Appl. Math. Inf. Sci. 6 No. 2S pp. 439S-444S (2012)

Applied Mathematics & Information Sciences

An International Journal

@ 2012 NSP
Natural Sciences Publishing Cor.

Point-Biserial Correlation Analysis of Fuzzy Attributes

Hao-En Chueh[1]

[1] Department of Information Management, Yuanpei University, 30015 HsinChu, Taiwan
Corresponding author: Email: hechueh@mail.ypu.edu.tw

Received June 28, 2011; Revised September 2, 2011; Accepted 13 September 2011
Published online: 1 March 2012

Abstract: Previous studies have presented correlation analyses among fuzzy attributes. This study proposes a correlation analysis method among crisp attributes and fuzzy attributes to investigate their linear relationships. If A and B are two attributes, A is a nominal-dichotomous attribute, and B is a fuzzy attribute. Thus, identifying the relationship between A and B is possible by using the point-biserial correlation analysis. The fuzzy point-biserial correlation analysis is defined and derived, using the membership grades of fuzzy attributes. Using the fuzzy point-biserial correlation analysis is feasible if A is a bimodal distribution, but not a real nominal-dichotomous attribute, such as crime youth versus non-crime youth, illiterate versus non-illiterate person, and mental deficiency versus normal intelligence. An experimental sample illustrates that our proposed method is simpler and faster than the traditional correlation analysis.

Keywords: Correlation Analysis, Point-Biserial Correlation Analysis, Nominal-Dichotomous Attribute, Fuzzy Attributes

1 Introduction

Correlation analysis of crisp attributes is often used to identify the relationships among attributes in databases. Various correlation analyses, defined on crisp attributes, have been discussed in conventional statistics [1, 5]. However, many recorded attributes in databases may be fuzzy [7, 8, 10] but useful, which cannot be described by crisp attributes, and must be generalized to fuzzy sets, for accurate data representation. Exploring the attributes is also necessary, and therefore, methods to investigate these fuzzy attributes are required.

Correlation analyses among fuzzy attributes are referred to as “fuzzy correlation analyses.” Previous studies [2, 3, 6, 9] have presented three fuzzy correlation analyses (fuzzy simple correlation analysis [2], fuzzy partial correlation analysis [3], and fuzzy semi-partial correlation analysis [6]) on Zadeh’s fuzzy sets.

The simple correlation coefficient between two fuzzy attributes is called a “fuzzy simple correlation coefficient” [2, 9], and is useful in discovering the strength of a relationship between two vague variables, or fuzzy attributes. If two fuzzy attributes are directly related, a positive score emerges. If two fuzzy attributes are inversely related, a negative score emerges. If two fuzzy attributes are not evidently related, the score is close to zero.

Fuzzy simple correlation coefficients provide a comprehensive understanding of the linear relationship between two fuzzy attributes. However, in some practical situations, fuzzy attributes, other than the two in question, are also responsible for the observed relationship, and may influence the relationship between the observed fuzzy attributes. Thus, solving this problem requires analyzing fuzzy partial correlations [3], to show the relationship between two fuzzy attributes, when the influences of other fuzzy attributes are removed from the observed fuzzy attributes.

In numerous fuzzy prediction models, the analysis of the fuzzy semi-partial correlation [6] is useful in choosing the predictor fuzzy attributes, to predict the criterion fuzzy attribute. The relationship between two fuzzy attributes is referred to as “the fuzzy semi-partial correlation analysis,” by the removal of the influence of other fuzzy attributes from one of the interested fuzzy attributes. Such a correlation is more complicated and has larger variety than the previous ones we have presented.

Previous studies have presented correlation analyses among fuzzy attributes. This study proposes a correlation analysis method among crisp attributes and fuzzy attributes to investigate their linear relationships. This paper generalizes the discussion of the linear relationship among fuzzy attributes to the linear relationship among crisp attributes and fuzzy attributes, measured by fuzzy point-biserial correlation coefficients.

If A and B are attributes, A is a nominal-dichotomous attribute, and B is a fuzzy attribute, which enables the utilization of the point-biserial correlation analysis to identify the relationship between A and B. If A is a bimodal distribution, but not a real nominal-dichotomous attribute, such as youth crime versus youth not involved in crime, illiteracy versus literacy, and mental deficiency versus normal intelligence, these attributes all belong to bimodal distribution. These situations also require the possible use of the fuzzy point-biserial correlation analysis.

