以看到方差描述的是一个随机变量的离散程度。根据定义，我们把方差写为：

$$D(X)=\int_{-\infty}^{\infty}[x-E(x)]^2 f(x)\mathrm{d}x \tag{1.14}$$

随机变量的方差可以写为：

$$D(X)=E(X^2)-[E(X)]^2 \tag{1.15}$$

**性质 1.2** *方差的性质*

- 设$C$为常数：$D(C)=0$
- 设$C$为常数，$X$为随机变量，有：

$$D(CX)=C^2D(X),\qquad D(X+C)=D(X)$$

- 设$X,Y$为两个随机变量，有：

$$D(X+Y)=D(X)+D(Y)+2E\{(X-E(X))(Y-E(Y))\}$$

若$X,Y$相互独立，则有：

$$D(X+Y)=D(X)+D(Y)$$

- $D(X)=0$的充要条件是$X$以概率为$1$取常数$E(X)$，即：

$$P\{X=E(X)\}=1$$

### 1.3.2 矩和协方差矩阵

矩在物理学中有广泛的应用，比如我们熟悉的力矩，电荷的分布等。力矩描述了力在空间上的分布；质量函数描述了质量在空间的分布等等，可以用下面的公式表征：

$$\mu_n=\int r^n\rho(r)\mathrm{d}r$$

而统计学中的矩也是类似，表征了概率密度函数的形状。类似的，我们也可以在数学上定义矩的概念:

**定义 1.7** 矩的概念

在数学上，矩是对函数的一种度量，是描述概率分布的一种方法。对于单变量分布，对于常数$c$的$k$阶矩，有：

$$\mu_k = \int (x-c)^k P(x) \mathrm{d}x \tag{1.16}$$

当$c=0, k=1$时，我们发现正是随机变量$X$的数学期望：

$$\mu_1 = \int x P(x) \mathrm{d}x = E(x)$$

我们把$c=0, k=k$的情况称为k阶原点矩。若$c=E(X), k=k$，则称为k阶中心距。我们看到方差公式正是我们的二阶中心矩。当我们有2个变量的时候可以定义混合矩的概念：

$$\mu_{kl} = \iint (x-c_x)^k (y-c_y)^k P(x,y) \mathrm{d}x \mathrm{d}y$$

同样的，当$c_x, c_y = 0$时，称为$k+l$阶混合矩：$E\{X^k Y^l\}$；当$c_x = E(X), c_y = E(Y)$时，称为$k+l$阶混合中心矩：$E\{X^k Y^l\}$。而随机变量$X, Y$的二阶混合中心矩则为协方差。

**定义 1.8** 协方差 随机变量$E\{(X-E(X))(Y-E(Y))\}$称为变量$X, Y$的协方差，记为$\mathrm{Cov}(X,Y)$：

$$\mathrm{Cov}(X,Y) = E\{[X-E(X)][Y-E(Y)]\} \tag{1.17}$$

协方差是变量误差的一种描述，衡量两个随机变量的相似性；而相关系数描述的是随机变量的相关性。若随机变量X,Y完全独立则有$\mathrm{Cov}(X,Y)=0$。

**定义 1.9** 相关系数

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}} \tag{1.18}$$

当二维随机变量的二阶中心矩存在：

$$\begin{aligned} c_{11} &= E\{[X_1 - E(X_1)]^2\} \\ c_{12} &= E\{[X_1 - E(X_1)][X_2 - E(X_2)]\} \\ c_{21} &= E\{[X_2 - E(X_2)][X_1 - E(X_1)]\} \\ c_{22} &= E\{[X_2 - E(X_2)]^2\} \end{aligned} \tag{1.19}$$

则矩阵

$$\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}$$

称为协方差矩阵。若有$n$维随机变量，

$$c_{ij} = \mathrm{Cov}(X_i, X_j), i, j = 1, 2, \cdots, n \tag{1.20}$$

则矩阵：

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nn} \end{bmatrix} \tag{1.21}$$

该矩阵是一个对称矩阵。在对角线上则为该变量的方差。

**定义 1.10** *矩母函数*

*定义矩母函数为：*

$$\psi(t) = E[e^{tX}] = \int e^{tX} dF(x) \tag{1.22}$$

对矩母函数求$n$次导可得：

$$\psi^n(t) = E[X^n e^{tX}] \tag{1.23}$$

当$t = 0$时，我们发现：

$$\psi^n(0) = E[X^n] \tag{1.24}$$

这正好对应了我们的n阶原点矩公式。

### 1.3.3 多元正态分布及协方差矩阵的直观理解

协方差矩阵描述随机变量的总体误差，表示随机变量之间的相似程度；而方差是协方差的一种特殊形式。协方差可以用多元正态分布直观理解其意义。对于一个边缘分布为正态分布的二维分布,其概率密度分布为：

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[ \frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1^2\sigma_2^2} + \frac{(x_2-\mu_2)^2)}{\sigma_2^2} \right] \right\} \tag{1.25}$$

我们知道二维正态分布的协方差为：$\mathrm{Cov}(X, Y) = \rho\sigma_1\sigma_2$ 其协方差矩阵为：

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

记

$$\mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

经过计算，我们发现：

$$\begin{aligned} &(\mathbf{X}-\mu)^T \mathrm{C}^{-1} (\mathbf{X}-\mu) \\ &= \frac{1}{2(1-\rho^2)} \left[ \frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1^2\sigma_2^2} + \frac{(x_2-\mu_2)^2)}{\sigma_2^2} \right] \end{aligned}$$

因此二维正态分布可以写为：

$$f(x_1, x_2) = \frac{1}{(2\pi)^{2/2}(\det \mathbf{C})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\}$$

我们发现N维正态分布是由协方差矩阵**C**规定的。当协方差**C**改变时，多维正态分布的函数形状也会依此改变。

——此处应该有图像——-

我们可以从二维正态分布扩展到n维正态分布：

$$f(X) = \frac{1}{(2\pi)^{n/2}(\det \mathbf{C})^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{X}-\boldsymbol{\mu})\right\} \tag{1.26}$$

这就是N维正态分布的一般形式。n维正态分布的性质有：

**性质 1.3** ***n*维正态分布的性质**

- 若$(X_1, X_2, \cdots, X_n)$服从$n$维正态分布，对于任意的$X_i (i = 1, 2, \cdots, n)$服从一维正态分布；若$X_1, X_2, \cdots, X_n$均服从正态分布且相互独立，则$(X_1, X_2, \cdots, X_n)$为服从$n$维正态分布。
- $n$正态维随机变量$(X_1, X_2, \cdots, X_n)$服从$n$维正态分布的充要条件是$X_1, X_2, \cdots, X_n$的线性组合：

$$l_1 X_1 + l_2 X_2 + \cdots + l_n X_n$$

- 若$(X_1, X_2, \cdots, X_n)$服从$n$维正态分布，而$Y_1, Y_2, \cdots, Y_n$与$X_1, X_2, \cdots, X_n$是线性变化，则$(Y_1, Y_2, \cdots, Y_n)$也服从$n$维正态分布。
- 若$(X_1, X_2, \cdots, X_n)$服从$n$维正态分布，则$X_1, X_2, \cdots, X_n$线性无关。

### 1.3.4 偏度和峰度

对于概率分布的形状可以用偏度系数和峰度系数表征。首先我们引入标准矩的概念：

**定义 1.11 标准矩** 一个概率分布的标准矩是经过标准化后的中心矩。标准化通常是将其除以标准差的过程，这样做可以使得标准矩对缩放和离散程度皆能保持一致，在比较不同概率分布的形状时更为方便。

$$\hat{\mu}_k = \frac{\mu_k}{\sigma^k} = \frac{\mathrm{E}[(X-\mu)^k]}{(\mathrm{E}[(X-\mu)^2])^{k/2}} \tag{1.27}$$

因为进行了标准化，因此标准矩具有缩放不变性。结合上一节对中心矩的定义，一阶标准矩即为随机变量标准化后的一阶中心距；因此一阶标准矩恒为0；二阶标准矩恒为1。而下面要定义的偏度则为三阶标准矩；峰度则为四阶标准矩。

**定义 1.12 偏度系数** 表征分布形态与平均值偏离程度，作为分布不对称的测度：

$$g_1 = \hat{\mu}_3 = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}[(X-\mu)^3]}{(\mathrm{E}[(X-\mu)^2])^{3/2}} \tag{1.28}$$

**定义 1.13 峰度系数** 表征分布形态图形顶峰的凸平度(即渐进于横轴的陡度):

$$g_2 = \hat{\mu}_4 = \frac{\mu_4}{\sigma^4} = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^4])^{4/2}} \tag{1.29}$$

## 1.4 常见分布及其数学特征

### 1.4.1 离散型分布

1. (0-1)分布:

$$P(X=k) = p^k(1-p)^{1-k}, \qquad 0<p<1, k=0,1 \tag{1.30}$$

0-1分布是最简单的分布，一个只有两种结果的随机现象即为0-1分布。其期望为$p$，方差为$p(1-p)$

2. 二项分布: 当有n次0-1分布时即为二项分布。

$$P(X=k) = \left( \begin{array}{c} n \\ k \end{array} \right) p^k(1-p)^{n-k} \tag{1.31}$$

