第5卷 第4期
1985年10月

数学研究与评论
JOURNAL OF MATHEMATICAL RESEARCH AND EXPOSITION

Vol.5 No.4
Oct. 1985

刀切不完全U统计量*

施锡铨

(武汉水利电力学院)

设θ为未知参数，$X_1, X_2, \cdots, X_N$为来自F_θ的N个独立同分布的观察值。$f(x_1, \cdots, x_m)$ $(m<N)$为m元对称可测函数，$Ef(X_1, \cdots, X_m)=\theta$。定义$U$统计量为

$$U_0=\binom{N}{M}^{-1}\sum_{1\le\alpha_1<\alpha_2<\cdots<\alpha_m\le N} f(X_{\alpha_1}, \cdots, X_{\alpha_m}). \tag{1}$$

如所周知，对于U统计量及它的正则函数，文[1]研究表明，刀切法是能获得近似置信区间的有效的非参数方法。

为了减少实用中许多不便之处，G. Blom[2] 于 1976 年提出：

定义1 从所有$f(X_{\alpha_1}, \cdots, X_{\alpha_m})$中适当选择$r$个对称核$f_i$，称统计量

$$U=\sum_{i=1}^{r} f_i/r \qquad r<\binom{N}{M} \tag{2}$$

为不完全U统计量(incomplete U-statistic)。

不完全U统计量对于N个观察值而言并非对称函数，刀切法对它是否有效是个有意义的问题。本文仅对 Brown[3] 等人所研究的一类特殊的不完全U统计量用刀切法的思想进行处理，得到与刀切U统计量同样的效果。

我们采用 Brown 的记号，$f(\cdot,\cdot)$为对称可测函数，

$$S_N=\sum_{C_k} f(X_i, X_j), \qquad \theta=Ef(X_1, X_2),$$

$$\sigma^2=\mathrm{Var} f(X_1, X_2)>0,\ \rho\sigma^2=\mathrm{Cov}\{f(X_1, X_2),\ f(X_1, X_3)\}, \tag{3}$$

其中$C_k=\{(i,j):\ 1\le i<j\le N$，每一$\le N$的整数恰出现$2k$次($k$固定)$\}$故$C_k$恰有$NK$项。$\hat{\theta}_N \triangleq \frac{1}{Nk}S_N$为不完全$U$统计量，易知

$$\mathrm{Var}\ S_N=Nk\sigma^2[1+2(2k-1)\rho].$$

定理1 若$Ef^2(X_i, X_j)<\infty$，那末

$$(Nk)^{\frac{1}{2}}(\hat{\theta}_N-\theta)\xrightarrow{\mathcal{L}} N(0, \sigma^2[1+2(2k-1)\rho]),$$

*1982年9月28日收到。

其中$\mathscr{L}$表示以分布收敛。证明详见[3]。

应用刀切法的思想，令$\hat{\theta}_{N-1}^{i}$为这样的估计量：它是基于$\hat{\theta}_N$中所含"对"(i,j)除去含x_i的那些"对"后所剩下来其余"对"的全体的不完全U统计量，即

$$\hat{\theta}_{N-1}^{i}=\frac{1}{(N-2)k}\sum_{(C_k-\text{含}x_i\text{的}k\text{个}(i,j))}f(X_i,X_j) \tag{4}$$

以$g(\hat{\theta}_N)$估计θ的正则函数$g(\theta)$(即在θ附近g'有界)，则定义新的含义下的"刀切估计量"为

$$Jg(\hat{\theta}_N)=Ng(\hat{\theta}_N)-\frac{N-1}{N}\sum_{i=1}^{N}g(\hat{\theta}_{N-1}^{i}). \tag{5}$$

