应用概率统计　第二十四卷
第四期　2008年8月

Chinese Journal of Applied Probability
and Statistics Vol.24 No.4 Aug. 2008

离散型最小和最大次序统计量相关性研究 *

梁冯珍　　史道济

(天津大学理学院, 天津, 300072)

摘　要

本文研究离散型随机变量之间的相关性度量, 讨论了最小次序统计量和最大次序统计量的渐近独立性, 给出了计算最小次序统计量和最大次序统计量的Kendall和Spearman秩相关系数的公式.

关键词: 相关性度量, 次序统计量, Kendall秩相关系数, Spearman秩相关函数.

学科分类号: O212.4.

§1. 引　　言

相关性和相关系数是统计、金融、保险理论中一个非常重要的概念. 在统计学中, 常常用两个变量之间的线性相关系数来度量两个变量之间的相关性, 但这种度量具有一定的局限性. 首先, 它只能度量两个变量之间的线性相关关系, 而不能度量两个变量之间的其他相关关系; 其次, 线性相关系数不仅依赖于变量的边缘分布, 而且依赖于联合分布; 第三, 在单调变换下, 线性相关系数的值常常会改变. 为了克服上述缺陷, 我们希望建立一种相关性度量, 能够满足理想的相关性度量应该具备的性质[1]. 本文采用文献[2]、[3]中介绍的度量相关性的两种方法, 来讨论离散型最小次序统计量和最大次序统计量之间的相关性. 对于连续型随机变量, 文献[4]已经讨论了它们之间的相关性.

§2. Copula及相关性度量的概念

我们首先给出Copula的定义.

定义 2.1　如果二元函数C满足条件

(1) 对任意的$(u,v)\in I^2=[0,1]^2$, 都有$0\le C(u,v)\le 1$;

(2) 对任意的$(u,v)\in I^2$, 都有$C(u,0)=C(0,v)=0$, $C(u,1)=u$, $C(1,v)=v$;

(3) 对任意的$u_1,u_2,v_1,v_2\in I$, 且$u_1\le u_2$, $v_1\le v_2$, 都有

$$V_C([u_1,u_2]\times[v_1,v_2])=C(u_2,v_2)-C(u_2,v_1)-C(u_1,v_2)+C(u_1,v_1)\ge 0,$$

则称函数C为Copula (相关结构).

*国家自然科学基金资助(70573077).
本文2005年8月11日收到, 2005年11月21日收到修改稿.

Sklar于1959年首先建立了随机向量(X,Y)的联合分布H和Copula C之间的关系, 他将相关结构从联合分布中提取出来, 使得Copula成为度量相关性的好方法.

定理 2.1 (Sklar定理[3]) 设随机向量(X,Y)的联合分布为$H(x,y)$, 边缘分布分别为$F(x)$和$G(y)$, 则存在Copula C, 使得对于任意的$(x,y)\in R^2$, 都有

$$H(x,y)=C(F(x),G(y)).$$

如果边缘分布函数是连续的, 则Copula C是唯一确定的; 否则, C在Ran$F*$Ran$G$上是唯一确定的. 称函数C为随机变量X与Y的相关结构函数(Copula).

根据Sklar定理, 如果已知随机变量X与Y的联合分布, 则可得到关于X与Y的所有条件分布和Copula. 反之, 如果已知两个变量之间的Copula和边缘分布, 就可以确定它们的联合分布.

当随机变量X与Y相互独立时, Copula为$C(u,v)=uv$, 我们称之为乘积Copula, 并记作$\Pi(u,v)$. 下面给出度量相关性的两种方法[2, 3].

