Hypothesis testing for means in connection with fuzzy rating scale-based data: Algorithms and applications

Article *in* European Journal of Operational Research · November 2015

DOI: 10.1016/j.ejor.2015.11.016

María Asunción Lubiano
University of Oviedo

Beatriz Sinova
University of Oviedo

Sara de la Rosa de Sáa
University of Oviedo

Maria Angeles Gil
University of Oviedo

Hypothesis testing for means in connection with fuzzy rating scale-based data: algorithms and applications

María Asunción Lubiano[a], Manuel Montenegro[a], Beatriz Sinova[a], Sara de la Rosa de Sáa[a,b], María Ángeles Gil[a1]

[a] *Departamento de Estadística e I.O. y D.M., Universidad de Oviedo,*
C/ Calvo Sotelo, s/n, 33007 Oviedo, Spain

[b] *Institut für Stochastik und Wirtschaftsmathematik, Technische Universität Wien,*
Wiedner Hauptstraße 8-10/E105, 1040 Wien, Austria

Abstract

The fuzzy rating scale was introduced as a tool to measure intrinsically ill-defined/ imprecisely-valued attributes in a free way. Thus, users do not have to choose a value from a class of prefixed ones (like it happens when a fuzzy semantic representation of a linguistic term set is considered), but just to draw the fuzzy number that better represents their valuation or measurement. The freedom inherent to the fuzzy rating scale process allows users to collect data with a high level of richness, accuracy, expressiveness, diversity and subjectivity, what is especially valuable for statistical purposes.

This paper presents an inferential approach to analyze data obtained by using the fuzzy rating scale. More concretely, the paper is to be focussed on testing different hypothesis about means, on the basis of a sound methodology which has been stated during the last years. All the procedures that have been developed to this aim will be presented in an algorithmic way adapted to the usual generic fuzzy rating scale-based data, and they will be illustrated by means of a real-life example.

Keywords: computational intelligence and information management, fuzzy sets, fuzzy rating scale, random fuzzy numbers, stochastics and statistics

1. Introduction

In connection with the statistical analysis of fuzzy-valued data, several methods have been suggested to test hypotheses about the population means of the random

[1] Corresponding author. Tel.:+34985103356, Fax:+34985103354, E-mail address: magil@uniovi.es

processes generating such data. More concretely, when these random processes are modeled by means of random fuzzy sets (or fuzzy random variables in Puri & Ralescu's sense, 1986) one can find the following testing methods for the Aumann-type fuzzy means:

- *one-sample* ones, to test the ('two-sided') null hypothesis that the population Aumann-type fuzzy mean of a random fuzzy set equals a given fuzzy value (see, for instance, Körner, 2000; Montenegro *et al.*, 2004; González-Rodríguez *et al.*, 2006b);
- *two-sample* ones, to test the ('two-sided') null hypothesis of the equality of the population Aumann-type fuzzy means of two random fuzzy sets being either *independent* (see Montenegro *et al.*, 2001) or *dependent* (see González-Rodríguez *et al.*, 2006a);
- *k-sample/ANOVA* ones, to test the ('two-sided') null hypothesis of the equality of the population Aumann-type fuzzy means of k random fuzzy sets which are either *independent* (see Gil *et al.*, 2006; González-Rodríguez *et al.*, 2012) or *dependent* (see Montenegro *et al.*, 2009).

For detailed reviews on the problem, one can see the paper by Colubi (2009), as well as the recent one by Blanco-Fernández *et al.* (2014a,b).

The above-mentioned testing methods have been developed to deal with general fuzzy-valued data, but most of the practical situations they apply to concern fuzzy number-valued data. In this respect, an important source of fuzzy number-valued data in real-life is associated with using (either consciously or not) the fuzzy rating scale to assess valuations/ratings to intrinsically ill-defined attributes like quality/satisfaction/perception/attitude/judgement...

To show the potentiality of these testing methods in analyzing data coming from the use of the fuzzy rating scale, this paper presents each of the methods in an algorithmic way, and illustrates their application on a real-life example involving several fuzzy datasets.

Although algorithms for different tests are to be first presented for general fuzzy values (irrespective of their shape and dimensionality), their steps will also be particularized to trapezoidal fuzzy data, as those we obtain when the usual fuzzy rating scale is considered. In this way, and to a certain extent, this paper aims to be a kind of 'instruction manual' for the use of testing techniques about means in dealing with (trapezoidal) fuzzy rating scale-based data.

For this purpose, Section 2 recalls the ideas and mechanism for the fuzzy rating scale and its application to valuate items in a questionnaire, along with some related tools and models. Section 3 presents the real-life example and how the fuzzy rating scale has been applied to get the fuzzy datasets to be analyzed later. Section 4 presents different tests for fuzzy means (one-sample, independent/dependent two-sample and independent/dependent ANOVA) by means of algorithms, including their particularization to trapezoidal fuzzy data; once each algorithm is described, it is applied to some of the datasets from the real-life example. In Section 5, as a matter of comparative analysis, some of the statistical conclusions are to be compared with those drawn from the use (for the same items) of a Likert scale or a numerical/fuzzy linguistic encoding. The paper ends with some concluding remarks about related open problems.

2. Preliminaries

By a fuzzy-valued datum we mean a fuzzy set on a finite-dimensional Euclidean space. More concretely, we will consider fuzzy vectors or fuzzy numbers having compact support set, that is,

Definition 2.1. *A (bounded)* ***fuzzy value*** *is a mapping* $\widetilde{U} : \mathbb{R}^p \to [0,1]$ *(with* $p \in \mathbb{N}$*) such that for each* $\alpha \in [0,1]$ *its* α*-level set*

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

(with cl = *topological closure) is a nonempty closed, convex and bounded set in* $\mathbb{R}^p$*.*

When $p = 1$*, the fuzzy value is referred to as a* ***fuzzy number****.*

A well-known key feature in applying fuzzy sets to model data is that for each $x \in \mathbb{R}^p$ the number $\widetilde{U}(x)$ is usually interpreted as the 'degree of compatibility' of x with the concept represented by (or the property describing) $\widetilde{U}$.

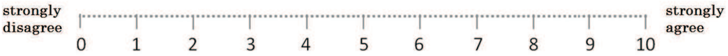
Fuzzy values are especially suitable to cope with the imprecision of human thought and experience in attributes such as quality, satisfaction, perception, attitude, and so on. The flexibility and expressiveness of fuzzy sets allow us to properly model and describe values from such attributes.

The ***fuzzy rating scale***, introduced by Hesketh *et al.* (1988), arises from interacting the abilities of fuzzy sets to formalize mathematically valuations which are intrinsically imprecise, to precisiate such valuations through a continuum, and to develop mathematical computations with the formalized valuations. On the basis of this scale, users establish their valuations (often responses to the items of a questionnaire) by drawing fuzzy numbers (usually a trapezoidal or, in particular, a triangular one) in a completely free way.

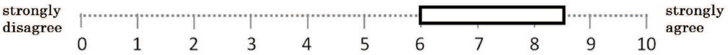
When open-ended valuations/responses are formalized in terms of arbitrary fuzzy numbers, they show a clear meaning. Nevertheless, they do not need to (and, actually, should not) be translated into words.

The guideline for drawing the fuzzy number that better expresses the valuation/response according to the fuzzy rating scale can be schematized as follows (see Hesketh *et al.*, 1988; De la Rosa de Sáa *et al.*, 2015):

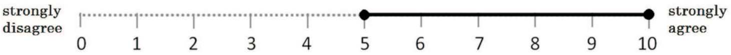
Step 1. A reference bounded interval/segment is first considered ($[0, 10]$ or $[0, 100]$ being the most frequent choices). The end-points are often labeled in accordance with their meaning referring to the degree of agreement, satisfaction, quality, and so on.



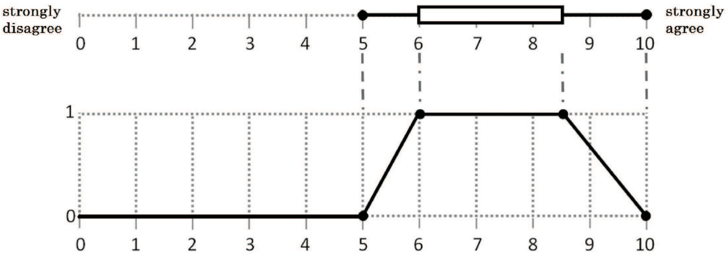
Step 2. The *core*, or 1-level set, associated with the response is determined. It corresponds to the interval (or singleton) of the real values (value) within the reference one which are (is) considered to be 'fully compatible' with the valuation/response.



Step 3. The *support* (or its closure or 0-level set), associated with the valuation/ response is determined. It corresponds to the interval of the real values within the reference one which are considered to be 'compatible to some extent' with the response.



Step 4. The two intervals are 'linearly interpolated' to get a trapezoidal fuzzy number



The freedom inherent to fuzzy rating scale-based data entails a gain of information and accuracy, which is certainly crucial for statistical purposes. In this regard, Hesketh *et al.* (2011) have indicated the need for statistical techniques to analyze fuzzy data. Although a few studies have been carried out to analyze fuzzy rating scale-based data (e.g., Hesketh *et al.*, 1988; Hesketh & Hesketh, 1994; Hesketh *et al.*, 1995; Takemura, 1999, 2007), these studies have been descriptive ones based on either on defuzzification processes or the end-points of the core 0- and 1-level sets.

