# Measuring the dissimilarity between the distributions of two random fuzzy numbers

María Asunción Lubiano, María Ángeles Gil, Beatriz Sinova,
María Rosa Casals, and María Teresa López

**Abstract** In a previous paper the fuzzy characterizing function of a random fuzzy number was introduced as an extension of the moment generating function of a real-valued random variable. Properties of the fuzzy characterizing function have been examined, among them, the crucial one proving that it unequivocally determines the distribution of a random fuzzy number in a neighborhood of 0. This property suggests to consider the empirical fuzzy characterizing function as a tool to measure the dissimilarity between the distributions of two random fuzzy numbers, and its expected descriptive potentiality is illustrated by means of a real-life example.

## 1 Introduction

The formalization of random fuzzy numbers as Borel-measurable fuzzy number-valued mappings associated with a probability space, this one modeling a random experiment, allows us to properly refer to its induced distribution as well as to the independence of random fuzzy numbers. Nevertheless, although the existence of such an induced distribution is clear (and it can be easily determined in the sample case), there is not a sound general concept which enables us to develop some probabilistic and statistical results we have in the real-valued case, like the distribution function of a real-valued random variable. Moreover, there are not exact or approximated models widely applicable and realistic enough for the induced distribution.

In Sinova *et al.* [9] a function characterizing the induced distribution of a random fuzzy number has been defined. This function aims to extend the moment generating function of a real-valued random variable (and, there-

Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, 33071 Oviedo
lubiano@uniovi.es,magil@uniovi.es,sinovabeatriz@uniovi.es,rmcasals@uniovi.es,
mtlopez@uniovi.es

fore, there are just a few distributions for which it does not exist) and it is based on the Aumann-type mean of a random fuzzy number. Since the extension preserves the convenient characterizing ability of the moment generating function, one can think of using it to measure to some extent whether the (induced) distributions of two random fuzzy numbers coincide or not. More concretely, we can consider to state a measure of the dissimilarity of such distributions.

This paper aims to empirically analyze the descriptive behaviour of this measure by means of a real-life example. The derived descriptive conclusions will be compared with some inferential ones which have been recently drawn. Some open problems will be finally proposed.

## 2 Preliminaries

Fuzzy sets, and particularly fuzzy numbers, are very suitable to cope with the imprecision of different real-life data, especially those coming from human thought and experience in variables like quality perception, satisfaction, opinion, etc.

**Definition 1.** A mapping $\widetilde{U} : \mathbb{R} \to [0,1]$ is said to be a (bounded) *fuzzy number* if its $\alpha$-levels

$$\widetilde{U}_\alpha = \begin{cases} \{x \in \mathbb{R} : \widetilde{U}(x) \geq \alpha\} & \text{if } \alpha \in (0,1] \\ \mathrm{cl}\{x \in \mathbb{R} : \widetilde{U}(x) > 0\} & \text{if } \alpha = 0 \end{cases}$$

(with cl denoting the topological closure) are nonempty compact intervals for all $\alpha \in [0,1]$. The class of (bounded) fuzzy numbers will be denoted by $\mathcal{F}_c^*(\mathbb{R})$.

To deal with fuzzy numbers in this paper we should consider the extension of the sum and product by a scalar as well as that for the exponential function, which will be supposed to be based on Zadeh's extension principle [10] and coincides level-wise with the usual interval arithmetic and function image (see Nguyen [7]).

**Definition 2.** Let $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c^*(\mathbb{R})$ and $\gamma \in \mathbb{R}$. The *sum* of $\widetilde{U}$ and $\widetilde{V}$ is the fuzzy number $\widetilde{U} + \widetilde{V}$ such that

$$(\widetilde{U}+\widetilde{V})_\alpha = \text{Minkowski sum of } \widetilde{U}_\alpha \text{ and } \widetilde{V}_\alpha = [\inf \widetilde{U}_\alpha + \inf \widetilde{V}_\alpha, \sup \widetilde{U}_\alpha + \sup \widetilde{V}_\alpha].$$

The *product* of $\widetilde{U}$ *by the scalar* $\gamma$ is the fuzzy number $\gamma \cdot \widetilde{U}$ such that

$$(\gamma \cdot \widetilde{U})_\alpha = \gamma \cdot \widetilde{U}_\alpha = \begin{cases} \left[\gamma \inf \widetilde{U}_\alpha, \gamma \sup \widetilde{U}_\alpha\right] & \text{if } \gamma \in [0, \infty) \\ \left[\gamma \sup \widetilde{U}_\alpha, \gamma \inf \widetilde{U}_\alpha\right] & \text{otherwise.} \end{cases}$$

