Kolmogorov-Smirnov Two Sample Test with Continuous Fuzzy Data

Pei-Chun Lin, Berlin Wu, and Junzo Watada

Abstract. The Kolmogorov-Smirnov two-sample test (K-S two sample test) is a goodness-of-fit test which is used to determine whether two underlying one-dimensional probability distributions differ. In order to find the statistic pivot of a K-S two-sample test, we calculate the cumulative function by means of empirical distribution function. When we deal with fuzzy data, it is essential to know how to find the empirical distribution function for continuous fuzzy data. In our paper, we define a new function, the weight function that can be used to deal with continuous fuzzy data. Moreover we can divide samples into different classes. The cumulative function can be calculated with those divided data. The paper explains that the K-S two sample test for continuous fuzzy data can make it possible to judge whether two independent samples of continuous fuzzy data come from the same population. The results show that it is realistic and reasonable in social science research to use the K-S two-sample test for continuous fuzzy data.

1 Introduction

The Kolmogorov-Smirnov two sample test (K-S two-sample test) is a goodness-of-fit test, which is used to determine whether the two underlying distributions differ. It

Pei-Chun Lin
Graduate School of Information, Production and Systems, Waseda University, 2-7 Hibikino, Wakamatsu, Kitakyushu 808-0135 Japan
e-mail: peichunpclin@gmail.com

Berlin Wu
Department of Mathematical Sciences, National Chengchi University, Taipei, 116, Taiwan
e-mail: berlin@nccu.edu.tw

Junzo Watada
Graduate School of Information, Production and Systems, Waseda University, 2-7 Hibikino, Wakamatsu, Kitakyushu 808-0135 Japan
e-mail: watada@waseda.jp

V.-N. Huynh et al. (Eds.): Integrated Uncertainty Management and Applications, AISC 68, pp. 175–186.
springerlink.com

is usual to call the Kolmogorov-Smirnov two-sample test as Smirnov test (Smirnov, 1939) while the Kolmogorov test is sometimes called the Kolmogorov-Smirnov one-sample test. In our paper, we discuss only Kolmogorov-Smirnov two-sample test, as our purpose here is to test whether two independent samples have been drawn from the same population. The two-sample test is one of the most useful nonparametric methods for comparing two samples, as it is sensitive to differences in both the location and the shape of the empirical cumulative distribution functions of the two samples. Other tests, such as the median test, the Mann-Whitney test, or the parametric *t* test, may also be appropriate (Conover, 1971). However, while these tests are sensitive to differences between the two means or medians, they may not detect other types of differences, such as differences in variances. One of the advantages of two-tailed tests is that such tests consistently reflect all types of differences between two distribution functions. Although many papers have discussed the powerful K-S two-sample test (see discussion in Dixon, 1954; Epstein, 1955; Schroer and Trenkler, 1995), they all simulated them under known distributions. However, sometimes vague information is given when describing data in natural language. When we want to deal with fuzzy data, the underlying distribution of the fuzzy data is not known. It is not easy to put such information into statistical terms. Therefore, we must establish techniques to handle such information and knowledge.

In this paper, we propose a method to judge whether two continuous fuzzy data samples have been draw from the same population. We use the K-S two-sample test to deal with this problem. However, the K-S two-sample test is concerned with real numbers. In order to manipulate continuous fuzzy data by means of the K-S two-sample test, we must find a method to classify all the continuous fuzzy data. Accordingly, we propose some new rules to classify and rank continuous fuzzy data. Several ranking methods have previously been proposed for fuzzy numbers; for instance, Chen (Cheng, 1998) used the distance between fuzzy numbers and compared data to find the largest distance. Moreover, in such the same way as Kaufmann and Gupta (Kaufmann and Gupta, 1988), Liou and Wang (Liou and Wang, 1992) use a membership function to rank fuzzy numbers. Yager (Yager, 1981) proposes a method of ranking fuzzy numbers using a centroid index. Although there are many ways to rank fuzzy numbers, all the methods are based on the central point. Any such method will lose some information about continuous fuzzy data. Given this consideration, we use a weight function to rank fuzzy numbers. The weight function includes both central point and radius, which can be used to classify all continuous fuzzy data. When we use this information, the K-S two-sample test with continuous fuzzy data can be found out.

