Chapter 4 AN INNOVATIVE APPROACH ON FUZZY CORRELATION COEFFICIENT WITH INTERVAL DATA

4.1 INTRODUCTION

Correlation measurement and causality analysis between two variables are very important topics in the science research work. For any two random variables, we use *Person's correlation coefficient* estimator to compute the degree of correlation/dependence from sample data. However, in the practical case, since uncertain or incomplete factors might interfere with the data collection so that the observed single-valued (real number) samples cannot describe the true situations of the sample from the population. To record the characteristics of such events, sometimes we use the interval data to represent the samples. For example, when we consider the daily temperature, we would like to know the temperature range for a day, and we record it as an interval form [*low*, *high*] by the low and high temperatures of the day.

Now the problem comes out: how do we calculate the correlation coefficient of interval data? It seems not easy to extend the evaluation method for *Person's correlation coefficient* straightly via the interval arithmetic. Hence, we need to consider applying the concept of the fuzzy operations to investigate the correlation of two variables. Nevertheless, it is a challenging work to define the correlation for interval data via the concept of fuzzy set theory as well as to evaluate the interrelation of fuzzy sets.

In the literatures, Bustince and Burillo [6], Chiang and Lin [10], Gerstenkorn and Manko [13], Hong and Hwang [17], and Yu [35] have discussed the correlation coefficient with fuzzy numbers. Liu and Kao [21] applied the extension principle of the fuzzy theory to calculate the correlation coefficient for a sample set of n-independent pairs of fuzzy observations. Nguyen and Wu [23] pointed out that intervals have the fuzzy characteristic due to their uncertainty. Hasuike and Ishii [14] considered the future return as a fuzzy number and proposed portfolio selection problems including them. Shinkai [25] proposed the evaluation method based on fuzzy decision. Shimakawa [24] present a proposal of extension fuzzy reasoning method based on the extension principle. Wang, *et al* [29] investigated three fuzzy inference models and then proposed fuzzy reasoning algorithms based on type-2 fuzzy sets.

In this chapter, we propose an innovative approach to calculate the fuzzy correlation coefficient with respect to interval data. The proposed approach considers intervals to have the fuzzy characteristic and then chooses a suitable membership function to find the approximate interval of the correlation coefficient interval. The approximate interval is called as fuzzy correlation coefficient. This approach is different from the traditional method which uses the samples to directly calculate the correlation coefficients, since the correlation coefficient is merely a single value.

The fuzzy correlation coefficient calculated from interval data will not only could explain the data integrity but also provide a *robustic* correlation coefficient which gives an objective description of the correlation between two variables. Finally, we would discuss the proposed fuzzy correlation coefficient algorithms, and also we would demonstrate its practical effectiveness through the experiment analysis of two examples on education study and nature science respectively.

4.2 CORRELATION COEFFICIENT

4.2.1 **Traditional Correlation Coefficient**

The traditional correlation coefficient has been extensively used in many practical applications, such as market survey [4] and the investigations of the correlation between advertising cost and sale volume [21].The correlation coefficient represents the interrelation of two variables, *X* and *Y*, and is generally denoted by ρ in the traditional statistics [3]. In practical applications, the correlation coefficient ρ is generally estimated by the sample correlation coefficient r, which is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{4.1}$$

where (x_i, y_i), i = 1, 2, …, n, is the *i*th sample, $\bar{x} = \sum_{i=1}^{n} x_i \Big/ n$ and $\bar{y} = \sum_{i=1}^{n} y_i \Big/ n$ are the sample means of *X* and *Y* respectively. Observe that $-1 \le r \le 1$. If $r > 0$, *X* and *Y* are positively correlated; if $r < 0$, *X* and *Y* are negatively correlated; and if $r = 0$, *X* and *Y* are uncorrelated.

4.2.2 **Correlation Coefficient Interval**

For every i = 1, 2, …, n, let $\boldsymbol{Iy}_i = [y_{Li}, y_{Ui}]$ be the observed data of variable $\boldsymbol{IY}$,where y_{Li} and y_{Ui} are respectively the lower boundary and upper boundary of $\boldsymbol{Iy}_i$. The operations of interval apply to equation (4.1) seems reasonable. Indeed, the sample correlation coefficient *r* should be an interval. But there often appears a trap of miscalculation in the practical calculations, especially in the case that the element 0 is contained in an interval [15]. As a result, the following will define the criteria for evaluating the correlation coefficient for interval data.

