Advances in Applied Mathematics 应用数学进展, 2017, 6(4), 508-514
Published Online July 2017 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2017.64061

Hans汉斯

The Study of the Correlation between Extreme Value and Mean Value of the Student Achievement Based on Copula Function

Hong Chang, Fuxia Xu*

School of Science, Tianjin Polytechnic University, Tianjin
Email: 13512948647@163.com, *xfx0910@126.com

Received: Jun. 25th, 2017; accepted: Jul. 15th, 2017; published: Jul. 18th, 2017

Abstract

Based on the semiparametric estimation method of nonparametric kernel density and the square and Euclidean distance with empirical Copula functions, the Gumbel Copula model is established for the extreme and mean value of the student achievement. We rely on the Gumbel Copula function to investigate the correlation between the highest score, the lowest score and the average score. It turns out that the Kendall rank correlation coefficient and the upper tail correlation coefficient are slightly higher than the highest points.

Keywords

Copula Function, Kerneldensity Estimation, Semiparametric Estimation, Rank Correlation, Tail Correlation

基于Copula函数的学生成绩极值与均值的相关性研究

常　虹，徐付霞*

天津工业大学理学院，天津
Email: 13512948647@163.com, *xfx0910@126.com

收稿日期：2017年6月25日；录用日期：2017年7月15日；发布日期：2017年7月18日

*通讯作者。

文章引用：常虹，徐付霞．基于 Copula 函数的学生成绩极值与均值的相关性研究[J]．应用数学进展，2017，6(4): 508-514. https://doi.org/10.12677/aam.2017.64061

摘　要

本文通过基于非参数核密度的半参数估计法以及与经验Copula函数之间的平方欧式距离对学生成绩的极值与均值建立了Gumbel Copula模型。利用基于Gumbel Copula函数的相关性度量研究了学生成绩最高分、最低分与平均分的相关性，得到最低分与平均分的Kendall秩相关系数与上尾相关系数均稍高于最高分的结论。

关键词

Copula函数，核密度估计，半参数估计，秩相关，尾部相关

1. 引言

均值是描述数据集中位置的一个统计量，用来表示统计对象的一般水平。由于它直观、简明，所以在日常生活中经常被用到，如衡量一个国家或地区经济发展状况的人均 GDP，反映居民收入状况的人均可支配收入，还有在学校中经常被用来衡量学生学习成绩的平均分等。平均值具有非常好的数学性质，但是它对极端值比较敏感，偏向尾部较厚的方向[1]。研究极端值对均值的影响程度，或两者之间的相关性具有重要意义。

2. 问题的分析

极端值对均值的影响可以通过二者之间的相关性来衡量。如果相关性较强，则影响程度较大，否则，则认为影响程度较小。描述变量间相关性的度量方法有很多，其中 Pearson 线性相关系数是最常用的，但是它只能描述变量间的线性相关关系，对于非线性的关系并不能充分刻画。而 Copula 函数是用来研究相关性的一个非常有效的方法。梁冯珍，史道济(2008)通过 Copula 函数研究了离散型最小和最大次序统计量的相关性，讨论了最小次序统计量和最大次序统计量的渐近独立性[2]。徐付霞，董永权(2009)利用 Copula 函数的相关结构分析了泥石流地貌要素中流域面积与流域高差的相关性[3]。Copula 函数的出现使变量之间的相关性刻画更加趋于完善，因为它几乎包含了随机变量所有的相关信息[4]。本文将利用 Copula 函数来研究相关性。

首先，以班级为单位选取某高校学生的线性代数期末考试成绩作为研究对象，将每个班级中的最高分与最低分看作极端值，来研究极端成绩对平均成绩的影响。我们选取的班级包括材料、建筑、电信等 111 个理工科班级，11 个卓越班和 20 个 IT 班，然后计算这全部 142 个班级中，每个班级成绩的最高分、最低分以及平均分。部分计算结果见表 1。