2 Literature Review

2.1 Kolmogorov-Smirnov Two-Sample Test

To apply the Kolmogorov-Smirnov two-sample test (Siegel, 1988), we determine the cumulative frequency distribution for each sample of observations. We use

the same intervals for each distribution and we subtract one step function from the other for each interval. The test focuses on the largest of these observed deviations.

Let $S_m(X)$ be the observed cumulative distribution for one sample (of size m), that is, $S_m(X) = \frac{K}{m}$, where K is the number of data equal to or less than X. And let $S_n(X)$ be the observed cumulative distribution for the other sample (of size n), that is, $S_n(X) = \frac{K}{n}$. Now, the Kolmogorov-Smirnov two-sample test statistic is

$$D_{m,n} = max[S_m(X) - S_n(X)], \tag{1}$$

for a one-tailed test, and

$$D_{m,n} = max|S_m(X) - S_n(X)|, \tag{2}$$

for a two-tailed test. Note that equation (2) uses the absolute value.

In each case, the sampling distribution of $D_{m,n}$ is known. The probabilities associated with the occurrence of values as large as an observed $D_{m,n}$ under the null hypothesis H_0 (that the two samples have come from the same distribution) have been tabled in (Siegel, 1988). Actually, there are two sampling distributions, depending upon whether the test is one-tailed or two-tailed. Notice that for a one-tailed test we find the $D_{m,n}$ in the predicted direction (using eq. (1)), and for a two-tailed test we find the maximum absolute difference $D_{m,n}$ (using eq. (2)) irrespective of direction. This is because in the one-tailed test, H_1 means that population values from which one of the samples was drawn are stochastically larger than the population values from which the other sample was drawn, whereas in the two-tailed test, H_1 means simply that the two samples are from different populations.

Now, we show the steps in the use of the Kolmogorov-Smirnov two-sample test as follows. Here, we consider the situation of small samples.

(i) Arrange each of two groups of scores in a cumulative frequency distribution using the same intervals (or classifications) for both distributions. Use as many intervals as possible.
(ii) By subtraction, determine the difference between the two-sample cumulative distributions at each listed point.
(iii) Determine the largest of these differences, $D_{m,n}$. For a one-tailed test, $D_{m,n}$ is the largest difference in the predicted direction. For a two-tailed test, $D_{m,n}$ is the largest difference in either direction.
(iv) Determine the significance of the observed $D_{m,n}$ depending on the sample sizes and the nature of H_1. When m and n are both ≤ 25, Appendix Table L_I in (Siegel, 1988) is used for one-tailed test and Appendix Table L_{II} in (Siegel, 1988) is used for two-tailed test. In either table, the entry $m * n * D_{m,n}$ is used.
(v) If the observed value is equal to or larger than that given in the appropriate table for a particular level of significance, H_0 may be rejected in favor of H_1.

In the following subsection, we give definitions we will use in the next section.

2.2 Definitions

We can use the following definition to determine the central point and radiaus.

Definition 2.1. Moments and Center of Mass of a Planar Lamina (Larson, 2008)
Let f and g be continuous functions such that $f(x) \geq g(x)$ on [a,b], and consider the planar lamina of uniform density ρ bounded by the graphs of $y = f(x)$, $y = g(x)$, and $a \leq x \leq b$.

