一类模糊数据的相关系数研究

王忠玉[1]　　吴柏林[2]

1. 哈尔滨工业大学经济与管理学院，哈尔滨，150001

2. 应用数学系，台湾政治大学，台北，11605

摘要：在统计学中，经常用（皮尔森）相关系数表达两个变量之间线性关系的强度，并揭示关系方向。相关系数所处理的数据都是明确实数值，但当数据是模糊数据时，如何计算此类模糊数据的相关系数一直困扰着统计学研究者。本文研究当数据为模糊数据比如区间数据时，探讨这种模糊数据相关系数并提出一类模糊数据相关系数的定义，并考察将影响大学生数学成绩的量化因素当成模糊区间数据进行实证研究分析，得出更符合实际情况、合理的预期结果。同时，研究将模糊区间数据相关定义用于两种数据都是实数或其中一组数据是实数的情况，揭示出这种广义模糊相关系数可用于更广泛的应用领域。

关键词：模糊数据；区间数据；广义模糊数据的相关系数

The Study on the Generalized Fuzzy Correlation Coefficient

Zhongyu Wang[1], Berlin Wu [2]

1. School of Economics and Management of Harbin Institute of Technology, Harbin, 150001, China,

2. Department of Applied Mathematics, National Chengchi University, Taiwan, 11605

In statistics, the Pearson correlation coefficient used to represent linear relationship between the two variables, and to reveal the direction of between them. Traditionally, correlation coefficient deals with data which consist of crisp real value. But when the data are composed of fuzzy value, it is not feasible to use this traditional approach to figure out the fuzzy correlation coefficient. The present study investigates the fuzzy samples of interval data to find out the fuzzy correlation coefficient. This paper defines fuzzy correlation with interval data, and proposes broad formulas in order to adjust the coefficient more reasonably and deal with it more accurately. The empirical studies are used to illustrate the application of fuzzy correlations in some applications. Moreover, the formulas derived in this study can be applied to the conditions of either both values of the data are real value or one value of the data is real value. More related practical phenomenon can be explained by this generalized definition of coefficient.

Key words: Fuzzy data, interval data, generalized fuzzy correlation

1、人类认知和思维的模糊性

人类思维主要源自对自然现象和社会活动的认知意识，因而人类知识语言会因本身的主观意识、时间、环境和分析事情的角度不同而具有模糊性[1] [2]。如果想要了解某两个变量如 X 与 Y 两个现象之间的关系程度，一种最直接方法是先将(X, Y)一组数据的散布图画出来。考察X 与 Y 这两个变量之间呈现何种程度的关系，通常画出数据组散布图，查看 X 与 Y 之间的相关性。事实上，任意两个变量之间必定存在某种关系，具体来说包括正相关、负相关或统计无关。因此，测量关系程度的大小则是极为重要的。

在统计学上，使用皮尔森相关系数(Pearson's Correlation Coefficient)表达两个变量间线性关系的强度，同时也表达出关系方向。以往相关系数所处理的数据都是明确实数值，但当

数据是模糊数据比如区间时，就不适合运用传统方法计算模糊相关系数(fuzzy correlation coefficient)。

社会科学领域中，就搜集到的数据而言，尤其是关于人类的认知及自身活动，绝大部分都体现出模糊数据的特征，最近20年多来，许多研究者开始探索如何将模糊数学用于计算数据的类似性和相关性，如Ragin(2000) [3]和Smithson(1987) [4]就曾探究如何将模糊理论用于社会科学，林原宏(2004) [5]提出模糊相关系数即针对模糊性数据，衡量其类似性(similarity)和相关性的系数。

类似性是计算两个模糊数据（或模糊集合）的类似程度，相关性则是计算一组模糊数据样本，每个模糊样本点的两个模糊数据相关性。尽管文献中存在许多不同公式，如Liu与Kao(2002) [6]研究发现，已有模糊数据相关系数仍存在一些问题，有待进一步完善和发展。

本文研究将针对区间模糊样本数据值求得模糊相关系数，将区间型模糊数据分为离散型和连续型，并依据Liu等所提出相关系数方法先求得模糊相关系数，并对相关系数做适当调整，能使所求出相关系数更加精确。此公式也能用于两个数据为实数或其中一个数据值为实数的情况，可解释更多在实际应用中所发生的相关现象。

2、以往相关系数定义的不足之处

如果想了解 X 与 Y 两个现象之间的关系程度，一种最直接方法是，先将(X, Y)的数据散布图画出来。到底 X 与 Y 这两变量之间呈现何种程度的关系，通过数据散布图可以查看它们之间的相关性。事实上，任意两个变量之间必定有关系存在，包括正相关、负相关、或统计无关。因此，测量关系程度的大小是关注焦点。

