Empirical Sensitivity Analysis on the Influence of the Shape of Fuzzy Data on the Estimation of Some Statistical Measures

María Asunción Lubiano[1], Sara de la Rosa de Sáa[1], Beatriz Sinova[1,2], and María Ángeles Gil[1]

[1] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, 33007 Oviedo, Spain
{lubiano,delarosasara,sinovabeatriz,magil}@uniovi.es
[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Gent, Belgium

Abstract. This paper means an introduction to analyze whether the choice of the shape for fuzzy data in their statistical analysis can or cannot affect the conclusions of such an analysis. More concretely, samples of fuzzy data are simulated in accordance with different assumptions (distributions) concerning four relevant points (namely, those determining their core and support), and later, by preserving core and support, the 'arms' are changed by considering trapezoidal, Π-curves, and some LR fuzzy numbers. For the simulations obtained with each of the considered shapes, several characteristics have been estimated: Aumann-type mean, 1-norm and wabl/ldev/rdev medians and Fréchet's variance. A comparative analysis with the bias, mean squared distance and variance of the estimates is finally included.

Keywords: fuzzy data, estimation, statistical measures, sensitivity analysis.

1 Introduction

Along the last years a distance-based methodology has been developed to analyze fuzzy number-valued data from a statistical perspective (see Blanco-Fernández *et al.* [2] for a recent review). The methodology assumes that data are generated from random elements taking on fuzzy numbers values (the so-called random fuzzy numbers or -one dimensional- fuzzy random variables in Puri and Ralescu's sense [10]).

Almost all the already developed methods refer to the estimation or to the hypothesis testing about some summary measures of the distributions of the random elements producing fuzzy-valued data. These methods are mostly theoretically supported, but empirical studies have been also conducted either to corroborate some of their generally stated properties or as an alternative when formal general results or conclusions cannot be stated.

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*, Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_15

Most of these empirical developments have been based on simulations from random mechanisms leading to trapezoidal fuzzy number values. This assumption is often considered in practice to ease both the drawing and the computing processes (see, as a recent example the studies in De la Rosa de Sáa *et al.* [4]) although this is not at all mandatory from a formal viewpoint. Actually, Pedrycz [9], Grzegorzewski [6], [7], Grzegorzewski and Pasternak-Winiarska [8], Ban *et al.* [1], and others, have provided with different arguments to employ triangular or trapezoidal fuzzy numbers or approximations preserving ambiguity, expected interval, and so on.

An open problem that has been often commented in the papers related to the aforementioned distance-based methodology is that of discussing whether or not the shape of the fuzzy data influences the statistical conclusions. Since fuzzy data are essentially subjective in this respect, it is convenient to know whether this subjectivity can importantly affect the outputs from the methods.

This paper aims to analyze such a possible influence in which concerns the estimation of some summary measures, namely, three location ones (Aumann-type mean, and two L^1-type medians), and the Fréchet variance of the fuzzy dataset. For this purpose, simulations have been carried out from random mechanisms generating different types of fuzzy values, but data of different type sharing the core (i.e., the 1-level) and the closure of the support (0-level).

2 The Simulation Procedures

To analyze how sensitive the considered summary measures are w.r.t. changes in shape, the simulations we have carried out refer to the four key points characterizing the involved fuzzy numbers (more concretely, those determining their core and support). Six different shapes (T1 to T6, see Figure 1) based on the same four-tuple are separately employed. It is known that for any fuzzy number A there exist four numbers $a_1, a_2, a_3, a_4 \in \mathbb{R}$ and two functions $l_A, r_A : \mathbb{R} \to [0,1]$, where l_A is nondecreasing and r_A is nonincreasing, such that we can describe A with its membership function in the following manner,

$$A(x) = \begin{cases} 0 & \text{if } x < a_1 \\ l_A(x) & \text{if } a_1 \leq x < a_2 \\ 1 & \text{if } a_2 \leq x \leq a_3 \\ r_A(x) & \text{if } a_3 < x \leq a_4 \\ 0 & \text{if } a_4 < x. \end{cases}$$

The corresponding fuzzy numbers have been obtained by using different l_A and r_A functions: linear functions in T1 (trapezoidal fuzzy numbers), quadratic functions with T2 (Π-curves, see, for instance, [3]) and shape functions handling parametric monotonic Hermite-type interpolation in T3-T4 (LR fuzzy numbers using (2,2)-rational splines) and T5-T6 (LR fuzzy numbers using mixed exponential splines). For more details about the considered LR fuzzy numbers see, for instance, [13].

