应用概率统计　第二十四卷
第一期　2008年2月

Chinese Journal of Applied Probability
and Statistics Vol.24 No.1 Feb. 2008

关于经验分布的最优选择问题

包　文　清

(浙江师范大学数理与信息工程学院, 金华, 321004)

摘　要

本文运用统计决策知识探索了经验分布的最优选择问题. 作者借鉴风险函数的思想, 在平方损失的意义下引进平均平方距离标准, 并推导出该标准下最优新经验分布函数. 继而采用另一源于Minimax思想的最大一次距离标准, 在连续总体下对五种经验分布函数加以模拟比较分析, 得出新经验分布函数仍一致占优.

关键词: 次序统计量, 经验分布, 最大一次距离, 平均平方距离.

学科分类号: O212.2.

§1. 问题的提出

设$X_1, X_2, \cdots, X_n$是来自某总体的一个样本. 该样本的第i个次序统计量, 记为$X_{(i)}$, 它是如下的样本函数, 每当该样本得到一组观测值$x_1, \cdots, x_n$时, 将它们从小到大排列为$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, 其中第$i$个值$x_{(i)}$就是$X_{(i)}$的观察值. 我们称$(X_{(1)}, \cdots, X_{(n)})$为该样本的次序统计量. 通常定义经验分布$F_n(x)$为:

$$F_n(x) = \begin{cases} 0 & \text{当 } x \le x_{(1)}; \\ \dfrac{k}{n} & \text{当 } x_{(k)} < x \le x_{(k+1)},\ k = 1, 2, \cdots, n-1; \\ 1 & \text{当 } x_{(n)} < x. \end{cases} \tag{1.1}$$

显然, $F_n(x)$是一非减左连续函数, 且满足$F_n(-\infty) = 0$和$F_n(+\infty) = 1$. 然而, 中华人民共和国国家标准ISO5479: 1997《数据的统计处理和解释—正态性检验 Statistical Interpretation of Data—Normality Tests》中推荐使用另一个经验分布函数$G_n(x)$来取代经验分布$F_n(x)$用以作正态概率纸, 该分布函数为

$$G_n(x) = \begin{cases} 0 & \text{当 } x \le x_{(1)}; \\ \dfrac{k - 0.375}{n + 0.25} & \text{当 } x_{(k)} < x \le x_{(k+1)},\ k = 1, 2, \cdots, n-1; \\ \dfrac{n - 0.375}{n + 0.25} & \text{当 } x_{(n)} < x. \end{cases} \tag{1.2}$$

本文2004年2月18日收到, 2004年7月9日收到修改稿.

我们自然会提出这样的问题: 这两种经验分布到底哪一个更好地近似于正态分布函数? 更进一步我们可以考虑一般的函数形式:

$$H_n(x)=\begin{cases} a_0 & \text{当 } x\le x_{(1)};\\ a_k & \text{当 } x_{(k)}<x\le x_{(k+1)},\ k=1,2,\cdots,n-1\\ a_n & \text{当 } x>x_{(n)}. \end{cases} \tag{1.3}$$

不难看出: 前两种经验分布所对应的a_k分别为k/n和$(k-0.375)/(n+0.25)$ $(a_0=0)$. 另外, 在概率图研究[2]中，a_k还有多种选择, 其中较常见的两种是$a_k=(k-0.5)/n$ $(a_0=0)$和$a_k=k/(n+1)$, 这四种a_k共同点: 在n趋于∞时a_k和k/n同阶无穷小量, 这显然是经验分布函数的基本要求. 然而这些a_k的选择有何差异? 有无更优的选择? 本文第二节先提出两个比较经验分布的好坏标准, 第三节在一个标准下推导出最优经验分布函数, 第四节我们采用另一标准对五种经验分布进行模拟比较.