To develop statistics with fuzzy rating scale-based data, two operations with them are frequently considered. The usual fuzzy arithmetic with fuzzy values is the one based on Zadeh's extension principle (Zadeh, 1975), which in case of dealing with fuzzy numbers coincides level-wise with the usual interval arithmetic.

Let $\mathcal{F}_c(\mathbb{R})$ denote the space of (bounded) fuzzy numbers. Then,

Definition 2.2. *If $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c(\mathbb{R})$, then the* ***sum of*** $\widetilde{U}$ ***and*** $\widetilde{V}$ *is defined as the fuzzy number $\widetilde{U} + \widetilde{V} \in \mathcal{F}_c(\mathbb{R})$ such that for each $\alpha \in [0,1]$*

$$(\widetilde{U} + \widetilde{V})_\alpha = \left[\inf \widetilde{U}_\alpha + \inf \widetilde{V}_\alpha, \sup \widetilde{U}_\alpha + \sup \widetilde{V}_\alpha\right].$$

If $\widetilde{U} \in \mathcal{F}_c(\mathbb{R})$ and $\gamma \in \mathbb{R}$, the ***product of*** $\widetilde{U}$ ***by the scalar*** γ *is defined as the fuzzy number $\gamma \cdot \widetilde{U} \in \mathcal{F}_c(\mathbb{R})$ such that for each $\alpha \in [0,1]$*

$$(\gamma \cdot \widetilde{U})_\alpha = \gamma \cdot \widetilde{U}_\alpha = \left\{\gamma \cdot y \,:\, y \in \widetilde{U}_\alpha\right\} = \begin{cases} \left[\gamma \inf \widetilde{U}_\alpha, \gamma \sup \widetilde{U}_\alpha\right] & \textit{if } \gamma \geq 0, \\ \left[\gamma \sup \widetilde{U}_\alpha, \gamma \inf \widetilde{U}_\alpha\right] & \textit{if } \gamma < 0. \end{cases}$$

It is well-known that, except when $\widetilde{U}$ reduces to the indicator function of a singleton, $\widetilde{U} + (-1) \cdot \widetilde{U} \neq \mathbb{1}_{\{0\}}$ (indicator of the singleton $\{0\}$, which is the neutral element for the fuzzy sum). Consequently, it is not possible to establish a difference operator between fuzzy numbers that is always well-defined and preserves all the properties of the difference of real numbers. Many of the drawbacks associated with this non-linearity can be often overcome in statistical developments by using suitable metrics, like the parameterized L^2 one by Bertoluzza *et al.* (1995) (see Trutschnig *et al.*, 2009, for a deep and more general study about).

Definition 2.3. *Let $\theta > 0$ and assume $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c(\mathbb{R})$. The* ***mid/spr distance*** *(or* ***Bertoluzza et al.'s distance****) between $\widetilde{U}$ and $\widetilde{V}$ is given by*

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

with $\operatorname{mid} \widetilde{U}_\alpha = (\inf \widetilde{U}_\alpha + \sup \widetilde{U}_\alpha)/2$ *denoting the centre/mid-point of the corresponding level interval $\widetilde{U}_\alpha$ and* $\operatorname{spr} \widetilde{U}_\alpha = (\sup \widetilde{U}_\alpha - \inf \widetilde{U}_\alpha)/2$ *denoting the spread/radius of the level interval $\widetilde{U}_\alpha$.*

When two trapezoidal fuzzy numbers $\widetilde{U} = \mathrm{Tra}(a,b,c,d)$ and $\widetilde{U}' = \mathrm{Tra}(a',b',c',d')$ are considered, then $D_\theta(\widetilde{U}, \widetilde{U}')$ can be expressed in terms of the midpoints and radii of the support and core sets. More concretely, $D_\theta(\widetilde{U}, \widetilde{U}')$ equals

$$\sqrt{\frac{(\underline{m} - \underline{m}')^2 + (\overline{m} - \overline{m}')^2 + (\underline{m} - \underline{m}')(\overline{m} - \overline{m}') + \theta(\underline{s} - \underline{s}')^2 + \theta(\overline{s} - \overline{s}')^2 + \theta(\underline{s} - \underline{s}')(\overline{s} - \overline{s}')}{3}},$$

where $\underline{m} = (a+d)/2$, $\overline{m} = (b+c)/2$, $\underline{s} = (d-a)/2$ and $\overline{s} = (b-c)/2$.

The weighting parameter θ is frequently chosen to be $\theta = 1$ or $\theta = 1/3$, leading to the particular metrics

$$D_1(\widetilde{U}, \widetilde{V}) = \rho_2(\widetilde{U}, \widetilde{V}) = \sqrt{\frac{1}{2} \int_{[0,1]} \left(\left[\inf \widetilde{U}_\alpha - \inf \widetilde{V}_\alpha\right]^2 + \left[\sup \widetilde{U}_\alpha - \sup \widetilde{V}_\alpha\right]^2\right) d\alpha},$$

introduced by Diamond & Kloeden (1990), and

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

with $\widetilde{U}_{\alpha}^{[\lambda]} = \lambda \sup \widetilde{U}_{\alpha} + (1-\lambda) \inf \widetilde{U}_{\alpha}$, that is, weighing uniformly the squared distances between the convex linear combinations of the end-points of the different level intervals.

To develop a well-stated methodology to analyze fuzzy data, and especially when inferential targets are involved, we need a formal model for the random mechanism generating fuzzy number-valued data. This model should integrate randomness (to generate data) and fuzziness (because of the intrinsic nature of these data).

Random fuzzy sets (originally coined as fuzzy random variables by Puri & Ralescu, 1986) result in a well-defined and sound model within the probabilistic setting, which allows us to extend or preserve almost all the fundamentals from statistical inference. In this way, it should be highlighted that, although extending statistical methods to the analysis of fuzzy data is not at all a simple task because of the problems that will be described at the beginning of Section 4, with the use of random fuzzy sets one can properly and immediately refer to their (induced) distribution, parameters, independence, and so on. Furthermore, one can make use of the key notions in inferential statistics like the p-value of a test, its consistency, etc., without needing to introduce a special statistical setting to deal with fuzzy data.

Consider a random experiment which is mathematically modeled by means of a probability space $(\Omega, \mathcal{A}, P)$.

Definition 2.4. *A* ***random fuzzy number*** *(or one-dimensional random fuzzy set, for short* ***RFN****) associated with* $(\Omega, \mathcal{A}, P)$ *is a mapping* $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$ *such that for all* $\alpha \in [0,1]$ *the* α*-level mapping* $\mathcal{X}_\alpha$ *is a compact random interval (that is, for all* $\alpha \in [0,1]$ *the real-valued mappings* $\inf \mathcal{X}_\alpha$ *and* $\sup \mathcal{X}_\alpha$ *are random variables).*

Remark 2.1. Since a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$ is an RFN if and only if it is a Borel-measurable mapping with respect to a certain σ-field associated with D_θ (see, for instance, Colubi *et al.*, 2001), the Borel-measurability allows us to trivially induce from P the *distribution of the RFN* $\mathcal{X}$, and refer to the *independence of RFNs*.

In performing inferential analysis about the distribution of RFNs, the best known involved parameters are the Aumann-type mean value by Puri & Ralescu (1986) and the Fréchet-type variance by Lubiano *et al.* (2000).

Definition 2.5. *Let $(\Omega, \mathcal{A}, P)$ be the probability space modeling a random experiment and $\mathcal{X}$ be an associated RFN.*

*The **Aumann-type mean of $\mathcal{X}$** is the fuzzy number $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R})$, if it exists, such that for all $\alpha \in [0,1]$*

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

with E denoting the mean value of the corresponding real-valued random variable.

*The D_θ **Fréchet-type variance of $\mathcal{X}$** is the real number $\sigma^2_{\mathcal{X}}$, if it exists, given by*

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

3. Real-life example

In this section a real-life example is presented. This example will be considered in Section 4 to illustrate the potentiality of the approach to hypotheses testing for means.

For many years, questionnaires have been important tools for assessing attitudes regarding a lot of educational issues (see, as instances of some recent studies about, Rutkowski & Rutkowski, 2010; Haelermans & De Witte, 2012; Rutkowski *et al.*, 2013; Nikolaidis & Dimitriadis, 2014; Ballouard *et al.*, 2015; Peró *et al.*, 2015).

The example concerns the well-known *questionnaire TIMSS-PIRLS 2011*. The International Association for the Evaluation of the Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) have provided information about home supports and school environments for teaching and learning. For this purpose, questionnaires on *reading*, *mathematics* and *science* have been conducted on the same students, and also additional questionnaires having been filled by their parents, teachers and school management team. Having data on the same students makes it possible to perform valuable investigations and researchers can apply a variety of modeling techniques to explore these important issues.

