Zastosowanie metod funkcjonalnej analizy danych do wnioskowania na podstawie nieprecyzyjnych obserwacji

Przemysław Grzegorzewski

Wydział Matematyki i Nauk Informacyjnych, Politechnika Warszawska
Instytut Badań Systemowych, Polska Akademia Nauk

IV Kongres Statystyki Polskiej
Warszawa, 2–4.VII.2024

Motivations

Motivations

A general development of statistical theory

Statistical theory	$\mathcal{X}$	Θ	Dating to
Classical parametric inf.	$\mathbb{R}$	$\theta \subset \mathbb{R}$	1920s
Multivariate analysis	$\mathbb{R}^p$ $(n \gg p)$	$\theta \subset \mathbb{R}^k$ $(n \gg k)$	1940s
Nonparametrics	$\mathbb{R}^p$ $(n \gg p)$	A function space	1960s
High dimensional problems	$\mathbb{R}^p$ $(n < p)$	$\theta \subset \mathbb{R}^k$	2000s
Functional Data Analysis	A function space	$\mathbb{R}^k$ or a function space	1990s

Antonio Cuevas (2014)

Motivations

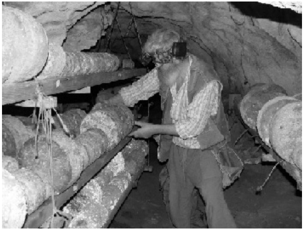
A general development of statistical theory

Statistical theory	$\mathcal{X}$	Θ	Dating to
Classical parametric inf.	$\mathbb{R}$	$\theta \subset \mathbb{R}$	1920s
Multivariate analysis	$\mathbb{R}^p$ $(n \gg p)$	$\theta \subset \mathbb{R}^k$ $(n \gg k)$	1940s
Nonparametrics	$\mathbb{R}^p$ $(n \gg p)$	A function space	1960s
High dimensional problems	$\mathbb{R}^p$ $(n < p)$	$\theta \subset \mathbb{R}^k$	2000s
Functional Data Analysis	A function space	$\mathbb{R}^k$ or a function space	1990s

Antonio Cuevas (2014)

- sound and images
- natural language
- imprecise data
- ...

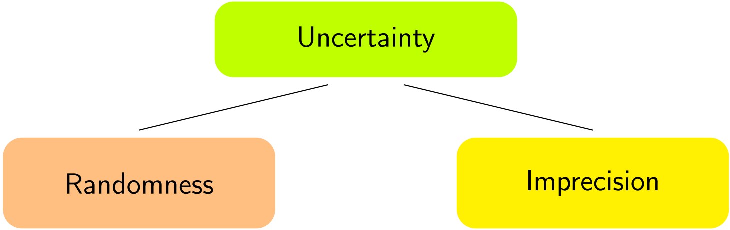
Example



The Gamonedo cheese is a kind of a blue cheese produced in Asturias. In quality control experts (tasters) express their perceptions about

- visual parameters (shape, rind, appearance),
- texture parameters (hardness and crumbliness),
- olfactory-gustatory parameters (smell intensity, smell quality, flavour intensity, flavour quality and aftertaste),
- an overall impression of the cheese.

(Gonzalez de Llano D., et al., 1992)

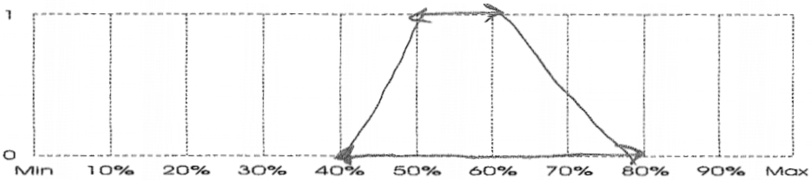

Uncertainty
Randomness
Imprecision

Outline

Imprecise data modeling

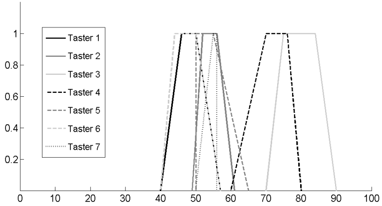
Example (cont.)

So far, the experts provide an ordinal number ranging from 1 to 7 to describe their perceptions about different cheese characteristics. Recently some of the tasters were proposed to express their subjective perceptions about the quality of the Gamonedo cheese by using fuzzy numbers.

