Considerations on the Use of FDA Methods in Statistical Inference with Fuzzy Data

Przemyslaw Grzegorzewski[1,2][0000−0002−5191−4123] and Anna Kozak[1][0000−0001−8523−3759]

[1] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland
[2] Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-447, Warsaw, Poland
przemyslaw.grzegorzewski,anna.kozak@pw.edu.pl

Abstract. We consider the prospects of Functional Data Analysis (FDA) methods for statistical inference based on fuzzy data. To make the FDA-based reasoning efficient, we introduce another representation of fuzzy numbers than the traditionally used one, namely ICr functions. To show that the suggested approach can be effective, we propose a new test designed in the spirit of FDA methods, which has been shown to have good statistical properties.

Keywords: Functional data analysis · Fuzzy data · Fuzzy numbers · Random fuzzy numbers · Credibility distribution · Goodness-of-fit test.

1 Introduction

Fuzzy set theory has proven to be extremely useful and efficient in mathematical uncertainty modeling and many practical applications. One such application area is the statistical analysis of imprecise data, usually modeled with random fuzzy numbers. Unfortunately, constructing statistical tools for reasoning based on fuzzy data does not directly generalize classical procedures to the fuzzy domain. As it turns out, the specificity of fuzzy random variables means that certain properties or operations we accept as natural for the crisp random variables do not apply to random fuzzy numbers. In particular, in contrast to the statistical analysis of crisp data, one should be aware of the following disadvantages typical for fuzzy numbers: (a) problems with subtraction and division of fuzzy numbers; (b) the lack of universally accepted total ranking between fuzzy numbers; (c) there are not yet realistic suitable models for the distribution of random fuzzy numbers; (d) there are not yet Central Limit Theorems for random fuzzy numbers that can be directly applied to making an inference.

The above-mentioned inconveniences force statisticians to give up some of the structures they are used to, but at the same time, encourage them to come up with new solutions and thus trigger their creative potential. For example, the problems with (a) resulted in using distances to avoid subtracting fuzzy numbers (see, e.g. [2]). Difficulties with ordering fuzzy numbers do not allow using rank tests, which we are trying to replace with other structures (e.g. [13]). Finally, problems with the distribution of fuzzy random variables, mentioned in (c) and (d), drew researchers' attention to bootstrap (e.g. [6, 16]) and permutation tests (e.g. [8–10, 12]).

Since we can identify each fuzzy set with its membership function, hence we can perceive a fuzzy sample as a set of functions. If so, the question immediately arises: Why fuzzy data are not treated as special functional data? Consequently, why aren't Functional Data Analysis (FDA) methods used to analyze fuzzy data?

These questions are not new, but it seems that we still have not considered the problem with sufficient intensity. In 2012 González-Rodríguez et al. published an interesting article with a significant title "Fuzzy data treated as functional data. A one-way ANOVA test approach" [7], but it didn't get the broader attention it deserved. A few years later, M.A. Gil in her, SMPS 2018 Tutorial asked this question again: "Can fuzzy data be treated as special functional data?". She responded by pointing out two opposing positions. "Directly NO: In applying functional arithmetic to handle elements in the space of (functional-valued) fuzzy numbers, one often moves out of the space and the fuzzy meaning is generally lost". On the other hand, she claimed that "Indirectly, YES: By using appropriate arithmetic and suitable metrics, fuzzy numbers can be identified with elements in a convex cone of the Hilbert space of functions and the arithmetic and metrics with fuzzy numbers with those in the Hilbert space of functions".

If so, let's ask again: Why, so far, FDA methods have not met with the wider interest of researchers dealing with the analysis of fuzzy data? It appears that this may have happened because in previous studies, their authors, although they used some FDA-related tools, they stuck to the LU-representation of fuzzy numbers. Perhaps things will change if we look at fuzzy data a little differently, through a slightly different representation of fuzzy numbers, one that better suits the nature of FDA methods.

Therefore, the main goal of this contribution is to indicate a slightly different representation of fuzzy numbers than those we are used to and to show, using the example of a certain statistical problem, that this representation enables us to apply the FDA technique effectively.

The paper is organized as follows: in Section 2 we recall basic notation and information related to fuzzy numbers. Then, in Section 3 we present in a few words what the FDA is all about. In Section we introduce another representation of fuzzy numbers that seems to be promising for further use in fuzzy FDA methods. Next, in Section 5 we propose a goodness-of-fit two-sample test for fuzzy data that combines the suggested representation with some FDA testing approach, while in Section 6 we show some results of the simulation study related to this test.