(i) The moments about the $x-$axis and $y-$axis are

$$M_x = \rho \int_a^b [\frac{(f(x)+g(x))}{2}][f(x)-g(x)]dx \tag{3}$$

$$M_y = \rho \int_a^b x[f(x)-g(x)]dx. \tag{4}$$

(ii) The center of mass $(\overline{x},\overline{y})$ is given by $\overline{x} = \dfrac{M_y}{m}$ and $\overline{y} = \dfrac{M_x}{m}$,
where $m = \rho \int_a^b [f(x)-g(x)]dx$ is the mass of the lamina.

In the common use, we always take $\rho = 1$.

3 Kolmogorov-Smirnov Two-Sample Test with Continuous Fuzzy Data

3.1 Empirical Distribution Function with Continuous Fuzzy Data

In order to provide the empirical distribution function for continuous fuzzy data, we must classify the continuous fuzzy data. We first define a weight function for continuous fuzzy data, and then use it to pursue a new classification. Thus, the empirical distribution function for the continuous fuzzy data can be found.

In order to correct the data accurately, we use the continuous revising to define the weight function as follows.

Definition 3.1. Weight function for continuous fuzzy data
The weight function of continuous fuzzy data $X_i \equiv (o_i, l_i)$ is defined as follows:

$$W_i \equiv W(o_i, l_i) = o_i[1 + ke^{-2l_i}], \forall i = 1,2,3,\ldots \tag{5}$$

where o_i is the central point, l_i is the radius with respect to o_i, and $k = max_i(o_i + l_i) - min_j(o_j - l_j), \forall i, j = 1,2,3\ldots$. We name k as weight constant.

Proposition 3.1. *Let $X_i = [a_i, b_i]$ be an interval value, then $o_i = \dfrac{a_i + b_i}{2}, l_i = \dfrac{b_i - a_i}{2}$, and $k = max_i b_i - min_j a_j, \forall i, j = 1,2,3\ldots$.*

Proof: It is trivial that $o_i = \frac{a_i + b_i}{2}$ and $l_i = \frac{b_i - a_i}{2}$.

Therefore, we have

$$\begin{aligned} k &= max_i(o_i + l_i) - min_j(o_j - l_j) \\ &= max_i(\frac{a_i + b_i}{2} + \frac{b_i - a_i}{2}) - min_j(\frac{a_j + b_j}{2} - \frac{b_j - a_j}{2}) \\ &= max_i b_i - min_j a_j, \forall i, j = 1, 2, 3 \ldots . \end{aligned}$$

Proposition 3.2. *Let* $X_i = [a_i, b_i, c_i]$ *be triangular fuzzy numbers, then* $o_i = \frac{a_i + b_i + c_i}{3}, l_i = \frac{c_i - a_i}{4}$, *and* $k = max_i(\frac{a_i + 4b_i + 7c_i}{12}) - min_j(\frac{7a_j + 4b_j + c_i}{12})$, $\forall i, j = 1, 2, 3 \ldots .$

Proof: By Definition 1, we let $\rho = 1$ and we can find that $o_i = \frac{M_y}{m}$.

When X_i is a triangular fuzzy number, its membership function is denoted as follows:

$$f(x) = \begin{cases} 0, & \text{x} < a \ \ and \ \ x > c \\ \frac{x-a}{b-a}, & \text{a} \le x \le b \\ \frac{c-x}{c-b}, & \text{b} \le x \le c \end{cases} .$$

Therefore, $M_y = 1\int_a^b x\frac{x-a}{b-a}dx + 1\int_b^c x\frac{c-x}{c-b}dx = \frac{1}{6}(c-a)(a+b+c)$ and $m = 1\int_a^b \frac{x-a}{b-a}dx + 1\int_b^c \frac{c-x}{c-b}dx = \frac{c-a}{2}$.

Hence,

$$o_i = \frac{M_y}{m} = \frac{\frac{1}{6}(c_i - a_i)(a_i + b_i + c_i)}{\frac{c_i - a_i}{2}} = \frac{a_i + b_i + c_i}{3}.$$

Moreover, the mean value theorem for definite integrals (George, 2005) enables us to find some points t in [a,c] such that

$$(c-a)f(t) = \int_a^c f(x)dx = \frac{c-a}{2}.$$

Therefore, $f(t) = \frac{1}{2}, \forall t \in [a, c]$.