最高分、最低分与平均分的频率分布直方图见图 1，它们各自的偏度、峰度以及正态性检验结果见表 2。其中正态性检验采用 Lillie test 检验[5]，显著性水平为 0.1。

由图 1 与表 2 可知，最高分、最低分与平均分的数据均不服从正态分布，而且也很难从其它常见分布中找到适合它们的分布[6]。因此，我们用非参数密度估计的核密度估计法来对这三个总体的密度函数

Table 1. The highest score, the lowest score and the average score of points in some classes
表 1. 部分班级成绩的最高分、最低分与平均分

序号	1	2	3	4	……	142	取值范围
最高分	99	92	91	95	……	70	[70,100]
最低分	29	10	30	26	……	8	[0,52]
平均分	62.5926	58.8148	64.2963	63.5556	……	39.2308	[36.84,75]

Table 2. Skewness, kurtosis and normality test results
表 2. 偏度、峰度以及正态性检验结果

	偏度	峰度	正态性检验 P 值	正态性检验结果
最高分	−0.7590	3.7990	0.0014	拒绝
最低分	0.0367	2.0530	0.0192	拒绝
平均分	−0.6032	3.6852	0.0861	拒绝

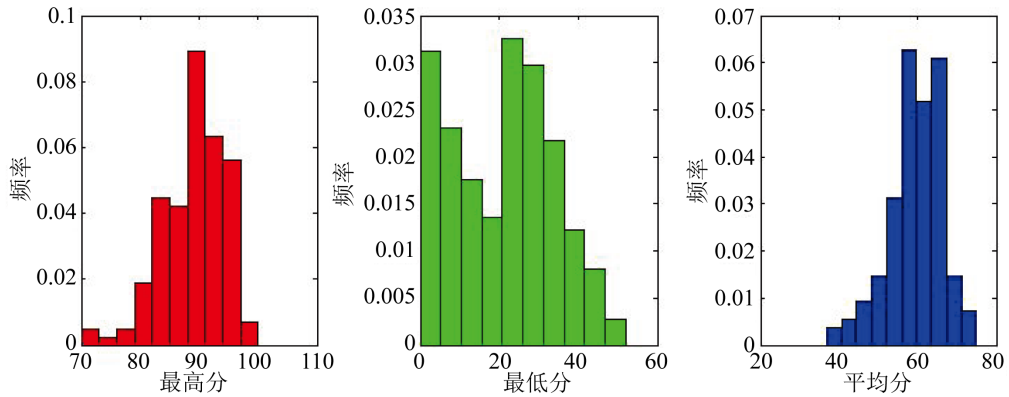

Figure 1. Frequency distribution histogram
图 1. 频率分布直方图

进行拟合，采用 Gaussian 核函数与默认的最佳带宽[7]，做出的核密度估计图与核分布估计图见图 2 与图 3。

比较图 1 与图 2 可知，求出的核密度估计曲线与三组成绩数据的频率直方图均附和较好，即对数据的拟合效果较好。因此，本文对三组成绩数据分布的拟合均采用核密度估计的结果。

3. Copula 函数模型的建立

3.1. 参数估计

首先，选用五个常用的 Copula 函数，分别为正态 Copula，t-Copula，Gumbel Copula，Clayton Copula 与 Frank Copula 作为本文研究的备选 Copula，这五个 Copula 函数可以充分刻画变量间对称与非对称、上尾或下尾相依的相关关系。它们的分布函数表达式及其参数取值范围见文[8]。

然后，我们对备选 Copula 函数中的未知参数进行估计。常用的参数估计方法有极大似然估计法(ML 估计)、分步估计法(IFM 估计)和半参数估计法(CML 估计)，其中半参数估计(CML 估计)又分为基于经验分布函数的标准极大似然估计和基于非参数核密度的极大似然估计[4]。