2 Fuzzy data

In real-life experiments and datasets, we often meet imprecise observations. A suitable mathematical model of such data is a family of fuzzy numbers.

Definition 1. *(cf. [4]) A fuzzy subset A of the real line $\mathbb{R}$ with a membership function $\mu : \mathbb{R} \to [0,1]$ is a **fuzzy number** if it satisfies the following properties:*

(1) A is normal (i.e. $\exists x_0 \in \mathbb{R}$ such that $\mu_A(x_0) = 1$),
(2) A is fuzzy convex (i.e. $\mu(\lambda x_1 + (1-\lambda)x_2) \geqslant \min\{\mu(x_1), \mu(x_2)\}$ for any $x_1, x_2 \in \mathbb{R}$ and any $\lambda \in [0,1]$,
(3) μ is upper semicontinuous,
(4) the support of A, i.e. $\operatorname{supp}(A) = cl\{x \in \mathbb{R} : \mu(x) > 0\}$, is bounded (where cl stands for the closure operator).

When considering any two fuzzy sets, including fuzzy numbers, two crisp sets are of special interest, the **support**, defined above (which contains all elements of the universe of discourse that are compatible at some extent with the concept modeled by a given fuzzy set) and the **core**, containing all elements of the universe of discourse that surely belong to the fuzzy set under study (i.e. $\{x : \mu(x) = 1\}$). If A is a fuzzy number then, by Definition 1, $\operatorname{supp}(A)$ is a closed interval, and $\operatorname{core}(A)$ is not empty.

Let us denote by $\mathbb{F}(\mathbb{R})$ the space of all fuzzy numbers. Moreover, let $\mathbb{F}_c(\mathbb{R})$ denote a family of all so-called **continuous** fuzzy numbers, i.e. fuzzy numbers with continuous membership functions.

Each fuzzy number has two equivalent representations: the so-called **LR-representation** and **LU- representation**. Following the first one, the membership function μ of a fuzzy number A can be represented in the following form

$$\mu(x) = \begin{cases} L\left(\frac{b-x}{b-a}\right) & \text{if } a < x \leqslant b, \\ 1 & \text{if } b \leqslant x \leqslant c, \\ R\left(\frac{x-c}{d-c}\right) & \text{if } c \leqslant x < d, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $L, R : \mathbb{R} \to [0,1]$ are decreasing functions such that $L(0) = R(0) = 1$, $L(1) = R(1) = 0$, $L(x), R(x) < 1$ for all $x > 0$ and $L(x), R(x) > 0$ for all $x < 1$. Functions L and R are called the left and right **shape functions** (**sides** or **arms**), respectively. Hence, each fuzzy number in the LR-representation is specified completely by its support (i.e. $\operatorname{supp}(A) = [a,d]$), core (i.e. $\operatorname{core}(A) = [b,c]$), and its shape functions L and R. One can easily notice that if $L = R$ and $b - a = d - c$ then the corresponding fuzzy number is **symmetric**.

Following the LU-representation, a fuzzy number A with the membership function μ is completely characterized by a family of its α-cuts $\{A_\alpha\}_{\alpha \in [0,1]}$ defined as follows

$$A_\alpha = \begin{cases} \{x \in \mathbb{R} : \mu(x) \geqslant \alpha\} & \text{if } \alpha \in (0,1], \\ cl\{x \in \mathbb{R} : \mu(x) > 0\} & \text{if } \alpha = 0. \end{cases}$$

It is easily seen that each α-cut of a fuzzy number A is a nonempty compact interval $A_\alpha = [A^L_\alpha, A^U_\alpha]$, where $A^L_\alpha = \inf A_\alpha$ and $A^U_\alpha = \sup A_\alpha$ denote its lower and upper endpoint (border), respectively. Alternatively, each α-cut can be represented by its midpoint and radius (spread) given by

$$\mathrm{mid}A_\alpha = \frac{A_\alpha^L + A_\alpha^U}{2}, \quad \mathrm{spr}A_\alpha = \frac{A_\alpha^U - A_\alpha^L}{2},$$

instead of its endpoints (which are sometimes more convenient in processing fuzzy numbers, e.g. in defining a distance between fuzzy numbers, see [20]). Therefore, each fuzzy number in the LU-representation is specified completely by all lower and upper endpoints of the α-cuts (i.e. A_α^L, A_α^U, for $\alpha \in [0,1]$) or by the midpoints and spreads of all α-cuts (i.e. $\mathrm{mid}A_\alpha$ and $\mathrm{spr}A_\alpha$, for $\alpha \in [0,1]$).

