

在Pearson分布族总体中选取代表点的问题

费 荣 昌

(基础课部)

摘 要 本文研究如何在Pearson分布族总体中选取代表点,使它们尽可能多地保留总体的信息。证明了代表点的存在性,并给出了求代表点的算法和三个计算实例。

主题词 Pearson分布族总体;代表点;损失函数

0 引 言

如何从总体 X 选取 n 个代表点 $x_{n1}, x_{n2}, \dots, x_{nn}$,使它们尽可能多地保留总体的信息,即使损失函数

$$L(x_1, x_2, \dots, x_n) = \int_a^b \min_{1 \leq i \leq n} \left(\frac{x - x_i}{\sigma} \right)^2 \varphi(x) dx$$

在 $x_{n1}, x_{n2}, \dots, x_{nn}$ 达到最小值,其中 $\varphi(x)$ ($a < x < b$)是 X 的密度函数, σ^2 是 X 的方差。

关于这个问题,文[1]研究了正态分布的情形,文[2]研究了 Γ 分布和Weibull分布的情形。本文更一般地研究Pearson分布族代表点的选取问题。

Pearson K从某些一般性考虑出发,提出了一个微分方程

$$\varphi'(x) = \frac{x+c}{b_2x^2+b_1x+b_0} \varphi(x),$$

把满足这方程的密度函数 $\varphi(x)$ 挑选出来,构成一个分布族,后人把它称为Pearson分布族^[3]。概率统计中的许多重要分布都属于这个分布族,例如正态分布、 Γ 分布、 χ^2 分布和 F 分布等。

下面考虑Pearson分布族代表点的选取问题。

1 定 理

设 $\varphi(x)$ ($x > 0$)是密度函数,有

$$L(x_1, x_2, \dots, x_n) = \frac{1}{\sigma^2} \left[\int_0^{\frac{x_1+x_2}{2}} (x-x_1)^2 \varphi(x) dx + \int_{\frac{x_1+x_2}{2}}^{\frac{x_2+x_3}{2}} (x-x_2)^2 \varphi(x) dx \right]$$

本文1989年4月4日收到

$$+ \cdots + \int_{\frac{x_{n-1} + x_n}{2}}^{+\infty} (x - x_n)^2 \varphi(x) dx \Big].$$

为求 $x_{n1}, x_{n2}, \dots, x_{nn}$, 使 $L(x_1, x_2, \dots, x_n)$ 在 $x_{n1}, x_{n2}, \dots, x_{nn}$ 达到最小值, 可解方程组

$$\frac{\partial L}{\partial x_1} = 0, \quad \frac{\partial L}{\partial x_2} = 0, \quad \dots, \quad \frac{\partial L}{\partial x_n} = 0,$$

$$\int_0^{\frac{x_1 + x_2}{2}} (x - x_1) \varphi(x) dx = 0, \quad (1)$$

$$\int_{\frac{x_1 + x_2}{2}}^{\frac{x_2 + x_3}{2}} (x - x_2) \varphi(x) dx = 0, \quad (2)$$

.....

$$\int_{\frac{x_{n-1} + x_n}{2}}^{+\infty} (x - x_n) \varphi(x) dx = 0. \quad (n)$$

定理 1 $x_{11} = EX$.

定理 2 设 $\varphi(x) (x > 0)$ 是 Pearson 分布族的密度函数, 如果 b_0, b_1, b_2 满足下列条件之一:

$$(i) \quad -\frac{1}{3} < b_2 < 0, \quad D \geq 0, \quad b_1 + \sqrt{D} \leq 0, \quad (ii) \quad -\frac{1}{3} < b_2 < 0, \quad D < 0;$$

$$(iii) \quad b_2 = 0, \quad b_1 < 0, \quad b_0 \leq 0;$$

$$(iv) \quad b_2 = b_1 = 0, \quad b_0 < 0,$$

其中 $D = b_1^2 - 4b_0b_2$. 则对 $x_{n-1} > 0$, 方程(n)有唯一解 $x_n^* = x_n^*(x_{n-1})$, 且 $x_n^* > x_{n-1}$.