Definition 4.1 *A particular correlation coefficient of (X, **IY**);* r_{pj}

Given j=1,2,3,…., let $\{(x_i, y_{ij}) | \, x_i \in R, y_{ij} \in \boldsymbol{IY}_i \text{ and } i = 1,2,...,n.\}$ *be a collection of sample data of (X, **IY**), then the particular correlation coefficient of (X, **IY**) is given by*

$$r_{pj} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_{ij} - \bar{y}_j)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_{ij} - \bar{y}_j)^2}}, \tag{4.2}$$

where $\bar{x} = \sum_{i=1}^{n} x_i \Big/ n$, $\bar{y}_j = \sum_{i=1}^{n} y_{ij} \Big/ n$.

A particular correlation coefficient, r_{pj}, of (*X*,***IY***) indicates there are many correlation coefficient values. When we gather all correlation coefficients, such as $\left\{ r_{pj} \middle| \; j = 1, 2, ... \right\}$, the definition of correlation coefficient interval can be defined as follows.

Definition 4.2 *Correlation coefficient interval of (X, **IY**); **r***

Let $r_L = \inf_j \left\{ r_{pj} \middle| j = 1, 2, 3, \ldots \right\}$ *and* $r_U = \sup_j \left\{ r_{pj} \middle| j = 1, 2, 3, \ldots \right\}$*, then the correlation coefficient interval of (X, IY) is given by* $\boldsymbol{r} = [r_L, r_U]$.

Property 4.1 *If* ***r*** *=* $[r_L, r_U]$ *is defined as in Definition 3.2, then* ***r*** *is a subinterval of* $[-1, 1]$.

<u>*Proof*</u>: By Cauchy-Schwarz inequality, we obtain

$$\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_{ij} - \bar{y}_j)\right)^2 \le \sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 ,$$

for $x_i \in R, y_{ij} \in \boldsymbol{IY}_i$, $i = 1,2,..., n$, and j = 1, 2, ….

Because $r_{pj}{}^2 = \dfrac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_{ij} - \bar{y}_j)\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2} \le 1$,

then $-1 \le r_{pj} \le 1$ for every j and $-1 \le r_L \le r_U \le 1$.

Consequently, $\boldsymbol{r} = [r_L, r_U] \subseteq [-1, 1]$. ■

According to the operations of interval for inequality (*c.f.* Introduction to interval FAQ from Dominque Faudot), we have the next definition., we have the next definition.

Definition 4.3 *Positive correlation and negative correlation*

Let $\boldsymbol{r} = [r_L, r_U]$ *be a correlation coefficient interval.*

(1) If $r_U > r_L \ge 0$*, then there is a positive correlation between X and **IY**.*

(2) If $r_L < r_U \le 0$*, then there is a negative correlation between X and **IY**.*

(3) If $r_L = r_U = 0$*, then there is no correlation between X and **IY**.*

(4) If $r_L < 0 < r_U$*, then we cannot judge whether X and **IY** have correlations.*

Example 4.1

A healthcare center provides blood pressure data recorded by monitoring 10 patients of different ages for a year, as listed in Table 4.1. Age (x_i) is a scalar variable, blood pressure ($\boldsymbol{IY_i}$) is an interval variable, and $\{y_{ij} \in \boldsymbol{IY}_i | i = 1,2,...,10\}$, j = 1, 2, 3, 4. are four sets of real-valued data from interval data $\{\boldsymbol{IY}_i | i = 1,2,...,10\}$.

Table 4.1 The ages and the blood pressure data of 10 patients

Patient	***Age*** **(x_i)**	***Blood pressure*** **($\boldsymbol{IY_i}$)**	y_{i1}	y_{i2}	y_{i3}	y_{i4}
1	35	[110,124]	110	114	117	124
2	36	[106,128]	106	113	117	128
3	43	[110,128]	110	115	119	128
4	48	[108,132]	108	115	120	132
5	56	[112,144]	112	122	128	144
6	61	[123,153]	123	132	138	153
7	67	[131,151]	131	137	141	151
8	70	[135,155]	135	141	145	155
9	74	[149,163]	149	153	156	163
10	80	[140,166]	140	148	153	166

In Table 4.1, y_{i1} in the fourth column is the lower boundary of $\boldsymbol{IY_i}$. The upper boundaries of $\boldsymbol{IY_i}$ are y_{i4} which is listed in the seventh column. In the sixth column, y_{i3} is the middle points of $\boldsymbol{IY_i}$. As well as y_{i2} in the fifth column are values lying between y_{Li} and y_{Ui}.