The membership grades of fuzzy attributes are defined and derived by using the fuzzy point-biserial correlation analysis. The rest of this paper is organized as follows: Section 2 presents several fuzzy correlation analyses, Section 3 illustrates the developed and experimented definition of the fuzzy point-biserial correlation analysis, and finally, Section 4 offers the conclusion.

2 Fuzzy Correlation Analyses

The simple correlation coefficient of fuzzy attributes is referred to as "the fuzzy simple correlation coefficient." Previous studies have proposed numerous methods to evaluate the fuzzy simple correlation coefficient [2, 9].

Suppose fuzzy attributes A and $B \subseteq F$, where F is a fuzzy space. Fuzzy attributes A and B are defined on the domain of a crisp universal set X, with membership functions μ_A and μ_B. Fuzzy attributes A and B can then be expressed as:

$$A = \{(x, \mu_A(x)) \mid x \in X\} \tag{2.1}$$

$$B = \{(x, \mu_B(x)) \mid x \in X\} \tag{2.2}$$

, where μ_A, $\mu_B : X \to [0,1]$.

Assume that $((x_1, \mu_A(x_1), \mu_B(x_1)), (x_2, \mu_A(x_2), \mu_B(x_2)), \cdots, (x_n, \mu_A(x_n), \mu_B(x_n)))$ is a random sample drawn from a crisp universal set X. The fuzzy simple correlation coefficient [2] between fuzzy attributes A and B is then defined by:

$$r_{A,B} = s_{A,B} \Big/ \sqrt{s_A^2 \cdot s_B^2} \tag{2.3}$$

, where

$$s_{A,B} = \frac{\sum_{i=1}^{n} (\mu_A(x_i) - \overline{\mu_A}) \cdot (\mu_B(x_i) - \overline{\mu_B})}{n-1} \tag{2.4}$$

is the covariance of fuzzy attributes A and B. $s_A^2 = s_{A,A}$, $s_B^2 = s_{B,B}$ are the variances of A and B, respectively.

The values of the fuzzy membership functions μ_A and μ_B lie between [0, 1], but the value of the fuzzy simple correlation coefficient $r_{A,B}$ is constrained between [-1, 1], which demonstrates not only the relationship degree between the fuzzy attributes, but also whether these two attributes are positively or negatively related. Properties of the fuzzy simple correlation coefficient are stated as follows [2]:

If $|r_{A,B}|$ is close to 1, then the fuzzy attributes A and B are highly related.

If $|r_{A,B}|$ is close to 0, then the fuzzy attributes A and B are barely related.

If $r_{A,B} > 0$, then the fuzzy attributes A and B are positively related.

If $r_{A,B} < 0$, then the fuzzy attributes A and B are negatively related.

If $r_{A,B} = 0$, then the fuzzy attributes A and B have no relationship at all.

In spite of the fact that values of the fuzzy membership function are constrained between in [0,1], the value of the fuzzy simple correlation coefficient lies between in [-1,1], which will show us not only the degree of the relationship between the fuzzy sets, but also the fact whether these two sets are positively or negatively related.

In several practical situations, fuzzy attributes, other than the two in question, are also responsible for the observed relationship, and these attributes may influence the relationship between the observed attributes. Thus, developing the analysis of the fuzzy partial correlation [3] is required, to show the relationship between two fuzzy attributes when the influences of other fuzzy attributes are partialed out from the observed fuzzy attributes.

Assume that simple correlation coefficients are computed between pairs of the fuzzy attributes A, B, and C, namely, $r_{A,B}$, $r_{B,C}$, and $r_{A,C}$. The fuzzy partial correlation coefficient between fuzzy attributes A and B, when the effects of the fuzzy attribute C on fuzzy sets A and B are partialed out, is defined as:

$$r_{A,B\bullet C}=\frac{r_{A,B}-r_{A,C}\cdot r_{B,C}}{\sqrt{1-(r_{A,C})^2}\cdot\sqrt{1-(r_{B,C})^2}} \quad (2.5)$$

Properties of the fuzzy partial correlation coefficient are stated as follows [3]:

The fuzzy partial correlation coefficient $r_{A,B\bullet C}$ is equal to the fuzzy partial correlation coefficient $r_{B,A\bullet C}$.

