Journal of Hyperstructures 6 *(2)* (*2017*), *147-155.*
ISSN: 2322-1666 print/2251-8436 online

KOLMOGOROV-SMIRNOV TWO-SAMPLE TEST IN FUZZY ENVIRONMENT

FERESHTEH MOMENI, BAHRAM SADEGHPOUR GILDEH AND GHOLAMREZA HESAMIAN

Abstract. Kolmogorov-Smirnov two-sample test, is a common test for fitting statistical population model. Statistic of this test is defined based on the empirical distribution function and so, sorting sample observations plays a key role in determination of the empirical distribution function. In this paper, a new approach to generalize the Kolmogorov-Smirnov two-sample test has been provided, where the sample observations is defined as imprecise numbers and hypotheses testing are precisely defined. To do this, first, a new method for ranking fuzzy numbers using $D_{p,q}$ metric was proposed. We used this metric for separating fuzzy data to separate classes and then placed fuzzy data in certain classes. Then, we have defined an extension of the empirical distribution function and similar to the classic case, calculated Kolmogorov-Smirnov two-sample test statistic and accomplished to make decision about accepting or rejecting the null hypothesis as completely exact. Finally, with a numerical example the proposed approach was evaluated and compared.

Key Words: Fuzzy data, Kolmogorov-Smirnov test, Goodness of fit test, $D_{p,q}$-ranking method, empirical distribution function.

2010 Mathematics Subject Classification: Primary: 13A15; Secondary: 13F30, 13G05.

Received: 27 August 2016, Accepted: 18 September 2016. Communicated by Ahmad Yousefian Darani;

∗Address correspondence to Fereshteh Momeni; E-mail: momeni@iaubeh.ac.ir.

1. Introduction

Non-parametric methods in statistics are analytical approaches which consider the least possible assumptions about the statistical population compared to parametric methods [3]. Classic non-parametric assumptions test are based on accurate data. But, in practice we are dealing with a lot of subjects that, the data is not precis. In such a case, to have the necessary tools to analyze the data, classical existing approaches should be appropriately expanded to fuzzy environment.

Kolmogorov-Smirnov test is one of the most common model fitness tests. Kolmogorov-Smirnov two-sample test is used to assess differences between two independent statistical populations based on only one variable. One of the most important tools in determining the statistics of Kolmogorov-Smirnov two-sample test is the use of empirical distribution function in independent statistical two samples. The criterion to determine the empirical distribution function, is use of precise data. But, in practice we confront with a condition that, two independent random samples observations are reported as imprecise. Therefore, to determine the empirical distribution function, it should be extended to fuzzy environment. Hesamian and Chachi [5] suggested a new approach for Kolmogorov-Smirnov two-sample test in the case that data are observations of fuzzy random variables. In their approach, new concepts of fuzzy cumulative distribution function and fuzzy empirical distribution function were presented based on fuzzy random variables and then, statistic of Kolmogorov-Smirnov two-sample test was extended to fuzzy environment. Finally, at certain significant level, definition of fuzzy p-value has been provided and used to decide the degree of credit for accepting or rejecting the null hypothesis.

Also, Lin et al. [6] studied Kolmogorov-Smirnov two-sample test for continuous interval data. In their approach, a new weight function for a continuous interval data was presented and used to put each observation in separate classes. Then, on the basis of certain classes, they proposed a new definition for the empirical distribution function. Finally, similar to the classic case, the test statistic was accurately calculated and decision on accepting or rejecting the null hypothesis of the test was made.

In this paper, inspired by Lin et al.'s approach, Kolmogorov-Smirnov two-sample test was evaluated based on fuzzy data. In the proposed approach, definition of the distance between two fuzzy numbers was used to assign independent sample observations in separate classes, then the empirical distribution function in two independent random samples with

respect to each class of fuzzy data was defined. Finally, similar to the classic mode, accurate test statistic was calculated and significantly, decisions about accepting or rejecting the null hypothesis were made.
The article is organized as follows: in section 2, we recall some definitions of fuzzy number and ranking fuzzy numbers based on $D_{p,q}-$metric. In section 3, we proposed a new definition of empirical distribution function based on fuzzy observations. In section 4, we considered the problem of kolmogorov-Smirnov two sample test in fuzzy environment. In section 5, we utilize this test within numerical example. Section 6 concludes the paper.