§2. 最大一次距离与平均平方距离

当然, a_k, $k=0,1,2,\cdots,n$选择的优劣取决于经验分布$H_n(x)$与总体分布$F(x)$的近似程度. 一种容易想到的方法: 计算$d_n(H)=\sup|H_n(x)-F(x)|$, 根据其值的大小衡量优劣, 值越小愈佳. 这是因为$H_n(x)$取$F_n(x)$时, $d_n(H)$就是柯尔莫哥洛夫统计量D_n, 根据格里汶科定理有$D_n\xrightarrow{\text{a.s.}}0$; 对于一般的$H_n(x)$有如下结论:

定理 2.1 对(1.1), (1.3)给出的$F_n(x)$, $H_n(x)$, 若$\sup|H_n(x)-F_n(x)|\to 0$, 则$d_n(H)\xrightarrow{\text{P}}0$.

证明:

$$\begin{aligned} &|F(x)-H_n(x)|\le|F(x)-F_n(x)|+|F_n(x)-H_n(x)|,\\ &\sup|F(x)-H_n(x)|\le\sup|F(x)-F_n(x)|+\sup|F_n(x)-H_n(x)|,\\ &d_n(H)\le D_n+\sup|H_n(x)-F_n(x)|. \end{aligned}$$

根据格里汶科定理即可证得. □

定义 2.1 设$x_1,\cdots,x_n$是来自总体分布为$F(x)$的样本, $H_n(x)$由(1.3)式给出, 称$d_n(H)=\sup|H_n(x)-F(x)|$为$H_n(x)$与$F(x)$的最大一次距离.

我们很容易证明本文中提到的五种经验分布(第五种将在下节给出)均满足定理2.1的条件. 使用最大一次距离或期望来度量接近程度是直观的, 且有意义. 但也有两个问题: 其一, 该距离只考虑了极端点的取值, 而没从整体上对接近程度进行考察. 这类似于决策问题中的Minimax思想, 是偏保守的; 其二, 它的计算是困难的, 没有显式解, 很难由它导出进一步的结论.

鉴于此, 我们借鉴决策论中的风险函数的思想, 给出如下距离的定义:

定义 2.2 记号意义与定义2.2相同, 称$L(H_n,F)=\int(H_n-F)^2\mathrm{d}F$为经验分布函数$H_n(x)$关于$F(x)$的平方距离.

由定义2.2给出的距离也是直观的, 易于理解的. 但是$L(H_n,F)$依赖于样本的取值, 它是一个随机变量. 为此我们引出如下平均平方距离.

定义 2.3 在定义2.2的记号下, 称$R(H_n)$为$H_n(x)$关于$F(x)$的平均平方距离. 即

$$R(H_n)=\mathsf{E}[L(H_n,F)]=\int\cdots\int L(H_n,F)\mathrm{d}P_n(x_1,\cdots,x_n).$$

其中, $P_n(x_1,\cdots,x_n)$为$(X_{(1)},\cdots,X_{(n)})$的联合分布函数.

不难看出: 定义2.3给出的平均平方距离$R(H_n)$更全面地度量了$H_n(x)$与$F(x)$的差距. 它的值是与$F(x)$无关的(见下一节), 也就是说它仅仅是$a_0,a_1,\cdots,a_n$的函数. 从而关于它的最优化问题可以讨论, 我们将在下一节给出$a_0,a_1,\cdots,a_n$的最优选择.

§3. 平均平方距离下最优经验分布函数

在导出我们的结论之前, 先不加证明引入如下结论:

引理 3.1 设Y服从$(0,1)$上的均匀分布, $Y_{(i)}$为样本量n的样本的第i个次序统计量, 那么,

$$\mathsf{E}(Y_{(i)})=\frac{i}{n+1},\qquad \mathsf{E}[(Y_{(i)})^2]=\frac{i(i+1)}{(n+1)(n+2)}.$$

引理 3.2 设连续的随机变量X的分布函数为$F(x)$, 那么, $F(X)$服从$(0,1)$上的均匀分布.