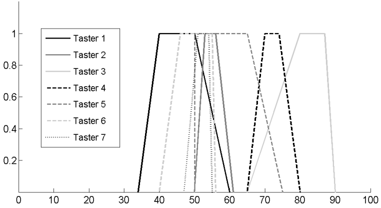


Opinion of a taster given by means of a trapezoidal fuzzy set

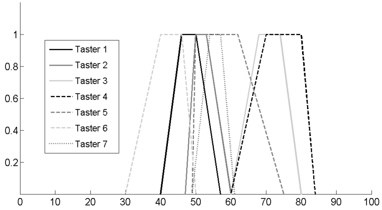
(Ramos-Guajardo A.B., et al., 2019)



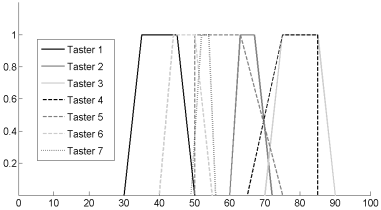
Appearance



Smell quality



Flavour quality



Overall impression

Fuzzy data in statistical context

Fuzzy data in statistical context

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. $\mathrm{supp}(A) = cl\{x \in \mathbb{R} : \mu(x) > 0\}$ is bounded (where cl stands for the closure operator).

Each fuzzy number has two equivalent representations: the so-called **LR-representation** and **LU- representation**.

LR-representation

The membership function μ of a fuzzy number A can be represented in the following form

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

where $L, R : \mathbb{R} \to [0, 1]$ denote decreasing functions such that $L(0) = R(0) = 1$, $L(1) = R(1) = 0$, $L(x), R(x) < 1$, $\forall x > 0$ and $L(x), R(x) > 0$, $\forall x < 1$.

Hence, A can be specified completely by its core (i.e. $\mathrm{core}(X) = [b, c]$), support (i.e. $\mathrm{supp}(X) = [a, d]$), and functions L, R, called the left and right shape functions (sides), respectively.

LU-representation

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

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

Thus 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 $A^U_\alpha = \sup A_\alpha$.

LU-representation

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

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

Thus each α-cut of a fuzzy number A is a nonempty compact interval $A_\alpha = [A_\alpha^L, A_\alpha^U]$, where $A_\alpha^L = \inf A_\alpha$ and $A_\alpha^U = \sup A_\alpha$.

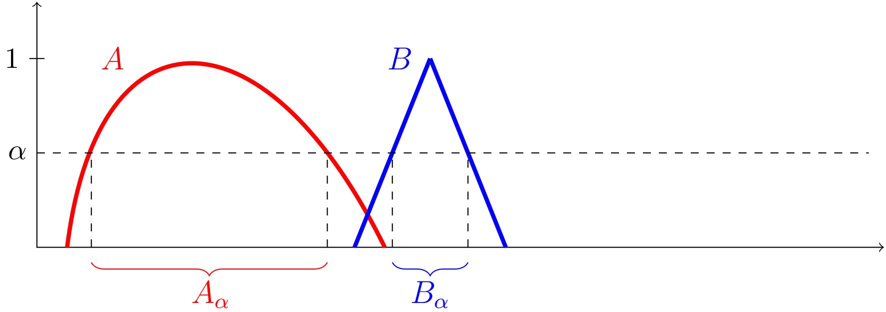
Alternatively, each α-cut can be represented by its midpoint and radius given by

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

instead of its endpoints.

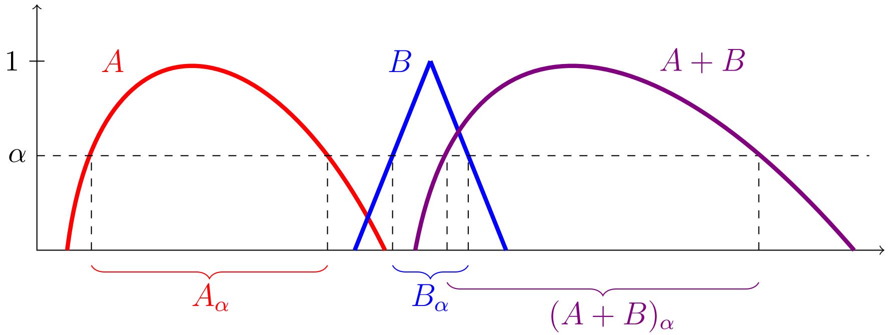
Arithmetic in $\mathbb{F}(\mathbb{R})$

$$\big(A+B\big)_\alpha = \big[\,\inf A_\alpha + \inf B_\alpha,\, \sup A_\alpha + \sup B_\alpha\big], \quad \forall \alpha \in [0,1]$$

