统计计算

贾金柱

Contents

前言

章节号	章节名	学时
第一章	均匀随机数的产生和检验	4
第二章	常用分布函数和分位数的计算	8
第三章	非均匀随机数的产生	4
第四章	随机模拟方法	4
第五章	EM算法	8
第六章	MCMC	8
第七章	Bootstrap 方法	4
第八章	统计方法的求解	8

作业20%,期中20%, 期末60%. 作业不得以任何理由迟交。

Chapter 1

均匀随机数的产生

1.1 概论

首先看一则新闻。该新闻可以从以下的网址查询：[http://news.bitauto.com/others/20110127/0805286622.html]

首轮购车摇号1.76万购车指标5秒摇出

……

10:11 抽取种子号现场大屏幕亮起，开始摇号，初始状态为6个0。第一位代表走近计算机，点击键盘，产生了本次的第一位种子数字“0”。两分钟后，6个种子数正式产生，为“040815”。摇号程序中，种子数是一个初值。借助这个初值，根据相关的函数公式，可以通过计算机等快速演算工具，循环测算出具体的中签编码。

10:14 抽1.76万指标购车指标摇号正式开始，经过5秒钟计算机数据处理，17600个购车编码被抽出。其中，编码号为5163100187648的尹明位列表格第一位，成为“摇号第一人”。随后，摇号结果现场刻盘。

……

本章将介绍均匀随机数的产生方法，为以后其他随机数的产生以及随机模拟方法做准备。

1.1.1 基本概念和理论

定义1 (随机数). 设随机变量 $\eta \sim F(x)$, 则称随机变量 η 的随机抽样序列 $\{\eta_i, i = 1\ldots, n\}$ 为分布$F(x)$ 的随机数。

例：若$\eta \sim N(\mu, \sigma^2)$,则称来自$\eta$ 的随机抽样序列$\{\eta_i, i = 1\ldots, n\}$ 为正态分布随机数；若$\eta \sim$ 指数分布,则称来自η 的随机抽样序列$\{\eta_i, i = 1\ldots, n\}$ 为指数分布随机数；$\eta \sim U[a, b]$,则称来自η 的随机抽样序列$\{\eta_i, i = 1\ldots, n\}$ 为$[a, b]$区间均匀分布随机数；

本章重点介绍$[0, 1]$ 区间均匀分布的随机数的产生和检验方法。虽然$[0, 1]$ 区间上的均匀分布

十分简单，但是产生这个简单分布的随机数对使用随机模拟方法解决问题是十分重要的。许多其他形式分布分布（如正态分布、指数分布、二项分布等）的随机数都可以从[0, 1]区间的均匀分布的随机数，经过变换得到。

为方便, 本书使用如下版本的反函数：

定义2. 设 $F(x)$ 是一个单调（无需严格单调）增函数，定义其反函数

$$F^{-1}(x) = \inf_t \{t : F(t) \geq x\}.$$

当$F(X)$ 是一个严格单调的函数时，上面定义的反函数等同于一般的反函数。例：$F(x)$ 是二项分布 $P(X = 1) = 1 - P(X = 0) = p, 0 < p < 1$ 的分布函数。那么 $F(x)$ 单调上升，但不是严格单调。

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$$

$$F^{-1}(x) = \begin{cases} -\infty & x = 0 \\ 0 & 0 < x \leq 1 - p \\ 1 & 1 - p < x \leq 1 \end{cases}$$

下面定理总结一下$F^{-1}(x)$ 的性质:

定理1. $F(x)$ 是随机变量X 的分布函数，那么

(1) $F^{-1}(x)$ 单调增（不一定严格增）

(2) $F^{-1}(F(x)) \leq x$

(3) $F(F^{-1}(x)) \geq x$

(4) $F^{-1}(y) \leq x$ *iff* $y \leq F(x)$

Proof. Homework. □

下面定理是将一般随即变量与均匀随机变量联系起来的理论基础。

定理2. *(1)* 若 ξ 的分布函数为 $F(x) := P(\xi \leq x)$, 如果$F(F^{-1}(x)) = x$，则 $F(\xi) \sim U[0, 1]$.

(2) 若 $X \sim U[0, 1]$，则 $F^{-1}(X)$ 的分布函数为 $F(x)$.

Proof. (1)

当$x \in [0, 1]$, $P(F(\xi) \leq x) = P(\xi \leq F^{-1}(x)) = F(F^{-1}(x)) = x$.

当$x<0$, $P(F(\xi)\le x)=0$.

当$x>1$, $P(F(\xi)\le x)=1$.

(2)由定理 1, $P(F^{-1}(X)\le x)=P(X\le F(x))=F(x)$. □

由定理 2，我们可以从 $[0,1]$ 均匀随机数出发，得到任意分布的随机数。设 $\eta_i, i=1,\ldots,n$ 是来自$U[0,1]$ 的随机数,令 $\zeta_i=F^{-1}(\eta_i)$, 则$\zeta_i, i=1,\ldots,n$ 就是分布为 $F(x)$ 的随机数。

将定理 2的(1) 和(2) 结合起来，我们可以从一个（连续随机变量）分布的随机数出发，得到另外一个分布的随机数。

推论1. 若 ξ 的分布函数为 $G(x):=P(\xi\le x)$, 且$G(G^{-1}(x))=x$；$F(x)$ 是一个分布函数，则 $F^{-1}(G(\xi))\sim F(x)$.

下面定理给出如何使用抛硬币的方法产生均匀分布的随机数。

定理3. 设 X_i *i.i.d.* $\sim Bernoulli(\frac{1}{2}), i=1,\ldots$:

$$P(X_i=0)=P(X_i=1)=\frac{1}{2},$$

令

$$\begin{aligned}\eta &= \frac{X_1}{2}+\frac{X_2}{2^2}+\ldots+\frac{X_k}{2^k}+\ldots\\ &= 0.X_1X_2\ldots X_k\ldots \text{ (用二进制小数表示)}\end{aligned}$$

则 $\eta\sim U(0,1)$.

Proof. 显然 $\eta\in[0,1)$. 以下证明η 在 [0,1) 中均匀分布.只需证任给$[a,b)\subset[0,1)$, 有$P(a\le\eta<b)=b-a$. (为什么?)

首先看一个简单的情形：$[a,b)=[\frac{1}{4},\frac{3}{4})$. 因为

$$\begin{aligned}\left[\frac{1}{4},\frac{3}{4}\right) &= \left[\frac{1}{4},\frac{2}{4}\right)\cup\left[\frac{2}{4},\frac{3}{4}\right)\\ &= \left[0.01,0.0\dot{1}\right)\cup\left[0.10,0.10\dot{1}\right)\\ &= \{X_1=0\&X_2=1\}\cup\{X_1=1\&X_2=0\},\end{aligned}$$

所以，

$$\begin{aligned} P\left[\frac{1}{4},\frac{3}{4}\right) &= P(\{X_1=0\&X_2=1\}\cup\{X_1=1\&X_2=0\}) \\ &= \frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2} \\ &= \frac{1}{2}=\frac{3}{4}-\frac{1}{4}. \end{aligned}$$

对于一般的 $[a,b)$, 因为a,b可以由有限位小数逼近，而对于有限位小数的情况，可用类似于上述方法证明。 □

上述定理给出了产生均匀随机数的精确方法。但是利用计算机并不能得到均匀随机数，因为计算机提供的数字总是有限的。需要寻求其他方法或者近似方法。

1.1.2 产生随机数的一般方法

常用的产生随机数的方法有三种：

1. **手工方法。**比如：抛硬币、掷骰子、抽签、摇号等。这种方法简单易行，技术含量低，在民间广泛采用。缺点是如果需要大量的随机数，执行效率较低。

2. **物理方法。**比如放射性物质放射出的粒子数等。该方法的优点是：产生的随机数质量好；缺点是：技术含量太高，需要有专门的物质和仪器。

3. **数学方法。**按照某一递推公式$\eta_n = f(\eta_{n-1},\eta_{n-2},\ldots,\eta_{n-k})$ 产生数列$\eta_1,\eta_2,\ldots,\eta_n,\ldots$. 当$n$充分大时，这一数列具有均匀分布随机变量的独立抽样序列的性质，这一数列称为伪随机序列。本方法的优点是可重复产生大量的伪随机数。缺点是产生的随机数不是真正的随机数，只是近似的随机数。本章重点介绍采用数学方法产生伪随机数列。

1.2 均匀随机数的产生

1.2.1 伪随机数

前面简单介绍了产生伪随机数列的数学原理，可以看出产生的数列是一个确定性的数列，因此不可能是真正意义的随机数列。所以把这种数列成为伪随机数列。虽然如此，如果计算方法经过精心的设计，可以产生看起来是相互独立的$[0,1]$区间均匀分布的随机数列，并且可以通过一系列的统计检验（如:独立性、均匀性等）。可以说，只要具有均匀分布随机数的一些统计性质，我们就可以把伪随机数作为真正的随机数使用。

平方取中法。 1946 年，冯 ·诺依曼等首次提出平方取中法：从一个4位数η_0开始，平方后得到8 位数。取出8位数中间的4位作为η_1，重复以上过程得到一个伪随机数列：$\{\frac{\eta_i}{1000}, i = 1, 2, \ldots\}$. 例如：$\eta_0 = 6031$, 将它平方得36372961, 再取中间四位作为η_1, 即3729. 依此类推，可得$9054, 9749, 0430, 1849, \ldots$.

此方法的特点:简单。缺点：均匀性不好，且数列很快趋于0，数列的长度也难以确定。结论：目前已经没有使用价值。

一个好的均匀随机数发生器(产生随机数的数学方法)应当具备以下几点:

1. 产生的数列要具有均匀总体随机样本的统计学性质,如分布的均匀性,抽样的随机性, 数列间的独立性等.
2. 产生的数列要有足够长的周期
3. 产生的数列要速度快，占用计算机内存少，具有完全可重复性。

本节重点介绍三种随机数发生器: 线性同余法、组合同余法和反馈位移寄存器方法。他们是目前较流行的也是多数统计学家认为较好的随机数发生器。也介绍一下matlab 和R 中使用的随机数发生器"Mersenne-Twister": From Matsumoto and Nishimura (1998). A twisted GFSR with period $2^{19937} - 1$ and equidistribution in 623 consecutive dimensions (over the whole period). The 'seed' is a 624-dimensional set of 32-bit integers plus a current position in that set.

1.2.2 线性同余发生器(LCG)

LCG 的全称为Linear Congruence Generator。Lehmer 在1951年提出的。此方法利用数论中的同余运算来产生随机数，故称为同余发生器。它包括混合同余发生器和乘同余发生器。

（一）同余与线性同余法

首先介绍一下与同余有关的一些概念和性质。

定义3. 设 a, b, M 为整数，$M > 0$, 若$a - b$ 为M 的整数倍，则称a 与b 关于模M 同余；记为$a \equiv b(mod\ M)$; 否则称a 与b 关于模M 不同余；记为$a \not\equiv b(mod\ M)$ 。

注：同余顾名思义，两个数同余表示他们除以一个正整数之后所得的余数相同。

例子：$1 \equiv 11(mod\ 10)$，$11 \equiv 1(mod\ 11)$

同余具有以下性质：

1. 对称性：若$a \equiv b(mod\ M)$，则$b \equiv a(mod\ M)$
2. 传递性：$a \equiv b(mod\ M)$, $b \equiv c(mod\ M)$, 则$a \equiv c(mod\ M)$
3. 若$a_i \equiv b_i(mod\ M), i = 1, 2$, 则$a_1 \pm a_2 \equiv b_1 \pm b_2(mod\ M)$, $a_1 a_2 \equiv b_1 b_2(mod\ M)$.

4. 可约分: 若$aC \equiv bC(mod\ M)$, 则$a \equiv b(mod\frac{M}{(M,C)})$, 其中$(M,C)$ 表示M 和C 的最大公因子。

Proof. 性质1和性质2 十分简单。我们首先证明性质3的第二条：$a_1a_2 \equiv b_1b_2(mod\ M)$。设$a_1 = k_1M + c_1, b_1 = k_2M + c_1; a_2 = k_3M + c_2, b_2 = k_4M + c_2$, 那么$a_1a_2 \equiv c_1c_2, b_1b_2 \equiv c_1c_2$ 所以$a_1a_2 \equiv b_1b_2$.

接着我们证明性质4– 可约分性. 设$aC = k_1M + d, bC = k_2M + d$, 那么$a = k_1M/C + d/C, b = k_2M/C + d/C$. $a - b = k_3M/C = k_3 * M/(M,C) * (M,C)/C = k_3 * k_4/k_5$. 注意到$k_4$ 和k_5 互质，所以k_3/k_5 是一个整数，记为k. 那么$a - b = k * k_4 = k * M/(M,C)$, 所以$a \equiv b(mod\frac{M}{(M,C)})$. □

LCG 方法的一般递推公式为

$$\begin{cases} 初值x_0 & \\ x_n & = (ax_{n-1} + c)(modM), \\ r_n & = x_n/M \end{cases} \tag{1.1}$$

显然由(1.1)式得到的$x_n(n = 1, 2, \ldots)$ 满足$0 \le x_n < M$. 从而$r_n \in [0, 1)$. 应用递推公式(1.1)产生均匀随机数时，式中参数a, c, x_0, M 的选取十分关键。

例

$$\begin{cases} 初值x_0 = 1 & \\ x_n & = (5x_{n-1} + 1)(mod\ 10) \end{cases}$$

由此递推公式，我们得到的数列$\{x_n\}$ 为$6, 1, 6, 1, \ldots\ T = 2$.

$$\begin{cases} 初值x_0 = 1 & \\ x_n & = (5x_{n-1} + 1)(mod\ 8) \end{cases}$$

由此递推公式，我们得到的数列$\{x_n\}$ 为$6, 7, 4, 5, 2, 3, 0, 1, 6, 7 \ldots\ T = 8$.

定义4 (周期)**.** 对初值x_0,同余法$x_n = (ax_{n-1} + c)(modM)$产生的数列$\{x_n, n = 1, 2, ldots\}$其重复数之间的最短长度，称为此初值下$LCG$的周期,记为$T$. 若是$T = M$, 则称为满周期。

前面的两个例子中，第一个不是满周期，第二个是满周期。

（二）混合同余法（混合式LCG）

式(1.1)中的参数$C > 0$ 时的LCG方法称为混合同余法。

（三）乘同余法(积式发生器)

式(1.1)中的参数$C = 0$ 时的LCG方法称为乘同余法（multiplier congruence generator）。

混合式同余法的周期：当同余法的参数a, c, x_0 和M 满足一定的条件时，同余法产生的数列可以达到满周期。

定理4 (满周期)**.** 如果下列条件都满足，则由式(1.1)产生的数列可达到满周期。

(1). c 与M 互素（即它们的公因数只有*1*）

(2). 对M 的任一个素因子P, $a \equiv 1(modP)$，即$a-1$应被P 整除。

(3). 如果4 是M 的因子，则$a \equiv 1(mod4)$.

在实际应用中,常取$M = 2^L$. 为满足定理的条件(1), 可以取$c = 2\beta + 1$; 显然4 是M 的一个因子，所以由条件(3), 取$a = 4\alpha + 1$; 2 是M 的唯一素因子，所以，$a = 4\alpha + 1$ 满足条件(2).

$$\begin{cases} \text{种子}x_0\text{任意非负整数} & \\ x_n & = ((4\alpha+1)x_{n-1} + c(2\beta+1))\,(mod 2^L), \\ r_n & = x_n/2^L \end{cases} \tag{1.2}$$

乘同余法的周期：一般来讲，乘同余发生器达不到满周期。下面的定理给出乘同余发生器的最大周期。

引理1 (最大周期)**.** 设乘同余发生器：

$$\begin{cases} \text{初值}x_0 < M, (x_0, M) = 1 & \\ x_n & = (ax_{n-1})(modM), \end{cases} \tag{1.3}$$

则使得$a^V \equiv 1(modM)$ 的最小正整数V 是乘同余发生器(1.6) 的周期。

Proof. 设乘同余发生器的周期为T。因为$x_V \equiv ax_{V-1} \equiv a^2x_{V-2} \equiv \ldots \equiv a^Vx_0(modM)$.注意到$a^V \equiv 1$, 所以$x_V \equiv x_0(modM)$. 因为$x_0 < M$, 所以$x_V = x_0$. 由周期的定义，$T \leq V$.

现假设$T < V$. 由周期的定义，必存在$1 \leq i < j \leq V$,使得$x_j \equiv x_i(modM)$. 类似于前面的讨论，$x_j \equiv a^jx_0 \equiv a^ix_0 \equiv x_i(modM)$. 由同余的性质(4)（约分）, $a^j \equiv a^i(modM)$. 又因为$a^V = a^ja^{V-j} \equiv a^ia^{V-j}(modM) \equiv 1(modM)$. 显然，$V-(j-i) < V$, 这与$V$ 是满足$a^V \equiv 1(modM)$ 的最小正整数V 矛盾。

□

注意：在积式发生器中，参数a 和 M 要求互素。否则产生的数列在若干步之后可能退化为 0 。

下面我们通过一个例子将乘同余法与混合同余法联系起来。

例：设乘同余发生器为

$$\begin{cases} \text{初值}x_0 = 4b+1 & \\ x_n & = (8\alpha+5)x_{n-1}(mod 2^L), \end{cases} \tag{1.4}$$

则 $x_n = 4x_n^* + 1$. x_n^* 满足：

$$\begin{cases} 初值x_0^* = b \\ x_n^* \quad = \quad (8\alpha+5)x_{n-1}^* + (2\alpha+1)(mod 2^{L-2}), \end{cases} \tag{1.5}$$

由定理4 知， 上面的混合同余法达到满周期 $T = 2^{L-2}$. 又因为$x_n \in \{4m+1 : 4m+1 < 2^L\}$, 而满足 $4m+1 < 2^L$ 的整数共有 2^{L-2} 个， 所以 $\{x_n\}$ 达到最大周期 $T = 2^{L-2}$. 整数列$\{x_n\}$ 是由0 到$2^L - 1$ 之间仅1/4 的数集合$\{1, 5, 9, \ldots, 2^L - 3\}$ 重新排列而成。

下面给出乘同余法达到满周期的充要条件：

定理5 (乘同余最大周期)． *1.* 当$M = 2^L (L \geq 4)$, x_0 为奇数时，则取$a \equiv 3 or 5 (mod 8)$, 且最大周期$T = 2^{L-2}$;

2. 当$M = 10^s (s \geq 5)$, x_0 不是2 或者*5* 的倍数时，则取$a (mod 200)$ 等于以下*32* 个值之一：*3*，*11*，*13*，*17*，*21*，*...*, 且$T = 5 \times 10^{s-2}$.

3. 当$M = p$*(*p 为素数*)*，则取a 为M 的素元，且可得最大周期 $T = M - 1$.

算法：溢出原理：当$M = 2^L$时，利用乘同余算法产生数列x_n 时，需要计算以M为模的余数：$x_n = (ax_{n-1})(mod M)$. 若$ax_{n-1} < M$, 则$x_n = ax_{n-1}$; 若$ax_{n-1} \geq M$, 记

$$ax_{n-1} = \sum_{i=0}^{k} \alpha_i 2^i (\alpha_k = 1, K \geq L)$$

, 这时

$$ax_{n-1} = \sum_{i=0}^{L-1} \alpha_i 2^i + 2^L \sum_{i=L}^{k} \alpha_i 2^{i-L}$$

所以，$x_n = \sum_{i=0}^{L-1} \alpha_i 2^i$.

如果取L 为计算机中整数的尾数字长，比如$L = 31$ 或者$L = 15$ 等。因为计算机可存放的最大整数为$2^L - 1$. 当整数相乘后，如果$ax_{n-1} = \sum_{i=0}^{k} \alpha_i 2^i > 2^L - 1 (= k \geq L)$, 这时将数值$ax_{n-1}$存入计算机的贮存单元时，就会‘溢出’，而保留的$L$位数值正好是$x_n$.这就是利用溢出产生余数的原理。

归一运算：若$x_n = \sum_{i=0}^{k} \alpha_i 2^i (k < L)$, 则$\frac{x_n}{2^L} = \sum_{i=0}^{k} \alpha_i 2^{i-L}$. 归一运算相当于把小数点移到最左边位置。利用底层语言如汇编编写程序运算速度会非常快。

素数模乘同余法 在以上的乘同余发生器中，适当选取参数，可使周期达到$T = 2^{L-2}$.但是这个周期离满周期还有很大的距离。如果挑选好的M,如M 是素数，则可以得到周期为$M - 1$的数列。这样的发生器成为素数模乘同余发生器。

一些基本概念:

设M, a 为正整数，$(a, M) = 1$.

定义5 (a 对M的阶数（或次数）). 称满足$a^V \equiv 1(modM)$ 的最小整数V为a 对模M的阶数（或次数），简称为a 的阶数（或次数）。

由引理1知，a的次数就是该乘同余法的周期。

定义6 (素元). 若a 对素数模M的阶数V满足：$V = M - 1$,则称a 为M 的素元（或原根）。

注：素元存在但可以不唯一。

例: 已知$M = 2^3 - 1$ 是一个素数；取$a = 5$. 因为

$$\begin{aligned} &5^1 = 5 \not\equiv 1(mod7), \qquad 5^2 = 25 \not\equiv 1(mod7), \\ &5^3 = 125 \not\equiv 1(mod7), \quad 5^4 = 625 \not\equiv 1(mod7), \\ &5^5 = 3125 \not\equiv 1(mod7), \quad 5^6 = 16255 \equiv 1(mod7). \end{aligned}$$

所以$a = 5$ 对模$M = 7$的阶数为6。5 是7的素元。

$$\begin{cases} \text{初值} x_0 \\ x_n \quad = \quad (5x_{n-1})(mod7), \end{cases} \tag{1.6}$$

周期为6。

素数模乘同余发生器中参数的选择：

1. M 取为小于2^L 的最大素数
2. 选a为M 的素元，这样可以保证$T = M - 1$
3. 一般要求a适当取大，且二进制表示尽可能无规律

$x_n = perm\{1, 2, \ldots, M-1\}.T = M - 1.$

下面给出两个良好的素数模乘同余发生器：

$L = 35$, 小于2^{35} 的最大素数$M = 2^{35} - 31 = 34359738337$, 取$a = 5^5 = 3125$.

$L = 32$, 小于2^{32} 的最大素数$M = 2^{32} - 1 = 2147483647$, 取$a$为以下四个数中任何一个：$7^5 = 16807, 397204094, 764261123, 630360016$.

算法：设$M = 2^L - g$ 为小于2^L的最大素数，递推公式为

$$x_n = ax_{n-1}(mod2^L - g),$$

令$z_n = ax_{n-1}(mod2^L)$，计算z_n时可以利用溢出原理。记$k = [ax_{n-1}/2^L]$,则

$$x_n = \begin{cases} z_n + kg & z_n + kg < 2^L - g \\ z_n + kg - (2^L - g) & z_n + kg \geq 2^L - g \end{cases}$$

1.2.3 反馈位移寄存器

LCG 的缺点：(1) 长生的均匀随机数做为m 维均匀随机向量时相关性较大；（2）周期T与计算机字长有关。在整数的尾数字长为L的计算机上，周期不会超过2^L的均匀随机数列。1965年，Tausworthe 发明了FSR (Feedback Shift Register) 方法以克服以上缺点。

(一) FSR 介绍

Tausworthe(1965) 提出的FSR方法，用线性反馈递推公式：

$$\alpha_k = c_p\alpha_{k-p} + c_{p-1}\alpha_{k-p+1} + \ldots + c_1\alpha_{k-1}(mod2), \tag{1.7}$$

对寄存器中的二进制码α_k 做递推运算，其中p 是给定的正整数，$c_p = 1$, $c_i = 0, 1$ $(i = 1, 2, \ldots, p-1)$ 为给定常数。

给定初值$(\alpha_{-p+1}, \alpha_{-p+2}, \alpha_0)$, 由(1.7) 产生的0 或1 组成一个二进制数列$\{\alpha_n\}$. 截取数列$\{\alpha_n\}$中连续的L位构成一个L位的二进制数；接着截取L位，又形成一个整数，以此类推，可得：

$$x_1 = (\alpha_1, \alpha_2, \ldots, \alpha_L)_2$$

$$\ldots$$

$$x_n = (\alpha_{(n-1)L+1}, \alpha_{(n-1)L+2}, \ldots, \alpha_{nL})_2$$

令$r_n = x_n/2^L (n = 1, 2, \ldots,)$，则$\{r_n\}$即为FSR方法产生的均匀随机数。

(二) 等价公式

1971 年Toothill, Robinson 和Adams 给出FSR 方法另一个表达式：

$$\begin{cases} u_0 &= 1 \\ u_k &= xu_{k-1}(modx^p + x^q + 1), \end{cases} \tag{1.8}$$

其中u_k 是次数$< p$ 且系数为0或1的多项式。而且$p > q > 0$ 为正整数。例如：$p = 5, q = 2$，我们

有

$$\begin{aligned} u_0 &= 1 \\ u_1 &= x \\ u_1 &= x^2 \\ u_1 &= x^3 \\ u_1 &= x^4 \\ u_1 &= x^2+1 \\ u_1 &= x^3+x \\ \ldots &\quad \ldots \quad \ldots. \end{aligned}$$

多项式u_k可以用系数$(\alpha_0^{(k)}, \alpha_1^{(k)}, \ldots, \alpha_{p-1}^{(k)})$表示:

$$u_k = \alpha_0^{(k)} + \alpha_1^{(k)} x + \ldots + \alpha_{p-1}^{(k)} x^{p-1}.$$

下面来看多项式系数之间的关系:

$$\begin{aligned} u_{n+p} &= x u_{n+p-1} (mod\ x^p + x^q + 1) \\ &= x^n u_p \\ &= x^n (u_q + u_0) \\ &= u_{q+n} + u_n \end{aligned}$$

所以多项式之间的系数满足以下关系：

$$w_{n+p} = w_{n+q} \oplus w_p, n = 1, 2, \ldots.$$

这跟前面的FSR 递推公式在系数$c_p = 1, c_q = 1$, 其余全为0，是等价的。这是因为当系数$c_p = 1, c_{p-q} = 1$, 其余全为0时，

$$\alpha_k = \alpha_{k-p} \oplus \alpha_{k-p+q}.$$

令$k = n + p$, 我们有

$$\alpha_{n+p} = \alpha_n \oplus \alpha_{n+q}.$$

(三) 算法

我们利用上面的公式：

$$\alpha_{n+p} = \alpha_n \oplus \alpha_{n+q}.$$

给出FSR的算法。取$L=p$, 令$y_n=(\alpha_n,\alpha_{n+1},\ldots,\alpha_{n+p-1})_2$ $y_{n+1}=(\alpha_{n+p},\alpha_{n+p+1},\ldots,\alpha_{n+2p-1})_2$.
注意到两种等价的关系，可以将y_n 和y_{n+1} 表示为两个多项式，其系数为：$w_n=(\alpha_n,\alpha_{n+1},\ldots,\alpha_{n+p-1})_2$ $w_{n+p}=(\alpha_{n+p},\alpha_{n+p+1},\ldots,\alpha_{n+2p-1})_2$ 记$w_{n+q}=(\alpha_{n+q},\alpha_{n+q+1},\ldots,\alpha_{n+q+p-1})_2$, 则

$$w_{n+p}=w_{n+q}\oplus w_p, n=1,2,\ldots.$$

可以用一个图表示：

$$\underbrace{\alpha_n,\alpha_{n+1},\ldots,[\alpha_{n+q},\ldots,\alpha_{n+p-1}}_{w_n}\ \underbrace{\alpha_{n+p},\ldots,\alpha_{n+p+q-1}],\ldots,\alpha_{n+2p-1}}_{w_{n+p}}$$

可以按照如下的两个步骤由y_n 得到y_{n+1}.

