RANDOM FUZZY SETS: A MATHEMATICAL TOOL TO DEVELOP STATISTICAL FUZZY DATA ANALYSIS

A. BLANCO-FERNÁNDEZ, M. R. CASALS, A. COLUBI, N. CORRAL, M. GARCÍA-BÁRZANA, M. A. GIL, G. GONZÁLEZ-RODRÍGUEZ, M. T. LÓPEZ, M. A. LUBIANO, M. MONTENEGRO, A. B. RAMOS-GUAJARDO, S. DE LA ROSA DE SÁA AND B. SINOVA

Abstract. Data obtained in association with many real-life random experiments from different fields cannot be perfectly/exactly quantified. Often the underlying imprecision can be suitably described in terms of fuzzy numbers/values. For these random experiments, the scale of fuzzy numbers/values enables to capture more variability and subjectivity than that of categorical data, and more accuracy and expressiveness than that of numerical/vectorial data. On the other hand, random fuzzy numbers/sets model the random mechanisms generating experimental fuzzy data, and they are soundly formalized within the probabilistic setting. This paper aims to review a significant part of the recent literature concerning the statistical data analysis with fuzzy data and being developed around the concept of random fuzzy numbers/sets.

1. Introduction

In [53] the father of Fuzzy Logic, Professor Lotfi A. Zadeh, entitled a discussion to an invited paper in the journal Technometrics by stating that "Probability Theory and Fuzzy Logic are complementary rather than competitive". This title/sentence guide, among many others in the literature (see, for instance, Viertl and Hareter [51], for a methodological approach to the statistical analysis of fuzzy data, Ramezanzadeh *et al.* [42], for a more concrete problem involving random fuzzy data and Taheri and Kelkinnama [47] for a quite recent fuzzy linear regression analysis), the study in this paper.

Invited Paper: Received in November 2011

Key words and phrases: Distances between fuzzy numbers/values, Fuzzy numbers/values, Fuzzy arithmetic, Random fuzzy numbers/sets, Statistical methodology.

This paper is dedicated to our (retired) former Head of the Department and scientific father/grandfather/great-grandfather of us, Professor Pedro Gil, who introduced us in the knowledge of Fuzzy Sets. The research in this paper has been partially supported by the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-01 and MTM2009-09440-C02-02, the COST Action IC0702 (especially through the Short Terms Scientific Missions spent by Blanco, Colubi, García-Bárzana, González-Rodríguez, Ramos, de la Rosa de Sáa and Sinova), the FPU Grant AP2009-1197 and the Ayuda de Investigación 2011 de la Fundación Banco Herrero (Sinova), the FPI Grant BES2010-032172 (García-Bárzana) and the Contract CP-PA-11-SMIRE from the Principality of Asturias-Universidad de Oviedo (de la Rosa de Sáa). Their financial support is gratefully acknowledged.

In this way, this paper combines ideas, concepts and results from both theories. However, the paper is not trying either to compare them or to replace any of them by the other. Random fuzzy sets take advantages of the skills of both theories and join them for a kind of consortium. More concretely, the two types of underlying uncertainties (namely, randomness and fuzziness) have been modeled by using their own tools, and they have been integrated into the model of random fuzzy sets (or fuzzy random variables in Puri and Ralescu's sense) with a different mission: randomness concerns the *generation* of experimental data, whereas fuzziness concerns the *nature of the information* provided by the data.

Statistics is the science of collecting, organizing, analyzing and interpreting experimental data. Available data in performing random experiments are traditionally supposed to be able to be expressed by using the scale of real numbers (or, more generally, the scale of vectorial values with real-valued components).

Nevertheless, in the real world we can find valuations/perceptions/judgements/ratings/classifications/etc. associated with random experiments and leading to data which cannot be appropriately expressed by using real/vectorial values. Usually these data are statistically treated and analyzed as categorical ones, but the methods for the statistical analysis can only exploit a limited information in most of the cases. An alternative approach to deal with many of these data is based on expressing them by using the richer and wider scale of fuzzy numbers (or, more generally, the scale of fuzzy values). The suitability of this fuzzy scale is to be commented and illustrated after recalling the concept of fuzzy number/value.

The aim of this paper is to overview the main basic concepts and ideas around random fuzzy numbers/sets and summarize some of the recent statistical developments to analyze experimental fuzzy data. In Section 2 we will motivate the use of, and we will present the key concepts and some guidelines to describe or collect, fuzzy data. In Section 3 we will recall the concept of random fuzzy number/set and related preliminaries. In Section 4 we will present a summary on the already existing statistical methodology to analyze fuzzy data on the basis of random fuzzy numbers/sets, and we will illustrate one of the methods with a real-life example. Finally, we will comment about open problems and future directions.

2. Motivating and Modeling Experimental Fuzzy Data

A basic notion in dealing with imprecise data produced by a random experiment is that of a fuzzy number (or more generally, that of fuzzy value). In this section we first recall this notion, and motivate its application with some real-life examples. Finally, we abstract the ideas in the examples to state some guidelines to express and collect experimental fuzzy data.

2.1. The Scale of Fuzzy Numbers/Values.

The concepts of fuzzy numbers and fuzzy values are formalized as follows:

Definition 2.1. A *fuzzy number* is a mapping $\widetilde{U} : \mathbb{R} \to [0,1]$ so that for each $\alpha \in (0,1]$ the *α-level set* $\widetilde{U}_\alpha = \{x \in \mathbb{R} : \widetilde{U}(x) \geq \alpha\}$ is a nonempty compact interval. A *fuzzy value* is a mapping $\widetilde{U} : \mathbb{R}^p \to [0,1]$ (with $p \in \mathbb{N}$) so that for each

$\alpha \in (0,1]$ the α-level set $\widetilde{U}_\alpha = \{x \in \mathbb{R}^p : \widetilde{U}(x) \geq \alpha\}$ is a nonempty compact convex set of $\mathbb{R}^p$.

The *space of fuzzy numbers* will be denoted by $\mathcal{F}_c(\mathbb{R})$, and the *space of fuzzy values* will be denoted by $\mathcal{F}_c(\mathbb{R}^p)$.

Definition 2.2. A fuzzy number/fuzzy value is said to be *bounded* if the 0-*level set* $\widetilde{U}_0 = \text{closure}\{x \in \mathbb{R}^p : \widetilde{U}(x) > 0\}$ is a nonempty compact set. The *space of bounded fuzzy numbers* will be denoted by $\mathcal{F}_c^*(\mathbb{R})$, and the *space of bounded fuzzy values* will be denoted by $\mathcal{F}_c^*(\mathbb{R}^p)$.

Equivalently, fuzzy numbers and values can be formalized as $[0,1]$-valued upper semicontinuous functions with convex bounded α-levels for all $\alpha \in (0,1]$, and attaining at least once the maximum value 1 (i.e., normal fuzzy sets).

Of course, real/vectorial/interval/set-valued data can be viewed as particular fuzzy numbers or values, by simply identifying them with the associated indicator functions.

To support its applicability, it is relevant to look at the interpretation of fuzzy numbers/values. Thus, a fuzzy number/value $\widetilde{U}$ models an ill-defined quantity or property on $\mathbb{R}/\mathbb{R}^p$, so that for each $x \in \mathbb{R}/\mathbb{R}^p$ the value $\widetilde{U}(x)$ can be interpreted as the 'degree of compatibility' of x with the property 'defining' $\widetilde{U}$ (or 'degree of membership' of x to $\widetilde{U}$).

In dealing with imprecise-valued data, like those coming from valuations/perceptions/judgements/ratings/classifications/etc., the scale of fuzzy numbers/values

- is much more expressive and mathematically manageable, and its flexibility allows to capture better the inherent diversity, variability (and hence subjectivity), than that of categories (or associated ranks in case of ordinal data), which are usually constrained to a reduced list of possible values or labels;
- it enables a more accurate description of the real perceptions/ratings/etc., and captures better the intrinsic imprecision than the scale of real/vectorial values;
- the usual arithmetic for fuzzy numbers and values (as shown later) pays attention to the 'location' and 'shape' of values (which are crucial for the meaning, characterization and application of fuzzy values);
- 'distances' can be defined between fuzzy numbers/values which take into account both 'location' and 'shape'.

2.2. **Motivating Real-life Examples.** As real-life examples motivating the use of fuzzy numbers we will consider the next ones:

Example 2.3. Hesketh and collaborators (see [20, 21]) have conducted the psychometric random experiment of rating people's perception on different aspects concerning an occupation.

Due to randomness, perceptions vary from respondent to respondent. Moreover, perceptions are intrinsically imprecise-valued and they could be properly expressed in terms of fuzzy numbers.

Instead of considering a single-point rating system, which may not accurately reflect the extent to which a respondent is prepared to consider a range of possibilities around a preferred/compatible point, experts have considered that the fuzzy scale captures respondent's preparedness to endorse a range of options while retaining information about the 'most preferred' point (which was referred to as the '$\vee$' pointer). Responses were stated on $[0, 100]$ (from low to high).

Based on previous psychometric studies, five anchors were developed to measure prestige and three to measure sex-type. Figure 1 reproduces the form that respondents have filled in the experiment by Hesketh *et al.* [20].

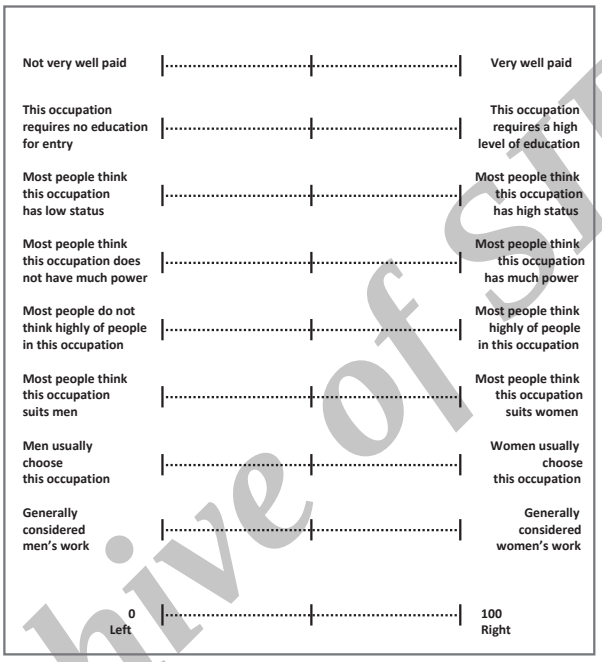

FIGURE 1. Verbal Anchors Used on the Five Prestige and Three Sex-Type Scales (Hesketh *et al.* [20])

Respondents were then asked how they thought people generally viewed each occupation in relation to the scale. The opportunity was used to ensure that respondents understood that the fuzzy rating represented their estimate of how people generally view occupations. It should be noted that the fuzzy numbers scale allows fuzzy ratings to be elicited from respondents who have no knowledge of the mathematics underpinning Fuzzy Logic.

Given these directions in Figure 2 we can see the response provided by a respondent to the question "How attractive do you find a (given) occupation?" (cf., Hesketh *et al.* [21])

This response could be interpreted that 70 is the '$\vee$' point (or preferred perception), 45 to 85 is the 'acceptable area' around the 'preferred perception' (or interval of points which are compatible to a greater or lesser extent with respondent's perception) and the compatibilities for the remaining points have being obtained by using 'linear interpolation', leading to the triangular fuzzy number $\mathrm{Tri}(45, 70, 85)$.