In the case where there are two points, say t_1 and t_2, such that

$$f(t_1) = f(t_2) = \frac{1}{2}, \forall t_1, t_2 \in [a, c].$$

This results in $t_1 = \frac{a+b}{2}$ and $t_2 = \frac{b+c}{2}$.

There also exists a rectangle with the same area as $\frac{c-a}{2}$. Hence $2l = t_2 - t_1 = \frac{c-a}{2}, l = \frac{c-a}{4}$.

When we have o_i and l_i, the weight constant k is

$$\begin{aligned} k &= max_i(o_i + l_i) - min_j(o_j - l_j) \\ &= max_i(\frac{a_i+b_i+c_i}{3} + \frac{c_i-a_i}{4}) - min_j(\frac{a_i+b_i+c_i}{3} - \frac{c_i-a_i}{4}) \\ &= max_i(\frac{a_i+4b_i+7c_i}{12}) - min_j(\frac{7a_j+4b_j+c_i}{12}), \forall i, j = 1,2,3\ldots \end{aligned}$$

Proposition 3.3. *Let* $X_i = [a_i, b_i, c_i, d_i]$ *be trapezoidal fuzzy numbers, then* $o_i = \frac{(c_i+d_i)^2 - (a_i+b_i)^2 + a_i b_i - c_i d_i}{3[(c_i+d_i)-(a_i+b_i)]}$, $l_i = \frac{(c_i+d_i)-(a_i+b_i)}{4}$, *and* $k = max_i(o_i + l_i) - min_j(o_j - l_j), \forall i, j = 1,2,3\ldots.$

Proof: By Definition 1, we let $l = 1$ and we can find that $o_i = \frac{M_y}{m}$.

When X_i is a trapezoidal fuzzy number, its membership function is denoted as follows:

$$f(x) = \begin{cases} 0, & \text{x} < aandx > d \\ \frac{x-a}{b-a}, & \text{a} \le x \le b \\ 1, & \text{b} \le x \le c \\ \frac{d-x}{d-c}, & \text{c} \le x \le d \end{cases}.$$

Therefore, $M_y = 1\int_a^b x\frac{x-a}{b-a}dx + 1\int_b^c x1dx + 1\int_c^d x\frac{d-x}{d-c}dx = \frac{1}{6}[(c+d)^2 - (a+b)^2 + (ab-cd)]$ and $m = 1\int_a^b \frac{x-a}{b-a}dx + 1\int_b^c 1dx + 1\int_c^d \frac{d-x}{d-c}dx = \frac{1}{2}[(c+d) - (a+b)]$.

Hence,

$$o_i = \frac{M_y}{m} = \frac{\frac{1}{6}[(c+d)^2 - (a+b)^2 + (ab-cd)]}{\frac{1}{2}[(c+d)-(a+b)]} = \frac{[(c+d)^2 - (a+b)^2 + (ab-cd)]}{3[(c+d)-(a+b)]}.$$

Moreover, the mean balue theorem for fefinite integrals (George, 2005) enables us to find some points t in [a,d] such that

$$(d-a)f(t) = \int_a^d f(x)dx = \frac{(c+d)-(a+b)}{2}.$$

Therefore, $f(t) = \frac{(c+d)(a+b)}{2(d-a)}, \forall t \in [a,d]$.

In the case where there are two points, say t_1 and t_2, such that

$$f(t_1) = f(t_2) = \frac{(c+d)(a+b)}{2(d-a)}, \forall t_1, t_2 \in [a,d].$$

We can also find a rectangle with the same area as $\frac{(c+d)(a+b)}{2}$.

Hence, $2l = t_2 - t_1 = \frac{(c+d)(a+b)}{2}$ and $l = \frac{(c+d)(a+b)}{4}$.