1. 由w_n 计算w_{n+p} 的前$p-q$ 个元素

$\alpha_{n+p}=\alpha_n\oplus\alpha_{n+q}$

$\alpha_{n+p+1}=\alpha_{n+1}\oplus\alpha_{n+q+1}$

...

$\alpha_{n+p+(p-q-1)}=\alpha_{n+p-q-1}\oplus\alpha_{n+p-1}$ 2. 由w_n 和w_{n+p} 的前$p-q$ 个元素计算w_{n+p} 的后q 个元素

$\alpha_{n+p+(p-q)}=\alpha_{n+p-q}\oplus\alpha_{n+p}$

$\alpha_{n+2p-1}=\alpha_{n+p-1}\oplus\alpha_{n+p+q-1}$

(四) GFSR

GFSR 将上述公式一般化，直接得到如下的简单公式：

$$y_{n+1}=y_n+\oplus y_{n+q}.$$

GFSR 产生均匀随机数的一般步骤：

1. 先产生p 的随机整数：$y_1,y_2,\ldots,y_p\in(0,2^p-1)$

2. 由$y_{n+p}=y_n\oplus y_{n+q}$ 依次产生y_{n+p}

3. $r_n=y_n/2^p$.

1.2.4 组合发生器

组合发生器是用两个随机数数发生器合成一个随机数发生器。具体来说，1. 用第一个随机数发生器产生k 个随机数。这k个随机数顺序的存放在一个数组（矢量）$T=(t_1,t_2,\ldots,t_k)$ 中。置n =1.

2. 用第二个随机数发生器产生一个随机整数1 $\le j\le k$.

3. 令$x_n = t_j$; 然后再利用地一个随机数发生器产生一个随机数y. 令$t_j = y$, 置$n = n + 1$.

4. 重复2 和3 得到随机数列$\{x_n\}$.

1.3 均匀随机数的检验

统计检验的一般方法：首先假设总体具有某种统计特征，然后由样本值检验这个假设是否可信，这种方法称为假设检验。具体步骤为：

1. 提出假设H_0。如：总体分布为$U(0,1)$

2. 选取合适的统计量T, 并求出T 在H_0 成立时的分布；

3. 给定显著水平α 如0.05，确定拒绝域W，使得

$$P(T \in W) = \alpha$$

4. 由观测值计算统计量T

5. 做统计推断。如果$T \in W$则拒绝H_0; 否则不拒绝。

在均匀随机数的检验中，主要使用两类常用的统计量：

(1) 根据中心极限定理得到近似正太分布统计量

设$\eta_1, \eta_2, \ldots, \eta_n$ i.i.d. 服从$F(x)$, 且$E(\eta_i) = \mu$, $var(\eta_i) = \sigma^2$. 记$\bar{\eta} = \frac{1}{n}\sum_{i=1}^{n}\eta_i$, 则

$$U = \frac{\sqrt{n}(\bar{\eta} - \mu)}{\sigma}$$

以$N(0,1)$ 为极限分布。

(2) χ^2 统计量

设总体η的简单子样$\eta_i, (i = 1, \ldots, n)$按照一定规则分为不相交的$m$ 个组，记第i个组的观测频数为$n_k (k = 1, \ldots, m)$. 若随机变量η属于第i 组的概率为p_i, 则理论频数为$\mu_i = np_i$, 由n_i, μ_i 构造统计量：

$$V = \sum_{k=1}^{m} \frac{(n_i - \mu_i)^2}{\mu_i}$$

渐近服从$\chi^2(f)$ 分布，自由度$f = m - \ell - 1$, ℓ 是附加在概率分布p_i上独立约束条件的个数(参数个数)。当$f > 30$，$U = \sqrt{2V} - \sqrt{2f - 1} \approx N(0,1)$.

1.3.1 参数检验

均匀随机数的参数检验是检验由某个发生器产生的随机数列$\{r_i\}$ 的均值、方差或者各阶矩等与均匀分布的理论值是否有显著的差异。

若$R \sim U(0,1)$, 则$E(R)=\frac{1}{2}, E(R^2)=\frac{1}{3}, var(R)=\frac{1}{12}$. 若$R_1, R_2, \ldots, R_n$ 是均匀总体R的简单随机抽样。记

$$\bar{R}=\frac{1}{n}\sum_{i=1}^{n} R_i, \bar{R^2}=\frac{1}{n}\sum_{i=1}^{n} R_i^2, s^2=\frac{1}{n}\sum_{i=1}^{n}(R_i-\frac{1}{2})^2,$$

则

$$\begin{aligned} E(\bar{R})&=\tfrac{1}{2}, \quad var(\bar{R})=\tfrac{1}{12n} \\ E(\bar{R^2})&=\tfrac{1}{3}, \quad var(\bar{R^2})=\tfrac{4}{45n} \\ E(s^2)&=\tfrac{1}{12}, \quad var(s^2)=\tfrac{1}{180n} \end{aligned}$$

设$r_1, r_2, \ldots, r_n$ 是某个发生器产生的随机数。则在$\{r_i\}$ i.i.d. 服从$U(0,1)$ 的假设下，统计量：

$$u_1=\frac{\bar{r}-E(\bar{r})}{\sqrt{var(\bar{r})}}=\sqrt{12n}(\bar{r}-\frac{1}{2})$$

$$u_2=\frac{\bar{r^2}-1/3}{\sqrt{4/(45n)}}=1/2\sqrt{45n}(\bar{r^2}-\frac{1}{3})$$

$$u_3=\frac{s^2-1/12}{\sqrt{1/(180n)}}=\sqrt{180n}(s^2-\frac{1}{12})$$

根据CLT，以上三个统计量都渐近服从$N(0,1)$, 给定显著性水平α后，找到分位点λ使得

$$P(|u_i|>\lambda)=\alpha.$$

否定域为$W_i=\{|u_i|>\lambda.\}$

1.3.2 均匀性检验

该检验又称为频数检验。用来检验由某个随机数发生器产生的随机数的经验频率与理论频率是否有显著差异。

（一）χ^2 检验

将$[0,1]$ 区间分为m个小区间，以$[\frac{i-1}{m}, \frac{i}{m}), i=1, \ldots, m$ 表示第i个小区间。设随机数列$r_j, j=1, \ldots, n$ 落入第i个小区间的数目为$n_i, i=1, \ldots, m.$ 由均匀分布的性质$\mu_i=E(n_i)=\frac{n}{m}$,统计量

$$V=\sum_{i=1}^{m}\frac{(n_i-\mu_i)^2}{\mu_i}=\frac{m}{n}\sum_{i=1}^{m}(n_i-\frac{n}{m})^2$$

渐进服从$\chi^2(m-1)$

（二）K-S检验该检验是检验连续分布的拟合性检验。它检验样本的经验分布函数与总体的分布函数间是否有显著差异。

记$r_1,\ldots,r_n$的经验分布函数为F_n. $F_n(x)=\hat{P}(X\le x),\underline{F_n}(x)=\hat{P}(X<x)$,

$$D_n^+=\max_i[F_n(r_i)-F(r_i)]$$

$$D_n^-=\max_i[F(r_i)-\underline{F_n}(r_i)]$$

统计量$D_n=\max\{D_n^+,D_n^-\}$ 服从K-S 分布。

注意到$F_n(r_{(i)})=\frac{i}{n},\underline{F_n}(r_{(i)})=\frac{i-1}{n}$,

$$D_n^+=\max_i[i/n-F(r_i)]=\max_i[\frac{i}{n}-r_{(i)}]$$

$$D_n^-=\max_i[F(r_i)-\underline{F}(r_i)]=\max[r_{(i)}-\frac{i-1}{n}]$$

（三）序列检验

序列检验实际上用于多为均匀分布的均匀性检验；它也间接的检验序列的独立性。

将$r_1,r_2,\ldots,r_{2n}$配对构成二维随机样本：

$$v_1=(r_1,r_2),v_2=(r_3,r_4),\ldots,$$

将正方形分成k^2个等面积的小正方形，$\mu_{ij}=\frac{n}{k^2}$.

$$V=\frac{k^2}{n}\sum_i\sum_j(n_{ij}-\frac{n}{k^2})^2$$

渐近服从$\chi^2(k^2-1)$.

1.3.3 独立性检验

（一）相关系数检验I 考虑j阶自相关系数：

$$\rho(j)=\frac{\frac{1}{n-j}\sum_{i=1}^{n-j}(r_i-\bar{r})(r_{i+j}-\bar{r})}{\frac{1}{n}\sum_{i=1}^{n}(r_i-\bar{r})^2}$$

当$n-j$充分大，且$\rho=0$成立时，$u_j=\rho(j)\sqrt{n-j}$ 渐近$N(0,1)$.

（二）相关系数检验II j 阶自相关系数还可以这样定义：

$$\rho(j)=\frac{\frac{1}{n}\sum_{i=1}^{n-j}(r_i-\bar{r})(r_k-\bar{r})}{\frac{1}{n}\sum_{i=1}^{n}(r_i-\bar{r})^2}$$

其中$k=i+j(modn)$.

记$C_j = \frac{1}{n}\sum_{i=1}^{n} r_i r_k$, 则：$E(C_j) = 1/4$, $var(C_j) = \frac{13}{144n}$.

$$T = \frac{C_j - 1/4}{\sqrt{\frac{13}{144n}}}$$

渐近服从$N(0,1)$.

（三）列联表检验

方法等同于前面的序列检验。

（四）游程检验

上升游程：将数列$\{r_i\}$分为许多个子序列，使得每一个字序列的都是一个上升序列，则称每一个子序列为一个上升游程。

例如：

$$0.855, |0.108, 0.26, |0.032, 0.123, |0.055, 0.545, 0.642, 0.870, |0.0104.$$

可以分为5个上升游程，其长度分别为1,2,2,4,1.

记游程长度为1,2,3,4,5和≥6的游程数目分别为$g_1, g_2, g_3, g_4, g_5, g_6$,则

$$Q_n = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} a_{ij}(g_i - nb_i)(g_j - bn_j)$$

渐近服从$\chi^2(6)$. a, b都是固定的常数，$n > 4000$.

1.3.4 组合规律检验

该检验把n个随机数按一定规律进行组合，检验观测值的各种组合规律是否与理论值之间有显著差异。

(一) 扑克检验将随机数列$\{r_i, i = 1, \ldots, n\}$ 8 个分为一组。记a_i 为r_i的第一位八进制数，得到8个数字$A = \{a_1, a_2 \ldots, a_8\}, a_i \in \{0, 1, 2, \ldots, 7\}$. 8个不同的数字对应8个不同的花色，$A$有如下几种可能：

1. $A_1 = \{A$ 为单色$\}$
2. $A_2 = \{A$ 为二色$\}$
3. $A_3 = \{A$ 为三色$\}$
8. $A_4 = \{A$ 为八色$\}$

$p_i = P(A_i)$ $\mu_i = np_i$ 统计量

$$V = \sum_{i=1}^{8} \frac{(\mu_i - m_i)^2}{\mu_i}$$

渐近服从$\chi^2(7)$.

注：可以考虑一般情况，有时候相邻时间需要合并。

(二) 配套检验

从均匀随机数列$\{r_i\}$中的第一个随机数开始，把它小数点后的第一位数字（如取s 位进制，s=8,10或16）记下来，略去其第一位数字已经出现过的随机数，知道用ℓ个随机数配齐全部s个不同的数字$0,1,2,\ldots,s-1$为止，作为一套。配齐一套所需的随机数的个数L是一个随机变量。

如果$\{r_i\}$相互独立，服从$U(0,1)$. 记X_k为配完第k个数字后，再配第$k+1$个不同数字时，所需的随机数个数；显然$X_0=1$, $X_1,X_2,\ldots,X_{s-1}$ 相互独立且服从几何分布。记：

$$P(X_k=j)=q_k^{j-1}p_k, j=1,2,\ldots,$$

其中$p_k=\frac{s-k}{s}, q_k=1-p_k=\frac{k}{s}$.

$$E(X_k)=\frac{1}{p_k}=\frac{s}{s-k}, var(X_k)=\frac{q_k}{p_k^2}=\frac{sk}{(s-k)^2}.$$

因为$L=X_0+X_1+\ldots+X_{s-1}$, 故有

$$E(L)=s(1+\frac{1}{2}+\ldots+\frac{1}{s}),$$

$$var(L)=s\left(\frac{1}{(s-1)^2}+\frac{2}{(s-2)^2}+\ldots+\frac{s-2}{2^2}+\frac{s-1}{1}\right).$$

根据上述描述，我们可以构造统计量

$$\frac{\frac{L_1+L_2+\ldots+L_N}{N}-E(L)}{\sqrt{var(L)/N}}$$

渐近正态分布$N(0,1)$.

记n_l 为配齐一套用l 个随机数的观测数（$l=s,s+1,\ldots,s+m$）,重新安排一下，有

$$\frac{L_1+L_2+\ldots+L_N}{N}=\frac{1}{N}\sum_{l=s}^{s+m}ln_l, N=\sum_{l=s}^{s+m}n_l$$

1.3.5 无连贯性检验

设随机数列$\{r_i, i=1,\ldots,n\}$ 按照前后顺序排列。把n 个数按大小分为两类或者k类，是否各类数字的出现没有连贯的现象？或者数列中各数字有连贯上升或者连贯下降的现象？例如把$\{r_i\}$按照一定规律分成两类，分别记为a,b,

$$a,a|,b,b,b|,a|,b|,a,a,a|,b|,a$$

我们把位于异类元素之间的同类元素称为一个连，连中包含同类元素的个数称为连的长度（简称连长）记为L.总连数记为T，T 可以分解为：

$$T = \sum_l T_l,$$

其中，T_l 为长度为l 的连的个数。T 和L 构成进行检验的统计量。

(一) 正负连检验令$u_i = r_i - \frac{1}{2}$，把u_i 按照正负分为两类，组成正负两类连。由均匀随机数列的均匀性，独立性假设，可得

$$E(T) = \frac{n+1}{2}, var(T) = \frac{n-1}{4}$$

$$P(L = k) = \frac{1}{2^k}, k = 1, 2, \ldots$$

Proof. Homework. □

由此可构成正态统计量或χ^2统计量。

(二)升降连检验

令$u_i = r_i - r_{i-1}$得序列$\{u_i, i = 2, 3, \ldots n\}$, 把$\{u_i\}$ 按正负分为两类，表示随机数的增减及其长度的变化规律。例如序列

$$2, 8, 9, 4, 3, 2, 5, 6, 7, 4$$

$$+ + - - - + + + -$$

有一个长度为2的上升连，接着长度为3的下降连，然后是长度为3的上升连，最后是一个长度为1的下降连。即：

$$T_1 = 1, T_2 = 1, T_3 = 2, T_4 = 0, \ldots.$$

总连数$T = \sum T_i = 4$.

当r_i i.i.d. 服从$U(0, 1)$ 时，

$$E(T) = \frac{2n-1}{3}, var(T) = \frac{16n-29}{90}.$$

Chapter 2

常用分布函数和分位数的计算

本章介绍常用分布函数$F(x)$ 和分位数x_p的计算方法。他们在假设检验中计算$p$value 和确定拒绝域中起到关键的作用。

2.1 常用分布的分布函数

(一)概念

定义7. 设X是以随机变量，则称函数

$$F(x) = P(X \leq x), -\infty < x < \infty$$

为随机变量X 的分布函数。

当X是连续型时，设密度函数为$f(x)$, 则

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

当X 是离散型时，设概率分布为

$$P(X = x_i) = p_i, i = 1, 2, , \ldots,$$

则

$$F(x) = \sum_{x_i \leq x} p_i.$$

分布函数的计算是积分或者级数的计算。

定义8 (分位数)**.** 设X是连续型随机变量。若存在数值x_p 满足

$$F(x_p) = P(X \leq x_p) = p,$$

其中$p \in [0,1]$, 则称x_p 为X 的p 分位数（或分位点）。

记$g(x) = F(x) - p$, 求分位数则可以转化成求方程$g(x) = 0$ 的根。

(二) 常用连续型分布的分布函数

(1) 均匀分布

$$U(x|a,b) = \int_a^x \frac{1}{b-a} dt, a \leq x \leq b$$

$$E(X) = \frac{a+b}{2}, var(X) = \frac{(b-a)^2}{12}$$

(2) 正态分布

$$\Phi(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, -\infty < x < \infty$$

$$E(X) = \mu, var(X) = \sigma^2.$$

(3)指数分布

$$Exp(x|\mu,\lambda) = \int_\mu \lambda e^{-\lambda(t-\mu)} dt, x \geq \mu,$$

$$E(X) = \mu + \frac{1}{\lambda}, var(X) = \frac{1}{\lambda^2}$$

(4)伽玛分布

$$G(x|a,b) = \frac{b^a}{\Gamma(a)} \int_0^x t^{a-1} e^{-bt} dt, a > 0, b > 0; x \geq 0$$

其中，$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt$;

$$E(X) = \frac{a}{b}, var(X) = \frac{a}{b^2}.$$

(5)Beta 分布

$$Beta(x|a,b) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt, a > 0, b > 0, 0 < x < 1.$$

其中

$$B(a,b) = \int_0^\infty t^{a-1}(1-t)^{b-1} dt = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

$$E(X)=\frac{a}{a+b}, Var(X)=\frac{ab}{(a+b)^2(a+b+1)}.$$

(6)$\chi^2(n)$ 分布

$$H(x|n)=\frac{1}{2^{n/2}\Gamma(n/2)}\int_0^x t^{n/2-1}e^{-t/2}dt, n \text{ 为正整数}, x>0$$

$$E(X)=n, var(X)=2n$$

$Gamma(n/2,1/2)$

(7) t 分布

$$t(x|n)=\frac{1}{\sqrt{n}B(\frac{1}{2},\frac{n}{2})}\int_{-\infty}^x (1+\frac{t^2}{n})^{-\frac{n+1}{2}}dt$$

n 为正整数,$-\infty<x<\infty$.

$$E(X)=0(n>1), var(X)=\frac{n}{n-2}, n>2$$

$$n=1, \text{ Cauchy distribution.}$$

(8) F 分布

$$F(x|m,n)=\frac{(\frac{m}{n})^{\frac{m}{2}}}{B(\frac{m}{2},\frac{n}{2})}\int_0^x t^{\frac{m}{2}-1}(1+\frac{mt}{n})^{-\frac{m+n}{2}}dt$$

$$E(x)=\frac{n}{n-2}(n>2), var(X)=\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}(n>4)$$

Homework.

(三) 常用的离散型分布

(1) 二项分布

$$B(x|n,p)=\sum_{i=0}^{[x]}\left(\begin{array}{c} n \\ x \end{array}\right)p^iq^{n-i}$$

$$E(X)=np, var(X)=np(1-p).$$

(2) 泊松分布

$$P(x|\lambda)=\sum_{i=0}^{[x]}\frac{\lambda^i}{i!}e^{-\lambda}(0!=1)$$

$$E(X)=var(X)=\lambda$$

(3) 几何分布

$$P(X=x)=pq^{x-1}, x=1,2,\ldots.$$

$$E(X) = \frac{1}{p}, var(X) = \frac{q}{p^2}.$$

(4) 负二项分布

每次实验成功的概率为$p, q = 1 - p$, 记成功k次所需实验的次数为$X + k$[前$X + k - 1$次成功了$k - 1$ 次，最后一次（第X+k次）成功], 那么X 服从负二项分布。

$$P(X = x) = \begin{pmatrix} x + k - 1 \\ k - 1 \end{pmatrix} p^{k-1} q^x p, x = 0, 1, 2, \ldots.$$

$$E(X) = \frac{kq}{p}, var(X) = \frac{kq}{p^2}.$$

(四) 分布之间的关系

1. 许多分布和正态分布有关：

(1) 若$U_1, U_2 \sim U(0, 1)$ 且相互独立，令

$$\begin{cases} X_1 &= \sqrt{-2\ln(U_1)}\cos(2\pi U_2), \\ X_2 &= \sqrt{-2\ln(U_1)}\sin(2\pi U_2), \end{cases}$$

则$X_1, X_2 \sim N(0, 1)$且独立。Homework。

(2) $t(n)$ 近似分布为$N(0, 1)$当n 充分大的时候；(3) X_1, X_2 i.i.d. $\sim N(0, 1)$, 则X_1/X_2 服从Cauchy 分布

$$f(x) = \frac{1}{\pi(1 + x^2)}, -\infty < x < \infty.$$

(4) X_1, X_2 i.i.d. $\sim N(0, 1)$, 则$\sqrt{X_1^2 + X_2^2}$ 服从瑞利分布

$$f(x) = xe^{-\frac{x^2}{2}}, x > 0.$$

2. 许多分布和Beta 分布有关：

(1)F分布和Beta 分布

$$F(x|m, n) = Beta(\frac{mx}{n + mx}|\frac{m}{2}, \frac{n}{2})$$

(2) t分布和Beta 分布

$$T(t|n) = \begin{cases} 1 - \frac{1}{2}Beta(\frac{n}{n+t^2}|\frac{n}{2}, \frac{1}{2}), t > 0 \\ \frac{1}{2}Beta(\frac{n}{n+t^2}|\frac{n}{2}, \frac{1}{2}), t \le 0 \end{cases}$$

(3) $Beta(1, 1) = U(0, 1)$

(4) 次序统计量$U_{(1)},\ldots,U_{(n)}$ 服从Beta 分布：

$$U_{(k)} \sim \beta(k, n-k+1), k = 1, 2, \ldots, n$$

$$U_{(1)} \sim \beta(1, n),$$

$$U_{(n)} \sim \beta(n, 1),$$

练习。

(5) $X_1 \sim Gamma(a, 1), X_2 \sim Gamma(b, 1)$，且相互独立，则

$$\frac{X_1}{X_1 + X_2} \sim Beta(a, b).$$

(5) $X_1 \sim \chi^2(m), X_2 \sim \chi^2(n)$，且相互独立，则

$$\frac{X_1}{X_1 + X_2} \sim Beta(\frac{m}{2}, \frac{n}{2}).$$

3. 许多分布和χ^2 分布有关：

(1) χ^2 分布是特殊的Gamma 分布：

$$\chi^2(n) = Gamma(\frac{n}{2}, \frac{1}{2})$$

(2) $X_1 \sim \chi^2, X_2 \sim \chi^2$ 且独立, 则

$$\frac{X_1/m}{X_2/n} \sim F(m, n)$$

(3) $X \sim N(0, 1), Y \sim \chi^2(n)$,则

$$\frac{X}{\sqrt{\frac{Y}{n}}} \sim t(n)$$

(4)$T \sim t(n)$,则

$$T^2 \sim F(1, n)$$

(5)Gamma 分布求和仍是Gamma 分布（有条件–第二参数固定）特别的：$X_1, X_2, \ldots, X_n \sim Gamma(1, \lambda) = exp(\lambda)$,则

$$\sum_{i=1}^{n} X_i \sim Gamma(n, \lambda).$$

(6) $R \sim U(0,1)$, 则

$$-\ln R \sim exp(1)(Gamma(1,1)).$$

二项分布、泊松分布与Gamma分布和χ^2分布之间的关系。Homework.

$$Bi(x|b,p) = \begin{cases} Beta(1-p|n-[x],[x]+1), & 0 \le x \le n \\ 0, & x < 0 \\ 1, & x > n \end{cases}$$

$$P(x|\lambda) = 1 - H(2\lambda|2([x]+1)).$$

2.2 分布函数的一般算法

本节介绍计算连续型随机变量分布函数的一般方法。

2.2.1 积分的近似算法

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

(一)等距内插求积分公式(Newton-Cotes)

该方法用于$\int_a^b f(x)dx$ 的近似计算。

已知$f(x)$在$n+1$个点$x_0, x_1, \ldots, x_n$上的值$f(x_i)$. 用多项式$L_n(x)$ 来近似$f(x)$, 即

$$f(x) = L_n(x) + R_n(x),$$

其中$L_n(x)$ 为n次多项式，R_n 为误差函数。

考虑等节点的情况。$x_k = a + kh, k = 0, 1, \ldots, n, h = (b-a)/n$. 取$L_n$ 为Lagrange 插值多项式，即

$$L_n(x) = \sum_{j=0}^{n} \frac{(x-x_0)\ldots(x-x_{j-1})(x-x_{j+1})\ldots(x-x_n)}{(x_j-x_0)\ldots(x_j-x_{j-1})(x_j-x_{j+1})\ldots(x_j-x_n)} f(x_j)$$

记$w(x) = \Pi_{j=0}^{n}(x-x_j)$, 则上式分子是

$$\frac{w(x)}{(x-x_j)},$$

分母是$w'(x)|x = x_j$. 所以

$$L_n(x) = \sum_{j=0}^{n} \frac{w(x)}{(x-x_j)w'(x_j)} f(x_j).$$

于是

$$\int_a^b f(x)dx \approx \int_a^b L_n(x) = \sum_{j=0}^{n} A_j f(x_j),$$

其中

$$A_j = \int_a^b \frac{w(x)}{(x-x_j)w'(x_j)}dx,$$

A_j 与$f(x)$ 无关，只要节点x_j 和n 确定，它就完全确定且有

$$A_j = (b-a)C_j^{(n)},$$

$$C_j^{(n)} = \frac{(-1)^{n-j}}{nj!(n-j)!}\int_0^n \frac{t(t-1)\dots(t-n)}{t-j}dt$$

$$\int_a^b \approx (b-a)\sum_{j=0}^{n} C_j^{(n)} f(a+hj).$$

$C_j^{(n)}$ 称为Newton-Cotes 系数，可以事先计算出来。下面计算几个特例：

(1) $n = 1 x_0 = a, x_1 = n$ $C_0^{(1)} = -\int_0^1 (t-1)dt = \frac{1}{2}, C_1^{(t)} = \int_0^1 tdt = \frac{1}{2}$. 梯形公式:

$$\int_a^b = \frac{b-a}{2}[f(x_0) + f(x_1)].$$

(2)$n = 2, x_0 = a, x_2 = b, x_1 = \frac{a+b}{2}$. 记$h = \frac{b-a}{2}$.可以得到抛物线公式(Simpson公式)：

$$\int_a^b f(x)dx \approx \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)].$$

(二) 高斯型求积公式

前面介绍的等距插值公式，满足下面这个性质：

$$\int_a^b f(x)dx \approx \sum_{j=0}^{n} A_j f(x_j),$$

A_j 是不依赖于$f(X)$ 的常数。上式等式成立，如果$f(x)$ 是不高于n次的多项式，但是如果$f(x)$是$n+1$次或者更高的多项式的时候，等式不成立。

下面我们将介绍一种新的方法，使得对于更高阶的多项式，等式也是成立的。这种方法的代价就是,插值点$x_1, \dots, x_n$的位置不是固定的，需要计算。

基本想法是这样的：定义一个多项式

$$w_n(x) = \Pi_{i=1}^{n}(x - x_i),$$

易知$w(x_i)=0$. 对于任意一个多项式$P_k(x)=q(x)w_n(x)+r(x)$.

$$\int_a^b P_k(x)dx=\int_a^b q(x)w_n(x)dx+\int_a^b r(x)dx.$$

因为$r(x)$ 是小于n 次的多项式，所以n 个点完全可以确定$r(x)$, 故使用Lagrange 插值公式，可以拟合$r(x)$，从而$\int_a^b r(x)dx=\sum_{i=1}^n A_j r(x_j)$. 如果$\int_a^b q(x)w_n(x)dx\equiv 0$, 那么

$$\int_a^b P_k(x)dx=\sum_{i=1}^n A_j r(x_j)=\sum_{i=1}^n A_j P_k(x_j).\ \text{注意到}w_n(x_j)=0.$$

如何确定$x_1,x_2,\ldots,x_n$ 呢？事实上，$q(x)$ 如果是不高于$n-1$ 的多项式，即$q(x)=a_0+a_1x+\ldots a_{n-1}x^{n-1}$. 下面的方程组（$n$ 个未知数，n 个方程）：

$$\int_a^b x^k w_n(x)=0, k=0,1,\ldots,n-1,$$

确定n 个节点$x_1,x_2,\ldots,x_n$.