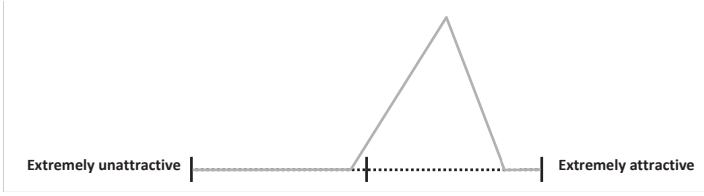

FIGURE 2. Example of a Response to the Question About How Attractive the Respondent Finds a Given Occupation (Hesketh *et al.* [21])

Example 2.4. Colubi and González-Rodríguez from the SMIRE Research Group, and other collaborators from the INDUROT (Research Institute of the University of Oviedo, Spain), have conducted the environmental random experiment of rating the quality and other aspects of trees in a reforestation which was performed more than two decades ago in Valle del Huerna (an area between the provinces of Asturias and León, in the North of Spain). For more details, see Colubi [3], González-Rodríguez *et al.* [14].

Due to randomness, ratings vary from tree to tree (and also from expert to expert). Moreover, ratings of the quality are intrinsically imprecise-valued and they could be properly expressed in terms of fuzzy numbers.

Instead of considering usual Likert's 1-5 or 1-7 codings, environmental experts at the INDUROT have been informed about the possibility of rating quality by using bounded fuzzy numbers on $[0, 100]$ (0 meaning the lowest quality, 100 meaning the highest one). To ease the graphical representation, environmental experts have been recommended to draw trapezoidal fuzzy numbers in a form like that in Figure 3, by stating for each qualified tree the 1-level (or closed interval of values which are viewed as being 'fully compatible' with their rating of the quality of the tree), the 0-level (or closed interval of values such that all those in the corresponding open interval are viewed as being 'compatible to some extent' with their rating of the quality of the tree), and finally the two closed intervals have been 'interpolated' by using linear interpolation to build the trapezoidal fuzzy rating.

Figure 4 shows the rating of the quality of a birch (*Betula celtiberica* species) provided by an expert which means that the expert considers that 35 to 40 are fully compatible with her/his rating, 27.5 to 42.5 are compatible to a greater or lesser extent, and the remaining values have a gradual (in accordance with a linear graduation) degree of compatibility with her/his rating, leading to the trapezoidal fuzzy number $\mathrm{Tra}(27.5, 35, 40, 42.5)$.

Although we can think, especially from a theoretical perspective, about fuzzy values in higher dimension, most of real world examples to which the studies in the paper would be directly applicable will be one-dimensional ones.

2.3. **Collecting Experimental Fuzzy Data: Guidelines to Express Them.** By abstracting the ideas in describing data in the two former examples, we can state some guidelines which are easy to explain and friendly to understand and handle by practitioners (cf., González-Rodríguez *et al.* [16]). Given an imprecise datum (opinion/valuation/rating/perception/etc.) which is to be described by means of a fuzzy value, the steps to follow in accordance with the suggested guidelines are:

(1) Practitioners (experts/researchers/...) first state the *general support* (or set of values which could be considered as a 'reference range', in which 0-levels of all the data will be included in) as the set of values which could be *potentially compatible with all the data*.
(2) Practitioners (experts/researchers/...) state the 0-*level* as the set of values which are considered to be *compatible with the given datum to some extent*.
(3) Practitioners (experts/researchers/...) state the 1-*level* as the set of values which are considered to be *fully compatible with the given datum*.
(4) These two *level sets are finally 'interpolated'* to get a fuzzy value.

To guarantee practitioners to be free to make their descriptions, the suggested hints are not constrained to a particular shape for the fuzzy values. Nevertheless,

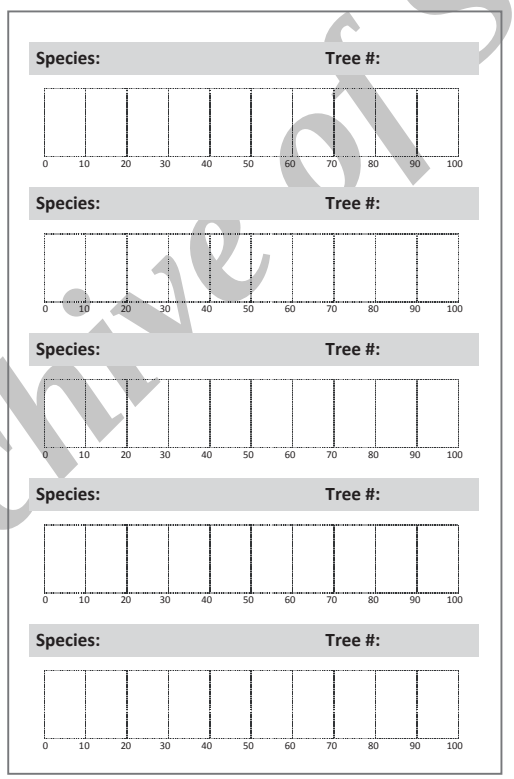

FIGURE 3. Form to be Filled by Environmental Experts to Rate the Quality of Threes From a Reforestation

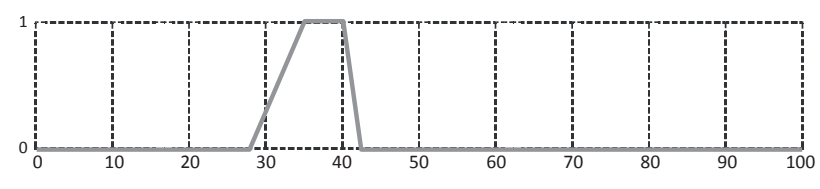

FIGURE 4. Example of a Rating of the Quality of a Birch in a Reforestation Study

the type of interpolation indicated in (4) is often chosen in case $p = 1$ from a list of manageable functions (e.g., linear, S-curves, Z-curves, and so on).

3. Statistics with Fuzzy Data: Preliminaries

In performing statistics with fuzzy data within a probabilistic setting (so that all the posterior methodology is formally sound and well-supported) there will be three key tools, namely, the arithmetic between fuzzy numbers/values, the distances between them, and the model for the random mechanism generating fuzzy data. In this section we will present these three tools.

3.1. Arithmetic with Fuzzy Data.

The elementary arithmetic operations required for the statistical fuzzy data analysis are the sum and the product by scalars. These two operations can be approached either by applying directly Zadeh's (also called the maximum-minimum) extension principle [52] or, equivalently and based on the results by Nguyen [38], as the level-wise extension of the usual set-valued arithmetic.

Given two fuzzy values $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c(\mathbb{R}^p)$, the *sum of* $\widetilde{U}$ *and* $\widetilde{V}$ is defined as the fuzzy value $\widetilde{U} + \widetilde{V} \in \mathcal{F}_c(\mathbb{R}^p)$ such that

$$(\widetilde{U} + \widetilde{V})(t) = \sup_{(y,z)\in\mathbb{R}^p\times\mathbb{R}^p \,:\, y+z=t} \min\left\{\widetilde{U}(y), \widetilde{V}(z)\right\}$$

or, equivalently, for each $\alpha \in [0, 1]$:

$$(\widetilde{U} + \widetilde{V})_\alpha = \text{Minkowski sum of } \widetilde{U}_\alpha \text{ and } \widetilde{V}_\alpha = \left\{y + z \,:\, y \in \widetilde{U}_\alpha, z \in \widetilde{V}_\alpha\right\}.$$

Given a fuzzy value $\widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p)$ and a real number γ, the *product of* $\widetilde{U}$ *by the scalar* γ is defined as the fuzzy value $\gamma \cdot \widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p)$ such that

$$(\gamma \cdot \widetilde{U})(t) = \sup_{y\in\mathbb{R}^p \,:\, y=\gamma t} \widetilde{U}(y) = \begin{cases} \widetilde{U}\left(\frac{t}{\gamma}\right) & \text{if } \gamma \neq 0 \\ \\ \mathbf{1}_{\{0\}}(t) & \text{if } \gamma = 0 \end{cases}$$

or, equivalently, for each $\alpha \in [0, 1]$:

$$(\gamma \cdot \widetilde{U})_\alpha = \gamma \cdot \widetilde{U}_\alpha = \left\{\gamma \cdot y \,:\, y \in \widetilde{U}_\alpha\right\},$$

which corresponds to consider level-wise the natural product of a set by a scalar.

Remark 3.1. The space of fuzzy values endowed with Zadeh's arithmetic, $(\mathcal{F}_c(\mathbb{R}^p), +, \cdot)$, has *not a linear* (but a semilinear-conical) *structure*. This is due to the fact that, whatever $\widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p)$ may be, then $\widetilde{U} + (-1) \cdot \widetilde{U} \neq \mathbf{1}_{\{\mathbf{0}\}}$ (where the indicator $\mathbf{1}_{\{\mathbf{0}\}}$ is the neutral element for the fuzzy sum) but in the very special case in which $\widetilde{U}$ reduces to the indicator function of a singleton.

A relevant consequence from the nonlinearity is that *there is no 'difference operation'* between fuzzy values which is simultaneously well-defined and preserving the main properties of the difference between real/vectorial values in connection with the sum. In fact, there exists a difference notion (Hukuhara's one [22]) satisfying the last condition, but it cannot be defined for many fuzzy values.

On the other hand, it should be pointed out that although fuzzy values are formalized as $[0, 1]$-valued functions, one cannot treat directly fuzzy data as functional ones, in the way which they are usually handled in Functional Data Analysis. This is due to the fact that the above-presented arithmetic does not coincide with the usual arithmetic with functions, and when we apply the functional arithmetic on $\mathcal{F}_c(\mathbb{R}^p)$ outputs are quite often out of this space and the fuzzy meaning is always lost.

3.2. **Metrics Between Fuzzy Values.** The two last concerns have been substantially overcome in developing statistics with fuzzy data by incorporating suitable distances between these data. On one hand, distances will allow to 'translate' the equality of fuzzy values (which in the case of real values is frequently expressed in terms of their difference being equal to 0) into the distance between these values being equal to 0. On the other hand, appropriate distances will allow us also to 'identify' fuzzy data with functional ones through the so-called support function (see Puri and Ralescu [39, 40]).

In the literature one can find many useful metrics between fuzzy numbers and a few ones between fuzzy values. Valuable references on this point can be found, for instance, in Klement *et al.* [23], Bertoluzza *et al.* [2], Diamond and Kloeden [7], Körner and Näther [25] and more recently, Trutschnig *et al.* [49] and Sinova *et al.* [46].

Regarding L^2 type metrics, they become quite convenient in connection with mean values (as we will see later) as well as in connection with the extension of the Least Squares approach to deal with fuzzy data. Among these metrics a generalized family of them have been introduced recently by Trutschnig *et al.* [49] taking Bertoluzza *et al.*'s and Körner and Näther's ones as inspiration.