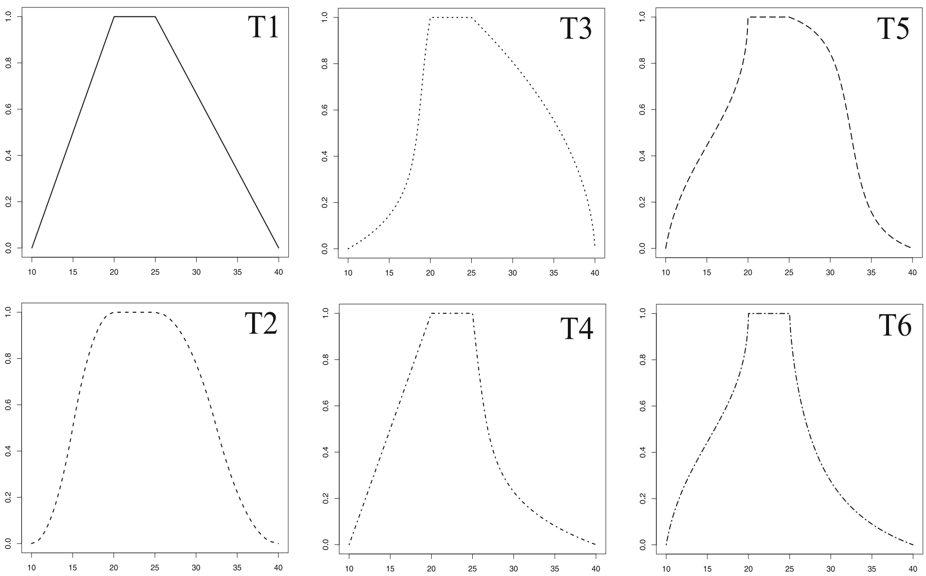

Fig. 1. Six types of fuzzy numbers sharing core and support and differing in shape. On the left, trapezoidal (top) and *Π*-curve (bottom), along with four different *LR* fuzzy numbers on the middle and the right

For each of these six shapes, some simulations studies have been conducted, generating the corresponding fuzzy numbers in two different ways:

Step 1. A sample of fuzzy numbers of the given shape has been obtained by simulating from

- four real-valued random variables X_i ($i = 1, 2, 3, 4$), defining a random fuzzy number $\mathcal{X}$ in Puri and Ralescu's sense, namely, $X_1 = (\inf \mathcal{X}_1 + \sup \mathcal{X}_1)/2$, $X_2 = (\sup \mathcal{X}_1 - \inf \mathcal{X}_1)/2$, $X_3 = \inf \mathcal{X}_1 - \inf \mathcal{X}_0$, $X_4 = \sup \mathcal{X}_0 - \sup \mathcal{X}_1$ (whence $\inf \mathcal{X}_0 = X_1 - X_2 - X_3$, $\inf \mathcal{X}_1 = X_1 - X_2$, $\sup \mathcal{X}_1 = X_1 + X_2$, $\sup \mathcal{X}_0 = X_1 + X_2 + X_4$);
- In the FIRST STUDY (similar to some ones considered by Sinova *et al.*, see [11], [12]), the sample size is $n = 100$ and two cases related to these four random variables X_i have been considered: one in which X_i are independent (CASE 1) and another one in which they are dependent (CASE 2). More specifically, CASE 1 assumes that
  - •• $X_1 \sim \mathcal{N}(0, 1)$ and $X_2, X_3, X_4 \sim \chi^2_1$, all of them being independent

  whereas CASE 2 assumes that
  - •• $X_1 \sim \mathcal{N}(0, 1)$ and $X_2, X_3, X_4 \sim 1/(X_1^2 + 1)^2 + 0.1 \cdot \chi^2_1$, where χ^2_1 is supposed to be independent of X_1, and the three involved χ^2_1 being independent.
- In the SECOND STUDY (which follows the ideas by De la Rosa de Sáa *et al.* [4] in developing comparative studies in connection with questionnaires based on the fuzzy rating scale, using the referential [0,10]), the

simulation strategy has mimicked the human behavior by considering a finite mixture of three different procedures. Concretely, 100000 fuzzy values have been generated in the following way:

- 5% of the data have been obtained by first considering a simulation from a simple random sample of size 4 (X_1, X_2, X_3, X_4) from a beta population $X \sim \beta(1,1)$, later scaling it in [0,10] and finally considering the ordered sample $(X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)})$.
- 35% of the data have been obtained considering a simulation of four random variables X_i as follows:
  $$\begin{aligned} &X_1 \sim \beta(1,1), \\ &X_2 \sim \text{Uniform}\big[0, \min\{1/10, X_1, 1-X_1\}\big], \\ &X_3 \sim \text{Uniform}\big[0, \min\{1/5, X_1 - X_2\}\big], \\ &X_4 \sim \text{Uniform}\big[0, \min\{1/5, 1-X_1-X_2\}\big]; \end{aligned}$$
- 60% of the data have been obtained considering a simulation of four random variables X_i as follows:
  $$\begin{aligned} &X_1 \sim \beta(1,1), \\ &X_2 \sim \begin{cases} \text{Exp}(200) & \text{if } X_1 \in [0.25, 0.75] \\ \text{Exp}(100 + 4\,X_1) & \text{if } X_1 < 0.25 \\ \text{Exp}(500 - 4\,X_1) & \text{otherwise} \end{cases} \\ &X_3 \sim \begin{cases} \gamma(4,100) & \text{if } X_1 - X_2 \geq 0.25 \\ \gamma(4, 100 + 4\,X_1) & \text{otherwise} \end{cases} \\ &X_4 \sim \begin{cases} \gamma(4,100) & \text{if } 1 - X_1 - X_2 \leq 0.75 \\ \gamma(4, 500 - 4\,X_1) & \text{otherwise.} \end{cases} \end{aligned}$$

Step 2. $N = 1000$ replications of *Step 1* in the first study have been considered and the 100000 fuzzy values from the second study have been divided randomly (and without replacement) into 1000 samples of size $n = 100$. So in both studies, there are 1000 available samples of size $n = 100$.

Step 3. The population summary measures have been approximated on the basis of 35.000 replications.

Step 4. The estimates have been complemented with the average distance-based bias along the 1000 samples, and some other associated mean errors.

Distances have been computed by considering three different metrics: the L^2 metric ρ_2, the L^1 metric ρ_1 (see Diamond and Kloeden [5]) and the L^1 metric $\mathscr{D}_1$ (a particular case of that introduced by Sinova *et al.* [11]), where for fuzzy numbers $\widetilde{U}, \widetilde{V}$ they are given by

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

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

$$\mathscr{D}_1(\widetilde{U}, \widetilde{V}) = |\text{wabl}(\widetilde{U}) - \text{wabl}(\widetilde{U})| + \frac{1}{2}\int_{[0,1]} \left[|\text{ldev}\, \widetilde{U}_\alpha - \text{ldev}\, \widetilde{V}_\alpha| + |\text{rdev}\, \widetilde{U}_\alpha - \text{rdev}\, \widetilde{V}_\alpha|\right] d\alpha,$$

with $\text{wabl}(\widetilde{U}) = \int_{[0,1]} (\inf \widetilde{U}_\alpha + \sup \widetilde{U}_\alpha)\, d\alpha/2$, $\text{ldev}\, \widetilde{U}_\alpha = \text{wabl}(\widetilde{U}) - \inf \widetilde{U}_\alpha$, $\text{rdev}\, \widetilde{U}_\alpha = \sup \widetilde{U}_\alpha - \text{wabl}(\widetilde{U})$.

The outputs for this first simulation study have been collected in Table 1 for the mean errors in estimating the summary measures and in Figure 2 and Table 2 for their estimates.