通过上面的描述我们知道，对于积分$\int_a^b f(x)dx$, 通过选取合适的节点$x_1,x_2,\ldots,x_n$，可以使得

$$\int_a^b f(x)dx=\sum_{i=1}^n A_j f(x_j)$$

对于任意的一个不高于$2n-1$的成立。

特别的，$a=-1,b=1,n=2$时，易得：

$$\int_{-1}^1 f(x)dx=f\left(-\frac{1}{\sqrt{3}}\right)+f\left(\frac{1}{\sqrt{3}}\right)$$

更一般的，考虑积分

$$\int_a^b p(x)f(x)dx.$$

采用和前面的讨论类似的思路，对于任意的多项式$P_k(x)$,

$$P_k(x)=p(x)q(x)w_n(x)+r(x)$$

如果$\int_a^b p(x)q(x)w_n(x)dx\equiv 0$, 那么

$$\int_a^b p(x)P_k(x)dx=\int_a^b p(x)r(x)dx=\sum_{i=1}^n A_j r(x_j)=\sum_{i=1}^n A_j P_k(x_j).$$

即：

$$\int_a^b p(x)f(x)dx = \sum_{i=1}^{n} A_j f(x_j),$$

对于任意的阶数小于$2n-1$ 的多项式都成立。

下面介绍几个常用的高斯型求积公式：

(1)Gauss-Legendre 求积公式$p(x)=1$,

$$\int_{-1}^{1} f(x)dx \approx \sum_{k=1}^{n} A_k f(x_k).$$

$x_k, k=1,\ldots,n$ 是Legendre 多项式L_n 的跟：

$$\begin{aligned} L_0(x) &= 1, \\ L_1(x) &= x, \\ L_2(x) &= \tfrac{1}{2}(3x^2-1), \\ \cdots\cdots &\cdots \cdots\cdots\cdots, \\ L_n(x) &= \tfrac{1}{2^n n!}\tfrac{d^n}{dx^n}[(x^2-1)^n] \end{aligned}$$

$$A_k = \int_{-1}^{1} \frac{L_n(x)}{(x-x_k)L_n'(x)}dx = \frac{2}{(1-x_k^2)[L_n'(x_k)]^2}$$

(2) Gauss-Laguerre 求积公式

当$p(x)=e^{-x}$, 区间为$[0,\infty)$, 求积公式为

$$\int_0^{\infty} e^{-x} f(x)dx \approx \sum_{1}^{n} A_k f(x_k),$$

而节点是Laguerre 多项式的根：

$$L_n(x) = e^x \frac{d^n}{dx^n}(e^{-x}x^n).$$

$$A_k = \frac{(n!)^2}{x_k[L_n'(x)]^2}.$$

(3) Gauss-Hermite 求积公式

当$p(x)=e^{-x^2}$, 区间为$(-\infty,\infty)$, 求积公式为

$$\int_0^{\infty} e^{-x^2} f(x)dx \approx \sum_{1}^{n} A_k f(x_k),$$

而节点是Hermite 多项式的根：

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2}).$$

$$A_k = \frac{2^{n+1} n! \sqrt{\pi}}{[H_n'(x)]^2}.$$

2.2.2 函数逼近法

(一) 有理函数逼近(Padé 逼近) Padé 逼近是以幂级数展开为基础的，设$f(x)$ 在$|x| \leq 1$ 内可以展成幂级数

$$f(x) = \sum_{k=0}^{\infty} c_k x^k.$$

又设$m \geq n$ 为非负整数：

$$P_m(x) = \sum_{j=0}^{m} a_j x^j, \quad Q_n(x) = \sum_{k=0}^{n} b_k x^k.$$

用有理式

$$R_{mn}(x) = \frac{P_m(x)}{Q_n(x)}$$

来近似$f(x)$ 称为Padé 近似。

利用

$$\sum_{j=0}^{m} a_j x^j = \left(\sum_{k=0}^{\infty} c_k x^k \right) \left(\sum_{k=0}^{n} b_k x^k \right)$$

建立$m+n$ 个线性方程构成的方程组，可以求得系数$a_j, j = 0, \ldots, m$ 和$b_k, k = 0, \ldots, n$.

(二) 连分式逼近

定义9. 形如

$$b_0 + \cfrac{a_1}{b_1 + \cfrac{a_2}{b_2 + \cfrac{a_3}{\ddots + \cfrac{a_n}{b_{n-1} + \cfrac{a_n}{b_n}}}}}$$

的表达式称为n节连分式。为书写方便，有时把连分式写成

$$b_0 + \frac{a_1}{b_1} + \frac{a_2}{b_2} + \cdots + \frac{a_n}{b_n}$$

连分式有两种算法：从后面向前递推和从前面向后递推。

从后面向前递推：

$$q_n = b_n; q_{k-1} = b_{k-1} + \frac{a_k}{q_k}, k = n, n-1, \ldots, 2, 1$$

q_0 就是连分式的值。

从前面向后递推: 记$\frac{A_0}{B_0}=\frac{b_0}{1}$, $\frac{A_1}{B_1}=b_0+\frac{a_1}{b_1}=\frac{b_0b_1+a_1}{b_1}$, 则有递推公式

$$\frac{A_k}{B_k}=\frac{b_kA_{k-1}+a_kA_{k-2}}{b_kB_{k-1}+a_kB_{k-2}}, k=2,3,4,\ldots.$$

Proof. 用归纳法。假设结论对$k\le m$ 成立，当$k=m+1$ 时，

$$\frac{A_{m+1}}{B_{m+1}}=\frac{(b_m+\frac{a_{m+1}}{b_{m+1}})A_{m-1}+a_mA_{m-2}}{(b_m+\frac{a_{m+1}}{b_{m+1}})B_{m-1}+a_mB_{m-2}}=\frac{A_m+\frac{a_{m+1}}{b_{m+1}}A_{m-1}}{B_m+\frac{a_{m+1}}{b_{m+1}}B_{m-1}}=\frac{b_{m+1}A_m+a_{m+1}A_{m-1}}{b_{m+1}B_m+a_{m+1}B_{m-1}}$$

□

对于任意一形如

$$\sum_{i=0}^{\infty}c_ix^i/\sum_{j=0}^{\infty}d_jx^j$$

的函数，都可以采用以下方法化为连分式函数：

$$\begin{aligned}\frac{\sum_{i=0}^{\infty}c_ix^i}{\sum_{i=0}^{\infty}d_ix^i} &= \frac{c_0}{d_0}+\frac{\sum_{i=0}^{\infty}c_ix^i}{\sum_{i=0}^{\infty}d_ix^i}-\frac{c_0}{d_0}\\ &= \frac{c_0}{d_0}+\frac{\sum_{i=0}^{\infty}(c_id_0-d_ic_0)x^i}{\sum_{i=0}^{\infty}d_0d_ix^i}\\ &= \frac{c_0}{d_0}+\frac{x}{\sum_{i=0}^{\infty}d_0d_ix^i/\sum_{i=0}^{\infty}(c_{i+1}d_0-d_{i+1}c_0)x^i}\\ &=\cdots.\end{aligned}$$

设$F(x)$ 为分布函数，如果有

$$F(x)=\sum_{i=0}^{n}c_ix^i,$$

通过把$F(x)=\sum_{i=0}^{n}c_ix^i/1$ 化为连分式，用有限节连分式作为$F(x)$ 的近似式。这种方法就是连分式逼近法。

可以证明：

$$\sum_{i=0}^{\infty}c_ix^i=\frac{a_0}{1}\,\_\,\frac{a_1x}{1+a_1x}\,\_\,\frac{a_2x}{1+a_2x}\,\_\,\cdots$$

其中$a_0=c_0, a_i=c_i/c_{i-1}, i=1,2,\ldots.$

Proof. 我们用归纳法证明

$$\sum_{i=0}^{n} c_i x^i = \frac{a_0}{1} \underset{-}{} \frac{a_1 x}{1+a_1 x} \underset{-}{} \frac{a_2 x}{1+a_2 x} \underset{-}{} \cdots \underset{-}{} \frac{a_n x}{1+a_n x}.$$

设$n \le k$成立。当$n = k+1$时

$$\begin{aligned}
\frac{a_0}{1} \underset{-}{} \frac{a_1 x}{1+a_1 x} \underset{-}{} \frac{a_2 x}{1+a_2 x} \underset{-}{} \cdots \underset{-}{} \frac{a_k x}{1+a_k x} \underset{-}{} \frac{a_{k+1} x}{1+a_{k+1} x} &= \frac{a_0}{1} \underset{-}{} \frac{a_1 x}{1+a_1 x} \underset{-}{} \frac{a_2 x}{1+a_2 x} \underset{-}{} \cdots \underset{-}{} \frac{a_k x}{1+a_k x - \frac{a_{k+1} x}{1+a_{k+1} x}} \\
&= \frac{a_0}{1} \underset{-}{} \frac{a_1 x}{1+a_1 x} \underset{-}{} \frac{a_2 x}{1+a_2 x} \underset{-}{} \cdots \underset{-}{} \frac{a_k x}{a_k x + \frac{1}{1+a_{k+1} x}} \\
&= \frac{a_0}{1} \underset{-}{} \frac{a_1 x}{1+a_1 x} \underset{-}{} \frac{a_2 x}{1+a_2 x} \underset{-}{} \cdots \underset{-}{} \frac{a_k x}{a_k x + \frac{1}{1+a_{k+1} x}} \\
&= \frac{a_0}{1} \underset{-}{} \frac{a_1 x}{1+a_1 x} \underset{-}{} \frac{a_2 x}{1+a_2 x} \underset{-}{} \cdots \underset{-}{} \frac{a_k (1+a_{k+1} x) x}{a_k (1+a_{k+1} x) x + 1} \\
&= \frac{a_0}{1} \underset{-}{} \frac{a_1 x}{1+a_1 x} \underset{-}{} \frac{a_2 x}{1+a_2 x} \underset{-}{} \cdots \underset{-}{} \frac{\tilde{a}_k x}{\tilde{a}_k x + 1}
\end{aligned}$$

由假设可以得到

$$\text{上式} = \sum_{i=0}^{k} \tilde{c}_i x_i,$$

其中$\tilde{c}_i = \Pi_{i=0}^{i} a_i$. 所以，$\tilde{a}_i = c_i, i = 1, 2 \ldots, k-1$. $\tilde{c}_k = a_{k-1} * \tilde{a}_k = c_{k-1} * a_k (1 + a_{k+1} x) = c_{k-1} * c_k / c_{k-1} * (1 + c_{k+1}/c_k x) = c_k + c_{k+1} x$. □

用连分式代替级数的一个好处是能够减少计算量。例：计算$sin(\alpha x) = \alpha x - \frac{1}{3!}(\alpha x)^3 + \frac{1}{5!}(\alpha x)^5 - \cdots$. 如果使用多项式逼近

$$sin(\alpha x) \approx x[A + x^2 (B + C x^2)]$$

需要使用　四次乘法　+　两次加法。

如果使用连分式近似

$$sin(\alpha x) \approx a_0 (x - \frac{a_1}{x + \frac{a_2}{x}})$$

$$a_0 = -\frac{7}{3}\alpha, a_1 = \frac{200}{7\alpha^2}, a_2 = \frac{20}{\alpha^2}.$$

需要两次除法+ 两次加法+ 一次乘法。

例：计算$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$. $c_k = \frac{1}{k!}, a_0 = 1, a_k = \frac{c_k}{c_{k-1}} = \frac{1}{k}$.

$$e^x = \frac{1}{1} \underset{-}{} \frac{x}{1+x} \underset{-}{} \frac{1/2x}{1+1/2x} \underset{-}{} \cdots.$$

2.2.3 利用分布函数之间的关系

利用分布函数之间的关系，由一些分布的分布函数，可以得到另外一些分布的分布函数。

2.3 计算分位数的一般方法

$$F(x_p) = p.$$

$$f(x) = F(x) - p = 0 \text{ 的根.}$$

2.3.1 方程求根的迭代算法

(一) 二分法

(二) 牛顿法（切线法）

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) = 0$$

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

(三) 割线法

2.3.2 分位数的迭代算法

(一) 分位数的一个展开式$x_p = F^{-1}(p)$. 指数分布的p分位数：$F(x) = 1 - e^{-\lambda x} = p$, $x_p = -\frac{1}{\lambda}\ln(1-p)$.

对于一般的F,反函数没有明显的表达式。但是我们可以利用Taylor 展开，得到$F^{-1}(p)$ 的展开式。具体步骤如下：

记$\xi(x) = F(x) - F(x_0)$, 则$\xi(x_0) = 0$. 由此可以得到$\xi(x)$ 的反函数$x = x(\xi)$, 且$x(0) = x_0$.

现在我们将$x(\xi)$ 在$\xi = 0$ 处Taylor 展开，即

$$x(\xi) = x(0) + \sum_{k=1}^{\infty}\left(\frac{1}{k!}\frac{d^k x(\xi)}{d\xi^k}|\xi = 0\right)(\xi)^k.$$

因为$F(x_p) = p$, 所以当$\xi = p - F(x_0)$时，$x(\xi) = x_p$. 将$\xi = p - F(x_0)$ 带入上式，有

$$x_p = x(0) + \sum_{k=1}^{\infty}\left(\frac{1}{k!}\frac{d^k x(\xi)}{d\xi^k}|\xi = 0\right)(p - F(x_0))^k.$$

现在我们来计算$\frac{d^k x(\xi)}{d\xi^k}$：

$$\frac{dx}{d\xi} = 1/\frac{d\xi}{dx} = \frac{1}{f(x)} \equiv \frac{C_1(x)}{f(x)} \text{ 其中} C_1(x) = 1$$

$$\frac{d^2x}{d\xi^2} = (\frac{1}{f(x)})' * 1/\frac{d\xi}{dx} = \frac{-f'(x)}{f^3(x)} = \frac{C_1'(x) + C_1(x)g(x)}{f^2(x)} \equiv \frac{C_2(x)}{f(x)} \text{ 其中} g(x) = \frac{f'(x)}{f(x)}$$

一般的，

$$\frac{d^k x(\xi)}{d\xi^k} = \frac{C_k(x)}{f^k(x)},$$

$$C_{k+1} = kC_k(x)g(x) + C_k'(x), k = 1, 2,$$

$$\frac{d^k x(\xi)}{d\xi^k}|\xi = 0 = \frac{C_k(x_0)}{f^k(x_0)}$$

$$x_p = x(0) + \sum_{k=1}^{\infty} \left(\frac{1}{k!} \frac{C_k(x_0)}{f^k(x_0)} \right) (p - F(x_0))^k.$$

2.4 一些特殊分布函数和分位数的计算

2.4.1 正态分布

(一) 几个基本公式

定义10. 称函数

$$erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt (x > 0)$$

为误差函数；称函数

$$erfc(x) = 1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt (x > 0)$$

为余误差函数；

显然

$$\Phi(x) = \begin{cases} 0.5 \left(1 + erf(\frac{x}{\sqrt{2}})\right), & x \geq 0; \\ 0.5 \left(1 - erf(\frac{-x}{\sqrt{2}})\right), & x < 0. \end{cases} \tag{2.1}$$

利用分部积分可以得到$\Phi(x)$ 的两个级数展开式：

$$\Phi(x) = \frac{1}{2} + \phi(x) \left[x + \frac{x^3}{3} + \frac{x^5}{3 \cdot 5} + \cdots + \frac{x^{2k+1}}{(2k+1)!!} + \cdots \right]$$

$$\Phi(x) = 1 - \phi(x)\left[\frac{1}{x} - \frac{1}{x^3} + \frac{3}{x^5} + \cdots + (-1)^k \frac{(2k-1)!!}{x^{2k+1}} + \cdots\right]$$

(二) $\Phi(x)$ 的计算方法

$\Phi(x) = 1 - \Phi(-x)$. 只需考虑$x > 0$ 的情形。

(1) 使用连分式计算

利用将函数化为连分式的一般方法，可以将$\Phi(x)$化为如下两个连分式；

$$\Phi_1(x) = \frac{1}{2} + \frac{\phi(x)x}{1} \underset{-}{} \frac{x^2}{3} \underset{+}{} \frac{2x^2}{5} \underset{+}{} \cdots \underset{+}{} (-1)^k \frac{kx^2}{2k+1} \underset{+}{} \cdots$$

$$\Phi_2(x) = 1 - \frac{\phi(x)}{x} \underset{+}{} \frac{1}{x} \underset{+}{} \frac{2}{x} \underset{+}{} \cdots \underset{+}{} \frac{k}{x} \underset{+}{} \cdots$$

截有限节作为$\Phi(x)$ 的近似式：

$$\Phi(x) = \begin{cases} \Phi_1(x), & 0 \le x \le 3; \\ \Phi_2(x), & x > 3. \end{cases}$$

$n = 28$时，精度可达10^{-12}.

(2) 用误差函数的近似公式计算：

$$erf(x) \approx 1 - (1 - \sum_{i=1}^{6} a_i x^i)^{-16}$$

精度可达1.3×10^{-7}

$$erf(\frac{x}{\sqrt{2}}) \approx 1 - (1 - \sum_{i=1}^{4} b_i x^i)^{-4}$$

精度可达2.5×10^{-4}.s

（三）分位数u_p 的计算

(1) 用u_p 的近似公式

(2) 利用分位数的展开式算法

2.4.2 Beta 分布

Beta 分布的递推算法

回顾Beta 分布的密度函数为

$$beta(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, x \in [0,1]$$

定义一个新的函数

$$U(x|a,b) = \frac{1}{B(a,b)} x^{a}(1-x)^{b}, x \in [0,1]$$

Beta 分布的分布函数$Beta(x|a,b)$ 有以下的递推公式：

$$\begin{cases} Beta(x|a+1,b) &= Beta(x|a,b) - \frac{1}{a}U(x|a,b) \\ Beta(x|a,b+1) &= Beta(x|a,b) + \frac{1}{b}U(x|a,b) \\ U(x|a+1,b) &= \frac{a+b}{a}xU(x|a,b) \\ U(x|a,b+1) &= \frac{a+b}{b}(1-x)U(x|a,b) \end{cases} \tag{2.2}$$

Homework.

利用$Beta$ 分布的分布函数计算t分布，F 分布，二项分布时，参数a,b都是正整数或者$\frac{1}{2}$的整数倍。这时，上式 (2.2) 的初值选取只有一下四种情况：

(1) $a = \frac{1}{2}, b = \frac{1}{2}$

$$U(x|\frac{1}{2},\frac{1}{2}) = \frac{1}{\pi}\sqrt{x(1-x)}$$

$$Beta(x|\frac{1}{2},\frac{1}{2}) = 1 - \frac{2}{\pi}\tan^{-1}\sqrt{\frac{1-x}{x}}$$

Proof.

$$\begin{aligned} Beta(x|\frac{1}{2},\frac{1}{2}) &= \frac{1}{B(\frac{1}{2},\frac{1}{2})}\int_0^x t^{-\frac{1}{2}}(1-t)^{-\frac{1}{2}}dt \\ &= (t = sin^2 u) \quad \frac{1}{B(\frac{1}{2},\frac{1}{2})}\int_0^{u_0} \frac{1}{\sin u \cos u} 2\sin u \cos u du \\ &= \frac{2u_0}{B(\frac{1}{2},\frac{1}{2})}. \end{aligned}$$

注意到$\sin^2 u_0 = x$, 如果$x = 1$, $u_0 = \frac{\pi}{2}$, 所以$B(\frac{1}{2},\frac{1}{2}) = \pi$. 所以上式等于

$$\frac{2}{\pi}\arcsin\sqrt{x} = 1 - \frac{2}{\pi}\arctan\sqrt{\frac{1-x}{x}}.$$

□

(2) $a = \frac{1}{2}, b = 1$

$$U(x|\frac{1}{2},1)=\frac{1}{2}\sqrt{x}(1-x)$$

$$Beta(x|\frac{1}{2},1)=\sqrt{x}$$

(3) $a=1,b=\frac{1}{2}$

$$U(x|1,\frac{1}{2})=\frac{1}{2}x\sqrt{(1-x)}$$

$$Beta(x|1,\frac{1}{2})=1-\sqrt{1-x}$$

(4) $a=1,b=1$

$$U(x|1,1)=x(1-x)$$

$$Beta(x|1,1)=x$$

练习。

2.4.3 χ^2 分布

χ^2 分布的递推算法

$$\begin{cases} Gamma(x|n) &= Gamma(x|n-2)+2f(x|n), \\ f(x|n) &= \frac{x}{n-2}f(x|n-2) \end{cases}$$

其中，$f(x|n)=\frac{1}{2\Gamma(\frac{n}{2})}(\frac{x}{2})^{\frac{n}{2}-1}e^{-\frac{x}{2}}$. 递推初值为：

$$\begin{cases} Gamma(x|1) &= 2\Phi(\sqrt{x})-1, f(x|1)=\frac{1}{\sqrt{2\pi x}}e^{-\frac{x}{2}}, \\ Gamma(x|2) &= 1-e^{-\frac{x}{2}}, f(x|2)=\frac{1}{2}e^{-\frac{x}{2}}. \end{cases}$$

注意到$\Gamma(\frac{1}{2})=\sqrt{\pi}$. 这可以由

$$B(a,b)=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

得到。

2.4.4 Gamma 分布

回顾Gamma 分布的密度函数

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

注意到如果$X \sim \Gamma(a,b)$, 则$Y = bX \sim \Gamma(a,1)$.

Γ 分布的递推公式.

$$G(x|a+1) = G(x|a) - \frac{1}{a} V(x|a)$$

$$V(x|a+1) = \frac{x}{a} V(x|a)$$

$$V(x|a) = \frac{1}{\Gamma(a)} x^a e^{-x}$$

Γ 分布的连分表达式可以采用和正态分布类似的方法，将Gamma 分布使用分部积分得到基数展开，然后使用连分式计算。

2.4.5 t 分布, F 分布,二项分布和泊松分布

这些分布可以使用Beta分布或者χ^2 分布计算

Chapter 3

非均匀随机数的产生

3.1 产生非均匀随机数的一般方法

(一) 直接抽样法（反函数法）

由前面的定理我们知道，如果$R \sim U[0,1]$, 那么$F^{-1}(R) \sim F(x)$, 其中，$F(x)$ 为任一分布函数。

特别的，如果$\xi \in [a,b] \sim F(x)$, 那么在计算$F^{-1}(x)$时，只需考虑$F(x), x \in [a,b]$的反函数。

例：产生$U[a,b]$ 的均匀随机数。

$$F(x) = \int_a^x \frac{1}{b-a} dt = \frac{x-a}{b-a}, \forall x \in [a,b].$$

$$F^{-1}(y) = a + (b-a)y, \forall y \in (0,1).$$

所以$\xi = (b-a)R + a$ 服从$U(a,b)$.

例：产生指数分布的随机数。$f(x) = \lambda e^{-\lambda x}, x \in [0,\infty)$

$$F(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}, \forall x \in [0,\infty)$$

$$F^{-1}(y) = -\frac{1}{\lambda}\log(1-y), \forall y \in (0,1).$$

所以，$\xi = -\frac{1}{\lambda}\log(1-R) =_d -\frac{1}{\lambda}\log(R)$ 服从指数分布。

例：离散随机变量的随机数的产生。

$$P(\xi = x_i) = p_i, i = 1, 2, \ldots$$

$$F(x) = \sum_{x_i \le x} p_i$$

是一个阶梯函数。根据$F^{-1}(x)$的定义：

$$F^{-1}(y) = \inf_x \{x : F(x) \ge y\},$$

所以

$$F^{-1}(y) = x_i, F(x_{i-1}) < y \le F(x_i), i = 1, 2, \ldots$$

这里$F(x_0) = 0$. 产生离散分布的随机数的直接抽样法如下：

1. 产生$R \sim U(0, 1)$

2. 取$\xi = x_i$, 若$F(x_{i-1}) < R \le F(x_i)$.

如果把y轴单独抽出来看，这实际上是一个Stick breaking 方法。

1. 由$p_k, k = 1, 2, \ldots$把长度为一的区间$[0, 1]$依次分为长度为$p_1, p_2, \ldots$的小区间$(F(x_{i-1}), F(x_i)]$;

2. 产生R, 如果R 落入第k个小区间$(F(x_{k-1}), F(k_i)]$, 令$\xi = x_k$.

直接抽样法的缺点：

如果$F(x)$的反函数很难计算，将很难使用此方法。

(二) 变换抽样法

X 具有密度函数$f(x)$, $Y = g(X)$是随机变量X 的函数，又设$x = g^{-1}(y) = h(y)$存在且具有一阶连续导数。则$Y = g(X)$的密度函数为：

$$p(y) = f(h(y))|h'(y)|.$$

例：$Y = R^\alpha, \alpha > 0$,则$h(y) = y^{1/\alpha}, h'(y) = 1/\alpha y^{1/\alpha - 1}$.

$$p(y) = \frac{1}{\alpha} y^{1/\alpha - 1} \quad Beta(1/\alpha, 1).$$

设随机向量(X, Y)具有二维联合密度$f(x, y)$, 考虑变换

$$U = g_1(X, Y), V = g_2(X, Y)$$

如果该变换具有一些好的性质：

1. 反变换存在唯一

$$x = h_1(u, v), y = h_2(u, v).$$

2. 反变换的一阶偏岛存在

3. 函数变换的雅各比行列式不为0

则(U,V) 具有联合密度

$$f[h_1(u,v),h_2(u,v)]|J|,$$

其中J 为雅各比行列式：

$$\begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

由此可以得到，正态分布随机数产生的变换抽样法：若$U_1,U_2\sim U(0,1)$ 且相互独立，令

$$\begin{cases} X_1 &= \sqrt{-2\ln(U_1)}\cos(2\pi U_2), \\ X_2 &= \sqrt{-2\ln(U_1)}\sin(2\pi U_2), \end{cases}$$

(X_1,X_2) i.i.d. $N(0,1)$.

例：χ^2分布的变换抽样法:

$$R_1^2+R_2^2+\ldots R_n^2\sim\chi^2(n).$$

(三) 值序抽样法

定理6. 设$X_1,X_2,\ldots,X_n$ *i.i.d.* 密度函数为$f(x)$, 分布函数为$F(x)$，次序统计量记为

$$X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)},$$

则$(X_{(1)}\le X_{(2)}\le\cdots\le X_{(n)})$的联合密度函数为

$$g(y_1,\ldots,y_n)=n!f(y_1)f(y_2)\cdots f(y_n)I_{y_1\le y_2\le\cdots y_n}$$

X_ℓ 的密度函数为

$$\frac{n!}{(\ell-1)!(n-\ell)!}[F(x)]^{l-1}[1-F(x)]^{n-l}f(x)$$

X_ℓ 的分布函数为

$$\frac{n!}{(\ell-1)!(n-\ell)!}\int_0^{F(x)}t^{l-1}[1-t]^{n-l}dt$$

特别的，$X_i,i=1,\ldots,n$ *i.i.d.* $U(0,1)$, 则$X_{(\ell)}$ 的密度为

$$\frac{n!}{(\ell-1)!(n-\ell)!}x^{l-1}(1-x)^{n-l},x\in(0,1)\quad[Beta(\ell,n-\ell+1)]$$

一般的，取 $n=a+b-1$, 则$X_{(a)}\sim Beta(a,b)$.