其中$\binom{n}{k} = \frac{n!}{k!(n-k)!}$为二项式系数。均值：$np$; 方差：$np(1-p)$

3. 泊松分布:

$$P(X=k) = \frac{\lambda^k e^{-\lambda}}{k!}, k=0,1,2,\ldots \tag{1.32}$$

当二项分布的n很大而p很小时候，泊松分布即为二项分布的近似。均值：$\lambda$; 方差: $\lambda$。这是最重要的离散分布。

### 1.4.2 连续型分布

首先介绍下$\Gamma$ 函数：

$$\Gamma(z) = \int_0^{+\infty} t^{z-1}e^{-t}\mathrm{d}t \tag{1.33}$$

这个函数在下面会经常用到。

1. 均匀分布:

$$f(x) = \begin{cases} \frac{1}{b-a} & a<x<b \\ 0 & \text{otherwise} \end{cases} \tag{1.34}$$

2. 指数分布:

$$f(x)=\begin{cases}\frac{1}{\theta}e^{-x/\theta} & x>0\\ 0 & \text{otherwise}\end{cases} \tag{1.35}$$

指数分布有一个非常重要的性质是无记忆性。

3. 正态分布:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.36}$$

$f(x)$关于$\mu$对称；$f(\mu)=\max[f(x)]=\frac{1}{\sqrt{2\pi}\sigma}$。均值为$\mu$；方差为$\sigma^2$。当$\mu=0$，$\sigma=1$时的正态分布称为标准正态分布；对于一个正态分布：$X\sim N(\mu,\sigma^2)$，则有变量：

$$Z=\frac{X-\mu}{\sigma}\sim N(0,1)$$

变量Z称为标准化变量。正态分布又称为高斯分布，多维高斯分布在天文中也是非常重要的，很多参数的估计开始都要考虑一个多维正态分布。

4. Beta分布:

$$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} \tag{1.37}$$

其中：$0\le x\le 1,\ \alpha>0,\ \beta>0,\ \Gamma(z)=\int_0^{+\infty}t^{z-1}e^{-t}\mathrm{d}t$。其期望为：$\frac{a}{a+b}$，方差为：$\frac{ab}{(a+b)^2(a+b+1)}$
Beta分布是从伯努利事件建模的出来的，Beta分布描述了一个事物出现的所有可能性的大小分布。

5. Gamma分布:

$$f(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x} \tag{1.38}$$

其中：$x>0,\ \alpha>0,\ \beta>0$; 数学期望为：$\frac{1}{\lambda}$; 方差为：$\frac{n}{\lambda^2}$

6. Inv-Gamma分布:

$$f(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-(\alpha+1)}e^{-\frac{\beta}{x}} \tag{1.39}$$

其中：$x>0,\ \alpha>0,\ \beta>0$

7. $\chi^2$分布:

$$f_k(x)=\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}} \tag{1.40}$$

等价$\alpha=k/2,\beta=1/2$的Gamma分布

8. Inv-$\chi^2$分布:

$$f(x)=\frac{2^{-\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)}x^{-\left(\frac{k}{2}+1\right)}e^{-\frac{1}{2x}} \tag{1.41}$$

等价$\alpha=k/2,\beta=1/2$的Inv-Gamma分布

9. Scaled Inv-$\chi^2$分布:

$$f(x)=\frac{\frac{k}{2}^{-\frac{k}{2}}s^k}{\Gamma\left(\frac{k}{2}\right)}x^{-\left(\frac{k}{2}+1\right)}e^{-\frac{ks^2}{2x}} \tag{1.42}$$

等价$\alpha=k/2,\beta=ks^2/2$的Inv-Gamma分布。

这些分布各有各的作用，也有其数学上和现实的意义。

## 1.5 概率密度的传递

如果我们需要从一个已知的变量x分布转换为另外变量y的分布（而一般这种变换是非线性的），这时候需要概率密度的传递公式。

对于随机变量X的概率密度函数$f_X(x)$已知的情况下，需要得到随机变量Y的概率密度函数，而随机变量X和Y有一个非线性变换的关系：$x=g(y)$；也就是说函数$g$是函数$f$的反函数。则概率密度函数$f(x)$转变为$\widetilde{f}(y)=f(g(y))$。这时候对于随机变量X和随机变量Y有关系：

$$p_X(x)\delta x\simeq p_Y(y)\delta y \tag{1.43}$$

因此对于随机变量Y的概率密度分布为：

$$\begin{aligned}p_Y(y)&=p_X(x)\left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|\\&=p_X[g(y)]|g'(y)|\end{aligned} \tag{1.44}$$

若对随机变量X和随机变量Y不是一维变量，而是多维变量$X$和$Y$，其反函数$g(y)$与随机变量$X$的概率密度函数$f$一一对应，则有等式：

$$p_X(x_i)=\sum_i p_Y(y_i)$$

这时候我们的误差传递公式可以写为：

$$p_{\mathbf{Y}}(y)=|\mathbf{J}|p_X[g(y)] \tag{1.45}$$

其中$\mathbf{J}$为Jacobian矩阵：

$$\mathbf{J}=\begin{bmatrix}\frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_1}{\partial y_n}\\ \vdots & \ddots & \vdots\\ \frac{\partial x_n}{\partial y_1} & \cdots & \frac{\partial x_n}{\partial y_n}\end{bmatrix} \tag{1.46}$$

## 1.6 从大数定律，中心极限定理到数理统计

### 1.6.1 大数定理

**定理 1.1** *弱大数定理，辛钦大数定理*

设$X_1, X_2, \cdots$是独立同分布的，且数学期望为$E(x_k) = \mu$，前$n$个变量的算术平均为$\frac{1}{n}\sum_{k=1}^{n} X_K$，则对于任意的的$\varepsilon > 0$有：

$$\lim_{n\to\infty} P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} X_K - \nu\right| < \varepsilon\right\} = 1 \tag{1.47}$$

**定理 1.2** 强大数定理
若$X_1, X_2, \cdots$独立同分布，且都有均值$\mu$，则：

$$P\{\lim_{n\to\infty}(X_1 + \cdots + X_n)/n = \mu\} = 1 \tag{1.48}$$

### 1.6.2 中心极限定理

**定理 1.3** 中心极限定理
若$X_1, X_2, \cdots$独立同分布，且都有均值$\mu$和方差$\sigma^2$，则：

$$\lim_{n\to\infty} P\left\{\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \le a\right\} = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\mathrm{d}x \tag{1.49}$$

大数定理告诉我们只要采样n足够大就能逼近期待值；而中心极限定理说明了如果采样n逼近无穷，会呈现一个正态分布。这也是正态分布在统计学上非常重要的原因之一。大数定理和中心极限定理是概率论与数理统计的桥梁，当我们样本足够大的时候可以反映一个事物的统计性质，这也为下面从随机抽样到数理统计建设起一条桥梁。

### 1.6.3 随机抽样

我们在进行随机试验的时候，很多情况下可以用数字表示，或者是可以量化的结果。在随机试验中，我们把所有可能观察值称为**总体**，而每一个观察值称为**个体**；总体中包含的个体数目称为**容量**：根据容量的有限或者无限，又可以区分为**有限总体**和**无限总体**。如果我们在一个分布F中抽取，我们可以定义样本：

**定义 1.14** 样本 设$X$是从分布函数$F$中抽取的随机变量。若$X_1, X_2, \cdots, X_n$是具有同一分布函数$F$（或概率密度$f$），相互独立的随机变量，则称为$X_1, X_2, \cdots, X_n$为$u$分布函数$F$得到的容量为$n$的简单随机样本，将成为样本。$n$个随机样本的值$x_1, x_2, \cdots, x_n$称为样本值。

因此样本$X_1, X_2, \cdots, X_n$的概率密度为：

$$f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) \tag{1.50}$$

## 1.7 参数估计

### 1.7.1 点估计

### 1.7.2 最大似然估计

### 1.7.3 区间估计

### 1.7.4 置信区间

## 1.8 假设检验

假设检验的出发点是：小概率事件不可能发生。由于我们对一个总体数据没有办法进行完全抽样统计，只能抽出一部分样本来估计总体情况。因此假设检验就是提出一个**原假设**$H_0$，然后对其假设进行统计推断，做出接受假设$H_0$还是拒绝假设$H_0$。首先假设$H_0$是正确的时，若抽样得到的样本d导致了小概率事件的发生，则拒绝原假设$H_0$，否之接受假设$H_0$。

当接受假设$H_0$时，则拒绝假设$H_1$；当拒绝假设$H_0$时，则为接受假设$H_1$。由于我们的假设不可能是完全正确的，这时候就会出现两类错误：第一类错误是**弃真错误：当假设$H_0$真时拒绝原假设**；第二类错误是**取伪错误：当假设$H_0$假时接受原假设**。

### 1.8.1 显著性检验

# Chapter 2

# 贝叶斯统计

## 2.1 贝叶斯定理

贝叶斯定理和贝叶斯参数估计是在计算天文学课上的笔记，缺少例子解释。这部分尽量找一些比较容易理解的例子。上来就直接讲贝叶斯定理和贝叶斯参数估计是很晦涩难懂的。

贝叶斯公式首先我们回忆下条件概率的定义，条件B下A的概率为：

$$P(A|B) = \frac{P(AB)}{P(B)}$$

我们反过来，在A条件下B的概率为：

$$P(B|A) = \frac{P(AB)}{P(A)}$$

我们发现A,B的联合概率分布是相等量。因此移项可得：

$$P(A|B)P(B) = P(B|A)P(A)$$

如果我们很容易得到B条件下A的概率，那么我们要算A条件下B的概率就可以用上面的公式：

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

若有n个样本，则贝叶斯公式为：

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_{j=1}^{n} P(A|B_i)P(B_j)} = \frac{P(A|B_i)P(B_i)}{P(A)} \tag{2.1}$$