The fuzzy partial correlation coefficient $r_{A,B\bullet C}$ or $r_{B,A\bullet C}$ is not defined, if there exists a linear relationship between fuzzy sets A and C, that is, $r_{A,C}=1$. Similarly, $r_{A,B\bullet C}$ or $r_{B,A\bullet C}$ is not defined, if $r_{B,C}=1$.

If $r_{A,C}=0$, then $r_{A,B\bullet C}=r_{B,A\bullet C}$. Similarly, if $r_{B,C}=0$, then $r_{A,B\bullet C}=r_{B,A\bullet C}$.

If $r_{A,C}=0$ and $r_{B,C}=0$, then $r_{A,B\bullet C}=r_{A,B}=r_{B,A}$ $=r_{B,A\bullet C}$.

If there is no relationship between A and B, that is, if $r_{A,B}=0$, this does not necessarily mean that there is no partial relationship between fuzzy sets A and B when the influence of fuzzy set C is removed.

Next, we discuss the fuzzy semi-partial correlation, the correlation between two fuzzy attributes when the influences of other fuzzy attributes are removed from only one of the two interested fuzzy attributes [6].

Assume that the simple correlation coefficients, $r_{A,B}$, $r_{B,C}$ and $r_{A,C}$ of fuzzy attributes A, B and C have been calculated according to formula (2.3). Then the semi-partial correlation coefficient between the fuzzy attributes A and B, with the influence of fuzzy attribute C removed from the fuzzy attribute B, can be calculated as:

$$r_{A,(B\bullet C)}=\frac{r_{A,B}-r_{A,C}\cdot r_{B,C}}{\sqrt{1-(r_{B,C})^2}} \quad (2.6)$$

Similarly, the semi-partial correlation $r_{B,(A\bullet C)}$ can be calculated as:

$$r_{B,(A\bullet C)}=\frac{r_{A,B}-r_{A,C}\cdot r_{B,C}}{\sqrt{1-(r_{A,C})^2}} \quad (2.7)$$

$r_{B,(A\bullet C)}$ illustrates the relationship between the fuzzy attributes A and B, with the influence of the fuzzy attribute C removed from fuzzy attribute A.

Properties of the fuzzy semi-partial correlation coefficient are stated as follows [6]:

There are six different fuzzy semi-partial correlation coefficients among three fuzzy attributes A, B and C, say $r_{A,(B\bullet C)}$, $r_{B,(A\bullet C)}$, $r_{B,(C\bullet A)}$, $r_{C,(B\bullet A)}$, $r_{A,(C\bullet B)}$ and $r_{B,(A\bullet B)}$, although there are only three different fuzzy partial correlation coefficients among the same attributes, say $r_{A,B}$, $r_{B,C}$ and $r_{A,C}$.

The fuzzy semi-partial correlation coefficient, $r_{B,(A\bullet C)}$, is different from $r_{A,(B\bullet C)}$ because of their different meanings. $r_{A,(B\bullet C)}$ means the relationship between the fuzzy attributes A and B with the influence of fuzzy attribute C removed from the fuzzy attribute B, but comparatively, $r_{B,(A\bullet C)}$ means the relationship between the fuzzy attributes A and B with the influence of fuzzy attribute C removed from the other fuzzy attribute, A.

When $r_{B,C}$ =1, the semi-partial correlation coefficient between the fuzzy attributes A and B with the influence of fuzzy attribute C removed from the fuzzy attribute B, $r_{A,(B\bullet C)}$, is not defined. Similarly, when $r_{A,C}$ =1, $r_{B,(A\bullet C)}$ is also not defined.

If there is no linear relationship between A and C, then the relationship between the fuzzy attributes A and B with the influence of fuzzy attribute C removed from the fuzzy attribute B is the same as the relationship between the fuzzy attributes A and B with the influence of fuzzy attribute C removed from both of the two fuzzy attributes A and B. That means, when $r_{A,C}$ =0, $r_{A,(B\bullet C)}=r_{A,B\bullet C}$. Similarly, when $r_{B,C}$ =0, then $r_{B,(A\bullet C)}$ is the same as $r_{B,A\bullet C}$.