2. Fuzzy numbers and distance between them

It may happen that a sample used for decision making consists of observations that are not necessarily crisp. In order to describe the vagueness of data, the notion of a fuzzy number is introduced by Dubois and Prade (1983) [1].

2.1. **Fuzzy numbers.** A fuzzy subset $\widetilde{A}$ of the universal set χ is defined by its membership function $\mu_{\widetilde{A}} : \chi \to [0,1]$, with the set $\text{supp}(\widetilde{A}) = \{x \in \chi : \ \mu_{\widetilde{A}}(x) > 0\}$, the support of $\widetilde{A}$. In this work, $\mathbb{R}$ (the real line) is considered as the universal set. It is denoted by $\widetilde{A}[\alpha]$, the α-cut of the fuzzy subset $\widetilde{A}$ of $\mathbb{R}$, defined for every $\alpha \in (0,1]$, by $\widetilde{A}[\alpha] = \{x \in \mathbb{R}, \ \mu_{\widetilde{A}}(x) \geq \alpha\}$, and $\widetilde{A}[0]$ is the closure of $\text{supp}(\widetilde{A})$. The fuzzy subset $\widetilde{A}$ of $\mathbb{R}$ is called a fuzzy number for every $\alpha \in (0,1]$, if the set $\widetilde{A}[\alpha]$ is a non-empty compact interval. Such an interval is denoted by $\widetilde{A}[\alpha] = \left[\widetilde{A}_\alpha^-, \widetilde{A}_\alpha^+\right]$, where $\widetilde{A}_\alpha^- = \inf\{x : \ x \in \widetilde{A}[\alpha]\}$ and $\widetilde{A}_\alpha^+ = \sup\{x : \ x \in \widetilde{A}[\alpha]\}$. The set of all fuzzy numbers is denoted by $\mathcal{F}(\mathbb{R})$.

One of the most popular types of fuzzy number, to be considered in this work, is the recalled trapezoidal fuzzy number $\widetilde{A} = (a^l, a^c, a^s, a^r)_T$ whose membership function is given by

$$\mu_{\widetilde{A}}(x) = \begin{cases} 0, & x < a^l \\ \frac{x-a^l}{a^c-a^l}, & a^l \leq x < a^c \\ 1, & a^c \leq x < a^s \\ \frac{a^r-x}{a^r-a^s}, & a^s \leq x \leq a^r \\ 0, & x > a^r \end{cases}$$

If $a^c = a^s$, it is called a triangular fuzzy number and denoted by $\widetilde{A} = (a^l, a^c, a^r)_T$. For more detailed information regarding fuzzy numbers, see [7].

2.2. **$D_{p,q}$-Distance between two fuzzy numbers.** The $D_{p,q}$-distance, indexed by parameters $1 \leq p \leq \infty$ and $0 \leq q \leq 1$, between two fuzzy numbers $\widetilde{A}$ and $\widetilde{B}$ is a nonnegative function given as follows:

$$D_{p,q}(\widetilde{A}, \widetilde{B}) = \begin{cases} \left[(1-q)\int_0^1 |A_\alpha^- - B_\alpha^-|^p d\alpha + q\int_0^1 |A_\alpha^+ - B_\alpha^+|^p d\alpha\right]^{\frac{1}{p}}, & p < \infty \\ (1-q)\sup\limits_{0<\alpha\leq 1}(|A_\alpha^- - B_\alpha^-|) + q\inf\limits_{0<\alpha\leq 1}(|A_\alpha^+ - B_\alpha^+|), & p = \infty \end{cases}$$

The analytical properties of $D_{p,q}$ depends on the first parameter p, while the second parameter q is the weighted one. $(\mathcal{F}(\mathbb{R}), D_{p,q})$ is a complete metric space. If there is no reason for distinguishing any side of the fuzzy numbers, $D_{p,\frac{1}{2}}$ is recommended.