From Definition 4.1, we obtain $r_{p1} = 0.92$, $r_{p2} = 0.95$, $r_{p3} = 0.97$, and $r_{p4} = 0.98$. All of them are positive correlations. It should be noted that the calculation of the particular correlation coefficient r_{pj} is dependent on $\{(x_i, y_{ij}) \in (X, \boldsymbol{IY}_i) | i = 1, 2, \ldots, 10.\}$ for j = 1, 2, 3, 4. We only list four groups of real values from interval data. ♦

4.3 FUZZY CORRELATION COEFFICIENT

The use of fuzzy sets, as mathematical models for fuzzy data, and their associated applications spreads out to many fields in science and technology, especially in computational intelligence. Here we only have statistics in mind. Consequently, a basic definition on fuzzy number is present as Definition 2.1. Fuzzy numbers are special fuzzy quantities. In nature, an interval is a fuzzy number. Dealing with fuzzy data problems from fuzzy viewpoint was first proposed by Tanaka, *et al* [27]. There are generally two data types of fuzzy samples. One is represented by a real-valued variable pair (*x*, *y*), where *x* is a real value and *y* is the observed real value. The other is represented by a fuzzy variable pair (*x*, *Y*), where *Y* is the observed fuzzy number instead of a real number. This chapter synthesizes the above two data types to represent fuzzy samples by variables (*X*, ***IY***), where *X* denotes a real value and ***IY*** denotes an interval data. By use of the fuzzy characteristic of interval, we can calculate the correlation coefficients of interval data so as to understand the interrelation between two variables.

4.3.1 An approximation approach for a correlation coefficient interval

It is an interesting problem how to find r_L and r_U in Definition 4.2. As described in Example 4.1, a correlation coefficient r_{pj} can be evaluated as long as $\{(x_i, y_{ij}) | \, i = 1,2,\ldots,n\}$ is given. Moreover, determining y_{ij} from IY_i for every i = 1, 2, …, n, is the most important task. But it is impossible to determining y_{ij} from IY_i in view of the fact computing correlation coefficient interval $\boldsymbol{r} = [r_L, r_U]$ is difficult. Therefore, an approximation is proposed for estimating $\boldsymbol{r}$. Due to its fuzzy characteristic, the approximation is called the fuzzy correlation coefficient.

Since intervals have the fuzzy property, let $\boldsymbol{IY}_{i,\alpha} = [y_{Li,\alpha}, y_{Ui,\alpha}]$ be the α–level set $\boldsymbol{IY}_{i,\alpha}$ of $\boldsymbol{IY}_i$ for $\alpha \in [0,1]$ and i =1, 2, …, n. When the membership function of the interval data $\boldsymbol{IY}_i$ is decided, such as Z-type, Λ-type, Π-type or S-type, $\boldsymbol{IY}_{i,\alpha} = [y_{Li,\alpha}, y_{Ui,\alpha}]$ is attained easily as the parameter α is given. About the selection of the membership functions, Zimmermann [37] proposed several types of the membership functions for references. Next, we select $0 \le \lambda_j \le 1$ and let $y_{i,\alpha}(\lambda_j) = (1-\lambda_j)y_{Li,\alpha} + \lambda_j y_{Ui,\alpha}$ for the reason that an interval should be convex, then $y_{i,\alpha}(\lambda_j) \in \boldsymbol{IY}_{i,\alpha}$. The particular correlation coefficient in Definition 4.1 can be rewritten as

$$r_\alpha(\lambda_j) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_{i,\alpha}(\lambda_j) - \bar{y}_\alpha(\lambda_j))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_{i,\alpha}(\lambda_j) - \bar{y}_\alpha(\lambda_j))^2}}, \tag{3}$$

where $\bar{x} = \sum_{i=1}^{n} x_i \big/ n$ and $\bar{y}_\alpha(\lambda_j) = \sum_{i=1}^{n} y_{i,\alpha}(\lambda_j) \big/ n$. Therefore, the fuzzy correlation coefficient for (*X*, ***IY***) can be calculated as follows.

Algorithm of calculating the fuzzy correlation coefficient

Step 1: *Set up $0 \le \alpha_1 < \alpha_2 < \ldots < \alpha_s \le 1$ according to the membership function of the interval data **IY**$_i$, i =1, 2, ..., n.*

Step 2: *For each i =1, 2, ..., n, determine the α–level sets $\boldsymbol{IY}_{i,\alpha_t}$, t = 1, 2, ..., s.*

Step 3: *Choose k+1 points from* [0, 1] *such that* $0 = \lambda_0 < \lambda_1 < \lambda_2 < \ldots < \lambda_k = 1$, *then calculate* $y_{i,\alpha_t}(\lambda_j) = (1-\lambda_j)y_{Li,\alpha_t} + \lambda_j y_{Ui,\alpha_t}$.