综合引理3.1, 3.2, 我们容易得出:

推论 3.1 设连续的随机变量X的分布函数为$F(x)$, $X_{(i)}$为样本量n的样本的第i个次序统计量, 则

$$\mathsf{E}[F(X_{(i)})]=\frac{i}{n+1},\qquad \mathsf{E}[F(X_{(i)})]^2=\frac{i(i+1)}{(n+1)(n+2)}.$$

有了这些准备知识, 我们可以给出如下一些结果. 令$X_{(0)}=-\infty$, $X_{(n+1)}=\infty$.

定理 3.1 设连续的随机变量X的分布函数为$F(x)$, 对给定的$a_i,i=0,1,\cdots,n$, $R(H_n,F)$的分布不依赖于总体分布$F(x)$.

证明:

$$\begin{aligned} L(H_n,F) &= \int [H_n(x)-F(x)]^2 \mathrm{d}F(x) \\ &= \sum_{i=0}^{n} \int_{X_{(i)}}^{X_{(i+1)}} [a_i - F(x)]^2 \mathrm{d}F \\ &= \sum_{i=0}^{n} \Big[a_i^2 F(x) - a_i F(x)^2 + \frac{1}{3} F(x)^3 \Big]_{X_{(i)}}^{X_{(i+1)}} \\ &= \sum_{i=0}^{n} \{ a_i^2 [F(X_{(i+1)}) - F(X_{(i)})] - a_i [F(X_{(i+1)})^2 - F(X_{(i)})^2] \\ &\quad + \frac{1}{3} [F(X_{(i+1)})^3 - F(X_{(i)})^3] \}. \end{aligned}$$

不难推出

$$\sum_{i=0}^{n} \frac{1}{3} [F(X_{(i+1)})^3 - F(X_{(i)})^3] = \frac{1}{3},$$

所以,

$$L(H_n,F) = \sum_{i=0}^{n} \{ a_i^2 [F(X_{(i+1)}) - F(X_{(i)})] - a_i [F(X_{(i+1)})^2 - F(X_{(i)})^2] \} + \frac{1}{3}.$$

显然, $L(H_n,F)$的值只与$F(X)$次序统计量的观察值有关, 而$F(X)$服从$(0,1)$上均匀分布, 所以, $L(H_n,F)$的值只与$(0,1)$上均匀分布的次序统计量的观察值有关, 其分布不依赖于总体分布. 证毕. □

定理 3.2 设连续的随机变量X的分布函数为$F(x)$, $a_i = (i+1)/(n+2)$, $i = 0,1,\cdots,n$对应的H_n的平均平方距离最小.

证明: 因为

$$R_n = \mathsf{E}[L(H_n,F)].$$

所以,

$$\begin{aligned} \frac{\partial R_n}{\partial a_i} &= 2\mathsf{E}\Big[\int_{X_{(i)}}^{X_{(i+1)}} (a_i - F(x)) \mathrm{d}F \Big] \\ &= 2a_i [\mathsf{E}(X_{(i+1)}) - \mathsf{E}(X_{(i)})] - [\mathsf{E}(F(X_{(i+1)})^2) - \mathsf{E}(F(X_{(i)})^2)]. \end{aligned}$$

令$\partial R_n/\partial a_i = 0$, $i = 0,1,\cdots,n$, 利用推论3.1, 有

$$\begin{aligned} a_i &= \frac{1}{2} \cdot \frac{\mathsf{E}[F(X_{(i+1)})]^2 - \mathsf{E}[F(X_{(i)})^2]}{\mathsf{E}(X_{(i+1)}) - \mathsf{E}(X_{(i)})} \\ &= \frac{1}{2} \cdot \Big\{ \Big[\frac{(i+1)(i+2)}{(n+1)(n+2)} - \frac{i(i+1)}{(n+1)(n+2)} \Big] \Big/ \frac{1}{n+1} \Big\} \\ &= \frac{i+1}{n+2}. \end{aligned}$$