In 2011, the Spanish Institute of Educational Evaluation (INEE) has commissioned some members of our Department of Statistics, Operations Research and DM at the University of Oviedo in Spain to develop an analysis with data collected through some of the TIMSS/PIRLS questionnaires conducted in Spanish schools (see Corral-Blanco *et al.*, 2013, for a summary of conclusions). These questionnaires have been standard ones in what concern responses, since most of the involved items had to be answered according to the 4-point Likert scale given by A1 = DISAGREE A LOT, A2 = DISAGREE A LITTLE, A3 = AGREE A LITTLE and A4 = AGREE A LOT.

Conclusions from these studies have been certainly interesting. Anyway, our colleagues have been wondering whether adapting these questions with the fuzzy rating scale approach would yield somewhat different statistical conclusions. To corroborate such an intuitive idea, an introductory inferential analysis has been performed. For this purpose nine items have been chosen from the Student questionnaires, the items being displayed in Table 1.

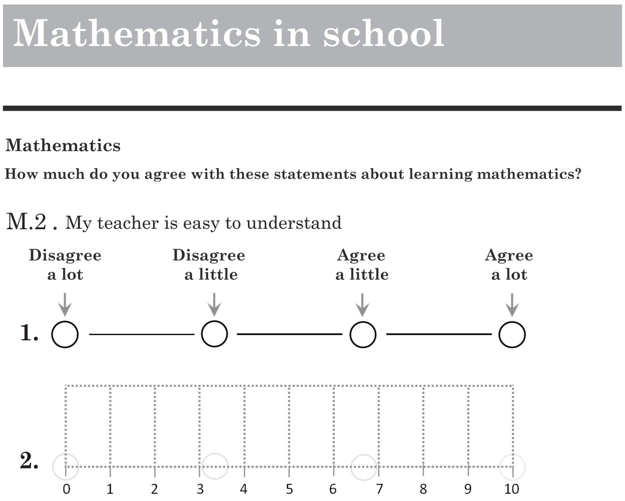
	READING IN SCHOOL
$R.1$	I like to read things that make me think
$R.2$	I learn a lot from reading
$R.3$	Reading is harder for me than any other subject
	MATHEMATICS IN SCHOOL
$M.1$	I like mathematics
$M.2$	My teacher is easy to understand
$M.3$	Mathematics is harder for me than any other subject
	SCIENCE 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

Table 1: Questions selected from the TIMSS-PIRLS 2011 Student Questionnaires

These nine items have been originally conceived to be answered in accordance with the four-point Likert scale $\{\mathrm{A1}, \mathrm{A2}, \mathrm{A3}, \mathrm{A4}\}$. Of course, an immediate way to incorporate Fuzzy Logic in analyzing these items is to handle them by means of a *fuzzy linguistic scale* (see, for instance, Zadeh, 1975; Tong & Bonissone, 1980; Pedrycz, 1989; Herrera *et al.*, 1998, 2008; Lalla *et al.*, 2008; Roszkowska & Wachowicz, 2015; Yan & Ma, 2015).

Nevertheless, from a statistical viewpoint the fuzzy rating scale adds a clear diversity and subjectivity which cannot be captured through a fuzzy linguistic one and definitely provides a much richer information. For this reason, the questionnaire form involving these nine items, along with a few more ones about students' support resources at home, has been adapted to allow a double-type response, namely, the original Likert and a fuzzy rating scale-based one (see Figure 1 for Question $M.2$, and see http://bellman.ciencias.uniovi.es/SMIRE/FuzzyRatingScaleQuestionnaire-SanIgnacio.html for the full paper-and-pencil and computerized versions).

In this way, each of the nine questions in Table 1 is assumed to be filled in accordance with both the 4-point Likert and the fuzzy rating scales with reference interval $[0, 10]$ (Spain uses a 10-point grading scale for elementary and high schools, so it could mean a proper simile for students). To ease the relationship between the two scales, four light circular marks have been placed indicating the usual numerical encoding of the four Likert categories with respect to the interval $[0, 10]$.



Mathematics in school

Mathematics

How much do you agree with these statements about learning mathematics?

M.2 . My teacher is easy to understand

Figure 1: Example of the double-response form to a question

The questionnaire has been conducted on 69 fourth grade students from Colegio San Ignacio (Oviedo-Asturias, Spain). Although a first analysis has been considered in Gil *et al.* (2015), this paper aims to enlarge, detail and complete it by paying attention to testing about means procedures and results. Some students have used the computer-administered format, whereas the others have filled the paper-and-pencil one. The training of the 9-year-old children has taken up to fifteen minutes, and was essentially based on the notion of trapezium they already knew, making them identify the upper base with the total compatibility degree and the lower one with the compatibility to some extent, and then legs joining both bases. In spite of the short training and, because of the students being nine-year-old, the rather naive mathematical background of the respondents, *there have been few more missing responses with the Likert than with the fuzzy rating scale.* This has been a first interesting conclusion: of course, *fuzzy rating scale-based questionnaires are not trivial to respond and there is a need for a certain training, but the fuzzy spirit seems not to be difficult to follow.*

The complete dataset for the study to be developed in Section 4 can be found in http://bellman.ciencias.uniovi.es/SMIRE/FuzzyRatingScaleQuestionnaire-SanIgnacio.html.

4. Testing hypotheses about means

When one attempts to develop statistics with fuzzy rating scale-based data, some distinctive features with respect to the real-valued case should be taken into account, namely,

- one cannot make use of a difference operator which is well-defined and preserves all the properties from the real-valued case;
- there is no ‘universally accepted’ total order between fuzzy numbers;

- there are not realistic and wide models for the distributions of random fuzzy numbers;
- there are not Central Limit Theorems for RFNs which can be directly applied for inferential purposes (thus, although CLTs for random fuzzy sets can be found in the literature - see, for instance, Wu, 1999, and Krätschmer, 2002 - either restrictive conditions for the values the random fuzzy sets take on have been assumed or the limit Gaussian random element cannot be always guaranteed to be identifiable with a random fuzzy set).

At this point, the involvement of suitable metrics like D_θ plays a crucial role in overcoming most of the associated drawbacks. This is mainly due to the fact that thanks to these metrics the space of fuzzy numbers endowed with the fuzzy arithmetic can be isometrically embedded *via* the support function (Puri & Ralescu, 1985) into a convex cone of a Hilbert space of functions endowed with the usual functional arithmetic and a certain metric (see González-Rodríguez *et al.*, 2012, for details about). This embedding entails a convenient identification of fuzzy numbers with functions in a Hilbert space, allowing us

- to apply the bootstrapped Central Limit Theorem for generalized spaces by Giné & Zinn (1990), when we deal with the means of random fuzzy sets; thus, while the Central Limit Theorem for generalized spaces could not be directly applied because of not having guarantees for the limit elements to take values within the convex cone the space of fuzzy numbers is embedded into, the use of bootstrap techniques along with the convexity of the cone circumvent such problems;
- to particularize several statistical methods from the so-called Functional Data Analysis; although at first glance all Functional Data Analysis methods could be particularized, special caution should be paid in being certain that the particularization does not lead us out of the cone the space of fuzzy numbers is embedded into.

Moreover, we can always try to develop *ad hoc* procedures (see, for instance, Montenegro *et al.*, 2001, 2004; Gil *et al.*, 2006; González-Rodríguez *et al.*, 2009).

Testing methods for (Aumann-type) means with fuzzy (vector or number-valued) data have been obtained by first developing asymptotic results and later approximating these results by means of the bootstrapped central limit theorem. In ANOVA's case, the most recent approach has consisted in developing a functional data analysis approach and later particularize it also thanks to the bootstrapped central limit theorem. The procedures obtained in this way make use of the same fundamentals that methods for real-valued data. Thus, most of the statistical conclusions are based on the computation of the p-value, which make full sense in this random fuzzy sets-based methodology. The *p-value* can be intuitively viewed as a kind of indicator of the sustainability of the null hypothesis (the smaller the p-value, the stronger evidence against the hypothesis). Furthermore, the obtained methods have been proved to be consistent, and they can be suitably applied even with small to moderate sample sizes.

As, already commented, the one-sample, two-sample and k-sample procedures are to be recalled in an algorithmic way, and each of the steps allowing it will be particularized to the usual fuzzy rating scale-based data, i.e., trapezoidal ones. Each of the procedures will be illustrated with the analysis of some of the items in the considered real-life example (in which θ has been chosen to equal $1/3$).

It should be remarked that, although the fuzzy rating scale has been stated on the basis of triangular/trapezoidal fuzzy numbers, this assumption is not essential for the hypotheses testing methods we describe and apply in this section. However, such a simple shape, which can be easily explained through the notions of triangle and trapezium, makes the training easier to understand almost irrespective of the background of respondents, as the considered real-life example corroborates. Furthermore, their use is also supported by arguments provided by Pedrycz (1994), Grzegorzewski (2008), Grzegorzewski & Pasternak-Winiarska (2011), Ban *et al.* (2011), and others, who have soundly discussed how triangular and trapezoidal fuzzy numbers can be considered to describe or approximate fuzzy data. Recently, an additional supporting argument has been given through an empirical sensitivity analysis which has been carried out about the effect of the shape of fuzzy data on several statistical measures (see Lubiano *et al.*, 2015).