Step 4: *Compute* $r_{\alpha_t}(\lambda_j)$ *by Eq.* (3).

Step 5: *Let* $r_{L,\alpha_t} = \inf_j \{ r_{\alpha_t}(\lambda_j) | j = 0,1,2,\ldots,k. \}$ *and* $r_{U,\alpha_t} = \sup_j \{ r_{\alpha_t}(\lambda_j) | j = 0,1,2,\ldots,k. \}$.

Step 6: *The fuzzy correlation coefficient which is denoted as* $\hat{\boldsymbol{r}}$ *is given by* $\hat{\boldsymbol{r}} = \bigcup_{t=1}^{s}[r_{L,\alpha_t}, r_{U,\alpha_t}] = [\inf_t r_{L,\alpha_t}, \sup_t r_{U,\alpha_t}]$.

Remark 4.1

It is trivial that $\hat{\boldsymbol{r}} \subseteq \boldsymbol{r}$ from Definition 4.1 and Definition 4.2.

4.3.2 **Comparison with the Traditional Correlation Coefficient**

Apparently, every sample (x_i, y_{ij}) will fall inside the interval sample (x_i, $\boldsymbol{IY}_i$). Thus, the traditional correlation coefficient r calculated by real-valued samples $\{(x_i, y_i) \mid i = 1,2,..., n.\}$ must belong to the correlation coefficient interval $\boldsymbol{r}$ which is obtained by the interval samples $\{(x_i, \boldsymbol{IY}_i) \mid i = 1, 2, ..., n.\}$, i.e. $r \in \boldsymbol{r} = [r_L, r_U]$. However, the correlation coefficient r may not fall inside the fuzzy correlation coefficient $\hat{\boldsymbol{r}} = [\inf_t r_{L,\alpha_t}, \sup_t r_{U,\alpha_t}]$. Since $\hat{\boldsymbol{r}}$ is calculated by the subset of real-valued samples $\{(x_i, y_{i,\alpha_t}(\lambda_j)_i) \mid j = 0, 1, 2..., k; i = 1, 2, ..., n.\}$ which are chosen by every α–level set $\boldsymbol{IY}_{i,\alpha_t} = [y_{Li,\alpha_t}, y_{Ui,\alpha_t}]$ for $\alpha_t \in [0, 1]$, $\hat{\boldsymbol{r}}$ is a conservative approximate correlation coefficient interval. Accordingly, it could happen that r is not contained in $\hat{\boldsymbol{r}}$. Nevertheless, evaluating the correlation coefficient for interval data is more objective than calculating the conventional correlation coefficient for real-valued data. From the definitions defined above, we can make several important conclusions between the fuzzy correlation coefficient and the traditional correlation coefficient as follows.

Subconclusion 1:

For two highly correlated variables, both the distributions of numeric data and interval data exhibit a high degree of correlation and the correlation coefficient of these two data type are close to ±1. That is, $|r| \approx 1$, $|\inf_t r_{L,\alpha_t}| \approx 1$ and $|\sup_t r_{U,\alpha_t}| \approx 1$. Example 4.1 obviously illustrates a highly correlated case.

Subconclusion 2:

When $\inf_t r_{L,\alpha_t} \le r \le \sup_t r_{U,\alpha_t}$, $\hat{\boldsymbol{r}}$ can suggest which point within the interval has the highest correlation coefficient $\sup_t r_{U,\alpha_t}$ (or the smallest correlation coefficient $\inf_t r_{L,\alpha_t}$). Therefore, the fuzzy correlation coefficient $\hat{\boldsymbol{r}}$ can be considered as a more *robustic* correlation estimator between two variables.

Subconclusion 3:

When $0 < \inf_t r_{L,\alpha_t} \le \sup_t r_{U,\alpha_t} < r$ or $r < \inf_t r_{L,\alpha_t} \le \sup_t r_{U,\alpha_t} < 0$, we had better reexamine the sample data and then design a new membership function for the interval data to compute the fuzzy correlation coefficient $\hat{r}$. On the other hand, we will check the drawbacks of traditional definition of correlation estimator, say if we should use the nonparametric correlation technique, for example *Spearman correlation coefficient* to estimate the more *robustic* correlation value.