For more details and sound arguments justifying the connection between the space of fuzzy values and Hilbert spaces which will be considered hereafter, readers can see González-Rodríguez *et al.* [16]. Let $\mathcal{H}$ = Hilbert space of the L^2-type real-valued functions defined on $\mathbb{S}^{p-1} \times (0, 1]$ with respect to λ_p and λ (with λ_p = normalized Lebesgue measure on $\mathbb{S}^{p-1}$ = unit sphere in $\mathbb{R}^p$, and λ = Lebesgue measure on $(0, 1]$). Let $\mathcal{F}_c^2(\mathbb{R}^p) = \{\widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p) : s_{\widetilde{U}} \in \mathcal{H}\}$, where $s_{\widetilde{U}}$ is the support function of $\widetilde{U}$ (see Puri and Ralescu [40]) which can be defined by

$$s_{\widetilde{U}} = \operatorname{mid} s_{\widetilde{U}} + \operatorname{spr} s_{\widetilde{U}},$$

where $\operatorname{mid} s_{\widetilde{U}}(u,\alpha)$ denotes the mid-point/center of the projection of $\widetilde{U}_\alpha$ over the direction $u \in \mathbb{S}^{p-1}$, $\operatorname{spr} s_{\widetilde{U}}(u,\alpha)$ denotes the spread/radius of the projection of $\widetilde{U}_\alpha$ over the direction u, $\mathbb{S}^{p-1}$ = unit sphere in $\mathbb{R}^p$. Of course, $\mathcal{F}_c^*(\mathbb{R}^p) \subset \mathcal{F}_c^2(\mathbb{R}^p)$, the last one being much wider than the first one. On $\mathcal{F}_c^2(\mathbb{R}^p)$ we can define

Definition 3.2. Let $\theta \in (0,+\infty)$ and let φ be an absolutely continuous probability measure on $([0,1],\mathcal{B}_{[0,1]})$ with the mass function being positive in $(0,1)$. Then, the θ,φ-*distance* is defined as the mapping $D_\theta^\varphi : \mathcal{F}_c^2(\mathbb{R}^p)\times\mathcal{F}_c^2(\mathbb{R}^p) \to [0,+\infty)$ such that it associates with $\widetilde{U},\widetilde{V} \in \mathcal{F}_c^2(\mathbb{R}^p)$ the value $D_\theta^\varphi(\widetilde{U},\widetilde{V})$ such that

$$\left(D_\theta^\varphi(\widetilde{U},\widetilde{V})\right)^2 = \int_{(0,1]}\int_{\mathbb{S}^{p-1}} \left[\operatorname{mid} s_{\widetilde{U}}(u,\alpha) - \operatorname{mid} s_{\widetilde{V}}(u,\alpha)\right]^2 d\lambda_p(u)\, d\varphi(\alpha)$$
$$+\theta \int_{(0,1]}\int_{\mathbb{S}^{p-1}} \left[\operatorname{spr} s_{\widetilde{U}}(u,\alpha) - \operatorname{spr} s_{\widetilde{V}}(u,\alpha)\right]^2 d\lambda_p(u)\, d\varphi(\alpha).$$

Remark 3.3. Due to the meaning of $\operatorname{mid} s_{\widetilde{U}}$ and $\operatorname{spr} s_{\widetilde{U}}$, for each level the choice of θ allows us to weight the effect of the deviation between spreads (which could be intuitively translated into the difference in 'shape' or 'imprecision') in contrast to the effect of the deviation between mid's (which can be intuitively translated into the difference in 'location').

On the other hand, the choice of φ enables to weight the relevance of different levels (i.e., the degree of 'imprecision', 'subjectivity', 'variability',...), and this measure has no stochastic but weighting mission.

From an interpretational perspective and for practical purposes, because of being the most frequent situation in statistical analysis of real-life fuzzy data, it is interesting to examine the particular case of fuzzy numbers, that is, $p = 1$, and to show alternative expressions of the D_θ^φ metric for some choices of θ.

In general, in case $p = 1$ the metric D_θ^φ can be expressed in terms of the squared Euclidean distances between the centers (mids) and the squared Euclidean distances between the radius (spreads) of the interval level sets of the involved fuzzy numbers. More precisely, given two fuzzy numbers $\widetilde{U},\widetilde{V} \in \mathcal{F}_c^2(\mathbb{R})$

$$D_\theta^\varphi(\widetilde{U},\widetilde{V}) = \sqrt{\int_{[0,1]} \left(\left[\operatorname{mid}\widetilde{U}_\alpha - \operatorname{mid}\widetilde{V}_\alpha\right]^2 + \theta\cdot\left[\operatorname{spr}\widetilde{U}_\alpha - \operatorname{spr}\widetilde{V}_\alpha\right]^2\right) d\varphi(\alpha)}.$$

If $\theta = 1$, D_θ^φ is equivalent to weighting only and uniformly the two squared Euclidean distances between the extreme points of the level sets (i.e., the δ_2 metric by Diamond and Kloeden [7]), so that

$$D_1^\varphi(\widetilde{U},\widetilde{V}) = \sqrt{\int_{[0,1]} \left(\frac{1}{2}\left[\inf\widetilde{U}_\alpha - \inf\widetilde{V}_\alpha\right]^2 + \frac{1}{2}\left[\sup\widetilde{U}_\alpha - \sup\widetilde{V}_\alpha\right]^2\right) d\varphi(\alpha)}.$$

If $\theta = 1/3$, D_θ^φ is equivalent to weighting uniformly all the squared Euclidean distances between the convex linear extreme points of the level sets, so that

$$D_{1/3}^\varphi(\widetilde{U},\widetilde{V}) = \sqrt{\int_{[0,1]} \left(\int_{[0,1]} \left[\widetilde{U}_\alpha^{[\nu]} - \widetilde{V}_\alpha^{[\nu]}\right]^2 d\lambda(\nu)\right) d\varphi(\alpha)}$$

with

$$\widetilde{U}_\alpha^{[\nu]} = \nu \sup \widetilde{U}_\alpha + (1-\nu) \inf \widetilde{U}_\alpha.$$

More generally, if $\theta \in (0,1]$, then (see Gil *et al.* [11], Trutschnig *et al.* [49]) there exist a weighting measure W formalized as a nondegenerate probability measure on $([0,1], \mathcal{B}_{[0,1]})$ with $\int_{[0,1]} dW(\nu) = .5$ and $\theta = \int_{[0,1]} (2\nu - 1)^2 \, dW(\nu)$, such that

$$D_\theta^\varphi(\widetilde{U}, \widetilde{V}) = \sqrt{\int_{[0,1]} \left(\int_{[0,1]} \left[\widetilde{U}_\alpha^{[\nu]} - \widetilde{V}_\alpha^{[\nu]} \right]^2 dW(\nu) \right) d\varphi(\alpha)},$$

which coincides with Bertoluzza *et al.*'s metric [2].

The θ, φ-metric holds several valuable metric properties, and it allows us to establish a Rådstrom-type isometry enabling us to identify each fuzzy number with a functional data and to connect one-to-one arithmetics and metrics.

Theorem 3.4. *Let $\theta \in (0, +\infty)$ and let φ be an absolutely continuous probability measure on $([0,1], \mathcal{B}_{[0,1]})$ with the mass function being positive in $(0,1)$. Then, D_θ^φ satisfies that*

i) D_θ^φ is an L^2-type metric on $\mathcal{F}_c^2(\mathbb{R}^p)$.

ii) $\left(\mathcal{F}_c^2(\mathbb{R}^p), D_\theta^\varphi\right)$ is a separable metric space.

iii) The support function $s : \mathcal{F}_c^2(\mathbb{R}^p) \to \mathcal{H}$ (with $s(\widetilde{U}) = s_{\widetilde{U}}$) states an isometrical embedding of $\mathcal{F}_c^2(\mathbb{R}^p)$ with the fuzzy arithmetic and D_θ^φ onto a closed convex cone of $\mathcal{H}$ with the functional arithmetic and an appropriate distance which can be found detailed in [16].

Remark 3.5. An immediate and crucial implication from *iii*) in the later theorem is that any fuzzy value $\widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}^p)$ can be identified by the corresponding function $s_{\widetilde{U}}$, and this identification is accompanied by the correspondences between the arithmetics and L^2 metrics. Consequently, data in the setting of fuzzy values with the fuzzy arithmetic and the metric D_θ^φ can be systematically translated into data in the setting of functional values with the functional arithmetic and the metric based on the associated norm. In this way, although fuzzy data should not be treated directly as functional data, they can be treated as functional data by considering the identification *via* the support function.

As a relevant implication for statistical purposes, several developments in Functional Data Analysis could be particularized to fuzzy data by using the adequate identifications and correspondences. However, it should be guaranteed that the resulting elements/outputs are not out of the cone $s\left(\mathcal{F}_c^2(\mathbb{R}^p)\right)$. Otherwise, *ad hoc* techniques should be developed, as we will show in the next section.

Regarding L^1 type metrics, we have recently considered some well-known ones and introduced a generalized one in exploring a more robust statistical analysis (cf., Sinova *et al.* [45, 46]).

3.3. **Random Fuzzy Numbers/Sets.** Random fuzzy numbers (or, more generally, random fuzzy sets) is a well-stated and supported model within the probabilistic setting for the random mechanisms generating fuzzy data. They integrate

randomness and fuzziness, so that the first one affects the generation of experimental data, whereas the second one affects the nature of experimental data which are assumed to be intrinsically imprecise.

Because of being a soundly established model within the probabilistic setting, especially in which concerns its formalization as random elements (i.e., as Borel-measurable mappings), most of the ideas, concepts, and tools in the statistical analysis of real-valued data make sense in handling experimental fuzzy data when they are assumed to be generated by a random fuzzy set. It has been our main policy, which will be shortly summarized in Section 4, to preserve as many as possible of the ideas, concepts and tools from traditional statistics, and to incorporate ideas, concepts and tools from Fuzzy Logic to model data and handle them mathematically without affecting the stochastic terms.

Random fuzzy sets have been often referred to in the literature as *fuzzy random variables in Puri and Ralescu's sense*, the notion being introduced in [41]. As mappings defined from the sample space of a probability space modeling a random experiment to the space of real numbers/values, random fuzzy sets associate a fuzzy value with each experimental outcome and fit many real-life classification/qualification processes associated with valuations/ratings/opinions/judgements leading to data which can be properly described by means of fuzzy values.

The notion of random fuzzy set can be formalized in several equivalent ways. Thus,

Definition 3.6. Given a probability space $(\Omega, \mathcal{A}, P)$, a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$ is said to be a *random fuzzy set* (in particular, in case $p = 1$ we can name it a *random fuzzy number*), for short RFS (RFN in case $p = 1$), if (see Puri and Ralescu [41]) for all $\alpha \in (0, 1]$ the α-level set-valued mapping

$$\mathcal{X}_\alpha : \Omega \to \mathcal{K}_c(\mathbb{R}^p) = \{\textit{nonempty compact convex sets of } \mathbb{R}^p\},\ \omega \mapsto (\mathcal{X}(\omega))_\alpha\,,$$

is a compact convex random set (that is, $\mathcal{X}_\alpha$ is a Borel measurable mapping w.r.t. $\mathcal{A}$ and the Borel σ-field generated by the topology induced by Hausdorff metric on $\mathcal{K}_c(\mathbb{R}^p)$).

Alternatively, it can be formalized in different equivalent ways like, for instance (see Colubi *et al.* [4], González-Rodríguez *et al.* [16]),

Theorem 3.7. *Given a probability space $(\Omega, \mathcal{A}, P)$ and a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$, $\mathcal{X}$ is an RFS if and only if*

i) $\mathcal{X}$ is a Borel measurable mapping w.r.t. $\mathcal{A}$ and the Borel σ-field generated by the topology induced by the metric D_θ^φ on $\mathcal{F}_c^2(\mathbb{R}^p)$.

ii) $s_\mathcal{X} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random element, that is, a Borel measurable mapping w.r.t. $\mathcal{A}$ and the Borel σ-field generated by the topology induced by the metric associated through the isometrical embedding in Theorem 3.4.

