多分格相关与皮尔逊相关的蒙特卡罗仿真

吴瑞林 王建中 袁克海

(北京航空航天大学 心理与行为研究所，北京 100191) (圣母大学 心理系，南湾，美国 46556)

摘 要：为了解对于有序分类数据，使用多分格相关系数估计数据间线性依存关系的精确性，运用蒙特卡罗仿真法考察样本量的大小、分类数的多少、相关程度的高低以及数据分布的偏度对多分格相关系数估计精确度的影响；同样条件下计算皮尔逊相关系数并与之相比较. 在实验设计的条件下，多分格相关系数的表现非常稳定，绝大多数情况下给出的估计值都比皮尔逊相关系数更精确. 两种相关系数受样本量的影响都很有限，且估计值的偏差会明显随着分类数的增多而变小；但数据的偏态分布对皮尔逊相关系数的精确度有很大影响.

关 键 词：相关分析；蒙特卡洛仿真法；数据处理与分析；多元分析

中图分类号：O 212.4

文献标识码：A **文 章 编 号**：1001-5965(2009)12-1507-04

Monte Carlo simulation of polychoric correlation and Pearson correlation coefficient

Wu Ruilin Wang Jianzhong

(Institute of Psychology and Behavior, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

Yuan Kehai

(Department of Psychology, University of Notre Dame, Notre Dame 46556, United States)

Abstract: The more accurate estimates were obtained via the polychoric correlation coefficient rather than traditional Pearson correlation coefficient in multivariate analysis for ordinal categorical data. The statistic model and estimators of the polychoric correlation were introduced. Then a Monte Carlo simulation was conducted to discuss the influence of sample size, category number, correlation degree, and data distribution on the precision of polychoric correlation estimate. The simulation results show that the polychoric correlation coefficient is more robust, and more precise than Pearson correlation coefficient in the most of the simulation setting. To both two correlation estimation approaches, sample size is not an influential factor and the bias has explicit decrease when adding the number of category. The skew distribution would distort the Pearson correlation; however it has a very limited influence on the polychoric correlation.

Key words: correlation theory; Monte Carlo methods; data processing; multivariant analysis

在科学研究，特别是管理科学和心理科学的研究中，有一些事物很难在一个连续轴上给出精准的测量，却可以很容易的将其定位到一组有等级关系的离散变量上，这样的离散数据被称为有序分类数据(ordinal categorical data). 早在100多年前，皮尔逊就注意到这类数据的存在，并提出用皮尔逊积矩相关系数(Pearson product moment correlation coefficient)来计算连续数据间的线性依存关系，而使用多分格相关系数(polychoric correlation coefficient)估计有序分类数据间的关系. 但由于计算技术的限制，对该问题的研究在相当长的一段时期内处于停滞状态. 直到上

收稿日期：2009-03-03

作者简介：吴瑞林(1980—)，男，陕西汉中人，博士生，wuruilin@cse.buaa.edu.cn.

世纪70年代,多分格相关系数的理论和估计方法才在计算机技术的支撑下得到快速发展.在多分格相关系数被统计学家深入研究之前,皮尔逊相关由于使用简单、便于理解的优点,已被广泛的应用于各类研究;尽管用它估计有序分类数据间的相关会产生不可忽略的误差[1].

本文将用蒙特卡罗仿真法(Monte Carlo method)考察在不同条件下,皮尔逊相关和多分格相关估计有序分类数据间相关的情况,并对两种方法估计值的偏差进行比较,以验证多分格相关系数的优势,确定其适用范围,并考察样本量、分类数、相关程度和样本分布等因素的影响程度.