先验：$P(B)$

似然: $P(A|B)$

后验: $P(B|A)$

证据（归一化）: $P(A)$

这个公式是不是很难懂。在这里举个例子：有三个门，有一头牛两头羊。小明想养牛耕地。小明选择了其中一扇门，主持人会打开一个一定是羊的门，这时候小明可以改变自己的选择，玩家要不要改变自己的策略呢？

当第一次选择时候，三个门里面有牛的概率均为1/3，因此不换门获得牛的概率为1/3；开门后我后面得到牛的概率不是1/2吗？等等，这其实有点不对，因为主持人知道了三个门分别是什么。由于主持人一定会打开羊的门（主持人知道），如果换门有三种情况：

- 选择的是羊，换门得到牛
- 选择的是羊，换门得到牛
- 选择的是牛，换门得到羊

那这么算下来换门的获胜的概率是2/3！我们从贝叶斯的角度上看，当主持人没有开门的时候，假设我获得牛的概率为$P(cowA) = 1/3$; 当主持人打开门的时候，假设牛在A门，那主持人一定会选择B或C门打开,这时候主持人打开B门的概率为：$P(openB|cowA) = 1/2$；如果牛在C门后，那么主持人只会选择把B打开，$P(OpenB|cowC) = 1$；这时候我们可以计算主持人打开门B之后牛在A门的后验:

$$p(cowC|openB) = \frac{p(openB|cowA)p(cowA)}{p(openB|cowA)p(cowA) + p(openB|cowC)p(cowC)} = 1/3$$

由于牛在A门和C门属于互斥事件，因此在打开门B之后牛出现在C门的概率为2/3.因此换门是更好的选择。从感性上理解，由于主持人打开了B门，也就是假设出现在A门（我们选的门）时候的的概率p(cowA)是1/2；而假设牛出现的C门的概率能达到1，因此有没有牛的先验概率已经从主持人的行为中“学习”到了。贝叶斯定理的精髓在于让已经得到的信息表达在先验概率之中，这也非常像我们人脑学习知识的过程。比如我们发现云很多时候大下雨情况非常多，这就是我们的先验，我们可以把自己的认识加载到先验之中，达到学习的目的。

贝叶斯公式含义：通过数据推算模型参数的概率。即：

$$P(\text{Model}(\theta)|\text{Data}) = P(\text{Data}|\text{Model}(\theta))P(\theta) \tag{2.2}$$

贝叶斯统计的优势：将这个某种程度上是主观性的信息明确表达在先验概率中，而不是隐藏在没有明确指出的假设中; 让数据说话，减少主观性的先验概率。

## 2.2 贝叶斯单参数估计

### 2.2.1 二项分布估计

**无信息先验**

- 似然：

$$p(y|\theta) = \begin{pmatrix} n \\ y \end{pmatrix} \theta^y (1-\theta)^{n-y} \tag{2.3}$$

- 先验：均匀分布
- 后验：

$$p(\theta|y) \propto \theta^y (1-\theta)^{n-y} \sim Beta(y+1, n-y+1) \tag{2.4}$$

- 预测：

$$Pr(\widetilde{y}=1|y) = \int_0^1 Pr(\widetilde{y}=1|\theta, y) = \int_0^1 \theta p(\theta|y) d\theta = E(\theta|y) \tag{2.5}$$

**有信息先验**

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

$$p(y|\theta) = \begin{pmatrix} n \\ y \end{pmatrix} \theta^y (1-\theta)^{n-y} \tag{2.6}$$

- 后验：

$$p(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \sim Beta(\alpha+y, \beta+n-y)$$

- 后验期望：$E(\theta|y) = \frac{\alpha+y}{\alpha+\beta+n}$
- 先验期望：$E(y) = \frac{\alpha}{\alpha+\beta}$

  当$n \to \infty$: $E(\theta|y) \to y/n$

  数据很大的时候可以用正态分布近似后验分布

### 2.2.2 正态分布参数估计

1. **已知方差求均值**

- 似然：

$$p(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y-\theta)^2\right] \sim N(\theta, \sigma^2) \tag{2.7}$$

- 先验：

$$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right) \sim N(\mu_0, \tau_0^2) \tag{2.8}$$

- 后验：

$$p(\theta|y) \propto \exp\left(-\frac{1}{2\tau_1^2}(\theta-\mu_1)^2\right) \sim N(\mu_1, \tau_1) \tag{2.9}$$

其中：

$$\mu_1 = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{1}{\sigma^2}y}{\frac{1}{\tau_0^2} + \frac{1}{\sigma^2}}$$

$$\frac{1}{\tau_1^2} = \frac{1}{\tau_0^2} + \frac{1}{\sigma^2}$$

精度：方差的倒数

- 预测：

$$\begin{aligned} p(\widetilde{y}|y) &= \int p(\widetilde{y}|\theta)p(\theta|y)d\theta \\ &\propto \int \exp\left(-\frac{(\widetilde{y}-\theta)^2}{2\sigma^2}\right)\exp\left(-\frac{(\theta-\mu_1)^2}{2\tau_1^2}\right)d\theta \\ E(\widetilde{y}|\theta) &= \theta \\ D(\widetilde{y}|\theta) &= \sigma^2 + \tau_1^2 \end{aligned} \tag{2.10}$$

- 多个相互独立数据：

$$
\begin{aligned}
p(\theta \mid y) &\propto p(\theta) p(y \mid \theta) \\
&= p(\theta) \prod_{i=1}^{n} p\left(y_i \mid \theta\right) \\
&\propto \exp\left(-\frac{1}{2\tau_0^2}\left(\theta-\mu_0\right)^2\right) \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}\left(y_i-\theta\right)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\left[\frac{1}{\tau_0^2}\left(\theta-\mu_0\right)^2+\frac{1}{\sigma^2} \sum_{i=1}^{n}\left(y_i-\theta\right)^2\right]\right)
\end{aligned} \tag{2.11}
$$

可得：

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

其中：

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

$$
\frac{1}{\tau_n^2} = \frac{1}{\tau_0^2} + \frac{n}{\sigma^2}
$$

若$n \to +\infty$，$\tau_0$不变，则：$\theta|y \sim N(\bar{y}, \sigma^2/n)$

若$\tau \to +\infty$，$n$ 不变，则：$\theta|y \sim N(\bar{y}, \sigma^2/n)$

1. **已知均值求方差**

由于有$n$个服从$\sim N(\theta, \sigma^2)$的分布，因此：

- 似然：

$$
p(y|\sigma^2) \propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right) = (\sigma^2)^{-n/2} e^{-\frac{n}{2\sigma^2}v} \tag{2.12}
$$

其中：

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

- 共轭先验：$p(\sigma^2) \propto (\sigma^2)^{-\alpha+1} e^{\beta/\sigma^2} \sim \text{Inv} - \chi^2(\upsilon_0, \sigma_0^2)$
- 后验：

$$
\begin{aligned}
p(\sigma^2|y) &\propto p(\sigma^2)p(y|\sigma^2) \\
&\propto \left(\frac{\sigma_0^2}{\sigma^2}\right)^{\upsilon_0/2+1} \exp\left(-\frac{\upsilon_0\sigma_0^2}{2\sigma^2}\right) \cdot (\sigma^2)^{-n/2} \exp\left(-\frac{n}{2}\frac{\upsilon}{\sigma^2}\right) \\
&\propto (\sigma^2)^{-((n+\upsilon_0)/2+1)} \exp\left(-\frac{1}{2\sigma^2}(\upsilon_0\sigma_0^2 + n\upsilon)\right) \\
&\sim \text{Inv} - \chi^2(\upsilon_0 + n, \frac{\upsilon_0\sigma_0^2 + n\upsilon}{\upsilon_0 + n})
\end{aligned}
\tag{2.13}
$$

## 2.3 贝叶斯多参数模型

### 2.3.1 多参数模型处理

1. **置之不理**

2. **边缘化**

$$
p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2 \tag{2.14}
$$

用贝叶斯公式展开：

$$
p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)
$$

将$\theta_2$边缘化积分，得到$\theta_1$的后验分布。

3. **平均化**

$$
p(\theta_1|y) = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2
$$

### 2.3.2 无信息先验的正态分布

- 先验：

$$
p(\mu, \ln\sigma^2) \sim U(\mu, \ln\sigma^2) \tag{2.15}
$$

或者先验写为：

$$
p(\mu, \sigma^2) \sim \frac{1}{\sigma^2}
$$

- 似然（**有$n$次观测**）：

$$p(y|\mu,\sigma^2)=\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right)$$

- 联合后验：

$$p(\mu,\sigma^2|y)\propto\sigma^{-n-2}\exp\left(-\frac{1}{2\sigma^2}[(n-1)s^2+n(\bar{y}-\mu)^2]\right)$$

其中：

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

- 边缘后验$p(\sigma^2|y)$：

$$\begin{aligned}p(\sigma^2|y)&\propto\int p(\mu,\sigma^2|y)d\mu\\&\propto(\sigma^2)^{-\frac{n+1}{2}}\exp\left(-\frac{(n-1)s^2}{2\sigma^2}\right)\\&\sim\text{Inv}-\chi^2(n-1,s^2)\end{aligned}\tag{2.16}$$