On the basis of these equivalences a relevant implication can be drawn: due to the Borel-measurability of RFSs, one can properly refer in this setting to notions like the *distribution induced by an RFS*, the *stochastic independence of RFSs*, and so on, which will be crucial terms in formalizing the statistical analysis.

Remark 3.8. It should be pointed out that Kwakernaak [27, 28] and Kruse and Meyer [26] defined a concept of fuzzy random variable in case $p = 1$, which is in fact equivalent from a mathematical perspective to Puri and Ralescu's one. However, the concept was introduced to model an essentially different mechanism, in which data are supposed to be generated from a real-valued random variable associated with a random experiment, but these values not being crisply but fuzzily perceivable.

This essential difference is not just a matter of motivation, but the main issue is that it strongly affects the aim of the statistics to be developed around. In this way, although distributions and parameters could be defined in some senses in connection with the fuzzy random variable through Zadeh's extension principle, the objective of statistical developments refer usually to the distribution and parameters of the underlying original (and imperfectly perceived) real-valued random variable. The objective of statistical developments in the next section will only refer to the distribution and parameters of the random fuzzy set, since either there is no underlying real-valued random variable behind the process -as happens when we deal with judgements, valuations, ratings, and so on- or the interest is just to be focussed on the fuzzy perception.

In analyzing fuzzy data two main types of summary measures/parameters may be distinguished:

- fuzzy-valued summary measures, like the mean value of an RFS or the median of an RFN as measures for the central tendency of their distributions;
- real-valued summary measures, like the variance of an RFS as a measure for the mean error/dispersion of the distributions of the RFS, or the covariance as a measure of the (absolute) 'linear' dependence/association of RFSs.

The most commonly used definition for the mean value of an RFS is the Aumann-type one introduced by Puri and Ralescu [41], which is formalized as follows:

Definition 3.9. Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated RFS $\mathcal{X}$ such that $s_{\mathcal{X}} \in L^1(\Omega, \mathcal{A}, P)$, the (Aumann-type) *mean value* of $\mathcal{X}$ is the fuzzy value $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c^2(\mathbb{R}^p)$ such that for all $\alpha \in (0, 1]$

$$\left(\widetilde{E}(\mathcal{X})\right)_\alpha = \text{Aumann integral of } \mathcal{X}_\alpha$$
$$= \left\{ \int_{\mathbb{R}^p} X(\omega)\, dP(\omega) \text{ for all } X : \Omega \to \mathbb{R}^p, X \in L^1(\Omega, \mathcal{A}, P), X \in \mathcal{X}_\alpha \text{ a.s. } [P] \right\}.$$

Equivalently, the mean value of $\mathcal{X}$ is the fuzzy value $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c^2(\mathbb{R}^p)$ such that

$$s_{\widetilde{E}(\mathcal{X})} = E(s_{\mathcal{X}}).$$

If $\mathcal{X}$ is an RFN such that $s_{\mathcal{X}} \in L^1(\Omega, \mathcal{A}, P)$, we have that for each $\alpha \in (0, 1]$

$$\left(\widetilde{E}(\mathcal{X})\right)_\alpha = \left[E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)\right].$$

Due to the properties of the support function and the Hilbertian random elements, the mean value of an RFS satisfies the usual properties of linearity and it is the Fréchet's expectation w.r.t. D_θ^φ, which corroborates the fact that it is a central tendency measure. In this way,

Proposition 3.10. *If $\mathcal{X}$ is an RFS associated with the probability space $(\Omega, \mathcal{A}, P)$, and the distribution of $\mathcal{X}$ is degenerate at $\widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}^p)$ (i.e., $\mathcal{X} \overset{\text{a.s. } [P]}{=} \widetilde{U}$), then*

$$\widetilde{E}(\mathcal{X}) = \widetilde{U}.$$

Proposition 3.11. *$\widetilde{E}$ is affine equivariant (i.e., equivariant under 'linear' transformations), that is, if $\gamma \in \mathbb{R}$, $\widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}^p)$ and $\mathcal{X}$ is an RFS associated with the probability space $(\Omega, \mathcal{A}, P)$ and such that $s_{\mathcal{X}} \in L^1(\Omega, \mathcal{A}, P)$, then*

$$\widetilde{E}(\gamma \cdot \mathcal{X} + \widetilde{U}) = \gamma \cdot \widetilde{E}(\mathcal{X}) + \widetilde{U}.$$

Proposition 3.12. *$\widetilde{E}$ is additive (i.e., equivariant under the sum of RFSs), that is, for RFSs $\mathcal{X}$ and $\mathcal{Y}$ associated with the same probability space $(\Omega, \mathcal{A}, P)$ and such that $s_{\mathcal{X}}, s_{\mathcal{Y}} \in L^1(\Omega, \mathcal{A}, P)$, we have that*

$$\widetilde{E}(\mathcal{X} + \mathcal{Y}) = \widetilde{E}(\mathcal{X}) + \widetilde{E}(\mathcal{Y}).$$

Proposition 3.13. *$\widetilde{E}$ is coherent with the usual fuzzy arithmetic, in the sense that if $\mathcal{X}$ is an RFS associated with the same probability space $(\Omega, \mathcal{A}, P)$ and such that the set of the RFS values is finite or countable, that is, $\mathcal{X}(\Omega) = \{\widetilde{x}_1, \ldots, \widetilde{x}_m, \ldots\} \subset \mathcal{F}_c^2(\mathbb{R}^p)$, then*

$$\widetilde{E}(\mathcal{X}) = P\left(\{\omega \in \Omega : \mathcal{X}(\omega) = \widetilde{x}_1\}\right) \cdot \widetilde{x}_1 + \ldots + P\left(\{\omega \in \Omega : \mathcal{X}(\omega) = \widetilde{x}_m\}\right) \cdot \widetilde{x}_m + \ldots$$

Proposition 3.14. *$\widetilde{E}$ is the 'Fréchet expectation' of $\mathcal{X}$ w.r.t. D_θ^φ, that is,*

$$\widetilde{E}(\mathcal{X}) = \arg \min_{\widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}^p)} E\left(\left[D_\theta^\varphi(\mathcal{X}, \widetilde{U})\right]^2\right),$$

so that the mean is the fuzzy value leading to the lowest mean squared D_θ^φ-distance (or error) w.r.t. the RFS distribution, and this corroborates the fact that it is a central tendency measure.

Although there are other definitions for the mean value of an RFS, the Aumann-type one is coherent with the considered arithmetic (as outlined in Proposition 3.13), and it is supported by *Strong Laws of Large Numbers* for RFSs (cf., Colubi *et al.* [6], Terán [48], and others). The mean value is the almost sure limit (w.r.t. different metrics) of the 'sample fuzzy mean', and the result for the metric D_θ^φ could be also derived from that for Hilbert space-valued random elements. Thus,

Proposition 3.15. *Let $(\Omega, \mathcal{A}, P)$ be a probability space, $\mathcal{X} : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$ an associated RFS such that $s_{\mathcal{X}} \in L^1(\Omega, \mathcal{A}, P)$ and $\{\mathcal{X}_n\}_n$ a sequence of pairwise independent RFSs being identically distributed as $\mathcal{X}$ (i.e., $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ being a simple random sample from $\mathcal{X}$ for each $n \in \mathbb{N}$). If $\overline{\mathcal{X}_n}$ denotes the 'sample fuzzy mean', that is, $\overline{\mathcal{X}_n} = \frac{1}{n} \cdot (\mathcal{X}_1 + \ldots + \mathcal{X}_n)$, then, $\lim_{n \to \infty} D_\theta^\varphi\left(\overline{\mathcal{X}_n}, \widetilde{E}(\mathcal{X})\right) = 0 \quad a.s.\,[P]$.*

Conversely, if $\{\mathcal{X}_n\}_n$, with $\mathcal{X}_n : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$, is a sequence of pairwise independent RFSs which are identically distributed as an RFS $\mathcal{X}$, and there exists $\widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}^p)$ so that $\lim_{n \to \infty} D_\theta^\varphi\left(\overline{\mathcal{X}_n}, \widetilde{U}\right) = 0$ a.s. $[P]$, then, $s_{\mathcal{X}} \in L^1(\Omega, \mathcal{A}, P)$ and $\widetilde{U} = \widetilde{E}(\mathcal{X})$.

As for the real-valued case, the mean value of an RFN is very sensitive to changes in values and to extreme values (see Sinova *et al.* [45] for some empirical analysis about). In this respect, when a more robust measure for the central tendency is wanted one can consider (among other approaches) the extension of the median of a real-valued random variable. A key problem in carrying such an extension is derived from the fact that there is no universally acceptable total ordering between fuzzy numbers. Consequently, the extension cannot be based on the approach to the median as a middle position value, but the notion can be extended as a value minimizing a mean L^1 distance to the distribution of the RFN. By following this approach, Sinova *et al.* [45] have introduced the following notion for the median of an RFN

Definition 3.16. Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated RFN $\mathcal{X}$ with values in $\mathcal{F}_c^*(\mathbb{R})$, the (1-norm inf/sup-type) *median* of $\mathcal{X}$ is the fuzzy value $\widetilde{\mathrm{Me}}(\mathcal{X}) \in \mathcal{F}_c^*(\mathbb{R}) =$ such that for all $\alpha \in (0,1]$

$$\left(\widetilde{\mathrm{Me}}(\mathcal{X})\right)_\alpha = \left[\mathrm{Me}(\inf \mathcal{X}_\alpha), \mathrm{Me}(\sup \mathcal{X}_\alpha)\right],$$

where in case either $\mathrm{Me}(\inf \mathcal{X}_\alpha)$ or $\mathrm{Me}(\sup \mathcal{X}_\alpha)$ are not unique the usual criterion of selecting the mid-point of the interval of medians is applied.

Due to the properties of the median of a real-valued random variable, the 1-norm inf/sup-type median of an RFN preserves the main properties for the median of random variables and minimizes the mean L^1 distance associated with the 1-norm inf/sup-type, that is, it is a central tendency measure. In this way (see Sinova *et al.* [45]),

Proposition 3.17. *If $\mathcal{X}$ is an RFN associated with the probability space $(\Omega, \mathcal{A}, P)$, and the distribution of $\mathcal{X}$ is degenerate at $\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R})$, then*

$$\widetilde{\mathrm{Me}}(\mathcal{X}) = \widetilde{U}.$$

Proposition 3.18. *$\widetilde{\mathrm{Me}}$ is affine equivariant, that is, if $\gamma \in \mathbb{R}$, $\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R})$ and $\mathcal{X}$ is an RFN associated with the probability space $(\Omega, \mathcal{A}, P)$, then*

$$\widetilde{\mathrm{Me}}(\gamma \cdot \mathcal{X} + \widetilde{U}) = \gamma \cdot \widetilde{\mathrm{Me}}(\mathcal{X}) + \widetilde{U}.$$

Proposition 3.19. *$\widetilde{\mathrm{Me}}(\mathcal{X})$ minimizes the mean ρ_1-distance w.r.t. the distribution of $\mathcal{X}$, that is,*

$$E\left(\rho_1(\mathcal{X}, \widetilde{\mathrm{Me}}(\mathcal{X}))\right) = \min_{\widetilde{U} \in \mathcal{F}_c^*(\mathbb{R})} E\left(\rho_1(\mathcal{X}, \widetilde{U})\right),$$

where

$$\rho_1(\widetilde{U}, \widetilde{V}) = \frac{1}{2} \int_{(0,1]} \left(\left| \inf \widetilde{U}_\alpha - \inf \widetilde{V}_\alpha \right| + \left| \sup \widetilde{U}_\alpha - \sup \widetilde{V}_\alpha \right| \right) d\alpha.$$

The last proposition indicates that the median is a fuzzy value leading to the lowest mean ρ_1-distance (or error) w.r.t. the RFN distribution, and this corroborates the fact that it is a central tendency measure. In fact, Sinova *et al.* [45] have introduced the median of an RFN as any fuzzy number minimizing the mean

ρ_1 distance w.r.t. the distribution of the RFN, the value in Definition 3.16 being a convenient easy-to-use solution for the minimization problem and showing good properties supporting its statistical robustness.