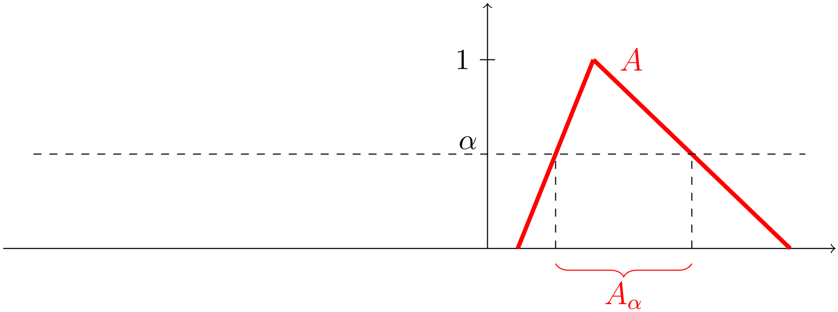


Arithmetic in $\mathbb{F}(\mathbb{R})$

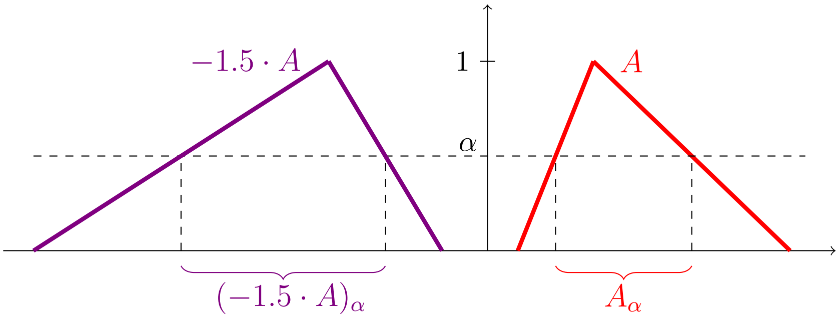
$$\big(A+B\big)_\alpha = \big[\inf A_\alpha + \inf B_\alpha,\, \sup A_\alpha + \sup B_\alpha\big], \quad \forall \alpha \in [0,1]$$



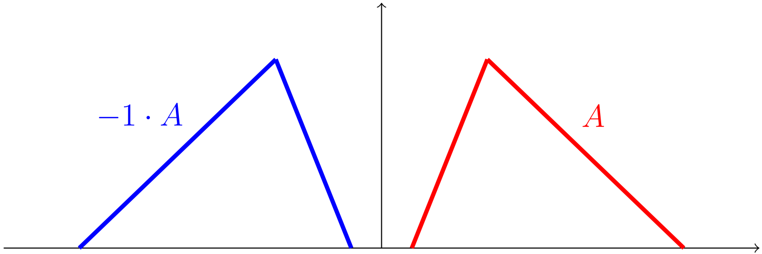
$$(\theta \cdot A)_\alpha = \begin{cases} [\theta \inf A_\alpha,\, \theta \sup A_\alpha] & \text{if } \theta > 0 \\ [\theta \sup A_\alpha,\, \theta \inf A_\alpha] & \text{if } \theta < 0 \end{cases}, \qquad \forall \alpha \in [0,1]$$



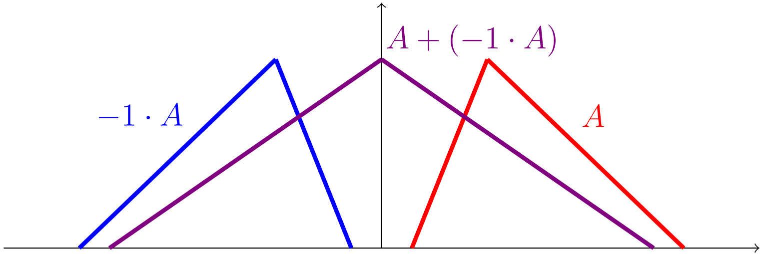
$$\left(\theta \cdot A\right)_\alpha = \begin{cases} \left[\theta \inf A_\alpha,\ \theta \sup A_\alpha\right] & \text{if } \theta > 0 \\ \left[\theta \sup A_\alpha,\ \theta \inf A_\alpha\right] & \text{if } \theta < 0 \end{cases}, \qquad \forall \alpha \in [0,1]$$



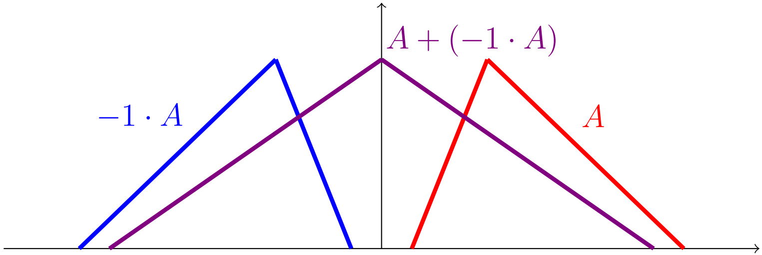
Note $(\mathbb{F}(\mathbb{R}), +, \cdot)$ has not linear but semilinear structure since $A + (-1 \cdot A) \neq \mathbb{1}_{\{0\}}$.



Note $(\mathbb{F}(\mathbb{R}), +, \cdot)$ has not linear but semilinear structure since $A + (-1 \cdot A) \neq \mathbb{1}_{\{0\}}$.