1 统计模型与假设

多分格相关系数的理论假设认为,观察到的不连续的有序分类变量的背后,是连续的潜变量.假设观察到两个变量 x 和 y,这两个变量都是有序分类变量,且分别生成自两个不可观察到的连续变量 ξ 和 η.ξ 和 η 服从二元正态分布,x 和 y 分别有 r 个和 s 个分类.x 和 ξ 的关系可用式(1)表示:

$$x=\begin{cases}1 & a_0<\xi<a_1\\ 2 & a_1\leqslant\xi<a_2\\ \vdots & \vdots\\ r & a_{r-1}\leqslant\xi<a_r\end{cases} \tag{1}$$

式(1)表示潜变量 ξ 落在某一范围内时,就可以观察到相应的观测值.

y 和 η 的关系也可以用类似的方程表示,并使用 $b_0,b_1,\cdots,b_s$ 替换 a_i 作为 η 变量的边界值.这里,a_i 和 b_i 被称为"临界点",且 $a_0=b_0=-\infty$,$a_r=b_s=+\infty$.如果直接对变量 x 和 y 求皮尔逊相关,得到的是两个观察变量间的相关 ρ_{xy},但实际上希望获得的是两个潜变量间的相关 $\rho_{\xi\eta}$(下面直接用 ρ 表示).

现在,假设变量 x 和 y 分别被两个临界点切分为3个分类,且背后的潜变量服从二元正态分布,就可以用图1来描述这一关系.95%的数据会落在图中的椭圆之内,整个坐标空间被4个"临界点"分为了9个区域,每个区域内有 n_{ij} 个观察到的数据,比如,观察值 $x=2$ 和 $y=3$ 的样本数量为 n_{23}.推广到一般情况,对于两个变量的有序分类数据,可以把样本总结为一张 r 行×s 列的列连表,表格中每个单元格的值是对应取值的样本数量.

由于无法获知 ξ 和 η 的真实度量刻度,只能

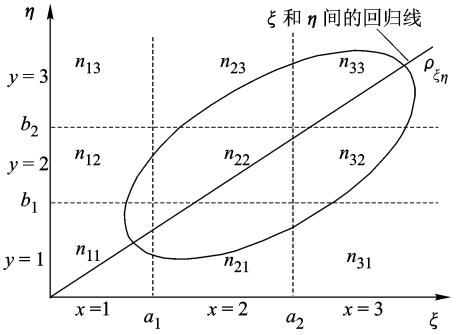


图1 多分格相关系数概念示意图

假设它们服从 $\mu_\xi=\mu_\eta=0$,$\sigma_\xi^2=\sigma_\eta^2=1$ 的标准二元正态分布,其概率密度函数 $\phi(\xi,\eta;\rho)$ 可用如下方程表示:

$$\phi(\xi,\eta;\rho)=\frac{1}{2\pi(1-\rho^2)^{1/2}}\exp\left[\frac{-1}{2(1-\rho^2)}(\xi^2-2\rho\xi\eta+\eta^2)\right] \tag{2}$$

那么,观察到 $x=x_i$,$y=y_i$ 的可能性就为

$$\pi_{ij}=\int_{a_{i-1}}^{a_i}\int_{b_{i-1}}^{b_i}\phi(\xi,\eta;\rho)\mathrm{d}\eta\mathrm{d}\xi \tag{3}$$

2 估计方法

下面的问题就是如何利用方程(2)和(3)以及观察到的样本分布来估计相关系数 $\rho_{\xi\eta}$.一种方法是联合极大似然(joint maximum likelihood)估计法,这也是最常见的参数估计方法;另一种是在联合极大似然法基础上发展出来的两步法(two-step estimator).

如果用 n_{ij} 表示实际观察到 $x=x_i$,$y=y_j$ 的样本数,则总样本数为 $N=\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}$,那么样本的对数似然函数为

$$l=\ln L=\sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}\ln\pi_{ij} \tag{4}$$

要想获得相关系数 ρ 的极大似然估计值,就需要对 l 求 ρ 的一阶导数并使其等于零.对 l 求一阶导数得到的表达式为

$$\frac{\mathrm{d}l}{\mathrm{d}\rho}=\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{\pi_{ij}}[\phi(a_i,b_j;\rho)-\phi(a_{i-1},b_j;\rho)-\phi(a_i,b_{j-1};\rho)+\phi(a_{i-1},b_{j-1};\rho)] \tag{5}$$

需要注意的是,方程(5)中不仅相关系数 ρ 是未知参数,临界点 a_i 和 b_j 也同样未知.所以,在估计 ρ 时,a_i 和 b_j 需要同时被估计.文献[1-3]分别详细描述过如何利用牛顿-拉夫逊迭代(Newton-Raphsom method)进行极大似然估计.