另一方面,

$$\begin{aligned}\frac{\partial^2 R_n}{\partial a_i^2} &= 2\mathsf{E}\Big[\int_{X_{(i)}}^{X_{(i+1)}} f(x)\mathrm{d}x\Big]\\ &= 2\mathsf{E}[F(X_{(i+1)}) - F(X_{(i)})] \geq 0.\end{aligned}$$

因此, 当$a_i = (i+1)/(n+2)$, $i = 0, 1, \cdots, n$, R_n的值最小. □

定理3.1给出了平均平方距离下的最有优经验分布函数, 我们把它记为:

$$M_n(x) = \begin{cases} \dfrac{1}{n+2} & \text{当 } x \leq x_{(1)};\\ \dfrac{k+1}{n+2} & \text{当 } x_{(k)} < x \leq x_{(k+1)},\ k = 1, 2, \cdots, n-1;\\ \dfrac{n+1}{n+2} & \text{当 } x_{(n)} < x. \end{cases} \tag{3.1}$$

至此, 在本文中一共提到五种经验分布函数, 在五种选择中, k/n是经验所得, 很直观; $(k-0.5)/n = (1/2)\cdot[k/n + (k-1)/n]$和$k/(n+1) = \mathsf{E}[F(X_{(k)})]$是简单估计值; $(k-0.375)/(n+0.25)$来自计算比较; $(k+1)/(n+2)$为推理论证结果. 所以经验分布发展经历了从感性到理性过程. $M_n(x)$形式不失简单, 计算简易; 而且有经验分布函数的共性: 不依赖于总体分布, a_k也与k/n为同阶无穷小量.

进一步我们计算各种经验分布函数对应的平均平方距离. 利用定义2.3和推论3.1, 我们有

$$\begin{aligned}R_n &= \mathsf{E}[L(H_n, F)]\\ &= \sum_{i=0}^{n}\{a_i^2\mathsf{E}[F(X_{(i+1)}) - F(X_{(i)})] - a_i\mathsf{E}[F(X_{(i+1)})^2 - F(X_{(i)})^2]\} + \frac{1}{3}\\ &= \sum_{i=0}^{n}\Big\{a_i^2\cdot\frac{1}{n+1} - 2a_i\frac{i+1}{(n+1)(n+2)}\Big\} + \frac{1}{3}.\end{aligned}$$

当$a_i = (i+1)/(n+2)$, $i = 0, 1, \cdots, n$时,

$$\begin{aligned}R_n(M) &= \mathsf{E}[L(M_n, F)]\\ &= \sum_{i=0}^{n}\Big[\frac{(i+1)^2}{(n+1)(n+2)^2} - \frac{2(i+1)^2}{(n+1)(n+2)^2}\Big] + \frac{1}{3}\\ &= -\frac{2n+3}{6(n+2)} + \frac{1}{3}\\ &= \frac{1}{6(n+2)}.\end{aligned}$$

这表明最优经验分布函数与$F(x)$的平均平方距离最小值$1/[6(n+2)]$, 其中n为样本容量. 类似地可以计算出其它四种的R_n值, 现列于表1中, 为有直观的认识, 我们还给出了$n = 5$时R_n具体取值.

表1 五种经验分布的R_n值

序号	a_i	R_n	R_5
1	i/n	$1/(6n)$	0.0333
2	$(i-0.375)/(n+0.25), (a_0=0)$	$(2n^3+3n^2+4n)/[12(n+1)(n+2)(n+1/4)^2]$	0.0248
3	$(i-0.5)/n, (a_0=0)$	$(2n^2+9n-2)/[12n(n+1)(n+2)]$	0.0369
4	$i/(n+1)$	$(n+2)/[6(n+1)^2]$	0.0324
5	$(i+1)/(n+2)$	$1/[6(n+2)]$	0.0238

因为

$$\frac{2n^3+3n^2+4n}{12(n+1)(n+2)(n+0.5)^2}\sim\frac{1}{6n}(n\to\infty),\qquad \frac{2n^2+9n-2}{12n(n+1)(n+2)}\sim\frac{1}{6n}(n\to\infty),$$