- 边缘后验$p(\mu|y)$：

$$\begin{aligned}p(\mu|y)&=\int_0^\infty p(\mu,\sigma^2|y)d\sigma^2\\&\propto\left[1+\frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2}\\&\sim t_{n-1}(\bar{y},s^2/n)\end{aligned}$$

- 预测后验分布

$$p(\widetilde{y}|y)=\iint p(\widetilde{y}|\mu,\sigma^2,y)p(\mu,\sigma^2|y)d\mu d\sigma^2$$

### 2.3.3 共轭先验分布

- 先验分布：

$\mu|\sigma^2\sim N(\mu_0,\sigma^2/\kappa_0)$

$\sigma^2\sim\text{Inv}-\chi^2(\nu_0,\sigma_0^2)$

- 联合先验分布

$$\begin{aligned} p(\mu, \sigma^2) &= p(\mu|\sigma^2)p(\sigma^2) \\ &\propto N - Inv - \chi^2(\mu_0, \sigma_0^2/\kappa_0 \ ; \ \nu_0, \sigma_0^2) \end{aligned}$$

- 似然分布

$$p(y|\mu, \sigma^2) = \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right)$$

- 联合后验分布

$$\begin{aligned} p\left(\mu, \sigma^2 \mid y\right) &\propto \sigma^{-1}\left(\sigma^2\right)^{(\nu_0/2+1)} e^{-\frac{1}{2\sigma^2}\left[\nu_0\sigma_0^2 + \kappa_0(\mu-\mu_0)^2\right]}\left(\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\bar{y}-\mu)^2\right]} \\ &\propto \mathrm{N} - \mathrm{Inv} - \chi^2\left(\mu_n, \sigma_n^2/\kappa_n; \nu_n, \sigma_n^2\right) \end{aligned} \tag{2.17}$$

其中参数为：

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\mu_0 + \frac{n}{\kappa_0 + n}\bar{y} \tag{2.18a}$$

$$\kappa_n = \kappa_0 + n \tag{2.18b}$$

$$\nu_n = \nu_0 + n \tag{2.18c}$$

$$\nu_n\sigma_n^2 = \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}\left(\bar{y} - \mu_0\right)^2 \tag{2.18d}$$

- 条件后验分布$p(\mu|\sigma^2, y)$

$$\left(\mu \mid \sigma^2, y\right) \sim \mathrm{N}\left(\mu_n \frac{\sigma^2}{\kappa_n}\right)$$

- 方差边缘后验分布$p(\sigma^2|y)$

$$\left(\sigma^2 \mid y\right) \sim \mathrm{Inv} - \chi^2\left(\nu_n, \sigma_n^2\right)$$

- 均值边缘后验分布$p(\mu|y)$

$$\begin{aligned} p(\mu \mid y) &\propto \left[1 + \frac{\kappa_n\left(\mu - \mu_n\right)^2}{\nu_n\sigma_n^2}\right]^{-(\nu_n+1)/2} \\ &= t_{\nu_n}\left(\mu \mid mu_n, \sigma_n^2/\kappa_n\right) \end{aligned} \tag{2.19}$$

## 2.4 层次化贝叶斯模型

### 2.4.1 参数化先验分布

1. 先验分布[1]

   先验分布是由某个未知参数的分布$\phi$给出：

$$p(\theta|\phi)=\prod_{j=1}^{J}p(\theta_j|\phi)\qquad \theta=(\theta_1,\theta_2,\cdots)$$

   边缘化：

$$p(\theta)=\int\left(\prod_{j=1}^{J}p(\theta_j|\phi)\right)p(\phi)d\phi$$

2. 联合先验分布 $p(\phi,\theta)=p(\theta|\phi)p(\phi)$

3. 超先验分布：$p(\phi)$

4. 后验分布

   - 联合后验：$p(\phi,\theta|y)\propto p(y|\theta)p(\theta|\phi)p(\phi)$
   - 条件后验：$p(\theta|\phi,y)$
   - 边缘后验：$p(\phi|y)$

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

5. 层次化贝叶斯完整表述

$$\begin{aligned}p(\phi,\theta|y)&\propto p(y|\phi,\theta)p(\phi,\theta)\\&=p(y|\theta)p(\phi,\theta)\\&=p(y|\theta)p(\theta|\phi)p(\phi)\end{aligned}\tag{2.20}$$

   由于likelihood中$p(y|\phi,\theta)$只取决于$\theta$，因此超参数$\phi$只通过参数$\theta$影响y。

6. 层次化贝叶斯计算步骤

   - 写出联合后验分布$p(\theta,\phi|y)$：即超先验分布，总体分布和似然分布的乘积
   - 确定条件后验分布$p(\theta|\phi,y)$：

$$p(\theta|\phi,y)=\prod_{j=1}^{J}p(\theta_j|\phi,y)$$

- 边缘化给出$\phi$的贝叶斯估计

### 2.4.2 二项分布的分层贝叶斯模型

1. $y_i$先验（组内模型） $y_j\sim Bin(n_j,\theta_j)$
2. $\theta_j$先验（组间模型）：$\theta_j\sim Beta(\alpha,\beta)$
3. 联合先验： $p(\alpha,\beta,\theta)=p(\alpha,\beta)p(\theta|\alpha,\beta)$
4. 似然： $p(y|\theta,\alpha,\beta)$
5. 联合后验： $p(\theta,\alpha,\beta|y)$

$$\begin{aligned}p(\theta,\alpha,\beta|y)&\propto p(\alpha,\beta)p(\theta|\alpha,\beta)p(y|\theta,\alpha,\beta)\\&=p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}(1-\theta_j)^{\beta-1}\prod_{j=1}^{J}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j}\end{aligned}\tag{2.21}$$

6. 条件后验： $p(\theta|\alpha,\beta,y)$ ：单参数模型给定的后验

$$\begin{aligned}p(\theta\mid\alpha,\beta,y)&=\prod_{j=1}^{J}\frac{\Gamma\left(\alpha+\beta+n_j\right)}{\Gamma\left(\alpha+y_j\right)\Gamma\left(\beta+n_j-y_j\right)}\theta_j^{\alpha+y_j-1}\left(1-\theta_j\right)^{\beta+n_j-y_j-1}\\&\sim\prod_{j=1}^{J}Beta(\alpha+y_j,\beta+n_j-y_j)\end{aligned}\tag{2.22}$$

7. 边缘后验：$p(\alpha,\beta|y)$

$$\begin{aligned}p(\alpha,\beta\mid y)=&\frac{p(\theta,\alpha,\beta\mid y)}{p(\theta\mid\alpha,\beta,y)}\\&\propto\frac{p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_j^{\alpha-1}\left(1-\theta_j\right)^{\beta-1}\prod_{j=1}^{J}\theta_j^{y_j}\left(1-\theta_j\right)^{n_j-y_j}}{\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta+n_j)}{\Gamma(\alpha+y_j)\Gamma(\beta+n_j-y_j)}\theta_j^{\alpha+y_j-1}\left(1-\theta_j\right)^{\beta+n_j-y_j-1}}\\=&p(\alpha,\beta)\prod_{j=1}^{J}\frac{\Gamma(\alpha+\beta)\Gamma\left(\alpha+y_j\right)\Gamma\left(\beta+n_j-y_j\right)}{\Gamma(\alpha)\Gamma(\beta)\Gamma\left(\alpha+\beta+n_j\right)}\end{aligned}\tag{2.23}$$

### 2.4.3 正态分布的分层贝叶斯模型

1. 数据结构

假设$J$个独立试验，每个实验都由$\theta_j$给出其参数估计，估计$n_j$个$i.i.d$正态分布的数据点$y_{ij}$，每个点方差为$\sigma^2$：

$$y_{ij}|\theta_j \sim N\left(\theta_j, \sigma^2\right) \quad i = 1, \ldots, n_j \quad j = 1, \ldots, J$$

样本均值（充分统计量）：

$$\bar{y}_{\cdot j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$$

样本均值的分布

$$\bar{y}_{\cdot j} \sim N(\theta_j, \sigma_j^2)$$

样本方差：

$$\sigma_j^2 = \frac{\sigma^2}{n_j}$$

样本的均值是从$\theta$中估计，样本方差是从$\sigma$中估计。

相当于 $\theta$ 的似然分布：

$$\bar{y}_{\cdot j}|\theta \sim N(\theta_j, \sigma_j^2)$$

由于 $\sigma$ 是已知的，下面所有的分布都是在 $\sigma$ 已知情况下成立。

2. **层次化模型：无信息先验**

$\theta$是从参数$(\mu, \tau)$中抽取：

$$p\left(\theta_1, \ldots, \theta_J \mid \mu, \tau^2\right) = \prod_{j=1}^{J} p\left(\theta_j \mid \mu, \tau^2\right)$$

边缘化：

$$p\left(\theta_1, \ldots, \theta_J\right) = \iint \prod_{j=1}^{J} \left[p\left(\theta_j \mid \mu, \tau^2\right)\right] p\left(\mu, \tau^2\right) d\mu d\tau$$

- **先验和似然**

组内抽样：$\bar{y}_{\cdot j} \sim N(\theta_j, \sigma_j^2)$

$\theta$的先验（组内模型）：$\theta|\mu,\tau \sim N(\mu,\tau^2)$

$\mu$的先验： $p(\mu,\tau) = p(\mu|\tau)p(\tau) \propto p(\tau)$

$\theta_j$似然分布：$p(y|\theta) \sim (\bar{y}_{\cdot j}|\theta) \sim N(\theta_j, \sigma_j^2)$