定义 2.2 设随机向量(X,Y)的联合分布函数为H, $(X_1,Y_1),(X_2,Y_2),(X_3,Y_3)$与$(X,Y)$独立同分布, 则称

$$\mathsf{P}\{(X_1-X_2)(Y_1-Y_2)>0\}-\mathsf{P}\{(X_1-X_2)(Y_1-Y_2)<0\}$$

为随机向量(X,Y) (或Copula C)的Kendall秩相关系数或Kendall's τ, 记作$\tau(X,Y)$或$\tau(C)$. 称

$$3[\mathsf{P}\{(X_1-X_2)(Y_1-Y_3)>0\}-\mathsf{P}\{(X_1-X_2)(Y_1-Y_3)<0\}]$$

为随机向量(X,Y) (或Copula C)的 Spearman 秩相关系数或 Spearman's ρ, 记作$\rho(X,Y)$或$\rho(C)$.

Kendall秩相关系数和Spearman秩相关系数具有性质[2, 3]: (1)在单调递增变换下, 秩相关系数保持不变; (2)当两个变量相互独立时, 秩相关系数均为0; (3)当两个变量完全正相关时, 秩相关系数均为1; 当两个变量完全负相关时, 秩相关系数均为-1. 总之, 它们满足理想相关性度量的性质.

当随机向量(X,Y)为连续型时, 可以证明[2], Kendall秩相关系数和Spearman秩相关系数完全由它们的Copula决定, 而与变量的边缘分布无关. 但当随机向量(X,Y)为离散型时, 由于Copula函数的不惟一性, 给计算Kendall和Spearman秩相关系数都带来许多麻烦. 这里我们避开Copula, 用定义2.2计算两个秩相关系数.

§3. 离散型次序统计量的相关性

设$X_1,\cdots,X_n$是独立同分布的随机变量, 共同分布为F, 记

$$X_{(1)}=\min\{X_1,\cdots,X_n\},\qquad X_{(n)}=\max\{X_1,\cdots,X_n\},$$

则$X_{(1)}$和$X_{(n)}$的分布函数[5]分别为

$$F_1(x) = 1 - (1 - F(x))^n, \qquad F_n(y) = F^n(y),$$

$X_{(1)}$和$X_{(n)}$的联合分布函数[5]为

$$F_{1,n}(x,y) = \begin{cases} F^n(y) - (F(y) - F(x))^n, & x < y; \\ F^n(y), & x \geq y. \end{cases}$$

令$u = F_1(x)$, $v = F_n(y)$, 则根据Sklar定理, 得随机变量$X_{(1)}$和$X_{(n)}$之间的Copula为

$$C_n(u,v) = \begin{cases} v - (v^{1/n} + (1-u)^{1/n} - 1)^n, & v^{1/n} + (1-u)^{1/n} > 1; \\ v, & v^{1/n} + (1-u)^{1/n} \leq 1. \end{cases}$$

函数$C_n(u,v)$具有下面的性质:

定理 3.1　对任意的$u, v \in [0,1]$, 有 $\lim\limits_{n\to\infty} C_n(u,v) = \Pi(u,v)$.

证明:　因为对于Clayton copula

$$C_\theta^C(u,v) = \max\{(u^{-\theta} + v^{-\theta} - 1)^{1/\theta}, 0\}$$

有结论[6] $\lim\limits_{\theta\to 0} C_\theta^C(u,v) = uv$, 而$X_{(1)}$和$X_{(n)}$的Copula $C_n(u,v)$与Clayton copula C_θ^C之间具有关系式

$$C_n(u,v) = v - C_{-1/n}^C(1-u, v).$$

所以

$$\lim_{n\to\infty} C_n(u,v) = v - (1-u)v = uv = \Pi(u,v).$$

□

该定理说明, 当样本容量n很大时, 最小次序统计量$X_{(1)}$和最大次序统计量$X_{(n)}$是渐近独立的.