When $r_{A,C}$ =0 and $r_{B,C}$ =0, $r_{A,(B\bullet C)}=r_{A,B}=r_{B,(A\bullet C)}$. That is, the two semi-partial correlation coefficients will be the same as the simple correlation coefficient in this particular situation.

If $r_{A,B}$ =0, then whether the influence of fuzzy attribute C is removed from fuzzy attribute A or B, it is not necessarily to discuss the semi-partial relationship between fuzzy attributes A and B.

The value of a fuzzy semi-partial correlation coefficient is usually smaller than value of a fuzzy partial correlation coefficient.

$r_{A,(B\bullet C)}=r_{A,B\bullet C}\sqrt{1-(r_{A,C})^2}$, the value of fuzzy simple correlation coefficient, $r_{A,B}$, lies in $[-1,1]$, and thus the value of $\sqrt{1-(r_{A,C})^2}$ lies in $[0,1]$, so the value of fuzzy semi-partial correlation coefficient, $r_{A,(B\bullet C)}$, is smaller than the value of fuzzy partial correlation coefficient, $r_{A,B\bullet C}$.

Also, we can see that sign of a fuzzy semi-partial correlation coefficient is the same as sign of a fuzzy partial correlation coefficient.

3 Fuzzy Point-Biserial Correlation Analysis

A correlation analysis method among crisp sets and fuzzy sets is proposed. Assume that two attributes, A and B, are present. A is a nominal-dichotomous attribute, where one value is a_1, and another value is a_2. $B \subseteq F$ is a fuzzy attribute, where F is a fuzzy space. The fuzzy attribute B is defined on the domain of a crisp universal set X, with the membership function μ_B. The fuzzy attribute B can then also be expressed with formula (2.2).

Assume that $((x_1, A(x_1), \mu_B(x_1)), (x_2, A(x_2), \mu_B(x_2)), \cdots, (x_n, A(x_n), \mu_B(x_n)))$ is a random sample drawn from a crisp universal set X. The fuzzy point-biserial correlation coefficient [1, 5] between the crisp attribute A and the fuzzy attribute B is then defined by:

$$r_{A,B} = \frac{\overline{\mu_{B1}} - \overline{\mu_{B2}}}{SD} \cdot \sqrt{p_1 \cdot p_2} \tag{3.1}$$

, where $\overline{\mu_{B1}}$ and $\overline{\mu_{B2}}$ are the means of the fuzzy attribute B when $A(x_i) = a_1$ and $A(x_i) = a_2$, respectively. p_1 is the percentage of $A(x_i) = a_1$, and $p_2 = 1 - p_1$ is the percentage of $A(x_i) = a_2$. SD is the standard deviation of the fuzzy attribute B.

Table 1. Degree of intelligence and gender of 15 students

Student	Gender	Degree of Intelligence
x_1	M	0.65
x_2	F	0.70
x_3	M	0.31
x_4	F	0.49
x_5	F	0.80
x_6	M	0.50
x_7	F	0.35
x_8	M	0.10
x_9	M	0.81
x_{10}	F	0.69
x_{11}	F	0.78
x_{12}	M	0.55
x_{13}	F	0.77
x_{14}	F	0.90
x_{15}	M	0.42

For example, let X ={all the students in a university} when a researcher wishes to learn the correlation between the degree of intelligence and student gender of the university in question. All the students on campus cannot be measured, and therefore, a sample of 15 students is taken at random from the campus, $x_1, x_2, \cdots, x_{15}$ [2]. Assume that the crisp attribute A is gender where a_1 is M, a_2 is F, and the fuzzy attribute B is the degree of intelligence. Table 1 shows the degree of intelligence, and the gender of 15 randomly selected students.

$$p_1 = 7/15 = 0.4667, \tag{3.2}$$

$$p_2 = 1 - 8/15 = 0.5334, \tag{3.3}$$

$$\overline{\mu_{B1}} = 3.34/7 = 0.4771, \tag{3.4}$$

$$\overline{\mu_{B2}} = 5.48/8 = 0.6850, \tag{3.5}$$

$$SD = \sqrt{\frac{5.8936 - \frac{(8.82)^2}{15}}{15}} = 0.21717, \tag{3.6}$$

$$r = \frac{0.4771 - 0.6850}{0.21717} \cdot \sqrt{(0.4667) \cdot (0.5334)} = -0.47764. \tag{3.7}$$

If we use 1 to represent M in gender, and use 0 to represent F in gender, and then use the product-moment correlation formula, to compute the correlation coefficient between the degree of intelligence and the sex, we can then obtain the same value.