Although membership functions of a fuzzy number may assume different shapes, some families of fuzzy numbers play a dominant role in considerations. In particular, the most often used are **trapezoidal fuzzy numbers** with the shape functions $L(x) = R(x) = \max\{0, 1 - x\}$. **Triangular fuzzy number** are special cases of trapezoidal fuzzy numbers with $b = c$. For more details on fuzzy numbers, their types, characteristics, and approximations we refer, e.g., to [1, 5].

Some authors limit the concept of fuzzy numbers only to those with a single-element core (i.e. satisfying $b = c$, by Definition 1), and call the other **fuzzy intervals**.

In our contribution, we restrict our attention to continuous fuzzy numbers having single-element cores, and, to avoid any misunderstanding, we will call such fuzzy numbers **regular fuzzy numbers**. A family of all regular fuzzy numbers will be denoted by $\mathbb{F}_r(\mathbb{R})$.

Since further on we consider fuzzy samples that are realizations of fuzzy random variables, let us also define a random fuzzy number. Indeed, statistical inference based on imprecise data requires a model that combines two kinds of uncertainty present in such data: imprecision (modeled by fuzzy sets) and randomness (expressed by probability theory). This brings us to the concept of a **fuzzy random variable**, also known as a **random fuzzy number** introduced by Puri and Ralescu [17].

Definition 2. *Let $(\Omega, \mathcal{A}, P)$ be a probability space. A mapping $X : \Omega \to \mathbb{F}(\mathbb{R})$ is a random fuzzy number if for all $\alpha \in [0,1]$ the α-level function is a compact random interval.*

3 A few words on FDA

By **Functional Data Analysis (FDA)** we mean all theoretical methods and practices relating to situations when the available data are not real numbers or vectors but **functions**. Thus, FDA usually refers to statistical problems where the available data consists of a sample of functions $\mathbb{x} = (x_1, \ldots, x_n)$, where $x_i = x_i(t)$, for each $i = 1, \ldots, n$, is defined on a compact interval of the real line, e.g. on the unit interval $[0, 1]$.

In FDA we usually assume that the sample space $\mathcal{X}$ is a real separable Banach space with some norm $||\cdot||$. Therefore, our sample data are observations drawn from an $\mathcal{X}$-valued random element X (i.e. a measurable function) defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Separability ensures that a linear combination of $\mathcal{X}$-valued random elements is again a random element. However, quite often a structure of (separable) Hilbert space, with associated inner product $\langle\,,\,\rangle$, is needed for $\mathcal{X}$.

Therefore, two standard choices for the sample space $\mathcal{X}$ are $\mathcal{C}[0,1]$, i.e. the Banach space of real continuous functions $x : [0,1] \to \mathbb{R}$ endowed with the supremum norm $||\cdot|| = \sup_t |x(t)|$, and the Hilbert space $L^2[0,1]$ of square integrable real functions on $[0,1]$ endowed with the usual inner product $\langle x, y\rangle = \int_0^1 x(t)y(t)\,dt$.

For more details on the FDA, we refer the reader to famous monographs [14, 18]. A brief overview of the FDA can be found in [3].

Now let's return to fuzzy numbers and ask which of the representations presented in Section 2 better fits the FDA requirements. Seemingly, both could be used, but as previous attempts show, it has not been possible to obtain results that would be interesting enough. Therefore, in the next section, we will propose yet another representation of fuzzy numbers, one that, in our opinion, fits the needs of the FDA perfectly.

4 ICr functions

When considering what we can do to make FDA methods more applicable in statistical inference based on fuzzy data, we found that another representation of fuzzy numbers than the two discussed in Section 2 would be useful. When thinking about finding such a representation that would have properties favorable from the FDA's point of view, we brought to mind a certain structure proposed by Liu [15] several years ago, called the credibility distribution.

Let $A \in \mathbb{F}(\mathbb{R})$ denote any fuzzy number with a membership function $\mu(x)$ given by (1), with shape functions L and R, $\mathrm{supp}(A) = [a, d]$ and $\mathrm{core}(A) = [b, c]$. Before defining the credibility distribution of A we have to extend the sides L and R of A to the real domain as follows

$$L_{ext}(x) = \begin{cases} 0 & \text{if } a < x, \\ L\left(\frac{b-x}{b-a}\right) & \text{if } a \leqslant x < b, \\ 1 & \text{if } b \leqslant x \end{cases} \tag{2}$$

$$R_{ext}(x) = \begin{cases} 1 & \text{if } x \leqslant c, \\ R\left(\frac{x-c}{d-c}\right) & \text{if } c < x \leqslant d, \\ 0 & \text{if } d < x. \end{cases} \tag{3}$$

Obviously, $\mu(x) = L_{ext}(x) - \big[1 - R_{ext}(x)\big]$ for any $x \in \mathbb{R}$.