Note $(\mathbb{F}(\mathbb{R}), +, \cdot)$ has not linear but semilinear structure since $A + (-1 \cdot A) \neq \mathbb{1}_{\{0\}}$.



Moreover, the Minkowski difference does not satisfy, in general, the addition/subtraction property that $(A + (-1)B) + B = A$.

To overcome this problem the Hukuhara difference was proposed

$$C := A -_{Hu} B \quad \text{if and only if} \quad B + C = A$$

Although now the desired properties $A -_{Hu} A = \mathbb{1}_{\{0\}}$ is satisfied, the Hukuhara difference does not always exist.

To overcome this problem the Hukuhara difference was proposed

$$C := A -_{Hu} B \quad \text{if and only if} \quad B + C = A$$

Although now the desired properties $A -_{Hu} A = \mathbb{1}_{\{0\}}$ is satisfied, the Hukuhara difference does not always exist.

Remark

There is no such difference in $\mathbb{F}(\mathbb{R})$ that

- $(A - B) + B = A$ is satisfied $\forall A, B \in \mathbb{F}(\mathbb{R})$,
- the operation is well-defined $\forall A, B \in \mathbb{F}(\mathbb{R})$.

To overcome this problem the Hukuhara difference was proposed

$$C := A -_{Hu} B \quad \text{if and only if} \quad B + C = A$$

Although now the desired properties $A -_{Hu} A = \mathbb{1}_{\{0\}}$ is satisfied, the Hukuhara difference does not always exist.

Remark

There is no such difference in $\mathbb{F}(\mathbb{R})$ that

- $(A - B) + B = A$ is satisfied $\forall A, B \in \mathbb{F}(\mathbb{R})$,
- the operation is well-defined $\forall A, B \in \mathbb{F}(\mathbb{R})$.

Remark

Similar problems arise in the case of division.

Let λ denote a normalized measure associated with a continuous distribution with support in $[0,1]$ and let $\theta > 0$. Then for any $A, B \in \mathbb{F}(\mathbb{R})$ we define a metric D_γ^λ as follows

$$D_\gamma^\lambda(A,B) = \left[\int_0^1 \left[(\mathrm{mid} A_\alpha - \mathrm{mid} B_\alpha)^2 + \gamma (\mathrm{spr} A_\alpha - \mathrm{spr} B_\alpha)^2 \right] d\lambda(\alpha) \right]^{1/2}$$

(Gil et al., 2002; Trutschnig et al., 2009)

Typical choices:
$\gamma = 1$ or $\gamma = \frac{1}{3}$; $\lambda = \ell =$ the Lebesgue measure on $[0,1]$.

$(\mathbb{F}(\mathbb{R}), D_\gamma^\lambda)$ is a separable metric space and for each fixed λ all metrics D_γ^λ are topologically equivalent.

Fuzzy random variables

Fuzzy random variables

Fuzzy random variables (random fuzzy numbers) integrate **randomness** (associated with data generation) and **fuzziness** (associated with data nature).

Definition (Puri M.L., Ralescu D., 1986)

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

In other words, X is a fuzzy random variable if and only if X is a Borel measurable function w.r.t. the Borel σ-field generated by the topology induced by D_γ^λ.

Note

In contrast to the statistical analysis of numerical data one should be aware of the following problems typical for fuzzy data:

- problems with subtraction and division of fuzzy numbers;
- the lack of universally accepted total ranking between fuzzy numbers;
- there are not yet realistic suitable models for the distribution of random fuzzy numbers;
- there are not yet Central Limit Theorems for random fuzzy numbers that can be directly applied for making inference.

Note

In contrast to the statistical analysis of numerical data one should be aware of the following problems typical for fuzzy data:

- problems with subtraction and division of fuzzy numbers;
- the lack of universally accepted total ranking between fuzzy numbers;
- there are not yet realistic suitable models for the distribution of random fuzzy numbers;
- there are not yet Central Limit Theorems for random fuzzy numbers that can be directly applied for making inference.

Conclusion

No straightforward generalizations of the classical parametric/nonparametric statistical tests for fuzzy data exist.

A few words on FDA

A few words on FDA

Functional Data Analysis (FDA) – all theoretical methods and practice 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 $[0, 1]$.

Ramsay J.O., Silverman B.W., *Functional Data Analysis*, Springer, 2005.

Cuevas A., *A partial overview of the theory of statistics with functional data*, Journal of Statistical Planning and Inference, 147 (2014), 1–23.

We 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.

Very often a structure of (separable) Hilbert space, with associated inner product $\langle \, , \, \rangle$, is needed for $\mathcal{X}$.

Two standard choices for the sample space $\mathcal{X}$ are $\mathcal{C}[0,1]$, 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 $\langle x, y\rangle = \int_0^1 x(t)y(t)\, dt$.

