卡方分布密度函数与分布函数的渐近展开

陈　刚[1],王梦婕[2]

(1. 南通职业大学基础课部,江苏 南通 226007)
(2. 加拿大百年理工学院商学院,多伦多 M1K 5E9)

[**摘要**]　通过对 χ^2 分布概率密度函数的自变量进行标准化变换,将其展开成如下形式:$\sqrt{2n}\chi^2(x;n)=\left[1+\frac{r_1(t)}{\sqrt{n}}+\frac{r_2(t)}{n}+\frac{r_3(t)}{n\sqrt{n}}+\frac{r_4(t)}{n^2}\right]\varphi(t)+o\left(\frac{1}{n^2}\right)$,其中 n 为自由度,$\varphi(t)$ 为标准正态分布的密度函数,$r_i(t)(1\leqslant i\leqslant 4)$ 均为关于 t 的多项式. 从该展开式得到 χ^2 分布密度函数的一个近似计算公式. 进一步建立 $\varphi(t)$ 的幂系数积分递推关系,得到 χ^2 分布函数的渐近展开式. 最后通过数值计算验证了这些结果在实际应用中的有效性.

[**关键词**]　χ^2 分布,概率密度函数,分布函数,渐近展开,标准化变换
[**中图分类号**]O211　[**文献标志码**]A　[**文章编号**]1001-4616(2014)03-0039-05

Asymptotic Expansions of the Probability Density Function and the Distribution Function of Chi-Square Distribution

Chen Gang[1], Wang Mengjie[2]

(1. Basic Course Department, Nantong Vocation University, Nantong 226007 China)
(2. School of Business, Centennial College, Toronto M1K 5E9, Canada)

Abstract: Through the transformation of the independent variable of χ^2 distribution probability density function, degree of freedom of which is n, the equation can be expanded as follows: $\sqrt{2n}\chi^2(x;n)=f(t;n)=\left[1+\frac{r_1(t)}{\sqrt{n}}+\frac{r_2(t)}{n}+\frac{r_3(t)}{n\sqrt{n}}+\frac{r_4(t)}{n^2}\right]\varphi(t)+o\left(\frac{1}{n^2}\right)$, here, $\varphi(t)$ is a density function of standard normal distribution; $r_i(t)$ is a 3i order polynomial of t$(1\leqslant i\leqslant 4)$. An approximate formula can be obtained from the expansion of the distribution density function. We further establish the integral recurrence relations of the power coefficients of the standard normal density function and obtain the asymptotic expansion of the distribution function of χ^2. Finally, the effectiveness of these results in practical application was verified by the numerical calculations.

Key words: χ^2 distribution, probability density function, distribution function, asymptotic expansion, standard transformation

t 分布和 χ^2 分布都是概率统计中重要的抽样分布,与最常用的正态分布有着深刻的联系. 文献[1]给出了 t 分布概率密度函数的渐近展开式,本文拟进一步对 χ^2 分布情形进行讨论.

1　预备知识

n 个独立标准正态变量的平方和,构成服从于 $\chi^2(n)$ 分布的随机变量 χ^2,其密度函数[2]为

$$\chi^2(x;n)=\begin{cases}\dfrac{1}{2^{n/2}\Gamma(n/2)}e^{-\frac{x}{2}}x^{\frac{n}{2}-1}, & x>0\\ 0 & x\leqslant 0\end{cases}$$

收稿日期:2014-01-16.
基金项目:江苏省自然科学基金(BK20141326)、江苏省高等教育教学改革研究课题重点项目(2011JSJG085).
通讯联系人:陈刚,副教授,研究方向:数理统计. E-mail:ntcggycg@163.com.

与标准正态分布、t 分布有所不同,χ^2 的密度函数不具有对称性,因此,首先需要对变量作标准化处理. 根据中心极限定理思想,将 χ^2 变量标准化,当 n 充分大时将近似地服从标准正态分布 $N(0,1)$[2]. 由于 $E(\chi^2)=n$, $\mathrm{Var}(\chi^2)=2n$,所以作变换 $t=\dfrac{x-n}{\sqrt{2n}}$,则分布函数为

$$P(\chi^2<x)=\int_{-\infty}^{x}\chi^2(x;n)\,\mathrm{d}x=\int_{-\infty}^{t}\sqrt{2n}\cdot\chi^2(\sqrt{2n}t+n;n)\,\mathrm{d}t.$$