For instance, for trapezoidal fuzzy numbers $\widetilde{A} = (a_1, a_2, a_3, a_4)_T$ and $\widetilde{B} = (b_1, b_2, b_3, b_4)_T$ the above distance with $p = 2$ and $q = \frac{1}{2}$ is calculated as:

$$D_{2,\frac{1}{2}}(\widetilde{A}, \widetilde{B}) = \sqrt{\frac{1}{6}\left[\sum_{i=1}^{4}(b_i - a_i)^2 + \sum_{i\in\{1,3\}}(b_i - a_i)(b_{i+1} - a_{i+1})\right]}. \tag{2.1}$$

2.3. **Ranking fuzzy numbers based on $D_{p,q}$-distance.** In this section, the method for ranking the fuzzy numbers is defined based on $D_{p,q}$ metric. In this method, first, a quantity of greater (or less) base from the set of data that can be presumed accurately (or in accurately) is considered and called B (or $\widetilde{B}$). Then, the distance between the set of fuzzy numbers and B (or $\widetilde{B}$) is calculated using the relation (2.1). Finally, the fuzzy numbers are ranked by comparing the obtained minute $D_{p,q}$ quantities. The more (less) the quantity of $D_{p,q}$ is, the higher (lower) the fuzzy number is ranked.

Example 2.1. Let $\widetilde{A} = (0.4, 0.5, 1)_T$, $\widetilde{B} = (0.4, 0.7, 1)_T$ and $\widetilde{C} = (0.4, 0.9, 1)_T$ be two a set of fuzzy numbers. First, the basic quantity is considered greater than fuzzy numbers as B= 1.1 in order to ranking fuzzy numbers set. Thus, by using relation (2.1), we obtain distance between every one of fuzzy number and B as the following:

$D_{2,\frac{1}{2}}(\widetilde{A}, 1.1) = \sqrt{\frac{1}{6}\left[0.49 + 0.72 + 0.01 + 0.42 + 0.06\right]} = 0.53,$

$D_{2,\frac{1}{2}}(\widetilde{B}, 1.1) = 0.44,$
$D_{2,\frac{1}{2}}(\widetilde{C}, 1.1) = 0.35.$
Thus, it can be concluded $\widetilde{A} < \widetilde{B} < \widetilde{C}$.

3. Empirical distribution function with fuzzy data

In this section, we generalize the empirical distribution function to fuzzy environment. For this purpose, the separate and independent classes were defined in the case that observations of a random sample was as fuzzy numbers. Then, the concept of separate classes will be used to determine the empirical distribution function.

Definition 3.1. Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with continuous distribution function F_X and the fuzzy values $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ were observed rather than the crisp data $x_1, x_2, \ldots, x_n$. we can say, $\widetilde{x}_i$'s belong to different classes (Gilvenko-Cantelli class [2],[4]) When $D_{2,\frac{1}{2}}(\widetilde{x}_i, \widetilde{B})$'s have different values on which $\widetilde{B}$ is desired value smaller than the fuzzy observations. Similarly, if $D_{2,\frac{1}{2}}(\widetilde{x}_i, \widetilde{B}) = D_{2,\frac{1}{2}}(\widetilde{x}_j, \widetilde{B})$ for every $i \neq j$, then say fuzzy observation classes of $\widetilde{x}_i$ and $\widetilde{x}_j$ will be the same.

Definition 3.2. [6] Random variables X and Y with fuzzy values $\widetilde{x}$ and $\widetilde{y}$ are independent, when fuzzy observations $\widetilde{x}$ and $\widetilde{y}$ belong to two different classes.

Definition 3.3. [6] Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with continuous distribution function F_X and the fuzzy values $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ were observed rather than the crisp data $x_1, x_2, \ldots, x_n$. If $\widetilde{x}_{(1)}, \widetilde{x}_{(2)}, \ldots, \widetilde{x}_{(n)}$ be order fuzzy observations corresponds to the random sample $X_1, X_2, \ldots, X_n$, , then the generalized empirical distribution function is defined as follows:

$$S_n(c) = \frac{1}{n} \sum_{i=1}^{n} I_c(\widetilde{x}_i), \quad c \in \mathcal{C} \tag{3.1}$$

where,

$$I_c(\widetilde{x}_i) = \begin{cases} 1, & \widetilde{x}_i \in \mathcal{C} \\ 0, & \widetilde{x}_i \notin \mathcal{C}. \end{cases}$$

In the next section, using the definition of generalized empirical distribution function, Kolmogorov-Smirnov two-sample test is extended in fuzzy environment.