Definition 3. *(Liu [15]) The* ***credibility distribution*** *of $A \in \mathbb{F}(\mathbb{R})$ is a function $\Upsilon : \mathbb{R} \to [0,1]$ defined by*

$$\Upsilon(x) = \frac{1}{2}\Big(L_{ext}(x) + \big[1 - R_{ext}(x)\big]\Big), \quad \forall x \in \mathbb{R}. \tag{4}$$

Looking for a justification for Definition 3 one may notice that Liu [15] defined the credibility distribution as the average of the possibility and necessity functions (see [22]), i.e.

$$\Upsilon(x) = \frac{1}{2}\Big(\mathrm{Pos}(x) + \mathrm{Nec}(x)\Big),$$

where

$$\mathrm{Pos}(x) = \sup_{t \leqslant x} \mu(t) = L_{ext}(x),$$
$$\mathrm{Nec}(x) = 1 - \sup_{t > x} \mu(t) = 1 - R_{ext}(x).$$

It is worth noting here that ten years later Stefanini and Guerra [19] considered a slightly more general construction, called **λ-Average Cumulative Function**, of which the credibility distribution is a special case. Indeed, the λ-Average Cumulative Function $\Psi^{(\lambda)} : \mathbb{R} \to [0, 1]$, where $\lambda \in [0, 1]$, is defined as follows

$$\Psi^{(\lambda)}(x) = (1 - \lambda)L_{ext}(x) + \lambda\big[1 - R_{ext}(x)\big]. \tag{5}$$

Hence $\Upsilon(x) = \Psi^{(\lambda)}(x)$ if and only if $\lambda = 1/2$.

A credibility distribution, although it has some interesting properties (is nondecreasing, takes values in the unit interval, etc.), it is still not the desired representation we are looking for. This is mainly because, in a given sample created by fuzzy numbers, each observation has a potentially different support.

Although we could continue our considerations at the same level of generality adopted so far, i.e. for any fuzzy numbers, in the following we will limit ourselves only to regular fuzzy numbers. We do so mainly to avoid some mathematical difficulties. But it is also easily justified because these types of fuzzy numbers are the most common in practice. Thus, for the regular fuzzy numbers the following theorem holds (cf. [23]).

Theorem 1. *$A \in \mathbb{F}_r(\mathbb{R})$ if and only if its credibility distribution $\Upsilon(x)$ is strictly increasing on $\{x \in \mathbb{R} : 0 < \Upsilon(x) < 1\}$. Moreover, $A \in \mathbb{F}_r(\mathbb{R})$ if and only if it has a unique inverse credibility distribution*

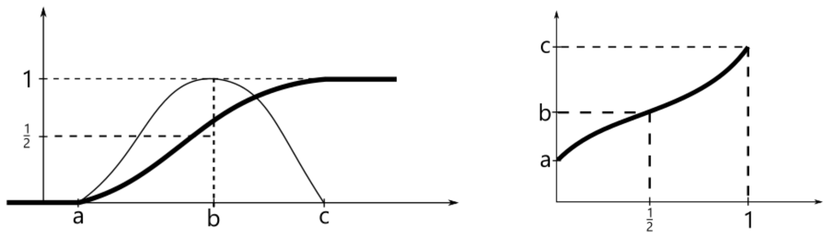
$$\upsilon(\alpha) := \Upsilon^{-1}(\alpha) \tag{6}$$

and $\upsilon(\alpha)$ is continuous and strictly increasing for $\alpha \in [0, 1]$.

Further on the **inverse credibility distribution**, defined by (6), will be abbreviated as the **ICr function**. An example of a regular fuzzy number and the corresponding credibility distribution is given in Figure 1.