下面将对概率密度函数 $f(t;n)=\sqrt{2n}\cdot\chi^2(\sqrt{2n}t+n;n)$ 作渐近展开,其表达式为:

$$f(t;n)=\begin{cases}\dfrac{\sqrt{2n}}{2^{n/2}\Gamma(n/2)}\mathrm{e}^{-\frac{1}{2}(\sqrt{2n}t+n)}(\sqrt{2n}t+n)^{\frac{n}{2}-1}, & t>-\sqrt{\dfrac{n}{2}}\\ 0, & t\leqslant-\sqrt{\dfrac{n}{2}}\end{cases}\tag{1}$$

引理 1(Stirling 公式)　设 $\Gamma(a)=\int_0^{+\infty}x^{a-1}\mathrm{e}^{-x}\mathrm{d}x\ (a>0)$,则

$$\ln\Gamma(a)=\ln\sqrt{2\pi}+\left(a-\frac{1}{2}\right)\ln a-a+\frac{B_1}{1\cdot2}\frac{1}{a}-\frac{B_2}{3\cdot4}\frac{1}{a^3}+\cdots+(-1)^{k-1}\frac{B_k}{(2k-1)\cdot2k}\frac{1}{a^{2k-1}}+\cdots,$$

其中常数 $B_k(k=1,2,\cdots)$ 是伯努利数[3],特别地,$B_1=\dfrac{1}{6}$, $B_2=\dfrac{1}{30}$.

该引理及其证明参见文献[4].

引理 2　对于式(1)给出的函数 $f(t;n)$,当 $t>-\sqrt{\dfrac{n}{2}}$ 时,有展开式

$$\ln f(t;n)=-\ln\sqrt{2\pi}-\frac{t^2}{2}+\frac{A}{\sqrt{n}}+\frac{B}{n}+\frac{C}{n\sqrt{n}}+\frac{D}{n^2}+o\left(\frac{1}{n^2}\right)\tag{2}$$

其中:$A=\sqrt{2}\left(\dfrac{t^3}{3}-t\right)$, $B=-\dfrac{t^4}{2}+t^2-\dfrac{1}{6}$, $C=2\sqrt{2}\left(\dfrac{t^5}{5}-\dfrac{t^3}{3}\right)$, $D=-\dfrac{2t^6}{3}+t^4$.

证明　函数 $f(t;n)$ 由 5 个因子的乘除构成,将它的对数分解为 5 项之和,即

$$\ln f(t;n)=\frac{1}{2}\ln(2n)-\frac{n}{2}\ln2-\frac{1}{2}(\sqrt{2n}t+n)-\ln\Gamma\left(\frac{n}{2}\right)+\left(\frac{n}{2}-1\right)\ln(\sqrt{2n}t+n)\tag{3}$$

将式(3)等号右边的第 4 项根据引理 1 展开,归并 n^{-2} 的高阶无穷小,得到

$$\ln\Gamma\left(\frac{n}{2}\right)=\ln\sqrt{2\pi}+\left(\frac{n}{2}-\frac{1}{2}\right)\ln\frac{n}{2}-\frac{n}{2}+\frac{1}{6n}+o\left(\frac{1}{n^2}\right)\tag{4}$$

第 5 项可按对数的幂级数展开:

$$\left(\frac{n}{2}-1\right)\ln(\sqrt{2n}t+n)=\left(\frac{n}{2}-1\right)\left[\ln n+\ln\left(1+t\sqrt{\frac{2}{n}}\right)\right]=\left(\frac{n}{2}-1\right)\ln n+\left(t\sqrt{\frac{n}{2}}-\frac{t^2}{2}+\frac{t^3}{3}\sqrt{\frac{2}{n}}-\frac{t^4}{2}\cdot\frac{1}{n}+\frac{t^5}{5}\cdot\frac{2\sqrt{2}}{n\sqrt{n}}-\frac{t^6}{3}\cdot\frac{2}{n^2}\right)+\left(-t\sqrt{\frac{2}{n}}+t^2\cdot\frac{1}{n}-\frac{t^3}{3}\cdot\frac{2\sqrt{2}}{n\sqrt{n}}+t^4\cdot\frac{1}{n^2}\right)+o\left(\frac{1}{n^2}\right)\tag{5}$$

将式(4)、(5)两式代入式(3),消去含 n 的非无穷小项,归并 n^{-2} 的高阶无穷小,即得式(2),其中的 A, B, C, D 都是关于 t 的多项式,合并同类项后即得本引理所列的结果.