On the other hand, the median is the almost sure limit w.r.t. ρ_1 of the 'sample fuzzy median', as stated in [45], in accordance with which

Proposition 3.20. *Let $\mathcal{X}$ be an RFN associated with a probability space $(\Omega, \mathcal{A}, P)$ and satisfying for each α that $\mathrm{Me}\big(\inf \mathcal{X}_\alpha\big)$ and $\mathrm{Me}\big(\sup \mathcal{X}_\alpha\big)$ are actually unique. Let $\{\mathcal{X}_n\}_n$ be a sequence of pairwise independent RFNs being identically distributed as $\mathcal{X}$. If $\widehat{\widetilde{\mathrm{Me}}(\mathcal{X})}_n$ denotes the 'sample fuzzy median' of the simple random sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ from $\mathcal{X}$, then we have that*

$$\lim_{n\to\infty} \rho_1\Big(\widehat{\widetilde{\mathrm{Me}}(\mathcal{X})}_n, \widetilde{\mathrm{Me}}(\mathcal{X})\Big) = 0 \quad a.s.\,[P].$$

In formalizing the variance of an RFS the Fréchet's approach will be considered (see Lubiano *et al.* [33], Körner and Näther [25]). In this approach the variance is conceived as a measure of the 'error' in approximating/estimating the values of the RFS through the corresponding mean value, this error being quantified in terms of a squared metric. A real-valued quantification of the 'dispersion' of an RFS w.r.t. its fuzzy mean will enable to compare random elements, populations, samples, estimators, etc. by simply ranking real numbers. By considering the θ, φ-metric

Definition 3.21. Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated RFS $\mathcal{X}$ such that $s_{\mathcal{X}} \in L^2(\Omega, \mathcal{A}, P)$, the (θ, φ)-*Fréchet variance* of $\mathcal{X}$ is defined to be the real number

$$\sigma^2_{\mathcal{X}} = E\left(\left[D^{\varphi}_{\theta}\Big(\mathcal{X}, \widetilde{E}(\mathcal{X})\Big)\right]^2\right)$$

or, equivalently,

$$\sigma^2_{\mathcal{X}} = \mathrm{Var}(s_{\mathcal{X}}) = \mathrm{Var}(\mathrm{mid}\, s_{\mathcal{X}}) + \theta\, \mathrm{Var}(\mathrm{spr}\, s_{\mathcal{Y}}).$$

Due to the properties of the support function and the Hilbertian random elements, the (θ, φ)-Fréchet variance of an RFS satisfies the usual properties for this concept. In this way,

Proposition 3.22. *$\sigma^2_{\mathcal{X}} \geq 0$ with $\sigma^2_{\mathcal{X}} = 0$ if, and only if, there exists $\widetilde{U} \in \mathcal{F}^2_c(\mathbb{R}^p)$ such that $\mathcal{X} = \widetilde{U}$* a.s. $[P]$.

Proposition 3.23. *If $\gamma \in \mathbb{R}$, $\widetilde{U} \in \mathcal{F}^2_c(\mathbb{R}^p)$ and $\mathcal{X}$ is an RFS associated with the probability space $(\Omega, \mathcal{A}, P)$ and such that $s_{\mathcal{X}} \in L^2(\Omega, \mathcal{A}, P)$, then*

$$\sigma^2_{\gamma\cdot\mathcal{X}+\widetilde{U}} = \gamma^2 \cdot \sigma^2_{\mathcal{X}}.$$

Proposition 3.24. *For RFSs $\mathcal{X}$ and $\mathcal{Y}$ associated with the same probability space $(\Omega, \mathcal{A}, P)$ and such that $s_{\mathcal{X}}, s_{\mathcal{Y}} \in L^2(\Omega, \mathcal{A}, P)$ and being independent, we have that*

$$\sigma^2_{\mathcal{X}+\mathcal{Y}} = \sigma^2_{\mathcal{X}} + \sigma^2_{\mathcal{Y}}.$$

The covariance for two RFSs could be defined by using the ideas in Körner and Näther [25]. Thus, since the space $\mathcal{H}$ has linear structure, then the covariance can be defined on it, although external operations are required leading finally to the following equivalent notion:

Definition 3.25. If $\mathcal{X}$, $\mathcal{Y}$ are RFSs such that $s_{\mathcal{X}}$, $s_{\mathcal{Y}} \in L^2(\Omega, \mathcal{A}, P)$, the (θ, φ)-*covariance between* $\mathcal{X}$ and $\mathcal{Y}$ is defined as the real number

$$\sigma_{\mathcal{X},\mathcal{Y}} = \mathrm{Cov}(s_{\mathcal{X}}, s_{\mathcal{Y}}) = \mathrm{Cov}(\mathrm{mid}\, s_{\mathcal{X}}, \mathrm{mid}\, s_{\mathcal{Y}}) + \theta\, \mathrm{Cov}(\mathrm{spr}\, s_{\mathcal{X}}, \mathrm{spr}\, s_{\mathcal{Y}}).$$

The covariance of two RFSs satisfies the following useful properties:

Proposition 3.26. *$\sigma_{\mathcal{X},\mathcal{X}} = \sigma^2_{\mathcal{X}}$ and $\sigma_{\mathcal{X},\mathcal{Y}} = \sigma_{\mathcal{Y},\mathcal{X}}$.*

Proposition 3.27. *For RFSs $\mathcal{X}$ and $\mathcal{Y}$ associated with the same probability space $(\Omega, \mathcal{A}, P)$ and such that $s_{\mathcal{X}}$, $s_{\mathcal{Y}} \in L^2(\Omega, \mathcal{A}, P)$ and being independent, we have that*

$$\sigma_{\mathcal{X},\mathcal{Y}} = 0.$$

Proposition 3.28. *If $\mathcal{X}$ and $\mathcal{Y}$ are RFSs associated with the probability space $(\Omega, \mathcal{A}, P)$ and such that $s_{\mathcal{X}}$,$s_{\mathcal{Y}} \in L^2(\Omega, \mathcal{A}, P)$, then*

$$\sigma^2_{\mathcal{X}+\mathcal{Y}} = \sigma^2_{\mathcal{X}} + \sigma^2_{\mathcal{Y}} + 2\,\sigma_{\mathcal{X},\mathcal{Y}}.$$

Proposition 3.29. *If $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$ are RFSs associated with the probability space $(\Omega, \mathcal{A}, P)$ and such that $s_{\mathcal{X}}$,$s_{\mathcal{Y}}$, $s_{\mathcal{Z}} \in L^2(\Omega, \mathcal{A}, P)$, then*

$$\sigma_{\mathcal{X}+\mathcal{Y},\mathcal{Z}} = \sigma_{\mathcal{X},\mathcal{Z}} + \sigma_{\mathcal{Y},\mathcal{Z}}.$$

Remark 3.30. Although the covariance preserves many of the properties for real-valued random variables, it should be pointed out that properties like those related to linear transformations of the RFSs are not satisfied in general. In this way, $\sigma_{a\cdot\mathcal{X},\mathcal{Y}} \neq a\,\sigma_{\mathcal{X},\mathcal{Y}}$.

4. Statistics with Fuzzy Data: an RFSs-based Methodology

In developing Statistics with fuzzy data there are several key distinctive features which should be pointed out, namely,

- the lack of a 'difference' between fuzzy values which is simultaneously well-defined and preserves the main properties of the difference between real numbers;
- the lack of a universally accepted total ordering between fuzzy data.

Furthermore, when Statistics are based on the concept of RFS, some additional problems arise, like

- the lack of realistic general 'parametric' families of probability distribution models for RFSs;
- the lack of Central Limit Theorems for RFSs which are directly applicable for inferential purposes (actually, there exist some CLTs for RFSs according to which the normalized distance sample-population fuzzy mean converges in law to the norm of a Gaussian random element but with values often out of the cone).

To overcome these drawbacks a crucial role will be played by the use of appropriate metrics between fuzzy data, like the θ, φ-distance. Another crucial role for the inferential approach to Statistics will be played by the existence of CLTs for Hilbert space-valued random elements, and mainly the bootstrapped CLTs ones (see, for instance, Giné and Zinn [13]).

In this section we will present a brief overview of some of the inferential statistical developments with fuzzy data based on RFSs. The aim of these developments will be to draw conclusions about the distribution of the involved RFSs over populations on the basis of the information supplied by samples of (fuzzy) observations (often referred to as realizations of random samples) from these RFSs.

4.1. **Estimating Parameters/Measures of the Distribution of RFSs.** One of the relevant inferential problems is to estimate the parameters or measures associated with the distribution of an RFS on the basis of the information provided by a sample of independent data from it, that is, a realization from a simple random sample from the RFS.

As for the real/vectorial-valued case, this can be done by considering either 'point' or 'region' estimation. In this respect, one should take into account that parameters or measures associated with the distribution of an RFS are usually fuzzy- or real-valued, so that the terms 'point' and 'region' estimation should be understood as fuzzy and fuzzy region estimation.

The parameters receiving the most attention has been the Aumann-type fuzzy mean and the θ, φ-Fréchet variance for which concerning the *'point' estimation* it has been proved (sometimes for the case $p = 1$ although it can be easily extended, cf., Lubiano *et al.* [30, 31, 32], García *et al.* [9]) that

Proposition 4.1. *Let $\mathcal{X}$ be an RFS associated with a probability space $(\Omega, \mathcal{A}, P)$. Consider a simple random sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ of size n (i.e., $\mathcal{X}_1, \ldots, \mathcal{X}_n$ i.i.d. RFSs) from $\mathcal{X}$. Then,*

i) if $s_{\mathcal{X}} \in L^1(\Omega, \mathcal{A}, P)$, the sample fuzzy mean satisfies that
- *it is an 'unbiased' fuzzy-valued estimator of $\widetilde{E}(\mathcal{X})$ (i.e., the fuzzy mean of $\overline{\mathcal{X}_n}[\cdot]$ over the space of all possible samples of n independent observations from $\mathcal{X}$ equals $\widetilde{E}(\mathcal{X})$);*
- *for most of the metrics D on $\mathcal{F}_c^2(\mathbb{R}^p)$, $\overline{\mathcal{X}_n}$ is a 'strongly consistent' fuzzy estimator of $\widetilde{E}(\mathcal{X})$ (that is, the sequence of random variables $\left\{D\left(\overline{\mathcal{X}_n}, \widetilde{E}(\mathcal{X})\right)\right\}_n$ converges a.s. to 0);*

ii) if $s_{\mathcal{X}} \in L^2(\Omega, \mathcal{A}, P)$, the sample θ, φ-Fréchet variance satisfies that
- *it is a 'biased' (although asymptotically unbiased) real-valued estimator of $\sigma^2_{\mathcal{X}}$, since for sample size n the mean of the θ, φ-Fréchet variance over the space of all possible samples of n independent observations from $\mathcal{X}$ equals $(n-1)\,\sigma^2_{\mathcal{X}}/n$;*
- *it is a 'strongly consistent' estimator of $\sigma^2_{\mathcal{X}}$ (i.e., the sequence of the sample θ, φ-Fréchet variances converges a.s. to $\sigma^2_{\mathcal{X}}$).*

The preceding results have been examined for both finite and general populations. Similar results have been discussed for other parameters/measures of RFNs,

like the 1-norm inf/sup median (see Sinova *et al.* [45]) or the fuzzy- and real-valued inequality indices or measures of the relative dispersion (see, for instance, Alonso *et al.* [1], López-García *et al.* [29], Gil *et al.* [10]).