从仿真的结果[4-5]看,联合极大似然估计法

获得的估计值非常接近真实值,但由于它需要同时对相关系数和临界点进行估计,当变量的分类较多时,计算开销明显增大.为了减少计算,文献[1]建议使用临界点的边际累积比率(cumulative marginal proportion)来取代其估计值,然后只对相关系数 ρ 进行迭代估计,该方法被称为"两步法".本仿真中将采用这种估计方法.

观察变量 x 和 y 间皮尔逊相关系数 r_{xy} 的计算方法已为大家所熟知,其公式为

$$r_{xy}=\frac{\sum(X-\overline{X})(X-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}} \tag{6}$$

3　仿真实验

为了验证多分格相关系数在有序分类数据上的估值准确性,本文进行了两个仿真实验.

3.1　实验 1:样本量和分类数的影响

实验 1 考察在正态分布条件下,分类数目和样本量对两种相关系数估计的影响.该实验中,选择 3 个相关系数 0.20,0.50 和 0.80 作为真实值,分别代表低、中、高 3 种不同程度的相关;分类数据的分类数目选择了常见的 2 点、3 点、5 点和 7 点 4 种;样本量 N 选择了从 100 到 500,每 100 为一个间隔,共 5 种样本量;这样一共存在 3×4×5=60 种情况.

对于每种情况,首先在正态分布下生成 N 个连续潜变量向量$[\xi,\eta]$,然后根据分类数目设置等间隔(为保证数据呈正态分布)的临界值切分连续潜变量,从而生成 N 个有序分类变量向量$[x,y]$,接下来对样本求皮尔逊相关系数和多分格相关系数.为了减小随机误差的影响,对于每种条件重复进行 100 次实验.最后求两种相关系数 100 次估计值的平均数,并计算平均值与预先设定的真实值之差的绝对值,这个差被称为偏差(bias).

仿真结果见表 1,其中列出了 60 种实验条件下,两种相关系数 100 次重复估值的平均数,还给出了估计值的偏差.

可以发现,除两种个别情况外(表 1 中,这两种情况的数据用粗体字突出显示),皮尔逊相关系数估计值的偏差都高于多分格相关的偏差,也就是说多分格相关系数估计值更为准确.而且,除了低度相关且分类数为 2 的情况外,其它条件下,多分格相关的偏差基本都比皮尔逊相关的偏差小一个数量级以上.当分类数小于等于 5 个的时候,皮尔逊相关系数的偏差占到了真实值的 10% 左右或以上,这样的偏差应该说是比较明显的.而同样情况下,多分格相关估计值的偏差则较小,一般只占到真实值的 10% 以下,最小的不足 1%.

表 1 中两种相关系数的偏差也可以转换为折线图,限于篇幅,这里只给出中度相关($\rho=0.50$)的情况,如图 2 所示.图 2a 表示皮尔逊相关系数的偏差,对应的图 2b 为多分格相关的偏差.

表 1　正态分布下皮尔逊相关系数与多分格相关系数的比较(重复次数为 100)