2　卡方分布概率密度函数的渐近展开

定理 1　当 $t>-\sqrt{\dfrac{n}{2}}$ 时,函数(1)有渐近展开式

$$f(t;n)=r(t;n)\varphi(t)+o\left(\frac{1}{n^2}\right),\tag{6}$$

其中:$\varphi(t)$ 是标准正态分布的密度函数;$r(t;n)=1+\dfrac{r_1(t)}{\sqrt{n}}+\dfrac{r_2(t)}{n}+\dfrac{r_3(t)}{n\sqrt{n}}+\dfrac{r_4(t)}{n^2}$; $r_1(t)=\sqrt{2}\left(\dfrac{1}{3}t^3-t\right)$, $r_2(t)=\dfrac{1}{9}t^6-\dfrac{7}{6}t^4+2t^2-\dfrac{1}{6}$, $r_3(t)=\sqrt{2}\left(\dfrac{1}{81}t^9-\dfrac{5}{18}t^7+\dfrac{47}{30}t^5-\dfrac{37}{18}t^3+\dfrac{1}{6}t\right)$, $r_4(t)=\dfrac{1}{486}t^{12}-\dfrac{13}{162}t^{10}+\dfrac{341}{360}t^8-\dfrac{1031}{270}t^6+\dfrac{151}{36}t^4-\dfrac{1}{3}t^2+\dfrac{1}{72}$.

证明　在式(2)中记:$M=\frac{A}{\sqrt{n}}+\frac{B}{n}+\frac{C}{n\sqrt{n}}+\frac{D}{n^2}+o\left(\frac{1}{n^2}\right)$,则由引理2知

$$\ln f(t;n)=-\ln\sqrt{2\pi}-\frac{t^2}{2}+M=\ln\varphi(t)+M,$$

因此$f(t;n)=\varphi(t)\mathrm{e}^M=\varphi(t)\left(1+M+\frac{M^2}{2}+\frac{M^3}{6}+\frac{M^4}{24}+\cdots\right)$.

展开括号内M的乘方,归并n^{-2}的高阶无穷小,易知

$$M^2=\frac{A^2}{n}+\frac{2AB}{n\sqrt{n}}+\frac{B^2+2AC}{n^2}+o\left(\frac{1}{n^2}\right),M^3=\frac{A^3}{n\sqrt{n}}+\frac{3A^2B}{n^2}+o\left(\frac{1}{n^2}\right),M^4=\frac{A^4}{n^2}+o\left(\frac{1}{n^2}\right),$$

代入上面的展开式,合并n的同阶次项,即得

$$\sqrt{2n}\chi^2(x;n)=f(t;n)=\left[1+\frac{r_1(t)}{\sqrt{n}}+\frac{r_2(t)}{n}+\frac{r_3(t)}{n\sqrt{n}}+\frac{r_4(t)}{n^2}\right]\varphi(t)+o\left(\frac{1}{n^2}\right)$$

其中:$r_1(t)=A,r_2(t)=B+\frac{A^2}{2},r_3(t)=C+AB+\frac{A^3}{6},r_4(t)=D+\frac{B^2+2AC}{2}+\frac{A^2B}{2}+\frac{A^4}{24},x=\sqrt{2n}t+n$. 显然$r_i(t)(i=1,2,3,4)$都是关于$t$的多项式,将引理2中关于$A,B,C,D$的表达式代入,经整理化简,便可得到本定理所列的$r_i(t)(i=1,2,3,4)$的具体表达式.

3　卡方分布函数的近似计算公式

对定理1的公式(6)取极限,可得到$\chi^2(n)$分布渐近于正态分布的直接证据:

$$\lim_{n\to\infty}\sqrt{2n}\chi^2(\sqrt{2n}t+n;n)=\varphi(t).$$

但定理1给出的不仅仅是极限结果,而是更为精细的渐近展开式,具有很好的实际应用价值[5],因为从中可以得到$\chi^2(n)$分布密度函数的近似计算公式:

$$\sqrt{2n}\chi^2(x;n)\approx\left[1+\frac{r_1(t)}{\sqrt{n}}+\frac{r_2(t)}{n}+\frac{r_3(t)}{n\sqrt{n}}+\frac{r_4(t)}{n^2}\right]\varphi(t),\tag{7}$$