- **联合后验**：$p(\theta,\phi|y)$

$$\begin{aligned} p(\theta,\mu,\tau|y) &\propto p(\mu,\tau)p(\theta|\mu,\tau)p(y|\theta) \\ &\propto p(\mu,\tau)\prod_{j=1}^{J} p(\theta_j|\mu,\tau^2)\prod_{j=1}^{J} p(\bar{y}_{\cdot j}|\theta_j,\sigma_j^2) \end{aligned} \tag{2.24}$$

其中：$\bar{y}_{\cdot j} \sim N(\theta_j, \sigma_j^2)$

可以忽略只依赖 $y$ 和 $\sigma_j$ 的参数，因为其已知。

- **$\theta$条件后验**：$p(\theta_j \mid \mu,\tau,y_{\cdot,j})$

$$\theta_j \mid \mu,\tau,y_{\cdot,j} \sim N\left(\hat{\theta}_j, V_j\right)$$

其中：

$$\hat{\theta}_j = \frac{\frac{1}{\sigma_j^2}\bar{y}_{\cdot j} + \frac{1}{\tau^2}\mu}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

$$V_j = \frac{1}{\frac{1}{\sigma_j^2} + \frac{1}{\tau^2}}$$

- **超参数边缘后验**：$p(\mu,\tau \mid y)$

$$p(\mu,\tau \mid y) \propto p(\mu,\tau)p(y \mid \mu,\tau)$$

对于正态分布：

$$\bar{y}_{\cdot j} \sim N(\mu, \tau^2 + \sigma^2)$$

因此：

$$p(\mu,\tau \mid y) \propto p(\mu,\tau)\prod_{j=1}^{J} p\left(\bar{y}_{\cdot j} \mid \mu, \tau^2 + \sigma_j^2\right) \tag{2.25}$$

$$\bar{y}_{\cdot j} \mid \mu, (\tau^2 + \sigma_j^2) \sim N\left(\mu, \tau^2 + \sigma_j^2\right) \tag{2.26}$$

- **给定$\tau$下$\mu$的边缘后验分布：** $p(\mu \mid \tau, y)$从单参数模型中得出的结论

$$\mu \mid \tau, y \sim N\left(\hat{\mu}, V_\mu\right) \tag{2.27}$$

$$\hat{\mu} = \frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2} \bar{y}_{\cdot j}}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}$$

$$V_\mu^{-1} = \sum_{j=1}^{J} \frac{1}{\sigma_j^2 + \tau^2}$$

- $\tau$的后验：$p(\tau \mid y)$

$$\begin{aligned} p(\tau \mid y) &= \frac{p(\mu, \tau \mid y)}{p(\mu \mid \tau, y)} \\ &\propto \frac{p(\tau) \prod_{j=1}^{J} N\left(\bar{y}_{\cdot j} \mid \mu, \sigma_j^2 + \tau^2\right)}{N\left(\mu \mid \hat{\mu}, V_\mu\right)} \\ &\propto p(\tau) V_\mu^{1/2} \prod_{j=1}^{J} \left(\sigma_j^2 + \tau^2\right)^{-1/2} \exp\left(-\frac{\left(\bar{y}_{\cdot j} - \hat{\mu}\right)^2}{2\left(\sigma_j^2 + \tau^2\right)}\right) \end{aligned} \tag{2.28}$$

从后往前采样，即先算出$\tau$的后验，然后依次采样算出$\mu$的后验，超参数联合后验，$\theta$的条件后验，联合后验等。

## 2.5 贝叶斯回归

### 2.5.1 线性贝叶斯回归

对于有n组观测数据，一般记为：

$$y = (y_1, y_2, \cdots, y_n)^T \tag{2.29}$$

解释变量记为：

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \tag{2.30}$$

我们看到一共有n行，也就是一共采样到n组数据；有k列，也就是说一组解释数据有(k-1)个变量。一般把第一列设置为1，因为我们在进行线性回归时有一个常数。回归系数$\beta = (\beta_1, \beta_2, \cdots, \beta_k)^T$。这样回归

方程可以写为：

$$E(y_i|\beta, X) = \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \qquad i = 1, \cdots, n \tag{2.31}$$

用矩阵表示为：

$$E(y|\beta, X) = X\beta \tag{2.32}$$

对于常规的线性回归情形，一般的我们认为给定的回归残差为一个正态分布，其回归方差为：

$$D(y_i|\theta, X) = \sigma^2$$

对于所有的i组数据都相同。因此我们的回归参数$\theta = (\beta, \sigma^2)$我们记：

$$p(y|\beta, \sigma^2, X) \sim N(X\beta, \sigma^2 I) \tag{2.33}$$

其中I是一个$n \times n$的单位矩阵。对于正态回归模型,我们的先验可以写为：

$$p(\beta, \sigma^2|X) \sim \sigma^{-2} \tag{2.34}$$

### 2.5.2 更一般的贝叶斯回归

### 2.5.3 有先验信息的贝叶斯回归

## 2.6 贝叶斯模型选择

**定义 2.1** 贝叶斯因子
在观测数据$d$中选择两个模型假设$H_1$和$H_2$，其参数为$\theta_1$和$\theta_2$，贝叶斯因子为：

$$\mathcal{B}_{12} = \frac{p(d|H_1(\theta_1))}{p(d|H_2(\theta_2))} \tag{2.35}$$

下面把贝叶斯因子（*Bayes factor*）缩写为$BF$。

当$\mathcal{B}_{12} > 1$时支持假设$H_1$，反之支持假设$H_2$。当然BF也有相应的判据，下面会慢慢介绍。

**定义 2.2** *odds ratio*
对于事件$E$和事件$E$的补集$E^c$，其概率比值：

$$\mathcal{O}(E) = \frac{P(E)}{P(E^c)} \tag{2.36}$$

称为*odds ratio*（下面缩写为$OR$）。从数据分析的角度上说，有两个模型假设$H_1$和$H_2$，$OR$为：

$$\mathcal{O}_{12} = \frac{p(H_1|d)}{p(H_2|d)} \tag{2.37}$$

OR和BF的关系是什么呢？结合BF公式2.35我们把OR展开：

$$\begin{aligned}\mathcal{O}_{12} &= \frac{p(H_1|d)}{p(H_2|d)} \\ &= \frac{p(d|H_1)p(H_1)}{p(d|H_2)p(H_2)} \\ &= \text{BF} \cdot \frac{p(H_1)}{p(H_2)} \\ &= \text{BF} \cdot \text{prior odds}\end{aligned} \tag{2.38}$$

因此Odds ratio是Bayes factor乘先验比。

对于模型选择，有一个非常著名的定理：**奥卡姆剃刀**。这个定理的概括来说就是：**如无必要，勿增实体**。

## 2.7 费舍尔信息矩阵Fisher Information

$$\Gamma_{ij}(\theta) := E_\theta \left\{ \frac{\partial \ln p(x;\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln p(x;\boldsymbol{\theta})}{\partial \theta_j} \right\} \tag{2.39}$$

# Chapter 3

# 傅里叶变换

# Chapter 4

# 随机过程

# Chapter 5

# MCMC与采样方法

## 5.1 蒙特卡罗法 Monte Carlo Method

### 5.1.1 随机采样和接受-拒绝采样

蒙特卡罗法是通过概率模型的随机抽样进行随机抽样的方法。假设概率分布已知，通过概率分布得到随机样本，并通过得到的随机样本得到概率分布的随机性质，因此蒙特卡洛方法的核心是随机抽样。接下来我们介绍**接受-拒绝采样**。

已知概率密度分布为$f(x)$，但是这个概率密度分布复杂，各个变量并不独立，无法直接采样或者积分，因此可以通过蒙特卡罗方法进行抽样，得到样本$X$，得到其随机分布。我们在这介绍**接受-拒绝采样**。我们需要一个辅助的**建议分布**，记为$q(x)$。这个建议分布可以产生我们的候选样本，但建议分布要满足：

$$c * q(x) \geq f(x)$$

之后我们对样本按照建议分布$q(x)$进行抽样，得到样本$x^*$，同时对均匀分布$U(0,1)$进行抽样，得到$u$。之后计算$\frac{f(x)}{c*q(x)}$（这个值一定在0~1之间，对应图1的绿色部分比例），若

$$u \leq \frac{f(x^*)}{c * q(x^*)}$$

则$x^*$接受作为样本，否则拒绝。

怎么理解这个过程呢？简单来说就是我们先对建议分布$q(x)$的概率密度进行采样，因为这个比我们的$f(x)$更容易采样。假如这个分布很复杂，维度很高，直接算的话浪费计算资源，因此要先用一个简单的建议分布$q(x)$进行采样，得到建议的采样，但是这建议分布的采样终究不是我们需要的采样，所以我们需要在利用均匀分布$U(0,1)$，由我们算出来的$f(x^*)/cq(x^*)$进行接受或者拒绝。在图1中，按照

图 5.1:

绿色比例进行接受。如果在第$x_i^*$个抽样刚好落在中间红色区域比较大的点，那么拒绝的概率就高，反之绿色部分的比例更大，则我们接受的概率就越高。

听到这是不是有点迷糊了？别着急！我们看看图1，在是不是$x^*$处红色部分占比越大，与目标分布$f(x)$相差就越远了？所以我们在这里就必须剔除一些点了，不然远离我们的真实分布了！这时候可能会有其他疑问了，那在图1两端概率很小时候岂不是都接受率很高？是的，但是那两端概率很低呀！因为我们在使用建议分布抽样的时候概率那里的点已经是很少了，所以我们不用拒绝很多样本点也就和目标分布类似了。所以这时候就要用到一个均匀分布$U(0,1)$，在该点上随机生成一个$u$，然后按照(1.1)则接受，否则拒绝。