In connection with the *'region' estimation*, when the parameters/measures considered have been real-valued ones, limit theorems based on the Large Sample Theory have been stated and approximate confidence intervals have been constructed. In case of fuzzy-valued ones, the way to proceed is not an immediate one since the notion of 'confidence region' cannot be directly applied.

An approach to this respect has been recently suggested by González-Rodríguez *et al.* [19] to estimate the Aumann-type fuzzy mean of an RFN, although the idea behind could be applied to other parameters/measures. Thus, for a given confidence coefficient $\tau \in (0,1)$ the *confidence ball of* $\widetilde{E}(\mathcal{X})$ *with respect to* D_θ^φ is defined to be given by

$$CR_\tau = \left\{ \widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}) \,:\, D_\theta^\varphi(\overline{\mathcal{X}_n}, \widetilde{U}) \leq \delta_\tau \right\},$$

where δ_τ should satisfy the following coverage condition:

$$P\left(D_\theta^\varphi(\overline{\mathcal{X}_n}, \widetilde{E}(\mathcal{X})) \leq \delta_\tau \right) \geq \tau.$$

Due to the lack of realistic general parametric models for RFNs, it is not possible to find in general a δ_τ fulfilling the coverage condition. Nevertheless we may choose δ_τ as the τ-quantile of the distribution of $D_\theta^\varphi(\overline{\mathcal{X}_n}, \widetilde{E}(\mathcal{X}))$, which could be approximated by the corresponding bootstrap τ-quantile. In fact, given a simple random sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ from $\mathcal{X}$, González-Rodríguez *et al.* [19] have suggested to proceed as follows:

Algorithm 4.2.

Step 1: Fix the confidence coefficient $\tau \in (0,1)$, and the number B of bootstrap replications.

Step 2: Obtain B bootstrap samples $(\mathcal{X}^*_{[b]1}, \ldots, \mathcal{X}^*_{[b]n})$ with $b \in \{1, \ldots, B\}$ from the simple random sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$, and for each b compute its corresponding sample mean, $\overline{\mathcal{X}^*_{[b]n}}$.

Step 3: Compute for each $b \in \{1, \ldots, B\}$ the distance between the sample mean and each bootstrap sample mean, $d_b^* = D_\theta^\varphi\left(\overline{\mathcal{X}_n}, \overline{\mathcal{X}^*_{[b]n}}\right)$.

Step 4: Choose δ_τ as one of the τ-quantiles of the sample $(d_1^*, \ldots, d_B^*)$ (that is, choose δ_τ so that at least $100\,\tau$ % of the computed distances are smaller or equal than δ_τ and at least $100(1-\tau)$ % of the computed distances are greater or equal than δ_τ).

In most of the developed studies above, simulations have been performed to show the empirical accuracy and suitability of the introduced methods.

4.2. **Testing Hypotheses About Parameters/Measures of the Distribution of RFSs.** The problem of testing hypotheses on the distribution of a random element is for sure the most discussed one in Inferential Statistics. This problem

often involves estimation either as a first step or as the basis to develop testing procedures.

The aim of the problem when it is referred to RFSs is to conclude whether a given hypothesis (called null hypothesis) about the distribution(s) of the RFS(s) could be accepted or should be rejected on the basis of the information provided by a sample of independent data from it, that is, a realization from a simple random sample from the RFS(s).

As for the real/vectorial-valued case, hypotheses could either concern parameters/measures of the distribution of the RFS(s) or concern the distribution itself. So far, the available literature on testing hypothesis about the distribution of RFS(s) consider null hypotheses related to parameters/measures of its distribution.

In case the parameter is fuzzy-valued, 'two-sided' hypotheses (understood and formalized as equalities of two or more fuzzy values) have been considered. The one-sided hypotheses make not a general sense in terms of inequalities between fuzzy data, due to the lack of a universally accepted ranking between them. Some research is being currently developed in the direction of comparing fuzzy values beyond the equality, although this research has not been yet published.

The problem of *testing null two-sided hypotheses about the population mean(s) of RFS(s)* has received a deep attention along the last years. In this subsection we will summarize some of the already published developments, and we will detail one of them for a more clear explanation on the methodology and used tools.

The *one-sample case of testing about the mean of an RFS* takes $H_0^{[1s]} : \widetilde{E}(\mathcal{X}) = \widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}^p)$ as the null hypothesis, where $\mathcal{X} : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$ is supposed to be an RFS for which the 'population' fuzzy mean $\widetilde{E}(\mathcal{X})$ exists, and $\widetilde{U}$ is a pre-specified fuzzy value. To test $H_0^{[1s]}$ a realization of a simple random sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ (i.e., a sample of independent fuzzy observations/data) is considered.

To test $H_0^{[1s]}$, this equality of fuzzy values has been first translated into the equivalent equality of real numbers $D_\theta^\varphi\big(\widetilde{E}(\mathcal{X}), \widetilde{U}\big) = 0$. The methods which have been carried out are the following:

- Test T1: Test for 'normal' RFSs (see Montenegro *et al.* [35]), where the normality is understood in Puri and Ralescu's sense [40] (i.e, $\mathcal{X} = \widetilde{V} + \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\widetilde{V} \in \mathcal{F}_c^*(\mathbb{R}^p)$). Although it is an exact and easy-to-apply method, the assumption for the RFS to be normal in Puri and Ralescu's sense is quite restrictive and unrealistic.
- Test T2: Asymptotic test for general RFSs (see Körner [24], Montenegro *et al.* [35]) based on the CLTs for generalized space-valued random elements. Although the method is not that restrictive, the asymptotic distributions of the statistics usually involve unknown parameters, and large sample sizes would be required.
- Test T3: Bootstrap test for general RFSs (see Montenegro *et al.* [35] for the simple RFSs and González-Rodríguez *et al.* [18] for general ones). Simulation studies have shown that estimating eigenvalues/covariance function in asymptotic methods in Test 2 entails a substantial loss of precision

w.r.t. the nominal significance level of the test. The use of D_θ^φ and the Generalized Bootstrapped CLT by Giné and Zinn [13] enables to consider bootstrap techniques. Simulations also have shown empirically that for small/medium samples the bootstrap method usually outperforms the asymptotic one, whereas for large sample sizes the improvement is not so remarkable, but the bootstrap approach still provides the best approximation to the nominal significance level.

Furthermore, the probability of rejecting the null hypothesis under alternative assumptions converges to 1 as $n \to \infty$ (i.e., both TEST 2 and TEST 3 are consistent).

In fact, given a realization $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of a simple random sample $(\mathcal{X}_1, \ldots, \mathcal{X}_n)$ from $\mathcal{X}$, González-Rodríguez *et al.* [18] have recommended to proceed as follows:

Algorithm 4.3.

Step 1: Compute for the available sample of fuzzy data the value of the statistic

$$T_n(\widetilde{x}_1, \ldots, \widetilde{x}_n) = \left[D_\theta^\varphi \left(\frac{1}{n} \cdot [\widetilde{x}_1 + \ldots + \widetilde{x}_n], \widetilde{U} \right) \right]^2 \Big/ \widehat{S_n^2}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$$

where

$$\widehat{S_n^2}(\widetilde{x}_1, \ldots, \widetilde{x}_n) = \frac{1}{n-1} \sum_{i=1}^{n} \left[D_\theta^\varphi \left(\widetilde{x}_i, \frac{1}{n} \cdot [\widetilde{x}_1 + \ldots + \widetilde{x}_n] \right) \right]^2 .$$

Step 2: Fix the bootstrap population to be $\{\widetilde{x}_1, \ldots, \widetilde{x}_n\}$, and obtain a realization $(\widetilde{x}_1^*, \ldots, \widetilde{x}_n^*)$ of the simple random sample $(\mathcal{X}_1^*, \ldots, \mathcal{X}_n^*)$ from the bootstrap population.

Step 3: Compute the value of the bootstrap statistic

$$T_n^*(\widetilde{x}_1^*, \ldots, \widetilde{x}_n^*) = \frac{\left[D_\theta^\varphi \left(\frac{1}{n} \cdot [\widetilde{x}_1^* + \ldots + \widetilde{x}_n^*], \frac{1}{n} \cdot [\widetilde{x}_1 + \ldots + \widetilde{x}_n] \right) \right]^2}{\widehat{S_n^2}(\widetilde{x}_1^*, \ldots, \widetilde{x}_n^*)}.$$

Step 4: Steps 2 and 3 should be repeated a pre-fixed large number B of times to get a set of B estimates, denoted by $\{T_n^{*(1)}, \ldots, T_n^{*(B)}\}$.

Step 5: Compute the bootstrap p-value as the proportion of values in $\{T_n^{*(1)}, \ldots, T_n^{*(B)}\}$ being greater than $T_n(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

To illustrate the preceding algorithm we will apply it to a real-life example which has been analyzed also in previous papers for different purposes (see, for instance, Sinova *et al.* [45]).

Example 4.4. In most of academic institutions it is a common practice to perform surveys among students to evaluate their satisfaction or to rate the level of different courses which are delivered at them. For this purpose questionnaires are designed to gather their students' opinions and judgements. Most of these questionnaires are based on a pre-specified response format, often related to a Likert scale.

In agreement with the policy which has been argued along this paper, a survey was carried out during the II Summer School of the European Centre for Soft Computing (Mieres, Spain) in July 2008. For each course, students attending it, who in this case were familiar with fuzzy numbers because of the courses belonging to a specialized teaching program, were inquired to represent their opinion/valuation about some aspects of each course.

A form similar to the one in Example 2.4 was supplied for students to fill. To ease the drawing of the fuzzy numbers the use of trapezoidal numbers $\mathrm{Tra}(\mathrm{i}_0, \mathrm{i}_1, \mathrm{s}_1, \mathrm{s}_0)$ were suggested. One of the questions to be answered referred to the 'motivation of the course' and the answers from the 29 students attending it were collected in Table 1.

Stud.	i_0	i_1	s_1	s_0	Stud.	i_0	i_1	s_1	s_0	Stud.	i_0	i_1	s_1	s_0
1	50	60	70	80	**11**	80	90	90	100	**21**	56	60	64	70
2	34	40	41	46	**12**	10	30	40	60	**22**	30	40	40	50
3	21	23	34	40	**13**	65	70	70	75	**23**	10	20	20	30
4	70	80	90	100	**14**	20	30	30	40	**24**	60	65	75	80
5	50	60	70	80	**15**	60	70	70	80	**25**	70	76	84	90
6	75	80	90	100	**16**	44	47	53	71	**26**	80	90	90	100
7	70	74	86	90	**17**	60	70	80	90	**27**	55	65	74	80
8	52	60	60	64	**18**	50	60	70	80	**28**	70	80	100	100
9	50	55	60	70	**19**	60	67	72	80	**29**	69	100	100	100
10	60	70	80	90	**20**	90	100	100	100					

TABLE 1. Answers on the Question 'Motivation of the Course' Provided by a Sample of 29 Students Attending It

Let $\mathcal{X}$ denote the RFN 'motivation of the considered course' defined on the population Ω of potential students for the course, and consider the null hypothesis

$$H_0 : \widetilde{E}(\mathcal{X}) = \mathrm{Tra}(50, 60, 70, 80)$$

which is to be tested on the basis of the available sample of fuzzy data supplied by the $n = 29$ students and gathered in Table 1.