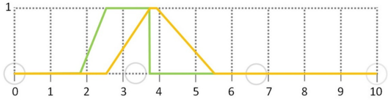
Can fuzzy data be treated as special functional data?

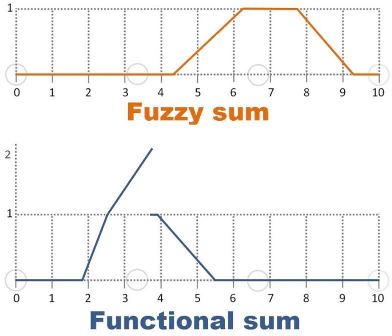
Can fuzzy data be treated as special functional data?

"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".

Can fuzzy data be treated as special functional data?

"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".





Maria Angeles Gil, SMPS 2018 Tutorial: *Random fuzzy sets and statistics with imprecise-valued data*

"INDIRECTLY, YES: By using appropriate arithmetic and suitable metrics, fuzzy numbers can be identified with elements in a convex cone of a Hilbert space of functions and the arithmetic and metrics with fuzzy numbers with those in the Hilbert space of functions".

Maria Angeles Gil, SMPS 2018 Tutorial: *Random fuzzy sets and statistics with imprecise-valued data*

González-Rodríguez G., Colubi A., Gil M.A., *Fuzzy data treated as functional data. A one-way ANOVA test approach*, Comp. Stat. Data Anal. 56 (2012), 943–955.

How can we apply FDA methods for fuzzy data?

How can we apply FDA methods for fuzzy data?

Firstly, we need a useful representation of fuzzy numbers.

Then we can try FDA with FNs.

ICr functions

ICr functions

Denoted by $\mathbb{F}(\mathbb{R})$ the space of all continuous fuzzy numbers, i.e. fuzzy numbers with continuous membership functions.

ICr functions

Denoted by $\mathbb{F}(\mathbb{R})$ the space of all continuous fuzzy numbers, i.e. fuzzy numbers with continuous membership functions.

Let us extend the sides L and R of $A \in \mathbb{F}(\mathbb{R})$ to the real domain as follows

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

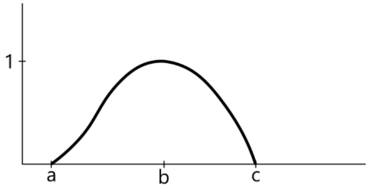
$$R_{ext}(x) = \begin{cases} 1 & \text{if } \ x \leqslant c, \\ R\left(\frac{x-c}{d-c}\right) & \text{if } \ c < x \leqslant d, \\ 0 & \text{if } \ d < x. \end{cases}$$

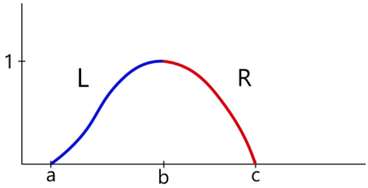
Obviously, $\mu(x) = L_{ext}(x) - \left[1 - R_{ext}(x)\right] \ \forall x \in \mathbb{R}$.

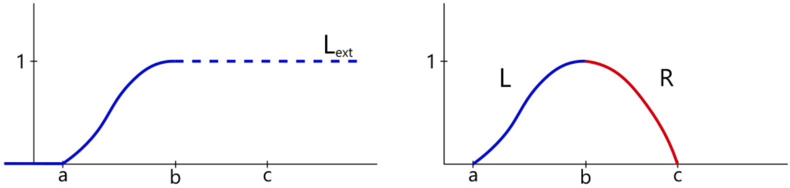
Definition (Liu, 2007)

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

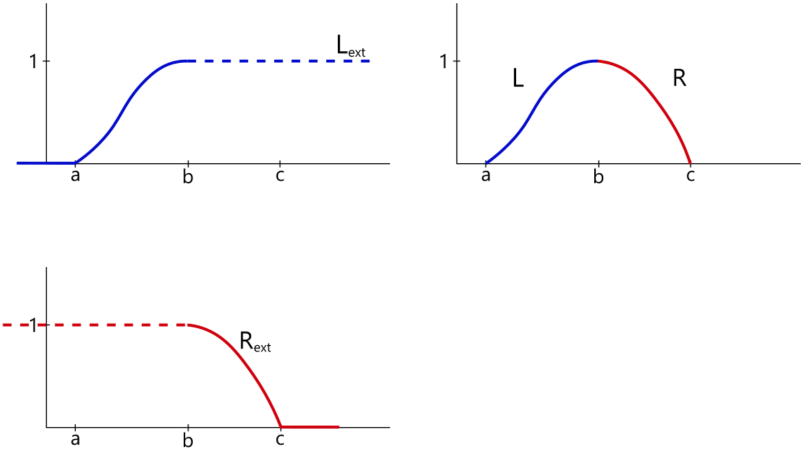
$$\Upsilon(x) = \frac{1}{2}\Big(L_{ext}(x) + \big[1 - R_{ext}(x)\big]\Big), \quad \forall x \in \mathbb{R}.$$