注意到若$\hat{\theta}_N$为U统计量，则两种刀切的意义恰为一致。

不失一般性，我们假定$\theta=0$，并设$I=(-\Delta,\Delta)(\Delta>0)$为$\theta=0$的邻域，我们有下述结论：

引理1 若$\sigma^2>0$，σ^2,ρ有限，则成立

$$p\{\hat{\theta}_N,\hat{\theta}_{N-1}^{i}(i=1,2,\cdots,N)\in I\text{同时发生}\}\to 1\quad(N\to\infty).$$

定理2 若$\sigma^2>0,\sigma^2\rho$有限，g为正则函数，那末

$$\sqrt{Nk}\{Jg(\hat{\theta}_N)-g(0)\}\to_{\mathscr{L}}N\{0,\sigma^2[1+2(2k-1)\rho][g'(0)]^2\},$$

引理1与定理2的证明方法与[1]中刀切U统计量有关证明类同。只是在证明中应注意到$\hat{\theta}_{N-1}^{i}$与通常刀切法中相应量之间的不同性以及下述关系式即可：

$$\hat{\theta}_{N-1}^{i}=\frac{Nk}{(N-2)k}\left\{\hat{\theta}_N-\frac{1}{Nk}[f(X_i,X_{\alpha_1})+\cdots+f(X_i,X_{\alpha_{2k}})]\right\}.$$

我们知道，在通常刀切法作用下，如果Tukey[4]猜测成立，那末可以利用虚拟值与一阶刀切的关系式来估计刀切方差。而这里由于刀切形式的改变，原来的那种关系式就不一定随机收敛于方差，我们有下面的引理：

引理2 如果$E|f(X_i,X_j)|^4<\infty$，那末

$$(N-1)k\sum_{i=1}^{N}(\hat{\theta}_{N-1}^{i}-\hat{\theta}_N)^2\to_{P}\sigma^2[2+2(2k-1)\rho]\quad(N\to\infty). \tag{6}$$

证明
$$(N-1)k\sum_{i=1}^{N}(\hat{\theta}_{N-1}^{i}-\hat{\theta}_N)^2=(N-1)k\left\{\frac{2\sum_{C_k}f^2(X_i,X_j)}{(N-2)NK^2}\right.$$

$$\left.+\frac{2(N-4)\sum_{D_k}f(X_i,X_j)f(X_i,X_k)}{N(N-2)^2k^2}-\frac{8}{N(N-2)^2k^2}\sum_{E_k}f(X_i,X_j)f(X_k,X_l)\right\}$$

$$=\frac{N-1}{N-2}\cdot 2R_N^{(1)}+\frac{(N-1)(N-4)}{(N-2)^2}\cdot 2(2k-1)R_N^{(2)}-\frac{4(N-1)(Nk-4k+1)}{(N-2)^2}R_N^{(3)}, \tag{7}$$

其中C_k如(3)式所定义，为Nk项。D_k表示$\{(i,j),(i,k)\}$的集合：i,j,k互不相同，$(i,j),(i,k)\in C_k$，共有$N[k(2k-1)]$项。E_k表示$\{(i,j),(k,l)\}$的集合，i,j,k,l互不相同，$(i,j),(k,l)\in C_k$，共有$\frac{1}{2}Nk(Nk-4k+1)$项。

显见$R_N^{(1)}$，$R_N^{(2)}$，$R_N^{(3)}$为不完全U统计量，经计算得

$$ER_N^{(1)}=\sigma^2,\quad ER_N^{(2)}=\rho\sigma^2,\quad ER_N^{(3)}=0.$$

且 $$\mathrm{Var}\ R_N^{(1)}\leqslant\frac{1}{(Nk)^2}\cdot 2Nk\cdot Ef^4(X_1,X_2)=O\left(\frac{1}{Nk}\right),$$

故(7)式右边第一项 $\frac{N-1}{N-2}\cdot 2R_N^{(1)}\xrightarrow{P}2\sigma^2$. $(N\to\infty)$.