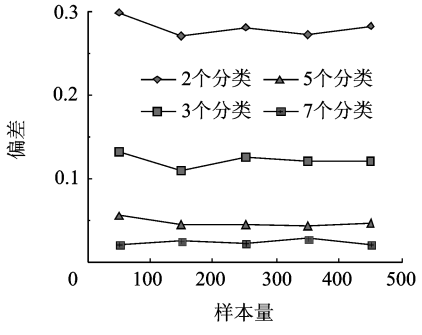
相关程度	样本量 N	2 个分类				3 个分类				5 个分类				7 个分类			
		$\overline{\hat{\rho}}_{ppc}$	B_{ppc}	$\overline{\hat{\rho}}_{pcc}$	B_{pcc}	$\overline{\hat{\rho}}_{ppc}$	B_{ppc}	$\overline{\hat{\rho}}_{pcc}$	B_{pcc}	$\overline{\hat{\rho}}_{ppc}$	B_{ppc}	$\overline{\hat{\rho}}_{pcc}$	B_{pcc}	$\overline{\hat{\rho}}_{ppc}$	B_{ppc}	$\overline{\hat{\rho}}_{pcc}$	B_{pcc}
低度相关 $\rho=0.20$	100	0.09	0.11	0.26	0.06	0.15	0.05	0.21	0.01	0.18	0.02	0.19	0.01	0.19	0.02	0.20	0.00
	200	0.07	0.13	0.28	0.08	0.16	0.04	0.21	0.01	0.18	0.02	0.20	0.00	0.19	0.01	0.20	0.00
	300	**0.07**	**0.13**	**0.33**	**0.13**	0.15	0.05	0.22	0.02	0.18	0.02	0.20	0.00	0.19	0.01	0.20	0.00
	400	0.07	0.13	0.32	0.12	0.15	0.05	0.20	0.00	0.18	0.02	0.20	0.01	0.18	0.02	0.19	0.01
	500	**0.09**	**0.11**	**0.34**	**0.14**	0.15	0.05	0.23	0.03	0.19	0.02	0.20	0.00	0.20	0.01	0.21	0.01
中度相关 $\rho=0.50$	100	0.20	0.30	0.52	0.02	0.37	0.13	0.49	0.01	0.45	0.05	0.48	0.02	0.48	0.02	0.51	0.01
	200	0.23	0.27	0.52	0.02	0.39	0.11	0.50	0.00	0.46	0.05	0.50	0.00	0.47	0.03	0.50	0.00
	300	0.22	0.28	0.53	0.03	0.38	0.13	0.50	0.00	0.46	0.04	0.50	0.00	0.48	0.02	0.50	0.00
	400	0.23	0.27	0.53	0.03	0.38	0.12	0.50	0.00	0.46	0.04	0.50	0.00	0.47	0.03	0.50	0.01
	500	0.22	0.28	0.53	0.03	0.38	0.12	0.50	0.00	0.45	0.05	0.50	0.00	0.48	0.02	0.50	0.00
高度相关 $\rho=0.80$	100	0.47	0.33	0.81	0.01	0.64	0.16	0.81	0.01	0.74	0.07	0.81	0.01	0.77	0.03	0.80	0.00
	200	0.48	0.32	0.80	0.00	0.64	0.16	0.81	0.01	0.73	0.07	0.80	0.00	0.77	0.03	0.80	0.00
	300	0.47	0.33	0.79	0.01	0.64	0.17	0.80	0.00	0.73	0.07	0.80	0.00	0.77	0.03	0.80	0.00
	400	0.48	0.32	0.79	0.01	0.64	0.16	0.81	0.01	0.74	0.07	0.80	0.00	0.77	0.03	0.80	0.00
	500	0.48	0.32	0.79	0.01	0.64	0.16	0.81	0.01	0.73	0.07	0.80	0.00	0.77	0.03	0.80	0.00

注:$\overline{\hat{\rho}}_{ppc}$ 表示使用皮尔逊相关获得估计值的均值;$\overline{\hat{\rho}}_{pcc}$ 表示用多分格相关获得估计值的均值;B_{ppc} 表示 $\overline{\hat{\rho}}$ 与真实值之差的绝对值,即皮尔逊相关系数的偏差;B_{pcc} 表示 $\overline{\hat{\rho}}$ 的偏差;粗体字表示多分格相关系数的偏差高于皮尔逊相关.

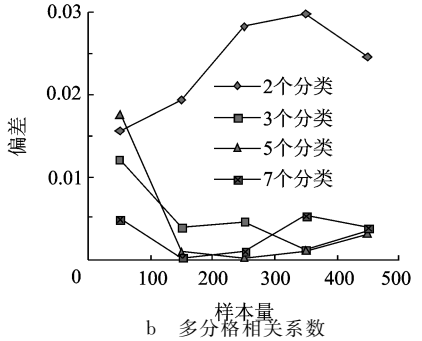
非常明显,皮尔逊相关估计值的偏差随着分类数目的增多而下降,4 条折线在图 2a 中明确的分为 4 个层次;但图中的折线并没有随样本量的增大而明显的减小或扩大,这说明样本量对皮尔逊相关系数估计的影响并不显著.