1
a
b
c

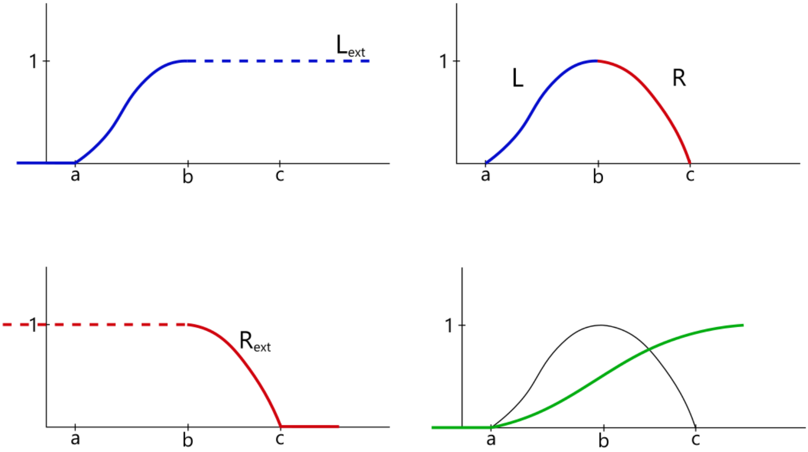

1
L
R
a
b
c



1
L
ext
a
b
c
1
L
R
a
b
c



1
L_{ext}
a
b
c
1
L
R
a
b
c
1
R_{ext}
a
b
c



1
a
b
c
L_ext
1
L
R
a
b
c
1
R_ext
a
b
c
1
a
b
c

Note

Liu (2007) defined the credibility distribution as the average of the possibility and necessity functions, i.e.

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

where

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

Note

Liu (2007) defined the credibility distribution as the average of the possibility and necessity functions, i.e.

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

where

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

Stefanini and Guerra (2017) considered the so-called λ-**Average Cumulative Function** (where $\lambda \in [0,1]$) defined as follows

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

Let $\mathbb{F}_r(\mathbb{R})$ denote a family of **regular** continuous fuzzy numbers, i.e. fuzzy numbers with continuous membership functions having single-element cores.

Let $\mathbb{F}_r(\mathbb{R})$ denote a family of **regular** continuous fuzzy numbers, i.e. fuzzy numbers with continuous membership functions having single-element cores.

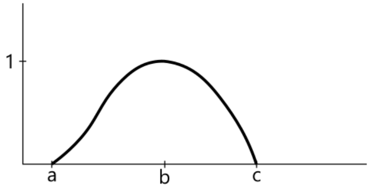
Theorem

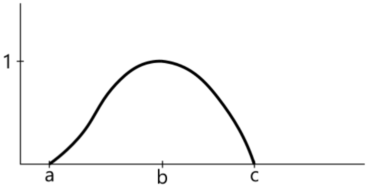
$A \in \mathbb{F}_r(\mathbb{R})$ if and only if its credibility distribution $\Upsilon(x)$ is strictly increasing on $\{x \in \mathbb{R} : 0 < \Upsilon(x) < 1\}$.

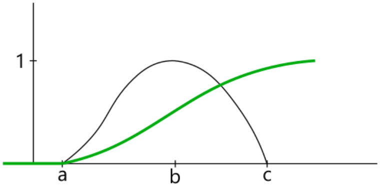
Moreover, $A \in \mathbb{F}_r(\mathbb{R})$ if and only if it has a unique **inverse credibility distribution** (abbreviated as ICr function)

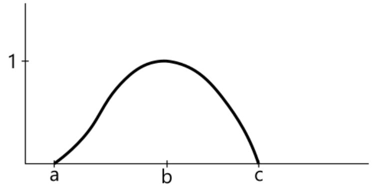
$$v(\alpha) := \Upsilon^{-1}(\alpha)$$

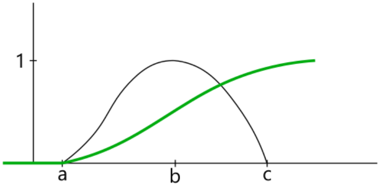
and $v(\alpha)$ is continuous and strictly increasing for $\alpha \in [0, 1]$.