When we have o_i and l_i, the weight constant k is

$$k = max_i(o_i + l_i) - min_j(o_j - l_j), \forall i, j = 1,2,3\ldots$$

Definition 3.2. Fuzzy classification

If $W_i < W_j, \forall i \neq j$, we say that X_i and X_j are in different classes. In particular, X_i is the class before X_j. Moreover, if $W_i = W_j, \forall i \neq j$, we say that X_i and X_j are in the same class.

Definition 3.3. Identical independence of continuous fuzzy data

If $W_i \neq W_j, \forall i \neq j$, we say that X_i and X_j are identical independent by the choose of k (weight constant). Otherwise, X_i and X_j are dependent.

Definition 3.4. Empirical distribution function with continuous fuzzy data

Let $X_1, X_2, \ldots, X_n$ be n continuous fuzzy data. We can use the weight function to separate X_i into different class $\mathscr{C}_i$, which are called Glivenko-Cantelli classes (see discussion in Gaenssler and Stute, 1979; Gine and Zinn, 1984; Serfling, 1980). If X_i and X_j are in different classes, then we say that X_i and X_j are identically independent for $i \neq j$. Moreover, we have the order statistic of X_i (assume that they are in different classes), denoted as

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)} \tag{6}$$

Hence, the empirical distribution function can be generalized to a set $\mathscr{C}$ to obtain an empirical measure indexed by c.

$$S_n(c) = \frac{1}{n}\sum_{i=1}^{n} I_c(X_i), c \in \mathscr{C}, \tag{7}$$

where I_c is the indicator function denoted by

$$I_c(x_i) = \begin{cases} 1, \ \mathrm{x}_i \in \mathscr{C}, \\ 0, \ \mathrm{x}_i \notin \mathscr{C} \end{cases}, \forall i = 1,2,\ldots n. \tag{8}$$

Now, when we have those definitions, we can proceed to study the Kolmogorov-Smirnov two-sample test with continuous fuzzy data.

3.2 Kolmogorov-Smirnov Two-Sample Test with Continuous Fuzzy Data

Procedure for using K-S two-sample test for continuous fuzzy data (Two-tailed test) in small samples:

(i) Samples: Let $X_m$and Y_n be two samples with continuous fuzzy data. X_i has size m and Y_j has size n. Combining all observations, we have $N = m + n$ pieces of data. A value of the weight function W_i can be found that will let us distribute X_m and Y_n into different classes $\mathscr{C}_i$ (maybe in the same class). The number of classes is less than or equal to N. Moreover, the two empirical distribution functions of X_m and Y_n can be found individually.
(ii) Hypothesis: Two samples have the same distribution H_0.
(iii) Statistics: $D_{m,n} = max|S_m(X) - S_n(X)|$.
(iv) Decision rule: Under significance level α. Appendix Table L_{II} (Siegel, 1988) is used.

4 Empirical Studies

Example 1. A Japanese dining hall manager planned to introduce new boxed lunch services and decided to take a survey to investigate what price for a boxed lunch would be acceptable to male and female customers. A sample was randomly selected of 20 customers (10 males and 10 females) who resided around this dining hall in the city of Taipei. The investigator asked them, how many dollars they would be willing to spend (can answer with interval) for a boxed lunch in a Japanese dining hall. The answers are shown in Table 1.

Table 1 The Price Which will be Acceptable by Males and Females

Males	[60,70]	[70,90]	[50,80]	[50,60]	[80,100]	[70,90]	[50,80]	[50,70]	[65,95]	[50,100]
Females	[50,60]	[60,70]	[80,100]	[90,120]	[90,100]	[55,75]	[70,90]	[100,120]	[80,120]	[90,120]

First, we distributed male answers and female answers into different classes. We had to find the weight values and compare them. Moreover, we had to determine which class they belong to. The calculation was done as Table 2.