(四) 舍选抽样法

上述各方法都可以成为直接抽样法。接着我们介绍舍选抽样法。该方法不是对所有产生的随机数都录用，而是采用一种筛选的方法，满足一定条件的随机数才被录用。例如：产生圆内的均匀随机数。

下面我们介绍几种常用的舍选抽样法。(1) 设随机变量Z的分布密度为$p(z)$, 其有上界函数$M(z)$:

$$p(z) \leq M(z),$$

且

$$\int_{-\infty}^{\infty} M(x) < \infty.$$

令$f(x) = M(x)/C$, 如果密度为$f(x)$ 的随机变量很易抽取，则我们首先抽取

$$X \sim f(x)$$

接着我们给定一个规则决定选还是舍弃此随机数。具体的规则如下：选取一个均匀随机数$R \sim U[0,1]$, R 与X 独立. 如果$R \leq \frac{p(X)}{M(X)}$, 则接受此随机数，令$Z = X$,否则舍弃。可以证明$Z \sim p(z)$.

Proof.

$$\begin{aligned} P(Z \leq z) &= P(X \leq z | R \leq \frac{p(X)}{M(X)}) \\ &= P(X \leq z, R \leq \frac{p(X)}{M(X)}) / P(R \leq \frac{p(X)}{M(X)}) \end{aligned}$$

分子是

$$P(X \leq z, R \leq \frac{p(X)}{M(X)}) = \int_{-\infty}^{z} [\int_{0}^{\frac{p(x)}{M(x)}} dr] f(x)dx = \int_{-\infty}^{z} \frac{p(x)}{M(x)} M(x)/C dx = \int_{-\infty}^{z} p(x)/C dx$$

分母是$\int_{-\infty}^{\infty} p(x)/C dx = 1/C$

所以

$$P(Z \leq z) = \int_{-\infty}^{\infty} p(x)dx$$

□

一般的结果：

定理7. *1.* $X \sim f(x)$

2. $Y \sim g(x)$ $[G(x)]$ 且X 与Y 独立

3. 接受X 如果$Y \leq h(X)$*:* $Z = X$*;* 否则放弃X*.*

则Z的密度为

$$p(z) \propto f(z)G(h(z)).$$

Proof.

$$P(Z \leq z) = P(X \leq z | Y \leq h(X)) \propto P(X \leq z, Y \leq h(X)) = \int_{-\infty}^{z} G(h(x))f(x)dx.$$

$$p(z) \propto f(z)G(h(z)).$$

□

例：试产生密度函数为$f(x)$ 的随机数$\xi \in [a, b]$ 且$\sup_x \in [a, b] f(x) = f_0 < \infty$.

取$M(x) = f_0 I_{x \in [a,b]}$.

1. 产生均匀随机数$X \sim U[a, b]$

2. 产生$R \sim U(0, 1)$

3. 若$R \leq p(X)/f_0$, 令$Z = X$,否则放弃X.

几何意义：

称

$$p_0 = P(R \leq p(x)/f_0) = \frac{1}{f_0(b-a)}$$

为舍选抽样法的效率。引入随机变量T 表示需要产生多少个X 才可以被接受一次。则T 服从“成功一次的概率为p_0”的几何分布：

$$P(T = k) = (1 - p_0)^{k-1} p_0$$

$$E(T) = \frac{1}{p_0}.$$

例：$Beta(a, b)$ 的随机数。密度为：

$$f(x) = \frac{1}{B(a, b)} x^{a-1}(1 - x)^{b-1}, x \in [0, 1]$$

$$x = \frac{a-1}{a+b-2}$$

时, $f(x)$ 达到最大值D.

$a = 4, b = 5$时， $D \approx 2.35, p_0 = \frac{1}{D(1-0)} \approx 0.4255$. 效率为42.55%, 平均2.35 个$X$ 接受一个。

例：设$Z \sim p(z), z \in [0,2]$,已知

$$0.3 \leq p(z) \leq \frac{z+1}{2}, z \in [0,2].$$

怎样产生随机数Z??

取$M(x) = \frac{x+1}{2}$, 则$f(x) = \frac{M(x)}{2} = \frac{x+1}{4}$

1. 产生$R_1 \sim U(0,1)$, 则$X = \sqrt{1+8R_1} - 1 \sim f(x)$

2. 产生$R_2 \sim U(0,1)$且独立于R_1

3. 如果$R_2 \leq 0.3/M(X)$ 则接受$Z = X$, else

4. 如果$R_2 \leq p(x)/M(X)$, 则接受$Z = X$, 否则放弃X.

例：

$$a - \frac{b(x-s)}{h} \leq g(x) \leq b - \frac{b(x-s)}{h}, x \in (s, s+h)$$

1. $R_1, R_2 \sim U[0,1]$, $U = R_{(1)}, V = R_{(2)}$

2. 如果$V \leq a/b$ 接受$X = s + hU$, else

3. 如果$V \leq U + \frac{1}{b}g(s+hU)$ 接受$X = s + hU$, 否则放弃X。

X 的密度函数为$g(x)$.

Homework.

舍选法II:

如果

$$p(z) = Lh(z)f(z),$$

其中$L > 1, 0 \leq h(z) \leq 1$, $f(z)$ 为一密度函数。则Z 的抽样过程可以简化为：

1. 产生$X \sim f(x)$

2. 产生$R \sim U[0,1]$

3. 如果$R \leq h(X)$, 则接受X: $Z = X$;否则放弃X.

Proof. 取$M(z) = Lf(z)$, 则由$M(z)$的到的密度为$f(z)$.

$R \leq \frac{p(z)}{M(z)} = h(z)$. □

例：

$$p(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}}, x \in [0, \infty)$$

$$p(x) = \sqrt{\frac{2e}{\pi}} \cdot e^{-\frac{(x-1)^2}{2}} \cdot e^{-x}$$

$f(x) = e^{-x}$ 为指数分布，$F(x) = \int_0^x e^{-t}dt = 1 - e^{-x}$.

$h(x) = e^{-\frac{(x-1)^2}{2}}$.

1. $R_1 \sim U[0,1]$, $X = F^{-1}(R_1) = -\log(1-R_1) =_d -\log(R_1)$
2. $R_2 \sim U(0,1)$
3. 如果$R_2 \le e^{-\frac{(X-1)^2}{2}} \iff \frac{(X-1)^2}{2} \le -\log R_2$,接受$Z = X$; 否则放弃$X$.

舍选法III：

如果

$$p(z) = L\int_{-\infty}^{h(z)} g(z,y)dy,$$

则可以这样抽取Z:

1. $(X,Y) \sim g(x,y)$
2. 如果$Y \le h(X)$, 接受X: $Z = X$, 否则放弃X.

Proof.

$$P(Z \le z) = P(X \le z | Y \le h(X)) \propto P(X \le z, Y \le h(X)) = \int_{-\infty}^{z}[\int_{-\infty}^{h(x)} g(x,y)]dydx$$

□

(五) 复合抽样法

$$F(x) = \sum_j p_j F_j(x).$$

1. 产生随机整数J, $P(J=j) = p_j$
2. 产生$X \sim F_J(x)$. 则$X \sim F(x)$.

Proof.

$$P(X \le x) = \sum_j P(X \le x | J = j)P(J = j)$$

□

(六) 近似抽样法选读。P 141 - P 145.

3.2 常用连续分布的抽样方法

(一) 正态分布(1) 近似分布

(2) Box-Muller(1958) 变换抽样法：

$$\begin{cases} X_1 &= \sqrt{-2\ln(U_1)}\cos(2\pi U_2), \\ X_2 &= \sqrt{-2\ln(U_1)}\sin(2\pi U_2), \end{cases}$$

(3) 修正的变换抽样法。

注意到$U_2 \sim U[0,1]$,那么$\pi U_2 \sim U[0,\pi]$. 半圆内的一点(X,Y) 与$\alpha = U_2\pi$ 的关系为：$\sin\alpha = \frac{Y}{\sqrt{X^2+Y^2}}$ $\cos\alpha = \frac{X}{\sqrt{X^2+Y^2}}$ 从而$\sin 2\alpha = \frac{2XY}{X^2+Y^2}$ $\cos 2\alpha = \frac{X^2-Y^2}{X^2+Y^2}$

综上，我们有产生$N(0,1)$ 的随机数的方法：

1. 产生相互独立的均匀随机数r_1, r_2, r_3

2. 计算$u_1 = 2r_2 - 1$, $u_2 = r_3$

3. If $u_1^2 + u_2^2 <= 1$

$$\begin{cases} X_1 &= \sqrt{-2\ln(r_1)}\frac{u_1^2-u_2^2}{u_1^2+u_2^2}, \\ X_2 &= \sqrt{-2\ln(r_1)}\frac{2u_2u_2}{u_1^2+u_2^2}, \end{cases}$$

另一种类似方法：

1. $V_1 = 2r_1 - 1, V_2 = 2r_2 - 1$

2. If $W \le 1$, 其中$W = V_1^2 + V_2^2$, 则令$Y = \sqrt{\frac{-2\log W}{W}}, X_1 = V_1Y, X_2 = V_2Y$

其他近似方法，略去。

(二) 指数分布

密度函数

$$f(x) = \lambda e^{-\lambda x}, \lambda > 0; x \ge 0$$

$$F(x) = 1 - e^{-\lambda x}$$

$$F^{-1}(y) = -\frac{1}{\lambda}\log(1-y), \text{ 由此得到直接法}$$

显然，如果$X \sim Exp(\lambda)$, $\lambda X \sim Exp(1)$.

(三)Gamma 分布

(四) Beta 分布

高效舍选抽样法。利用不等式：

$$(\frac{x}{A})^A(\frac{1-x}{B})^B C^C \le \exp[-2C(x-\frac{A}{C})^2],$$

其中，$A = a-1, B = b-1, C = A+B$.

$a = 4, b = 5$, 利用这个上界可以得到效率为$0.8983 > 0.4255$.

(五) χ^2 、t、F 分布

3.3 常用离散分布的抽样方法

(一) 二项分布$X_i \sim Bernoulli(p)$

$$\sum_{i=1}^{n} X_i \sim Binomial(n,p)$$

$n < 38$ 时，可采用直接抽样X_i,然后求和的方法。

引理2. 设a,b 为正整数，$a+b-1=n$. 又设S 服从$\beta(a,b)$.

1. 如果$S \le p$, 设$Y \sim Binomial(b-1, \frac{p-S}{1-S})$, 令$X = Y + a$;

2. 如果$S > p$, 设$Y \sim Binomial(a-1, \frac{p}{S})$, 令$X = Y$.

则$X \sim Binomial(n,p)$.

(二) 泊松分布

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

注意到泊松过程中，泊松分布和指数分布之间的关系：

(三)几何分布和负二项分布

3.4 随机向量的抽样方法

Chapter 4

随机模拟方法

Chapter 5

EM 算法

5.1 EM 算法

EM 算法全称是Expectation-Maximum Algorithm. 其主要目的是用于求MLE. 当数据含有缺失数据，或者虽然没有缺失数据，但是引进隐变量（可以看成是缺失数据）之后，似然函数变得简单，这时候，使用EM 算法会使得MLE 求解变得简单。

首先我们来看一个例子。基因连锁模型（Rao, 1973）：观测到197 个动物，根据其属性，可以将这些动物分为4 类，而且模型假定这四类占的概率分别是：

$$\frac{1}{2}+\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}.$$

用一个表表示观测及概率分布：

125	18	20	34
$\frac{1}{2}+\theta$	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{\theta}{4}$

Table 5.1: 基因连锁模型及观测

似然函数是：

$$\ell(\theta|X) \propto p(X|\theta) \propto (\frac{1}{2}+\theta)^{125}(\frac{1}{4}(1-\theta))^{18+20}(\frac{\theta}{4})^{34} \propto (2+\theta)^{125}(1-\theta)^{18+20}(\theta)^{34}$$

θ 的极大似然估计是

$$\hat{\theta} = \arg\max_{\theta} \log \ell(\theta) = \arg\max_{\theta} \left[125\log(2+\theta) + (18+20)\log(1-\theta) + 34\log(\theta)\right].$$

求解θ 可以使用Newton 法。

现在考虑引进一个隐变量，将表5.1 第一个格子分成两个格子，见下表：

$125-x_2$	x_2	18	20	34
$\frac{1}{2}$	θ	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{\theta}{4}$

Table 5.2: 引进隐变量后，基因连锁模型及观测

如果隐变量(x_2) 知道了，那么似然函数是：

$$\ell(\theta) = (\theta)^{x_2+34}(1-\theta)^{18+20},$$

这时候，可以很容易的求出

$$\hat{\theta} = \frac{x_2+34}{x_2+34+18+20}.$$

EM 算法就是一个迭代的方法，每一次迭代，将x_2 填补出来，然后很容易的计算$\hat{\theta}$.

5.1.1 EM 算法的推导

用X 代表观测数据，Z 代表缺失数据（或者隐变量），θ 是待估计的参数。观测数据的边缘分布是

$$P(X|\theta) = \sum_z P(X|Z,\theta)P(Z|\theta), \text{ 如果}Z \text{ 是连续的，用积分代替求和.}$$

这时候，对数似然是

$$\begin{aligned} L(\theta) &= \log\left[\sum_z P(X|Z,\theta)P(Z|\theta)\right] \\ &= \log\left[\sum_z P(Z|X,\theta_n)\frac{P(X|Z,\theta)P(Z|\theta)}{P(Z|X,\theta_n)}\right] \\ &\geq \sum_z P(Z|X,\theta_n)\log\left[\frac{P(X|Z,\theta)P(Z|\theta)}{P(Z|X,\theta_n)}\right] \\ &\triangleq \ell(\theta|\theta_n) \end{aligned}$$

从上面的不等式可以得到如下性质：

1. $L(\theta) \geq \ell(\theta|\theta_n)$.

2.

$$\begin{aligned}
\ell(\theta_n|\theta_n) &= \sum_z P(Z|X,\theta_n)\log\left[\frac{P(X|Z,\theta_n)P(Z|\theta_n)}{P(Z|X,\theta_n)}\right] \\
&= \sum_z P(Z|X,\theta_n)\log\left[\frac{P(X,Z|\theta_n)P(X|\theta_n)}{P(Z,X|\theta_n)}\right] \\
&= \sum_z P(Z|X,\theta_n)\log\left[P(X|\theta_n)\right] \\
&= \log P(X|\theta_n) = L(\theta_n)
\end{aligned}$$

所以，任何使得$\ell(\theta|\theta_n)$增大的θ将会使$L(\theta_{n+1}) \geq L(\theta_n)$. EM 算法寻求$\theta_{n+1} = \arg\max_\theta \ell(\theta|\theta_n)$.

$$\begin{aligned}
\theta_{n+1} &= \arg\max_\theta \ell(\theta|\theta_n) \\
&= \arg\max_\theta \sum_z P(Z|X,\theta_n)\log\left[\frac{P(X|Z,\theta)P(Z|\theta)}{P(Z|X,\theta_n)}\right] \\
&= \arg\max_\theta \sum_z P(Z|X,\theta_n)\log\left[\frac{P(X,Z|\theta)}{P(Z|X,\theta_n)}\right] \\
&= \arg\max_\theta \sum_z P(Z|X,\theta_n)\log\left[P(X,Z|\theta)\right] \\
&= \arg\max_\theta E_{Z|X,\theta_n}\log\left[P(X,Z|\theta)\right]
\end{aligned}$$

综上所述，从初始值θ_n出发，计算θ_{n+1} 的具体步骤可以总结下面两部：

1. E 步：计算条件期望$E_n(\theta) \triangleq E_{Z|X,\theta_n}\log P(X,Z|\theta)$ (完全似然的期望)

2. M 步：计算上一步所得期望的最大值点：$\theta_{n+1} = \arg\max_\theta E_n(\theta)$.

例子：基因连锁模型（续）。前面已经介绍了基因连锁模型，引进隐变量后，观测及其概率分布见下表：

x_1	x_2	x_3	x_4	x_5
$125-x_2$	x_2	18	20	34
$\frac{1}{2}$	θ	$\frac{1}{4}(1-\theta)$	$\frac{1}{4}(1-\theta)$	$\frac{\theta}{4}$

Table 5.3: 引进隐变量后，基因连锁模型及观测

完全似然为

$$f(\theta) = \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}.$$

1. E 步：$E_{Z|X,\theta_n}[(x_2+x_5)\log\theta + (x_3+x_4)\log(1-\theta)] = (E(x_2|X,\theta_n)+x_5)\log\theta + (x_3+x_4)\log(1-\theta)$. 注意到$x_2 \sim Binomial(x_1+x_2, \frac{\theta}{2+\theta})$. 所以，$E_n(\theta) = ((x_1+x_2)\frac{\theta_n}{2+\theta_n}+x_5)\log\theta + (x_3+x_4)\log(1-\theta)] = (E(x_2|X,\theta_n)+x_5)\log\theta + (x_3+x_4)\log(1-\theta)$.

2. M 步：

$$\theta_{n+1} = \arg\max_{\theta} E_n(\theta) = \frac{(x_1+x_2)\frac{\theta_n}{2+\theta_n} + x_5}{(x_1+x_2)\frac{\theta_n}{2+\theta_n} + x_5 + x_3 + x_4}$$

例：Gaussian Mixture Model.

$x_i \sim_{i.i.d.} pN(\mu_1, \sigma_1^2) + (1-p)N(\mu_2, \sigma_2^2), \mu_1 \neq \mu_2, i = 1, 2, \ldots, n.$

5.2 指数族的EM 算法

本节考虑完全似然是指数族，定义如下：

$$p(Z, X|\theta) = \phi(Z, X)\psi(\xi(\theta)) \exp\{\xi(\theta)^T t(Z, X)\}.$$

这时候称$t(Z, X)$ 为充分统计量。

我们知道，θ 的极大似然估计就是最大化下面的对数似然：

$$L(\theta) = \log \psi(\xi(\theta)) + \xi(\theta)^T t(Z, X).$$

现在看一下，EM 算法，其E 步是：

$$E_n(\theta) = E_{Z|X,\theta_n}(L(\theta)) = \log \psi(\xi(\theta)) + \xi(\theta)^T E_{Z|X,\theta_n} t(Z, X).$$

注意到，上式和前面的对数似然不一样的地方，就是使用$E_{Z|X,\theta_n} t(Z, X)$ 代替了充分统计量$t(Z, X)$.

所以，对于指数族，EM 算法重点就是求充分统计量的条件期望，然后用该条件期望代替充分统计量计算θ_{n+1}.

例：（基因连锁模型。）完全似然是：

$$(\theta)^{x_2+x_5}(1-\theta)^{x_3+x_4} = \exp\{(x_2+x_5)\log\theta + (x_3+x_4)\log(1-\theta)\}$$

所以，充分统计量是$(x_2 + x_5, x_3 + x_4)$ 且θ 的MLE 是

$$\theta = \frac{x_2+x_5}{x_2+x_5+x_3+x_4}.$$

所以由前面的指数族的EM 算法知，M 步为：

$$\theta_{n+1} = \frac{E(x_2|X, \theta_n) + x_5}{E(x_2|X, \theta_n) + x_5 + x_3 + x_4},$$

其中$E(x_2|X,\theta_n)=(x_1+x_2)\frac{\theta_n}{2+\theta_n}$

例：含有缺失数据的列联表。(X_1,X_2)

5.3 利用EM 算法计算MLE 的方差

经典MLE 方法有一个重要的结论：在正则条件下，θ 的MLE $\hat{\theta}$ 有如下性质：

$$\sqrt{n}(\hat{\theta}-\theta)\Rightarrow N(0,I^{-1}),$$

其中I 是Fisher 信息阵，定义为:

$$I(\theta)=-E\left(\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right).$$

所以，为了计算MLE 的方差，就要计算

$$-\frac{\partial^2}{\partial\theta^2}\log f(x|\theta),$$

因为上式只和观测数据有关，我们称上式为观测信息。

计算观测信息的方法可以有许多种，比如数值微分的方法。本节介绍更精确的方法，使用EM 算法的结果计算。一个重要的结果：

定理8.

$$-\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)=E_{Z|x,\theta}\left\{-\frac{\partial^2}{\partial\theta^2}\log f(x,Z|\theta)\right\}-E_{Z|X,\theta}\left\{-\frac{\partial^2}{\partial\theta^2}\log f(Z|x,\theta)\right\}.$$

上面的结果又称为缺失信息原理，用语言描述为：观测信息= 完全信息− 缺失信息。

Proof.

$$f(Z,X|\theta)=f(X|\theta)f(Z|X,\theta),$$

$$\log f(Z,X|\theta)=\log f(X|\theta)+\log f(Z|X,\theta),$$

$$-\frac{\partial^2}{\partial\theta^2}\log f(Z,X|\theta)=[-\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)]+[-\frac{\partial^2}{\partial\theta^2}\log f(Z|X,\theta)],$$

两边求期望，

$$E_{Z|X,\theta}[-\frac{\partial^2}{\partial\theta^2}\log f(Z,X|\theta)]=E_{Z|X,\theta}[-\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)]+E_{Z|X,\theta}[-\frac{\partial^2}{\partial\theta^2}\log f(Z|X,\theta)],$$

即：

$$E_{Z|X,\theta}[-\frac{\partial^2}{\partial\theta^2}\log f(Z,X|\theta)] = [-\frac{\partial^2}{\partial\theta^2}\log f(X|\theta)] + E_{Z|X,\theta}[-\frac{\partial^2}{\partial\theta^2}\log f(Z|X,\theta)].$$

□

缺失信息或许不是很容易计算，接下来我们介绍Louis 方法。该方法证明缺失信息等于完全得分的（条件）方差：

定理9 (Louis' Method)**.**

$$E_{Z|X,\theta}\left\{-\frac{\partial^2}{\partial\theta^2}\log f(Z|X,\theta)\right\} = var\left[\frac{\partial\log f(Z,X|\theta)}{\partial\theta}|X,\theta\right].$$

为证明Louis 方法，我们首先介绍一个重要引理, 即：观测得分等于完全得分的期望：

引理3.

$$\frac{\partial\log f(X|\theta)}{\partial\theta} = E\left[\frac{\partial\log f(Z,X|\theta)}{\partial\theta}|X,\theta\right]$$

Proof.

$$f(Z,X|\theta) = f(X|\theta)f(Z|X,\theta),$$

所以，

$$\log f(X|\theta) = \log f(Z,X|\theta) - \log f(Z|X,\theta).$$

$$\frac{\partial}{\partial\theta}\log f(X|\theta) = \frac{\partial}{\partial\theta}\log f(Z,X|\theta) - \frac{\partial}{\partial\theta}\log f(Z|X,\theta).$$

两边对Z 求期望，有

$$\frac{\partial}{\partial\theta}\log f(X|\theta) = E_{Z|X,\theta}\left[\frac{\partial}{\partial\theta}\log f(Z,X|\theta)\right] - E_{Z|X,\theta}\left[\frac{\partial}{\partial\theta}\log f(Z|X,\theta)\right],$$

其中

$$\begin{aligned}
E_{Z|X,\theta}\left[\frac{\partial}{\partial\theta}\log f(Z|X,\theta)\right] &= E_{Z|X,\theta}\left[\frac{\frac{\partial}{\partial\theta}f(Z|X,\theta)}{f(Z|X,\theta)}\right] \\
&= \int\left[\frac{\frac{\partial}{\partial\theta}f(Z|X,\theta)}{f(Z|X,\theta)}\right]f(Z|X,\theta)dZ \\
&= \int\left[\frac{\partial}{\partial\theta}f(Z|X,\theta)\right]dZ \\
&= \frac{\partial}{\partial\theta}\int\left[f(Z|X,\theta)\right]dZ \\
&= 0
\end{aligned}$$

□

有了上述引理，接下来我们证明Louis 方法。

Proof. 由引理, 知

$$\frac{\partial \log f(X|\theta)}{\partial \theta} = E\left[\frac{\partial \log f(Z,X|\theta)}{\partial \theta}|X,\theta\right].$$

所以，两边求导，有

$$\begin{aligned}\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} &= \frac{\partial}{\partial \theta}E\left[\frac{\partial \log f(Z,X|\theta)}{\partial \theta}|X,\theta\right] \\ &= \frac{\partial}{\partial \theta}\int \frac{\partial \log f(Z,X|\theta)}{\partial \theta} f(Z|X,\theta)dZ \\ &= \int \frac{\partial^2 \log f(Z,X|\theta)}{\partial \theta^2} f(Z|X,\theta)dZ \\ &\quad + \int \frac{\partial \log f(Z,X|\theta)}{\partial \theta}\left[\frac{\partial f(Z|X\theta)}{\partial \theta}\right]^T dZ.\end{aligned}$$

所以

$$\int \frac{\partial \log f(Z,X|\theta)}{\partial \theta}\left[\frac{\partial f(Z|X\theta)}{\partial \theta}\right]^T = -\int \frac{\partial^2 \log f(Z,X|\theta)}{\partial \theta^2} f(Z|X,\theta)dZ - \left[-\frac{\partial^2 \log f(X|\theta)}{\partial \theta^2}\right].$$

由定理8, 上式右边等于$E_{Z|X,\theta}\left\{-\frac{\partial^2}{\partial \theta^2}\log f(Z|X,\theta)\right\}$. 所以为证明Louis 方法，只需证明

$$\int \frac{\partial \log f(Z,X|\theta)}{\partial \theta}\left[\frac{\partial f(Z|X,\theta)}{\partial \theta}\right]^T dZ = var\left[\frac{\partial \log f(Z,X|\theta)}{\partial \theta}|X,\theta\right].$$

$$
\begin{aligned}
\int \frac{\partial \log f(Z,X|\theta)}{\partial\theta}\left[\frac{\partial f(Z|X,\theta)}{\partial\theta}\right]^T dZ &= \int \frac{\partial \log f(Z,X|\theta)}{\partial\theta}\left[\frac{\partial \log f(Z|X\theta)}{\partial\theta}\right]^T f(Z|X,\theta)dZ \\
&= \int \frac{\partial \log f(Z,X|\theta)}{\partial\theta}\left[\frac{\partial[\log f(Z,X|\theta)-\log f(X|\theta)]}{\partial\theta}\right]^T f(Z|X,\theta)dZ \\
&= \int \frac{\partial \log f(Z,X|\theta)}{\partial\theta}\left[\frac{\partial \log f(Z,X|\theta)}{\partial\theta}\right]^T f(Z|X,\theta)dZ \\
&\quad -\int \frac{\partial \log f(Z,X|\theta)}{\partial\theta}\left[\frac{\partial \log f(X|\theta)}{\partial\theta}\right]^T f(Z|X,\theta)dZ \\
&= \int \frac{\partial \log f(Z,X|\theta)}{\partial\theta}\left[\frac{\partial \log f(Z,X|\theta)}{\partial\theta}\right]^T f(Z|X,\theta)dZ \\
&\quad -\int \frac{\partial \log f(Z,X|\theta)}{\partial\theta} f(Z|X,\theta)dZ \left[\frac{\partial \log f(X|\theta)}{\partial\theta}\right]^T \\
&= \int \frac{\partial \log f(Z,X|\theta)}{\partial\theta}\left[\frac{\partial \log f(Z,X|\theta)}{\partial\theta}\right]^T f(Z|X,\theta)dZ \\
&\quad -\int \frac{\partial \log f(Z,X|\theta)}{\partial\theta} f(Z|X,\theta)dZ \left[\int \frac{\partial \log f(Z,X|\theta)}{\partial\theta} f(Z|X,\theta)dZ\right]^T
\end{aligned}
$$

(上式是由引理5.3 得到的。)

$$
= var\left[\frac{\partial \log f(Z,X|\theta)}{\partial\theta}|X,\theta\right].
$$

最后一个等式是方差的定义。 □

将上述的结果整理一下，便有：

$$
I_o = I_c - I_m = -\int \frac{\partial^2 \log f(X,Z|\theta)}{\partial\theta^2} p(Z|X,\theta)dZ - var\left[\frac{\partial \log f(X,Z|\theta)}{\partial\theta}|X,\theta\right].
$$