To simplify notations along the following subsections, if $\widetilde{U}$ is a trapezoidal fuzzy number, the midpoint and radii of the support $\widetilde{U}_0$ and core $\widetilde{U}_1$ will be denoted by

$$\underline{m}(\widetilde{U}) = \operatorname{mid}\widetilde{U}_0,\ \overline{m}(\widetilde{U}) = \operatorname{mid}\widetilde{U}_1,\ \underline{s}(\widetilde{U}) = \operatorname{spr}\widetilde{U}_0,\ \overline{s}(\widetilde{U}) = \operatorname{spr}\widetilde{U}_1.$$

Furthermore, if $\widetilde{x}_i$ is a trapezoidal fuzzy datum we will denote:

$$\underline{m}_i = \underline{m}(\widetilde{x}_i),\ \overline{m}_i = \overline{m}(\widetilde{x}_i),\ \underline{s}_i = \underline{s}(\widetilde{x}_i),\ \overline{s}_i = \overline{s}(\widetilde{x}_i).$$

4.1. One-sample test for the mean with fuzzy rating scale-based data

To test the null hypothesis that the Aumann-type mean of an RFN $\mathcal{X}$ equals a given fuzzy number $\widetilde{U}$, one can consider the *bootstrapped algorithm* by Colubi (2009) approximating the *one-sample test for the mean of an RFN* (see Körner, 2000; Montenegro *et al.*, 2004; González-Rodríguez *et al.*, 2006b), which is now algorithmically summarized.

Let $\widetilde{\boldsymbol{x}}_{\boldsymbol{n}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ be a sample of independent observations from the RFN $\mathcal{X}$, and let $\widetilde{U} \in \mathcal{F}_c(\mathbb{R})$. Then, the algorithm to test the null hypothesis $H_0 : \widetilde{E}(\mathcal{X}) = \widetilde{U}$ (or, equivalently, $H_0 : D_\theta\big(\widetilde{E}(\mathcal{X}), \widetilde{U}\big) = 0$) proceeds as follows:

Step 1. Compute the value of the statistic

$$T_n = \frac{A_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}, \widetilde{U})}{B_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}})} = \frac{\left[D_\theta\big(\frac{1}{n}\cdot(\widetilde{x}_1+\ldots+\widetilde{x}_n), \widetilde{U}\big)\right]^2}{\frac{1}{n-1}\sum_{i=1}^n \left[D_\theta\big(\widetilde{x}_i, \frac{1}{n}\cdot(\widetilde{x}_1+\ldots+\widetilde{x}_n)\big)\right]^2}.$$

If all sample data $\widetilde{x}_i$ are trapezoidal, then the sample mean is also trapezoidal, and

$$\begin{aligned} A_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}, \widetilde{U}) = {} & \left[\frac{1}{n}\sum_{l=1}^n \underline{m}_l - \underline{m}(\widetilde{U})\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n \overline{m}_l - \overline{m}(\widetilde{U})\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n \underline{m}_l - \underline{m}(\widetilde{U})\right]\cdot\left[\frac{1}{n}\sum_{l=1}^n \overline{m}_l - \overline{m}(\widetilde{U})\right] \\ & + \theta\left(\left[\frac{1}{n}\sum_{l=1}^n \underline{s}_l - \underline{s}(\widetilde{U})\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n \overline{s}_l - \overline{s}(\widetilde{U})\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n \underline{s}_l - \underline{s}(\widetilde{U})\right]\cdot\left[\frac{1}{n}\sum_{l=1}^n \overline{s}_l - \overline{s}(\widetilde{U})\right]\right), \\ B_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}) = {} & \frac{1}{n-1}\sum_{i=1}^n\left(\left[\frac{1}{n}\sum_{l=1}^n(\underline{m}_l - \underline{m}_i)\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n(\overline{m}_l - \overline{m}_i)\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n(\underline{m}_l - \underline{m}_i)\right]\cdot\left[\frac{1}{n}\sum_{l=1}^n(\overline{m}_l - \overline{m}_i)\right]\right) \\ & + \frac{\theta}{n-1}\sum_{i=1}^n\left(\left[\frac{1}{n}\sum_{l=1}^n(\underline{s}_l - \underline{s}_i)\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n(\overline{s}_l - \overline{s}_i)\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n(\underline{s}_l - \underline{s}_i)\right]\cdot\left[\frac{1}{n}\sum_{l=1}^n(\overline{s}_l - \overline{s}_i)\right]\right). \end{aligned}$$

Step 2. Fix the bootstrap population to be $\big\{\widetilde{x}_1, \ldots, \widetilde{x}_n\big\}$.

Step 3. Obtain a sample of independent observations from the bootstrap population, $\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n}} = (\widetilde{x}^*_1, \ldots, \widetilde{x}^*_n)$.

Step 4. Compute the value of the bootstrap statistic

$$T^*_n = \frac{A_n\Big(\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n}}, \widetilde{\boldsymbol{x}}_{\boldsymbol{n}}\Big)}{B_n(\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n}})}.$$

If all sample data $\widetilde{x}_i$ are trapezoidal, then the bootstrap data are also trapezoidal and the sample means (original and bootstrap) are too, whence

$$\begin{aligned} A_n\Big(\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n}}, \widetilde{\boldsymbol{x}}_{\boldsymbol{n}}\Big) = {} & \left[\frac{1}{n}\sum_{l=1}^n(\underline{m}^*_l - \underline{m}_l)\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n(\overline{m}^*_l - \overline{m}_l)\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n(\underline{m}^*_l - \underline{m}_l)\right]\cdot\left[\frac{1}{n}\sum_{l=1}^n(\overline{m}^*_l - \overline{m}_l)\right] \\ & + \theta\left(\left[\frac{1}{n}\sum_{l=1}^n(\underline{s}^*_l - \underline{s}_l)\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n(\overline{s}^*_l - \overline{s}_l)\right]^2 + \left[\frac{1}{n}\sum_{l=1}^n(\underline{s}^*_l - \underline{s}_l)\right]\cdot\left[\frac{1}{n}\sum_{l=1}^n(\overline{s}^*_l - \overline{s}_l)\right]\right). \end{aligned}$$

Step 5. *Steps 3* and *4* should be repeated a large number B of times to get a set of estimates, denoted by $\{t^*_1, \ldots, t^*_B\}$.

Step 6. Compute the bootstrap p-value as the proportion of values in $\{t^*_1, \ldots,\, t^*_B\}$ which are greater than T_n.

To illustrate the application of this test we consider the following:

Example 4.1. Consider the fuzzy rating scale dataset associated with the responses to item $M.2$ ('My math teacher is easy to understand') in the real-life example in Section 3 for the $n = 66$ students who responded to this question using both scales.

Assume that we aim to test whether the (population) Aumann-type mean of the responses equals the usual fuzzy linguistic encoding of the two higher values in a 4-point Likert scale, namely, $\widetilde{U}3$ (encoding A3, see Table 2 and also Figure 2 later explained), $\widetilde{U}4$ (encoding A4, see Table 2 and Figure 2), or its average $.5 \cdot (\widetilde{U}3 + \widetilde{U}4)$ (see Table 2).

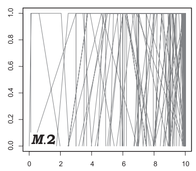
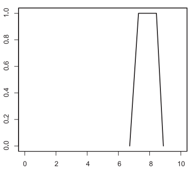
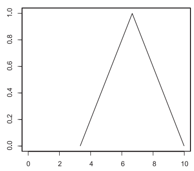
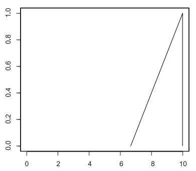
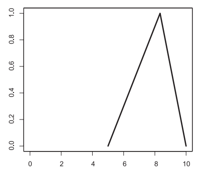
FRS sample data	FRS sample mean	hypothetical mean $\widetilde{U}3$	hypothetical mean $\widetilde{U}4$	hypothetical mean $.5 \cdot (\widetilde{U}3 + \widetilde{U}4)$
				
p-value		$\mathbf{.000}^{***}$	$\mathbf{.000}^{***}$	.097

Table 2: On the top, the sample fuzzy rating scale-based responses and sample mean (on the left), along with three hypothetical means (on the right) in connection with Question $M.2$; on the bottom, the corresponding p-values in testing the hypothetical means ($^{***}p < .001$)

The p-values obtained after applying the preceding algorithm (with $B = 1000$) indicate that the two fuzzy linguistic terms for A3 and A4 should be rejected at any significance level as the population mean response to $M.2$. Evidences against are not as categorical in case we consider as the hypothetical mean the average of A3 and A4 (in the usual fuzzy arithmetic sense), since the p-value equals .097.

Example 4.2. Consider, separately, the fuzzy rating scale datasets associated with the responses to item $M.2$ for the 43 students who have responded to the question by using the computerized form and the 23 students filling the paper-and-pencil one (see sample datasets and means on the top of Table 3).

First, we aim to test whether the (population) Aumann-type mean of the responses equals the original trapezoidal fuzzy number which appears in the computerized form as the default answer (i.e., a pattern to illustrate the type of valid answers and to ease the response: by only moving its four 0- and 1-levels end-points it is possible to get the wanted trapezoidal fuzzy answer). Such default answer is the trapezoidal fuzzy number $\mathrm{Tra}(2.5, 3.75, 6.25, 7.5)$. This test could indicate, in case of significant differences for the computer users, that the pattern can make their responses being close to it. In both cases, the p-values obtained from the preceding algorithm (with $B = 1000$) equal 0, so the pattern trapezoidal fuzzy number in the computerized version should be rejected at any significance level as the population mean response to $M.2$ for both groups of students.