设(X,Y)是离散型随机向量, 联合概率分布列为

$$\mathsf{P}\{X = x_i, Y = y_j\} = p_{ij}, \qquad i,j = 0,1,\cdots,$$

分布函数为

$$H(x,y) = \sum_{i,j} p_{ij} I_{\{x_i \leq x, y_j \leq y\}}.$$

定理 3.2　设离散型随机向量(X,Y)的分布函数为$H(x,y)$, 边缘分布函数分别为$F(x)$和$G(y)$, 则

$$\tau(X,Y) = \mathsf{E}H(X,Y) + \mathsf{E}H(X,Y_{-1}) + \mathsf{E}H(X_{-1},Y) + \mathsf{E}H(X_{-1},Y_{-1}) - 1, \quad (3.1)$$

$$\begin{aligned} \rho(X,Y) = {} & 3\{\mathsf{E}[F(X)G(Y)] + \mathsf{E}[F(X)G(Y_{-1})] \\ & + \mathsf{E}[F(X_{-1})G(Y)] + \mathsf{E}[F(X_{-1})G(Y_{-1})] - 1\}. \end{aligned} \quad (3.2)$$

其中变量X_{-1}, Y_{-1}的含义为: 当$X = x_i$, $Y = y_j$时, $X_{-1} = x_{i-1}$, $Y_{-1} = y_{j-1}$.

证明: 设离散型随机向量$(X_1,Y_1),(X_2,Y_2),(X_3,Y_3)$与$(X,Y)$独立同分布, 因为

$$\begin{aligned}
\tau(X,Y) &= \mathsf{P}\{(X_1-X_2)(Y_1-Y_2)>0\}-\mathsf{P}\{(X_1-X_2)(Y_1-Y_2)<0\}\\
&= \mathsf{P}\{X_1-X_2>0,Y_1-Y_2>0\}+\mathsf{P}\{X_1-X_2<0,Y_1-Y_2<0\}\\
&\quad -\mathsf{P}\{X_1-X_2>0,Y_1-Y_2<0\}-\mathsf{P}\{X_1-X_2<0,Y_1-Y_2>0\}\\
&= \sum_{i,j}\mathsf{P}\{X_1=x_i,Y_1=y_j\}\mathsf{P}\{X_2<x_i,Y_2<y_j\}\\
&\quad +\sum_{i,j}\mathsf{P}\{X_1=x_i,Y_1=y_j\}\mathsf{P}\{X_2>x_i,Y_2>y_j\}\\
&\quad -\sum_{i,j}\mathsf{P}\{X_1=x_i,Y_1=y_j\}\mathsf{P}\{X_2<x_i,Y_2>y_j\}\\
&\quad -\sum_{i,j}\mathsf{P}\{X_1=x_i,Y_1=y_j\}\mathsf{P}\{X_2>x_i,Y_2<y_j\}\\
&= \sum_{i,j}p_{ij}\{H(x_{i-1},y_{j-1})+[1-F(x_i)-G(y_j)+H(x_i,y_j)]\}\\
&\quad -\sum_{i,j}p_{ij}\{[1-G(y_j)-H(x_{i-1},y_j)]+[1-F(x_i)-H(x_i,y_{j-1})]\}\\
&= \sum_{i,j}p_{ij}[H(x_i,y_j)+H(x_i,y_{j-1})+H(x_{i-1},y_j)+H(x_{i-1},y_{j-1})]-1\\
&= \mathsf{E}H(X,Y)+\mathsf{E}H(X,Y_{-1})+\mathsf{E}H(X_{-1},Y)+\mathsf{E}H(X_{-1},Y_{-1})-1.
\end{aligned}$$

所以, (3.1)式成立. 同理可证(3.2)式. □

定理3.2适用于任何两个离散型随机变量之间的秩相关系数的计算. 但对于次序统计量来说, 秩相关系数有着更简单的计算公式.