其中I_o 是观测信息，I_c 是完全信息，I_m 是缺失信息。

例：（基因连锁模型。）完全似然是

$$
f(X,Z|\theta) = \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}.
$$

所以$\frac{\partial \log f(X,Z|\theta)}{\partial\theta} = \frac{x_2+x_5}{\theta} - \frac{x_3+x_4}{1-\theta}$, $-\frac{\partial^2 \log f(X,Z|\theta)}{\partial\theta^2} = \frac{x_2+x_5}{\theta^2} + \frac{x_3+x_4}{(1-\theta)^2}$. $I_c = E(\frac{x_2+x_5}{\theta^2} + \frac{x_3+x_4}{(1-\theta)^2}|X,\theta) = \frac{E(X_2|X,\theta)+34}{\theta^2} + \frac{18+20}{(1-\theta)^2}$, $I_m = var(\frac{x_2+x_5}{\theta} - \frac{x_3+x_4}{1-\theta}|X,\theta) = var(\frac{x_2}{\theta}|X,\theta)$.

$$
\sqrt{var(\hat\theta)} = \sqrt{1/I_o} = \sqrt{1/[\frac{E(X_2|X,\theta)+34}{\theta^2} + \frac{18+20}{(1-\theta)^2} - var(\frac{x_2}{\theta}|X,\theta)]}_{\theta=\hat\theta} = 0.05
$$

5.4 PIEM

令$Z=(Z_1,Z_2)$其中Z_1 表示将在PIEM算法的E步填补的那部分缺失数据，Z_2表示不填补的缺失数据。图3, 给出填补组2个体的X_1和填补组3的X_1和X_2 (即Z_1)，得到的单调数据模式。如果变量X_k被观测，那么填补$X_1,\ldots,X_{k-1}$的缺失值，构成单调数据模式。观测数据和填补数据的似然函数可以因子化为

$$L(\phi|Y,Z_1)=\left[\prod_{k=1}^{3}\prod_{i\in\text{group }k} f(x_{1i},x_{2i};\phi_{12})\right]\times\left[\prod_{k=2}^{3}\prod_{i\in\text{group }k} f(x_{3i}|x_{1i},x_{2i};\phi_{3|12})\right]$$

$$\times\left[\prod_{i\in\text{group }3} f(x_{4i}|x_{1i},x_{2i},x_{3i};\phi_{4|123})\right],$$

其中x_{ji}表示个体i的变量X_j的值。如果参数ϕ_{12}, $\phi_{3|12}$和$\phi_{4|123}$是相异的，那么，每个因子相应于一个完全数据的似然，这些参数的MLE可以分别对各因子求最大得到。

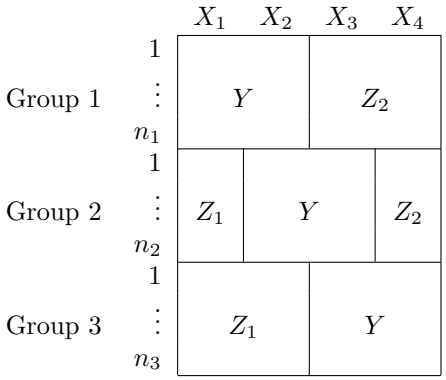

图3. 部分填补得到的单调数据模式

例：X_1 和X_2 是二值变量。假定(X_1,X_2) 服从$p_{ij}=Pr(X_1=i,X_2=j)$, 样本量是n。假设观测模式是$T=\{\{X_1,X_2\},\{X_1\},\{X_2\}\}$. 即有三组观测: $G_1=\{X_1,X_2\}$ 两个变量都观测到; $G_2=\{X_1\}$ 只有X_1 被观测到；$G_3=\{X_2\}$ 只有X_2 被观测到。观测数据显示在下表5.4 中, 样本量$n=520$. $n_{12}(ij)$'s, $n_1(i)$'s 和$n_2(j)$'s 代表观测频数，$n_{12}(ij)$'s 代表完全观测的数据, $n_1(i)$'s 代表只有X_1 被观测到。$n_2(j)$'s 代表只有X_2 被观测到。

Table 5.4: Observed Data

	$n_{12}(ij)$		
	$j=1$	$j=2$	$n_1(i)$
$i=1$	5	4	300
$i=2$	2	1	200
$n_2(j)$	5	3	

PIEM 算法：

Table 5.5: MLEs of parameters p_{ij}

	$\hat{p}(ij)$	
	$j=1$	$j=2$
$i=1$	0.3402	0.2633
$i=2$	0.2700	0.1265

- the P-E step:

$$\hat{n}_2(ij) = \frac{\hat{p}^{(t)}(ij)}{\hat{p}^{(t)}(+j)} n_2(j);$$

- the P-M step:

$$\hat{p}^{(t+1)}(ij) = \frac{n_{12}(ij) + \hat{n}_2(ij)}{n_{12}(i+) + \hat{n}_2(i+)} \cdot \frac{n_{12}(i+) + n_1(i) + \hat{n}_2(i+)}{n_{12}(++) + n_1(+) + n_2(+)}.$$

5.5 Louis' Turbo EM 算法

注意到Newton-Raphson 的迭代为

$$\theta^{n+1} = \theta^n + \left[-\frac{\partial^2 \log p(Y|\theta)}{\partial \theta^2}|_\theta^n \right] \frac{\partial \log p(Y|\theta)}{\partial \theta}|_\theta^n$$

这里的目的是利用EM 类型的量来实现Newton-Raphson 迭代，达到在众数附近的二阶收敛速度。

注意到由Louis' 方法，

$$-\frac{\partial^2 \log p(Y|\theta)}{\partial \theta^2} = -\int \frac{\partial^2 \log p(Y,Z|\theta)}{\partial \theta^2} p(Z|Y,\theta) dZ - var\left\{ \frac{\partial \log p(Y,Z|\theta)}{\partial \theta} | Y, \theta \right\}$$

$$\begin{aligned} \frac{\partial \log p(Y|\theta)}{\partial \theta} &= \int \frac{\partial \log p(Y,Z|\theta)}{\partial \theta} p(Z|Y,\theta) dZ \\ &= \frac{\partial \int \log p(Y,Z|\theta) p(Z|Y,\theta) dZ}{\partial \theta} \end{aligned}$$

注意到

$$\frac{\partial \int \log p(Y,Z|\theta) p(Z|Y,\theta^{(m)}) dZ}{\partial \theta}|_{\theta=\theta_{EM}} = 0$$

所以，可以将

$$\frac{\partial \int \log p(Y,Z|\theta) p(Z|Y,\theta^{(m)}) dZ}{\partial \theta}|_{\theta=\theta_{EM}}$$

在θ^m 处Taylor 展开，有

$$0 = \frac{\partial \int \log p(Y, Z|\theta) p(Z|Y, \theta^{(m)}) dZ}{\partial \theta}|_{\theta=\theta^m} + \left[\frac{\partial^2 \log p(Y, Z|\theta)}{\partial \theta^2} p(Z|Y, \theta^{(m)})\right](\theta_{EM} - \theta^m)$$

即：

$$\frac{\partial \int \log p(Y, Z|\theta) p(Z|Y, \theta^{(m)}) dZ}{\partial \theta}|_{\theta=\theta^m} = -\left[\frac{\partial^2 \log p(Y, Z|\theta)}{\partial \theta^2} p(Z|Y, \theta^{(m)})\right](\theta_{EM} - \theta^{(m)})$$

Chapter 6

数据扩充算法

6.1 Bayes 方法

Bayes 方法认为所有的变量都是随机变量。人们根据观测到的数据，不断地对一些参数重新认识。设 观测数据的似然函数为$L(\theta|Y)$, 再设参数θ 的先验为$p(\theta)$. 则观测到数据 X 后，θ 的后验密度为：

$$p(\theta|X) = cp(\theta)L(\theta|X),$$

其中 c 为归一化因子：

$$c = 1/\int_{\Theta} p(\theta)L(\theta|X).$$

例：线性模型

$$y = X\beta + \epsilon,$$

其中，$y \in R^n$, $\beta \in R^p$, $X \in R^{n\times p}$是观测矩阵。$\epsilon \sim N(0, \sigma^2 I_n)$, σ 是未知参数。

$$f(y|\beta, \sigma^2) = (\frac{1}{\sqrt{2\pi}\sigma})^n \exp\{-\frac{\|y - X\beta\|_2^2}{2\sigma^2}\}$$

记

$$\hat{\beta} = (X^T X)^{-1} X^T y, v = n - p, \hat{y} = X\hat{\beta}$$

$$s^2 = (y - \hat{y})^T (y - \hat{y})/v,$$

则

$$\begin{aligned}(y - X\beta)^T (y - X\beta) &= [y - X\hat{\beta} + X\hat{\beta} - X\beta]^T [y - X\hat{\beta} + X\hat{\beta} - X\beta] \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) + (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}).\end{aligned}$$

所以，

$$f(y|\beta,\sigma^2)=(\frac{1}{\sqrt{2\pi}\sigma})^n\exp\{-\frac{1}{2\sigma^2}[vs^2+(\beta-\hat{\beta})^TX^TX(\beta-\hat{\beta})]\}$$

在先验$p(\sigma^2,\beta)\propto\sigma^{-2}$的情况下，有后验分布

$$p(\beta,\sigma^2|Y)=p(\sigma^2|s^2)p(\beta|\hat{\beta},\sigma^2),$$

其中$p(\sigma^2|s^2)$ 是vs^2/χ_v^2 的密度函数, $p(\theta|\hat{\theta},\sigma^2)$是$N(\hat{\theta},(X^TX)^{-1}\sigma^2)$ 的密度函数

证明：后验

$$p(\beta,\sigma^2|Y)\propto L(\theta,\sigma^2|Y)p(\beta,\sigma^2)\propto\left[\sigma^{-p/2}\exp(\frac{(\beta-\hat{\beta})^TX^TX(\beta-\hat{\beta})}{2\sigma^2})\right]\times\left[(\sigma^2)^{-(n-p)/2}\exp(\frac{-vs^2}{2\sigma^2})\times(\sigma^2)^{-1}\right].$$

第一个方括号与多元正态分布$N(\hat{\beta},\sigma^2(X^TX)^{-1})$ 有相同的形式；第二个方括号为σ^2 的函数，与vs^2/Z的密度具有相同的形式，其中Z服从分布χ_v^2, 贝叶斯方法在给定数据下，vs^2被看成常数。

注：χ_v^2分布密度为：$f(z)\propto z^{\frac{v}{2}-1}e^{-\frac{z}{2}}$, 所以$T=c/Z$ 的分布密度为：

$$f(t)\propto(c/t)^{\frac{v}{2}-1}e^{-\frac{c/t}{2}}|-c/t^2|\propto t^{-\frac{v}{2}-1}e^{-\frac{c}{2t}}$$

关于上述结果的一个应用。如果我们希望从后验密度$p(\beta,\sigma^2|Y)$ 中抽取样本，可以使用如下步骤：

1. 从χ_v^2 中抽取一个z^*, 并计算$\sigma_*^2=vs^2/z^*$

2. 从正态分布$N(\hat{\beta},\sigma^2(X^TX)^{-1})$ 抽取β^*.

那么，(θ^*,σ_*^2) 就是一个来自后验密度$p(\beta,\sigma^2|Y)$ 的样本。

6.2 无信息先验

考虑$y_1,y_2,\ldots,y_n\sim N(\mu,\sigma^2)$, 其中$\sigma^2$ 已知。似然可以写成

$$L(\theta|Y)=\exp\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\}$$

可以看到，数据告诉θ 的位置信息，不同的数据给出不同的位置，所以，在没有任何信息的时候，一个自然的先验就是θ 可以取任何值，即$p(\theta)=constant$.

一般的，如果能改写似然函数$L(\theta|Y)$ 为$g(\psi(\theta)-t(y))$ 的形式，那么这个似然函数是一个数

据变换的形式，不同的数据机和仅通过$t(y)$ 平移这个似然函数。所以在没有任何其他信息的时候，一个自然的先验就是

$$p(\psi(\theta)) \propto constant,$$

变换回θ 上，

$$p(\theta) \propto |\frac{\partial \psi(\theta)}{\partial \theta}|$$

例：正态总体，均值已知，方差未知。

$$L(\sigma^2|Y) = (\sigma^2)^{-n/2} \exp(\frac{-ns^2}{2\sigma^2}),$$

其中

$$s^2 = \sum (y_i - \theta)^2/n.$$

注意到

$$\begin{aligned} L(\sigma^2|Y) &\propto (\sigma^2/s^2)^{-n/2} \exp(\frac{-ns^2}{2\sigma^2}) \\ = \exp\left[\frac{-n}{2}(\log(\sigma^2) - \log(s^2)) - \frac{n}{2}\exp(-[\log(\sigma^2) - \log(s^2)])\right] \\ &= g(\log(\sigma^2) - \log(s^2)), \end{aligned}$$

其中$g(t) = \exp\left[\frac{-n}{2}t - \frac{n}{2}\exp(-t)\right]$. 根据前面的描述，$\log(\sigma^2)$ 可以选取任意值：即$p(\log(\sigma^2)) \propto constant$, 推出$p(\sigma^2) \propto |\frac{\partial \log(\sigma^2)}{\partial(\sigma^2)}| = \frac{1}{\sigma^2}$

Jeffereys prior:

$$p(\theta) \propto \sqrt{|I(\theta)|},$$

where

$$I(\theta) = E_X\left[-\frac{\partial^2 \log L(\theta|X)}{\partial \theta^2}\right]$$

例：

$$f(x|\sigma) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}},$$

$$I(\sigma) = E[(\frac{d}{d\sigma}\log f(x|\sigma))^2] = E\left(\frac{(x-\mu)^2 - \sigma^2}{\sigma^3}\right)^2 = \frac{3\sigma^4 - 2\sigma^4 + \sigma^4}{\sigma^6} = \frac{2}{\sigma^2}$$

所以，$p(\sigma) = \frac{1}{\sigma}$, 从而

$$p(\sigma^2) = \frac{1}{\sigma^2}.$$

计算Poison 分布、二项分布的Jeffereys prior. (作业)

Jefferys Prior 的直观意义：目标：求θ 的prior。先考虑θ 的变换$\tau = \tau(\theta)$. 易知：

$$\frac{\partial l(\tau|Y)}{\partial \tau} = \frac{\partial l(\theta|Y)}{\partial \theta}\frac{\partial \theta}{\partial \tau},$$

所以

$$-\frac{1}{n}[\frac{\partial l(\tau|Y)}{\partial \tau}]^2 = -\frac{1}{n}[\frac{\partial l(\theta|Y)}{\partial \theta}]^2[\frac{\partial \theta}{\partial \tau}]^2$$

$I(\tau) = I(\theta)[\frac{\partial \theta}{\partial \tau}]^2$, 如果取$\tau(\theta)$ 使得$\frac{\partial \theta}{\partial \tau} \propto I^{-1/2}(\theta)$, 则$I(\tau) = c$.

现在看，τ 的对数似然函数：

$$l(\tau|X) \approx l(\hat{\tau}|X) + \frac{1}{2}\frac{\partial^2 l(\theta|X)}{\partial \theta^2}(\tau - \hat{\tau})^2 \approx l(\hat{\tau}|X) + \frac{1}{2}c(\tau - \hat{\tau})^2$$

这时，由无信息先验的描述，$p(\tau) \propto constant$, 转到θ,有

$$p(\theta) \propto |\frac{d\tau}{d\theta}| \propto I^{1/2}(\theta).$$

6.2.1 共轭先验

如果似然函数是指数分布族，密度为

$$g(\theta)h(y)\exp\{\sum_j^m \phi_j(\theta)t_i(y)\},$$

样本量为 n, 那么先验密度的形式如下时

$$p(\theta) \propto [g(\theta)]^b \exp\{\sum_{j=1}^m \phi_j(\theta)a_j\},$$

后验密度与$p(\theta)$具有相同的形式，只是 $b, a_1, \ldots, a_m$ 替换为

$$b' = b + n, a_j' = a_j + \sum_{i=1}^n t_j(y_i), j = 1, \ldots, m.$$

例：Beta 分布是二项分布的共轭先验。二项分布的似然为

$$L(\theta|y) = \theta^y(1-\theta)^{n-y},$$

如果选取priot

$$p(\theta) \propto \theta^{a-1}(1-\theta)^{b-1},$$

则后验为

$$p(\theta|y) = \theta^{a+y-1}(1-\theta)^{n-y+b-1} = Beta(a+y, n-y+1).$$

同样的方法，可以知道

1. Gamma 可以作为Poisson 的先验。

2. Normal 可以作为Normal(σ已知) 的先验。

3. Inverse Gamma 可以作为Normal (μ 已知，σ 未知)的先验。

注：Inverse Gamma = 1/Gamma。

作业。

6.3 数据扩充算法

本章的目标是得到整个后验密度，而不仅仅是后验众数。

数据扩充算法的基本思想是，假设在给定Y 和Z 下，可以很容易的计算扩充数据的后验分布$p(\theta|Y,Z)$, 并从其抽样。为了得到观测后验$p(\theta|Y)$, 需要将Z积分掉，但是将Z积分掉，依赖于$p(\theta|Y)$, 所以这是一个计算$p(\theta|Y)$ 的迭代算法。

具体来讲，

$$p(\theta|Y) = \int p(\theta|Y,Z)p(Z|Y)dZ,$$

而

$$p(Z|Y) = \int p(Z|\theta,Y)p(\theta|Y)d\theta.$$

由前面的两个公式，我们可以得到计算后验分布的迭代方案：

a. 填补步：

a1. 给定初始分布$p(\theta|Y)$, 产生样本θ;

a2. 产生$z \sim p(Z|\theta,Y)$

重复上述步骤，可以得到$z_1, z_2, \ldots, z_m$,他们来自$p(Z|Y)$.

b. 后验更新步：

$$p(\theta|Y) = \frac{1}{m}\sum_{j=1}^{m} p(\theta|z_j, Y).$$

例：基因连锁。扩充后的后验分布：

$$p(\theta|Y,Z) = \theta^{x_2+x_5}(1-\theta)^{x_3+x_4}.$$

$$p(X_2|\theta,Y) = Binomial(125, \frac{\theta}{\theta+2}).$$

因此计算后验$p(\theta|Y)$的步骤为：

a. 填补步:

a1. 从目前观测后验的估计抽取θ

a2. 从$Binomial(125, \frac{\theta}{\theta+2})$ 抽取x_2

重复以上步骤m 次得到$x_2^i, i = 1, 2, \ldots, m.$

b. 后验更新步：

$$p(\theta|Y) = \frac{1}{m}\sum Beta(v_1^{(i)}, v_2^{(i)})(\theta),$$

其中，

$$v_1^{(i)} = x_2^{(i)} + x_5 + 1, v_2^{(i)} = x_3 + x_4 + 1.$$

重复以上步骤，直至收敛即可得到$p(\theta|Y)$ 的后验。

作业：使用R 或者 matlab 运用上述迭代方法计算 $p(\theta|Y)$, 然后在图上画出$\hat{p}(\theta|Y)$ 和 $p(\theta|Y)$.

例：二元正态协方差矩阵。假定表?? 中的数据为二元正态分布的观测，均值$\mu_1 = \mu_2 = 0$, 相关系数ρ 和方差σ_1^2, σ_2^2 未知。注意4对完全观测，其中两对相关为1，两对相关为-1，可以想象ρ 的后验分布不是单峰。

Table 6.1: 二元正态不完全数据

1	1	-1	-1	2	2	-2	-2	?	?	?	?
1	-1	1	-1	?	?	?	?	2	2	-2	-2

假设Σ 的先验分布为

$$p(\Sigma) \propto |\Sigma|^{-3/2},$$

则Σ的后验分布是逆Wishart 分布(Homework)。

因此，数据扩充算法为：

a. 填补步：

a1. 从目前观测后验的估计抽取Σ

a2. 产生缺失数据

$$x_2 \sim N(\rho\frac{\sigma_1}{\sigma_2}x_1, \sigma_2^2(1-\rho^2)),$$

$$x_1 \sim N(\rho\frac{\sigma_2}{\sigma_1}x_2, \sigma_1^2(1-\rho^2)).$$

b. 后验更新步：Σ 的后验密度等于逆Wishart 分布的混合分布。

作业：画出ρ 的直方图。

补充：什么是逆Wishart 分布？怎样从逆Wishart 分布抽样？

Wishart 分布：$W \sim Wishart_v(S)$, with $dim(W) = dim(S) = k \times k$.

$$p(W) \propto |S|^{-v/2}|W|^{(v-k-1)/2} \exp\left(-\frac{1}{2}tr(S^{-1}W)\right),$$

$v > 0$, S and W 对称正定矩阵。

$E(W) = vS$ 注：如果$X_1, X_2, \ldots, X_v$ i.i.d. 服从$N_k(0, S)$, 则

$$\theta = \sum X_i X_i^T \sim W_v(S)$$

对于一维情形，Wishart 分布是χ^2 分布。

逆Wishart 分布：$W \sim Inverse - Wishart_v(S^{-1})$, with $dim(W) = dim(S) = k \times k$.

$$p(W) \propto |S|^{v/2}|W|^{-(v+k+1)/2} \exp\left(-\frac{1}{2}tr(SW^{-1})\right),$$

$v > 0$, S and W 对称正定矩阵。

$$E(W) = (v - k - 1)^{-1}S.$$

Note: 如果$W \sim Wishart_v(S)$,则$W^{-1} \sim Inverse - Wishart_v(S^{-1})$

6.4 穷人的数据扩充算法

6.4.1 PMDA1

在数据扩充算法的填补步，需要抽取$Z \sim p(Z|Y)$. 完整的做法是这样的：

a1. 给定初始分布$p(\theta|Y)$, 产生样本θ;

a2. 产生$z \sim p(Z|\theta, Y)$

一种近似的方法是，

$$p(Z|Y) \approx p(Z|Y, \hat{\theta}),$$

其中$\hat{\theta} = \arg\max p(\theta|Y)$. 事实上，

$$p(Z|Y) = p(Z|Y, \hat{\theta})\left\{1 + O\left(\frac{1}{n}\right)\right\},$$

证明：首先我们证明一个引理。

引理4 (Laplace 方法(Tierney and Kadane, 1986))**.** 假设$-h(\theta)$ 一元, 光滑，有界，单峰的函

数，其最大值点是$\hat{\theta}$. 则

$$I = \int f(\theta)\exp[-nh(\theta)]d\theta \approx f(\hat{\theta})\sqrt{\frac{2\pi}{n}}\sigma\exp[-nh(\hat{\theta})] = \hat{I},$$

其中，

$$\sigma = \left[\frac{\partial^2 h}{\partial\theta^2}|_{\hat{\theta}}\right]^{-1/2}.$$

Laplace 方法的（不严格）证明：将$h(\theta)$ 在$\hat{\theta}$ 处展开，有：

$$\begin{aligned} I &\approx \int f(\theta)\exp\left[-n[h(\hat{\theta}) + h'(\hat{\theta})(\theta-\hat{\theta}) + \frac{(\theta-\hat{\theta})^2}{2}h''(\hat{\theta})]\right]d\theta \\ &\approx \exp[-n[h(\hat{\theta})]]\int f(\hat{\theta})\exp\left[\frac{-(\theta-\hat{\theta})^2}{2}nh''(\hat{\theta})]\right]d\theta \\ &= \exp[-n[h(\hat{\theta})]]\sqrt{2\pi[1/nh''(\hat{\theta})]}f(\hat{\theta})\int\frac{1}{\sqrt{2\pi[1/nh''(\hat{\theta})]}}\exp\left[\frac{-(\theta-\hat{\theta})^2}{2}nh''(\hat{\theta})]\right]d\theta \end{aligned}$$

特别的，

$$I = \hat{I}[1 + O(1/n)].$$

利用Laplace 方法，可以得到

$$E(g(\theta)|Y) = \frac{\int g(\theta)\exp\{-nh(\theta)\}d\theta}{\int\exp\{-nh(\theta)\}d\theta} \approx \frac{g(\hat{\theta}) * C}{1 * C}.$$

其中$\exp\{-nh(\theta)\} = L(\theta|Y)p(\theta)$, $C = \sqrt{\frac{2\pi}{n}}\sigma\exp[-nh(\hat{\theta})]$, $\hat{\theta}$ 为后验众数。

还有一种(二阶)近似方法：

$$\begin{aligned} E(g(\theta)|Y) &= \frac{\int g(\theta)\exp\{-nh(\theta)\}d\theta}{\int\exp\{-nh(\theta)\}d\theta} \\ &= \frac{\int\exp\{-nh(\theta) + \log(g(\theta))\}d\theta}{\int\exp\{-nh(\theta)\}d\theta} \\ &\approx \frac{\sigma^*\exp\{-nh^*(\hat{\theta}^*)\}}{\sigma\exp\{-nh(\hat{\theta})\}}. \\ &= \hat{I}_2 \end{aligned}$$

其中，$-nh^*(\theta) = -nh(\theta) + \log(\theta)$, $\theta^* = \arg\max -nh^*(\theta)$, $\sigma^* = [(h^*)''(\theta^*)]^{-1/2}$.