由于极大似然估计(ML 估计)和分步估计(IFM 估计)的精度依赖于边缘分布拟合的准确性，如果边缘分布的拟合不精确，则 ML 估计和 IFM 估计的精度会受到很大影响，而半参数估计(CML 估计)不需要对边缘分布进行拟合，可以避免因边缘分布拟合不精确带来的损失，因此本文在问题分析中的核密度估计

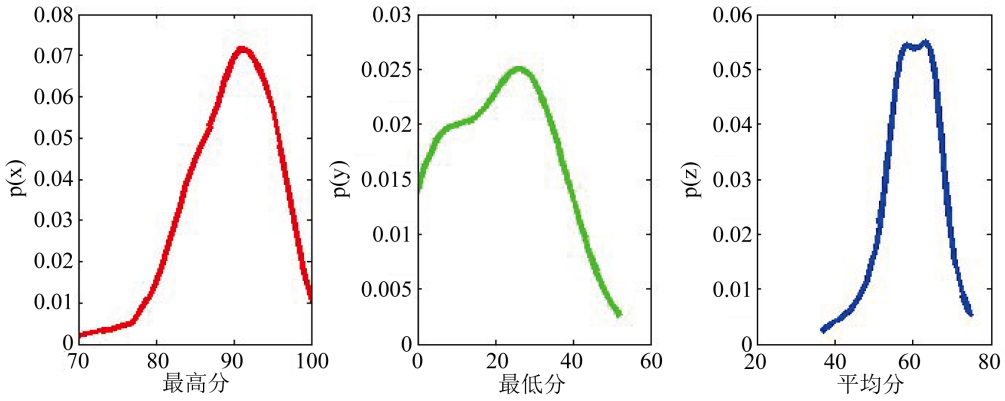

Figure 2. Kernel density estimation
图 2. 核密度估计图

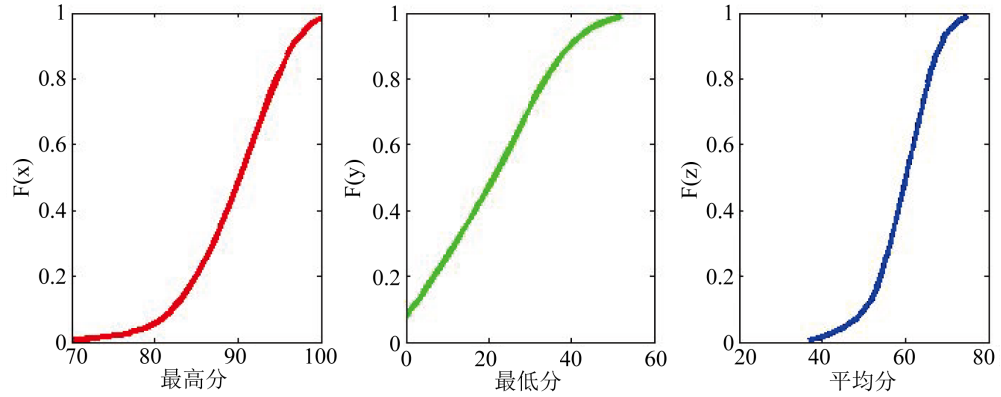

Figure 3. Estimation of nuclear distribution
图 3. 核分布估计图

结果的基础上，采用 CML 估计中基于非参数核密度的极大似然估计法来对 Copula 函数中的未知参数进行估计。即将已经计算出的在原始样本点处的核分布估计值代入 Copula 函数中，采用极大似然法估计函数中的未知参数。本文选取的五个 Copula 函数的参数估计结果分别见表 3 与表 4，其中表 3 表示最高分与平均分之间 Copula 函数的参数估计结果，表 4 表示最低分与平均分之间 Copula 函数的参数估计结果。

从参数估计结果可以看出，五个 Copula 函数的参数估计值均在参数取值范围之内[8]。

3.2. 模型的选取

根据经验 Copula 函数与 Copula 函数的平方欧式距离 d^2 来选择合适的 Copula，距离 d^2 越小，表示 Copula 函数的拟合程度越好。平方欧式距离 d^2 的计算公式为

$$d^2 = \sum_{i=1}^{n} \left| \hat{C}(u_i, v_i) - C(u_i, v_i) \right|^2 \tag{1}$$