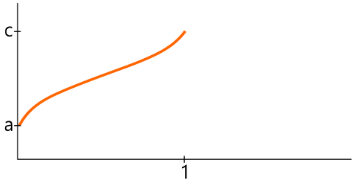

1
a
b
c


1
a
b
c


1
a
b
c


1
a
b
c


1
a
b
c


c
a
1

FDA in hypothesis testing with fuzzy data

FDA in hypothesis testing with fuzzy data

Let $\mathbb{X} = (X_1, \ldots, X_n)$ and $\mathbb{Y} = (Y_1, \ldots, Y_m)$ denote independent samples of i.i.d. random fuzzy numbers. We want to verify

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

FDA in hypothesis testing with fuzzy data

Let $\mathbb{X} = (X_1, \ldots, X_n)$ and $\mathbb{Y} = (Y_1, \ldots, Y_m)$ denote independent samples of i.i.d. random fuzzy numbers. We want to verify

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

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

FDA in hypothesis testing with fuzzy data

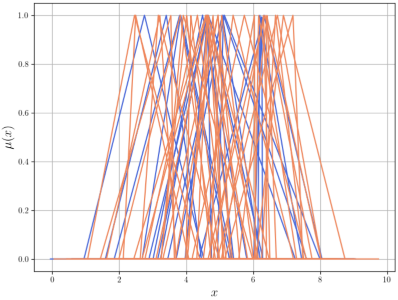
Let $\mathbb{X} = (X_1, \ldots, X_n)$ and $\mathbb{Y} = (Y_1, \ldots, Y_m)$ denote independent samples of i.i.d. random fuzzy numbers. We want to verify

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

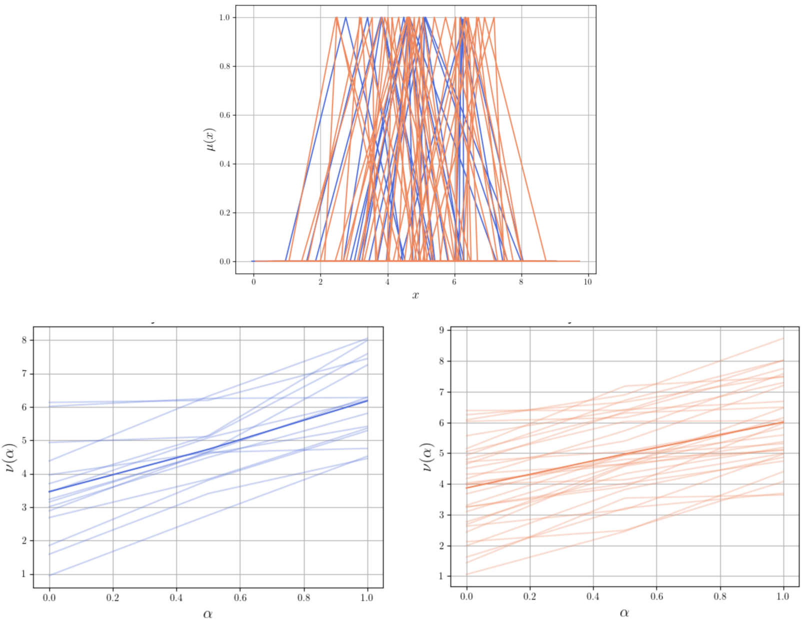
Let $\mathbb{x} = (x_1, \ldots, x_n)$ and $\mathbb{y} = (y_1, \ldots, y_m)$ denote the experimental realizations of two independent random fuzzy samples, i.e. $x_i \in \mathbb{F}_r(\mathbb{R})$ $\forall i = 1, \ldots, n$ and $y_j \in \mathbb{F}_r(\mathbb{R})$ $\forall j = 1, \ldots, m$.

Firstly, we construct a credibility distribution for each observation, i.e. $\big(\Upsilon_{x_1}(t), \ldots, \Upsilon_{x_n}(t)\big)$ and $\big(\Upsilon_{y_1}(t), \ldots, \Upsilon_{y_m}(t)\big)$, $t \in \mathbb{R}$.

Then, we determine the ICr functions for both samples, i.e. $\upsilon_x = \big(\upsilon_{x_1}(\alpha), \ldots, \upsilon_{x_n}(\alpha)\big)$ and $\upsilon_y = \big(\upsilon_{y_1}(\alpha), \ldots, \upsilon_{y_m}(\alpha)\big)$, $\alpha \in [0, 1]$, where $\upsilon_{x_i}(\alpha) := \Upsilon^{-1}_{x_i}(\alpha)$ and $\upsilon_{y_j}(\alpha) := \Upsilon^{-1}_{y_j}(\alpha)$.



1.0
0.8
0.6
0.4
0.2
0.0
$\mu(x)$
0
2
4
6
8
10
x



$\mu(x)$
x
$\nu(\alpha)$
α
$\nu(\alpha)$
α

Consider the following test statistic for the given experimental data

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

where

$$\overline{\upsilon}_x(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \upsilon_{x_i}(\alpha), \quad s_x^2(\alpha) = \frac{1}{n-1} \sum_{i=1}^{n} \Big[\upsilon_{x_i}(\alpha) - \overline{\upsilon}_x(\alpha)\Big]^2,$$

$$\overline{\upsilon}_y(\alpha) = \frac{1}{m} \sum_{j=1}^{m} \upsilon_{y_j}(\alpha), \quad s_y^2(\alpha) = \frac{1}{m-1} \sum_{j=1}^{m} \Big[\upsilon_{y_j}(\alpha) - \overline{\upsilon}_y(\alpha)\Big]^2,$$

and let

$$t_0 = \sup_{\alpha \in [0,1]} T(\alpha).$$

Now, starting from the initial dataset we will design a specific permutation procedure.