Comparison among W_i, results in the following inequality:

$W(X_4)=W(Y_1)<W(X_8)<W(X_3)=W(X_7)<W(Y_6)<W(X_1)=W(Y_2)<W(X_{10})<W(X_9)<W(X_2)=W(X_6)=W(Y_7)<W(X_5)=W(Y_3)<W(Y_5)<W(Y_9)<W(Y_4)=W(Y_{10})<W(Y_8)$.

Here, we take $k = max_i b_i - min_j a_j = 120 - 50 = 70, \forall i, j = 1, 2, \ldots, 20$.

Table 2 The Weight Values and Classes

	$[a_i,b_i]$	o_i	l_i	W_i	$\mathscr{C}_i$
X_1	[60,70]	65	5	$65[1+ke^{-10}]$	5
X_2	[70,90]	80	10	$80[1+ke^{-20}]$	8
X_3	[50,80]	65	15	$65[1+ke^{-30}]$	3
X_4	[50,60]	55	5	$55[1+ke^{-10}]$	1
X_5	[80,100]	90	10	$90[1+ke^{-20}]$	9
X_6	[70,90]	80	10	$80[1+ke^{-20}]$	8
X_7	[50,80]	65	15	$65[1+ke^{-30}]$	3
X_8	[50,70]	60	10	$60[1+ke^{-20}]$	2
X_9	[65,95]	80	15	$80[1+ke^{-30}]$	7
X_{10}	[50,100]	75	25	$75[1+ke^{-50}]$	6
Y_1	[50,60]	55	5	$55[1+ke^{-10}]$	1
Y_2	[60,70]	65	5	$65[1+ke^{-10}]$	5
Y_3	[80,100]	90	10	$90[1+ke^{-20}]$	9
Y_4	[90,120]	105	15	$105[1+ke^{-30}]$	12
Y_5	[90,100]	95	5	$95[1+ke^{-10}]$	10
Y_6	[55,75]	65	15	$65[1+ke^{-30}]$	4
Y_7	[70,90]	80	15	$80[1+ke^{-30}]$	8
Y_8	[100,120]	110	15	$110[1+ke^{-30}]$	13
Y_9	[80,120]	100	20	$100[1+ke^{-40}]$	11
Y_{10}	[90,120]	105	15	$105[1+ke^{-30}]$	12

From the above, we have 13 classes. Now, we went on to find the cumulative distributions of X_i and Y_j.

Table 3 The cumulative distributions of X_i and Y_j

$\mathscr{C}_i$	1	2	3	4	5	6	7	8	9	10	11	12	13
$S_{10}(X)$	.1	.2	.4	.4	.5	.6	.7	.9	1	1	1	1	1
$S_{10}(Y)$	.1	.1	.1	.2	.3	.3	.3	.4	.5	.6	.7	.9	1
$\lvert S_{10}(X)-S_{10}(Y)\rvert$	0	.1	.3	.2	.2	.3	.4	.5	.5	.4	.3	.1	0

From Table 3, the test statistic was obtained:

$$D = max|S_{10}(X) - S_{10}(Y)| = 0.5.$$

at a significance level $\alpha = 0.05$, $mnD = 10 * 10 * (0.5) = 50 < 60$ (Appendix Table L_{II} (Siegel, 1988)). Since the observed value did not exceeds the critical value, we did not reject H_0. We conclude that males and females have the same interval of the acceptable price of a boxed lunch.

Example 2. With the rest of the procedure as illustrated in Example 1. The investigator asked them in the following questions: 1. In which price range (an interval)

would they be willing to spend for a lunch box in a Japanese dining hall? 2. In which price range (real numbers) will not they buy it? We can collect those data and get trapezoidal fuzzy numbers. The answers are shown in Table 4.