其中 $\hat{C}(u_i, v_i)$ 表示经验 Copula 函数，$C(u_i, v_i)$ 表示 Copula 函数。

经验 Copula 函数的图像见图 4，五个 Copula 函数与经验 Copula 函数的平方欧式距离 d^2 见表 5 与表 6，其中表 5 表示最高分与平均分的 Copula 函数的平方欧式距离，表 6 表示最低分与平均分的 Copula 函数的平方欧式距离。

由表 5 可以看出，Gumbel Copula 函数对最高分与平均分的 Copula 函数的拟合效果是最好的，所以

Table 3. Parameter estimation results of Copula functions with the highest score and the average score
表 3. 最高分与平均分的 Copula 函数的参数估计结果

Copula 函数	正态 Copula	t-Copula	Gumbel Copula	Clayton Copula	Frank Copula
参数估计值	0.5015	0.5504	1.5103	0.7641	3.4849

Table 4. Parameter estimation results of Copula functions with the lowest score and the average score
表 4. 最低分与平均分的 Copula 函数的参数估计结果

Copula 函数	正态 Copula	t-Copula	Gumbel Copula	Clayton Copula	Frank Copula
参数估计值	0.5179	0.5827	1.5958	0.8253	3.7651

Table 5. Square Euclidean distance of Copula functions with the highest score and the average score
表 5. 最高分与平均分的 Copula 函数的平方欧式距离

Copula 函数	正态 Copula	t-Copula	Gumbel Copula	Clayton Copula	Frank Copula
距离 d^2	0.1562	0.1325	0.1263	0.2835	0.1492

Table 6. Square Euclidean distance of Copula functions with the lowest score and the average score
表 6. 最低分与平均分的 Copula 函数的平方欧式距离

Copula 函数	正态 Copula	t-Copula	Gumbel Copula	Clayton Copula	Frank Copula
距离 d^2	0.0872	0.0739	0.0571	0.1776	0.0915

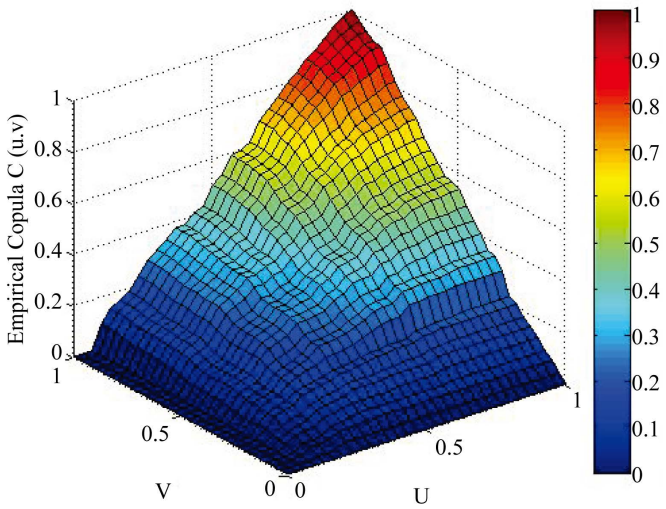

Figure 4. The distribution function of Empirical Copula
图 4. 经验 Copula 分布函数图

我们选用 Gumbel Copula 函数来研究最高分与平均分的相关性。具体地，描述学生成绩最高分与平均分之间相关关系的 Gumbel Copula 函数为

$$\hat{C}^{up}(\hat{u},\hat{v}) = \exp\left\{-\left[(-\ln\hat{u})^{1.5103} + (-\ln\hat{v})^{1.5103}\right]^{1/1.5103}\right\} \tag{2}$$