4. Kolmogorov-Smirnov two sample test with fuzzy data

Suppose $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are two independent random sample of two statistical population with continuous cumulative distribution functions F_X and F_Y, respectively. Now, let that observations of two random samples are considered as fuzzy quantities $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ and $\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_m$ rather than the crisp data $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$, respectively. If we are interested in comparing the two distributions of given populations, then test hypotheses can be defined as follows:

$$\begin{cases} H_0 : F_X(\widetilde{x}) = F_Y(\widetilde{y}), \\ H_1 : F_X(\widetilde{x}) \neq F_Y(\widetilde{y}). \end{cases}$$

A common way to test above hypotheses is Kolmogorov-Smirnov two-sample test. To perform this test, first, we sort the fuzzy observations of pooled samples of $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ and $\widetilde{y}_1, \widetilde{y}_2, \ldots, \widetilde{y}_m$. To sort fuzzy observations of pooled samples, we will use $D_{p,q}$ ranking method described in section 2.3. Then, based on $D_{2,\frac{1}{2}}(\widetilde{x}_i, \widetilde{B})$ values for $i = 1, 2, \ldots, N$, in which $N = n + m$, fuzzy observations of two random samples are classified to separate classes (class of the some observations may be identical). The number of classes will be less or equal N. Finally, based on generalized empirical distribution functions of X and Y, statistic of Kolmogorov-Smirnov two-sample test is determined as follows:

$$D = n.m.D_{n,m} = n.m. \max |S_m(\widetilde{y}) - S_n(\widetilde{x})|. \tag{4.1}$$

Note that D will be a crisp value. Therefore, in Significant level of δ, the null hypothesis is accepted when the observed $p-$value from table of Kolmogorov-Smirnov two-sample test (table I, p.581 of [3]) is larger than δ. Otherwise, the null hypothesis will be rejected.

5. Numerical example

The manager of a Japanese cafeteria plans to introduce a new lunch service pack. For this purpose, the head of the hall decides to conduct an investigation about the proper price of this lunch package between consumers. For this study, a random sample of 20 clients (10 male and 10 female) of customers who were staying near the dining hall was surveyed and, they suggested the appropriate price of this lunch package approximately in dollars [6]. The results of this survey, which is triangular fuzzy numbers have been recorded in Table 1.

TABLE 1. The price which will be acceptable by customers

Male		Female	
$\widetilde{x}_1 = (60, 65, 70)_T$	$\widetilde{x}_2 = (70, 75, 90)_T$	$\widetilde{y}_1 = (50, 55, 60)_T$	$\widetilde{y}_2 = (55, 65, 75)_T$
$\widetilde{x}_3 = (70, 85, 90)_T$	$\widetilde{x}_4 = (50, 70, 80)_T$	$\widetilde{y}_3 = (70, 80, 90)_T$	$\widetilde{y}_4 = (60, 65, 70)_T$
$\widetilde{x}_5 = (50, 70, 80)_T$	$\widetilde{x}_6 = (50, 60, 70)_T$	$\widetilde{y}_5 = (100, 110, 120)_T$	$\widetilde{y}_6 = (80, 90, 100)_T$
$\widetilde{x}_7 = (50, 55, 60)_T$	$\widetilde{x}_8 = (65, 80, 95)_T$	$\widetilde{y}_7 = (80, 95, 120)_T$	$\widetilde{y}_8 = (90, 110, 120)_T$
$\widetilde{x}_9 = (80, 95, 100)_T$	$\widetilde{x}_{10} = (50, 85, 100)_T$	$\widetilde{y}_9 = (90, 110, 120)_T$	$\widetilde{y}_{10} = (90, 95, 100)_T$

TABLE 2. $D_{2,\frac{1}{2}}(., 45)$ and classes

Fuzzy Data	$D_{2,\frac{1}{2}}(\widetilde{x}_i,45)$	c_i	Fuzzy Data	$D_{2,\frac{1}{2}}(\widetilde{y}_i,45)$	c_i
$\widetilde{x}_1$	20.2	3	$\widetilde{y}_1$	10.4	1
$\widetilde{x}_2$	37.9	9	$\widetilde{y}_2$	20.2	3
$\widetilde{x}_3$	24.2	4	$\widetilde{y}_3$	45.5	11
$\widetilde{x}_4$	10.4	1	$\widetilde{y}_4$	63.1	15
$\widetilde{x}_5$	47.9	12	$\widetilde{y}_5$	50.1	13
$\widetilde{x}_6$	33	6	$\widetilde{y}_6$	20.8	5
$\widetilde{x}_7$	24.2	4	$\widetilde{y}_7$	35.5	7
$\widetilde{x}_8$	16.1	2	$\widetilde{y}_8$	65.3	16
$\widetilde{x}_9$	36.1	8	$\widetilde{y}_9$	53.8	14
$\widetilde{x}_{10}$	38	10	$\widetilde{y}_{10}$	63.1	15