定理 3 在定理 2 的假设条件下, 对 $x_1 < x_{n-1,1}$, 方程组(1) - (n-1)有唯一解 $x_i = x_i(x_1)$ ($i = 2, 3, \dots, n$).

根据上述定理, 求方程组(1) - (n)的近似解的算法如下^{[1][2]}:

①取初始值 $x_1 = x_{n-1,1}/(n+1)$ (当 $n = 2$ 时, 由定理 1, x_{11} 已知), 令 $AL = 0$, $AR = x_{n-1,1}$, 则由定理 3 知 $x_{n1} \in (AL, AR)$.

②因 $x_1 < x_{n-1,1}$, 由定理 2 和定理 3, 对 x_1 , 方程组(1) - (n)有唯一解 $x_2, \dots, x_n, x_n^*$, 可用牛顿法依次求各方程的近似根而得.

③给定精度 ε (本文取 $\varepsilon = 0.0001$), 则可能出现下述三种情况:

(i) 如果 $|x_n^* - x_n| < \varepsilon$, 则认为 $x_1, x_2, \dots, x_n$ 就是所要求的 $x_{n1}, x_{n2}, \dots, x_{nn}$.

(ii) 如果 $x_n^* \geq x_n + \varepsilon$, 则说明初始值 x_1 太小, 令 $AL = x_1$, 再取 $x_1 = (AL + AR)/2$ 作为新的初始值, 转向②.

(iii) 如果 $x_n^* \leq x_n - \varepsilon$, 则说明初始值 x_1 太大, 令 $AR = x_1$, 再取 $x_1 = (AL + AR)/2$ 作为新的初始值, 转向②。

因迭代过程中区间 (AL, AR) 的长度每次缩小一半, 所以上述算法是收敛的。

2 定理的证明

定理1的证明是容易的, 定理3的证法见[2], 为了证明定理2, 先建立两个引理。

引理1 设 $\varphi(x)$ 是 Pearson 分布族的密度函数, 如果 b_0, b_1, b_2 满足下列条件之一:

$$(i) \quad -\frac{1}{2} < b_2 < 0; \quad (ii) \quad b_2 = 0, b_1 < 0; \quad (iii) \quad b_2 = b_1 = 0, b_0 < 0.$$

那末

$$\varphi(x) = o(x^{-2}).$$

证 因 $\varphi(x)$ 是 Pearson 分布族的密度函数, 有

$$\varphi'(x) = \frac{x+c}{b_2x^2+b_1x+b_0} \varphi(x).$$

当 $-\frac{1}{2} < b_2 < 0$ 时,

$$\varphi(x) = \begin{cases} k_1 (b_2x^2+b_1x+b_0)^{\frac{1}{2b_2}} \cdot \left(\frac{2b_2x+b_1-\sqrt{D}}{2b_2x+b_1+\sqrt{D}} \right)^{\left(C - \frac{b_1}{2b_2} \right) \frac{1}{\sqrt{D}}}, & D > 0, \\ k_2 (2b_2x+b_1)^{\frac{1}{b_2}} \cdot \exp \left\{ - \left(C - \frac{b_1}{2b_2} \right) \cdot \frac{2}{2b_2x+b_1} \right\}, & D = 0, \\ k_3 (b_2x^2+b_1x+b_0)^{\frac{1}{2b_2}} \cdot \exp \left\{ \left(C - \frac{b_1}{2b_2} \right) \frac{2}{\sqrt{-D}} \arctg \frac{2b_2x+b_1}{\sqrt{-D}} \right\}, & D < 0, \end{cases}$$

其中 $D = b_1^2 - 4b_0b_2$.