以往线性相关系数通常用 ρ 表示，代表两个变量 X 与 Y 的相关程度，也就是将相关系数定义为

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \qquad (1)$$

当 ρ > 0 时，称 X 与 Y 之间为直线正相关；当 $\rho < 0$，则称 X 与 Y 之间为直线负相关；当 ρ = 0 时，则称 X 与 Y 为之间没有线性相关存在或统计不相关。不过，要求相关系数，必须知道变异数 $\sigma_X{}^2$、$\sigma_Y{}^2$ 和它们之间的协方差 Cov (X, Y)。在实际应用上，不易得到。因此，用样本相关系数 r_{xy} 估计 ρ，即

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad (2)$$

其中 (x_i, y_i) 表示第 i 对观测值，i=1, 2, 3, …, n； $\bar{x}$ 与 $\bar{y}$ 分别为其样本平均数。

在社会科学研究中尤其是人类活动，区间模糊数据随处可见，在某些情况下，人们无法确定现有信息真正的值是什么。

现在，考察下面的区间模糊数据。区间型模糊数据是一种具有均匀隶属度函数的模糊数据，本文用闭区间符号“[ ]”来表示。当 $a, b \in$R 且 $a < b$，则$[a, b]$表示区间型模糊数据，将 a 称为$[a, b]$下界，将 b 称为$[a, b]$上界；当 a=b，则$[a, b] = [a, a] = [b, b] = a = b$ 表示实数 a（或 b）。同理，实数 k 亦可表示为 $[k, k]$。当$[a, b]$为区间型模糊数据时，设 $c_0 = \frac{a+b}{2}$，

$s_0=\frac{b-a}{2}$ 分别表示其中心及半径，也可将区间型模糊数据表示成：$[c_0;s_0]\Rightarrow[c_0-s_0,c_0+s_0]=[a,b]$。用 l = b-a 表示此区间长度。

举例来说，某个工人的工资年薪为 2.6 至 3.5 万元这样的区间型模糊数，工时 8 至 10 小时也是区间型模糊数，希望知道工资高低和工作时数的相关系数，应该要如何计算呢? 这类区间模糊数据，无法用以往相关系数公式分析。针对上述问题，对于区间型模糊数据，本文提出一类新的模糊数据相关系数公式。

3、一类新的模糊数据相关系数

首先，考虑 (x_i,y_i) 为第 i 对样本值，i=1, 2,…, n；x_i 与 y_i 均表示区间模糊数; $\bar{x}$ 及 $\bar{y}$ 分别表示其样本平均数。当研究的两个变量都是模糊数据时，分别对两个变量取得模糊区间 Ix_λ 与 Iy_λ，如图 1。

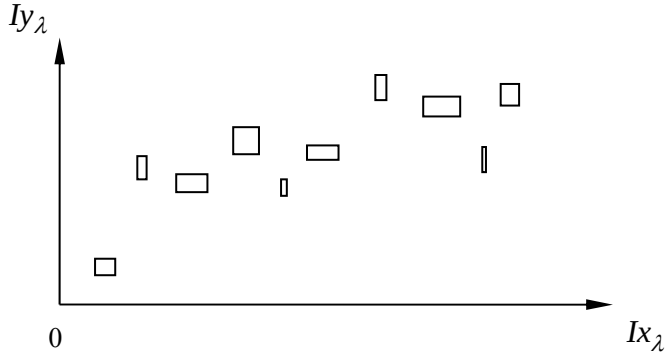

图 1 x_i 与 y_i 均为区间模糊数据的分布图

将区间型模糊数据（区间内均匀分配），分别对两个变量 X, Y 取各个样本区间中心点 x_i、y_i 作为代表值。当模糊数据为区间型模糊数据，用公式 1 分别对两变量 X, Y 取各样本，得模糊数据的重心值 x_i , y_i 当作代表值。针对相关系数值 r_{xy}，再考虑连续区间模糊数长度不一样，或区间隶属度不同，因此必须考虑区间的相关效应。若将两种相关系数等重相加，所得结果的相关系数可能有一边出现大于 1 或小于-1 情况。区间长度相关效应也不应该重于中心点相关效应。因此，为对区间数据进行合理修正相关系数，用公式(3)加以调整，得到更具有合理性的模糊数据相关系数。考虑取以 e 为底的自然对数 *ln* 函数进行转换。设连续区间样本 x_i 的长度 l_{x_i} 连续区间样本 y_i 的长度 l_{y_i}，则修正长度相关系数为

$$\delta=1-\frac{\ln(1+|r_l|)}{|r_l|};\ \text{其中}\ \ r_L=\frac{\sum_{i=1}^{n}(l_{x_i}-\bar{l}_x)(l_{y_i}-\bar{l}_y)}{\sqrt{\sum_{i=1}^{n}(l_{x_i}-\bar{l}_x)^2}\sqrt{\sum_{i=1}^{n}(l_{y_i}-\bar{l}_y)^2}} \qquad (3)$$