Table 1. Mean errors in the estimation of some summary measures with the first simulations (CASES 1 and 2) for the six different types of fuzzy numbers in Figure 1

CASE 1													
		ρ_2-Mean			ρ_1-Median			$\mathscr{D}_1$-Median			ρ_2-Variance		
Type	Error	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$
T1	Bias	0.004	0.004	0.005	0.008	0.008	0.009	0.007	0.008	0.011	0.028	0.028	0.028
	Variance	0.028	0.035	0.055	0.027	0.036	0.056	0.029	0.038	0.058	0.583	0.583	0.583
	MSE	0.028	0.035	0.055	0.028	0.036	0.056	0.029	0.038	0.058	0.584	0.584	0.584
T2	Bias	0.004	0.004	0.005	0.008	0.008	0.009	0.007	0.008	0.010	0.028	0.028	0.028
	Variance	0.028	0.034	0.054	0.027	0.035	0.055	0.029	0.037	0.058	0.576	0.576	0.576
	MSE	0.028	0.034	0.054	0.027	0.035	0.056	0.029	0.037	0.058	0.576	0.576	0.576
T3	Bias	0.005	0.005	0.005	0.008	0.008	0.008	0.008	0.008	0.011	0.022	0.022	0.022
	Variance	0.028	0.035	0.055	0.027	0.036	0.056	0.029	0.038	0.059	0.583	0.583	0.583
	MSE	0.028	0.035	0.055	0.027	0.036	0.056	0.030	0.038	0.059	0.583	0.583	0.583
T4	Bias	0.004	0.004	0.004	0.007	0.007	0.009	0.008	0.008	0.008	0.030	0.030	0.030
	Variance	0.027	0.033	0.051	0.026	0.034	0.053	0.026	0.034	0.051	0.562	0.562	0.562
	MSE	0.027	0.033	0.051	0.026	0.034	0.053	0.026	0.034	0.051	0.563	0.563	0.563
T5	Bias	0.004	0.004	0.005	0.008	0.008	0.009	0.008	0.008	0.010	0.027	0.027	0.027
	Variance	0.028	0.035	0.054	0.027	0.035	0.055	0.029	0.037	0.058	0.574	0.574	0.574
	MSE	0.028	0.035	0.054	0.027	0.035	0.055	0.029	0.037	0.058	0.575	0.575	0.575
T6	Bias	0.004	0.004	0.004	0.007	0.008	0.009	0.008	0.008	0.009	0.029	0.029	0.029
	Variance	0.027	0.033	0.051	0.026	0.034	0.052	0.026	0.033	0.051	0.558	0.558	0.558
	MSE	0.027	0.033	0.051	0.026	0.034	0.052	0.026	0.034	0.051	0.559	0.559	0.559
CASE 2													
		ρ_2-Mean			ρ_1-Median			$\mathscr{D}_1$-Median			ρ_2-Variance		
Type	Error	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$
T1	Bias	0.003	0.003	0.003	0.002	0.002	0.002	0.005	0.005	0.006	0.004	0.004	0.004
	Variance	0.011	0.013	0.020	0.005	0.007	0.012	0.021	0.026	0.041	0.023	0.023	0.023
	MSE	0.011	0.013	0.020	0.005	0.007	0.012	0.021	0.026	0.041	0.023	0.023	0.023
T2	Bias	0.003	0.003	0.003	0.002	0.002	0.002	0.005	0.005	0.006	0.004	0.004	0.004
	Variance	0.011	0.013	0.020	0.005	0.006	0.011	0.021	0.026	0.041	0.023	0.023	0.023
	MSE	0.011	0.013	0.020	0.005	0.006	0.011	0.021	0.026	0.041	0.023	0.023	0.023
T3	Bias	0.003	0.003	0.003	0.001	0.002	0.002	0.005	0.005	0.006	0.004	0.004	0.004
	Variance	0.011	0.013	0.019	0.006	0.008	0.014	0.021	0.025	0.040	0.023	0.023	0.023
	MSE	0.011	0.013	0.019	0.006	0.008	0.014	0.021	0.025	0.040	0.023	0.023	0.023
T4	Bias	0.003	0.003	0.003	0.002	0.002	0.003	0.004	0.004	0.005	0.004	0.004	0.004
	Variance	0.011	0.013	0.019	0.007	0.009	0.015	0.020	0.024	0.038	0.023	0.023	0.023
	MSE	0.011	0.013	0.019	0.007	0.009	0.015	0.020	0.024	0.038	0.023	0.023	0.023
T5	Bias	0.003	0.003	0.003	0.002	0.002	0.002	0.004	0.005	0.006	0.004	0.004	0.004
	Variance	0.011	0.013	0.019	0.005	0.007	0.012	0.021	0.025	0.040	0.023	0.023	0.023
	MSE	0.011	0.013	0.019	0.005	0.007	0.012	0.021	0.025	0.040	0.023	0.023	0.023
T6	Bias	0.003	0.003	0.003	0.002	0.002	0.003	0.004	0.004	0.005	0.004	0.004	0.004
	Variance	0.011	0.012	0.018	0.007	0.009	0.016	0.020	0.024	0.037	0.022	0.022	0.022
	MSE	0.011	0.012	0.018	0.007	0.009	0.016	0.020	0.024	0.037	0.023	0.023	0.023