Then, the bootstrap test in Algorithm 4.3 (with $\theta = 1/3, \varphi =$ Lebesgue measure on $[0, 1]$, and $B = 10000$) leads to a p-value (i.e., the minimum significance level at which the null hypothesis would be rejected) equal to .648, whence we can conclude that H_0 could be accepted at the most usual the significance levels. Figure 5 shows the sample mean value (on above) and the hypothetical population mean (on below) in this example.

The *two-sample case of testing about the means of two RFSs* takes $H_0^{[2s]} : \widetilde{E}(\mathcal{X}) = \widetilde{E}(\mathcal{Y})$ as the null hypothesis, where $\mathcal{X}, \mathcal{Y} : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$ are supposed to be RFSs for which the 'population' fuzzy means $\widetilde{E}(\mathcal{X})$ and $\widetilde{E}(\mathcal{Y})$ exist. To test $H_0^{[2s]}$ two (either independent or linked) realizations of simple random samples from the two RFSs are considered.

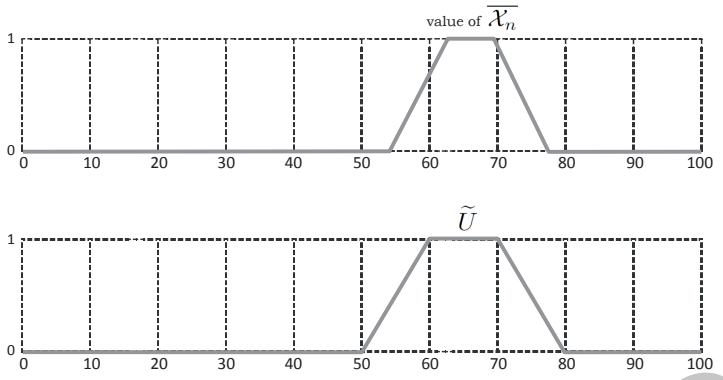

FIGURE 5. Sample Mean Value (on Above) and Hypothetical Population Mean (on Below) in Example 4.4

To test $H_0^{[2s]}$, this equality of fuzzy values has been first translated into the equivalent equality of real numbers $D_\theta^\varphi\big(\widetilde{E}(\mathcal{X}),\widetilde{E}(\mathcal{Y})\big)=0$. As for the one-sample case the developed methods have been the TEST T1-type (see Montenegro *et al.* [34]), the TEST T2-type for independent samples (see Montenegro *et al.* [34]) and the TEST T3-type for dependent samples (see González-Rodríguez *et al.* [17]). The conclusions which have been drawn are similar to those for the one-sample case, the bootstrap approach being the most convenient one.

Finally, the *k-sample case of testing about the means of k RFSs* takes $H_0^{[ks]}$: $\widetilde{E}(\mathcal{X}_1)=\ldots=\widetilde{E}(\mathcal{X}_k)$ as the null hypothesis, where $\mathcal{X}_1,\ldots,\mathcal{X}_k:\Omega\to\mathcal{F}_c^2(\mathbb{R}^p)$ are supposed to be RFSs for which the 'population' fuzzy means $\widetilde{E}(\mathcal{X}_i)$ exist. To test $H_0^{[ks]}$, k (either independent or dependent) realizations of simple random samples from the k RFSs are considered.

To test $H_0^{[ks]}$, this equality of fuzzy values has been first translated into the equivalent equality of real numbers $\sum_{i=1}^k\Big[D_\theta^\varphi\Big(\widetilde{E}(\mathcal{X}_i),\widetilde{E}\big(\frac{1}{k}\cdot[\mathcal{X}_1+\ldots+\mathcal{X}_k]\big)\Big)\Big]^2=0$. The developed methods have been the TEST T1-type (see Gil *et al.* [12]), the TEST T2- and TEST T3-type for independent samples from simple RFNs (see Gil *et al.* [12]), the TEST T2- and TEST T3-type for independent sample from general RFSs (see González-Rodríguez *et al.* [16]), and the TEST T3-type for dependent samples (see Montenegro *et al.* [36]). The conclusions which have been drawn are similar to those for the one- and two-sample case, the bootstrap approach being the most convenient one. The one-way ANOVA test for RFSs have been also extended to the factorial ANOVA (see Nakama *et al.* [37]).

On the other hand, concerning the testing of hypotheses about real-valued parameters, some methods have been recently stated in connection with the variances. In this way, the problem of *testing null hypotheses about the population Fréchet variances of RFS(s)* has been discussed in detail in the one-sample (see Ramos-Guajardo *et al.* [43]) and in the two- and k-sample cases (see Ramos-Guajardo *et al.* [44]).

4.3. **Other Statistical Developments.** Other statistical problems involving RFSs have been studied. Among them, the linear regression and the classification will be now shortly described.

Regarding the *linear regression analysis between two RFSs*, González-Rodríguez *et al.* [14] have faced the following problem:

Let $\mathcal{X}, \mathcal{Y} : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$ be two RFSs associated with the probability space $(\Omega, \mathcal{A}, P)$ and fulfilling the *fuzzy arithmetic-based Linear Model* $\mathcal{Y} = a \cdot \mathcal{X} + \varepsilon$, where $a \in \mathbb{R}$ and $\varepsilon : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$ being an RFS associated with the same probability space and such that $\widetilde{E}(\varepsilon|\mathcal{X}) = B \in \mathcal{F}_c^2(\mathbb{R}^p)$. Assume that the Fréchet variances for the three involved RFSs are finite, the two first ones not vanishing. Then, the population regression function corresponds to $\widetilde{E}(\mathcal{Y}|\mathcal{X}) = a \cdot \mathcal{X} + B$. It should be noted that under the linear model assumption the (random) Hukuhara difference $\mathcal{Y} -_{\mathrm{H}} a \cdot \mathcal{X}$ (where for each $\omega \in \Omega$, $a \cdot \mathcal{X}(\omega) + \mathcal{Y}(\omega) -_{\mathrm{H}} a \cdot \mathcal{X}(\omega) = \mathcal{Y}(\omega)$) is well-defined.

Let $\big((\mathcal{X}_1, \mathcal{Y}_1), \ldots, (\mathcal{X}_n, \mathcal{Y}_n)\big)$ be a random sample from $(\mathcal{X}, \mathcal{Y})$. The goal of the problem is to find the θ, φ-least-squares estimators of the parameters a and B of the considered linear model on the basis of the sample information, that is,

$$\begin{aligned} &\text{Minimize} \quad \frac{1}{n}\sum_{i=1}^{n} \left[D_\theta^\varphi(\mathcal{Y}_i, a \cdot \mathcal{X}_i + B)\right]^2, \\ &\text{subject to} \quad a \in A = \{a \in \mathbb{R} : \mathcal{Y}_i -_H a \cdot \mathcal{X}_i \text{ exists for all } i = 1, \ldots, n\}. \end{aligned}$$

Since sample data are supposed to be generated from the fuzzy arithmetic-based linear model, it would be possible to find an exact solution which can be found in detail in González-Rodríguez *et al.* [14].

A rather different approach and model for the so-called LR fuzzy numbers have been presented in Ferraro *et al.* [8]. Whereas in [14] the regression coefficients affect the explanatory RFS as a whole, in [8] linear models are separately considered for each of the terms characterizing the explanatory LR RFN. A different metric has been used by taking into account these characterizing terms, and solutions have been found for the problem.

In connection with the *classification of fuzzy data*, Colubi *et al.* [5] have recently developed a density-based criterion involving random fuzzy sets. The problem can be stated as follows: let $(\Omega, \mathcal{A}, P)$ be a probability space and assume that for each individual a fuzzy datum is observed, each individual belonging to one of k different categories $g_1, \ldots, g_k$. As available learning sample there is a group of n independent individuals along with the corresponding fuzzy data. The goal is to find a rule allowing us to classify a new individual in one of the k groups on the basis of the associated fuzzy datum. Given a fuzzy value $\widetilde{x} \in \mathcal{F}_c^2(\mathbb{R}^p)$, the suggested Ball-based classification criteria for fuzzy data in [5] can be summarized as follows:

Algorithm 4.5.

Step 1: Compute the distance between the datum to be classified $\widetilde{x}$ and the set of training fuzzy data, that is,

$$d_{j,g} = D_\theta^\varphi(\widetilde{x}, \mathcal{Y}_{j,g}) \quad \text{for all } j \in 1, \ldots, n_g \text{ and all } g \in \{1, \ldots, k\}.$$

Step 2: Fix a value for $\delta > 0$ and compute for each $g \in \{1, \ldots, k\}$,

$$n_{\delta,g} = \sum_{j=1}^{n_g} I_{[0,\delta]}(d_{j,g}).$$

Step 3: Estimate the membership probabilities $p_g = P(G = g|\mathcal{X} \in B(\widetilde{x};\delta))$ by means of

$$\widehat{P}(G = g|\mathcal{X} \in B(\widetilde{x};\delta)) = \frac{n_{\delta,g}}{\sum_{l=1}^{k} n_{\delta,l}},$$

$G : \Omega \to \{1,\ldots,k\}$ being the classification rule.

Step 4: Assign $\widetilde{x}$ to the group $g(\widetilde{x}) \in \{1,\ldots,k\}$ of highest estimated probability.

One of the main issues in this method is to find an appropriate way to fix the value for δ. A possible simple and suitable choice is to consider the maximum of the sample deviations in each group, trying to ensure that the balls are large enough to contain data points of at least one group.

5. Concluding Remarks

Random fuzzy sets have been shown to be a well-formalized notion within the probabilistic setting and using it to model imprecise data enables to preserve most of the key concepts and ideas in Statistical Reasoning. The methodology for statistical analysis of fuzzy data can be directly applied to analyze interval and set-valued data.

Most of the concepts and methods described in the paper can be computed and applied by using a recently developed R-package called SAFD (Statistical Analysis of Fuzzy Data), which has been designed by Lubiano and Trutschnig [50] to perform statistical computations with RFSs. It is being periodically updated.

Based on some tools from Probability and Fuzzy Set Theories, and more precisely on the basis of a characterizing fuzzy representation of real-valued random variables by González-Rodríguez *et al.* [15], an integral methodology to develop statistical inferences on the distributions of real-valued random variables has been derived. This methodology is based on the Aumann-type mean of the fuzzified random variable and it shows convenient properties (like strong consistency and others) because of being mean-based.

A lot of theoretical developments on both the statistical analysis of fuzzy data and the implications to distributions of real-valued random variables remain to be performed (e.g., regarding empirical studies: to consider a sensitivity analysis w.r.t. the choice of the metric for the first type of studies).

Acknowledgements. Authors of this paper are the members of the SMIRE Research Group (http://bellman.ciencias.uniovi/SMIRE) who in 2011 are affiliated to the University of Oviedo. Authors wish to thank Professors Reinhard Viertl and Syed Mohmoud Taheri for having invited them to prepare this survey to the Special Issue on 'Statistical Analysis in Fuzzy Environment' of the Iranian Journal of Fuzzy Systems. They also thank their colleagues at the SMIRE Group from the University 'La Sapienza' at Roma and the European Centre for Soft Computing at Mieres (Asturias, Spain) for their helpful discussions and contributions. The

authors thank their colleagues in the Department of Statistics, OR and DM in the University of Oviedo who have contributed to the topic with their research.