4.4 EMPIRICAL STUDIES

In this section, two examples are provided for illustrating how to evaluating the fuzzy correlation coefficient as well as what are the differences between the fuzzy correlation coefficient and the traditional correlation coefficient.

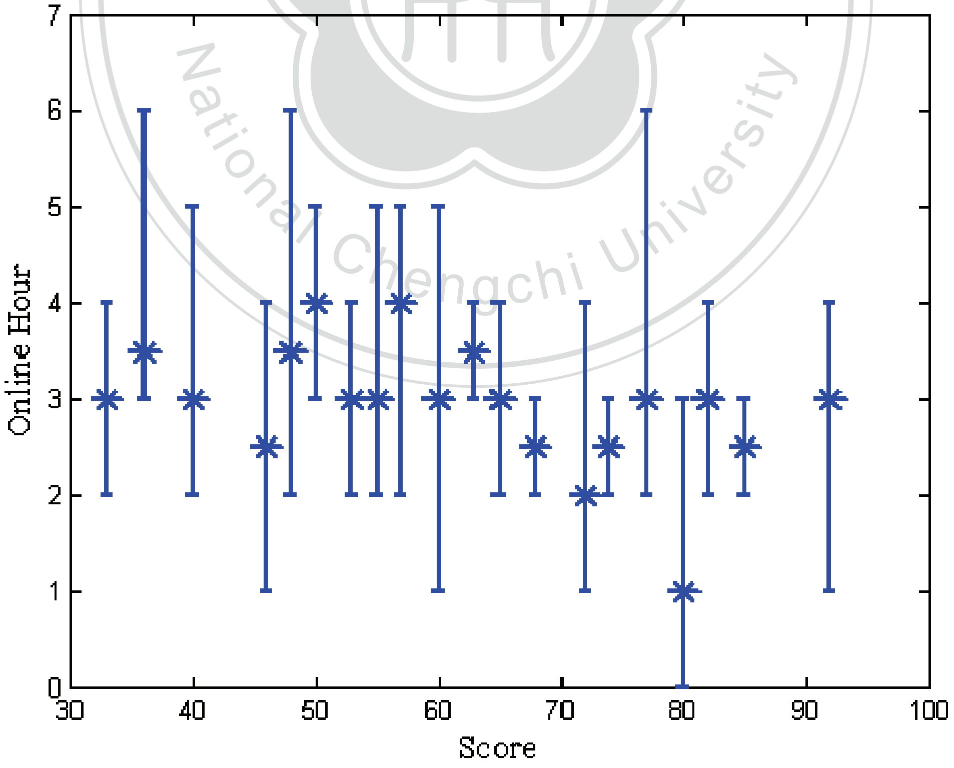

Figure 4.1 Two kinds of data for score and online hour.

4.4.1 **Calculus Score and Online Hour**

This example illustrates whether the Calculus scores have relations with the online hours. In the Department of Information Management, China University of Technology, 20 students who are never absent from the Calculus class before the mid-term exam are asked to fill out two kinds of surveys, one is to write down the range of online hours and the other is to mark the average of online hours per day. Hence, there are two kinds of data: interval data and real-valued data. Figure 4.1 shows two types of data for the Calculus mid-term scores. The blue lines are the interval data and the star marks are the real-valued data.

We select Λ-type and obtain $\boldsymbol{IY}_{i,\alpha_t} = [y_{Li,\alpha_t}, y_{Ui,\alpha_t}]$ for $\alpha_t = 0.0, 0.1, 0.3, 0.5, 0.7, 0.9$. Next, $\lambda_0 = 0$, $\lambda_1 = 0.1$, $\lambda_2 = 0.2$, …, $\lambda_9 = 0.9$, $\lambda_{10} = 1$ are picked, and $r_{\alpha_t}(\lambda_j)$, $j = 0, 1, 2, \ldots, 10$, are calculated by Eq. (3). Table 2 lists the values of $r_{\alpha_t}(\lambda_j)$ and $[r_{L,\alpha_t}, r_{U,\alpha_t}]$. Moreover, by the algorithm of calculating the fuzzy correlation coefficient, we can obtain the fuzzy correlation coefficient $\hat{\boldsymbol{r}} = \bigcup_{t=1}^{s}[r_{L,\alpha_t}, r_{U,\alpha_t}] = [-0.55, -0.40]$ in Table 4.2. With the same argument, we pick S-type and record the values of $r_{\alpha_t}(\lambda_j)$ and $[r_{L,\alpha_t}, r_{U,\alpha_t}]$ in Table 4.3, then $\hat{\boldsymbol{r}} = [-0.55, -0.40]$. Since $r_L = -0.55 < 0$ and $r_U = -0.40 < 0$ by Definition 4.3, there is a negative correlation between the Calculus scores and the online hours.