其中$x=\sqrt{2n}t+n$. 由此出发,进一步还可建立$\chi^2(n)$分布的分布函数的近似计算公式. 在分布函数的概率积分中作变换$x=\sqrt{2n}t+n$,可得

$$P\{\chi^2(n)<x\}=\int_{-\infty}^{x}\chi^2(x;n)\mathrm{d}x=\int_{-\infty}^{t}\sqrt{2n}\chi^2(\sqrt{2n}t+n;n)\mathrm{d}t\approx\int_{-\infty}^{t}\varphi(t)\mathrm{d}t+\frac{1}{\sqrt{n}}\int_{-\infty}^{t}r_1(t)\varphi(t)\mathrm{d}t+$$
$$\frac{1}{n}\int_{-\infty}^{t}r_2(t)\varphi(t)\mathrm{d}t+\frac{1}{n\sqrt{n}}\int_{-\infty}^{t}r_3(t)\varphi(t)\mathrm{d}t+\frac{1}{n^2}\int_{-\infty}^{t}r_4(t)\varphi(t)\mathrm{d}t.$$

为了计算这些积分,引入$\varphi(t)$的幂系数积分$I_k(t)=\int_{-\infty}^{t}t^k\varphi(t)\mathrm{d}t(0\leqslant k\leqslant 12)$. 根据定理1所示的多项式$r_i(t)(i=1,2,3,4)$的表达式,不难得到

$$\int_{-\infty}^{t}r_1(t)\varphi(t)\mathrm{d}t=\sqrt{2}\left[\frac{1}{3}I_3(t)-I_1(t)\right];$$

$$\int_{-\infty}^{t}r_2(t)\varphi(t)\mathrm{d}t=\frac{1}{9}I_6(t)-\frac{7}{6}I_4(t)+2I_2(t)-\frac{1}{6}I_0(t);$$

$$\int_{-\infty}^{t}r_3(t)\varphi(t)\mathrm{d}t=\sqrt{2}\left[\frac{1}{81}I_9(t)-\frac{5}{18}I_7(t)+\frac{47}{30}I_5(t)-\frac{37}{18}I_3(t)+\frac{1}{6}I_1(t)\right];$$

$$\int_{-\infty}^{t}r_4(t)\varphi(t)\mathrm{d}t=\frac{1}{486}I_{12}(t)-\frac{13}{162}I_{10}(t)+\frac{341}{360}I_8(t)-\frac{1031}{270}I_6(t)+\frac{151}{36}I_4(t)-\frac{1}{3}I_2(t)+\frac{1}{72}I_0(t).$$

现在计算积分$I_k(t)$. 注意到$\varphi'(t)=-t\varphi(t)$,可得$I_0(t)=\Phi(t),I_1(t)=-\varphi(t)$,其中$\Phi(t)$和$\varphi(t)$分别是标准正态变量的分布函数和密度函数(下同),

$$I_{k+1}(t)=-\int_{-\infty}^{t}t^k d\varphi(t)=-t^k\varphi(t)\Big|_{-\infty}^{t}+\int_{-\infty}^{t}kt^{k-1}\varphi(t)\mathrm{d}t=-t^k\varphi(t)+kI_{k-1}(t).$$

利用这个递推式,由$I_0(t)$和$I_1(t)$分别可递推得到$I_{2i}(t)(1\leqslant i\leqslant 6)$和$I_{2i+1}(t)(1\leqslant i\leqslant 4)$. 将它们代入上面

的积分结果,经整理化简,$\Phi(t)$均抵消,可得

$$\int_{-\infty}^{t} r_1(t)\varphi(t)\mathrm{d}t=\sqrt{2}R_1(t)\varphi(t),\ \int_{-\infty}^{t} r_2(t)\varphi(t)\mathrm{d}t=R_2(t)\varphi(t),$$

$$\int_{-\infty}^{t} r_3(t)\varphi(t)\mathrm{d}t=\sqrt{2}R_3(t)\varphi(t),\ \int_{-\infty}^{t} r_4(t)\varphi(t)\mathrm{d}t=R_4(t)\varphi(t),$$