The (induced) *image of* $\widetilde{U}$ *through the exponential function* is the fuzzy number $e^{\gamma\cdot\widetilde{U}}$ such that

$$(e^{\gamma\cdot\widetilde{U}})_\alpha = \begin{cases} \left[e^{\gamma \inf \widetilde{U}_\alpha}, e^{\gamma \sup \widetilde{U}_\alpha}\right] & \text{if } \gamma \in [0,\infty) \\ \left[e^{\gamma \sup \widetilde{U}_\alpha}, e^{\gamma \inf \widetilde{U}_\alpha}\right] & \text{otherwise.} \end{cases}$$

If a random experiment leads to data which can be suitably modeled in terms of fuzzy numbers, one should also properly model the random mechanism generating such data to analyze them in a rigorously established setting. The concept of random fuzzy number (or one-dimensional fuzzy random variable, as coined and introduced by Puri and Ralescu [8] is an appropriate model to formalize a random mechanism associating with each experimental outcome a fuzzy number. That is, random fuzzy numbers are mainly addressed to deal with the 'ontic' view of experimental fuzzy data (see Couso and Dubois [1]).

**Definition 3.** Let $\mathcal{K}_c(\mathbb{R})$ be the space of nonempty compact intervals. Given a probability space $(\Omega, \mathcal{A}, P)$, a *random fuzzy number* associated with it is a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R})$ such that for each $\alpha \in [0,1]$ the set-valued mapping $\mathcal{X}_\alpha : \Omega \to \mathcal{K}_c(\mathbb{R})$ (with $\mathcal{X}_\alpha(\omega) = \big(\mathcal{X}(\omega)\big)_\alpha$) is a compact random interval.

Equivalently, a *random fuzzy number* is a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^*(\mathbb{R})$ such that it is Borel-measurable w.r.t. the Borel $\sigma$-field generated on $\mathcal{F}_c^*(\mathbb{R})$ by the topology induced by several different metrics, like the 2-norm distance

$$\rho_2(\widetilde{U},\widetilde{V}) = \sqrt{\frac{1}{2}\int_{[0,1]} \left(\left[\inf \widetilde{U}_\alpha - \inf \widetilde{V}_\alpha\right]^2 + \left[\sup \widetilde{U}_\alpha - \sup \widetilde{V}_\alpha\right]^2\right) d\alpha}$$

by Diamond and Kloeden [2].

As we have already pointed out, the assumed Borel-measurability of random fuzzy numbers in the second equivalent definition allows us to trivially induce the distribution (from $P$) of a random fuzzy number.

A relevant measure in summarizing such an induced distribution is the mean value, which has been defined by Puri and Ralescu [8]) as follows:

**Definition 4.** Given a probability space $(\Omega, \mathcal{A}, P)$ and a random fuzzy number $\mathcal{X}$ associated with it, the *(population) Aumann-type mean value of* $\mathcal{X}$ is the fuzzy number $\widetilde{E}(\mathcal{X})$, if it exists, such that for each $\alpha \in [0,1]$

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

In particular, if one deals with a finite sample of observations from a random fuzzy number $\mathcal{X}$, say $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$, the corresponding *(sample) Aumann-type mean value* is the fuzzy number

$$\overline{\widetilde{\boldsymbol{x}}} = \frac{1}{n} \cdot (\widetilde{x}_1 + \cdots + \widetilde{x}_n)\,.$$

On the basis of the Aumann-type mean value of a random fuzzy number, one can formally extend the notion of moment generating function of a real-valued random variable as follows (see Sinova *et al.* [9]):

**Definition 5.** Given a probability space $(\Omega, \mathcal{A}, P)$ and a random fuzzy number $\mathcal{X}$ associated with it, the *(population) fuzzy characterizing function* of $\mathcal{X}$ is the mapping $\widetilde{M}_{\mathcal{X}}$ defined on a neighborhood of 0 that associates with each $t$ in the neighborhood the fuzzy number $\widetilde{M}_{\mathcal{X}}(t) = \widetilde{E}\left(e^{t\mathcal{X}}\right)$, if it exists. That is, for each $\alpha \in [0,1]$

$$\left(\widetilde{M}_{\mathcal{X}}(t)\right)_{\alpha} = \begin{cases} \left[E(e^{t \inf \mathcal{X}_\alpha}), E(e^{t \sup \mathcal{X}_\alpha})\right] & \text{if } t \geq 0 \\ \left[E(e^{t \sup \mathcal{X}_\alpha}), E(e^{t \inf \mathcal{X}_\alpha})\right] & \text{otherwise.} \end{cases}$$