To test whether the fuzzy point-biserial correlation coefficient is significantly different from 0, the following formula is employed [1, 5, 6]:

$$t = \frac{r - \rho}{\sqrt{\frac{1 - r^2}{N - 2}}} = \frac{-0.47764 - 0}{\sqrt{\frac{1 - (-0.47764)^2}{15 - 2}}} = -1.96021 \tag{3.8}$$

Let the degree of confidence be 95 %, that is, $\alpha = 0.05$, and therefore, according to the critical value from the statistics table, is calculated as follows:

$$t_{\frac{\alpha}{2}(N-2)} = t_{(0.025)(13)} = -2.160. \tag{3.9}$$

The value of r does not fall into the region of rejection, and therefore, should accept the null hypothesis. The test result above shows that no correlation exists between the degree of intelligence and the gender of students in this university.

Another experimental dataset used in this study resulted from the randomly sampled customer retention activities, and the responses of customers, of a telecom company in Taiwan, whose contracts were due to expire between June and July 2008 [4].

Table 2. The results of customer retention activities

CS	EM	BP	MM	EC
1	June	0 ~ 300	TM	0.18
2	June	0 ~ 300	DM	0.22
3	June	301 ~ 800	TM	0.48
4	June	301 ~ 800	DM	0.04
5	July	0 ~ 300	TM	0.17
6	July	0 ~ 300	DM	0.25
7	July	301 ~ 800	TM	0.48
8	July	301 ~ 800	DM	0.04

CS: Customer subgroup
EM: Expired Month of contract
BP: Bill Payment (NT$)
MM: Marketing Method
EC: Degree of willing to extend contract
TM: Telemarketing
DM: Direct mail

Among the customers whose contracts were due to expire in June and July 2008, 400 customers were randomly selected from each of the following groups: customers with monthly bills of NT$0 ~ NT$300 and customers with monthly bills of NT$301 ~ NT$800. Each group of 400 customers was then divided further into two subgroups of 200 customers each. Customer retention marketing programs were implemented using direct mail (indirect marketing) and telemarketing (direct marketing). This telecom company developed marketing programs to retain their customers. During this retention marketing process, customers could choose the marketing programs they wanted. Table 2 displays the results of the entire retention marketing process. The telecom company needs to know the correlation between the degree of "willing to extend contract" and other attributes in Table 2.

Assume that the crisp attribute A is the expired month, where a_1 is June, a_2 is July, and the fuzzy attribute B is the degree of "willing to extend contract".

$$p_1 = 4/8 = 0.5, \tag{3.10}$$

$$p_2 = 1 - 4/8 = 0.5, \tag{3.11}$$

$$\overline{\mu_{B1}} = 0.92/4 = 0.23, \tag{3.12}$$

$$\overline{\mu_{B2}} = 0.94/4 = 0.235, \tag{3.13}$$

$$SD = \sqrt{\frac{0.6362 - \frac{(1.86)^2}{8}}{8}} = 0.159589, \tag{3.14}$$

$$r = \frac{0.23 - 0.235}{0.159586} \cdot \sqrt{(0.5)\cdot(0.5)} = -0.01567 \tag{3.15}$$

Similarly, we can obtain the fuzzy point-biserial correlation coefficient between the degree of "willing to extend contract" and the bill payment. Assume that the crisp attribute C is bill payment, where c_1 is NT$0 ~ NT$300, c_2 is NT$301 ~ NT$800, and the fuzzy attribute B is also the degree of "willing to extend contract".

$$p_1 = 4/8 = 0.5, \tag{3.16}$$

$$p_2 = 1 - 4/8 = 0.5, \tag{3.17}$$

$$\overline{\mu_{B1}} = 0.82/4 = 0.205, \tag{3.18}$$

$$\overline{\mu_{B2}} = 1.04/4 = 0.26, \tag{3.19}$$

$$SD = \sqrt{\frac{0.6362 - \frac{(1.86)^2}{8}}{8}} = 0.159589, \tag{3.20}$$

$$r = \frac{0.205 - 0.26}{0.159586} \cdot \sqrt{(0.5)\cdot(0.5)} = -0.17232 \tag{3.21}$$