表 6 表明，Gumbel Copula 函数对最低分与平均分的 Copula 函数的拟合效果也是最好的，因此我们也选用 Gumbel Copula 函数来研究最低分与平均分的相关性。描述学生成绩最低分与平均分之间相关关系的 Gumbel Copula 函数为

$$\hat{C}^{lo}(\hat{u},\hat{v})=\exp\left\{-\left[(-\ln\hat{u})^{1.5958}+(-\ln\hat{v})^{1.5958}\right]^{1/1.5958}\right\} \tag{3}$$

相应的 Gumbel Copula 的密度函数和分布函数图见图 5 和图 6。

接下来我们通过选取的 Gumbel Copula 模型对数据进行相关性分析。

4. 相关性分析

我们知道，通过 Copula 函数不仅可以求变量间的秩相关系数，还可以求尾部相关系数。尾部相关性可以较好地描述极端事件发生时变量间的相互作用，即当一个随机变量大幅度增加或者大幅度减少时，另一个随机变量也发生大幅度增加或者大幅度减少的概率[9]。基于 Gumbel Copula 函数求解 Kendall 秩相关系数的公式为 $\tau=1-\frac{1}{\alpha}$，求解尾部相关系数的公式为 $\hat{\lambda}^{up}=2-2^{1/\alpha}$ 与 $\hat{\lambda}^{lo}=0$，其中 $\hat{\lambda}^{up}$ 与 $\hat{\lambda}^{lo}$ 分别表示上尾与下尾相关系数的估计值。由于 Gumbel Copula 函数具有明显的上尾相关性，因此我们可以通过它对数据进行上尾相关性分析。

将表 3 和表 4 第 4 列的参数估计值 $\hat{\alpha}$ 代入基于 Gumbel Copula 函数求解相关系数的表达式中，得到成绩间的 Kendall 秩相关性与上尾相关性情况见表 7。

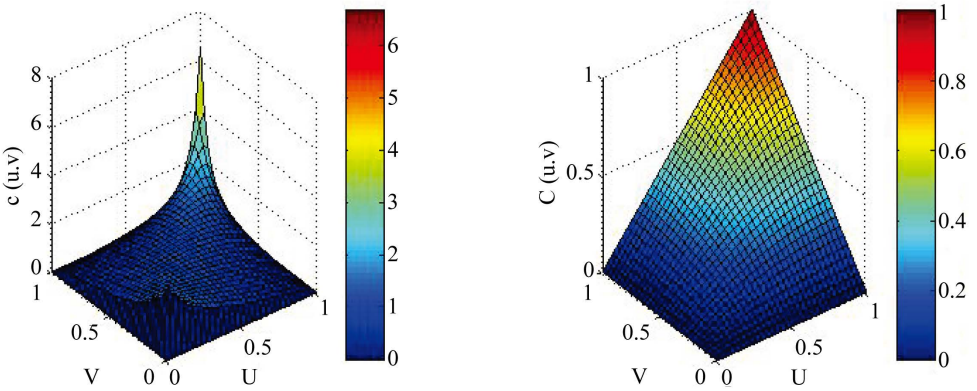

Figure 5. The density function graph and distribution function graph of Gumbel-Copula between the highest score and average score