Now, starting from the initial dataset we will design a specific permutation procedure.

Let $\mathbb{w} := \upsilon_x \uplus \upsilon_y$, where $\uplus$ stands for vector concatenation pooling the two samples into one, i.e.

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

Let $\mathbb{w}^*$ denote a permutation of the initial dataset $\mathbb{w}$.

Suppose, we take first n elements of $\mathbb{w}^*$ and assign them to sample υ_x^*, while the remaining m elements create the second sample υ_y^*.

Thus, it works like a random assignment of $N = n + m$ elements into two samples of the size n and m, respectively.

Next we calculate the corresponding value of the test statistic for $\upsilon_x^* = \big(\upsilon_{x_1}^*, \dots, \upsilon_{x_n}^*\big)$ and $\upsilon_y^* = \big(\upsilon_{y_1},^* \dots, \upsilon_{y_m}^*\big)$, i.e.

$$t^* = \sup_{\alpha\in[0,1]} T(\upsilon_x^*, \upsilon_y^*; \alpha) = \sup_{\alpha\in[0,1]} \frac{|\overline{\upsilon}_x^*(\alpha) - \overline{\upsilon}_y^*(\alpha)|}{\sqrt{\frac{1}{n}s_{x^*}^2(\alpha) + \frac{1}{m}s_{y^*}^2(\alpha)}}.$$

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

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

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

Other interesting tests:

- two-sample test based on

$$T(\upsilon_x, \upsilon_y) = \int_0^1 \left(\overline{\upsilon}_x(\alpha) - \overline{\upsilon}_y(\alpha)\right)^2 d\alpha$$

- k-sample test based on

$$\widetilde{T} = \sum_{i=1}^{k} T(\upsilon_{x_i}, \upsilon_w),$$

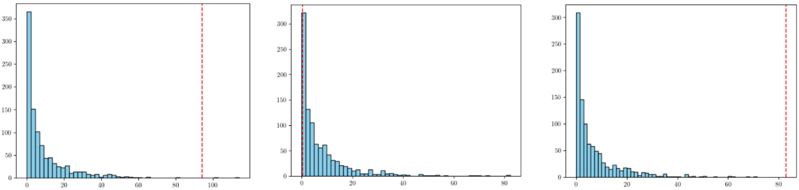
  where $\mathbb{w} := \upsilon_{x_1} \uplus \ldots \uplus \upsilon_{x_k}$

- two-sample test based on the modified band depth

Example (cont.)

We consider some data given in Ramos-Guajardo A.B. et al.(2019) to compare the opinions of the three experts about the overall impression of the Gamonedo cheese. We have three independent fuzzy samples of sizes $n_1 = 40$, $n_2 = 38$ and $n_3 = 42$, coming from the unknown distributions.

Opinion	Expert 1	Expert 2	Expert 3
1	$(65, 75, 85, 85)$	$(50, 50, 63, 75)$	$(60, 63, 67, 72)$
2	$(35, 37, 44, 50)$	$(39, 47, 52, 60)$	$(53, 58, 63, 68)$
3	$(66, 70, 75, 80)$	$(60, 70, 85, 90)$	$(43, 47, 54, 58)$
4	$(70, 74, 80, 84)$	$(50, 56, 64, 74)$	$(70, 76, 83, 86)$
5	$(65, 70, 75, 80)$	$(39, 45, 53, 57)$	$(54, 60, 65, 70)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$



Empirical null distribution of the permutation test with red vertical line indicating the test statistic value for experts 1 vs. 2; 1 vs. 3 and 2 vs. 3.

Conclusions and further research

Conclusions and further research

- Due to certain difficulties with fuzzy modeling, statistical tools for reasoning with imprecise data usually cannot be generalized straightforwardly from their classical prototypes.
- Some of those difficulties might be solved by applying another representation of fuzzy numbers, like the credibility distribution and its inverse function (ICr).
- It seems that ICr functions open wide prospects for the use of FDA methods in the analysis of fuzzy data.

11. International Conference on

Soft Methods in Probability and Statistics

September 3-6, 2024, in Salzburg (Austria)

XLIX Konferencja „Statystyka Matematyczna 2024”

Europejskie Centrum Edukacji Geologicznej w Chęcinach,
9 – 12 grudnia 2024 r.