On the basis of the outputs in Table 1 one can empirically conclude to some extent that the shape of the considered data scarcely affects the bias, variance and mean squared error of the summary measures estimates.

Table 2. Monte Carlo estimate of the Fréchet ρ_2-variance in the first simulations

Variance	T1	T2	T3	T4	T5	T6
CASE 1	3.629	3.547	3.587	3.402	3.565	3.402
CASE 2	1.268	1.262	1.254	1.223	1.258	1.219

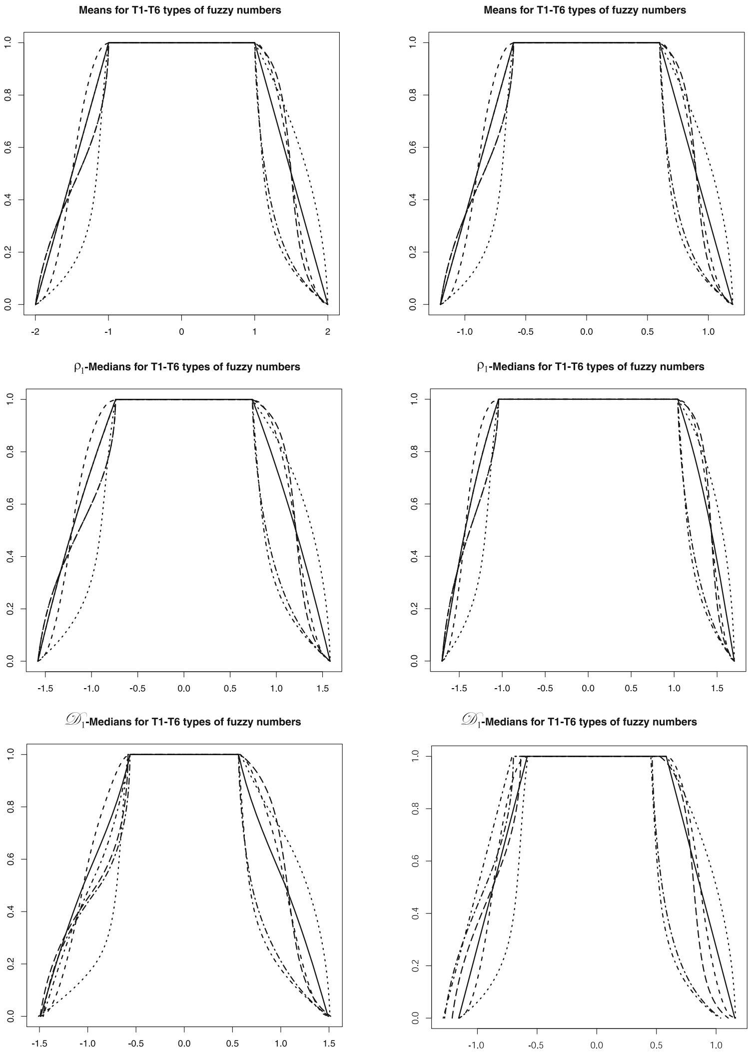

Fig. 2. Monte Carlo estimates of the (Aumann type) means and ρ_1- and $\mathscr{D}_1$-medians in CASE 1 (on the left) and CASE 2 (on the right) of the first simulations

The same happens for the estimates of the ρ_2-Fréchet variance in Table 2. The estimates of the location measures, graphically displayed in Figure 2, are more influenced by the shape of the involved fuzzy data. Nevertheless, the location estimates are indeed closer than the original data.