所以, 五种分布的R_n值为等价无穷小量, 也就是说大样本时五种分布的差别不大, 很容易得到证明它们都是样本量n的递减函数, 这点也可以从图1中得到验证. 在小样本情况下, 对表1的R_5值比较, (1), (2), (3), (4)的R_5值分别比(5)相应值大39.92%, 4.20%, 55.04%, 36.13%; 同样通过计算可得当$n\geq 12$时, (2)和(5)的R_n值几乎相等; 当$n\geq 19$时, (1)和(4)的R_n值几乎相等.

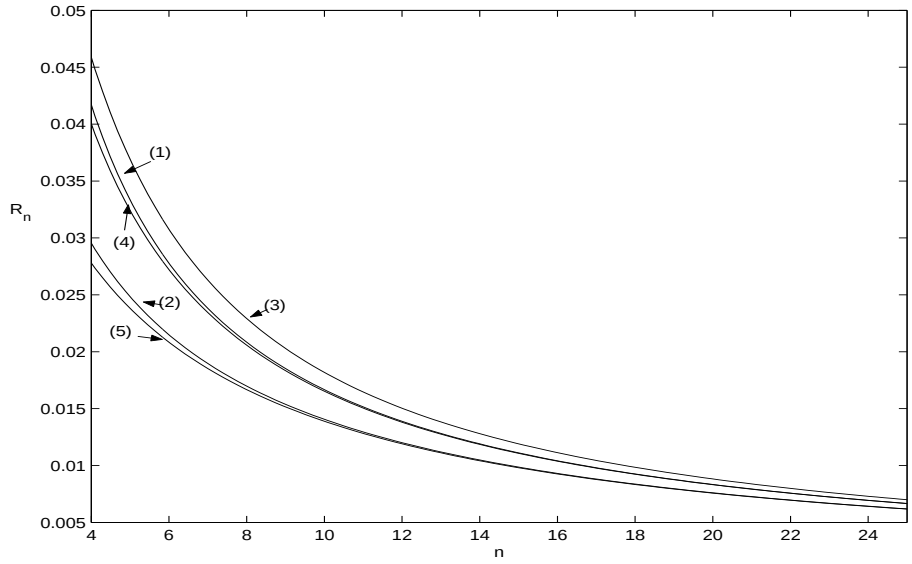

图1 五种分布的R_n函数曲线, 标号与表1序号相同

从图1可以看出: 以R_n标准, a_i选取优劣顺序为(5), (2), (4), (1), (3), 因此, 国际标准推荐经验分布函数$G_n(x)$来取代经验分布$F_n(x)$用以作正态概率纸有其合理性, 但是经验分布M_n的R_n相应值总是最小的, 其曲线也最平稳的, 在小样本时优势更明显. 总之, 经验分布M_n是最优选择.

§4. 最大一次距离下模拟比较

上节得出$M_n(x)$在平均平方距离下是最优的, 那么, 在最大一次距离d_n下, $M_n(x)$表现如何? 为了回答这问题, 我们对以上五种经验分布函数在连续总体分布条件下进行模拟比较. 设连续的随机变量X的分布函数为$F(x)$, 对给定的a_i, $i=0,1,\cdots,n$给定, 那么d_n的分布不依赖于总体分布$F(x)$ (易证). 不失一般性, 取总体分布为$N(0,1)$, 模拟步骤如下:

(1) 给定一自然数n, 通过MATLAB软件中normrnd()产生标准正态分布的n个随机数$X_1,X_2,\cdots,X_n$, 并用sort()对这n个数进行排序, 得到次序统计量$x_{(1)}\le x_{(2)}\le\cdots\le x_{(n)}$;

(2) 利用MATLAB软件中内部函数命令normcdf()求得给定点的累积函数$F(x_{(i)})=\Phi(x_{(i)})$;

(3) 计算$|a_i-F(x_{(i)})|$, $|a_{i-1}-F(x_{(i)})|$, $(i=1,\cdots,n)$的值, 得到的最大值就是d_n;

(4) 由于d_n是由随机样本$x_1,\cdots,x_n$值所决定的, 具有随机性. 为了避免随机性引起的误差, 重复进行(1) (2) (3)步N次, 得到N个d_n, 继而计算d_n的均值与方差.