In particular, if one deals with a finite sample of observations from a random fuzzy number $\mathcal{X}$, say $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$, the corresponding *empirical fuzzy characterizing function* is the mapping $\widehat{\widetilde{M}_{\widetilde{\boldsymbol{x}}}}$ associating with each $t$ in a neighborhood of 0 the fuzzy number

$$\widehat{\widetilde{M}_{\widetilde{\boldsymbol{x}}}}(t) = \frac{1}{n} \cdot \left(e^{t\widetilde{x}_1} + \cdots + e^{t\widetilde{x}_n}\right).$$

As shown in [9], the fuzzy characterizing function preserves most of the properties of the moment generating one in the real-valued case, but the one associated with the moment generation. However, it keeps the crucial property of characterization of the induced distribution of the associated random element, so that if $\mathcal{X}$ and $\mathcal{Y}$ are two random fuzzy numbers for which the fuzzy characterizing functions exist and coincide in a neighborhood of 0, then $\mathcal{X}$ and $\mathcal{Y}$ should be equally distributed.

In the next section, we are going to take advantage of this characterizing skill to state a descriptive measure for the dissimilarity between the sample distributions of two random fuzzy numbers.

## 3 A sample measure for the dissimilarity between the distributions of two fuzzy datasets

This section aims to state an index for the dissimilarity between the distributions of two fuzzy datasets. Due to the characterizing property, and being inspired by ideas in some statistics for the homogeneity of distributions in the real-valued case (see, for instance, Meintanis [5], Mora and Mora-López [6], who also suggest the correction in contrast to the measure in Lubiano *et al.* [3]), it seems plausible to consider in the current setting a statistic based

on distances between the sample fuzzy characterizing functions in a narrow neighborhood of 0.

In this way, for an arbitrarily fixed $\varepsilon > 0$:

**Definition 6.** The *$\varepsilon$-sample dissimilarity between the distributions of samples $\widetilde{\boldsymbol{x}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ and $\widetilde{\boldsymbol{y}} = (\widetilde{y}_1, \ldots, \widetilde{y}_m)$* is given by the index

$$\varrho_{n,m,\varepsilon}(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}}) = \frac{1}{\varepsilon}\sqrt{\frac{nm}{n+m}} \max_{t\in[-\varepsilon,\varepsilon]} \rho_2\left(\widehat{M_{\widetilde{\boldsymbol{x}}}}(t), \widehat{M_{\widetilde{\boldsymbol{y}}}}(t)\right).$$

In this section we are going to apply the preceding measure on a dataset from a real-life situation.

*Example.* The nine items displayed in Table 1 have been drawn from the TIMSS/PIRLS 2011 Student questionnaire. This questionnaire is conducted in many countries and it is to be responded by fourth grade students (nine to ten years old) in connection with some aspects about reading, math and science.

**Table 1** Items selected from the TIMSS-PIRLS 2011 Student Questionnaire

| | READING IN SCHOOL |
|---|---|
| $R.1$ | I like to read things that make me think |
| $R.2$ | I learn a lot from reading |
| $R.3$ | Reading is harder for me than any other subject |
| | MATHEMATICS IN SCHOOL |
| $M.1$ | I like mathematics |
| $M.2$ | My teacher is easy to understand |
| $M.3$ | Mathematics is harder for me than any other subject |
| | SCIENCE IN SCHOOL |
| $S.1$ | My teacher taught me to discover science in daily life |
| $S.2$ | I read about in my spare time |
| $S.3$ | Science is harder for me than any other subject |

These nine items have been originally designed to be answered in accordance with a 4-point Likert scale (DISAGREE A LOT, DISAGREE A LITTLE, AGREE A LITTLE, AGREE A LOT).

Recently, the questionnaire form involving these nine items, along with a few more ones about students' support resources at home, has been adapted to allow also a fuzzy rating scale-based one (see Figure 1 for Question $M.2$). For the full paper-and-pencil and computerized versions of the questionnaire, see http://bellman.ciencias.uniovi.es/SMIRE/FuzzyRatingScaleQuestionnaire-SanIgnacio.html.

The fuzzy rating scale (see, e.g., [3, 4]) has been designed with reference interval $[0, 10]$. The adapted questionnaire has been conducted on 69 fourth grade students from Colegio San Ignacio (Oviedo-Asturias, Spain). The complete dataset can be found in the webpage containing the forms.

Now we are going to examine whether the fuzzy rating scale-based responses seem or not to be affected by respondents' sex, filled form version and the fact that respondents have or not an individual bedroom at home.