In the case of the traditional survey whose data are real-valued, the sample correlation coefficient *r* is –0.44 by using Eq. (1). While the correlation coefficient *r* is compared to the fuzzy correlation coefficient $\hat{\boldsymbol{r}}$, $\hat{\boldsymbol{r}}$ can demonstrate a more *robustic* negative correlation coefficient than *r*. It implies the Calculus scores and the online hours should have some essential relations which cannot be conveyed by the traditional correlation coefficient *r*.

In Table 4.2, the correlation coefficient is $r_{\alpha_t}(\lambda_j) =$ –0.55 for λ_j = 0.6, 0.7 and α_t =

0.5, 0.7, 0.9. As $\alpha_t = 0.9$ and $\lambda_j \geq 0.6$, all of the correlation coefficients are −0.55. Because of $y_{i,\alpha_t}(\lambda_j) = (1-\lambda_j)y_{Li,\alpha_t} + \lambda_j y_{Ui,\alpha_t}$, for any i = 1, 2, ⋯, 20, the observed data $y_{i,\alpha_t}(\lambda_j)$, whose distribution is deviated to the right of $\boldsymbol{IY}_{i,\alpha_t}$, has the smallest correlation coefficient when Λ-type is chosen.

Table 4.2 Fuzzy correlation coefficient of score and online hour for Λ−type

λ_j / α_t	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	$[r_{L,\alpha_t}, r_{U,\alpha_t}]$
0.0	-.40	-.44	-.48	-.52	-.54	-.54	-.54	-.54	-.53	-.50	-.47	[-.54, -.40]
0.1	-.43	-.48	-.50	-.52	-.53	-.54	-.54	-.53	-.52	-.52	-.50	[-.54, -.43]
0.3	-.47	-.49	-.51	-.53	-.54	-.54	-.55	-.54	-.53	-.53	-.52	[-.55, -.47]
0.5	-.50	-.51	-.53	-.54	-.54	-.54	-.55	-.55	-.54	-.54	-.54	[-.55, -.50]
0.7	-.53	-.53	-.54	-.54	-.54	-.54	-.55	-.55	-.54	-.54	-.54	[-.55, -.53]
0.9	-.54	-.54	-.54	-.54	-.54	-.54	-.55	-.55	-.55	-.55	-.55	[-.55, -.54]

Table 4.3 Fuzzy correlation coefficient of score and online hour for S−type

λ_j / α_t	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	$[r_{L,\alpha_t}, r_{U,\alpha_t}]$
0.0	-.40	-.44	-.48	-.52	-.54	-.54	-.54	-.54	-.53	-.50	-.49	[-.54, -.40]
0.1	-.44	-.48	-.51	-.53	-.54	-.55	-.54	-.53	-.52	-.51	-.49	[-.55, -.44]
0.3	-.52	-.53	-.54	-.55	-.54	-.54	-.53	-.52	-.51	-.50	-.49	[-.55, -.49]
0.5	-.54	-.55	-.54	-.54	-.54	-.53	-.52	-.52	-.51	-.50	-.49	[-.55, -.49]
0.7	-.54	-.53	-.53	-.52	-.52	-.52	-.51	-.51	-.50	-.50	-.49	[-.54, -49]
0.9	-.51	-.51	-.50	-.50	-.50	-.50	-.50	-.50	-.49	-.49	-.49	[-.51, -.49]

While the S-type is decided, the correlation coefficient −0.55 appears only at (α_t, λ_j) = (0.1, 0.5), (0.3, 0.3), (0.5, 0.5) in Table 4.3. It is hard to determine the tendency of the observed data in the intervals $\boldsymbol{IY}_{i,\alpha_t}$, i = 1, 2, ⋯, 20. So, the S-type is not a good choice for $\boldsymbol{IY}_{i,\alpha_t}$, $i = 1, 2, \cdots, 20$. Whatever Λ-type or S-type is selected, all the fuzzy correlation coefficients are [−0.55, −0.40]. It implies the scores have a negative correlation with the online time. When the online time is longer, the calculus score is lower.