以下是对不同的n进行$N=10000$次试验, 经过模拟计算得到它们的均值及方差. 现列成如下表格:

表2 不同样本量d_n的均值

序号	a_i	5	10	15	19	25
1	i/n	0.3584	0.2592	0.2132	0.1907	0.1676
2	$(i-0.375)/(n+0.25),(a_0=0)$	0.3597	0.2616	0.2156	0.1934	0.1692
3	$(i-0.5)/n,(a_0=0)$	0.3676	0.2647	0.2190	0.1946	0.1699
4	$i/(n+1)$	0.3448	0.2543	0.2095	0.1889	0.1663
5	$(i+1)/(n+2)$	0.3077	0.2373	0.2020	0.1812	0.1621

表3 不同样本量d_n的方差

序号	a_i	5	10	15	19	25
1	i/n	0.0119	0.0063	0.0044	0.0034	0.0027
2	$(i-0.375)/(n+0.25),(a_0=0)$	0.0129	0.0068	0.0045	0.0037	0.0028
3	$(i-0.5)/n,(a_0=0)$	0.0136	0.0069	0.0047	0.0037	0.0028
4	$i/(n+1)$	0.0115	0.0061	0.0042	0.0035	0.0027
5	$(i+1)/(n+2)$	0.0076	0.0048	0.0037	0.0029	0.0024

事实上, 我们模拟计算了$\mathsf{E}(d_n)$和$\mathsf{Var}(d_n)$. 我们知道: 均值越小, 则说明经验分布与总体分布的差异越小, 近似于总体分布越好. 方差越小, 则说明这种差异波动比较小. 据此从图2, 3中得到: 在正态总体分布下以d_n为标准, 经典经验分布F_n (图标为(1))的d_n均

值与方差一致小于G_n (图标为(2))的相应值, 所以, 国际推荐使用经验分布函数G_n取代经验分布$F_n(x)$用以作正态概率纸有待进一步商榷. 从d_n均值与方差看来, 最优经验分布M_n仍然是五者中最优的. 以样本量$n = 5$为例, (1), (2), (3), (4)分别与(5)比较, d_5均值分别大16.47%, 16.69%, 19.95%, 13.36%, d_5方差分别大56.58%, 69.74%, 78.95%, 51.13%, 另外, 虽然五种分布函数d_n均值与方差都随样本量增加而减少, 从减少量看来, n从15增加到19时, (3)的d_n均值减少量为0.0206, 而(5)相应值为0.0208, (5)的d_n均值减少量稍大, 除此外, (5)的d_n均值与方差随样本量增加而减少的量最小, 即最优经验分布M_n在小样本下最平稳. 总之, 在连续总体分布下以d_n为标准, 最优经验分布M_n在小样本下仍然是五者中最优选择.