设$X_1,\cdots,X_n$是与X独立同分布的离散型随机变量, 概率分布列为

$$\mathsf{P}\{X=x_i\}=p_i,\qquad i=0,1,\cdots,$$

分布函数为$F(x)=\mathsf{P}\{X\le x\}=\sum\limits_i p_iI_{\{x_i\le x\}}$, 特别地, $F(x_k)=\sum\limits_{i=0}^{k}p_i$. 当$x=x_i$, $y=x_j$时, 随机变量$X_{(1)},X_{(n)}$以及$(X_{(1)},X_{(n)})$的分布函数值分别为

$$\begin{aligned}
&F_1(x_i)=1-(1-F(x_i))^n,\qquad i=0,1,\cdots;\\
&F_n(x_j)=F^n(x_j),\qquad j=0,1,\cdots;\\
&F_{1,n}(x_i,x_j)=\begin{cases}F^n(x_j)-[F(x_j)-F(x_i)]^n, & i<j;\\ F^n(x_j), & i\ge j.\end{cases}\qquad i,j=0,1,\cdots.
\end{aligned}$$

下面给出秩相关系数的计算方法.

定理 3.3 设$X_1,\cdots,X_n$是与X独立同分布的离散型随机变量, 概率分布列为

$$\mathsf{P}\{X=x_i\}=p_i,\qquad i=0,1,\cdots,$$

分布函数在点$x = x_j$的值为$F(x_j) = \sum_{i=0}^{j} p_i,\ j = 0, 1, \cdots$. 则$X_{(1)}$和$X_{(n)}$之间的Kendall秩相关系数为

$$\begin{aligned}\tau(X_{(1)}, X_{(n)}) &= 2\sum_{j=1}^{\infty}\sum_{i=0}^{j-1}\{[F(x_j) - F(x_i)]^n[F(x_{j-1}) - F(x_{i-1})]^n \\ &\quad - [F(x_j) - F(x_{i-1})]^n[F(x_{j-1}) - F(x_i)]^n\}. \qquad (3.3)\end{aligned}$$

证明: 令$a_j = F(x_j)$, $H(i,j) = H(x_i, x_j) = F_{1,n}(x_i, x_j)$, $p_{ij} = \mathsf{P}(X_{(1)} = x_i, X_n = x_j)$, 则

$$p_{ij} = H(i,j) + H(i-1, j-1) - H(i-1, j) - H(i, j-1),$$

根据(3.1)式得

$$\begin{aligned}\tau + 1 &= \sum_{i,j} p_{ij}(H(i,j) + H(i,j-1) + H(i-1,j) + H(i-1,j-1)) \\ &= \sum_{i,j}[H^2(i,j) + H^2(i-1,j-1) - H^2(i-1,j) - H^2(i,j-1) \\ &\quad + 2H(i,j)H(i-1,j-1) - 2H(i-1,j)H(i,j-1)] \\ &= \sum_{j=0}^{\infty}\sum_{i=0}^{j-1}\{[a_j^n - (a_j - a_i)^n]^2 + [a_{j-1}^n - (a_{j-1} - a_{i-1})^n]^2 \\ &\quad - [a_j^n - (a_j - a_{i-1})^n]^2 - [a_{j-1}^n - (a_{j-1} - a_i)^n]^2 \\ &\quad + 2[a_j^n - (a_j - a_i)^n][a_{j-1}^n - (a_{j-1} - a_{i-1})^n] \\ &\quad - 2[a_j^n - (a_j - a_{i-1})^n][a_{j-1}^n - (a_{j-1} - a_i)^n]\} \\ &\quad + \sum_{j=0}^{\infty}\{a_j^{2n} - [a_j^n - (a_j - a_{j-1})^n]^2 + 2a_j^n a_{j-1}^n - 2[a_j^n - (a_j - a_{j-1})^n]a_{j-1}^n\} \\ &= \sum_{j=0}^{\infty}\sum_{i=0}^{j-1}\{-2a_j^n(a_j - a_i)^n + (a_j - a_i)^{2n} - 2a_{j-1}^n(a_{j-1} - a_{i-1})^n + (a_{j-1} - a_{i-1})^{2n} \\ &\quad + 2a_j^n(a_j - a_{i-1})^n - (a_j - a_{i-1})^{2n} + 2a_{j-1}^n(a_{j-1} - a_i)^n - (a_{j-1} - a_i)^{2n} \\ &\quad + 2(a_j - a_i)^n(a_{j-1} - a_{i-1})^n - 2a_{j-1}^n(a_j - a_i)^n - 2a_j^n(a_{j-1} - a_{i-1})^n \\ &\quad - 2(a_j - a_{i-1})^n(a_{j-1} - a_i)^n + 2a_{j-1}^n(a_j - a_{i-1})^n + 2a_j^n(a_{j-1} - a_i)^n\} \\ &\quad + \sum_{j=0}^{\infty}\{2a_j^n(a_j - a_{j-1})^n - (a_j - a_{j-1})^{2n} + 2a_{j-1}^n(a_j - a_{j-1})^n\} \\ &= \sum_{j=0}^{\infty}\sum_{i=0}^{j-1}\{[(a_j - a_i)^{2n} - (a_j - a_{i-1})^{2n}] + [(a_{j-1} - a_{i-1})^{2n} - (a_{j-1} - a_i)^{2n}] \\ &\quad - 2a_j^n[(a_j - a_i)^n - (a_j - a_{i-1})^n] + 2a_{j-1}^n[(a_{j-1} - a_i)^n - (a_{j-1} - a_{i-1})^n] \\ &\quad + 2a_j^n[(a_{j-1} - a_i)^n - (a_{j-1} - a_{i-1})^n] - 2a_{j-1}^n[(a_j - a_i)^n - (a_j - a_{i-1})^n] \\ &\quad + 2(a_j - a_i)^n(a_{j-1} - a_{i-1})^n - 2(a_j - a_{i-1})^n(a_{j-1} - a_i)^n\} \\ &\quad + \sum_{j=0}^{\infty}\{2a_j^n(a_j - a_{j-1})^n - (a_j - a_{j-1})^{2n} + 2a_{j-1}^n(a_j - a_{j-1})^n\}\end{aligned}$$