The outputs for the second simulation study have been collected in Table 3 for the mean errors in estimating the summary measures and in Figure 3 and Table 4 for their estimates. On the basis of the outputs in Table 3 one can empirically conclude to some extent that the shape of the considered data does not strongly affect the bias, variance and mean squared error of the summary measures estimates.

Table 3. Mean errors in the estimation of the summary measures with the second simulations for the six different types of fuzzy numbers in Figure 1

		ρ_2-Mean			ρ_1-Median			$\mathscr{D}_1$-Median			ρ_2-Variance		
Type	Error	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$	ρ_1	ρ_2	$\mathscr{D}_1$
T1	Bias	0.002	0.002	0.003	0.009	0.009	0.012	0.005	0.005	0.005	0.016	0.016	0.016
	Variance	0.083	0.086	0.106	0.220	0.238	0.323	0.213	0.217	0.252	0.523	0.523	0.523
	MSE	0.083	0.086	0.106	0.220	0.238	0.323	0.213	0.217	0.252	0.524	0.524	0.524
T2	Bias	0.002	0.002	0.003	0.008	0.009	0.012	0.005	0.005	0.005	0.016	0.016	0.016
	Variance	0.083	0.085	0.105	0.220	0.237	0.322	0.213	0.217	0.253	0.524	0.524	0.524
	MSE	0.083	0.085	0.105	0.221	0.237	0.322	0.213	0.217	0.253	0.525	0.525	0.525
T3	Bias	0.002	0.002	0.002	0.010	0.011	0.013	0.005	0.005	0.006	0.018	0.018	0.018
	Variance	0.083	0.086	0.105	0.219	0.236	0.317	0.211	0.214	0.248	0.533	0.533	0.533
	MSE	0.083	0.086	0.105	0.219	0.236	0.318	0.211	0.214	0.248	0.534	0.534	0.534
T4	Bias	0.002	0.002	0.003	0.007	0.008	0.009	0.002	0.002	0.003	0.016	0.016	0.016
	Variance	0.084	0.086	0.105	0.218	0.234	0.316	0.219	0.221	0.254	0.537	0.537	0.537
	MSE	0.084	0.086	0.105	0.218	0.235	0.316	0.219	0.221	0.254	0.537	0.537	0.537
T5	Bias	0.002	0.002	0.003	0.009	0.009	0.012	0.004	0.004	0.004	0.017	0.017	0.017
	Variance	0.083	0.086	0.106	0.219	0.236	0.320	0.214	0.217	0.252	0.527	0.527	0.527
	MSE	0.083	0.086	0.106	0.219	0.237	0.320	0.214	0.217	0.252	0.527	0.527	0.527
T6	Bias	0.002	0.002	0.003	0.008	0.008	0.009	0.001	0.001	0.002	0.016	0.016	0.016
	Variance	0.084	0.086	0.105	0.217	0.234	0.315	0.218	0.221	0.253	0.539	0.539	0.539
	MSE	0.084	0.086	0.105	0.217	0.234	0.315	0.218	0.221	0.253	0.539	0.539	0.539