由于 $0<r_l<1$，故 δ 范围为 $0<\delta<0.3069$。

定义 1 模糊相关区间(取区间中心点与长度的方法)

设 c_{x_i} , c_{y_i} 表示 X ,Y 总体的模糊样本区间中心点, l_{x_i} , l_{y_i} 表示区间长度。r 是中心点相关系数，δ 表示修正长度相关系数。于是，将相关区间定义为

(i) $r\ge0,\ \ r_l\ge0,\ (r,\ min(1,r+\delta))$

(ii) $r\ge0,\ \ r_l<0,\ (r-\delta,\ r)$

(iii) $r<0,\ r_l \geq 0, (r,\ r+\delta)$

(iv) $r<0,\ r_l<0, (max(-1, r-\delta), r)$

4、实证应用分析

本节给出区间模糊数相关系数的实例应用，在下面例 3.1 中处理(i)两组变量均为实数；(ii)一组变量为连续型等距尺度区间模糊数据，另一组变量为实数，与(iii)两组变量皆为连续型等距尺度模糊数据的情形，例 4.1.3、4.2 与 4.3 均是两组变量皆为连续型等距尺度模糊数据，并利用定义 3.1 公式计算相关系数。

4.1 上网时间与数学成就

考察哈尔滨市某高校大学新生影响数学成绩因素，随机调查 10 名学生，利用模糊问卷的方式[7] ，决定各指标的重要性，问题是探讨学生数学分数与上网时间是否有内在的联系（即关系）。

4.1.1 学生数学分数（x_i）为实数，上网时间（y_i）亦为实数

表 1：数学平均成绩与平均上网时间

学生	数学平均分数 x_i	平均上网时间 y_i
A	88	1.25
B	83	6.75
C	67	4.5
D	92	2
E	45	16.5
F	72	15.5
G	70	16
H	90	2
I	88	1.5
J	83	5
r = - 0.79		

根据皮尔森相关系数计算公式，数学平均成绩和平均上网时间的相关系数为-0.79，也就是学生的平均成绩越高，上网时间越少。

4.1.2 学生数学分数（x_i）为实数，上网时间（y_i）为区间模糊数

假如想要知道数学平均分数和上网时间模糊区间之间的相关系数，这时搜集到的数据如表 2。

表 2：数学平均成绩与一周上网时间的模糊区间

学生	数学平均分数 x_i	上网时间模糊区间
A	88	[1.0, 1.5]
B	83	[6.5, 7.0]
C	67	[4, 5]
D	92	[1.5, 2.5]
E	45	[16, 17]
F	72	[15, 16]
G	70	[15, 17]
H	90	[1, 3]
I	88	[0, 3]
J	83	[4.5, 5.5]
r = -0.79, r_l=0, δ=0 模糊相关系数= -0.79		

在这个例子里，调查“学生每周上网时数的模糊区间”，并纪录“数学平均成绩”，计算相关

系数，因“学生每周上网时数的模糊区间”经过模糊统计，故为一组模糊数据，而“数学平均分数”为一组实数，用定义 3.1 得到模糊相关系数为－0.79，由此可知，当模糊相关系数其中一组为实数时，其相关系数会等于皮尔森相关系数。

4.1.3 学生数学分数（x_i）为区间模糊数，上网时间（y_i）为区间模糊数据

如果研究者搜集到学生一周之间的数学成绩分布，为了方便起见，以每十分为一区间，则表 1 的数学平均分数变为表 3，由此表可计算，数学分数和上网时间的模糊相关系数：

表 3：学生的数学成绩与一周上网时间的模糊区间

学生	数学分数 x_i	上网时间模糊区间
A	[80, 90]	[1.0, 1.5]
B	[80, 90]	[6.5, 7.0]
C	[60, 80]	[4, 5]
D	[90, 100]	[1.5, 2.5]
E	[40, 70]	[16, 17]
F	[70, 80]	[15, 16]
G	[60, 80]	[15, 17]
H	[80, 100]	[1, 3]
I	[80, 90]	[0, 3]
J	[80, 90]	[4.5, 5.5]
r = -0.79, r_l =0.10, δ =0.05		区间模糊相关=(-0.79, -0.74)

利用定义 1 可知，数学分数与上网时间的区间模糊相关为(-0.79, -0.74)，此相关系数呈现高度负相关的关系，也就是数学分数越高，上网时数越少，学生每周上网时数会对数学学业成绩产生负面的影响。