On the other hand, if the null hypothesis considers as the population mean the trapezoidal fuzzy number Tra$(5, 6.5, 7.5, 9)$, then the hypothesis would be rejected at many of the usual significance levels for the students filling the paper-and-pencil form, whereas evidences against such a null hypothesis are not that strong for the students using the computerized version. Actually, the two-sample test in the next section will show that the version of the questionnaire influences to some extent the response to $M.2$.

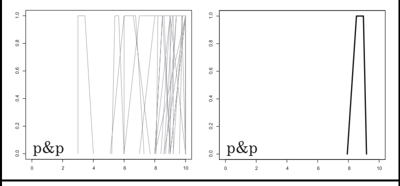
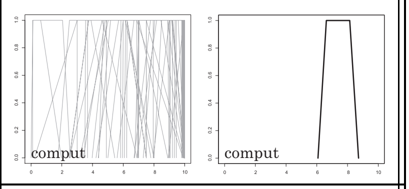
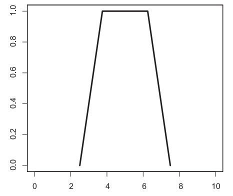
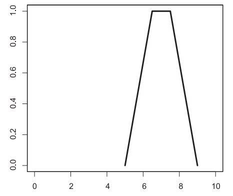
p-value	paper & pencil sample data and mean	computerized sample data and mean
		
hypothetical mean 	**.000*****	**.000*****
hypothetical mean 	**.010****	.267

Table 3: p-Values in testing the hypothetical means (on the left) for sample datasets for paper-and-pencil responses and computer responses (on the top) in connection with Question $M.2$ ($^{**}p < .01$, $^{***}p < .001$)

4.2. *Two-sample test for means with fuzzy rating scale-based data (independent samples)*

To test the null hypothesis of equality of the Aumann-type means of two RFNs $\mathcal{X}$ and $\mathcal{Y}$, one can consider the *bootstrapped algorithm* by Colubi (2009) approximating the *two-sample test for the means of two RFNs* for independent samples (see Montenegro *et al.*, 2001), which is now algorithmically summarized.

If $\mathcal{X}$ and $\mathcal{Y}$ are two independent RFNs, consider a sample of independent observations from $\mathcal{X}$, $\widetilde{\boldsymbol{x}}_{\boldsymbol{n_1}} = (\widetilde{x}_1, \ldots, \widetilde{x}_{n_1})$, and a sample of independent observations from $\mathcal{Y}$, $\widetilde{\boldsymbol{y}}_{\boldsymbol{n_2}} = (\widetilde{y}_1, \ldots, \widetilde{y}_{n_2})$.

Then, the algorithm to test the null hypothesis $H_0 : \widetilde{E}(\mathcal{X}) = \widetilde{E}(\mathcal{Y})$ (i.e., $H_0 : D_\theta\big(\widetilde{E}(\mathcal{X}), \widetilde{E}(\mathcal{Y})\big) = 0$) proceeds as follows:

Step 1. Compute the value of the statistic

$$T_{n_1,n_2} = \frac{A_{n_1,n_2}(\widetilde{\boldsymbol{x}}_{\boldsymbol{n_1}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{n_2}})}{\frac{B_{n_1}(\widetilde{\boldsymbol{x}}_{\boldsymbol{n_1}})}{n_1} + \frac{B_{n_2}(\widetilde{\boldsymbol{y}}_{\boldsymbol{n_2}})}{n_2}},$$

where

$$A_{n_1,n_2}(\widetilde{\boldsymbol{x}}_{\boldsymbol{n_1}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{n_2}}) = \left[D_\theta\Big(\frac{1}{n_1} \cdot (\widetilde{x}_1 + \ldots + \widetilde{x}_{n_1}), \frac{1}{n_2} \cdot (\widetilde{y}_1 + \ldots + \widetilde{y}_{n_2})\Big)\right]^2,$$

$$B_{n_1}(\widetilde{\boldsymbol{x}}_{\boldsymbol{n_1}}) = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left[D_\theta\Big(\widetilde{x}_i, \frac{1}{n_1} \cdot (\widetilde{x}_1 + \ldots + \widetilde{x}_{n_1})\Big)\right]^2,$$

$$B_{n_2}(\widetilde{\boldsymbol{y}}_{\boldsymbol{n_2}}) = +\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} \left[D_\theta\Big(\widetilde{y}_j, \frac{1}{n_2} \cdot (\widetilde{y}_1 + \ldots + \widetilde{y}_{n_2})\Big)\right]^2.$$

If all sample data $\widetilde{x}_i$ and $\widetilde{y}_j$ are trapezoidal, then the bootstrap data are also trapezoidal and the sample means are too, whence denoting the mid and spr for values $\widetilde{y}$'s by m' and s'

$$A_{n_1,n_2}(\widetilde{\boldsymbol{x}}_{\boldsymbol{n_1}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{n_2}}) = \left[\frac{1}{n_1}\sum_{i=1}^{n_1} \underline{m}_i - \frac{1}{n_2}\sum_{j=1}^{n_2} \underline{m}'_j)\right]^2 + \left[\frac{1}{n_1}\sum_{i=1}^{n_1} \overline{m}_i - \frac{1}{n_2}\sum_{j=1}^{n_2} \overline{m}'_j)\right]^2$$
$$+\theta\left[\frac{1}{n_1}\sum_{i=1}^{n_1} \underline{s}_i - \frac{1}{n_2}\sum_{j=1}^{n_2} \underline{s}'_j)\right]^2 + \theta\left[\frac{1}{n_1}\sum_{i=1}^{n_1} \overline{s}_i - \frac{1}{n_2}\sum_{j=1}^{n_2} \overline{s}'_j)\right]^2$$
$$+\left[\frac{1}{n_1}\sum_{i=1}^{n_1} \underline{m}_i - \frac{1}{n_2}\sum_{j=1}^{n_2} \underline{m}'_j)\right] \cdot \left[\frac{1}{n_1}\sum_{i=1}^{n_1} \overline{m}_i - \frac{1}{n_2}\sum_{j=1}^{n_2} \overline{m}'_j)\right] + \theta\left[\frac{1}{n_1}\sum_{i=1}^{n_1} \underline{s}_i - \frac{1}{n_2}\sum_{j=1}^{n_2} \underline{s}'_j)\right] \cdot \left[\frac{1}{n_1}\sum_{i=1}^{n_1} \overline{s}_i - \frac{1}{n_2}\sum_{j=1}^{n_2} \overline{s}'_j)\right].$$

Step 2. Fix the bootstrap populations to be as follows (to ensure that bootstrap populations fulfill the null hypothesis, one can add to each value in each sample the mean of the other one):

$$\Big\{\widetilde{x}_1 + \tfrac{1}{n_2} \cdot (\widetilde{y}_1 + \ldots + \widetilde{y}_{n_2}), \ldots, \widetilde{x}_{n_1} + \tfrac{1}{n_2} \cdot (\widetilde{y}_1 + \ldots + \widetilde{y}_{n_2})\Big\},$$

$$\Big\{\widetilde{y}_1 + \tfrac{1}{n_1} \cdot (\widetilde{x}_1 + \ldots + \widetilde{x}_{n_1}), \ldots, \widetilde{y}_{n_2} + \tfrac{1}{n_1} \cdot (\widetilde{x}_1 + \ldots + \widetilde{x}_{n_1})\Big\}.$$

Step 3. Obtain a sample of independent observations from each bootstrap population, say $\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n_1}} = (\widetilde{x}^*_1, \ldots, \widetilde{x}^*_{n_1})$ and $\widetilde{\boldsymbol{y}}^*_{\boldsymbol{n_2}} = (\widetilde{y}^*_1, \ldots, \widetilde{y}^*_{n_2})$.

Step 4. Compute the value of the bootstrap statistic

$$T^*_{n_1,n_2} = \frac{A_{n_1,n_2}(\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n_1}}, \widetilde{\boldsymbol{y}}^*_{\boldsymbol{n_2}})}{\frac{B_{n_1}(\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n_1}})}{n_1} + \frac{B_{n_2}(\widetilde{\boldsymbol{y}}^*_{\boldsymbol{n_2}})}{n_2}}.$$

Step 5. *Steps 3* and *4* should be repeated a large number B of times to get a set of B estimates, denoted by $\{t^*_1, \ldots, t^*_B\}$.

Step 6. Compute the bootstrap p-value as the proportion of values in $\{t^*_1, \ldots, t^*_B\}$ being greater than T_{n_1,n_2}.

To illustrate the application of this method we consider the following:

Example 4.3. Consider in connection with item $M.2$ the sample of the 43 students who have responded by using the computerized form and the sample of 23 students filling the paper-and-pencil format. Table 4 gathers the responses in accordance with both the fuzzy rating scale and the original 4-point Likert one.