其中:

$$R_1(t)=\frac{1}{3}-\frac{1}{3}t^2,R_2(t)=-\frac{1}{6}t+\frac{11}{18}t^3-\frac{1}{9}t^5,R_3(t)=\frac{1}{270}+\frac{23}{270}t^2-\frac{133}{270}t^4+\frac{29}{162}t^6-\frac{1}{81}t^8,$$

$$R_4(t)=\frac{1}{72}t-\frac{23}{216}t^3+\frac{883}{1080}t^5-\frac{463}{1080}t^7+\frac{14}{243}t^9-\frac{1}{486}t^{11}.$$

于是得到分布函数的近似公式:

$$P\{\chi^2(n)<x\}\approx\Phi(t)+\left[\frac{\sqrt{2}R_1(t)}{\sqrt{n}}+\frac{R_2(t)}{n}+\frac{\sqrt{2}R_3(t)}{n\sqrt{n}}+\frac{R_4(t)}{n^2}\right]\varphi(t),\tag{8}$$

其中 $x=\sqrt{2n}t+n$. 对此式取极限,可得到与中心极限定理一致的结果:

$$\lim_{n\to\infty}P\{\chi^2(n)<x\}=\Phi(t).$$

4　结论有效性的数值计算验证

对于每个取定的 x 或 t,式(7)、(8)的误差(等式左右两边之差的绝对值)分别记做 δ_n 和 Δ_n,通过对这两个误差的数值计算和分析,可以验证这两个近似计算公式的有效性.

上面的极限式表明,标准化后的随机变量 $\sqrt{2n}\chi^2$ 近似服从标准正态分布,根据 3σ 原理[2],在 $-3\leqslant t\leqslant 3$ 范围内对取定的 t,运用 Matlab 计算误差 δ_n 和 Δ_n. 限于篇幅,这里仅选取对应于 $t=0$、±1、±2、±3 时的曲线(见图 1 和图 2)作直观的描述,图形显示,误差 δ_n 和 Δ_n 都随 n 增大而减小且趋向于 0,并且当 n 大约从 30 开始,误差已控制在万分之五以内.

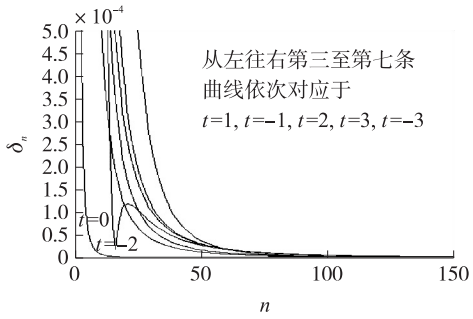

图 1　误差 δ_n 随 n 增大的变化趋势

Fig. 1　The tendency of error δ_n against n

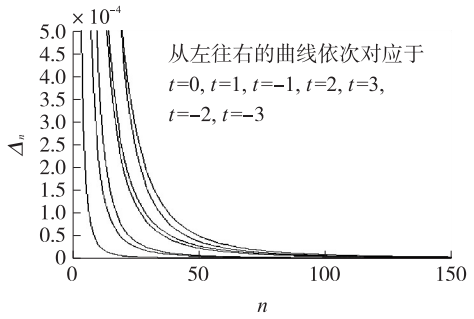

图 2　误差 Δ_n 随 n 增大的变化趋势

Fig. 2　The tendency of error Δ_n against n

由于分布函数可直接计算随机变量的概率,所以式(8)具有更高的实用价值. 下面对式(8)进行数值计算验证.

χ^2 分布的密度曲线在 $x=n-2$ 处取得峰值,概率计算要考虑包含或避开曲线的峰段,考虑峰段左右陡峭程度的不同[6],所以取 $t=\pm1$($x=n\pm\sqrt{2n}$). 式(8)左右两边的概率值,分别称为理论概率和近似概率,计算结果如表 1 所示.

从表中看出,随着 n 的增大,由近似公式得到的概率值能精确到 2 位、3 位、4 位小数,误差不断减小,近似公式的渐近速度比较快.

再来看临界值的情况. χ^2 分布的右侧临界值为 $\chi^2_\alpha(n)$,表示 $P(\chi^2(n)>\chi^2_\alpha(n))=\alpha$.

以 $n=20$,$\alpha=0.05$ 为例,理论临界值为 $\chi^2_{0.05}(20)=31.410$. 取 $x=31.410$,由 $x=\sqrt{2n}t+n$ 算得 t,代入式(8)右边并算得 $P=1-\Phi(t)-R(t,n)\varphi(t)$,然后调整 x 作一维搜索,使得 $P=0.05$,所得的 x 值为 $x=31.420\ 1$,这就是近似式(8)给出的临界值结果.