Example 1. Consider a triangular fuzzy number A with the following membership function

$$\mu(x) = \begin{cases} \frac{x-a}{b-a} & \text{if } a < x \leqslant b, \\ \frac{c-x}{c-b} & \text{if } b \leqslant x < c, \\ 0 & \text{otherwise}, \end{cases} \tag{7}$$



(a) A membership function and the corresponding credibility distribution

(b) The corresponding ICr function

Fig. 1: The membership function of a regular fuzzy number, its credibility distribution $\Upsilon(x)$, and the corresponding ICr function $\upsilon(\alpha)$.

where $a < b < c$. By (4) its credibility distribution is given by

$$\Upsilon_X(x) = \begin{cases} 0 & \text{if } x <\leqslant a, \\ \frac{x-a}{2(b-a)} & \text{if } a < x \leqslant b, \\ \frac{x+c-2b}{2(c-b)} & \text{if } b < x \leqslant c, \\ 1 & \text{if } x > c, \end{cases} \quad (8)$$

while the ICr function of A is given as follows

$$\upsilon_X(\alpha) = \begin{cases} 2(b-a)\alpha + a & \text{if } 0 \leqslant \alpha < 0.5, \\ 2(c-b)\alpha + 2b - c & \text{if } 0.5 \leqslant \alpha \leqslant 1. \end{cases} \quad (9)$$

It seems that the transformation of fuzzy data consisting of moving from the membership function to the ICr function brings many profits, in particular, those that will favor the use of FDA methods for statistical reasoning based on fuzzy data. Let us briefly list some basic advantages of the indicated representation using the ICr functions. Firstly, there is a one-to-one relationship between both representations of fuzzy data. Secondly, ICr functions have some analytical properties favoring the use of FDA methods, for instance, they are non-decreasing and their common support is the unit interval $[0, 1]$. Moreover, the ICr functions of the regular fuzzy numbers are continuous in $[0, 1]$, and a value of ICr for $1/2$ indicates always the mode of the corresponding membership function.

Credibility distributions and their inverses were applied in fuzzy programming [23] and in system reliability analysis [21]. However, to our knowledge, no one yet has tried to use them as a convenient device allowing us to apply FDA methods to fuzzy data.

5 Two-sample goodness-of-fit test based on ICr functions

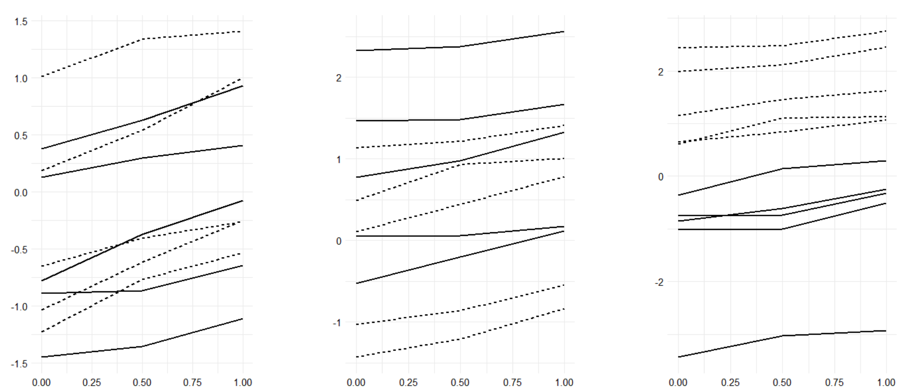
To show that FDA methods can be effectively used in fuzzy settings, we propose the construction of a two-sample goodness-of-fit test for fuzzy data, based on ICr functions.

Let $\mathbb{X} = (X_1, \ldots, X_n)$ and $\mathbb{Y} = (Y_1, \ldots, Y_m)$ denote two independent random fuzzy samples drawn from populations with unknown distributions. We want to verify the null hypothesis that both samples are identically distributed against the alternative that these two distributions differ significantly. Thus we consider the following testing problem

$$\begin{cases} H_0 : X \overset{d}{=} Y, \\ H_1 : \neg H_0. \end{cases}$$

Let $\mathbb{x} = (x_1, \ldots, x_n)$ and $\mathbb{y} = (y_1, \ldots, y_m)$ denote the experimental realizations of the considered independent random fuzzy samples, where $x_i \in \mathbb{F}_r(\mathbb{R})$, for $i = 1, \ldots, n$ and $y_j \in \mathbb{F}_r(\mathbb{R})$, for $j = 1, \ldots, m$.

In the first step, we determine the credibility distributions for each observation in both samples, and as a result we obtain $\boldsymbol{\Upsilon}_x = \big(\Upsilon_{x_1}(t), \ldots, \Upsilon_{x_n}(t)\big)$ and $\boldsymbol{\Upsilon}_y = \big(\Upsilon_{y_1}(t), \ldots, \Upsilon_{y_m}(t)\big)$, where $t \in \mathbb{R}$. Then we determine the ICr functions for each observation and as a result, we obtain two sets of functions that will be further used to perform the test, i.e. $\boldsymbol{\upsilon}_x = \big(\upsilon_{x_1}(\alpha), \ldots, \upsilon_{x_n}(\alpha)\big)$ and $\boldsymbol{\upsilon}_y = \big(\upsilon_{y_1}(\alpha), \ldots, \upsilon_{y_m}(\alpha)\big)$, $\alpha \in [0, 1]$, where $\upsilon_{x_i}(\alpha) := \Upsilon_{x_i}^{-1}(\alpha)$ and $\upsilon_{y_j}(\alpha) := \Upsilon_{y_j}^{-1}(\alpha)$.