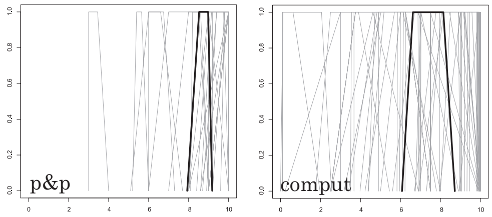
<table>
<tr><th>measurement scale</th><th>fuzzy rating</th><th>4-point Likert</th></tr>
<tr><td>sample data</td><td>
</td><td><table>
<tr><th>M.2</th><th>paper-and-pencil</th><th>computerized</th></tr>
<tr><td>A1</td><td>0</td><td>2</td></tr>
<tr><td>A2</td><td>1</td><td>3</td></tr>
<tr><td>A3</td><td>7</td><td>12</td></tr>
<tr><td>A4</td><td>15</td><td>26</td></tr>
</table></td></tr>
<tr><td>p-value</td><td>.014*</td><td>.572</td></tr>
</table>

Table 4: Fuzzy rating scale-based responses and sample means (in thick black line) to Question $M.2$, 4-point Likert responses to Question $M.2$, and the corresponding p-values in testing for the equality of means ($^*p < .05$)

If we aim to test whether there is or not a significant difference in the mean fuzzy rating scale response to $M.2$ depending of the filled version, we can apply the last algorithm, since the two samples correspond to independent populations. The use of this algorithm (with $B = 1000$) allows us to determine the bootstrap p-value which equals .014 and indicates that at many of the nominal significant levels one can consider there are significant differences in the mean response to $M.2$ depending on the version of the questionnaire students have filled. Actually, this conclusion is coherent with those in testing the second null hypothesis in Example 4.2.

POSED QUESTION	$R.1$	$R.2$	$R.3$	$M.1$	$M.2$	$M.3$	$S.1$	$S.2$	$S.3$
fuzzy rating scale p-value	.135	.051	.601	.311	**.014***	.069	**.032***	.615	.083
Likert p-value	.886	.452	.105	.950	.572	**.001****	.445	.707	.256

Table 5: Analyzing the influence of the filled version of the questionnaire on the fuzzy rating and Likert responses to items $R.1$ to $S.3$ ($^*p < .05$, $^{**}p < .01$)

To additionally illustrate the fact that conclusions can be very different if one considers the 4-point Likert scale (the originally considered for the whole TIMSS-PIRLS questionnaire) we will can make use of the well-known *two-sample Mann-Whitney-Wilcoxon test* for the Likert data and, by using any standard statistical package, we get a p-value equal to .572, so we can conclude the influence of the filled version seems not to be well-supported by the sample data.

If this analysis is enlarged to involve the nine questions in Table 1 we get the p-values in Table 5. Although the (paper & pencil *vs.* computerized) version of the questionnaire does not generally affect significantly most of the mean responses, conclusions are different depending on the considered measurement scale (see, for instance, the results for $M.2$, $M.3$ and $S.1$).

Similar analyses could be developed for the influence of either sex or the fact that students hold shared or individual room at home. All the p-values can be shown to be over .05 with both scales (for most of the items the p-values are much greater than .05).

4.3. Two-sample test for means with fuzzy rating scale-based data (linked samples)

To test the null hypothesis of equality of the Aumann-type means of two RFNs $\mathcal{X}$ and $\mathcal{Y}$, one can consider the *bootstrapped algorithm* by Colubi (2009) approximating the *two-sample test for the means of two RFNs* for linked samples (see González-Rodríguez *et al.*, 2006a), which is now algorithmically summarized.

If $(\mathcal{X}, \mathcal{Y})$ is a two-dimensional random fuzzy set (that is, a mapping from Ω to $\mathcal{F}_c(\mathbb{R}) \times \mathcal{F}_c(\mathbb{R})$ for which α-levels are compact convex random sets of $\mathbb{R}^2$), consider a sample of independent observations from it, $\big((\widetilde{x}_1, \widetilde{y}_1), \ldots, (\widetilde{x}_n, \widetilde{y}_n)\big)$. Denote $\widetilde{\boldsymbol{x}}_{\boldsymbol{n}} = (\widetilde{x}_1, \ldots, \widetilde{x}_n)$ and $\widetilde{\boldsymbol{y}}_{\boldsymbol{n}} = (\widetilde{y}_1, \ldots, \widetilde{y}_n)$.

Then, the algorithm to test the null hypothesis $H_0 : \widetilde{E}(\mathcal{X}) = \widetilde{E}(\mathcal{Y})$ (i.e., $H_0 : D_\theta\big(\widetilde{E}(\mathcal{X}), \widetilde{E}(\mathcal{Y})\big) = 0$) proceeds as follows:

Step 1. Compute the value of the statistic

$$\mathsf{T}_n = \frac{A_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{n}})}{C_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{n}})} = \frac{\left[D_\theta\Big(\frac{1}{n}\cdot(\widetilde{x}_1 + \ldots + \widetilde{x}_n), \frac{1}{n}\cdot(\widetilde{y}_1 + \ldots + \widetilde{y}_n)\Big)\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\left[D_\theta\Big(\widetilde{x}_i + \frac{1}{n}\cdot(\widetilde{y}_1 + \ldots + \widetilde{y}_n), \widetilde{y}_i + \frac{1}{n}\cdot(\widetilde{x}_1 + \ldots + \widetilde{x}_n)\Big)\right]^2}$$

If all sample data $\widetilde{x}_i$ and $\widetilde{y}_i$ are trapezoidal, then the bootstrap data are also trapezoidal and the sample means are too, whence

$$\begin{aligned} C_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}, \widetilde{\boldsymbol{y}}_{\boldsymbol{n}}) = {} & \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{l=1}^{n}(\underline{m}_i + \underline{m}_l - \underline{m}_i' - \underline{m}_l')\right]^2 + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{l=1}^{n}(\overline{m}_i + \overline{m}_l - \overline{m}_i' - \overline{m}_l')\right]^2 \\ & + \frac{\theta}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{l=1}^{n}(\underline{s}_i + \underline{s}_l - \underline{s}_i' - \underline{s}_l')\right]^2 + \frac{\theta}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{l=1}^{n}(\overline{s}_i + \overline{s}_l - \overline{s}_i' - \overline{s}_l')\right]^2 \\ & + \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{l=1}^{n}(\underline{m}_i + \underline{m}_l - \underline{m}_i' - \underline{m}_l')\right]\cdot\left[\frac{1}{n}\sum_{l=1}^{n}(\overline{m}_i + \overline{m}_l - \overline{m}_i' - \overline{m}_l')\right] \\ & + \frac{\theta}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{l=1}^{n}(\underline{s}_i + \underline{s}_l - \underline{s}_i' - \underline{s}_l')\right]\cdot\left[\frac{1}{n}\sum_{l=1}^{n}(\overline{s}_i + \overline{s}_l - \overline{s}_i' - \overline{s}_l')\right]. \end{aligned}$$

Step 2. Fix the bootstrap populations to be as follows (to ensure that bootstrap populations fulfill the null hypothesis, one can add to each value in each sample the mean of the other one):

$$\Big\{\big(\widetilde{x}_1 + \tfrac{1}{n}\cdot(\widetilde{y}_1 + \ldots + \widetilde{y}_n), \widetilde{y}_1 + \tfrac{1}{n}\cdot(\widetilde{x}_1 + \ldots + \widetilde{x}_n)\big), \\ \ldots, \big(\widetilde{x}_n + \tfrac{1}{n}\cdot(\widetilde{y}_1 + \ldots + \widetilde{y}_n), \widetilde{y}_n + \tfrac{1}{n}\cdot(\widetilde{x}_1 + \ldots + \widetilde{x}_n)\big)\Big\}.$$

Step 3. Obtain a sample of independent observations from each bootstrap population, say $\{(\widetilde{x}_1, \widetilde{y}_1)^*, \ldots, (\widetilde{x}_n, \widetilde{y}_n)^*\}$ and, for the sake of simplicity, denote $(\widetilde{x}_i^*, \widetilde{y}_i^*) = (\widetilde{x}_i, \widetilde{y}_i)^*$ and $\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}^* = (\widetilde{x}_1^*, \ldots, \widetilde{x}_n^*)$, $\widetilde{\boldsymbol{y}}_{\boldsymbol{n}}^* = (\widetilde{y}_1^*, \ldots, \widetilde{y}_n^*)$.

Step 4. Compute the value of the bootstrap statistic

$$\mathsf{T}_n^* = \frac{A_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}^*, \widetilde{\boldsymbol{y}}_{\boldsymbol{n}}^*)}{C_n(\widetilde{\boldsymbol{x}}_{\boldsymbol{n}}^*, \widetilde{\boldsymbol{y}}_{\boldsymbol{n}}^*)}.$$

Step 5. *Steps 3* and *4* should be repeated a large number B of times to get a set of B estimates, denoted by $\{\mathsf{t}_1^*, \ldots, \mathsf{t}_B^*\}$.

Step 6. Compute the bootstrap p-value as the proportion of values in $\{\mathsf{t}_1^*, \ldots, \mathsf{t}_B^*\}$ being greater than T_n.

To illustrate the application of this method we consider the following:

Example 4.4. If we aim to test the equality of mean responses to some pairs of items from Table 1, samples are clearly linked, so one can apply the last algorithm.