Tierney and Kadane(1986), Mosteller and Wallace (1964) 证明了

$$I = \hat{I}_2(1 + O(1/n^2)).$$

由以上近似，可以得到穷人的数据扩充算法：

a. 填补步：

抽取$z_1, \ldots, z_m \sim p(Z|Y, \hat{\theta})$

b. 后验更新步：

$$p(\theta|Y) = \frac{1}{m}\sum_{j=1}^{m} p(\theta|z_j, Y).$$

6.4.2 PMDA -精确方法

PMDA1 是一个近似方法，如果$p(Z|Y)$ 是很容易计算的（但是抽样不容易），可以使用重要性抽样的方法计算后验分布

$$p(\theta|Y) = \int p(\theta|Z, Y)p(Z|Y)dz = \int p(\theta|Z, Y)\frac{p(Z|Y)}{p(Z|Y, \hat{\theta})}p(Z|Y, \hat{\theta})dz$$

a. 填补步：

a1. 抽取$z_1, \ldots, z_m \sim p(Z|Y, \hat{\theta})$

a2. 计算权重：

$$w_j = \frac{p(z_j|Y)}{p(z_j|Y, \hat{\theta})}$$

b. 后验更新步：

$$p(\theta|Y) = \sum_{j=1}^{m} w_j p(\theta|z_j, Y) \Big/ \sum_{j=1}^{m} w_j \ .$$

注: $P(Z|Y) = \frac{P(Z,Y)}{P(Y)}$.

$$p(Z, Y) = \frac{p(Z, Y|\theta)p(\theta)}{p(\theta|Z, Y)}.$$

6.4.3 PMDA2

当$p(Z|Y)$ 不易计算的时候，我们可以使用二阶近似：

$$p(Z|Y) = \int p(Z|Y, \theta)p(\theta|Y)d\theta = \int p(Z|Y, \theta)\exp\{\log(p(\theta|Y))\}d\theta.$$

由二阶近似方法，有

$$= \left(\frac{|\Sigma^*|}{|\Sigma|}\right)^{1/2} \frac{\exp\{-nh^*(\theta^*)\}}{\exp\{-nh(\hat{\theta})\}}[1 + O(1/n^2)]$$

注意到:

$$\exp\{-nh(\theta)\} = p(\theta|Y) = \frac{p(\theta|Y, Z)p(Z|Y)}{p(Z|Y, \theta)}$$

$$\exp\{-nh^*(\theta)\} = p(\theta|Y)p(Z|Y,\theta) = p(\theta|Y,Z)p(Z|Y)$$

所以，$\theta^* = \arg\max p(\theta|Y,Z)$. 从而，

$$p(Z|Y) \propto |\Sigma^*|^{1/2} \frac{p(\theta^*|Y,Z)p(Z|Y)}{\frac{p(\hat{\theta}|Y,Z)p(Z|Y)}{p(Z|Y,\hat{\theta})}} = |\Sigma^*|^{1/2} \frac{p(\theta^*|Y,Z)p(Z|Y,\hat{\theta})}{p(\hat{\theta}|Y,Z)}$$

PDMA2:

a. 填补步：

a1. 抽取$z_1, \ldots, z_m \sim p(Z|\hat{\theta}, Y)$

a2. 计算权重：

$$\begin{aligned} w_j &= \frac{p(z_j|Y)}{p(z_j|Y,\hat{\theta})} \\ &\approx |\Sigma^*|^{1/2} \frac{p(\theta^*|Y,Z_j)}{p(\hat{\theta}|Y,Z_j)} \end{aligned}$$

b. 后验更新步：

$$p(\theta|Y) = \sum_{j=1}^{m} w_j p(\theta|z_j, Y) \Big/ \sum_{j=1}^{m} w_j \ .$$

6.5 一般的填补方法

本节假定数据满足以下模型：

$$Y_i = X_i^T \theta + \epsilon_i,$$

其中$\epsilon_i's$ i.i.d. 均值为0，方差σ^2, θ 是d 维参数向量，X_i 是已知常数的协变量的向量。

本节介绍在给定观测数据：

$$\begin{pmatrix} Y_1 \\ X_1 \end{pmatrix}, \cdots \begin{pmatrix} Y_{n_1} \\ X_{n_1} \end{pmatrix}, \begin{pmatrix} ? \\ X_(1) \end{pmatrix}, \ldots, \begin{pmatrix} ? \\ X_{(n_0)} \end{pmatrix},$$

产生n_0 个缺失响应值$Y_(1), \ldots, Y_{(n_0)}$ 的三种技术。

(1) Hot Deck 填补这种方法是用于X 是离散值的情形. 假设X 有b 种不同的取值，并且对应每一个取值，有几个响应值。将数据分为b 类：

$$\begin{pmatrix} Y_{11} \\ X_1 \end{pmatrix}, \cdots \begin{pmatrix} Y_{1,n_{11}} \\ X_1 \end{pmatrix}, \begin{pmatrix} ? \\ X_1 \end{pmatrix}_1, \ldots, \begin{pmatrix} ? \\ X_1 \end{pmatrix}_{n_{01}};$$

$$\vdots$$

$$\begin{pmatrix} Y_{b1} \\ X_b \end{pmatrix}, \cdots \begin{pmatrix} Y_{b,n_{1b}} \\ X_b \end{pmatrix}, \begin{pmatrix} ? \\ X_b \end{pmatrix}_1, \ldots, \begin{pmatrix} ? \\ X_b \end{pmatrix}_{n_{0b}},$$

其中$n_1 = \sum_{j=1}^{b} n_{1j}, n_0 = \sum_{j=1}^{b} n_{0j}$.

Hot Deck 方法：在每一类中，有放回的独立抽取观测到的响应，形成缺失响应的填补。

填补完全所有的b 类，即可得到一个完整的数据集。由此完整的数据集，可以计算$\hat{\theta}$.

重复以上过程m 次，得到多重填补(即填补多次)，并得到m 个θ 的估计：$\hat{\theta}^{(j)}$. Hot Deck 估计：

$$\hat{\theta}_{HD} = \frac{1}{m} \sum_{j} \hat{\theta}^{(j)}$$

Schenker and Welsh (1988) 给出了正则条件，在此条件下，

$$\hat{\theta}_{HD} - \theta \sim N(0, V).$$

(2) 简单残差填补（SRI）

此种填补和Hot Deck 类似，适用于X是连续值的情形。此时，不易将X分为b 类。

令$r_i = Y_i - X_i^T \hat{\theta}, i = 1, \ldots, n_1$, 其中$\hat{\theta}$ 是基于n_1 个完全数据(X, Y) 得到的最小二乘估计。SRI 从$r_1, \ldots, r_{n_1}$ 中，有放回的随机抽取n_0 个残差。然后，令

$$Y_{(i)} = X_{(i)}^T \hat{\theta} + r_{(i)}, i = 1, \ldots, n_0.$$

重复上述过程m 次，便得到m 个完全数据集，利用每一个数据集，可以得到m 个最小二乘估计$\hat{\theta}^{(j)}, j = 1, \ldots, m$,

$$\hat{\theta}_{SRI} = \frac{1}{m} \sum_{j} \hat{\theta}^{(j)}.$$

Schenker and Welsh (1988) 给出了正则条件，在此条件下，

$$\hat{\theta}_{SRI} - \theta \sim N(0, V).$$

(3) 正态填补

回顾本章开始时正态模型的结论：

在先验$p(\sigma^2,\beta)\propto\sigma^{-2}$的情况下，有后验分布

$$p(\beta,\sigma^2|Y)=p(\sigma^2|s^2)p(\beta|\hat{\beta},\sigma^2),$$

其中$p(\sigma^2|s^2)$ 是vs^2/χ^2_v 的密度函数, $p(\theta|\hat{\theta},\sigma^2)$是$N(\hat{\theta},(X^TX)^{-1}\sigma^2)$ 的密度函数。

其中，

$$s^2=\|y-\hat{\beta}\|^2/n.$$

由以上结论，我们可以得到，正态填补:

1. 从$(n_1-d)s_1^2/\chi^2_{n_1-d}$ 分布抽取σ_*^2

2. 从$N(\hat{\theta},\sigma_*^2(\sum_{i=1}^{n_1}X_i^TX_i)^{-1})$ 抽取θ^*;

3. 填补：

$$Y_{(i)}=X_{(i)}^T\theta^*+\sigma_*e_{(i)},i=1,\ldots,n_0$$

其中$e_{(i)}\sim N(0,1)$.

4. 计算$\hat{\theta}$.

重复以上过程得到m 个$\hat{\theta}$，正态填补估计为

$$\hat{\theta}_{Normal}=\frac{1}{m}\sum_{j=1}^{m}\hat{\theta}^{(j)}.$$

Schenker and Welsh (1988) 给出了正则条件，在此条件下，

$$\hat{\theta}_{Normal}-\theta\sim N(0,V).$$

6.6 不可忽略不响应

不可忽略不响应：不响应和响应变量Y 的取值是有关系的（换句话这种缺失不是随机的）。Rubin(1987) 考虑使用两种模型来处理这类问题：Mixture model and selection model.

6.6.1 Mixture Model – Without Followup Data

首先看一个正态混合模型。对应于响应个体，Y 服从均值为$X_i^T\beta_1+\alpha_1$ 和方差σ_1^2 的分布；对于不响应个体，Y 服从均值为$X_i^T\beta_0+\alpha_0$ 方差σ_0^2 的正态分布。为简单起见，假设$\sigma_1=\sigma_0=\sigma$.

注意到在无信息先验下，$(\alpha_1,\beta_1,\sigma|Y)$ 可以分解为逆卡方分布和正态分布。缺失数据的预测分布$p(Z_i|\alpha_0,\beta_0,\alpha_1,\beta_1,\sigma,Y)$ 服从$N(\alpha_0+X_i^T\beta_0,\sigma^2)$. 现在的问题是，怎样连接$(\alpha_0,\beta_0)$和$(\alpha_1,\beta_1)$?

给定$(\alpha_1, \beta_1, \sigma)$ 条件下，(α_0, β_0) 的条件分布是两个独立分布的乘积。$\beta_0|\alpha_1, \beta_1, \sigma \sim N(\beta_1, C_\beta^2 \beta_1 \beta_1^T)$. α_0 的条件边缘分布采用在$\bar{X}_1$ 处的Y 的均值来说明，$\eta_0 = \alpha_0 + \bar{X}_1^T \beta_0$:

$$\eta_0|\alpha_1, \beta_1, \sigma \sim N(\eta_1 = \alpha_1 + \bar{X}_1^T \beta_1, C_\eta^2 \eta_1^2).$$

注：η_1 是响应组的Y的均值。

注意到,如果$C_\eta = C_\beta = 0$, 以上描述的模型是一个可忽略模型。系数C_β 说明响应个体和不响应个体斜率的类似程度。Rubin(1987) 称C_β 为不响应个体和响应个体的回归系数的“变异性的先验系数”。令β_{1j} 和β_{0j} 表示β_1 和β_0 的第j 元素。模型假设β_{0j} 有95% 的把握落在区间$\beta_{1j}(1 \pm 1.96 C_\beta)$, 并且$\beta_0$ 的分布是以β_1 为中心。C_η 具有类似的解释。模型假设η_0 有95% 的把握落在区间$\eta_1(1 \pm 1.96 C_\eta)$, 并且$\eta_0$ 的分布是以η_1 为中心。

在以上的模型假定下，$Z, \eta_0, \beta_0, \eta_1, \beta_1, \sigma$ 的联合后验分布可以因子化为：

$$p(Z, \eta_0, \beta_0, \eta_1, \beta_1, \sigma|Y) = p(Z|\eta_0, \beta_0, \sigma, Y) p(\eta_0, \beta_0|\eta_1, \beta_1, \sigma, Y) p(\eta_1, \beta_1, \sigma|Y),$$

有了这样的分解，我们就可以填补缺失数据了：

第一步：抽取$\eta_1^*, \beta_1^*, \sigma_*^2$

1. 抽取σ_*^2 from inverse χ^2 分布：

$$(n_1 - d) s_1^2 / \chi_{n_1 - d}^2,$$

$$s_1^2 = \|Y_1 - \hat{Y}_1\|^2 / n_1$$

2. 抽取β from

$$N(\hat{\beta}, (\bar{X}_1^T \bar{X}_1)^{-1} \sigma_*^2),$$

where $\hat{\beta}$ 是β_1 的OLS 估计，X_1 是去掉截距的设计阵。

3. 抽取η_1 from

$$N(\bar{Y}_1, \sigma_*^2 / n_1).$$

第二步：抽取β_0^*, η_0^*

1. 从$N(\beta_1^*, C_\beta^2 \beta_1^* \beta_1^{*T})$ 抽取β_0^*

2. 从$N(\eta_1^*, C_\eta^2 \eta_1^2)$ 抽取η_0^*

第三步：抽取Z from

$$N(\eta_0^* - \bar{X}_1^T \beta_0^* + X_i^T \beta_0^*, \sigma_*^2)$$

重复以上三个步骤m 次，可以得到m 个填补的完全数据。

Brown (1990) 指出对于某些模型，该方法不能给出相合的点估计。

6.6.2 Mixture Model – With Followup Data

所谓追踪数据（Followup data）是指那些缺失响应的个体，通过追踪，又完成了一些观测。特别的，对于跟踪数据，假定

$$Y = \alpha_0 + X_f^T \beta_0 + \sigma\epsilon.$$

有了这些跟踪数据之后，就可以使用这些数据对未响应的个体进行填补。

$$p(Z, \beta_0, \alpha_0, \sigma|Y) = p(Z|\beta_0, \alpha_0, \sigma|Y)p(\beta_0, \alpha_0, \sigma|Y)$$

1. 抽取$\alpha_0^*, \beta_0^*, \sigma_*^2$
2. 抽取$Z \sim N(\alpha_0^* + X_{nf}^T \beta_0^*, \sigma_*^2)$.

也可以使用Hot Deck 方法填补未响应的缺失值。

6.6.3 Selection Model – Without Followup Data

Greenless, Reece, and Zieschang (1982) 引入了选择模型处理不可忽略不响应。他们依然假设正态模型：

$$Y_i = X_i^T \beta + \sigma\epsilon_i,$$

其中$\epsilon_i \sim N(0, 1), i.i.d.$

在选择模型中，假定回答的概率依赖于Y_i 和其他变量V_i:

$$P(R_i = 1|Y_i, V_i) = \frac{1}{1 + \exp(-\alpha - \gamma Y_i - V_i^T \delta)},$$

其中 $R_i = 1$ 表示个体 i 回答，$R_i = 0$ 表示丢失；V_i 为个体的一个p 维向量；α 和γ 是标量，δ 是一个p 维参数向量。

回答个体对似然函数的贡献为

$$L_i = \left\{ \frac{1}{1 + \exp(-\alpha - \gamma Y_i - V_i^T \delta)} \right\} \frac{1}{\sigma} \phi(\frac{Y_i - X_i^T \beta}{\sigma}),$$

其中$\phi(x)$ 为标准正态密度。不回答个体对似然的贡献为：

$$L_i = \int_{-\infty}^{\infty} \left\{ 1 - \frac{1}{1 + \exp(-\alpha - \gamma Y_i - V_i^T \delta)} \right\} \frac{1}{\sigma} \phi(\frac{Y_i - X_i^T \beta}{\sigma}) dY$$

整个样本的似然为所有L_i 的乘积。Greenless, Reece, and Zieschang (1982) 建议使用Newton 算法最大化这个似然，求$\alpha, \gamma, \beta, \delta$ 和σ.

下面的拒绝/接受算法（舍选法）可以得到多重填补：

1. 从$N(0,1)$ 抽取e_i

2. 计算

$$Y_i = X_i^T \hat{\beta} + \hat{\sigma} e_i$$

$$p(R = 0|Y_i, V_i) = 1 - 1/[1 + \exp(-\hat{\alpha} - \hat{\gamma} Y_i - V_i^T \hat{\delta})]$$

3. 从均匀分布$U(0,1)$ 抽取u.

4. If $p(R = 0|Y_i, V_i) \geq u$，则接受Y_i, 否则拒绝并返回第1步。

6.6.4 Selection Model – With Followup Data

类似无追踪数据方法：如果个体i 是回答这或者为追踪的不回答者，那么它们对似然的贡献跟前面一样。追踪的不响应对似然的贡献为：

$$L_i = \left\{1 - \frac{1}{1 + \exp(-\alpha - \gamma Y_i - V_i^T \delta)}\right\} \frac{1}{\sigma} \phi(\frac{Y_i - X_i^T \beta}{\sigma}),$$

6.7 进一步的重要性抽样方法

6.7.1 采用重要性抽样

注意到在数据扩充算法的填补步，需要从下面的混合分布抽取一个θ:

$$\frac{1}{m} \sum_{j=1}^{m} p(\theta|z_j, Y).$$

但是有时候，也许不能直接从$p(\theta|z, Y)$ 中抽样，这时候我们需要一种方法绕过这个问题。

以前的做法是，产生$z_1, z_2, \ldots, z_m$, 然后更新

$$p(\theta|Y) = \int p(\theta|Y, Z) p(Z|Y) dZ$$

在产生z_i 的时候，是利用如下等式产生的：

$$p(Z|Y) = \int p(Z|\theta, Y) p(\theta|Y) d\theta.$$

所以，产生z_i, 分为两步：

1. 产生$\theta \sim p(\theta|Y)$

2. 产生$z \sim p(Z|\theta, Y)$.

如过第一步不易产生，可以利用如下等式产生带有权重的样本：

$$p(Z|Y) = \int p(Z|\theta, Y)\frac{p(\theta|Y)}{I(\theta)}I(\theta)d\theta.$$

1. 从$I(\theta)$ 产生ϕ

2. 从$p(z|\phi, Y)$ 产生Z

3. 计算权重$w = \frac{p(\theta=\phi|Y)}{I(\phi)}$.

以后再计算$E_{\sim Z|Y}g(Z)$ 可以使用上面带有权重的样本Z,即

$$E[g(Z)|Y] \approx \frac{1}{m}\sum g(z_j)w_j.$$

这是因为

$$\begin{aligned}\frac{1}{m}\sum g(z_j)w_j \to E(g(\tilde{Z})W(\phi)) &= \int g(z)\frac{p(\phi|Y)}{I(\phi)}I(\phi)p(Z|\phi, Y)dZd\phi \\ &= \int g(z)p(\phi|Y)p(Z|\phi, Y)dZd\phi = E(g(Z)|Y)\end{aligned}$$

综上所述，我们得到重要性抽样的数据扩充算法：

a. 填补步

a1. 从$I(\theta)$ 产生ϕ

a2. 从$p(z|\phi, Y)$ 产生Z

a3. 计算权重$w = \frac{p(\theta=\phi|Y)}{I(\phi)}$.

b. 后验更新步

$$p(\theta|Y) = E[p(\theta|Y, Z)|Y] \approx \frac{\sum_{j=1}^{m} w_j p(\theta|Y, Z_j)}{\sum_{j=1}^{m} w_j}$$

6.7.2 序贯填补

Kong, Liu and Wong (1994) 提出序贯填补方法产生模拟缺失数据。

$$X = [X_1, \ldots, X_n] = [Y_1, Z_1, Y_2, Z_2, \ldots, Y_n, Z_n].$$

令缺失数据$z(j)$ 有n 个元素：$z_1(j), \ldots, z_n(j)$. 为了从$p(z|Y)$ 中抽到样本，这个算法从$p(z_1|y_1)$ 抽取z_1^* 并计算权重$w_1 = p(y_1)$，其中y_1 是y 的第一个元素。对于$t = 2, \ldots, n$, 顺序执行下面两步：

1. 抽取$z_t^* \sim p(z_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*, y_t)$

2. 计算$p(y_t|y_1, z_1^*, \ldots, y_{t-1}, z_{t-1}^*)$, 并令$w_t = w_{t-1}p_t$.

注:

$$p(X_t|X_1, \ldots, X_{t-1}) = \frac{p(X_1, \ldots, X_t)}{p(X_1, \ldots, X_{t-1})}$$

$$p(X_1, \ldots, X_t) = \frac{p(\theta|x_1, \ldots, x_t|\theta)\pi(\theta)}{p(\theta|x_1, \ldots, x_t)}$$

注意到每个z_t^* 的抽取依赖于先前抽取的$z_1^*, \ldots, z_{t-1}^*$, 所以z_t^* 的抽取是序贯的。上面两个步骤重复m 次，得到

$$\tilde{z}^*(1), \ldots, \tilde{z}^*(m)$$

和

$$w(1), \ldots, w(m),$$

其中

$$\tilde{z}^*(j) = (z_1^*(j), \ldots, z_n^*(j)), j = 1, \ldots, m$$

$$w(j) = p(y_1)\prod_{t=2}^{n} p(y_t|y_1, z_1^*(j), \ldots, y_{t-1}, z_{t-1}^*(j)).$$

然后，Kong, Liu and Wong (1994) 用下面的加权混合估计$p(\theta|Y)$:

$$\sum_{j=1}^{m} w(j)p(\theta|Y, z^*(j))/\sum_{j=1}^{m} w(j).$$

从第1步，可以看到$z^*(j)$ 来自于如下分布：

$$p^*(z^*(j)|Y) = p(z_1^*(j)|y_1) \times \prod_{t=2}^{n} p(z_t^*(j)|y_1, z_1^*(j), \ldots, y_{t-1}, z_{t-1}^*(j), y_t),$$

而不是$P(Z|Y)$. 所以，在计算$p(\theta|Y)$ 时，需要计算一个权重:

$$\begin{aligned}
p(z^*(j)|Y)/p^*(z^*(j)|Y) &= p(z^*(j)|Y)/[p(z_1^*(j)|y_1) \times \prod_{t=2}^{n} p(z_t^*(j)|y_1, z_1^*(j), \ldots, y_{t-1}, z_{t-1}^*(j), y_t)] \\
&= \frac{p(z^*(j), Y)}{p(Y)} \frac{p(y_1)}{p(y_1, z_1^*(j))} \prod_{t=2}^{n} \frac{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_{t-1}^*(j))}{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_t^*(j))} \\
&= \frac{p(z^*(j), Y)}{p(Y)} \frac{p(y_1)}{p(y_1, z_1^*(j))} \prod_{t=2}^{n} \frac{p(y_t|y_1, \ldots, y_{t-1}, z_1^*(j), \ldots, z_{t-1}^*(j))p(y_1, \ldots, y_{t-1}, z_1^*(j), \ldots, z_{t-1}^*(j))}{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_t^*(j))} \\
&= \frac{p(z^*(j), Y)}{p(Y)} \frac{p(y_1)}{p(y_1, z_1^*(j))} \prod_{t=2}^{n} p(y_t|y_1, \ldots, y_{t-1}, z_1^*(j), \ldots, z_{t-1}^*(j)) \prod_{t=2}^{n} \frac{p(y_1, \ldots, y_{t-1}, z_1^*(j), \ldots, z_{t-1}^*(j))}{p(y_1, \ldots, y_t, z_1^*(j), \ldots, z_t^*(j))} \\
&= \frac{p(z^*(j), Y)}{p(Y)} \frac{p(y_1)}{p(y_1, z_1^*(j))} \prod_{t=2}^{n} p(y_t|y_1, \ldots, y_{t-1}, z_1^*(j), \ldots, z_{t-1}^*(j)) \frac{p(y_1, z_1)}{p(y_1, \ldots, y_n, z_1^*(j), \ldots, z_n^*(j))} \\
&= \frac{w_j}{p(Y)}
\end{aligned}$$

因为$p(Y)$ 是一个常数，所以$w(j)/\sum w(j)$ 提供了正确的权重。具体细节及例子参见Sequential Imputationsand Bayesian Missing Data Problems by Augustine KONG，Jun S. Liu, and Wing Hung WONG (1994).

6.7.3 计算后验

在计算后验分布的更新的时候，需要计算

$$\frac{1}{m}\sum_{j=1}^{m} p(\theta|Y, Z_j).$$

在计算上面这个分布的时候，通常需要计算一个正则常数。Chen (1994) 提出了一个重要性抽样算法，避免了这个常数的计算。

注意到，对于任意的条件密度$w(\theta|Z, Y)$ 都有下式成立：

$$p(\theta^*|Y) = \int_\Theta \int_Z w(\theta|Z, Y)\frac{p(\theta^*, Z|Y)}{p(\theta, Z|Y)}p(\theta, Z|Y)dZd\theta,$$

这是因为Fubini 定理保证了上式双重积分等于下式

$$\int_Z p(\theta^*, Z|Y)[\int_\Theta w(\theta|Z, Y)d\theta]dZ = p(\theta^*|Y).$$

所以，Chen (1994) 建议使用下式近似$p(\theta^*|Y)$:

$$\frac{1}{n}\sum_{j=1}^{n} w(\theta_j|z_j, Y)\frac{p(\theta^*, Z_j|Y)}{p(\theta_j, Z_j|Y)},$$

其中$\theta_j, Z_j \sim p(\theta, Z|Y)$.

注意：这个时候$p(\theta, Z|Y)$ 的正则常数不必计算，因为这个常数在比率$\frac{p(\theta^*,Z_j|Y)}{p(\theta_j,Z_j|Y)}$的分子分母中消失了。

Chapter 7

MCMC

本章主要介绍两种MCMC方法：Gibbs 抽样和Metropolis 算法

7.1 Gibbs 抽样介绍

为了引出Gibbs 抽样，我们首先考虑一个修改的数据扩充算法“链数据扩充”。Gibbs 抽样其实就是链数据扩充算法的多元推广。

7.1.1 链数据扩充

首先考虑数据扩充中两个基本等式：

$$\text{后验等式：} p(\theta|Y) = \int_Z p(\theta|Y,Z))p(Z|Y)dZ,$$

$$\text{预测等式：} p(Z|Y) = \int_\Theta p(Z|Y,\theta)p(\theta|Y)d\theta.$$

考虑“一步” 数据填补的数据扩充算法：即在填补部，只抽取一次$p(Z|Y)$:

a1: Z^* 给定，抽取θ^* from $p(\theta|Z^*,Y)$

a2: θ^* 给定，抽取Z^* from $p(Z|\theta^*,Y)$.

重复以上步骤，即得到“一步” 数据填补的数据扩充算法。这就是链数据扩充算法。在“完整” 的数据扩充算法中，需要执行以上步骤，m 次，得到$z^{(1)},\ldots,z^{(m)}$, 然后更新后验$p(\theta|Y)$. 当$m=1$ 时，更新的后验就是$p(\theta|Z^*,Y)$.

对于链数据扩充算法，记第i 次循环(执行a1 和a2 为一个循环) 得到的θ 为$\theta^{(i)}$, Tanner and Wong (1987) 指出：

1. $\theta^{(1)},\ldots\theta^{(i)},\ldots$ 是一个马氏链。

2. 该马氏链的稳定分布为：$p(\theta|Y)$

$$\theta^{(i)} \to^d \theta \sim p(\theta|Y).$$

3.

$$\frac{1}{t}\sum_{i=1}^{t} f(\theta^(i)) \to^{a.s.} E_{p(\theta|Y)}(f(\theta)).$$

例：层次模型。(Hierarchical Models) 假定要估计d 个总体的均值$\theta = (\theta_1, \ldots, \theta_d)$. 已观测到$d$ 个独立的正态分布样本均值$Y = (Y_1, \ldots, Y_d)$. 给定θ_i 是来自$N(\theta_i, V_i)$ 的独立样本，并且V_i 是已知的。θ_i 的先验分布是独立共轭的，A 未知，给定A 下，θ_i 是来自$N(0, A)$ 的$i.i.d.$ 样本。超参数A 的分布密度是$f(A) = cA^{-1-q/2}\exp(-0.5\lambda/A)$, $\lambda > 0, q > 0$ 是已知常数。

易知，后验分布$p(\theta|Y, A)$ 是正态分布，其第j 个元素为

$$N((1 - B_j)Y_j, V_j(1 - B_j)),$$

$B_j = V_j/(V_j + A)$.