表 1　近似概率的计算结果及其误差

Table 1　The numerical result of approximate probability and its error

n	$t=1$			$t=-1$		
	理论概率	近似概率	误差 Δ_n	理论概率	近似概率	误差 Δ_n
5	0.852 5	0.853 6	0.001 14	0.128 9	0.131 9	0.002 98
10	0.847 5	0.847 7	0.000 22	0.146 8	0.147 2	0.000 42
15	0.845 6	0.845 7	0.000 08	0.151 4	0.151 5	0.000 14
20	0.844 6	0.844 7	0.000 04	0.153 4	0.153 5	0.000 07
25	0.844 0	0.844 1	0.000 02	0.154 6	0.154 6	0.000 04
30	0.843 6	0.843 6	0.000 02	0.155 4	0.155 4	0.000 02
35	0.843 3	0.843 3	0.000 01	0.155 9	0.155 9	0.000 02
40	0.843 1	0.843 1	0.000 01	0.156 3	0.156 3	0.000 01

类似地可计算其他 n 和 α 对应的近似临界值,计算结果如表 2 所示.

表 2　近似临界值的计算结果及其相对误差

Table 2　The numerical result of approximate critical value and its relative error

n	α	理论临界值	近似临界值	相对误差/%	n	α	理论临界值	近似临界值	相对误差/%
10	0.99	2.558	2.639 2	3.174	30	0.99	14.954	14.978 0	0.160
	0.95	3.940	3.931 8	0.208		0.95	18.493	18.492 6	0.002
	0.90	4.865	4.856 3	0.179		0.90	20.599	20.598 2	0.004
	0.10	15.987	15.977 6	0.059		0.10	40.256	40.255 0	0.002
	0.05	18.307	18.354 3	0.258		0.05	43.773	43.776 9	0.009
	0.01	23.209	22.964 9	1.052		0.01	50.892	50.881 2	0.021
20	0.99	8.260	8.305 9	0.556	40	0.99	22.164	22.179 1	0.068
	0.95	10.851	10.849 9	0.010		0.95	26.509	26.509 4	0.002
	0.90	12.443	12.440 2	0.023		0.90	29.051	29.049 9	0.004
	0.10	28.412	28.409 6	0.008		0.10	51.805	51.804 5	0.001
	0.05	31.410	31.420 1	0.032		0.05	55.758	55.760 5	0.004
	0.01	37.566	37.525 5	0.108		0.01	63.691	63.687 2	0.006

从表中看出,由近似公式得到的临界值,相对误差都很小,$n\geqslant 10$ 时均可用于实际. 尤其是 $n\geqslant 20$ 时,相对误差均小于千分之六,$n\geqslant 30$ 时更好,与理论数据基本无差别. 对于常用的显著性水平 $\alpha(0.05,0.95,0.10,0.90)$,$n\geqslant 10$ 时得到的临界值非常精确,用于实际计算可保无虞.

以上数值计算的检验体现出近似式(8)的可靠性,而且自由度 n 并不需要很大.

式(7)和(8)的近似计算可以运用 Matlab、SPSS 或者 r 等数学软件,也可在非常普及的 excel 环境中实施,这就更加体现了其应用的广泛性.

[参考文献]

[1] 丁邦俊. t-分布密度函数的渐近展开[J]. 数理统计与应用概率,1998(4):307-310.

[2] 茆诗松,程依明,濮晓龙. 概率论与数理统计教程[M]. 2 版. 北京:高等教育出版社,2011:283.

[3] 菲赫金哥尔茨 F M. 微积分学教程(第二卷第二分册)[M]. 北京大学高等数学教研室,译. 北京:人民教育出版社,1954:447-451.

[4] 菲赫金哥尔茨 F M. 微积分学教程(第二卷第三分册)[M]. 徐献瑜,冷生明,梁文骐,等译. 北京:人民教育出版社,1954:704-719.

[5] 陈刚,黄超,林金官. 具有高斯过程误差的函数型线性模型的统计诊断[J]. 应用数学与计算数学学报,2014(1):117-126.

[6] 朱建国. 一类次序统计量的数学期望[J]. 南通职业大学学报,2010(2):68-69.

[责任编辑:陆炳新]