In the correlation analysis for calculus score and online time, there is an evident difference between the fuzzy correlation coefficient calculated from the interval data of online time and the correlation coefficient from the average online time. The former presents a stronger negative correlation. Because the average online time is restricted to the form of single numerical values, they cannot totally represent the difference of daily online time. For example, the online time on holidays is normally longer than that on school days. Due to some reasons (such as not to be divulged), the average online time is deliberately underestimated. As a result, these artificial factors cause that the correlation of online time and calculus score is too low ($r = -0.44$). On the other hand, the online time in the form of intervals offers a more flexible record. Due to its fuzzy characteristic, more important information is retained to recover the deficiency of the average online time. Therefore, the fuzzy correlation coefficient shows a higher negative correlation of online time and calculus scores ($r_L = -0.55$).

4.4.2 **Temperature and Relative Humidity**

In the weather report, it is usually heard that tomorrow temperature in Taipei is 15 to 18 degrees, relative humidity 70%. Everybody knows the meaning of the temperature, but what is the relative humidity? Relative humidity is a term used to describe the amount of water vapor that exists in a gaseous mixture of air and water. Relative humidity is defined as the amount of water vapor in a sample of air compared to the maximum amount of water vapor the air can hold at any specific temperature in a form of 0 to 100%. Relative humidity is an important metric used in weather forecasts and reports, as it is an indicator of the likelihood of precipitation, dew, or fog.

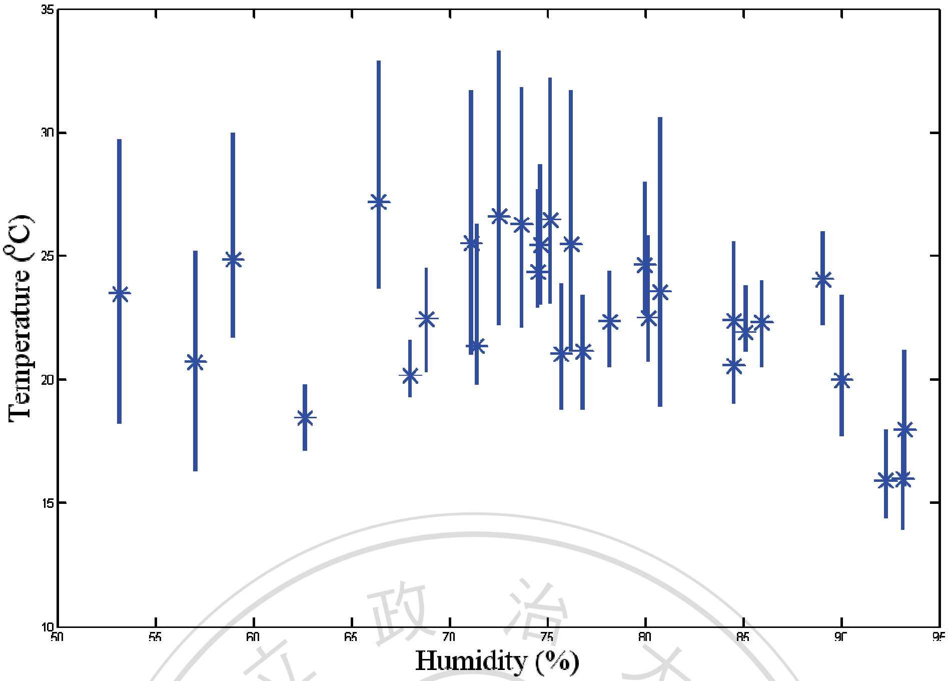

Figure 4.2 Temperature and relative humidity in April 2008.

In hot summer weather, it seems to increase the apparent temperature to humans (and other animals) by deterring the evaporation of perspiration from the skin as the relative humidity rises. However, is the temperature actually related to the relative humidity? If so, how is the temperature correlated to the relative humidity? Is the temperature higher when the relative humidity is higher? So we collect the data from the Central Weather Bureau to find the correlation between the temperature and the relative humidity. The interval data are composed of the daily highest temperature and the lowest temperature in Taipei from April 1, 2008 to April 30, 2008 which are described as in Section 2.5.2. And the daily average temperatures are real-valued data. The graph of these two types of data with the relatively humidity is shown in Figure 4.2. The solid lines symbolize the interval data of temperatures and the star marks represent the daily average temperatures.