分布$p(A|\theta, Y)$ 为逆卡方分布：

$$\frac{\lambda + \|\theta\|^2}{\chi^2_{d+q}}.$$

注意到

$$p(\theta|Y) = \int p(\theta|Y, A)p(A|Y)dA,$$

$$p(A|Y) = \int p(A|\theta, Y)p(\theta|Y)d\theta,$$

所以，这个时候，A 扮演一个缺失数据的角色。这个时候，链数据扩充算法为:

1. 从$p(\theta|Y, A) = N((1 - B_j)Y_j, V_j(1 - B_j))$ 抽取θ_j

2. 从$p(A|\theta, Y) = \frac{\lambda+\|\theta\|^2}{\chi^2_{d+q}}$ 抽取A

重复以上过程，直至收敛。

例：由条件分布取得边缘分布。已知$p(X|Y), p(Y|X)$. 求$p(X) or p(Y)$. 我们以求$p(X)$ 为例。这个时候我们视Y 为缺失数据。注意到如下两个等式：

$$p(X|\Omega) = \int_Y p(X|Y)p(Y)dY,$$

$$p(Y|\Omega) = \int_x p(Y|X)p(X)dX.$$

1. 从$p(X|Y)$ 抽取X

2. 从$p(Y|X)$ 抽取Y

重复以上过程，直至收敛。

7.1.2 多元链数据扩充–Gibbs 抽样

现在考虑链扩充算法的多元推广，称为Gibbs 抽样。给定起始点$(\theta_1^{(0)},\ldots,\theta_d^{(0)})$, 循环下面步骤：

1. 从$p(\theta_1|\theta_2^{(i)},\ldots,\theta_d^{(i)},Y)$ 抽取$\theta_1^{(i+1)}$
2. 从$p(\theta_2|\theta_1^{(i+1)},\theta_2^{(i)}\ldots,\theta_d^{(i)},y)$ 抽取$\theta_2^{(i+1)}$

⋮

3. 从$p(\theta_d|\theta_1^{(i+1)},\ldots,\theta_{d-1}^{(i+1)},Y)$ 抽取$\theta_d^{(i+1)}$

向量$(\theta^{(0)},\ldots,\theta^{(t)},\ldots)$ 是一个马氏链。转移概率是:

$$K(\theta',\theta)=p(\theta|\theta')=p(\theta_1|\theta_2',\ldots,\theta_d',Y)\times p(\theta_2|\theta_1,\theta_3',\ldots,\theta_d',Y)\times p(\theta_3|\theta_1,\theta_2,\ldots,\theta_d',Y)\times\ldots\times p(\theta_d|\theta_1,\ldots,\theta_{d-1},Y)$$

很多研究人员给出了如下结果成立的条件：

结果1：$(\theta_1^{(i)},\ldots,\theta_d^{(i)})$ 的联合分布收敛到$p((\theta_1,\ldots,\theta_d)|Y)$

结果2：$\frac{1}{n}\sum_{i=1}^n f(\theta^{(i)})\to^{a.s.}\int f(\theta)p(\theta|Y)d\theta,$ as $n\to\infty.$

结果3：$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(\theta^{(i)})-\int f(\theta)p(\theta|Y)d\theta\right)\Rightarrow N(0,\sigma_f^2).$

例：（基因连锁模型）

Gelfand and Smith (1990) 考虑了基因连锁例子的扩展。数据为$(14,1,1,1,5)$, 模型为

$$\left[(\frac{\theta}{4}+\frac{1}{8}),\frac{\theta}{4},\frac{\eta}{4},\frac{\eta}{4}+\frac{3}{8},\frac{1}{2}(1-\theta-\eta)\right]$$

扩充数据集合为

$$X=(X_1,\ldots,X_7)\sim Multinomial\left(22;\frac{\theta}{4},\frac{1}{8},\frac{\theta}{4},\frac{\eta}{4},\frac{\eta}{4},\frac{3}{8},\frac{1}{2}(1-\theta-\eta)\right)$$

缺失数据：$Z=[X_1,X_5].$

X_1	X_2	Y_2	Y_3	X_5	X_6	Y_5
$\frac{\theta}{4}$	$\frac{1}{8}$	$\frac{\theta}{4}$	$\frac{\eta}{4}$	$\frac{\eta}{4}$	$\frac{3}{8}$	$\frac{1}{2}(1-\theta-\eta)$

取先验分布Dirichlet(1,1,1) (共轭先验or flat prior)，注：$\theta\sim Dirichlet(\alpha_1,\alpha_2,\ldots,\alpha_k)$, 则

$$p(\theta)\propto\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}.$$

则易知

$$p(\theta|\eta,Y,Z)=(1-\eta)Beta(X_1+Y_2+1,Y_5+1),\theta\in[0,1-\eta]$$

说明：

$$p(\theta,\eta|Y,Z) \propto \theta^{X_1}\theta^{Y_2}\eta^{Y_3}\eta^{X_5}(1-\theta-\eta)^{Y_5} = \theta^{X_1+Y_2}\eta^{Y_3+X_5}(1-\theta-\eta)^{Y_5}.$$

$$p(\eta|\theta,Y,Z) = (1-\theta)Beta(Y_3+X_5+1,Y_5+1), \eta \in [0,1-\theta]$$

$$p(Z|\eta,Y,\theta) = P(X_1,X_5|\eta,Y,\theta) = Bi(Y_1,2\theta(1+2\theta)^{-1})Bi(Y_4,2\eta(3+2\eta)^{-1})$$

用Gibbs 抽样，重复5000 次，可以得到：

$$E(\eta|Y) = 0.123, E(\eta^2|Y) = 0.022, Var(\eta|Y) = 0.0065 E(\theta^2|Y) = 0.288, var(\theta|Y) = 0.018.$$

Homework.

更多的例子见Tanner. Tools for statistical inference.

7.1.3 评定Gibbs 链的收敛性

7.2 Metropolis 方法

Metropolis 算法开始是为了研究粒子（如原子中的电子）的大系统的平衡特性。这个算法广发的应用于统计物理的文献来模拟复杂系统。本节介绍如何使用Metropolis 算法构造平衡分布$\pi(x)$ 的Markov 链。首先简单回顾一下离散空间的Markov 链的一些基本知识。

7.2.1 离散空间的Markov 链的一些基本知识

假定状态集合为$S = \{s_1, s_2, \ldots, s_d\}$. 在时刻0, 过程从某个状态开始，单位时间内从状态s_i 转到s_j. 对于Markov 链，时刻$n+1$ 的状态仅依赖于时刻n 的状态。用p_{ij} 表示从状态s_i 转移到s_j 的概率，构成一个$d \times d$ 的矩阵，p_{ij}, $\sum_j p_{ij} = 1$.

一般的，一个具有转移阵P 的Markov 链有一个平稳分布当且仅当$\pi = \pi P$, 其中$\pi \in R^{1\times d}$. 如果$\pi_i p_{ij} = \pi_j p_{ji}$, 称这个链市可逆链。注意到：$\pi_i p_{ij} = \pi_j p_{ji}$ 蕴含$\pi = \pi P$, 这是因为

$$(\pi P)_j = \sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j.$$

所以，为了实现从π 抽样，我们只要能够构造一个以π 为平稳分布的可逆马氏链，然后从任意状态出发，直到链达到平衡分布，即可。Metropolis 等（1953）的贡献之一就是，提出了实

现平衡分布π, 构造P 的一个一般方法。

7.2.2 Metropolis 方法

首先描述一下离散情况下的思想，然后再介绍连续情况。令$Q = q_{ij}$ 是一个对称的转移矩阵。从状态i出发，以Q 的第i 行为概率随机的抽取状态s_j, 再以已知的概率a_{ij} 接受这个状态（从s_i 转移到s_j）, 否则，保持在状态s_i. 这个过程定义了一个Markov 链，其转移矩阵$p_{ij} = q_{ij}\alpha_{ij}$, $p_{ii} = 1 - \sum_{j\neq i} p_{ij}$。Metropolis 等（1953）选取

$$\alpha_{ij} = \begin{cases} 1 & \text{if } \pi_j/\pi_i \geq 1 \\ \pi_j/\pi_i & \text{if } \pi_j/\pi_i \leq 1 \end{cases}.$$

这个马氏链是一个可逆马氏链，因为

$$\begin{aligned} \pi_i p_{ij} &= \pi_i \min\left\{1, \frac{\pi_j}{\pi_i}\right\} q_{ij} \\ &= \min\left\{\pi_i, \pi_j\right\} q_{ij} \end{aligned}$$

$$\begin{aligned} \pi_j p_{ji} &= \pi_j \min\left\{1, \frac{\pi_i}{\pi_j}\right\} q_{ji} \\ &= \min\left\{\pi_i, \pi_j\right\} q_{ji} \\ &= \pi_i p_{ij} \end{aligned}$$

事实上，只要$\pi_i\alpha_{ij} = \pi_j\alpha_{ji}$, P 对应的马氏链就是可逆的。Baker (1965) 选取$\alpha_{ij} = \pi_j/(\pi_i + \pi_j)$.

在连续情况下，π 是密度函数，$f(x, y)$ 是对称的转移概率函数，即$f(x, y) = f(y, x)$, 那么Metropolis 算法为：

1. 如果目前链是在$X_n = x$, 那么从f(x,y) 产生下一个位置X_{n+1}的候选值y^*.

2. 以概率$\alpha(x, y^*) = \min\left\{\frac{\pi(y^*)}{\pi(x)}, 1\right\}$ 接受这个候选值，并且将链移到$X_{n+1} = y^*$. 否则，拒绝这个候选值，并令$X_{n+1} = x$.

注：Metropolis 方法仅要求一个与π 成比例的函数，因为正则常数可以在比率中$\frac{\pi(y^*)}{\pi(x)}$ 约去。

Hastings (1970) 扩展了Metropolis 算法，令

$$\alpha(x, y) = \begin{cases} \min\left\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\} & \text{if } \pi_j/\pi_i \geq 1 \\ 1 & \text{if } \pi(x)q(x, y) = 0 \end{cases},$$

例：考虑基因连锁模型。我们需要从后验分布

$$\pi(\theta) = (2+\theta)^{125}(1-\theta)^{38}\theta^{34}, \theta \in [0,1],$$

抽样。

注意到（1）正则常数是不需要计算的（2）抽样直接从观测后验抽样，不需要像Gibbs 抽样那样，需要抽取缺失数据。

可以选取$f(x,y) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(y-x)^2}{2\sigma^2}}$. 分别取$\sigma = 0.001, 0.05, 0.12$.

1. 如果目前链是在$X_n = x$, 那么从$N(x,\sigma^2)$ 产生下一个位置X_{n+1}的候选值y^*.

2. 以概率$\alpha(x,y^*) = \min\left\{\frac{\pi(y^*)}{\pi(x)}, 1\right\}$ 接受这个候选值，并且将链移到$X_{n+1} = y^*$. 否则，拒绝这个候选值，并令$X_{n+1} = x$.

注意：$\pi(x) = 0$, if $x \notin [0,1]$.

7.2.3 Gibbs 抽样和Metropolis 的关系

Gibbs 抽样是特殊的Metropolis 算法，其接受概率总是1。考虑在Gibbs 抽样中，θ^{t+1} from $p(\theta_1|\theta_2^t, \ldots, \theta_d^t|Y)$. 为什么总是接受这个新的point $(\theta_1^{t+1}, \theta_2^t, \ldots, \theta_d^t)$ 呢？

当前点是$\theta = (\theta_1^t, \theta_2^t, \ldots, \theta_d^t)$, 现在我们从$q(\theta,\theta') = p(\theta_1^{t+1}|\theta_2^t, \ldots, \theta_d^t, Y)$ 抽取了一个新的点$\theta' = (\theta_1^{t+1}, \theta_2^t, \ldots, \theta_d^t)$. 由Gibbs 抽样过程知道，$q(\theta',\theta) = p(\theta_1^t|\theta_2^t, \ldots, \theta_d^t, Y)$. 所以，由Hastings 方法，

$$\begin{aligned}
\alpha(\theta,\theta') &= \frac{p(\theta')q(\theta',\theta)}{p(\theta)q(\theta,\theta')} && (7.1)\\
&= \frac{p(\theta_1^t\theta_2^t, \ldots, \theta_d^t, Y)p(\theta_1^{t+1}|\theta_2^t, \ldots, \theta_d^t, Y)}{p(\theta_1^{t+1}\theta_2^t, \ldots, \theta_d^t, Y)p(\theta_1^t|\theta_2^t, \ldots, \theta_d^t, Y)} && (7.2)\\
&= 1 && (7.3)
\end{aligned}$$

7.2.4 Gibbs 和Metropolis 的混合算法

在Gibbs 算法中，需要从条件分布抽样。比如从$p(\theta_1|\theta_2^{(i)}, \ldots, \theta_d^{(i)})$ 抽样。Muller (1993) 建议用Metropolis 算法。令$\pi(\rho) = p(\rho|\theta_2^{(i)}, \ldots, \theta_d^{(i)})$.

a. 从分布$f(x-y) = f(y-x)$ 产生y^*, 其中x 是链的当前状态。

b. 以概率$\alpha(x,y^*) = \min(1, \pi(y^*)/\pi(x))$ 接受y^* 作为一个新状态；否则保持原来的状态x.

重复a, b T 次，$\theta^{(i+1)}$ 就是这个链的第T 个值。特别的，T 可以取1。

7.3 例子：Bayesian Variable Selection Using Gibbs-Based Methods.

我们着重介绍来自于Dellaportas et al (2002)的方法。

在变量选择问题中，任意一个模型m，可以由一个p 维的二值向量表示：$\gamma \in \{0,1\}^p$. 其中$\gamma_j = 0$ 表示第j 个变量不在模型中，$\gamma_j = 1$ 表示第j 个变量在模型中。

GLM 这样刻画：

$$g(E(Y)) = \sum_{j=0}^{p} \gamma_j X_j \beta_j \tag{7.4}$$

7.3.1 Prior distribution for variable selection in GLM

Zellner (1986) g-prior:

$$\beta_m|\sigma^2, m \sim N\left(\mu_{\beta_m}, c^2(X_m^T X_m)^{-1}\sigma^2\right).$$

关于Zellner's g-prior, 现在有一个直观的解释。现在设想有一个假想的data y^*, 它由GLM (7.4) 模型得到，设计阵是X_m. 现在我们将β 的先验设为这个假想data 的似然的指数(类似conjugate)，或者说从假想的data “借”了一些信息。

$$f(\beta_m|y^*, m) = [f(y^*|\beta_m, m)]^{1/c^2}$$

从上面的定义可知，这种先验相当于“借了” n/c^2 个data point。在正态模型，即$g(x) = x$ 并且noise 是正态分布时，上面的prior gives

$$\beta_m|\sigma^2, m \sim N\left((X_m^T X_m)^{-1} X_m' y^*, c^2(X_m^T X_m)^{-1}\sigma^2\right).$$

取$y^* = X_m \mu_{\beta_m}$ 即得Zellner's $g-$prior.

如果没有任何的先验信息，通常取$\mu_{\beta_m} = 0$. $c^2 = n$, 即“借” 一个data point。

对于其他模型，如logistic regression, poisson regression β 的先验可以类似得到。

在模型选择中，通常对模型使用无信息先验:

$$f(m) = \frac{1}{|\mathcal{M}|}, \text{ for all } m \in \mathcal{M}.$$

这等价于

$$\gamma_j \sim Bernoulli(1/2), \text{ for all } j = 1, 2, \ldots, p.$$

7.3.2 Gibbs variable selection (GVS)

为了实现GVS, 我们首先要指定先验分布。

$$f(\gamma, \beta) = f(\gamma) f(\beta|\gamma)$$

现在我们将β 分成两部分：$(\beta_\gamma, \beta_{\setminus\gamma})$, 其中$\beta_\gamma$ 是选进模型的β_j's, $\beta_{\setminus\gamma}$ 是没有选进模型的β_j's. 那么先验进一步表示为：

$$f(\gamma, \beta) = f(\gamma) f(\beta|\gamma) = f(\gamma) f(\beta_\gamma|\gamma) f(\beta_{\setminus\gamma}|\beta_\gamma, \gamma).$$

有了以上的先验的设定，就可以得到（联合）后验：

$$f(\beta_\gamma, \beta_{\setminus\gamma}, \gamma|y) = f(y|\beta, \gamma) f(\gamma) f(\beta_\gamma|\gamma) f(\beta_{\setminus\gamma}|\beta_\gamma, \gamma).$$

从而，得到完全条件分布：

$$f(\beta_\gamma|\beta_{\setminus\gamma}, \gamma, y) \propto f(y|\beta, \gamma) f(\beta_\gamma|\gamma) f(\beta_{\setminus\gamma}|\beta_\gamma, \gamma)$$

$$f(\beta_{\setminus\gamma}|\beta_\gamma, \gamma, y) \propto f(\beta_{\setminus\gamma}|\beta_\gamma, \gamma)$$

$$f(\gamma_j|\beta, \gamma_{\setminus j}, y) \sim Bernoulli(\frac{O_j}{1+O_j}),$$

其中

$$O_j = \frac{f(\gamma_j = 1, \gamma_{\setminus j}, \beta, y)}{f(\gamma_j = 0, \gamma_{\setminus j}, \beta, y)} = \frac{f(y|\gamma_j = 1, \gamma_{\setminus j}, \beta) f(\beta|\gamma_j = 1, \gamma_{\setminus j}) f(\gamma_j = 1, \gamma_{\setminus j})}{f(y|\gamma_j = 0, \gamma_{\setminus j}, \beta) f(\beta|\gamma_j = 0, \gamma_{\setminus j}) f(\gamma_j = 0, \gamma_{\setminus j})}.$$

在实际应用中，通常选取$f(\beta_{\setminus\gamma}|\beta_{\gamma,\gamma}) = f(\beta_{\setminus\gamma}|\gamma)$.

7.3.3 Posterior Inference

Bayesian variable selection 的主要目的是得到一个最大后验(MAP: maximum a posteriori)的模型。

后验概率可以这样得到：

$$\hat{P}(m|y) = \frac{1}{T-B} \sum_{t=B+1}^{T} I(m^t = m)$$

变量j 选进模型的概率:

$$\hat{P}(\gamma_j = 1|y) = \frac{1}{T-B}\sum_{t=B+1}^{T} I(\gamma_j^t = 1).$$

有时候，我们并不希望选出一个模型。这时候，可以考虑使用所有的模型，然后加权求平均。比如，如果希望得到预测ξ 的分布:

$$f(\xi|y) = \sum_{m\in\mathcal{M}} f(\xi|y,m)f(m|y).$$

7.3.4 Implementation in WinBugs

本节通过一个简单的正态模型来展示如何使用WinBugs 做MCMC 的计算。其他更多的模型和例子可以参见Ntzoufas et al(2000) "Stochastic search variable selection for log-linear models", Journal of Statistical Computation and Simulation, pp 23 – 38.

$n = 50, p = 15$, $X_j, j = 1, \ldots, p \sim_{i.i.d.} N(0,1)$.

$$Y_i \sim N(X_{i4} + X_{i5}, 2.5^2), i = 1, \ldots, n.$$

Data 和code 可见于http://stat-athens.aueb.gr/∼jbn/winbugs_book/.

Chapter 8

Bootstrap 方法

8.1 Bootstrap

目标： 通过一系列随机样本$x = (x_1, \ldots, x_n) \sim F$, 其中$F$ 未知，估计某一指定的随机变量$R(X, F)$ 的分布。

8.1.1 Bootstrap 方法

我们首先描述Bootstrap 方法：

1. 构造经验分布$\hat{F}$: $P(X = x_i) = \frac{1}{n}$;
2. 从经验分布$\hat{F}$ 独立抽取n 个样本点。记为

$$x_i^*, i = 1, 2, \ldots, n.$$

注意：x_i^*其实就是有放回的从$\{x_1, \ldots, x_n\}$ 抽取的。

3. 计算$R^* = R(x^*, \hat{F})$.

以上过程重复很多次，即可以得到观测R^*，从而得到R^* 的经验分布。用此经验分布近似$R(X, F)$ 的经验分布的方法，就是Bootstrap 方法。

注意到, 样本 x 观测到之后，R^* 的分布是可以精确计算出来的，不需要使用重抽样即可。这是因为x^* 的分布这个时候是已知的,它是 F^*.

例：考虑$x = (x_1, x_2, \ldots, x_n)$ i.i.d. from $Bernoulli(p)$. 现在关心的统计量是 bias：

$$R(x, F) = \bar{x} - P(X = 1).$$

标准的统计方法告诉我们：

$$E(R(x,F))=0, var(R(x,F))=\frac{p(1-p)}{n},$$

在实际计算中，常常取

$$\hat{var}(R(x,F))=\frac{\bar{x}(1-\bar{x})}{n},$$

现在我们来看Bootstrap 方法计算的E_* 和var_*. 注：我们使用E_* 和 var_* 代表由Bootstrap方法得到的期望和方差。

注意x_i 取值是1或者0。而且经验频率$\hat{P}(x_i=1)=\bar{x}$. 由 Bootstrap 抽样得到的统计量为：

$$R^*=R(X^*,\hat{F})=\bar{X}^*-\bar{x}$$

由X^*的抽样过程知道，它的均值和方差为：

$$E_*(R^*)=0, var_*(R^*)=\frac{\bar{x}(1-\bar{x})}{n}.$$

例：估计方差。$X\sim F$。考虑估计$var_F(X)$.使用如下的估计量：

$$t(X)=\sum_{i=1}^{n}(X_i-\bar{X})^2/(n-1).$$

现在我们的目标是估计下面的Bias 的分布：

$$R(X,F)=t(X)-var_F(X).$$

定义$\mu_k(F):=E_F(X-E_FX)^k$. 用$\hat{\mu}_k(\hat{F})$ 代表k 阶中心样本矩。

由Bootstrap样本得到的统计量为

$$R^*=R(X^*,\hat{F})=t(X^*)-var(\hat{F}).$$

可以证明，

$$E_*(R^*)=0,\quad var_*(R^*)=\frac{\hat{\mu}_4-[(n-3)/(n-1)]\hat{\mu}_2^2}{n}.$$

第一个式子很简单。如果y_i, i.i.d. from F with a finite variance σ^2, then $E(\sum_{i=1}^{n}(y_i-\bar{y})/(n-1))$ 是σ^2 的无偏估计。注意到$X_i^*\sim\hat{F}$, 所以有$E_*t(X^*)=var(\hat{F})$.

Homework.

在上面的两个例子中，因为要计算的统计量较为简单（都是多项式），所以可以直接计算。在实际的应用中，Bootstrap 方法的难点在于如何计算由Bootstrap样本得到的统计量的Bootstrap 期望和方差，更一般的，统计量的分布。

常用处理该难点的方法:

方法1. 像前两个例子中那样直接计算。

方法2. Monte Carlo 近似。 多次抽取 Bootstrap 样本，并计算$R(x^*, \hat{F})$。使用Histogram近似$R(X^*, \hat{F})$。

8.1.2 估计中位数

本小节，我们希望估计分布F 的中位数，记为$\theta(F)$. 用$t(X)$ 表示样本中位数，

$$t(X) = X_{(m)},$$

其中$X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ 是次序统计量，$n = 2m - 1$. 我们的目标是估计如下统计量的分布：

$$R(X, F) = t(X) - \theta(F).$$

观测数据为$x_1, \ldots, x_n$, bootstrap 抽样记为$X^* = (x_1^*, \ldots, x_n^*)$, 由此得到Bootstrap 统计量

$$R^* = R^*(X^*, \hat{F}) = X_{(m)}^* - x_{(m)}.$$

注意到对于任意的正整数$1 \le \ell < n$,

$$Prob_*(X_{(m)}^* > x_{(\ell)}) = Prob\left(Binomial(n, \frac{\ell}{n}) \le m - 1\right) = \sum_{j=0}^{m-1} (\ell/n)^j (1 - \ell/n)^{n-j}.$$

Homework.

所以，

$$Prob_*(R^* = x_{(\ell)} - x_{(m)}) = Prob_*(X_{(m)}^* > x_{(\ell-1)}) - Prob_*(X_{(m)}^* > x_{(\ell)}).$$

作为一个特例，$n = 13(m = 7)$, 下表给出R^* 的分布：

$\ell =$	2 or 12	3 or 11	4 or 10	5 or 9	6 or 8	7
	0.0015	0.0142	0.0550	0.1242	0.1936	0.2230.

8.1.3 判别分析中的错误率

本节讨论标准线性判别分析问题中的错误率估计问题。

数据：

$$X_i \sim_{i.i.d.} F, i = 1, 2, \ldots, m,$$

$$Y_j \sim_{i.i.d.} G, j = 1, 2, \ldots, n.$$

线性判别分析的任务：根据观测到的data，将空间分为两个区域A 和B. 对于新来的一个观测z, 如果$z \in A$, 就认为它来自F; 如果来自B, 就认为它来自G.

对于F, 一个显然的错误率

$$error_F := Prob\{X \in B\}.$$

的估计是

$$\widehat{error}_F = \frac{\#\{x_i \in B\}}{m}.$$

注意：在计算概率的时候，我们视B fixed.

我们关心的统计量是

$$R((X, Y), (F, G)) = error_F - \widehat{error}_F.$$

对于观测数据x 和y, 线性判别(LDA) 这样定义B：

$$B = \left\{ z : (\bar{y} - \bar{x})' S^{-1} (z - \frac{\bar{x} + \bar{y}}{2}) > \log \frac{m}{n} \right\},$$

其中$\bar{x} = \sum x_i/m, \bar{y} = \sum y_j/n$ and $S = \sum_i (x_i - \bar{x})(x_i - \bar{x})' + \sum_j (y_j - \bar{y})(y_j - \bar{y})'/(m + n)$.

在Gauss 分布下，我们推导一下LDA. 设F 和G 都是Gauss分布，且具有相同的方差。则，F 的density 是

$$f(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp[-\frac{1}{2}(x - \mu_x)^T \Sigma^{-1} (x - \mu_x)].$$

用π_A 代表属于A 的prior，π_B 代表属于B 的prior，那么

$$\begin{aligned}
\log\frac{P(z\in A|z)}{P(z\in B|z)} &= \log\frac{f_x(z)\pi_A}{f_y(z)\pi_B} \\
&= \left[\log(\pi_A)-\frac{1}{2}\log|\Sigma|-\frac{1}{2}(z-\mu_x)^T\Sigma^{-1}(z-\mu_x)\right] \\
&\quad -\left[\log(\pi_B)-\frac{1}{2}\log|\Sigma|-\frac{1}{2}(z-\mu_y)^T\Sigma^{-1}(z-\mu_y)\right] \\
&= \log(\frac{\pi_A}{\pi_B})+\frac{1}{2}[(z-\mu_y)^T\Sigma^{-1}(z-\mu_y)-(z-\mu_x)^T\Sigma^{-1}(z-\mu_x)] \\
&= \log(\frac{\pi_A}{\pi_B})+(\mu_x-\mu_y)^T\Sigma^{-1}z+\frac{1}{2}[\mu_y^T\Sigma^{-1}\mu_y-\mu_x^T\Sigma^{-1}\mu_x] \\
&= \log(\frac{\pi_A}{\pi_B})+(\mu_x-\mu_y)^T\Sigma^{-1}(z-\frac{\mu_x+\mu_y}{2})
\end{aligned}$$

最后一个等式成立，是因为

$$\begin{aligned}
x'Ax-y'Ay &= x'A(\frac{x+y}{2})+x'A(\frac{x-y}{2})-[y'A(\frac{x+y}{2})+y'A(\frac{y-x}{2})] \\
&= (x-y)'A(\frac{x+y}{2})+\frac{1}{2}[x'Ax-y'Ay].
\end{aligned}$$

所以$\frac{1}{2}[x'Ax-y'Ay]=(x-y)'A(\frac{x+y}{2})$.

所以，要判断$z\in B$, 就是

$$\log(\frac{\pi_A}{\pi_B})+(\mu_x-\mu_y)^T\Sigma^{-1}(z-\frac{\mu_x+\mu_y}{2})<0.$$

将各种估计带入上式，即有

$$B=\left\{z:(\bar{y}-\bar{x})'S^{-1}(z-\frac{\bar{x}+\bar{y}}{2})>\log\frac{m}{n}\right\}.$$

接下来，我们使用Bootstrap 方法，估计R 的分布。

1. 抽取$x_i^*\sim_{iid}\hat{F}$, i =1,..., m; $y_j^*\sim_{iid}\hat{G}, j=1,2,\ldots,n$.

2. 计算B^*

3. 计算Bootstrap 统计量

$$R^*((X^*,Y^*),(\hat{F},\hat{G}))=error_{\hat{F}}-\widehat{error}_F=\frac{\#\{x_i\in B^*\}}{m}-\frac{\#\{x_i^*\in B\}}{m}.$$

换句话讲，上式就是“真实错误率（对应真实（经验）分布）”－“估计的错误率（对应观测（Bootstrap 样本）分布）”。

以上过程重复许多次，得到R^* 的样本，用这些可以近似R^* 的分布，以此估计R 的分布。

现在考虑$F=N((-1/2,0)',I)$, $G=N((1/2,0)',I)$. $m=n=20$. 使用simulation 来测

试Bootstrap.

	Mean	SD
Error Rate	*	*
Bootstrap Expectation (E_*R^*)	*	*
Bootstrap SD ($SD_*(R^*)$)	*	*

Simulation 中，所有的重复次数均取100。

Homework.

实验发现，$E_*R^* \neq 0$. 这意味着$\widehat{error}_F$ 是一个有偏估计。所以一个更好的纠偏的估计是：$\widehat{error}_F + E_*(R^*)$.

8.1.4 回归分析

一般的回归模型是

$$X_i = g_i(\beta) + \epsilon_i, i = 1, 2, \ldots, n,$$

其中，$g(\cdot)$ 是已知的函数，含有未知的参数β; $\epsilon_i \sim_{i.i.d} F, i = 1, 2, \ldots, n$. $E_F(\epsilon_i) = 0$.

β 可以由最小二乘估计给出：

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} [x_i - g_i(\beta)]^2.$$

现在我们希望得到$\hat{\beta}$ 的分布。这是参数Bootstrap.