4.2 睡眠时间与数学成就

为了认识“影响数学成绩因素”，将 10 位学生“每天睡眠时间”做模糊问卷调查，并求模糊相关系数，将“每天睡眠时间”指标的问卷结果整理如表 4。

表 4：数学成绩与一周睡眠时间的模糊区间相关

学生	数学分数 x_i	睡眠时间模糊区间
A	[80, 90]	[8, 8.5]
B	[80, 90]	[7, 7.5]
C	[60, 80]	[9, 10.5]
D	[90, 100]	[8, 8.5]
E	[40, 70]	[6, 7.5]
F	[70, 80]	[10, 11]
G	[60, 80]	[7, 8]
H	[80, 100]	[8, 10]
I	[80, 90]	[6.5, 8]
J	[80, 90]	[7.5, 8.5]
r=0.12, r_l =0.61, δ = 0.22		区间模糊相关=(0.12,0.34)

呈现低度正相关的关系，故学生睡眠时数越多，数学成绩越好，但睡眠时间的影响不大，但仍有一些关系。

4.3 睡眠时间与上网时间

现在考察“每天学生睡眠时间”与“学生一周上网时间”是否有相关，两组都是区间模糊数据，经收集整理得到下面模糊区间。如表 5 所示：

表 5 ：睡眠时间模糊区间与上网时间模糊区间

学生	睡眠时间模糊区间	上网时间模糊区间

A	[8,8.5]	[1.0,1.5]
B	[7,7.5]	[6.5,7.0]
C	[9,10.5]	[4,5]
D	[8,8.5]	[1.5,2.5]
E	[6,7.5]	[16,17]
F	[10,11]	[15,16]
G	[7,8]	[15,17]
H	[8,10]	[1,3]
I	[6.5,8]	[0,3]
J	[7.5,8.5]	[4.5,5.5]
r=-0.05, r_I=0.60, δ=0.22		区间模糊相关=(-0.05,0.17)

上表显示出低度相关的关系，代表学生睡眠时数与上网时间并没有非常直接关系。

这里主要探讨影响数学成绩因子，并计算与数学成绩与其它因子之间的模糊相关系数，通过整理得到表 6，从表得知，睡眠时间、上网时间、与学生数学成绩之间的区间相关系数。对于学生数学成绩影响较大的为上网时间，学生上网时间越多，成绩越差。而睡眠时间则和数学成绩呈现低度正相关的情形，代表睡眠时间长，数学成绩会越好，但影响幅度较小。

表 6：睡眠时间、上网时间、与数学成绩之间的区间相关

	睡眠时间	上网时间	数学成绩
睡眠时间	1	(-0.05, 0.17)	(0.12, 0.34)
上网时间		1	(-0.79, -0.74)
数学成绩			1

五、结论

对于以往相关系数来说，由于模糊数据所取得的模糊线性相关系数所传递的相关关系更具有解释能力，本文探索并讨论连续型区间模糊数据，利用本文定义的相关系数，既可计算两组变量都是模糊数据情况，又可计算当两组变量都是实数或其中一组是实数的情况，这时模糊相关系数则退化成皮尔森相关系数，因此，本文提出的相关系数能适用于变量在不同情况的组合。

参考文献

[1] 王忠玉，吴柏林，《模糊数据统计学》[M]。哈尔滨：哈尔滨工业大学出版社，2008。

[2] 王忠玉，吴柏林，模糊数据均值方法及应用研究。[J] 统计与信息论坛, 2010, (25) 10, 13-17。

[3] Regin, C. C., *Fuzzy-Set social science.* Chicago: University of Chicago Press [M]. 2000, pp 4-25.

[4] Smithson, M. *Fuzzy Set analysis for behavioral and social sciences.* [M] New York. Springer-Verlag 1987. pp31-58

[5] 林原宏，模糊相关系数. [J] 教育研究月刊， 2004（第122期）, 122,148-149。

[6] Liu, S. T. , and Kao, C.(2002). Fuzzy measures for correlation of fuzzy numbers. [J] *Fuzzy Sets and Systems*, 128, 267-275.

[7] 王忠玉，吴柏林，模糊数据问卷调查表的设计及应用，[J] 经济研究导刊，2012，第 14 期, 174-178。

基金项目：黑龙江省哲学社科基金专项项目（12D069），中央高校基本科研业务费专项资金

资助项目（HIT. HSS. 201229）
通信地址：
王忠玉　wangzhy@hit.edu.cn
15945693806
哈尔滨工业大学经济与管理学院应用经济系，150001
哈尔滨市西大直街 92 号
王忠玉　简介，哈尔滨工业大学经济与管理学院，副教授、博士后，研究方向：经济计量学、模糊统计学、数理金融学等。

男，1963 年生人，黑龙江哈尔滨人。

吴柏林　简介，台湾政治大学应用数学系，教授、博士，研究方向：时间数列分析与预测、人工智能、模糊统计、市场调查与分析。

男，1956 年生人，台湾高雄人。