$$
\begin{aligned}
= & \sum_{j=0}^{\infty}\{(a_j-a_{j-1})^{2n}-a_j^{2n}+a_{j-1}^{2n}-2a_j^n[(a_j-a_{j-1})^n+a_j^n] \\
& +2a_{j-1}^n(-a_{j-1}^n)+2a_j^n(-a_{j-1}^n)-2a_{j-1}^n[(a_j-a_{j-1})^n-a_j^n]\} \\
& +2\sum_{j=0}^{\infty}\sum_{i=0}^{j-1}[(a_j-a_i)^n(a_{j-1}-a_{i-1})^n-(a_j-a_{i-1})^n(a_{j-1}-a_i)^n] \\
& +\sum_{j=0}^{\infty}\{2a_j^n(a_j-a_{j-1})^n-(a_j-a_{j-1})^{2n}+2a_{j-1}^n(a_j-a_{j-1})^n\} \\
= & \sum_{j=0}^{\infty}[(a_j-a_{j-1})^{2n}+a_j^{2n}-a_{j-1}^{2n}-2a_j^n(a_j-a_{j-1})^n-2a_{j-1}^n(a_j-a_{j-1})^n] \\
& -2\sum_{j=0}^{\infty}\sum_{i=0}^{j-1}[(a_j-a_{i-1})^n(a_{j-1}-a_i)^n-(a_j-a_i)^n(a_{j-1}-a_{i-1})^n] \\
& +\sum_{j=0}^{\infty}\{2a_j^n(a_j-a_{j-1})^n-(a_j-a_{j-1})^{2n}+2a_{j-1}^n(a_j-a_{j-1})^n\} \\
= & 1-2\sum_{j=0}^{\infty}\sum_{i=0}^{j-1}[(a_j-a_{i-1})^n(a_{j-1}-a_i)^n-(a_j-a_i)^n(a_{j-1}-a_{i-1})^n] \\
= & 1+2\sum_{j=1}^{\infty}\sum_{i=0}^{j-1}[(a_j-a_i)^n(a_{j-1}-a_{i-1})^n-(a_j-a_{i-1})^n(a_{j-1}-a_i)^n].
\end{aligned}
$$