所以假设我们抽了n个样本，对样本进行拒绝，就是要生成和判断n次$u$的取值，也就是对每个样本点进行计算和判断是否拒绝，完成一次拒绝-接受采样。

接受-拒绝采样的缺点也是有的，主要是接受率比较低，抽样效率低。

### 5.1.2 数学期望和蒙特卡罗积分

如果我们要算目标函数为$f(x)$，其概率密度为$p(x)$，我们记函数$f(x)$关于密度函数$p(x)$的数学期望为$E_{p(x)}[f(x)]$。我们按照密度函数$f(x)$独立抽取n个样本$x_1, x_2 \cdots, x_n$，之后计算样本的均值：

$$\hat{f}_n = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

作为$f(x)$的近似值。根据大数定理，当样本容量增大，样本均值以概率1收敛于数学期望。因此我们可以用上述方法得到我们的数学期望。

$$E_{p(x)}[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

而在$\mathcal{X}$上数学期望的积分形式为：

$$E(x) = \int_{\mathcal{X}} f(x)p(x)dx$$

如果我们的目标积分为：

$$h(x) = f(x)q(x)$$

我们可以写为：

$$\int_{\mathcal{X}} h(x)dx = \int_{\mathcal{X}} f(x)q(x)dx = E_{p(x)}[f(x)]$$

对于复杂的函数，可以给定一个概率密度函数$p(x)$，只要取

$$f(x) = \frac{h(x)}{p(x)}$$

就可以用$p(x)$进行抽样算出积分：

$$\int_{\mathcal{X}} h(x)dx = E_{p(x)}[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

因此我们可以先用$p(x)$抽样。然后再进行积分。

## 5.2 MCMC原理

### 5.2.1 MCMC原理

我们简单介绍了蒙特卡罗方法和马尔可夫链，接下来我们介绍马尔可夫链蒙特卡罗方法，下面简称MCMC方法。MCMC方法适用于随机变量多元的、密度函数是非标准形式的、随机变量不相互独立的情况。若存在维数太高的情况，直接抽样是不可能的，因为其数据量在指数级增长，MCMC就可以避免维数爆炸的问题。

假设多元随机变量$x = [x_1, x_2, x_3, \cdots]$，满足$x \in \mathcal{X}$且其概率密度为$p(x)$，$f(x)$是定义在$x \in \mathcal{X}$上的函数，我们的目标是获得概率分布$p(x)$的抽样以及$f(x)$的数学期望$E_{p(x)}[f(x)]$。在随机变量x的状态空间$\mathcal{S}$上满意遍历定理（上一章2.2马尔科夫链性质）的马尔可夫链$X = \{X_0, X_1, \cdots, X_t, \cdots\}$，当这个马尔可夫链平稳时的分布就是其抽样的目标分布$p(x)$。

怎么通俗地解释这个原理呢？首先构造一个马尔可夫链，在状态空间$\mathcal{S}$上进行随机游走。根据遍历原理，总有一个时刻m之后，这个马尔可夫链近于平稳分布，也就是在期望附近游走。假设我们随机游走了n步，取我们平稳分布之后的样本集合$\{x_{m+1}, x_{m+2}, x_{m+3}, \cdots, x_n\}$，就是我们目标抽样分布的结果。

这里又要唠叨一句，平稳分布只是各个状态的期望。我们用天气由于预报的例子来说，假设当天气平稳分布的时候，其平稳分布为$[0.7\ 0.3]^T$，我们之后观测天气的结果符合我们这个平稳分布，也就是说我们接下来100天中有70天是晴天，30天是雨天。当然这只是一个简单的假设模型罢了！

所以当到达平稳之后样本集合为$\{x_{m+1}, x_{m+2}, x_{m+3}, \cdots, x_n\}$就是我们目标分布的结果。在时刻m之前的时期我们称为**燃烧期** (burn-in)。

更晕的还在后边，我们怎么构造这样一个马尔可夫链？这里我们需要一个转移核（连续）或者转移矩阵（离散）。如何构造这转移核/矩阵，构成一个可逆的马尔可夫链，使得遍历定理成立是很关键

的。如果该马尔可夫链成立，由于遍历定理成立，因此初始值的选取最终会收敛到同一平稳分布；燃烧期之前的样本都要丢弃，因为燃烧期之前的样本都不是服从样本的分布。当然目前MCMC收敛的判断是经验性的。

MCMC方法比拒绝-接受采样更容易实现，虽然丢弃了燃烧器之前的样本，但其效率仍然比拒绝-接受采样的效率高。目前常用的MCMC 方法主要是Metropolis-Hasting算法（M-H算法）和吉布斯抽样。

### 5.2.2 MCMC算法

根据上面的介绍，MCMC方法可以是以下的步骤：

1. 在随机变量$x$的状态空间$\mathcal{S}$上构造一个满足遍历定理的马尔可夫链，使得其平稳分布为$p(x)$；
2. 在状态空间某一点$x_0$出发，构造随机游走，产生样本$x_0, x_1, x_2, \cdots, x_t, \cdots$；
3. 应用遍历定理，确定燃烧期m，求得函数$f(x)$的均值

$$\hat{E}f = \frac{1}{n-m}\sum_{i=m+1}^{n} f(x_i)$$

在这有几个问题：

1. 如何定义马尔可夫链
2. 如何确定收敛步骤
3. 如何确定迭代步数确保精度

## 5.3 Metropolis-Hastings采样

### 5.3.1 M-H采样原理

上一章我们讲了MCMC抽样的一些问题，这一章我们介绍MCMC的一种代表算法。这一小节我们介绍M-H采样的原理。

我们需要构造一条马尔可夫链；要构造一个转移核，使得平稳分布就是我们要的抽样分布。参考第二章2.2节的细致平衡：

$$p_{ji}\pi_j = p_{ij}\pi_i$$

当然要构造这个细致平衡条件很难。假设**目标分布**为$\pi(x)$，转移核为$q(i,j)$，通常我们只能得到这样的结果：

$$\pi(i)q(i,j) \neq \pi(j)q(j,i)$$

因此我们要构造一个分布能使得其分布是细致平衡。

假设我们通过建议分布$q(i,j)$中随机抽取一个**候选状态**$x_j$**(后面简写为**$j$**)**，我们可以在两端乘以一个$\alpha(i,j)$使得其变为平稳分布，即：

$$\pi(i)q(i,j)\alpha(i,j) = \pi(j)q(j,i)\alpha(j,i)$$

在这里$\alpha(i,j)$为**接受分布**。建议分布$q(i,j)$是马尔可夫链转移核，且该马尔可夫链是不可约的，同时这个分布是容易采样的。那这个接受分布怎么构建呢？

我们对$\pi(i)q(i,j)\alpha(i,j) = \pi(j)q(j,i)\alpha(j,i)$ 移项：

$$\alpha(i,j) = \frac{\pi(j)(j,i)}{\pi(i)q(i,j)}\alpha(j,i)$$

实际上我们需要把两边的接受分布扩大到1，这样接受率才会达到最大。如果$\alpha(j,i) = 1$，有：

$$\alpha(i,j) = \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}$$

相反的，如果上式$\alpha(i,j) > 1$，我们令$\alpha(i,j) = 1$：

$$1 = \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}\alpha(j,i)$$

即我们取$\alpha(i,j) = 1$。这时候接受分布$\alpha$扩大到最大。因此得到接受分布：

$$\alpha(i,j) = \min\left\{1, \frac{\pi(j)q(j,i)}{\pi(i)q(i,j)}\right\}$$

这时候的接受率最高，且容易证明这个转移核$q(i,j)\alpha(i,j)$是平稳分布的。之后我们从区间(0fi1)中均匀采样，得到一个随机数$u$，按照以下判断：

$$x_t = \begin{cases} x_j, & u \leq \alpha(i,j) \\ x_i, & u > \alpha(i,j) \end{cases}$$

决定其是否接受下一步。

图 5.2:

怎么理解这个接受分布呢？前面提到，当我们从状态$i$以概率$p(i,j)$转移到状态$j$的时候，不一定是平稳的，而加入$\alpha$之后就可以得到一个新的平稳的马尔可夫链。而我们是否要转移到下一步就是要考虑这个接受分布了。我们定性的解释这个问题。假设我们的建议分布$q(i,j) = q(j,i)$，我们的接受分布$\alpha = \min\{1, \frac{\pi_j}{\pi_i}\}$，如果$\alpha$太小，说明我们下一步抽取的$j$所对应的概率是远小于我们上一步抽取的概率$i$，所以要舍弃；而如果是1的话说明抽取的$j$很符合我们的目标分布$\pi$，因此可以更好地接近我们要抽样的分布。这个判断步骤类似于本文开头的拒绝-接受采样。$\alpha$类似于接受-拒绝采样的$\frac{f(x)}{c*q(x)}$。综上，其实这个转移核是要根据我们的抽样$i,j$共同确定的。

### 5.3.2 M-H采样算法

常见的采样有ptmcmc[2]