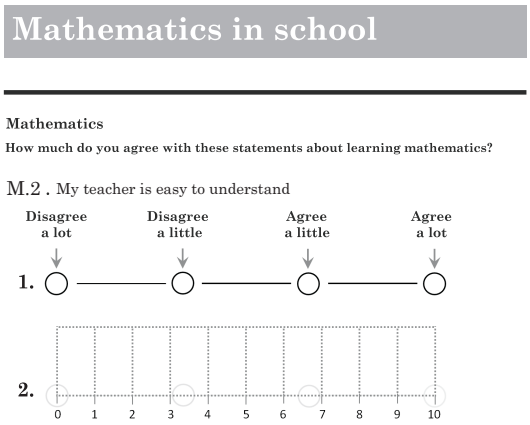


**Fig. 1** Example of the double-response form to an item

For this purpose, and for each of the three variables, we have first considered the (descriptive) dissimilarity index with $\varepsilon = .001, .01$ and .1 (a deeper and exhaustive discussion about the choice of $\varepsilon$ should be developed in the future). Secondly, as an alternative (albeit inferential) way to discuss such an influence, we have considered tests in Lubiano *et al.* [4] for the two-sample equality of independent means and compute the associated $p$-values when the chosen metric is $\rho_2$. Tables 2, 3 and 4 gather the outputs for the descriptive and inferential analyses.

**Table 2** $\varepsilon$-sample dissimilarity between the distributions of girls' and boys' samples for $\varepsilon = .001, .01, .1$ and $\rho_2$-based testing $p$-values for the equality of means

| Item | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .001$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .01$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .1$) | $\rho_2$ two-sample test $p$-values |
|---|---|---|---|---|
| $R.1$ | 0.3874 | 0.4056 | 0.665 | 0.502 |
| $R.2$ | 0.2397 | 0.2544 | 0.4759 | 0.702 |
| $R.3$ | 0.6087 | 0.6416 | 1.1206 | 0.425 |
| $M.1$ | 1.2692 | 1.3337 | 2.2487 | 0.049 |
| $M.2$ | 0.3713 | 0.39 | 0.658 | 0.574 |
| $M.3$ | 0.6207 | 0.6469 | 1.0211 | 0.49 |
| $S.1$ | 0.6784 | 0.7145 | 1.2232 | 0.275 |
| $S.2$ | 0.2754 | 0.2942 | 0.5738 | 0.687 |
| $S.3$ | 0.4223 | 0.4394 | 0.6851 | 0.606 |

As an attempt to analyze the coherence between the descriptive dissimilarity and the inferential testing for the equality of means outputs, we have computed Pearson's correlation coefficient $r$ between both series of outputs. In connection with sex we have that $r = -0.9567$ (if $\varepsilon = .001$), $r = -0.9574$ (if $\varepsilon = .01$), and $r = -0.9572$ (if $\varepsilon = .1$).

**Table 3** $\varepsilon$-sample dissimilarity between the distributions of paper-and-pencil respondents' sample and computerized respondents' sample for $\varepsilon = .001, .01, .1$ and $\rho_2$-based testing $p$-values for the equality of means

| Item | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .001$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .01$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .1$) | $\rho_2$ two-sample test $p$-values |
|---|---|---|---|---|
| $R.1$ | 1.0148 | 1.0606 | 1.678 | 0.065 |
| $R.2$ | 1.1045 | 1.1724 | 2.1556 | 0.029 |
| $R.3$ | 0.7244 | 0.7497 | 1.0904 | 0.366 |
| $M.1$ | 0.8622 | 0.9008 | 1.4245 | 0.176 |
| $M.2$ | 1.3347 | 1.4103 | 2.5025 | 0.01 |
| $M.3$ | 1.5316 | 1.6148 | 2.8161 | 0.062 |
| $S.1$ | 1.5403 | 1.6122 | 2.6124 | 0.016 |
| $S.2$ | 0.6827 | 0.7058 | 0.9985 | 0.292 |
| $S.3$ | 1.5221 | 1.5978 | 2.664 | 0.042 |

**Table 4** $\varepsilon$-sample dissimilarity between the distributions of respondents' sample with individual bedroom and respondents' sample with shared bedroom for $\varepsilon = .001, .01, .1$ and $\rho_2$-based testing $p$-values for the equality of means

| Item | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .001$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .01$) | $\varrho_{n,m,\varepsilon}$ ($\varepsilon = .1$) | $\rho_2$ two-sample test $p$-values |
|---|---|---|---|---|
| $R.1$ | 0.5859 | 0.6277 | 1.2509 | 0.294 |
| $R.2$ | 1.2238 | 1.3036 | 2.4909 | 0.013 |
| $R.3$ | 0.4755 | 0.4983 | 0.8005 | 0.543 |
| $M.1$ | 0.9392 | 0.9919 | 1.7486 | 0.188 |
| $M.2$ | 0.3153 | 0.3365 | 0.6604 | 0.685 |
| $M.3$ | 0.6548 | 0.6987 | 1.3606 | 0.46 |
| $S.1$ | 0.2659 | 0.2746 | 0.394 | 0.772 |
| $S.2$ | 0.5868 | 0.6063 | 0.8561 | 0.373 |
| $S.3$ | 0.8058 | 0.859 | 1.6633 | 0.366 |