Table 2 shows 16 different classes to show pooled samples. Generalized empirical distribution functions of samples obtained by $D_{2,\frac{1}{2}}$ ranking method for $B = 45$, have been shown in Table 3. Statistic of Kolmogorov-Smirnov two-sample test will be as follows:

$$D = 10(10)(\max|S_{10}(\widetilde{y}) - S_{10}(\widetilde{x})|) = 100(0.5) = 50$$

According to the table I (p.583 [3]), we have $p-value = 0.2$. Therefore, at the significant level of $\delta = 0.05$, the null hypothesis is accepted. In other words, it can be concluded that two independent random samples are from the identical distribution.

Remark 5.1. Lin et al. investigated the problem of testing this example, while two random samples data were considered as continuous fuzzy data. First, by applying a kind of weight function, continuous fuzzy data are sorted and then the data were put in the classes. Finally, defining the generalized empirical distribution function, statistic of Kolmogorov-Smirnov two-sample test has been calculated accurately and they decided about acceptance or rejection of the null hypothesis.

TABLE 3. The generalized empirical distribution functions

c_i	$S_{10}(\widetilde{x})$	$S_{10}(\widetilde{y})$	$\lvert S_{10}(\widetilde{x}) - S_{10}(\widetilde{y})\rvert$
1	0.1	0.1	0
2	0.2	0.1	0.1
3	0.3	0.2	0.1
4	0.5	0.2	0.3
5	0.5	0.3	0.2
6	0.6	0.3	0.3
7	0.6	0.4	0.2
8	0.7	0.4	0.3
9	0.8	0.4	0.4
10	0.9	0.4	0.5
11	0.9	0.5	0.4
12	1	0.5	0.5
13	1	0.6	0.4
14	1	0.7	0.3
15	1	0.9	0.1
16	1	1	0

It should be noted that the results of Lin et al. approach has exactly been consistent with the results of the proposed approach [6] .

6. Conclusion

In this article, a new method for ranking fuzzy numbers based on the $D_{p,q}$ metric was proposed which plays a key role in nonparametric statistics. One of the most common tests for the goodness of fit of a statistical model is the Kolmogorov-Smirnov test. For this reason, we have examined Kolmogorov-Smirnov two-sample test in the case that observations are imprecise which is applicable for differences of two statistical distributions in terms of a variable. To do this, first, according to $D_{p,q}$ values of fuzzy observation in pooled sample, fuzzy observations in two sample were separated to distinct observation classes. Then, empirical distribution function was generalized in fuzzy environment. Finally, Generalized empirical distribution function is employed to obtain statistic of Kolmogorov-Smirnov two-sample test. The proposed approach can be easily used for all non-parametric tests in the fuzzy environment which is based on the ranking of fuzzy observations.

References

[1] D. Dubois, H. Prade, *Ranking of fuzzy numbers in the setting of possibility theory*, Information Sciences, **30** (1983), 183-224.
[2] P. Gaenssler and W. Stute, *Emperical process: a survey of results for independent and identically distributed random variables*, Ann. Probab, **7** (1979), 193-243.
[3] J. D. Gibbons, S. Chakraborti, *Non-Parametric Statistical Inference*, New York: Marcel Dekker (2003).
[4] E. Gine and J. Zinn, *Some limit theorem for emperical measure (with discussion)*, Ann. Probab, **12** (1984), 929-989.
[5] G. Hesamian and Chachi, *Kolmogorov-Smirnov two sample test for fuzzy random variables*, Statistical Papers, **56** (2015), 61-82.
[6] P. Lin, B. Wu and J. Watada *Kolmogorov-Smirnov two sample test with continuous fuzzy data* Advances in intelligent and soft computing, **68** (2010), 175-186.
[7] K. H. Lee, *First Course on Fuzzy Theory and Applications*, Heidelberg: Springer (2005).

Fereshteh Momeni
Department of Statistics, Behshahr branch, Islamic Azad University, Behshahr, Iran
Email: momeni@iaubeh.ac.ir

Bahram sadeghpour Gildeh
Department of Statistics, Faculty of Mathematical Sciences,
Ferdowsi University of Mashhad , Iran
Email: sadeghpour@um.ac.ir

Gholamreza Hesamian
Department of Statistics, University of Payamenoor, 19395-3697, Tehran, Iran
Email: gh.hesamian@pnu.ac.ir