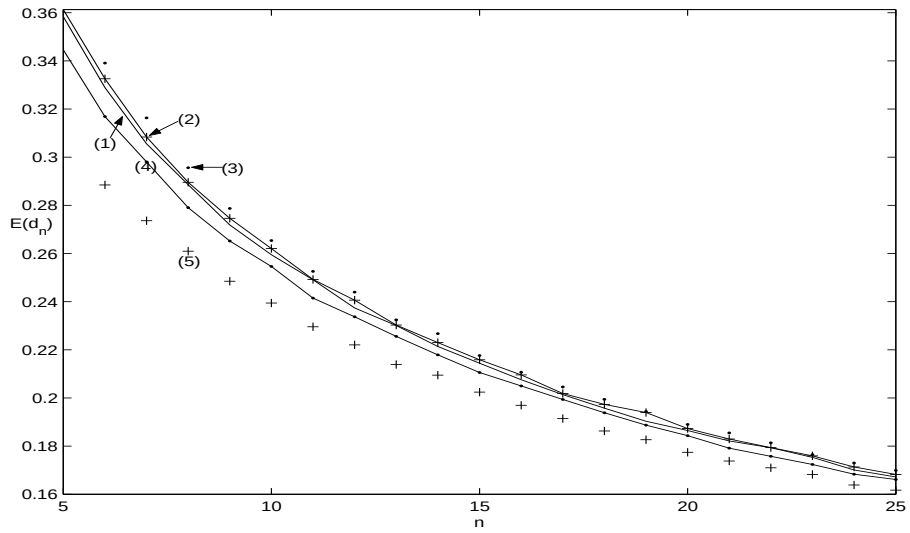

图2 五种分布的E(d_n)函数曲线, 标号与表2序号相同

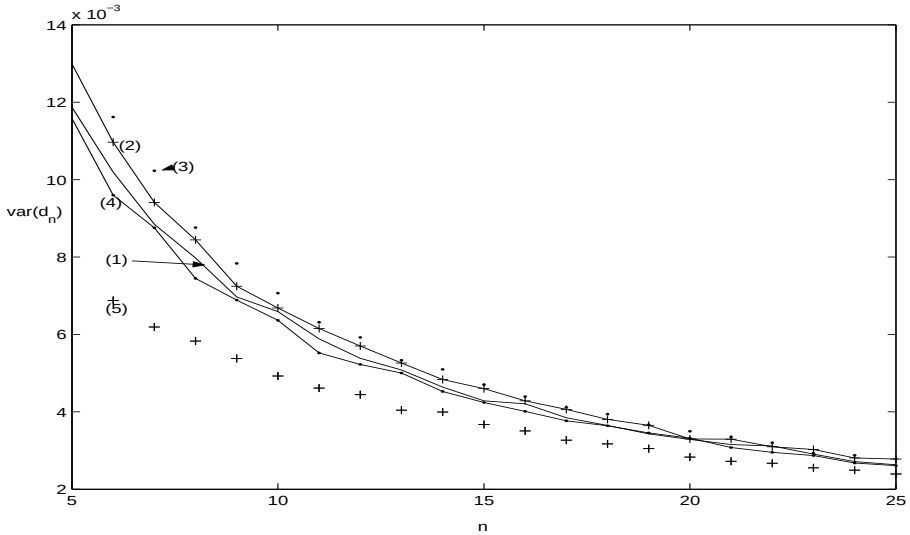

图3 五种分布的Var (d_n)函数曲线, 标号与表3序号相同。

参 考 文 献

[1] 陈希孺, 数理统计引论, 中国统计出版社, 1997.

[2] Lawless, J.F., *Statistical Models and Methods of Lifetime Data*, 中译本: 寿命数据中的统计模型与方法, 茆诗松, 濮晓龙, 刘忠译, 中国统计出版社, 1997.

[3] SAMUEL KOTZ, 吴喜之, 现代贝叶斯统计学, 中国统计出版社, 2000.

[4] 魏宗舒, 概率论与数理统计教程, 高等教育出版社, 1997.

[5] 茆诗松, 王静龙, 濮晓龙, 高等数理统计, 北京: 高等教育出版社, 德国: 斯普林格出版社, 1998.

A Note on Optimal Empirical Distribution

Bao Wenqing

(*College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua, 321004*)

To deal with optimal selection for empirical distribution, we propose "mean square distance" by introducing square loss function, and derive a "new" empirical distribution function, which is shown to be optimal. For comparisons of five empirical distribution functions, we adopt another measure in the sense of "Minimax", and simulation is used to illustrate the domination of the "new" empirical distribution function in continuous population.

Keywords: Ordered statistic, empirical distribution, maximum absolute distance, mean square distance.

AMS Subject Classification: 62G30.