当 $b_2 = 0, b_1 < 0$ 时,

$$\varphi(x) = k_4 (b_1x+b_0) \frac{b_1c-b_0}{b_1^2} e^{-\frac{x}{b_1}}.$$

当 $b_2 = b_1 = 0, b_0 < 0$ 时,

$$\varphi(x) = k_5 \exp \left\{ \frac{x^2}{2b_0} + \frac{cx}{b_0} \right\}.$$

显然都有

$$\varphi(x) = o(x^{-2}).$$

引理2 如果 $f(t)$ 在 $[a, +\infty)$ 具有二阶连续导数, 且

$$f(a) > 0, \quad f(+\infty) = 0, \quad f'(a) < 0, \quad f'(+\infty) = 0,$$

$$f''(t) \begin{cases} > 0, & a \leq t < t_0; \\ = 0, & t = t_0; \\ < 0, & t > t_0. \end{cases}$$

那末方程 $f(t) = 0$ 在 $(a, +\infty)$ 有唯一解。

证 显然 $f'(t)$ 在 t_0 有极大值。若 $f'(t_0) < 0$, 则当 $t > t_0$ 时, 因 $f''(t) < 0$, 有 $f'(t) < f'(t_0)$, 于是 $f'(+\infty) \leq f'(t_0)$, 与 $f'(+\infty) = 0$ 矛盾。若 $f'(t_0) = 0$, 则当 $t > t_0$ 时, $f'(t) < 0$, 且 $f'(t)$ 在 $[t_0, +\infty)$ 单调减少, 于是 $f'(+\infty) < 0$, 也与 $f'(+\infty) = 0$ 矛盾, 所以 $f'(t_0) > 0$ 。又因 $f'(t)$ 在 $[a, t_0]$ 单调增加, 在 $[t_0, +\infty)$ 单调减少, 且 $f'(a) < 0$, $f'(+\infty) = 0$, 所以在 (a, t_0) 有且仅有一点 t_0' , 使

$$f'(t) \begin{cases} < 0, & a \leq t < t_0'; \\ = 0, & t = t_0'; \\ > 0, & t > t_0'. \end{cases}$$

显然 $f(t)$ 在 t_0' 有极小值。若 $f(t_0') > 0$, 则当 $t > t_0'$ 时, 因 $f'(t) > 0$, 有 $f(t) > f(t_0')$, 于是 $f(+\infty) \geq f(t_0')$, 与 $f(+\infty) = 0$ 矛盾。若 $f(t_0') = 0$, 则当 $t > t_0'$ 时, $f(t) > 0$, 且 $f(t)$ 在 $[t_0', +\infty)$ 单调增加, 于是 $f(+\infty) > 0$, 也与 $f(+\infty) = 0$ 矛盾, 所以 $f(t_0') < 0$ 。又因 $f(t)$ 在 $[a, t_0']$ 单调减少, 在 $[t_0', +\infty)$ 单调增加, 且 $f(a) > 0$, $f(+\infty) = 0$, 所以在 (a, t_0') 有且仅有一点 t_0'' , 使

$$f(t) \begin{cases} > 0, & a \leq t < t_0''; \\ = 0, & t = t_0''; \\ < 0, & t > t_0''. \end{cases}$$

即方程 $f(t) = 0$ 在 $(a, +\infty)$ 有唯一解。

现在来证明定理 2。令 $x_n = 2t - x_{n-1}$, 则方程 (n) 化为

$$F(x_{n-1}, t) \triangleq \int_t^{+\infty} (x + x_{n-1} - 2t) \varphi(x) dx = 0.$$

只需证, 对 $x_{n-1} > 0$, 方程 $F(x_{n-1}, t) = 0$ 有唯一解 $t^* = t^*(x_{n-1})$, 且 $t^* > x_{n-1}$ 。

显然

$$F(x_{n-1}, x_{n-1}) = \int_{x_{n-1}}^{+\infty} (x - x_{n-1}) \varphi(x) dx > 0.$$

由 L'Hospital 法则和引理 1,

$$\lim_{t \rightarrow +\infty} \frac{\int_t^{+\infty} \varphi(x) dx}{\frac{1}{t}} = \lim_{t \rightarrow +\infty} t^2 \varphi(t) = 0,$$

知

$$F(x_{n-1}, +\infty) = 0.$$

因

$$F_{t'}'(x_{n-1}, t) = (t - x_{n-1})\varphi(t) - 2 \int_t^{+\infty} \varphi(x) dx,$$