(a) (b) (c)

Fig. 2: Three datasets illustrating different arrangements of ICr functions: $\boldsymbol{\upsilon}_x = \big(\upsilon_{x_1}(\alpha), \ldots, \upsilon_{x_n}(\alpha)\big)$ (solid lines) and $\boldsymbol{\upsilon}_y = \big(\upsilon_{y_1}(\alpha), \ldots, \upsilon_{y_m}(\alpha)\big)$ (dashed lines).

In Figure 2 we show three datasets illustrating various experimental situations. In Fig. 2 (a) we can see two well-mixed curves suggesting that there is no reason for rejecting H_0. The situation illustrated in Fig. 2 (b) is not conclusive at first glance – without a test, we cannot say whether samples come from the same distribution or not. Finally, samples in Fig. 2 (c) are so much separated that we are inclined to conclude that the samples come from different distributions.

Once we have properly prepared the data, we can start constructing the desired statistical test.

Given $\boldsymbol{\upsilon}_x = \big(\upsilon_{x_1}(\alpha), \dots, \upsilon_{x_n}(\alpha)\big)$ and $\boldsymbol{\upsilon}_y = \big(\upsilon_{y_1}(\alpha), \dots, \upsilon_{y_m}(\alpha)\big)$ we consider for each $\alpha \in [0,1]$ the following pointwise test statistic

$$T(\alpha) = T(\boldsymbol{\upsilon}_x, \boldsymbol{\upsilon}_y; \alpha) = \frac{|\overline{\upsilon}_x(\alpha) - \overline{\upsilon}_y(\alpha)|}{\sqrt{\frac{1}{n}s_x^2(\alpha) + \frac{1}{m}s_y^2(\alpha)}}, \quad \alpha \in [0,1], \tag{10}$$

where

$$\begin{aligned} \overline{\upsilon}_x(\alpha) &= \frac{1}{n}\sum_{i=1}^{n} \upsilon_{x_i}(\alpha), \quad s_x^2(\alpha) = \frac{1}{n-1}\sum_{i=1}^{n}\Big[\upsilon_{x_i}(\alpha) - \overline{\upsilon}_x(\alpha)\Big]^2, \\ \overline{\upsilon}_y(\alpha) &= \frac{1}{m}\sum_{j=1}^{m} \upsilon_{y_j}(\alpha), \quad s_y^2(\alpha) = \frac{1}{m-1}\sum_{j=1}^{m}\Big[\upsilon_{y_j}(\alpha) - \overline{\upsilon}_y(\alpha)\Big]^2, \end{aligned} \tag{11}$$

while the final value of our test statistic for given samples $\boldsymbol{\upsilon}_x$ and $\boldsymbol{\upsilon}_y$ is determined as follows

$$t_0 = t_0(\boldsymbol{\upsilon}_x, \boldsymbol{\upsilon}_y) = \sup_{\alpha \in [0,1]} T(\alpha). \tag{12}$$

To conclude whether to reject or not the null hypothesis we need either a critical value to be compared with (12) or the p-value calculated for the obtained value of our test statistic (12). In the proposed test, we will determine the p-value. For this purpose, we will use an approach typical of permutation tests.

Let $\boldsymbol{w} := \boldsymbol{\upsilon}_x \uplus \boldsymbol{\upsilon}_y$, where $\uplus$ stands for vector concatenation pooling the two samples into one, i.e.

$$w_i = \upsilon_{x_i} \text{ if } 1 \leqslant i \leqslant n \quad \text{and} \quad w_i = \upsilon_{y_{i-n}} \text{ if } n+1 \leqslant i \leqslant N,$$

where $N = n + m$. Let $\boldsymbol{w}^*$ denote a permutation of the initial dataset $\boldsymbol{w}$.

Suppose, we take first n elements of $\boldsymbol{w}^*$ and assign them to sample $\boldsymbol{\upsilon}_x^*$, while the remaining m elements create the second sample $\boldsymbol{\upsilon}_y^*$. Thus, it works like a random assignment of $N = n + m$ elements into two samples of the size n and m, respectively.