Table 6 gathers the responses in accordance with both the fuzzy rating scale and the original 4-point Likert one, and also collects the corresponding p-values, in testing the equality of mean responses for the two first items in reading, $R.1$ *vs* $R.2$, and the two first items in math, $M.1$ *vs* $M.2$.

These p-values being equal to 0 indicate that differences are unequivocally significant at any significance level one can usually consider, and whatever the involved measurement scale may be. By looking at the sample means (in black thick line), it can be concluded that: students are aware about how much one can learn from reading, although they do not like that much reading things that made them think; although students seem to like their math teacher more than they like math.

For the last two pairwise comparisons the conclusions coincide for fuzzy rating scale-based data and Likert-type ones. However, this not always happens, as one can see, for instance, in testing the equality of mean responses to items $R.1$ *vs* $S.1$ or $M.3$ *vs* $S.3$. The conclusions for the Likert responses are based on the well-known *two-sample sign test.*

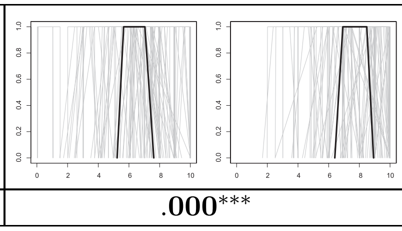
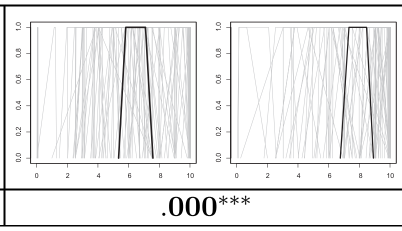
Testing equality of means	fuzzy rating scale data and means	4-point Likert scale data			
R.1 *vs* *R*.2		*R*.1	# ind.	*R*.2	# ind.
		A1	3	A1	0
		A2	13	A2	6
		A3	38	A3	24
		A4	13	A4	37
p-value	**.000*****	**.000*****			
M.1 *vs* *M*.2		*M*.1	# ind.	*M*.2	# ind.
		A1	3	A1	2
		A2	20	A2	4
		A3	20	A3	19
		A4	23	A4	41
p-value	**.000*****	**.000*****			

Table 6: p-Values in testing the equality of mean responses to items $R.1$ *vs* $R.2$, and $M.1$ *vs* $M.2$ ($^{***}p < .001$)

In accordance with the p-values in Table 7: there are no strong evidences against the equality of mean responses to items $R.1$ *vs* $S.1$, this assertion being better supported when Likert-type data are analyzed; there are no strong evidences against the equality of mean fuzzy rating scale-based responses to items $M.3$ *vs* $S.3$, whereas differences are significant at all the usual significance levels when Likert-type data are considered.

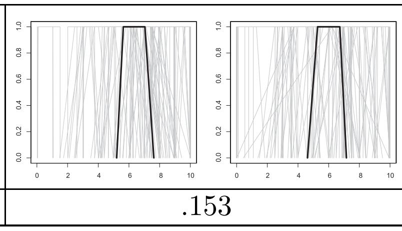
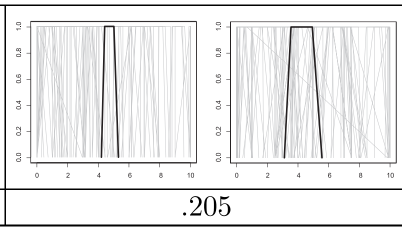
Testing equality of means	fuzzy rating scale data and means	4-point Likert scale data			
R.1 *vs* *S*.1		*R*.1	# ind.	*S*.1	# ind.
		A1	3	A1	7
		A2	13	A2	12
		A3	37	A3	28
		A4	12	A4	18
p-value	.153	.868			
M.3 *vs* *S*.3		*M*.3	# ind.	*S*.3	# ind.
		A1	13	A1	22
		A2	14	A2	19
		A3	10	A3	16
		A4	29	A4	9
p-value	.205	**.000*****			

Table 7: p-Values in testing the equality of mean responses to items $R.1$ *vs* $S.1$, and $M.3$ *vs* $S.3$ ($^{***}p < .001$)

4.4. One-way ANOVA test with fuzzy rating scale-based data (independent samples)

To test the null hypothesis of equality of the Aumann-type means of k RFNs, $\mathcal{X}_1, \ldots, \mathcal{X}_k$, one can consider the *bootstrapped algorithm* approximating the *multi-sample test for the means of k RFNs* (see Gil *et al.*, 2006, and also Colubi, 2009, and González-Rodríguez *et al.*, 2012, for recent updates for the case of independent samples), which is now algorithmically summarized.

If $\mathcal{X}_1, \ldots, \mathcal{X}_k$ are independent RFNs, consider a sample of independent observations $\widetilde{\boldsymbol{x}}_{\boldsymbol{j}} = (\widetilde{x}_{1j}, \ldots, \widetilde{x}_{n_j j})$ from $\mathcal{X}_j$, $j = 1, \ldots, k$, the k samples being also independent. Denote $\widetilde{\boldsymbol{x}}_{\boldsymbol{n_1+\ldots+n_k}} = (\widetilde{x}_{11}, \ldots, \widetilde{x}_{n_1 1}, \ldots, \widetilde{x}_{1k}, \ldots, \widetilde{x}_{n_k k})$, $\overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{j}}} = \frac{1}{n_j} \cdot (\widetilde{x}_{1j} + \ldots + \widetilde{x}_{n_j j})$ and $\overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{(-j)}}} = \overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{1}}} + \ldots + \overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{j-1}}} + \overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{j+1}}} + \ldots + \overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{k}}}$.

Then, the algorithm to test the null hypothesis that $H_0 : \widetilde{E}(\mathcal{X}_1) = \ldots = \widetilde{E}(\mathcal{X}_k)$ (which is equivalent to $H_0 : \sum_{j=1}^k \left[D_\theta\big(\widetilde{E}(\mathcal{X}_j), \widetilde{E}\big(\frac{1}{k}\cdot(\mathcal{X}_1+\ldots+\mathcal{X}_k)\big)\big)\right]^2 = 0$) proceeds as follows:

Step 1. Compute the value of the statistic

$$T_{n_1,\ldots,n_k} = \frac{\sum_{j=1}^k n_j \cdot A_{n_j,n_1+\ldots+n_k}(\widetilde{\boldsymbol{x}}_{\boldsymbol{j}}, \widetilde{\boldsymbol{x}}_{\boldsymbol{n_1+\ldots+n_k}})}{\sum_{j=1}^k \frac{n_j-1}{n_j} B_{n_j}(\widetilde{\boldsymbol{x}}_{\boldsymbol{j}})}$$

$$= \frac{\displaystyle\sum_{j=1}^k n_j \left[D_\theta \left(\tfrac{1}{n_j} \cdot (\widetilde{x}_{1j} + \ldots + \widetilde{x}_{n_j j}), \tfrac{1}{n_1+\ldots+n_k} \cdot (\widetilde{x}_{11} + \ldots + \widetilde{x}_{n_k k})\right)\right]^2}{\displaystyle\sum_{j=1}^k \frac{1}{n_j} \sum_{i=1}^{n_j} \left[D_\theta \left(\widetilde{x}_{ij}, \tfrac{1}{n_j} \cdot (\widetilde{x}_{1j} + \ldots + \widetilde{x}_{n_j j})\right)\right]^2}.$$

Step 2. Fix the bootstrap populations to be as follows: $\left\{\widetilde{x}_{1j} + \overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{(-j)}}}, \ldots, \widetilde{x}_{n_j j} + \overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{(-j)}}}\right\}$, for each $j = 1, \ldots, k$ (to get bootstrap populations fulfilling the null hypothesis).

Step 3. Obtain a sample of independent observations from each bootstrap population, say $\widetilde{\boldsymbol{x}}^*_{\boldsymbol{j}} = (\widetilde{x}^*_{ij}, \ldots, \widetilde{x}^*_{n_j j})$, $j = 1, \ldots, k$, and denote $\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n_1+\ldots+n_k}} = (\widetilde{x}^*_{11}, \ldots, \widetilde{x}^*_{n_1 1}, \ldots, \widetilde{x}^*_{1k}, \ldots, \widetilde{x}^*_{n_k k})$.

Step 4. Compute the value of the bootstrap statistic

$$T^*_{n_1,\ldots,n_k} = \frac{\sum_{j=1}^k n_j \cdot A_{n_j,n_1+\ldots+n_k}(\widetilde{\boldsymbol{x}}^*_{\boldsymbol{j}}, \widetilde{\boldsymbol{x}}^*_{\boldsymbol{n_1+\ldots+n_k}})}{\sum_{j=1}^k \frac{n_j-1}{n_j} B_{n_j}(\widetilde{\boldsymbol{x}}^*_{\boldsymbol{j}})}.$$

Step 5. Steps 3 and *4* should be repeated a large number B of times to get a set of B estimates, denoted by $\{t^*_1, \ldots, t^*_B\}$.

Step 6. Compute the bootstrap p-value as the proportion of values in $\{t_1^*, \ldots, t_B^*\}$ being greater than $T_{n_1,\ldots,n_k}$.