显然有

$$F_{t'}'(x_{n-1}, x_{n-1}) < 0, F_{t'}'(x_{n-1}, +\infty) = 0.$$

而

$$F_{t''}''(x_{n-1}, t) = (t - x_{n-1})\varphi'(t) + 3\varphi(t),$$

因

$$\varphi'(t) = \frac{t+c}{b_2 t^2 + b_1 t + b_0} \varphi(t),$$

有

$$\begin{aligned} F_{t''}''(x_{n-1}, t) &= \frac{(1+3b_2)t^2 + (3b_1+c-x_{n-1})t + 3b_0 - cx_{n-1}}{b_2 t^2 + b_1 t + b_0} \varphi(t) \\ &\triangleq \frac{g(t)}{h(t)} \varphi(t). \end{aligned}$$

当条件(i)满足时,

$$\begin{aligned} 1+3b_2 &> 0, \\ h(t) &= b_2 \left(t + \frac{b_1 - \sqrt{D}}{2b_2} \right) \left(t + \frac{b_1 + \sqrt{D}}{2b_2} \right), \end{aligned}$$

因

$$-\frac{b_1 - \sqrt{D}}{2b_2} \leq -\frac{b_1 + \sqrt{D}}{2b_2} \leq 0,$$

有

$$h(t) < 0 (t > 0).$$

当条件(ii)满足时,

$$1+3b_2 > 0, h(t) = b_2 \left[\left(t + \frac{b_1}{2b_2} \right)^2 - \frac{D}{4b_2^2} \right] < 0.$$

当条件(iii)满足时,

$$1+3b_2 > 0, h(t) = b_1 t + b_0 < 0 (t > 0).$$

当条件(iv)满足时,

$$1+3b_2 > 0, h(t) = b_0 < 0.$$

于是对二次多项式 $g(t)$ 有

$$g(x_{n-1}) = 3h(x_{n-1}) < 0, g(+\infty) = +\infty,$$

从而存在 $t_0 (> x_{n-1})$, 使

$$g(t) \begin{cases} < 0, & x_{n-1} \leq t < t_0, \\ = 0, & t = t_0, \\ > 0, & t > t_0, \end{cases}$$

即有

$$F_{tt}''(x_{n-1}, t) \begin{cases} > 0, & x_{n-1} \leq t < t_0; \\ = 0, & t = t_0; \\ < 0, & t > t_0. \end{cases}$$

由引理 2, 对 $x_{n-1} > 0$, 方程 $F(x_{n-1}, t) = 0$ 有唯一解 $t^* = t^*(x_{n-1})$, 且 $t^* > x_{n-1}$.

3 负指数分布、 χ^2 分布和F分布的代表点

对于负指数分布

$$\varphi(x) = \lambda e^{-\lambda x}, \quad x > 0, \quad \lambda > 0,$$

有

$$\varphi'(x) = -\lambda \varphi(x),$$

$$b_2 = 0, \quad b_1 = -1/\lambda, \quad b_0 = 0,$$

满足定理 2 的条件(iii).

对于 χ^2 分布

$$\varphi(x) = \frac{1}{2^{N/2} \Gamma(N/2)} x^{N/2-1} e^{-x/2}, \quad x > 0,$$

有

$$\varphi'(x) = \frac{N-2-x}{2x} \varphi(x),$$

$$b_2 = 0, \quad b_1 = -2, \quad b_0 = 0,$$

也满足定理 2 的条件(iii).

对于 F 分布

$$\varphi(x) = \frac{1}{B(N_1/2, N_2/2)} \left(\frac{N_1}{N_2} \right)^{N_1/2} x^{N_1/2-1} \left(1 + \frac{N_1}{N_2} x \right)^{-((N_1+N_2)/2)}, \quad x > 0.$$

有

$$\varphi'(x) = \frac{(N_1-2)N_2 - N_1(N_2+2)x}{2x(N_2+N_1x)} \varphi(x),$$

$$b_2 = -\frac{2}{N_2+2}, \quad b_1 = -\frac{2N_2}{N_1(N_2+2)}, \quad b_0 = 0,$$

当 $N_2 > 4$ 时满足定理 2 的条件(i).