Next, using formulas (10)–(12) we calculate the corresponding value of the test statistic for $\boldsymbol{\upsilon}_x^* = \big(\upsilon_{x_1}^*, \dots, \upsilon_{x_n}^*\big)$ and $\boldsymbol{\upsilon}_y^* = \big(\upsilon_{y_1},^* \dots, \upsilon_{y_m}^*\big)$, i.e.

$$T(\alpha) = T(\boldsymbol{\upsilon}_x^*, \boldsymbol{\upsilon}_y^*; \alpha) = \frac{|\overline{\upsilon}_x^*(\alpha) - \overline{\upsilon}_y^*(\alpha)|}{\sqrt{\frac{1}{n}s_{x^*}^2(\alpha) + \frac{1}{m}s_{y^*}^2(\alpha)}}, \tag{13}$$

$$t^* = \sup_{\alpha \in [0,1]} T^*(\alpha). \tag{14}$$

By repeating the whole procedure B times we obtain from (14) test statistic values t_b^*, for $b = 1, \dots, B$ permutations, to determine the approximate p-value

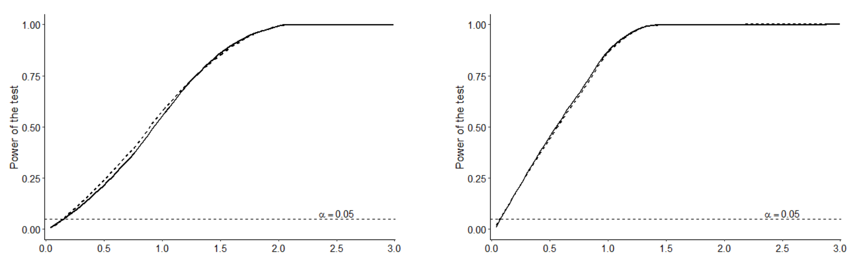
$$\text{p-value} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left(t_b^* \geqslant t_0\right), \tag{15}$$

where t_0 is the test statistic value (12) received for the original samples.

6 Simulation study

To examine some properties of the proposed goodness-of-fit test we conducted a simulation study. Below we show some of its results related to situations when both samples $\mathbb{X} = (X_1, \dots, X_n)$ and $\mathbb{Y} = (Y_1, \dots, Y_m)$ were triangular random fuzzy numbers. The cores of $\mathbb{X}$ and $\mathbb{Y}$ observations were generated from the normal distribution $\mathrm{N}(\xi_x, \sigma_x)$ and $\mathrm{N}(\xi_y, \sigma_y)$, respectively, while the distances between the core and both endpoints of the support were simulated from the uniform distribution.

Firstly, to examine the stability of the test size (i.e. the supremum of the probability of making a type I error) many repetitions of the test at 5% significance level performed under H_0 were considered. In each test, $B = 1000$ permutations were drawn and the empirical percentages of rejections under H_0 were determined. The results obtained both for equal and nonequal sample sizes showed that the size of our test is stable at the desired level.



(a) Sample sizes $n = m = 10$ (b) Sample sizes $n = m = 20$

Fig. 3: Power functions for the proposed test (solid lines) and the distance-based test [8] (dashed lines).

Next, we conducted a power study to compare the proposed test with the goodness-of-fit test based on the distance between sample averages obtained for both samples [8]. Starting from the situation when both samples come from the same distribution and then by enlarging the difference in location between samples, we estimated the probability of rejection of the null hypothesis H_0. Experiments carried out for samples of different sizes showed that the proposed test is generally equivalent to the distance-based test (some power curves obtained for the considered two tests are shown in Figure 3). This is a good sign for our new test because its competitor [8] is known for its good power, which many other studies confirm (e.g. [12, 13]).

7 Conclusions and further research

In this paper, we considered the prospects of using FDA methods for the statistical analysis of fuzzy data. We believe that the proposed new representation of fuzzy numbers, i.e. ICr functions, can significantly contribute to this goal. The test proposed in this paper, as an example of the use of FDA methods, obviously requires further research. But, it seems, that both it and other tools created in the spirit of FDA will allow us to expand the horizons of reasoning based on fuzzy data to new areas.