In connection with the filled format we have that $r = -0.8269$ (if $\varepsilon = .001$), $r = -0.8331$ (if $\varepsilon = .01$), and $r = -0.8664$ (if $\varepsilon = .1$). In connection with bedroom type for respondents we have that $r = -0.9437$ (if $\varepsilon = .001$), $r = -0.9426$ (if $\varepsilon = .01$), and $r = -0.9145$ (if $\varepsilon = .1$).

Consequently, there is a high linear relationship between both tools. Notice that the correlation coefficient is not expected to be exactly equal to $-1$, not only because we are using samples and linearity could be a restrictive assumption, but also because the dissimilarity index is related to the whole distribution whereas $p$-values concern only their means.

## 4 Conclusions and future directions

By looking at the outputs in Table 2, one can conclude both descriptively (through the dissimilarity measure) and inferentially (through the $p$-value) that sex affects the liking for mathematics (related to item $M.1$). Actually, $M.1$ is the only item among the 9 in the adapted questionnaire for which $\varrho_{n,m,.001} > 1$ and the $p$-value is lower than .05.

By looking at the outputs in Table 3, one can conclude that the version form affects (to a rather great extent) the response to items $R.1$, $R.2$, $M.2$, $M.3$, $S.1$ and $S.3$, for which $\varrho_{n,m,.001} > 1$ and the $p$-value is always lower or much lower than .07.

By looking at the outputs in Table 4, one can conclude that having or not an individual bedroom at home affects students' learning from reading (related to item $R.2$), for which $\varrho_{n,m,.001} > 1$ and the $p$-value is lower than .02.

On the other hand, the measure in this paper has been simply applied for descriptive purposes. Consequently, we cannot attempt to interpret the significance of the dissimilarity measure. It would be desirable to consider this measure in the near future to develop inferential methods (more concretely, for testing hypothesis about the homogeneity of the population distributions of two random fuzzy numbers).

**Acknowledgements** Authors are grateful to Colegio San Ignacio in Oviedo-Asturias (Spain) for allowing us to collect the data in the real-life example. The research in this paper has been partially supported by/benefited from Principality of Asturias Grant GRUPIN14-101, and the Spanish Ministry of Economy and Competitiveness Grants MTM2015-63971-P and MTM2013-44212-P. Their financial support is gratefully acknowledged.

## References

1. Couso I, Dubois D (2014) Statistical reasoning with set-valued information: Ontic *vs.* epistemic views. Int J Appr Reas 55(7):1502–1518
2. Diamond P, Kloeden P (1999) Metric spaces of fuzzy sets. Fuzzy Sets Syst 100:63–71
3. Lubiano MA, De la Rosa de Sáa S, Montenegro M, Sinova B, Gil, MA (2016) Descriptive analysis of responses to items in questionnaires. Why not using a fuzzy rating scale? Inform Sci (submitted, under review)
4. Lubiano MA, Montenegro M, Sinova B, De la Rosa de Sáa S, Gil, MA (2016) Hypothesis testing for means in connection with fuzzy rating scale-based data: algorithms and applications. Eur J Oper Res 251:918–929
5. Meintanis SG (2007) A KolmogorovSmirnov type test for skew normal distributions based on the empirical moment generating function. J Stat Plan Infer 137:2681–2688
6. Mora J, Mora-López L (2010) Comparing distributions with bootstrap techniques: An application to global solar radiation. Math Comp Simul 81:811–819
7. Nguyen HT (1978) A note on the extension principle for fuzzy sets. J Math Anal Appl 64:369–380
8. Puri ML, Ralescu DA (1986) Fuzzy random variables. J Math Anal Appl 114:409–422
9. Sinova B, Casals MR, Gil MA, Lubiano MA (2015) The fuzzy characterizing function of the distribution of a random fuzzy number. Appl Math Model 39(14):4044–4056
10. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning, Part 1. Inform Sci 8:199–249; Part 2. Inform Sci 8:301–353; Part 3. Inform Sci 8:43–80