Table 4 The Price which will be Acceptable by Males and Females

Males	[0,60,90,100]	[60,60,90,100]	[30,60,90,100]	[50,60,80,80]	[50,50,80,100]
	[50,50,80,80]	[55,65,75,80]	[50,60,80,80]	[50,70,160,160]	[40,60,120,150]
Females	[40,50,70,70]	[50,50,70,100]	[50,50,100,100]	[150,150,250,300]	[50,50,70,70]
	[50,70,80,80]	[40,40,90,150]	[50,50,70,80]	[50,60,150,200]	[50,70,150,200]

First, we classified male answers and female answers into different classes. We had to find the weight values and compare them. Moreover, we had to determine which class they belong to. The calculation was done as shown in Table 5.

Table 5 The Weight Values and Classes

	$[a_i,b_i]$	o_i	l_i	W_i	$\mathscr{C}_i$
X_1	[0,60,90,100]	$(2350/39)$	32.5	$(2350/39)[1+ke^{-65}]$	3
X_2	[60,60,90,100]	$(1630/21)$	17.5	$(1630/21)[1+ke^{-35}]$	13
X_3	[30,60,90,100]	$(208/3)$	25.0	$(208/3)[1+ke^{-50}]$	9
X_4	[50,60,80,80]	$(202/3)$	12.5	$(202/3)[1+ke^{-25}]$	6
X_5	[50,50,80,100]	$(845/12)$	20.0	$(845/12)[1+ke^{-40}]$	11
X_6	[50,50,80,80]	65	15.0	$65[1+ke^{-30}]$	5
X_7	[55,65,75,80]	$(480/7)$	8.75	$(480/7)[1+ke^{-17.5}]$	7
X_8	[50,60,80,80]	$(202/3)$	12.5	$(202/3)[1+ke^{-25}]$	6
X_9	[50,70,160,160]	104	50.0	$104[1+ke^{-100}]$	16
X_{10}	[40,60,120,150]	$(4730/51)$	42.5	$(4730/51)[1+ke^{-85}]$	15
Y_1	[40,50,70,70]	$(172/3)$	12.5	$(172/3)[1+ke^{-25}]$	2
Y_2	[50,50,70,100]	$(480/7)$	17.6	$(480/7)[1+ke^{-35}]$	8
Y_3	[50,50,100,100]	75	25.0	$75[1+ke^{-50}]$	12
Y_4	[150,150,250,300]	$(640/3)$	62.5	$(640/3)[1+ke^{-125}]$	19
Y_5	[50,50,70,70]	$(235/6)$	10.0	$(235/6)[1+ke^{-20}]$	1
Y_6	[50,70,80,80]	$(415/6)$	10.0	$(415/6)[1+ke^{-20}]$	10
Y_7	[40,40,90,150]	$(655/8)$	40.0	$(655/8)[1+ke^{-80}]$	14
Y_8	[50,50,70,80]	$(188/3)$	12.5	$(188/3)[1+ke^{-25}]$	4
Y_9	[50,60,150,200]	$(695/6)$	60.0	$(695/6)[1+ke^{-120}]$	17
Y_{10}	[50,70,150,200]	$(2720/23)$	57.5	$(2720/23)[1+ke^{-115}]$	18

Comparison among W_i results in the following inequality:

$W(Y_5)<W(Y_1)<W(X_1)<W(Y_8)<W(X_6)<W(X_4)=W(X_8)<W(X_7)<W(Y_2)<W(X_3)<W(Y_6)<W(X_5)<W(Y_3)<W(X_2)<W(Y_7)<W(X_{10})<W(X_9)<W(Y_9)<W(Y_{10})<W(Y_4)$

Here, we take $k=max_i(o_i+l_i)-min_j(o_j-l_j)=(\frac{640}{3}+62.5)-(\frac{2350}{39}-32.5)\approx 248.0769\ldots,\forall i,j=1,2,\ldots,20.$

From the above, we have 19 classes. Now, we went on to find the cumulative distributions of X_i and Y_j.