同理， $2(2k-1)R_N^{(2)}\cdot\frac{(N-1)(N-4)}{(N-2)^2}\xrightarrow{P}2(2k-1)\rho\sigma^2$, $(N\to\infty)$.

$$\frac{4(N-1)(Nk-4k+1)}{(N-2)^2}R_N^{(3)}\xrightarrow{P}0,\ (N\to\infty).$$

于是 $(N-1)k\sum_{i=1}^{N}(\hat\theta_{N-1}^{i}-\hat\theta_N)\xrightarrow{P}\sigma^2[2+2(2k-1)\rho]$.

令 $$s_g^2=(N-1)k\sum_{i=1}^{N}\left[g(\hat\theta_{N-1}^{i})-\frac{1}{N}\sum_{i=1}^{N}g(\hat\theta_{N-1}^{i})\right]^2,\qquad(8)$$

利用文[1]中定理6的证明方法，得到下述结论：

定理3　若$E|f|^4<\infty$，$Ef(X_1,X_2)=0$，g为实轴函数，在$\theta=0$的邻域内具有连续一阶导数，那末当$N\to\infty$时，

$$s_g^2\xrightarrow{P}[g'(0)]^2\sigma^2[2+2(2k-1)\rho].$$

现在我们还需估计$[g'(0)]^2\sigma^2$，取F_N为(i,j)的集合，恰含有$\left[\frac{N}{2}\right]$个对，其中所有i，j均不相同．例$(1,2)$，$(3,4)$，…，作$U_{\left[\frac{N}{2}\right]}=\frac{1}{\left[\frac{N}{2}\right]}\sum_{F_N}f(X_i,X_j)$．$U_{\left[\frac{N}{2}\right]}$事实上可视作“二维空间”中次数为1的$U$统计量，经过讨论可以得知，通常的刀切$U$统计量结果适用于$U_{\left[\frac{N}{2}\right]}$，因此利用刀切$U$统计量的已知结果，当$N\to\infty$时，有

$$s_T^2=\frac{1}{\left[\frac{N}{2}\right]-1}\sum_{i=1}^{\left[\frac{N}{2}\right]}\left[g_i^*-Jg(U_{\left[\frac{N}{2}\right]})\right]^2\xrightarrow{P}[g'(0)]^2\sigma^2,\qquad(9)$$

其中 $$g_i^*=\left[\frac{N}{2}\right]g\left(U_{\left[\frac{N}{2}\right]}\right)-\left(\left[\frac{N}{2}\right]-1\right)g\left(U^{i}_{\left[\frac{N}{2}\right]-1}\right).$$

综上所述，我们可以得到：

定理4　若g为正则函数，$\sigma^2>0$，σ^2，ρ有限，那末

$$\frac{\sqrt{Nk}\{Jg(\hat\theta_N)-g(\theta)\}}{(s_g^2-s_T^2)^{1/2}}\xrightarrow{\mathcal{L}}N(0,1).$$

这样，与刀切U统计量的结果一样，我们也能求得$g(\theta)$的近似置信区间．

参　考　文　献

[1] Arvesen, J., Jackknifing U-statistic, *Ann. Math. Satist.* 40., (1969), 2076—2100.

[2] Blom, G., Some properties of incomplete U-statietics, *Biometrika*, 63 (1976) No3. 573—580.

[3] Brown, B. M. and Kildea, D. G., Reduced U-statietice and the Hodges-Lehmann estimator, *Ann Math. Statist.*, 6 (1978), No. 4, 828—835.

[4] Tukey, J. W., Bias and confidence in not quite large samples (abstrsact), *Ann Math. statist.*, 29 (1958), 614.

Jackknifing Incomplete U-statistics

Shi Xi-quan

(The Wuhan Institute of Hydraulic and Electric Engineering)

In this paper, we propose jackknife only for a particular class of non-symmetric statistics, which is incomplete U-statistics of significance in practice. We find that if we deal with incomplele U-statistics using a slightly modified jackknife, then Tukey's conjecture will be true, and the consistent estimator of asymptotic variance is given. Thus we obtain an approximate confidence interval for reguluar function $g(\theta)$.