图 5. 最高分与平均分的 Gumbel-Copula 密度函数与分布函数图

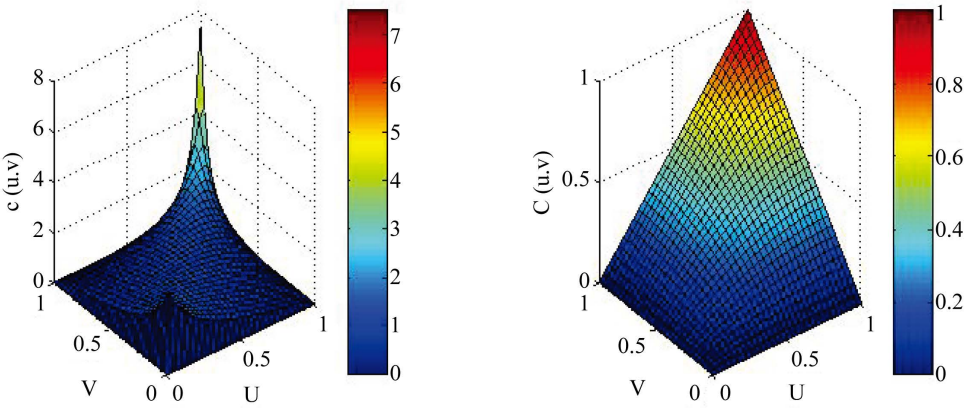

Figure 6. The density function graph and distribution function graph of Gumbel-Copula between the lowest score and average score

图 6. 最低分与平均分的 Gumbel-Copula 密度函数与分布函数图

Table 7. Correlation coefficient based on Gumbel Copula
表 7. 基于 Gumbel Copula 的相关系数

	Kendall 相关系数	上尾相关系数
最高分与平均分	0.3379	0.4176
最低分与平均分	0.3734	0.4560

通过比较基于 Gumbel Copula 函数的 Kendall 秩相关系数可知，最高分与平均分之间的秩相关系数低于最低分与平均分之间的秩相关系数，这说明最低分对平均分的影响程度大于最高分对平均分的影响程度。

另外，由表 7 可知，最高分与平均分之间的上尾相关系数为 0.4176，最低分与平均分之间的上尾相关系数为 0.4560，这说明最高分、最低分与平均分之间均存在着上尾相关性。我们知道，上尾相关性是指变量间同时出现大值的概率[10]，因此，当某个班级的平均分较高时，同时该班级的最高分与最低分也较高的概率分别为 0.4176 与 0.4560。由于最低分与平均分同为较大值的概率高于最高分与平均分同为较大值的概率，说明在平均分较高的情况下，要想继续提高平均分，提高最低分比提高最高分会更加有效。

5. 结论

通过本文对某高校工科学生的线性代数期末考试成绩的研究可知，不论是从描述全局相关性的 Kendall 秩相关系数来看，还是从描述局部相关性的上尾相关系数来看，最低分与平均分之间的相关系数均高于最高分与平均分之间的相关系数，即最低分对平均成绩的影响程度大于最高分对平均成绩的影响程度。

本文的研究结果可以推广到实际生活中去。例如，在教学中，如果想提高一个班级的平均成绩，老师可以重点提高分数较低的同学的成绩；在经济领域中，例如对于北上广这些人均 GDP 较高的比较发达的地区，可以重点通过扶贫来进一步提升地区发展水平。

参考文献 (References)

[1] 张晓宇, 徐付霞. 基于 Copula 理论的学生成绩平均值和中位数的分布特征研究[J]. 大学数学, 2016, 32(1): 56-60.

[2] 梁冯珍, 史道济. 离散型最小和最大次序统计量相关性研究[J]. 应用概率统计, 2008, 24(4): 381-387.

[3] 徐付霞, 董永权. 泥石流地貌要素的极值相关性[J]. 系统工程理论与实践, 2009, 29(2): 180-185.

[4] 李霞. COPULA 方法及其应用[M]. 北京: 经济管理出版社, 2014.

[5] 谢中华. MATLAB 统计分析与应用[M]. 北京: 北京航空航天大学出版社, 2010.

[6] 尹向飞. 基于混合正态分布的大学生考试成绩分布的拟合[J]. 统计与决策, 2007(8): 133-135.

[7] 吴喜之, 赵博娟. 非参数统计[M]. 第 3 版. 北京: 中国统计出版社, 2009: 175-180.

[8] Nelsen, R.B. (2006) An Introduction to Copulas. 2nd Edition, Springer, New York.

[9] Juri, A. and Wüthrich, M.V. (2002) Copula Convergence Theorems for Tail Events. Insurance Mathematics and Economics, **30**, 405-420.

[10] 李悦, 程希骏. 上证指数和恒生指数的 copula 尾部相关性分析[J]. 系统工程, 2006, 24(5): 88-92.