Table 6 The cumulative distributions of X_i and Y_j

$\mathscr{C}_i$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
$S_{10}(X)$	0	0	.1	.1	.2	.4	.5	.5	.6	.6	.7	.7	.8	.8	.9	1	1	1	1
$S_{10}(Y)$	.1	.2	.2	.3	.3	.3	.3	.4	.4	.5	.5	.6	.6	.7	.7	.7	.8	.9	1
$\lvert S_{10}(X) - S_{10}(Y)\rvert$	.1	.2	.1	.2	.1	.1	.2	.1	.2	.1	.2	.1	.2	.1	.2	.3	.2	.1	0

From Table 6, the test statistic was obtained in the following:

$$D = max|S_{10}(X) - S_{10}(Y)| = 0.3.$$

At a significance level $\alpha = 0.05$, $mnD = 10 * 10 * (0.3) = 30 < 60$ (Appendix Table L_{II} (Siegel, 1988)). Since the observed value did not exceeds the critical value, we did not reject H_0. We conclude that males and females have the same interval of the acceptable price of a boxed lunch.

5 Conclusions

In this paper, we studied the use of the K-S two-sample test with small samples of continuous fuzzy data. In order to identify the statistical pivot, we defined a new function, the weight function, which includes both central point and radius. The weight function can be used to classify all continuous fuzzy data. Moreover, we could divide fuzzy data samples into different classes. With this rule, the cumulative distribution function can be found out. Therefore, we could obtain the statistical pivot of K-S test with continuous fuzzy data. We also give an example of empirical studies, which showed that fuzzy hypothesis testing with soft computing is a realistic and reasonable approach to deal with continuous fuzzy data in the social science research.

However, we still can identify some open problems that require future investigation:

(i) How should we verify that the continuous fuzzy data are really separated each other? Moreover, can we say that they are independent?
(ii) For large samples, is this weight function still useful?
(iii) How is the sensitivity of the hypothesis test known with continuous fuzzy data?

References

1. Conover, W.J.: Practical nonparametric statistics, New York (1971)
2. Cheng, C.H.: A new approach for ranking fuzzy numbers by distance method. Fuzzy sets and systems 95(3), 307–317 (1998)

3. Dixon, W.J.: Power under normality of several nonparametric tests. The Annals of Mathematical Statistics 25(3), 610–614 (1954)
4. Epstein, B.: Comparison of Some Non-Parametric Tests against Normal Alternatives with an Application to Life Testing. Journal of the American Statistical Association 50(271), 894–900 (1955)
5. Gaenssler, P., Stute, W.: Empirical processes: a survey of results for independent and identically distributed random variables. Annals of Applied Probability 7(2), 193–243 (1979)
6. Gine, E., Zinn, J.: Some limit theorems for empirical measures (with discussion). Annals of Applied Probability 12(4), 929–989 (1984)
7. Thomas Jr., G.B.: 11th Thomas Calculus. Pearson Education, Inc., Boston (2005)
8. Kaufmann, A., Gupta, M.M.: Fuzzy mathematical models in engineering and management science. Elsevier Science Publishers BV, New York (1988)
9. Larson, R., Hostetler, R., Edwards, B.H.: Essential Calculus: Early Transcendental Functions. Houghton Mifflin Company, Boston (2008)
10. Liou, T.S., Wang, M.J.: Ranking fuzzy numbers with integral value. Fuzzy Sets and Systems 50(3), 247–255 (1992)
11. Serfling, R.J.: Approximation theorems of mathematical statistics, New York (1980)
12. Schroer, G., Trenkler, D.: Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples. Computational Statistics and Data Analysis 20(2), 185–202 (1995)
13. Siegel, S., Castellan, N.J.: Nonparametric statistics for the behavioral sciences, 2nd edn., New York (1988)
14. Smirnov, N.V.: Estimate of deviation between empirical distribution functions in two independent samples (Russian) Bulletin Moscow Univ. 2(2), 3–16 (1939)
15. Yager, R.R.: A procedure for ordering fuzzy subsets of the unit interval. Information Science 24(2), 143–161 (1981)