S-type is selected to record the values of $r_{\alpha_t}(\lambda_j)$ and $[r_{L,\alpha_t}, r_{U,\alpha_t}]$ in Table 4.4, then $\hat{\boldsymbol{r}} = [-0.44, -0.24]$. Because of $r_L = -0.44 < 0$ and $r_U = -0.24 < 0$, there is a negative correlation between the temperature and the relatively humidity by Definition 4.3. And the temperature whose distribution is deviated to the right of interval data has the smallest correlation coefficient −0.44 when S-type is chosen in Table 4.4. Namely, when temperature is higher, there exists a more notable negative correlation between temperature and relative humidity. Because r_L = −0.44 is close to zero, the temperature and the relative humidity have a weak negative correlation. While we use the traditional survey to evaluate the correlation coefficient of the average temperature and relatively humidity, the correlation coefficient r is −0.40 which is closer to zero than r_L is. Therefore, it means that the correlation of the temperature and the relative humidity is not significant.

Table 4.4 Fuzzy correlation coefficient of temperature and relative humidity for S–type

λj αt	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	$[r_{L,\alpha_t}, r_{U,\alpha_t}]$
0.0	-.24	-.28	-.32	-.35	-.37	-.39	-.41	-.42	-.43	-.44	-.44	[-.44, -.24]
0.1	-.28	-.31	-.34	-.36	-.38	-.40	-.41	-.42	-.43	-.44	-.44	[-.44, -.28]
0.3	-.35	-.36	-.38	-.39	-.40	-.41	-.42	-.43	-.43	-.44	-.44	[-.44, -.35]
0.5	-.39	-.40	-.41	0.41	-.42	-.43	-.43	-.43	-.44	-.44	-.44	[-.44, -.39]
0.7	-.42	-.42	-.43	-.43	-.43	-.43	-.44	-.44	-.44	-.44	-.44	[-.44, -42]
0.9	-.44	-.44	-.44	-.44	-.44	-.44	-.44	-.44	-.44	-.44	-.44	[-.44, -.44]

In the correlation analysis of temperature and relative humidity, the traditional correlation coefficient and the fuzzy correlation coefficient both infer a low correlation between two variables. In spite of this, the fuzzy correlation coefficient based on the S-type membership function could identify the degree of correlation between temperature and relative humidity. While the temperature is closer to the upper bound of the temperature range, the negative correlation of temperature and relative humidity is higher ($r_L = -0.44$). Contrarily, when it is closer to the lower bound of the temperature range, the negative correlation of temperature and relative humidity is lower ($r_U = -0.24$). Even though the correlation of these two variables is not high, the fuzzy correlation coefficient still could demonstrate the dynamic variation of the correlation between these two variables. This feature is the advantage of the fuzzy correlation coefficient.

4.5 CONCLUSIONS

Since the fuzzy correlation coefficient is the approximate interval of the correlation coefficient interval, the degree of relevance between two variables could be explained by the maximum and minimum correlation coefficients from the fuzzy correlation coefficient. Using the conventional correlation coefficient to explain the degree of correlation between two variables is simple and clear. However, it would be too subjective since the correlation between two variables is decided by a single real number while many random and ambiguous situations exist in a real case.

It cannot be denied that the traditional correlation coefficient is easy to show the degree of correlation between two variables. But the procedure of collecting numeric data could be affected by some unpredictable factors so as to impair the accuracy of the correlation coefficient. If we exploit this artificial accuracy to do causal analysis, it may lead to the deviation of causal judgment, the misleading of decision. This chapter proposes to use the interval data to avoid such risks to happen and discusses how to evaluate the correlation coefficient interval by means of the fuzzy aspect and the approximation method.

Some practical examples were given to compare the fuzzy correlation coefficient and the traditional correlation coefficient. And it is found that the fuzzy correlation coefficient provides a more objective judgment than the traditional correlation coefficient. The sample data should be collected with more care when using the fuzzy correlation coefficient. Although the approach proposed in this chapter performs the correlation coefficient for interval data, there are some problems still remained to be solved and some improvement can be done for further researches, which is described respectively as follows.

1. As a matter of fact, the correlation coefficient is a statistic for measuring the linear relationship of two variables X and Y. If the correlation coefficient is close to 0, we can only say there is no linear relationship between them. It usually can use a transformation to adjust the relationship of X and Y. The same argument may apply to X and $\boldsymbol{IY}$. How to transform the interval variable $\boldsymbol{IY}$ is an important issue for improving survey of the correlation coefficient interval.

2. Since there are so many observed data $y_i \in \boldsymbol{IY}_i$, $i = 1, 2, \ldots, n$, y_i can be considered as a random variable in $\boldsymbol{IY}_i$. If the distribution of $\boldsymbol{IY}_i$ can be determined, it will help to find further methods for evaluating the correlation coefficient interval.

3. The types of membership functions are various. Appropriate membership functions can advance the accuracy of the fuzzy correlation coefficient. Consequently, the choice of membership functions will be a worthy study.