Assume that the crisp attribute D is marketing method, where d_1 is telemarketing, d_2 is direct mail, and the fuzzy attribute B is still the degree of "willing to extend contract".

$$p_1 = 4/8 = 0.5, \tag{3.22}$$

$$p_2 = 1 - 4/8 = 0.5, \tag{3.23}$$

$$\overline{\mu_{B1}} = 1.31/4 = 0.3275, \tag{3.24}$$

$$\overline{\mu_{B2}} = 0.55/4 = 0.1375, \tag{3.25}$$

$$SD = \sqrt{\frac{0.6362 - \frac{(1.86)^2}{8}}{8}} = 0.159589, \tag{3.26}$$

$$r = \frac{0.3275 - 0.1375}{0.159586} \cdot \sqrt{(0.5)\cdot(0.5)} = 0.595279 \tag{3.27}$$

The fuzzy point-biserial correlation coefficient between the degree of "willing to extend contract" and expired month is equal to -0.01576; the fuzzy point-biserial correlation coefficient between the degree of "willing to extend contract" and the bill payment is equal to -0.17232; the fuzzy point-biserial correlation coefficient between the degree of "willing to extend contract" and marketing method is equal to 0.595279. According to the above analyses, the most important attribute of the retention marketing process is marketing method.

Previous research [4] used the data mining technique to construct a decision tree based churn retention model and reduce churn rates based on customer responses to retention activities performed by customer service centers. The constructed churn retention model confirms the key factors determined by fuzzy point-biserial correlation analysis.

4 Conclusions

Information regarding the correlations between attributes is a necessity in numerous data analysis tasks. Correlation analyses are commonly used when an analysis of the relationships between database attributes is required. Various correlation analyses among crisp attributes have been widely discussed in conventional statistics. Previous studies have also presented numerous correlation analyses among fuzzy attributes. This study is primarily concerned with correlations among crisp attributes and fuzzy attributes. Most researchers may select the product-moment correlation formula to compute the correlation among crisp sets and fuzzy sets. This paper presents a new method, called "fuzzy point-biserial correlation analysis" to solve this problem, and is a simpler and faster method than traditional correlation analyses.

References

[1] S. F. Arnold, *Mathematical Statistics*, Prentice- Hall, New Jersey. (1990).

[2] D. A. Chiang and N. P. Lin, Correlation of Fuzzy Sets, *Fuzzy Sets and Systems*, Vol. 102. (1999), 221-226.

[3] D. A. Chiang and N. P. Lin, Partial Correlation of Fuzzy Sets, *Fuzzy Sets and Systems*, Vol. 110. (2000), 209-215.

[4] H. E. Chueh, S. C. Lin and N. Y. Jan, Mining the Telecom Marketing Information to Optimizing the Customer Retention Strategies, *ICIC Express Letter*, Part B: Application, Vol. 2, No.1 (2011), 1-6.

[5] S. Dowdy and S. Wearden, *Statistics for Research*, John Wiley & Sons. (1983).

[6] N. P. Lin and H. E. Chueh, Fuzzy Semi-Partial Correlation Analysis, *WSEAS Transactions on Computers*, Vol. 5, No. 12. (2006), 2970-2976.

[7] G. J. Klir and T. A. Folger, *Fuzzy Sets, Uncertainty, and Information*, Prentice-Hall International. (1988).

[8] G. J. Klir and B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice-Hall International. (1995).

[9] C. Yu, Correlation of Fuzzy Numbers, *Fuzzy Sets and Systems*, Vol. 55. (1993), 303-307.

[10] L. A. Zadeh, Fuzzy sets, *Information and Control*, Vol. 8. (1965), 338-353.

Hao-En Chueh received the Ph.D. in Computer Science and Information Engineering from Tamkang University, Taiwan. He is currently an Assistant Professor with the Department of Information Management at Yuanpei University, Taiwan. His research interests are in the areas of fuzzy set theory, data dining, database system and its applications. He is now the Editor In Chief of International journal of Web & Semantic Technology (IJWesT).