于是都可按所给算法算出代表点, 下面给出三个例的计算结果.

负指数分布($\lambda = 1$)

n	L	x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	$\ln x$	x_{n9}	x_{n10}
1	1.0000	1.0000									
2	0.3524	0.5936	2.5936								
3	0.1797	0.4240	1.6112	3.6112							
4	0.1090	0.3301	1.1780	2.3652	4.3651						
5	0.0731	0.2703	0.9305	1.7784	2.9656	4.9656					
6	0.0524	0.2290	0.7697	1.4298	2.2777	3.4650	5.4650				
7	0.0394	0.1986	0.6565	1.1972	1.8574	2.7053	3.8926	5.8926			
8	0.0307	0.1753	0.5725	1.0305	1.5712	2.2313	3.0792	4.2665	6.2665		
9	0.0246	0.1570	0.5077	0.9048	1.3628	1.9035	2.5636	3.4115	4.5987	6.5986	
10	0.0202	0.1421	0.4560	0.8067	1.2039	1.6618	2.2025	2.8627	3.7106	4.8979	6.8979

 χ^2 分布($N = 4$)

n	L	x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}	x_{n9}	x_{n10}
1	1.0000	4.0000									
2	0.3565	2.6224	7.7366								
3	0.1836	2.0104	5.2100	10.0410							
4	0.1120	1.6538	4.0360	7.0188	11.7224						
5	0.0755	1.4168	3.3371	5.5372	8.4208	13.0490					
6	0.0544	1.2464	2.8670	4.6280	6.7436	9.5684	14.1458				
7	0.0410	1.1172	2.5263	4.0042	5.6904	7.7554	10.5403	15.0804			
8	0.0321	1.0155	2.2667	3.5459	4.9561	6.5973	8.6281	11.3840	15.8955		
9	0.0258	0.9330	2.0614	3.1931	4.4101	5.7793	7.3898	9.3956	12.1292	16.6178	
10	0.0211	0.8646	1.8946	2.9120	3.9859	5.1651	6.5062	8.0943	10.0808	12.7965	17.2662

F分布($N_1 = 10, N_2 = 10$)

n	L	x_{n1}	x_{n2}	x_{n3}	x_{n4}	x_{n5}	x_{n6}	x_{n7}	x_{n8}	x_{n9}	x_{n10}
1	1.0000	1.2500									
2	0.4351	0.9391	2.9533								
3	0.2468	0.7814	1.9652	4.5516							
4	0.1596	0.6822	1.5373	2.9294	6.1120						
5	0.1120	0.6125	1.2903	2.2401	3.8699	7.6537					
6	0.0830	0.5604	1.1267	1.8493	2.9224	4.7983	9.1848				
7	0.0640	0.5194	1.0091	1.5943	2.3889	3.5942	5.7195	10.7085			
8	0.0508	0.4863	0.9199	1.4134	2.0433	2.9186	4.2597	6.6360	12.2277		
9	0.0414	0.4587	0.8494	1.2776	1.7995	2.4826	3.4422	4.9210	7.5493	13.7434	
10	0.0344	0.4354	0.7921	1.1714	1.6177	2.1762	2.9159	3.9619	5.5795	8.4602	15.2561

参 考 文 献

- 1 方开泰, 贺曙东. 在正态总体中如何选择给定数目的代表点的问题. 应用数学学报, 1984; 3
- 2 张立民. 如何在母体中选取给定数目的代表点问题. 数理统计与应用概率, 1987; 2
- 3 林少宫. 基础概率与数理统计. 人民教育出版社, 1963

The Problem of Selecting Representative Points From Pearson Distributions Population

Fei Rongchang

(Dep. of Fundamental Course)

Abstract The problem of selecting representative points from Pearson Distributions Population with as much informations of the population as possible arise. These representative points were proved to be existed. A iterative procedure was given to find these points. Also three numerical examples were given to illustrate this procedure.

Subjectwords Pearson Distributions Population; Representative points; Lose function