References

[1] M. C. Alonso, T. Brezmes, M. A. Lubiano and C. Bertoluzza, *A generalized real-valued measure of the inequality associated with a fuzzy random variable*, Int. J. Approx. Reas., **26** (2001), 47–66.

[2] C. Bertoluzza, N. Corral and A. Salas, *On a new class of distances between fuzzy numbers*, Mathware & Soft Computing, **2** (1995), 71–84.

[3] A. Colubi, *Statistical inference about the means of fuzzy random variables: applications to the analysis of fuzzy- and real-valued data*, Fuzzy Sets and Systems, **160** (2009), 344–356.

[4] A. Colubi, J. S. Domínguez-Menchero, M. López-Díaz and D. A. Ralescu, *On the formalization of fuzzy random variables*, Information Sciences, **133** (2001), 3–6.

[5] A. Colubi, G. González-Rodríguez, M. A. Gil and W. Trutschnig, *Nonparametric criteria for supervised classification of fuzzy data*, Int. J. Approx. Reas., **52** (2011), 1272–1282.

[6] A. Colubi, M. López-Díaz, J. S. Domínguez-Menchero and M. A. Gil, *A generalized strong law of large numbers*, Prob. Theor. Rel. Fields, **114** (1999), 401–417.

[7] P. Diamond and P. Kloeden, *Metric spaces of fuzzy sets*, Fuzzy Sets and Systems, **100** (1999), 63–71.

[8] M. B. Ferraro, R. Coppi, G. González-Rodríguez and A. Colubi, *A linear regression model for imprecise response*, Int. J. Approx. Reas., **51** (2010), 759–770.

[9] D. García, M. A. Lubiano and M. C. Alonso, *Estimating the expected value of fuzzy random variables in the stratified random sampling from finite populations*, Information Sciences, **138** (2001), 165–184.

[10] M. A. Gil, M. López-Díaz and H. López-García, *The fuzzy hyperbolic inequality index associated with fuzzy random variables*, Eur. J. Oper. Res., **110** (1998), 377–391.

[11] M. A. Gil, M. A. Lubiano, M. Montenegro and M. T. López, *Least squares fitting of an affine function and strength of association for interval-valued data*, Metrika, **56** (2002), 97–111.

[12] M. A. Gil, M. Montenegro, G. González-Rodríguez, A. Colubi and M. R. Casals, *Bootstrap approach to the multi-sample test of means with imprecise data*, Comp. Stat. Data Anal., **51** (2006), 148–162.

[13] E. Giné and J. Zinn, *Bootstrapping general empirical measures*, Ann. Probab., **18** (1990), 851–869.

[14] G. González-Rodríguez, A. Blanco, A. Colubi and M. A. Lubiano, *Estimation of a simple linear regression model for fuzzy random variables*, Fuzzy Sets and Systems, **160** (2009), 357-370.

[15] G. González-Rodríguez, A. Colubi and M. A. Gil, *A fuzzy representation of random variables: an operational tool in exploratory analysis and hypothesis testing*, Comp. Stat. Data Anal., **51** (2006), 163–176.

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

[17] G. González-Rodríguez, A. Colubi, M. A. Gil and P. D'Urso, *An asymptotic two dependent samples test of equality of means of fuzzy random variables*, In: Proc. COMPSTAT'2006, (2006), `http://www.stat.unipg.it/iasc/Proceedings/2006/COMPSTAT/CD/145.pdf`.

[18] G. González-Rodríguez, M. Montenegro, A. Colubi and M. A. Gil, *Bootstrap techniques and fuzzy random variables: Synergy in hypothesis testing with fuzzy data*, Fuzzy Sets and Systems, **157** (2006), 2608–2613.

[19] G. González-Rodríguez, W. Trutschnig and A. Colubi., *Confidence regions for the mean of a fuzzy random variable*, In: Abstracts of IFSA-EUSFLAT 2009, `http://www.eusflat.org/publications/proceedings/IFSA-EUSFLAT_2009/pdf/tema_1433.pdf`.

[20] T. Hesketh, R. Pryor and B. Hesketh, *An application of a computerized fuzzy graphic rating scale to the psychological measurement of individual differences*, Int. J. Man-Machine Studies, **29** (1988), 21–35.

[21] B. Hesketh, T. Hesketh, J. I. Hansen and D. Goranson, *Use of fuzzy variables in developing new scales from the strong interest inventory*, J. Counseling Psychology, **42** (1995), 85–99.
[22] M. Hukuhara, *Intégration des applications measurables dont la valeur est un compact convexe*, Funkcial. Ekvac., **10** (1967), 205-223.
[23] E. P. Klement, M. L. Puri and D. A. Ralescu, *Limit theorems for fuzzy random variables*, Proc. R. Soc. Lond. A, **407** (1986), 171–182.
[24] R. Körner, *An asymptotic α-test for the expectation of random fuzzy variables*, J. Stat. Plann. Infer., **83** (2000), 331–346.
[25] R. Körner and W. Näther, O*n the variance of random fuzzy variables*, In: C. Bertoluzza, M. A. Gil and D. A. Ralescu, eds., Statistical Modeling, Analysis and Management of Fuzzy Data, Physica-Verlag, Heidelberg, (2002), 22–39.
[26] R. Kruse and K. D. Meyer, *Statistics with vague data*, D. Reidel Publishing Company, Dordrecht, 1987.
[27] H. Kwakernaak, *Fuzzy random variables-I. definitions and theorems*, Information Sciences, **15** (1978), 1–29.
[28] H. Kwakernaak, *Fuzzy random variables-II. algorithms and examples for the discrete case*, Information Sciences, **17** (1979), 253–278.
[29] H. López-García, M. A. Gil, N. Corral and M. T. López, *Estimating the fuzzy inequality associated with a fuzzy random variable in random samplings from finite populations*, Kybernetika, **34** (1998), 149–161.
[30] M. A. Lubiano, M. C. Alonso and M. A. Gil, *Statistical inferences on the S-mean squared dispersion of a fuzzy random variable*, In: B. de Baets, J. Fodor and L. T. Koczy, eds., Proceedings of EUROFUSE-SIC99, University of Veterinary Science, Budapest, (1999), 532–537.
[31] M. A. Lubiano and M. A. Gil, *Estimating the expected value of fuzzy random variables in random samplings from finite populations*, Stat. Pap., **40**(1999), 277–295.
[32] M. A. Lubiano, M. A. Gil and M. López-Díaz, *On the Rao-Blackwell theorem for fuzzy random variables*, Kybernetika, **35** (1999), 167–175.
[33] M. A. Lubiano, M. A. Gil, M. López-Díaz and M. T. López, *The $\overrightarrow{\lambda}$-mean squared dispersion associated with a fuzzy random variable*, Fuzzy Sets and Systems, **111** (2000), 307–317.
[34] M. Montenegro, M. R. Casals, M. A. Lubiano and M. A. Gil, *Two-sample hypothesis tests of means of a fuzzy random variable*, Information Sciences, **133** (2001), 89–100.
[35] M. Montenegro, A. Colubi, M. R. Casals and M. A. Gil, *Asymptotic and Bootstrap techniques for testing the expected value of a fuzzy random variable*, Metrika, **59** (2004), 31–49.
[36] M. Montenegro, M. T. López-García, M. A. Lubiano and G. González-Rodríguez, *A dependent multi-sample test for fuzzy means*, In: Abst. 2nd Workshop ERCIM WG Comput. & Statist, (2009), 102.
[37] T. Nakama, A. Colubi and M. A. Lubiano, *Factorial analysis of variance for fuzzy data*, In: Abst. CFE'10 & ERCIM'10, (2010), 88.
[38] H. T. Nguyen, *A note on the extension principle for fuzzy sets*, J. Math. Anal. Appl., **64** (1978), 369–380.
[39] M. L. Puri and D. A. Ralescu, *Differentials of fuzzy functions*, J. Math. Anal. Appl., **91** (1983), 552–558.
[40] M. L. Puri and D. A. Ralescu, *The concept of normality for fuzzy random variables*, Ann. Probab., **11** (1985), 1373–1379.
[41] M. L. Puri and D. A. Ralescu, *Fuzzy random variables*, J. Math. Anal. Appl., **114** (1986), 409–422.
[42] S. Ramezanzadeh, M. Memariani and S. Saati, *Data envelopment analysis with fuzzy random inputs and outputs: a chance-constrained programming approach*, Iranian Journal of Fuzzy Systems, **2** (2005), 21–29.
[43] A. B. Ramos-Guajardo, A. Colubi, G. González-Rodríguez and M. A. Gil, *One sample tests for a generalized Fréchet variance of a fuzzy random variable*, Metrika, **71** (2010), 185–202.

[44] A. B. Ramos-Guajardo and M. A. Lubiano, *K-sample tests for equality of variances of random fuzzy sets*, Comp. Stat. Data Anal., **56** (2012), 956–966.
[45] B. Sinova, M. A. Gil, A. Colubi and S. Van Aelst, *The median of a random fuzzy number. The 1-norm distance approach*, Fuzzy Sets and Systems, **200** (2011), 99-115.
[46] B. Sinova, S. de la Rosa de Sáa and M. A. Gil, *A generalized L^1-type metric between fuzzy numbers for an approach to central tendency of fuzzy data*, Information Sciences, under 2nd review.
[47] S. M. Taheri and M. Kelkinnama, *Fuzzy linear regression based on least absolutes deviations*, Iranian Journal of Fuzzy Systems, **9(1)** (2012), 121-140.
[48] P. Terán, *A strong law of large numbers for random upper semicontinuous functions under exchangeability conditions*, Statist. Prob. Lett., **65** (2003), 251–258.
[49] W. Trutschnig, G. González-Rodríguez, A. Colubi and M. A. Gil, *A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread*, Information Sciences, **179** (2009), 3964–3972.
[50] W. Trutschnig and M. A. Lubiano, *SAFD: statistical analysis of fuzzy data*, http://cran.r-project.org/web/packages/SAFD/index.html.
[51] R. Viertl and D. Hareter, *Fuzzy information and stochastics*, Iranian Journal of Fuzzy Systems, **1** (2004), 43–56.
[52] L. A. Zadeh, *The concept of a linguistic variable and its application to approximate reasoning, Part 1*, Information Sciences, **8** (1975), 199–249; *Part 2*, Information Sciences, **8** (1975), 301–353; *Part 3*, Information Sciences, **9** (1975), 43–80.
[53] L. A. Zadeh, *Discussion: probability theory and fuzzy logic are complementary rather than competitive*, Technometrics, **37** (1995), 271–276.

A. Blanco-Fernández, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: blancoangela@uniovi.es

M. R. Casals, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: rmcasals@uniovi.es

A. Colubi, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: colubi@uniovi.es

N. Corral, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: norbert@uniovi.es

M. García-Bárzana, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: martagb5@gmail.com

M. A. Gil*, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: magil@uniovi.es

G. González-Rodríguez, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: gil@uniovi.es

M.T. López Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: mtlopez@uniovi.es

M. Montenegro, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: mmontenegro@uniovi.es

M. A. Lubiano, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: lubiano@uniovi.es

A. B. Ramos-Guajardo, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: ramosana@uniovi.es

S. de la Rosa de Sáa, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: delarosasara@uniovi.es

B. Sinova, Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain
E-mail address: sinovabeatriz@uniovi.es

*Corresponding author