References

1. Ban, A.I., Coroianu, L., Grzegorzewski, P.: Fuzzy Numbers: Approximations, Ranking and Applications. Polish Academy of Sciences, Warsaw (2015)
2. Blanco-Fernández, A., Casals, M., Colubi, A., Corral, N., García-Bárzana, M., Gil, M.-A., González-Rodríguez, G., López, M., Lubiano, M., Montenegro, M., Ramos-Guajardo, A.-B., de la Rosa de Sáa, S., Sinova, B.: A distance-based statistical analysis of fuzzy number-valued data. International Journal of Approximate Reasoning **55**(7), 1487–1501 (2014)
3. Cuevas A.: A partial overview of the theory of statistics with functional data. Journal of Statistical Planning and Inference **147**, 1–23 (2014)
4. Dubois, D., Prade, H.: Operations on fuzzy numbers. International Journal of Systems Science **9**, 613–626 (1978)
5. Dubois, D., Prade, H.: Fuzzy numbers: an overview. Readings in Fuzzy Sets for Intelligent Systems, Morgan Kaufmann, pp. 112–148 (1993)
6. González-Rodríguez, G., Colubi, A., Gil, M.-A.: Bootstrap techniques and fuzzy random variables: Synergy in hypothesis testing with fuzzy data. Fuzzy Sets and Systems **157**(19), 2608–2613 (2006).
7. González-Rodríguez, G., Colubi, A., Gil, M.-A.: Fuzzy data treated as functional data. A one-way ANOVA test approach. Computational Statistics and Data Analysis **56**, 943–955 (2012)
8. Grzegorzewski, P.: Permutation k-sample goodness-of-fit test for fuzzy data. In: 2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8 (2020). https://doi.org/10.1109/FUZZ48607.2020.9177765
9. Grzegorzewski, P.: Two-sample dispersion problem for fuzzy data. In: Lesot, M.-J., Vieira, S., Reformat, M.-Z., Carvalho, J.-P., Wilbik, A., Bouchon-Meunier, B., Yager, R.-R. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 82–96. Springer (2020). https://doi.org/10.1007/978-3-030-50153-2_7
10. Grzegorzewski, P.: Two-sample test for comparing ambiguity in fuzzy data. In: 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–8 (2022). https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882757

11. Grzegorzewski, P.: Paired sample test for fuzzy data. In: García-Escudero, L.A., Gordaliza, A., Mayo, A., Lubiano Gomez, M.A., Gil, M.A., Grzegorzewski, Przemyslaw, Hryniewicz, O. (eds.) Building Bridges between Soft and Statistical Methodologies for Data Science, pp. 200–207. Springer (2023). https://doi.org/10.1007/978-3-031-15509-3_27
12. Grzegorzewski, P., Gadomska, O.: Nearest neighbor tests for fuzzy data. In: 2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE). pp. 1–6 (2021). https://doi.org/10.1109/FUZZ45933.2021.9494432
13. Grzegorzewski, P., Zacharczuk, M.: A new two-sample location test for fuzzy data, In: Massanet, S., Montes, S., Ruiz-Aguilera, D., González-Hidalgo, M. (eds), Fuzzy Logic and Technology, and Aggregation Operators (EUSFLAT 2023, AGOP 2023), Lecture Notes in Computer Science, vol. 14069, Springer, 2023, pp. 737–748.
14. Horváth, L., Kokoszka, P.: Inference for Functional Data with Applications. Springer (2012)
15. Liu B.: Uncertainty Theory. Springer (2007)
16. Montenegro, M., Colubi, A., Casals, M., Gil, M.-A.: Asymptotic and bootstrap techniques for testing the expected value of a fuzzy random variable. Metrika **59**(1), 31-49 (2004)
17. Puri, M.-L., Ralescu, D.-A.: Fuzzy random variables. Journal of Mathematical Analysis and Applications **114**(2), 409–422 (1986)
18. Ramsay, J.-O., Silverman, B.-W.: Functional Data Analysis. Springer (2005)
19. Stefanini, L., Guerra, M.-L.: On possibilistic representations of fuzzy intervals. Information Sciences **405**, 33–54 (2017)
20. Trutschnig W., González-Rodríguez, G., Colubi, A., Gil, M.-A.: A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Information Sciences **179**(23), 3964–3972 (2009)
21. Wang, K., Zhou, J., Ralescu D.-A.: Arithmetic operations for LR mixed fuzzy random variables via mean chance measure with applications. Journal of Intelligent & Fuzzy Systems **32**, 451–466 (2017)
22. Zadeh L.A.: Fuzzy sets as a basis for a theory of possibility. Fuzzy Sets and Systems **1**, 3–28 (1978)
23. Zhou, J., Yang, F., Wang, K.: Fuzzy arithmetic on LR fuzzy numbers with applications to fuzzy programming. Journal of Intelligent & Fuzzy Systems **30**, 71–87 (2016)