## 5.4 吉布斯采样

### 5.4.1 满条件分布

MCMC要抽样的函数一般都是多变量的联合概率分布$p(x) = p(x_1, x_2, \cdots, x_k)$，其中$x = (x_1, x_2, \cdots, x_k)^T$是$k$维变量。若条件概率分布：$p(x_I|x_{-I})$中出现了所有的变量$k$，其中:

$$x_I = \{x_i, i \in I\},\ x_{-I} = \{x_i, i \notin I\} \qquad I \subset K = \{1, 2, \cdots, k\}$$

那么称这个分布为**满条件分布**。

### 5.4.2 Gibbs采样原理

当然M-H采样有接受分布的存在，因此效率还是不够高。Gibbs采样可以避免这个问题。吉布斯采样的基本原理是从满条件概率分布出发，从满条件概率分布中抽样，得到一个样本序列。基本原理是：吉布斯抽样过程是在一个马尔可夫链上随机游走，平稳分布就是目标联合分布。接下来我们介绍Gibbs采样的细节。

我们考虑二维情况：假设有一个二维分布$p(x,y)$，我们发现：

$$p(x_1, y_1)p(y_2|x_1) = p(x_1)p(y_1|x_1)p(y_2|x_1)p(x_1, y_2)p(y_1|x_1) = p(x_1)p(y_2|x_1)p(y_1|x_1)$$

整理得：

$$p(x_1, y_1)p(y_2|x_1) = p(x_1, y_2)p(y_1|x_1)$$

图 5.3:

假设点$A$为$(x_1, y_1)$，点$B$为$(x_1, y_2)$，我们可以改写为：

$$p(A)p(y_2|x_1) = p(B)p(y_1|x_1)$$

也就是说当点A转移到点B的时候服从上述的马尔可夫链。而上式的条件概率就是我们的转移矩阵或者转移核。也就是说对维度$y$的满条件分布即为上述马尔可夫链的转移核或者转移矩阵。如果把二维扩展到$n$维，即可以得到$n$维的吉布斯抽样。假设建议分布是当前变量$x_j,\ j = 1, 2, \cdots, k$（也就是抽取第$j$维变量）的满条件分布：$q(x', x) = p(x'_j|x_{-j})$，这里的$x$指的是当前的抽样，$x'$指的是下一步的抽样。扩展到维的情多维：

$$p(x'_j, x_{-j})p(x'_j|x_{-j}) = p(x_j, x_{-j})p(x_j|x_{-j})$$

这时候的接受率$\alpha = 1$。因此吉布斯采样可以认识是M-H采样的一种特殊情况。这个建议分布就是我们的目标分布的满条件分布。（这里和上一章的符号有所变化，不过内容是一样的，只是为了方便表达。）下面证明如何得到接受率$\alpha = 1$：

根据M-H采样的接受分布公式，有：

$$q(x, x') = p(x'_j|x_{-j})$$

代入接受分布：

$$\alpha(x, x') = min\left\{1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\right\}$$

因为我们是抽取当前变量$j$，因此有：

$$\begin{aligned}\alpha(x, x') &= min\left\{1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\right\} \\ &= min\left\{1, \frac{p(x'_j, x_{-j})p(x'_j|x_{-j})}{p(x_j, x_{-j})p(x_j|x_{-j})}\right\} = 1\end{aligned} \tag{5.1}$$

之后抽取$k$维，循环$n$次，抛去燃烧期$m$，得到我们的抽样，计算均值。

### 5.4.3 Gibbs采样算法

## 5.5 Nested采样

对于evidence来说要进行高维积分，这是一个非常大的开销，因此还有另外的比较常用的MCMC积分：MultiNest 算法。一个模型的参数空间的活动的点先填充先验；

## 5.6 Savage-Dickey density ratio

我们发现Bayes Factor的计算比较困难，需要多重积分，因此使用**Savage-Dickey density ratio**的方式估计Bayes Factor。

我们考虑一个包含假设$\mathcal{H}_1$和假设$\mathcal{H}_2$的嵌套的模型，这个模型有共同参数$\theta$。假设$\mathcal{H}_1$有自己特有的参数A，可以通过设置$A=0$得到假设$\mathcal{H}_2$：

$$p(d|A=0,\theta;\mathcal{H}_1)=p(d|\theta;\mathcal{H}_2) \tag{5.2}$$

当A=0时，后验密度为：

$$\begin{aligned} p(A=0|d;\mathcal{H}_1) &= \int p(A=0,\theta|d;\mathcal{H}_1)\mathrm{d}^n\theta \\ &= \int \frac{p(d|A=0,\theta;\mathcal{H}_1)p(A=0)p(\theta)}{p(d|\mathcal{H}_1)}\mathrm{d}^n\theta \\ &= \int \frac{p(d|\theta;\mathcal{H}_2)p(A=0)p(\theta)}{p(d|\mathcal{H}_1)}\mathrm{d}^n\theta \\ &= \frac{p(A=0)}{p(d|\mathcal{H}_1)}\int p(d|\theta;\mathcal{H}_2)p(\theta)\mathrm{d}^n\theta \\ &= \frac{p(d|\mathcal{H}_2)}{p(d|\mathcal{H}_1)}p(A=0) \end{aligned} \tag{5.3}$$

因此Bayes Factor为：

$$\mathcal{B}_{12}=\frac{p(d|\mathcal{H}_1)}{p(d|\mathcal{H}_2)}=\frac{p(A=0)}{p(A=0|d;\mathcal{H}_1)} \tag{5.4}$$

相应的就是A=0时的先验和归一化后验之比。在写程序的时候$A=0$会导致许多的错误，因此一般会把A设置成一个非常小的数，比如在pulsar timing的探测中，会使用$\log_{10}A_{\mathrm{low}}=-18$，这个振幅远远低于pulsar的固有噪声。

## 5.7 Product space sampling

很多时候我们给出的后验采样是很难采样到A=0时的后验概率分布，因此需要**Product space sampling**。原理是通过增加一个超参数$n$进行odds ratio的计算。我们假设有n个待选的模型$H_i$，其中$i\in\{1,\cdots,n\}$。我们把Bayes定理写为：

$$p(\theta|d,H_n)=\frac{p(d|\theta,H_n)p(\theta|H_n)}{p(d|H_n)}\equiv\frac{\mathcal{L}\pi}{\mathcal{Z}} \tag{5.5}$$

其中$\mathcal{L}$为likelihood function，$\pi$为先验分布，$\mathcal{Z}$为的evidence，可以写为：

$$\mathcal{Z} = \int_{\text{all }\theta} \mathcal{L}(\theta)\pi(\theta)\mathrm{d}\theta \tag{5.6}$$

对于我们的假设（或者说模型）$H_i, i \in 1, \cdots, n$的后验概率为：

$$\begin{aligned} p(H_i|d) &= \frac{p(d|H_i)p(H_i)}{p(d)} \\ &= \frac{\int_{\theta_i} p(d|\theta_i, H_i)p(\theta_i)\mathrm{d}\theta_i p(H_i)}{p(d)} \\ &= \frac{\mathcal{Z}_i \pi_{H_i}}{p(d)} \end{aligned} \tag{5.7}$$

根据OR的定义[2.37]，我们的后验ORs可以写为：

$$\ln \mathcal{O}_{ji} = \mathcal{P}_{ji} = \ln\left[\frac{p(H_j|d)}{p(H_i|d)}\right] = \ln\left(\frac{\mathcal{Z}_j}{\mathcal{Z}_i}\right) + \ln\left(\frac{\pi_{H_j}}{\pi_{H_i}}\right) \tag{5.8}$$

当$\pi_{H_i} = \pi_{H_j}$时，BF=OR。注意这里的ORs和前面定义的ORs是互为倒数的。

这样计算ORs或者BF的时候要计算evidence，但这个计算成本太高了，我们可以试着转换为我们熟悉的MCMC采样。我们的假设模型$H_n (n \in \{1, \cdots, N\})$可以考虑合并为一个hypermodel。对于模型$H_n$来说，下表n就像一个开关。而各个模型的参数$\theta_n$可以合并为一个参数向量$\theta$。对于模型选择参数n可以连续化，再整数化转为整数参数n。

一般的，有两个不同的模型假设，其参数向量：$\theta_n$和$\theta_{n'}$，一般这两个参数的维度不一样。如果有$\theta_n \subset \theta_{n+1}$这种情况，一般会考虑nested model，并使用reversible-jump Markov chain Monte Carlo (RJMCMC)实现不同维度空间的转换。这里介绍另外的方法。我们需要知道假设N的先验是已知的，对参数空间$\theta$加入一个新的参数n，得到整体参数空间$(\theta, n)$，这个参数空间的先验已经确定下来了。这样我们就可以使用传统的MCMC采样方法或者是nested sampling。对于任意给定的参数n，联合参数空间可以划分为假设$H_n$的参数$\theta_n$和不属于其假设的参数$\phi_n$。也就是说：

$$p(d|n, \theta) = p(d|n, \theta) \tag{5.9}$$

在我们规定的hypermodel $\mathcal{H}$下会将参数$\theta_n$传递到假设$H_n$；而剩下的参数$\phi_n$将会被忽略，在参数空间上赋予一个常数值。如果参数空间维度比较大的情况下，很可能出现参数重叠的情况，这里我们考虑nested 模型。

当参数空间$(\theta, n)$获得到一组后验分布时，我们可以计算归一化后验分布：

$$p(n|d, \mathcal{H}) = \int p(\theta, n|d, \mathcal{H})\mathrm{d}\theta = \frac{1}{\mathcal{Z}_{\mathcal{H}}} \int \mathcal{L}(\theta, n)\pi(\theta, n)\mathrm{d}\theta \tag{5.10}$$