所以, 将$a_j=F(x_j)$代入上式, 则可得(3.3)式成立. □

定理 3.4 设$X_1,\cdots,X_n$是与X独立同分布的离散型随机变量, 概率分布列为

$$\mathsf{P}\{X=x_i\}=p_i,\qquad i=0,1,\cdots,$$

分布函数在点$x=x_j$的值为$F(x_j)=\sum_{i=0}^{j}p_i$, $j=0,1,\cdots$. 则$X_{(1)}$和$X_{(n)}$之间的Spearman秩相关系数为

$$
\begin{aligned}
& \rho(X_{(1)},X_{(n)}) \\
= & 3\sum_{j=2}^{\infty}\sum_{i=0}^{j-2}\{[\overline{F}^n(x_{i+1})-\overline{F}^n(x_{i-1})][F(x_j)-F(x_i)]^n[F^n(X_{j+1})-F^n(x_{j-1})]\} \\
& -3\sum_{j=0}^{\infty}\{[\overline{F}^n(x_j)-\overline{F}^n(x_{j-2})]F^n(x_{j-1})[F(x_j)-F(x_{j-1})]^n\}. \qquad (3.4)
\end{aligned}
$$

其中$\overline{F}(x)=1-F(x)$.

该定理的证明与定理3.3类似, 只需按部就班的计算即可.

例 设随机变量X服从两点分布$b(1,p)$, $X_1,\cdots,X_n$是来自总体X的样本, 由于$F(0)=1-p$, $F(1)=1$, 则根据(3.3)与(3.4)式得, $X_{(1)}$和$X_{(n)}$之间的Kendall秩相关系数和Spearman秩相关系数分别为

$$\tau(X_{(1)},X_{(n)})=2p^n(1-p)^n,\qquad \rho(X_{(1)},X_{(n)})=3p^n(1-p)^n.$$

用定义2.2和(3.1)、(3.2)式分别计算, 所得结果与上面的计算结果一致.

综上, 我们用具有良好性质的两种度量相关性的方法, 讨论了离散型最小次序统计量和最大次序统计量之间的相关性. 定理3.2给出了任何两个离散型随机变量之间的Kendall和Spearman秩相关系数的计算公式; 而定理3.3和定理3.4则分别给出了次序统计量$X_{(1)}$和$X_{(n)}$之间的Kendall和Spearman秩相关系数的计算公式, 虽然结论是以级数的形式给出, 但由于计算软件的普及和计算机运算速度的提高, 利用该结论计算秩相关系数还是简单易行的. 当然, 任何两个次序统计量之间的相关性度量还有待于进一步研究.

参 考 文 献

[1] 史道济, 相关系数与相关性, 统计科学与实践, **4**(2002), 22–24.

[2] Nelsen, R.B., *An Introduction to Copulas*, Springer, New York, 1999.

[3] Joe, H., *Multivariate Models and Dependence Concepts*, Chapman and Hall, London, 1997.

[4] Averous, J., Genest, C. and Kochar, S.C., On the dependence structure of order statistics, *Journal of Multivariate Analysis*, **94**(2005), 159–171.

[5] 茆诗松, 王静龙, 濮晓龙, 高等数理统计, 高等教育出版社, 1998.

[6] Cook, R.D., Johson, M.E., A family of distributions for modeling non-elliptically symmetric multivariate data, *J. Roy. Statist. Soc. B*, **43(2)**(1981), 210–218.

The Dependence Analysis between Minimum and Maximum Order Statistics

Liang Fengzhen Shi Daoji

(*Department of Mathematics, Institute of Science, Tianjin University, Tianjin, 300072*)

The dependence measures are discussed for the discrete random variables in this paper. It is proved that the asymptotic independence of the minimum and maximum of n i.i.d. random variables, and explicit expressions are given for the value of Kendall's and Spearman's rank correlation coefficient between minimum and maximum order statistics.

Keywords: Dependence measures, order statistics, Kendall's rank correlation coefficient, Spearman's rank correlation coefficient.

AMS Subject Classification: 62H20.