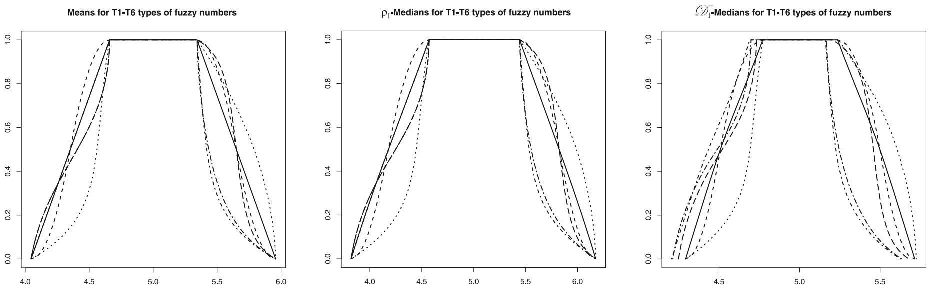

Fig. 3. Approximated estimates of the (Aumann type) means and ρ_1- and $\mathscr{D}_1$-medians for the second simulations

Table 4. Approximated estimate of the Fréchet ρ_2-variance in the second simulations

	T1	T2	T3	T4	T5	T6
Variance	7.921	7.902	7.950	7.971	7.926	7.983

The same happens for the estimates of the ρ_2-Fréchet variance in Table 4, although the shape difference influences slightly more than for the first study. The estimates of the location measures, graphically displayed in Figure 3, are more influenced by the shape of the involved fuzzy data, also slightly more than for the first simulations. Again, the location estimates are indeed closer than the original data.

As a clear extension of the study in this paper, it is a must to develop comparison concerning the influence on the power of hypothesis testing involving fuzzy data.

Acknowledgments. The research in this paper has been partially supported by the Severo Ochoa Grant BP12012 from the Principality of Asturias (de la Rosa de Sáa) and the FPU Grant AP2009-1197 from the Spanish Ministry of Education (Sinova). The support is gratefully acknowledged.

References

1. Ban, A., Coroianu, L., Grzegorzewski, P.: Trapezoidal approximation and aggregation. Fuzzy Sets Syst. 177, 45–59 (2011)
2. Blanco-Fernández, A., Casals, M.R., Colubi, A., Corral, N., García-Bárzana, M., Gil, M.A., González-Rodríguez, G., López, M.T., Lubiano, M.A., Montenegro, M., Ramos-Guajardo, A.B., De la Rosa de Sáa, S., Sinova, B.: A distance-based statistical analysis of fuzzy number-valued data. Int. J. Approx. Reas (2014), doi:10.1016/j.ijar.2013.09.020
3. Cox, E.: The fuzzy Systems Handbook. Academic Press, Cambridge (1994)
4. De la Rosa de Sáa, S., Gil, M.A., González-Rodríguez, G., López, M.T., Lubiano, M.A.: Fuzzy rating scale-based questionnaires and their statistical analysis. IEEE Trans. Fuzzy Syst (2014), doi:10.1109/TFUZZ.2014.2307895
5. Diamond, P., Kloeden, P.: Metric spaces of fuzzy sets. Fuzzy Sets Syst. 35, 241–249 (1990)
6. Grzegorzewski, P.: Trapezoidal approximations of fuzzy numbers preserving the expected interval - algorithms and properties. Fuzzy Sets Syst. 159, 1354–1364 (2008)
7. Grzegorzewski, P.: Fuzzy number approximation via shadowed sets. Inform. Sci. 225, 35–46 (2013)
8. Grzegorzewski, P., Pasternak-Winiarska, K.: Trapezoidal approximations of fuzzy numbers with restrictions on the support and core. In: Proc. 7th Conf. EUSFLAT-2011 and LFA-2011, pp. 749–756. Atlantis Press, Paris (2011)
9. Pedrycz, W.: Why triangular membership functions? Fuzzy Sets Syst. 64(1), 21–30 (1994)

10. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422 (1986)
11. Sinova, B., De la Rosa de Sáa, S., Gil, M.A.: A generalized L^1-type metric between fuzzy numbers for an approach to central tendency of fuzzy data. Inform. Sci. 242, 22–34 (2013)
12. Sinova, B., Gil, M.A., Colubi, A., Van Aelst, S.: The median of a random fuzzy number. The 1-norm distance approach. Fuzzy Sets Syst. 200, 99–115 (2012)
13. Stefanini, L., Sorini, L., Guerra, M.L.: Parametric representation of fuzzy numbers and applications to fuzzy calulus. Fuzzy Sets Syst. 157, 2423–2455 (2006)