$\mathcal{Z}_{\mathcal{H}}$是Hypermodel的evidence。先验可以写为：

$$\pi(\theta|n) = \pi(\theta_n|n)\pi(\phi_n)\pi(n) \tag{5.11}$$

为了让推导更加清晰，在这里展开得到关于参数空间$(\theta, n)$的后验分布：

$$p(n, \theta|d) = \frac{p(d|n, \theta_n)p(\theta_n|n)p(\phi_n|\theta_n, n)p(n)}{p(d)} \tag{5.12}$$

实际上分子第一项对应的是关于模型n的likelihood function；而分母项则是hypermodel的evidence $\mathcal{Z}_{\mathcal{H}}$。对参数$\theta$积分得到：

$$p(n|d, \mathcal{H}) = \frac{\pi(n)}{\mathcal{Z}_{\mathcal{H}}} \int \mathcal{L}(\theta_n)\pi(\theta_n|n)\mathrm{d}\theta \tag{5.13}$$

在这里我们利用先验的归一化：$\int \mathrm{d}\phi_n = 1$。我们注意到5.13是正是模型$\mathcal{H}_n$的evidence。因此我们可以得到：

$$\pi(n)\mathcal{Z}_n = \mathcal{Z}_{\mathcal{H}} p(n|d, \mathcal{H}) \tag{5.14}$$

因此ORs可以写为：

$$\ln \mathcal{O}_{ji} = \ln \left[\frac{p(n = j|d, \mathcal{H})}{p(n = i|d, \mathcal{H})}\right] \tag{5.15}$$

我们发现关于hypermodel的evidence就被约掉了，因此避免了繁杂的evidence的计算。这个方法的核心算法[3, 4]：

$$\theta_i \sim p(\theta_i|\phi_i, n, d) \propto \begin{cases} p(y|n, \theta_n)p(\theta_n|n), i = n \\ p(\theta_n|\phi_n, n), i \neq n, \end{cases} \tag{5.16}$$

$$k \sim p(n|\theta, d) \propto p(d|n, \theta_n)p(\theta_n|n)p(\phi_n|\theta_n, n)p(n) \tag{5.17}$$

当然在这里我们的evidence只是一个归一化常数，当我们计算ORs或者BF的时候，hypermodel的evidence会被除掉，因此可以不用计算evidence的积分，而采用传统的MCMC采样方法。

# Chapter 6

# 高斯随机过程

## 6.1 高斯随机过程及其统计描述

## 6.2 协方差函数与核Kernel

## 6.3 高斯混合模型

## 6.4 高斯学习

# Chapter 7

# 统计算法

# Chapter 8

# Gaussian process and likelihood model

Now that we have defined all of our noise terms, we will write down the likelihood function and go over some computational tricks to make it tractable.

First we will assume (and epmirical studies show this is a good assumption) that the measurement uncertainties are gaussian distributed with covariance matrix $N$. If we not write the residuals using the information above then we have

首先我们假设测量误差是一个协方差为 $N$ 的正态分布，时间残差$\delta\tau$可以写为：[5]

$$\delta\tau = M\epsilon + F_{\rm red}a_{\rm red} + M_{\rm DM}d_{\rm DM} + F_{\rm DM}a_{\rm DM} + Uj + n \tag{8.1}$$

这里$n$是白噪声项，可以写为$\langle nn^T\rangle = N$

where $n$ is the white noise with $\langle nn^T\rangle = N$.

There are two ways to think about the likelihood function, the Basis picture and the Kernel picture. In the Basis picture we explicitly model the residuals as above and then impose Gaussian priors on the amplitudes $\{\epsilon, a_{\rm red}, d_{\rm DM}, a_{\rm DM}, j\}$. In the kernel picture we just model the residuals using the full covariance matrix defined by the white noise and these Gaussian priors. In the Bayesian sense, the Kernel comes from marginalizing the basis picture likelihood over the basis function amplitudes.

贝叶斯方法先确定一个噪声模型，对于$\{\epsilon, a_{\rm red}, d_{\rm DM}, a_{\rm DM}, j\}$做高斯先验假设构建likelihood function。Kernel方法使用高斯先验的白噪声及其协方差矩阵构建likelihood function。

Now we have shown that there are certain physical or statistical reasons to assume certain Gaussian priors on $\{a_{\rm red}, a_{\rm DM}, j\}$; however, we have no such priors for the timing model or quadratic DM

parameters and would like to use uniform priors. This can be accomodated by using Gaussian priors with infinite variance (it works!, you'll see). With that in mind let us simplify the notation as

$\{a_{\rm red}, a_{\rm DM}, j\}$使用高斯先验分布有物理和统计上的原因。然而对于timing model或者是quadratic DM parameters我们没有类似的的先验分布，因此我们使用先验分布，先验分布可以使用一个方差无穷大的高斯先验来实现。这样可以简化我们的先验分布：

$$T = [M\ F_{\rm red}\ M_{\rm DM}\ F_{\rm DM}\ U]; \qquad b = \begin{bmatrix} \epsilon \\ a_{\rm red} \\ d_{\rm DM} \\ a_{\rm DM} \\ j \end{bmatrix} \qquad B = \begin{bmatrix} \infty & & & & \\ & \varphi & & & \\ & & \infty & & \\ & & & \varphi_{\rm DM} & \\ & & & & J \end{bmatrix}$$

and the residuals and covariance matrix are now

残差和协方差矩阵可以表示为：

$$\delta\tau = Tb + n$$

$$C = N + K = N + TBT^T$$

We define the vector of paramters that characterize the Gaussian priors (such as amplitude and spectral index of red noise or DM spectrum, and variance of ECORR) as $\phi$, so the matrix $B = B(\phi)$ and $N = N(\phi)$. In this case the likelihoods are

这里我们可以写出我们的参数向量。描述高斯先验分布的参数向量定义为$\phi$，以及矩阵$B = B(\phi)$ and $N = N(\phi)$

* Basis Picture:

$$p(\delta\tau|b,\phi) = \frac{\exp\left[-\frac{1}{2}(\delta\tau - Tb)^T N^{-1}(\delta\tau - Tb)\right]}{\sqrt{\det 2\pi N}} \frac{\exp\left[-\frac{1}{2}b^T B^{-1} b\right]}{\sqrt{\det 2\pi B}}$$

* Kernel Picture

$$p(\delta\tau|\phi) = \frac{\exp\left[-\frac{1}{2}\delta\tau^T (N+K)^{-1}\delta\tau\right]}{\sqrt{\det 2\pi(N+K)}}$$

In the Kernel picture we use the Woodbury Lemma to compute the inverse and determinant of $(N + TBT^T)$

$$\begin{aligned} (N + TBT^T)^{-1} &= N^{-1} - N^{-1}T\left(B^{-1} + T^T N^{-1} T\right)^{-1} T^T N^{-1} \\ \det\left(N + TBT^T\right) &= \det(N)\det(B)\det\left(B^{-1} + T^T N^{-1} T\right) \end{aligned} \tag{8.2}$$

In case you were worried about our infinite variance trick notice that $\det(B) = 1/\det(B^{-1})$ and all other dependence on $B$ comes in the form of an inverse.

For a given pulsar and noise model $B$ may be a $1,000 \times 1,000$ matrix whereas the full covariance matrix (N+K) could be on the order of $30,000 \times 30,000$. This means that we cut the computation time by over a factor of 1000!

# 参考文献


[1] Andrew Gelman, John B Carlin, Hal Steven Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. 2014. OCLC: 1063654237.

[2] Justin Ellis and Rutger van Haasteren. jellis18/ptmcmcsampler: Official release, October 2017.

[3] Simon J Godsill. On the Relationship Between Markov chain Monte Carlo Methods for Model Uncertainty. *Journal of Computational and Graphical Statistics*, 10(2):230–248, June 2001.

[4] S. Hee, W. J. Handley, M. P. Hobson, and A. N. Lasenby. Bayesian model selection without evidences: application to the dark energy equation-of-state. *Monthly Notices of the Royal Astronomical Society*, 455(3):2461–2473, January 2016.

[5] Z. Arzoumanian, A. Brazier, S. Burke-Spolaor, S. J. Chamberlin, S. Chatterjee, B. Christy, J. M. Cordes, N. J. Cornish, K. Crowter, P. B. Demorest, X. Deng, T. Dolch, J. A. Ellis, R. D. Ferdman, E. Fonseca, N. Garver-Daniels, M. E. Gonzalez, F. Jenet, G. Jones, M. L. Jones, V. M. Kaspi, M. Koop, M. T. Lam, T. J. W. Lazio, L. Levin, A. N. Lommen, D. R. Lorimer, J. Luo, R. S. Lynch, D. R. Madison, M. A. McLaughlin, S. T. McWilliams, C. M. F. Mingarelli, D. J. Nice, N. Palliyaguru, T. T. Pennucci, S. M. Ransom, L. Sampson, S. A. Sanidas, A. Sesana, X. Siemens, J. Simon, I. H. Stairs, D. R. Stinebring, K. Stovall, J. Swiggum, S. R. Taylor, M. Vallisneri, R. van Haasteren, Y. Wang, W. W. Zhu, and The NANOGrav Collaboration. THE NANOGRAV NINE-YEAR DATA SET: LIMITS ON THE ISOTROPIC STOCHASTIC GRAVITATIONAL WAVE BACKGROUND. *The Astrophysical Journal*, 821(1):13, April 2016.