注意，这里X认为是fixed. 随机性来自于$\epsilon \sim F$. 所以，Bootrstrap 抽样，要从$\hat{F}$ 去抽取。

$$\hat{F}: \text{ mass } \frac{1}{n} at \hat{\epsilon}_i = x_i - g_i(\hat{\beta}), i = 1, 2, \ldots, n.$$

Bootstrap 样本：

$$X_i^* = g_i(\hat{\beta}) + \epsilon_i^*, \epsilon_i^* \sim_{i.i.d.} \hat{F}, i = 1, 2, \ldots, n.$$

由以上Bootstrap 样本可以得到$\hat{\beta}^*$.

重复以上过程可以得到许多$\hat{\beta}^*$.以此可以近似$\hat{\beta}^*$ 的分布，此分布用来估计$\hat{\beta}$ 的分布。

我们使用一个简单的例子，来说明Bootstrap 在回归分析中的应用。

设$Y = X\beta + \epsilon$. $\hat{\beta} = (X'X)^{-1}X'Y$. 如果$X$的第一列是$\mathbf{1}$. 则误差$\hat{\epsilon}_i$的经验分布均值为0，方差为$\hat{\sigma} = mean(Y - X\hat{\beta})^2$. 所以，可以得到$\hat{\beta}^* = (X'X)^{-1}X'(X\hat{\beta} + \hat{\epsilon})$的均值和期望分别是：$E_*(\hat{\beta}^*) = \hat{\beta}, Cov_*(\hat{\beta}^*) = \hat{\sigma}^2(X'X)^{-1}$.

8.2 Jackknife

8.2.1 Jackknife 方法

$Y_1, \ldots, Y_n$ 是n *i.i.d.* 抽样。用$\hat{\theta}$ 表示由此n个样本得到的参数θ 的估计。将n 个样本分成g个Group, 每个group 含有h 个元素：$n = gh$, 特别的，$g = n, h = 1$. 令$\hat{\theta}_{-i}$ 是根据删除第i个Group 的数据得到的估计。定义(pseudo value)

$$\tilde{\theta}_i = g\hat{\theta} - (g-1)\hat{\theta}_{-i}, i = 1, \ldots, g.$$

Jackknife estimator is

$$\hat{\theta}_J = \frac{1}{g}\sum_{i=1}^{g}\tilde{\theta}_i = g\hat{\theta} - (g-1)\frac{1}{g}\sum_{i=1}^{g}\hat{\theta}_{-i}.$$

The estimate of variance of $\tilde{\theta}_J$ is

$$\widehat{var}(\hat{\theta}_J) = \frac{1}{g(g-1)}\sum_{i=1}^{g}(\tilde{\theta}_i - \hat{\theta}_J)^2 = \frac{g-1}{g}\sum_{i=1}^{g}(\hat{\theta}_{-i} - \hat{\theta}_{(\cdot)})^2$$

其中$\hat{\theta}_{(\cdot)} = \frac{1}{g}\sum_{i=1}^{g}\hat{\theta}_{-i}$.

例：$X_1, \ldots, X_n$ iid. from F. F 未知。现在考虑F 的期望。通常使用

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

估计F 的期望。

事实上，对于这个简单的例子，我们还可以给出$\bar{x}$ 的方差：

$$\hat{\sigma}^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

但是，如果是比较复杂的统计量，比如中位数，就很难给出方差的估计。Bootstrap和Jackknife 是这个例子的扩展，可以计算出任意统计量的方差。Bootstrap 方法我们已经了解过。现在我们首先使用前面介绍的Jackknife 方法计算Jackknife 估计和方差。如果没有任何说明，以后的Jackknife 中取$h = 1, g = n$.

1. 计算pseudo value

$$\bar{x}_{-i} = \frac{1}{n-1}(n\bar{x} - x_i)$$

$$\tilde{x}_i = n\bar{x} - (n-1)\bar{x}_{-i} = n\bar{x} - (n-1)\frac{1}{n-1}(n\bar{x} - x_i) = x_i$$

2. 计算Jackknife 估计：

$$\hat{x}_J = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i = \frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}.$$

3. 计算方差：

$$\hat{\sigma}_J^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{x}_i - \hat{x}_J)^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

可以看到Jackknife 估计的方差和前面给出的方差一样。Jackknife 的一个好处是，计算过程简单，可以对任意的估计量，计算出方差。

为了对比，我们这里也给出由Bootstrap 方法得到的estimator的方差。

回顾Bootstrap 方法：

1. 从经验$\hat{F}$抽取$x_1^*, \ldots, x_n^*$.

2. 由Bootstrap sample 计算Bootstrap 统计量:

$$\bar{X}^* = \frac{1}{n}\sum_{i=1}^{n}x_i^*.$$

3. 用Bootstrap 统计量的分布，近似estimator 的分布。所以，估计的方差是

$$var_*(\bar{X}^*) = var_*(\frac{1}{n}\sum_{i=1}^{n}x_i^*) = \frac{1}{n}var_*(x_i^*) = \frac{1}{n}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{n}{n-1}\sigma_J^2.$$

8.2.2 Why Jackknife?

给出估计的方差

可以通过一个简单的步骤，给出估计的方差。

剪偏

一个结论：如果

$$E(\hat{\theta}) = \theta + a_1/n + O(1/n^2),$$

那么

$$E(\hat{\theta}_J) = \theta + O(1/n^2).$$

例：

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}^2)$$

.

$E(\hat{\sigma}^2) = \frac{n-1}{n}\theta = \theta + (-\theta)/n$，没有高阶项，所以，经过Jackknife 后，偏差完全剪掉。

8.3 Bootstrap 和Jackknife 之间的关系

现在考虑$\theta(F)$. 样本是$x_1, \ldots, x_n$. 记估计为$\hat{\theta} = \theta(\hat{F})$. 用一个向量

$$P = (P_1, P_2, \ldots, P_n)$$

表示一个和为1的概率(非负)向量。用该向量重新定一个“经验分布”

$$\hat{F}(P): \text{ mass } P_i \text{ on } x_i, i = 1, 2, \ldots, n.$$

定义

$$\hat{\theta}(P) = \theta(\hat{F}(P)).$$

现在我们重新用另一种方式来考虑Bootstrap.

对于一次抽样，记$P_i^* = \#\{X_i^* = x_i\}/n$

$$P^* = (P_1^*, P_2^*, \quad \ldots, P_n^*).$$

至此，我们可以看得出来，Bootstrap 和Jackknife 在计算方差的时候，总体思路是一致的：

对于每一步，给样本$(x_1, x_2, \ldots, x_n)$ 一个权重:

在Bootstrap 中，权重是$(P_1^*, P_2^*, \ldots, P_n^*)$;

在Jackknife 中，权重是$1/(n-1) * (1, 1, \ldots, \underbrace{0}_{i}, \ldots, 1)$

8.4 Cross Validation

我们简单介绍一下与Bootstrap 和Jackknife 有关的Cross Validation. 首先介绍Leave One Out Cross Validation (LOO).

Cross validation 是为了计算prediction error 的。

假设我们有一个model。

$$y = g(x) + \epsilon.$$

现在有一些观测$(x_i, y_i), i = 1, 2, \ldots, n.$

LOO 的过程是这样的：

对于i = 1,2,. . . , n

1. 第i 步，去掉第i 个观测，拟合模型

2. 用上面拟合的模型，去预测第i 个观测，从而得到prediction error e_i

最后，prediction error 就是$\frac{1}{n}\sum_{i=1}^{n} e_i$.

例: 考虑一个简单的模型. 给定观测数据$x_1, \ldots, x_n \sim F$。现在给定一个新的观测$x_0 \sim F$. 我们使用$\bar{x}$ 预测它。这样，预测误差是

$$E(x_0 - \bar{x})^2 = \frac{n+1}{n}\sigma^2.$$

现在我们用LOO CV 计算这个预测误差：

1. 去除第i 个观测，fit 模型，得到预测值$\bar{x}_{-i}$.

2. 计算prediction error: $e_i = (x_i - \bar{x}_{(-i)})^2$

平均一下，得到prediction error:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_{(-i)})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \frac{n\bar{x} - x_i}{n-1})^2 = \frac{n}{(n-1)^2}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

K-fold CV: K-fold CV 是将Data 随机的分成K 份，然后使用和上面一样的方法估计prediction error. 不同的地方在于，LOO 是删除第i个观测，用剩余的data去fit model, CV 是删除第i个group，用剩余的data 去fit model。

对于i = 1,2,. . . , K

1. 第i 步，去掉第i 个group，拟合模型

2. 用上面拟合的模型，去预测第i 个group，从而得到prediction error e_i

最后，prediction error 就是$\frac{1}{n}\sum_{i=1}^{n} e_i$.

8.5 Asymptotic Theories

8.5.1 Bootstrapping the mean

$X_1, X_2, \ldots, X_n \sim_{iid.} F$. F 未知，均质为μ, 方差是σ^2，两个参数都未知. 传统估计μ 使用如下统计量

$$\mu_n = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

记样本方差是

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_n)^2.$$

中心极限定理告诉我们：

$$Q_n := \sqrt{n}(\mu_n - \mu)/s_n \Rightarrow N(0,1).$$

现在我们来看Bootstrap 是否也有类似结果。

用F_n 代表经验分布。用$X_1^*, \ldots, X_m^*$ 代表Bootstrap 重抽样。现在来看

$$Q_m^* = \sqrt{m}(\mu_m^* - \mu_n)/s_m^*,$$

其中

$$\mu_m^* = \frac{1}{m}\sum_{i=1}^{m} X_i^*, s_m^* = \frac{1}{m}\sum_{i=1}^{m}(X_i^* - \mu_m^*)^2.$$

定理10. 假设$X_1, X_2, \ldots$ *i.i.d.* $var(X_i) = \sigma^2 < \infty$. *Given* $(X_1, \ldots, X_n)$, 当n 和m 都趋于无穷时,

a. Conditional distribution $\sqrt{m}(\mu_m^* - \mu_n) \Rightarrow N(0, \sigma^2)$.

b. $s_m^* \to \sigma$ *in Conditional Probability: that is,* $\forall \epsilon > 0$,

$$P\left[|s_m^* - \sigma| > \epsilon | X_1, \ldots, X_n\right] \to 0, a.e.$$

8.5.2 Bootstrapping Regression Models

考虑线性回归模型：

$$Y = X\beta + \epsilon,$$

其中$Y \in \mathbb{R}^n, X \in \mathbb{R}^{n\times p}, \beta \in \mathbb{R}^p$ and $\epsilon \in \mathbb{R}^n$. 这里，$X$ 是满秩矩阵。ϵ 是随机误差项。

β 的估计可由最小二乘法得到：

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

这里要处理的问题是$\hat{\beta}$ 与β 差距是怎样的？Bootstrap 方法得到的结果和经典的大样本性质有怎样的关系？

基本假设：(1) X 是fixed.

(2) ϵ_i, i.i.d. from F. F unknown, mean 0，variance σ^2. σ^2 unknown.

(3) $\frac{1}{n}X^TX \to V$ which is positive definite.

经典的大样本理论告诉我们：

$$\sqrt{n}(\hat{\beta} - \beta) \Rightarrow N(0, \sigma^2 V^{-1}).$$

Bootstrap: 1. 计算$\hat{\epsilon} = Y - X\hat{\beta}$

2. 构造$\hat{F}$: put mass $\frac{1}{n}$ at $\hat{\epsilon} - mean(\hat{\epsilon})$

3. draw $\epsilon_1^*, \ldots, \epsilon_n^*$ i.i.d. from $\hat{F}$.

4. 计算$Y^* = X\hat{\beta} + \epsilon^*$

5. 计算$\hat{\beta}^*$.

定理11. 假设回归分析模型及基本假设*(1,2,3)*成立。当m 和n 都趋于无穷时

a. $\sqrt{m}(\hat{\beta}^*(m) - \hat{\beta}(n)) \to N(0, \sigma^2 V^{-1})$

b. $\hat{\sigma}_m^* \to \sigma, a.e.$

更多的例子及推导见"Some Asymptotic Theory for the Bootstrap" by Peter Bickel and David Freedman (1981).

8.5.3 Open Questions

对于线性回归模型，在$p < n$ 的条件下，David Freedman 考虑了用Bootstrap 方法给出参数估计的分布。问题是，在高维数据下，$p \gg n$ 的时候，Bootstrap 结果怎么样呢？参考文献：D. A. Freedman(1981). Bootstrapping Regression Models. Annals of Statistics, 1218-1228.

8.6 0.632 Bootstrap

目标：通过观测数据$(t_i, y_i), i = 1, \ldots, n$ 估计错误率，其中t_i 是预测变量，y_i 是响应。本节，y_i 是二值变量。

通常，人们使用CV 估计错误率。CV 可以给出一个关于错误率的近似无偏的估计。但是这种无偏性的代价就是方差较大。本节试图说明使用Bootstrap 方法可以大大地减少方差。

首先我们看一个例子：$i = 1, 2, \ldots, n = 20$. Independently draw (y_i, t_i) with

$$y_i = \begin{cases} 0, & 1/2; \\ 1, & 1/2. \end{cases} \quad t_i | y_i \sim N((y_i - 1/2, 0)^T, I).$$

用r 代表判别规则。损失定义为：

$$Q(y, r) = I_{y \neq r}.$$

对于一个新的观测(y_0, t_0), 判别损失就是

$$Q(y_0, r(t_0)).$$

True error rate 定义为：

$$Err = Err(X, F) = E_{0F}[Q(y_0, r(t_0))|(y_i, t_i), i = 1, 2, \ldots, n]$$

来衡量一个判别规则的好坏。

有时候也使用expected true error:

$$E_F[E_{0F}[Q(y_0, r(t_0))|(y_i, t_i), i = 1, 2, \ldots, n]]$$

Err 的一个自然的估计是：

$$e\bar{r}r = Err(x, \hat{F}) = E_{0\hat{F}}Q(x_0, X) = \frac{1}{n}\sum_{i=1}^{n} Q(y_i, r_x(t)).$$

通常$e\bar{r}r$ 得到的估计要偏小。这是因为训练样本X 使用了两次，一次用于fit model, 一次用于测试错误率。

CV 可以避免连续使用两次训练样本的问题。因为训练和评价模型的data不是一样的。Error rate 的LOO CV 估计是

$$\hat{Err}^{cv1} = \frac{1}{n}\sum_{i=1}^{n} Q[y_i, r_{x_{-i}}(t)] = \frac{1}{n}\sum_{i=1}^{n} Q[x_i, x_{-i}].$$

现在考虑对$Q(x_i, x_{-i})$ 做Bootstrap. 其中x_i 是fixed。这样做的好处之一是，将一个不连续的$Q(x_i, x_{-i})$ 转化成了一个光滑取值的函数。

我们使用

$$E_{\hat{F}_{-i}}[Q(x_i, x^*_{-i})]$$

来得到

$$E_F[Q(x_i, x_{-i})]$$

的Bootstrap 样本。

其中$\hat{F}_{-i}$ 是$x_1, x_2, \ldots, x_{i-1}, x_i, \ldots, x_n$ 的经验分布。

对每一i 都使用bootstrap 方法，我们就得到了Error rate 的LOO bootstrap 估计:

$$\widehat{Err}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} E_{\hat{F}_{(-i)}}[Q(x_i, x^*)].$$

这个估计是$\widehat{Err}^{cv1}$ 的一个平滑版本。此估计通过删除第i个样本点后的bootstrap 样本估计样本i 的误差。

可以用另一种方法解释$\widehat{Err}^{(1)}$. 它可以看作是Expected true error 的一个估计。Expected true error 的直接的Bootstrap 的估计是：

$$E_{\hat{F}}E_{0\hat{F}}Q(x_0^*, X^*).$$

注意到，x_0^* 在X^* 中的概率是：$1-(1-\frac{1}{n})^n = 0.632.$

但是这个估计常常低估错误率，所以改用

$$E_{\hat{F}_{(-\cdot)}}E_{0\hat{F}}Q(x_0, X)$$

Efron (1983) 提出了使用0.632 estimator：

$$\widehat{Err}^{.632} = 0.368\overline{err} + 0.632\widehat{Err}^{(1)}.$$

这是因为通常$\overline{err}$ 低估了误差，而$\widehat{Err}^{(1)}$ 往往高估了误差。所以要使用它们的平均来调整误差。

为什么要用0.632 呢？有0.632 的机会，x_0^* 和x^* 来自于相同的data set. 另外的0.368 的机会，使得$\widehat{Err}^{(1)}$ 高估了误差。这部分可以使用$\overline{Err}$ 来“纠正” 偏差。

Chapter 9

统计方法的求解

本章通过Ridge Regression 的求解过程，说明正则化方法中，怎样选择tuning parameter.

9.1 Ridge regression

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda\|\beta\|_2^2$$

求解：

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'Y.$$

当有截距项的时候：

$$\hat{\beta} = \arg\min_{\alpha,\beta} \|Y - X\beta - \alpha\|^2 + \lambda\|\beta\|_2^2$$

这个时候怎样求解呢？

可以首先将X, Y 中心化得到X_c, Y_c, 然后运行一个没有截距的ridge regression.

$$\hat{\beta}_c = \arg\min_{\beta} \|Y_c - X_c\beta\|^2 + \lambda\|\beta\|_2^2,$$

则

$$\hat{\beta} = \hat{\beta}_c, \hat{\alpha} = \bar{Y} - \bar{X}\hat{\beta}.$$

9.2 Selecting tuning parameters

在实际中，需要指定一个λ. 通常可以使用CV 来确定λ. 记首先确定一组λ: $(\lambda_1, \ldots, \lambda_m)$. 然后通过使用CV 从中间选取一个$\lambda$, 它使得prediction error 达到最小。

现在的问题是，如果有许多λ 要选择，对于每一个λ 都要做求逆运算。这样的运算不是有效

的。注意到对于所有的λ, 求逆运算过程中，eigen vector 是不变的，只是eigen value 变化了。这就为寻找更有效的计算方法提供了基础。

将X 做SVD 分解：

$$X_{n\times p} = U_{n\times k}S_{k\times k}V_{k\times p},$$

其中U 是X 的奇异值。S 和V 分别是做特征向量和右特征向量。

SVD 分解和特征分解之间的关系：

1. SVD 分解的对象是任意矩阵；

2. 特征分解的对象是对称矩阵;

3. SVD 分解得到的奇异值，都是正的。而特征分解不一定。

4. $X = USV, \quad X'X = V'SU'USV = V'S^2V, XX' = US^2U'.$

有了这些后，我们可以得到Ridge regression 的解：

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'Y = \left[V'(S^2 + \lambda I)^{-1}V\right]V'SU'Y = V'[(S^2 + \lambda I)^{-1}S]U'Y$$

注意：$VV' = I$, 即V 的各行是正交的。但是V 的各列不是正交的，$V'V \neq I$. 但是可以将V 的各行进行扩充，使得新的矩阵是一个列正交的矩阵。定义

$$\tilde{V} = \begin{bmatrix} V \\ V_0 \end{bmatrix},$$

使得

$$\tilde{V}'\tilde{V} = \begin{bmatrix} V \\ V_0 \end{bmatrix}' \begin{bmatrix} V \\ V_0 \end{bmatrix} = V'V + V_0'V_0 = I,$$

且

$$V_0V' = 0$$

即V_0 的各行和V的各行是正交的。实际上，V_0 是$X'X$ 的零特征根对应的特征向量。

有了这些准备之后，可以得到

$$(X'X + \lambda I)^{-1} = \begin{bmatrix} V \\ V_0 \end{bmatrix}' \begin{bmatrix} S^2 + \lambda I & 0 \\ 0 & \lambda I \end{bmatrix}^{-1} \begin{bmatrix} V \\ V_0 \end{bmatrix} = V'(S^2 + \lambda I)^{-1}V + \lambda V_0'V_0$$

从而

$$(X'X+\lambda I)^{-1}X' = [V'(S^2+\lambda I)^{-1}V + \lambda V_0'V_0]V'SU' = V'(S^2+\lambda I)^{-1}VV'SU'$$

9.3 the Lasso

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

如果$X'X = I$, 则上式转化成：

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\beta'\beta - Y'X\beta + \lambda\|\beta\|_1$$

求导，可得：

$$\beta_j - X_j'Y + \lambda = 0, \text{ if } \beta_j > 0$$

$$\beta_j - X_j'Y - \lambda = 0, \text{ if } \beta_j < 0$$

所以

$$\hat{\beta}_j = X_j'Y - \lambda \text{ if } X_j'Y > \lambda$$

$$\hat{\beta}_j = X_j'Y + \lambda \text{ if } X_j'Y < -\lambda$$

如果$|X'Y| < \lambda$, 那么上面两式都不可能成立，这个时候$\hat{\beta}_j = 0$

由上可知，合适的选择λ, 可以进行变量选择。

综上，

$$\hat{\beta}_j = \begin{cases} X_j'Y - \lambda & if \quad X_j'Y > \lambda \\ 0 & if \quad |X_j'Y| \le \lambda \\ X_j'Y + \lambda & if \quad X_j'Y < -\lambda \end{cases}$$

这称为soft-thresholding. 可以用下图表示：

Figure 9.1: Soft-thresholding

以上的解释，可以严格的用convex optimization 的理论得到。

Chapter 10

关于正态性的一些重要概念

在许多的统计问题中，假定样本来自正态总体。本章讨论以下如何检验样本是来自于正态总体。如果样本不是来自于正态总体，可以做怎样的变换，使得样本来自于正态总体。

10.1 正态性检验

10.1.1 χ^2 检验

本方法适用于检验任何分布。在这里，还是以正态分布为例，说明如何使用χ^2 检验法。

设$X_1, X_2, \ldots, X_n$ 来自某总体。现在我们检验H_0 : 样本来自于正态分布$N(\mu, \sigma^2)$.

具体做法：

1. 在实数轴上取m 个点: $t_1 < t_2 < \ldots < t_m$. 这些点把实数轴分成$m+1$ 段。

2. 用μ_i 表示落入每一段的样本的个数。

3. 在H_0 下，计算样本落入每个区间的概率：

$$p_i = \frac{1}{\sqrt{2\pi}\sigma} \int_{x \in I_i} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

4. 计算统计量：

$$V = \sum_{i=1}^{m+1} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{m+1} \frac{(\mu_i - np_i)^2}{np_i}$$

在原假设下，V 服从$\chi^2(m)$, 当n 充分大的时候。

注意：在实际应用中μ 和σ^2 是不知道的，所以需要先估计出来。所以自由度应该减少2, 即V 服从$\chi^2(m-2)$.

10.1.2 偏峰检验法（矩检验法）

设X 是一随即变量，称标准三阶中心矩$g_1 = \frac{E[X-E(X)]^3}{\sigma^3}$ 为X 的偏度；称标准四阶中心矩$g_2 = \frac{E[X-E(X)]^3}{\sigma^4}$ 为X 的峰度(其中σ^2 是X 的方差)。当X 为正态分布时，易知，偏度为0,峰度为3.

为检验样本$X_1, X_2, \ldots, X_n$ 是否来自一个正态分布，先计算样本的偏度和峰度的统计量：

$$G_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3},$$

$$G_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4},$$

其中$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,\ \ s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

由中心极限定理，当n 充分大的时候，G_1 和G_2 都近似正态分布，均值和方差分别是：

$$E(G_1) = 0, var(G_1) = \frac{6}{n}$$

$$E(G_2) = 3, var(G_1) = \frac{24}{n}$$

所以，如果

$$|G_1| \leq 1.96\sqrt{\frac{6}{n}}$$

或者

$$|G_2 - 3| \leq 1.96\sqrt{\frac{24}{n}}$$

有一个不成立，就拒绝原假设，认为总体不是来自于正态分布。注意，这里并没有考虑多重检验。实际上，如果考虑多重检验，该检验的水平不是0.05. 因为这样的检验，增加了犯一类错误的概率。

在对每一个检验都取水平0.05 的时候，即控制第一类错误的概率$\leq$ 0.05, 实际上的错误率是(这里假定两个检验是独立的，如果不独立的话(很多时候都是，如本例)，是一个难题)

$$P_{H_0}(H_{01} \text{ fail or } H_{02} \text{ fail }) = 1 - P_{H_0}(H_{01} \text{ suc and } H_{02} \text{ suc }) = 1 - 0.95^2 = 0.0975$$

一个保守的检验方法是Benferonni 检验。为了控制犯第一类错误的概率小于一个水平，比如0.05. 现在我们有m 个检验，则设定每一个检验的水平是$0.05/m$.

则实际的犯第一类错误的概率可以被控制：

$$P_{H_0}(\exists i, H_{0i} \text{ fail}) \leq mP_{H_0}(H_{0i} \text{ fail}) = m \times 0.05/m = 0.05.$$

10.1.3 $Q-Q$ 检验法

Quantile - Quantile 检验的基本原理：

假设样本来自正态总体$N(\mu, \sigma^2)$. 用经验分布($\hat{F}_n$)近似分布函数，有

$$P(X \leq x) = \Phi(\frac{x-\mu}{\sigma}) \approx \hat{F}_n(x),$$

从而$\frac{x-\mu}{\sigma} = \Phi^{-1}(F_n(x)) := u$, 即$x = \sigma u + \mu$.

如果把(u, x) 画在一个平面上，这些点将近似分布在一条直线上。

现在将样本$x_1, x_2, \ldots, x_n$ 排序，得到次序统计量：

$$x_{(1)}, \ldots, x_{(n)}.$$

在这些次序统计量处，易知$\hat{F}_n(x_{(i)}) = \frac{i}{n}$, 在实际应用中，通常用$\frac{i-0.5}{n}$ 代替$\frac{i}{n}$. 所以，$Q-Q$ plot 的步骤是：

1. 将数据从小到大排序
2. 计算经验分布：$p_i = \hat{F}_n(x_{(i)}) = \frac{i-0.5}{n}$(注意，如果有些$x$ 的值相同，则只取最大的i)
3. 计算Quantile $u_i = \Phi^{-1}(p_i)$
4. 画图$(u_i, x_{(i)})$
5. 计算u 和x 的相关系数

10.2 数据的变换

如果数据的正态性假设被否定。通常可以对数据变化，使得变换后的数据近似正态分布。下面我们介绍一个常见的变换– 幂变换法。

设$x_i > 0$ (如果不是，选取a, 使得$x_i + a > 0$), 令

$$Y_i = \begin{cases} (x_i^\lambda - 1)/(\lambda g_1^{\lambda-1}), & \lambda \neq 0 \\ g_1 \log x_i, & \lambda = 0. \end{cases},$$

其中$g_1 = (\prod_{i=1}^n x_i)^{1/n}$.

然后是要选择一个合适的参数λ, 使得变换后的数据具有正态性。λ 可以通过最大似然法获得。

实际上，上面的变换就是$Box-Cox$ 变换。g_1 的引入使得Jacobi 行列式为1。这是因为：$\lambda \neq 0$ 时，

$$\prod_{i=1}^{n} \frac{dY_i}{dx_i} = \prod_{i=1}^{n} \frac{\lambda x_i^{\lambda-1}}{\lambda g_1} = 1.$$

$\lambda = 0$ 时，也有

$$\prod_{i=1}^{n} \frac{dY_i}{dx_i} = \prod_{i=1}^{n} \frac{g_1}{x_i} = 1$$

这时候，假定$Y_i(\lambda)$ 服从正态分布，可以得到似然函数：

$$L(\mu, \sigma^2, \lambda) = (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2\sigma^2}(Y-\mu)'(Y-\mu)\}.$$

对固定的λ, 可以知道，μ 和σ^2 取下列值的时候，似然函数达到最大值：

$$\hat{\mu} = \bar{Y}(\lambda)$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum (Y_i - \bar{Y})^2$$

此时似然函数的最大值是

$$L = (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp(-\frac{n}{2})$$

对上式关于λ 求最大，即得到λ 参数。

$$\hat{\lambda} = \arg\max_{\lambda}(-\log\hat{\sigma}^2).$$