在图 2b 中,4 条折线没有出现如图 2a 那样

明确的层次,而是在很多地方有交叉.大体上2个分类的折线在上,3个分类的折线在中间,5个和7个分类的折线在最下面,但这种趋势并不严格.这可能是因为在3个及以上分类的情况下,多分格相关系数估计值的偏差本身就比较小(图2a和图2b纵坐标的单位量级并不相同),此时的误差很可能更多的来自于随机误差而不是方法引起的系统性误差,因此出现了折线交错的状态.与皮尔逊相关类似,多分格相关系数受样本量的影响也不明显,这两者间的关系变得毫无规律可循.



a 皮尔逊相关系数



b 多分格相关系数

图2 $\rho=0.50$时估计值偏差随样本量和分类数目变化的折线图

3.2 实验2:样本分布的影响

实验1是在数据正态分布的条件下进行的,相关系数估计的准确性实际上还会受到样本分布形态的影响.尤其是多分格相关系数的理论假设就是潜变量服从二元正态分布,那么非正态分布的有序分类数据会对多分格相关系数估计准确性的影响就必须被研究.下面设计了实验2来考察这一影响.

由于实验1已经发现样本量对相关系数估计的影响不明显,实验2就只选择样本量等于400.另外,分类数多少的影响在本实验里也不是考察重点,所以只选择了较为常用的5个和7个分类的形式.两个观测变量x和y被设定为4种分布条件:①同为正态;②偏度同为-1;③偏度同为1;④一个变量的偏度为1,另一个的偏度为-1.

实验2的结果列于表2.表2的数据显示,无论在5个分类,还是在7个分类的情况下,偏态分布对多分格相关的估计都没有很大影响.虽然在个别的分布条件下,其偏差有所增大;但最大偏差也只有0.01,不超过低度相关真实值0.20的5%.

反观皮尔逊相关系数,在数据呈偏态分布时,偏差的增加非常明显;特别是当两个变量朝着不同方向偏斜,即偏度为1/-1时,其误差急剧增大,估计值与真实值间相距甚远.即使将分类数目增大到7个,皮尔逊相关系数的偏差也没有明显的减小.

表2 非正态分布下皮尔逊相关系数与多分格相关系数的比较($N=400$,重复100次)

相关程度	样本分布	5个分类				7个分类			
		$\bar{\hat{\rho}}_{ppc}$	B_{ppc}	$\bar{\hat{\rho}}_{pcc}$	B_{pcc}	$\bar{\hat{\rho}}_{ppc}$	B_{ppc}	$\bar{\hat{\rho}}_{pcc}$	B_{pcc}
低度相关 $\rho=0.20$	①	0.18	0.02	0.20	0.01	0.18	0.02	0.19	0.01
	②	0.17	0.03	0.20	0.00	0.17	0.03	0.20	0.00
	③	0.16	0.04	0.20	0.00	0.16	0.04	0.20	0.00
	④	0.16	0.04	0.21	0.01	0.16	0.04	0.20	0.01
中度相关 $\rho=0.50$	①	0.46	0.04	0.50	0.00	0.47	0.03	0.50	0.01
	②	0.42	0.08	0.49	0.01	0.42	0.08	0.49	0.01
	③	0.43	0.07	0.50	0.00	0.43	0.07	0.50	0.00
	④	0.37	0.13	0.49	0.01	0.37	0.13	0.50	0.00
高度相关 $\rho=0.80$	①	0.74	0.07	0.80	0.00	0.77	0.03	0.80	0.00
	②	0.73	0.07	0.80	0.00	0.74	0.06	0.80	0.00
	③	0.73	0.07	0.80	0.00	0.74	0.06	0.80	0.00
	④	0.57	0.23	0.80	0.00	0.56	0.24	0.80	0.01

注:①同为正态;②偏度同为-1;③偏度同为1;④一个变量的偏度为1,另一个的偏度为-1.

4 结 论

综合2个仿真实验的结果,可以对皮尔逊和多分格两种相关系数得出以下4点结论:

1) 对于有序分类数据,多分格相关的估计值

(下转第1515页)