To illustrate the application of this method we consider the following:

Example 4.5. In the ANOVA study to be now described, the samples come from the levels of the factor 'mark taken by the students in the last examination of the subject-matter' (an additional question included in the conducted student's questionnaire). This factor has been assumed to act at 4 levels, associated with

$$\boldsymbol{G1} = \ [0,6], \quad \boldsymbol{G2} = (6,8], \quad \boldsymbol{G3} = (8,9], \quad \boldsymbol{G4} = (9,10].$$

The choice of these levels, which can influence on the final conclusions, has been based on attempting to get a rather balanced distribution of students among groups. The consider attribute has been the fuzzy rating scale-based response to Question $M.2$, in short FRS.

To additionally illustrate the fact that conclusions can be very different depending on the involved scale, we will also consider the 4-point Likert scale and two of its common encodings (namely,

- the usual numerical encoded one re-scaled to the reference interval $[0,10]$, in short NELikert (for which A1 $\equiv 0$, A2 $\equiv 10/3$, A3 $\equiv 20/3$ and A4 $\equiv 10$).
- the set of four terms with its usual fuzzy linguistic encoded semantics re-scaled to $[0,10]$, in short FLELikert (see Figure 2).

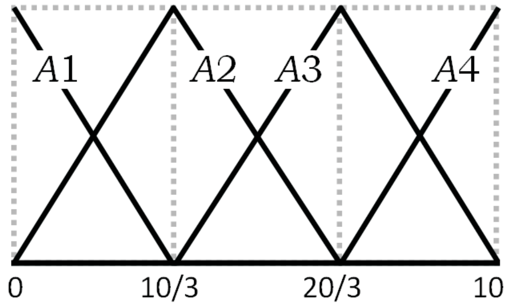


Figure 2: A usual fuzzy linguistic encoding of a 4-point Likert scale re-scaled to $[0,10]$

To analyze Likert-type data we have made use of the well-known *Kruskal-Wallis test* and, for the rest of data types the last algorithm (has been employed, leading to the p-values in Table 8.

The p-values displayed in Table 8 clearly shows that the mark taken in the last examination in Math definitely affects the fuzzy rating scale response to item $M.2$, whereas there is not a great evidence of this influence when Likert data are considered.

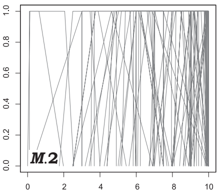
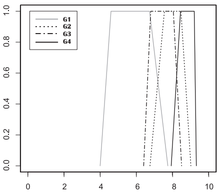
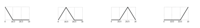
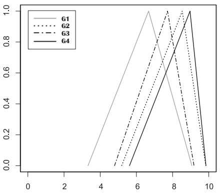
<table>
<tr><th>FRS
data</th><th>FRS
group means</th><th>FRS
p-value</th><th>Likert
data</th><th>KW-Likert
p-value</th></tr>
<tr>
<td>
</td>
<td>
</td>
<td>**.000***&#42;**</td>
<td><table>
<tr><th></th><th>A1</th><th>A2</th><th>A3</th><th>A4</th></tr>
<tr><td><b><i>G</i>1</b></td><td>0</td><td>2</td><td>3</td><td>2</td></tr>
<tr><td><b><i>G</i>2</b></td><td>0</td><td>1</td><td>6</td><td>11</td></tr>
<tr><td><b><i>G</i>3</b></td><td>2</td><td>0</td><td>5</td><td>9</td></tr>
<tr><td><b><i>G</i>4</b></td><td>0</td><td>1</td><td>4</td><td>14</td></tr>
</table></td>
<td>.167</td>
</tr>
<tr><th>FLELikert
data</th><th>FLELikert
group means</th><th>FLELikert
p-value</th><th>NELikert
data</th><th>NELikert
p-value</th></tr>
<tr>
<td><table>
<tr><th></th><th colspan="4"></th></tr>
<tr><td><b><i>G</i>1</b></td><td>0</td><td>2</td><td>3</td><td>2</td></tr>
<tr><td><b><i>G</i>2</b></td><td>0</td><td>1</td><td>6</td><td>11</td></tr>
<tr><td><b><i>G</i>3</b></td><td>2</td><td>0</td><td>5</td><td>9</td></tr>
<tr><td><b><i>G</i>4</b></td><td>0</td><td>1</td><td>4</td><td>14</td></tr>
</table></td>
<td>
</td>
<td><b>.002**</b></td>
<td><table>
<tr><th></th><th>0</th><th>10/3</th><th>20/3</th><th>10</th></tr>
<tr><td><b><i>G</i>1</b></td><td>0</td><td>2</td><td>3</td><td>2</td></tr>
<tr><td><b><i>G</i>2</b></td><td>0</td><td>1</td><td>6</td><td>11</td></tr>
<tr><td><b><i>G</i>3</b></td><td>2</td><td>0</td><td>5</td><td>9</td></tr>
<tr><td><b><i>G</i>4</b></td><td>0</td><td>1</td><td>4</td><td>14</td></tr>
</table></td>
<td><b>.006**</b></td>
</tr>
</table>

Table 8: ANOVAs tests on the effect of the mark taken in the last examination (acting at 4 levels/groups) on the response to Question $M.2$ in accordance with different scales ($^{**}p < .01$, $^{***}p < .001$)

This difference in the statistical conclusions seems to be due in this case not only to the scale, but partially to the applied statistical method, since when the ANOVA for fuzzy data is applied to encoded Likert data the influence is also significant for most of the usual nominal significance levels.

4.5. One-way ANOVA test with fuzzy rating scale-based data (linked samples)

To test the null hypothesis of equality of the Aumann-type means of k RFNs, $\mathcal{X}_1, \ldots, \mathcal{X}_k$, one can consider the *bootstrapped algorithm* approximating the *multi-sample test for the means of k RFNs* (see Montenegro *et al.*, 2009), which is now algorithmically summarized.

If $(\mathcal{X}_1, \ldots, \mathcal{X}_k)$ is a k-dimensional random fuzzy set (that is, a mapping from Ω to $\mathcal{F}_c(\mathbb{R}) \times \overset{(k \text{ times})}{\cdots} \times \mathcal{F}_c(\mathbb{R})$ for which α-levels are compact convex random sets of $\mathbb{R}^k$), consider a sample of independent observations from it, $\big((\widetilde{x}_{11}, \ldots, \widetilde{x}_{1k}), \ldots, (\widetilde{x}_{n1}, \ldots, \widetilde{x}_{nk})\big)$. Denote $\widetilde{\boldsymbol{x}}_{\boldsymbol{n\cdot k}} = (\widetilde{x}_{11}, \ldots, \widetilde{x}_{1k}, \ldots, \widetilde{x}_{n1}, \ldots, \widetilde{x}_{nk})$, $\widetilde{\boldsymbol{x}}_{\boldsymbol{j}} = (\widetilde{x}_{1j}, \ldots, \widetilde{x}_{nj})$, $\overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{j}}} = \frac{1}{n} \cdot (\widetilde{x}_{1j} + \ldots + \widetilde{x}_{nj})$ and $\overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{i.}}} = \frac{1}{k} \cdot (\widetilde{x}_{i1} + \ldots + \widetilde{x}_{ik})$.

Then, the algorithm to test the null hypothesis that $H_0 : \widetilde{E}(\mathcal{X}_1) = \ldots = \widetilde{E}(\mathcal{X}_k)$ (i.e., $H_0 : \sum_{j=1}^{k} \big[D_\theta\big(\widetilde{E}(\mathcal{X}_j), \widetilde{E}\big(\frac{1}{k} \cdot (\mathcal{X}_1 + \ldots + \mathcal{X}_k)\big)\big)\big]^2 = 0$) proceeds as follows:

Step 1. Compute the value of the statistic

$$\mathsf{T}_{n\cdot k} = \frac{\sum_{j=1}^{k} A_{n,n\cdot k}(\widetilde{\boldsymbol{x}}_{\boldsymbol{j}}, \widetilde{\boldsymbol{x}}_{\boldsymbol{n\cdot k}})}{\sum_{j=1}^{k} C_n\big((\widetilde{x}_{1j}, \ldots, \widetilde{x}_{nj}), (\overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{1.}}}, \ldots, \overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{n.}}})\big)}$$

$$= \frac{\sum_{j=1}^{k} \left[D_\theta\left(\frac{1}{n}\cdot(\widetilde{x}_{1j}+\ldots+\widetilde{x}_{nj}), \frac{1}{n\cdot k}\cdot(\widetilde{x}_{11}+\ldots+\widetilde{x}_{nk})\right)\right]^2}{\sum_{j=1}^{k}\frac{1}{n}\sum_{i=1}^{n}\left[D_\theta\left(\widetilde{x}_{ij}+\frac{1}{n\cdot k}\cdot(\widetilde{x}_{11}+\ldots+\widetilde{x}_{nk}), \frac{1}{n}\cdot(\widetilde{x}_{1j}+\ldots+\widetilde{x}_{nj})+\frac{1}{k}\cdot(\widetilde{x}_{i1}+\ldots+\widetilde{x}_{ik})\right)\right]^2}.$$

Step 2. Fix the bootstrap populations to be as follows:

$$\Big\{(\widetilde{x}_{11},\ldots,\widetilde{x}_{1k}),\ \ldots,(\widetilde{x}_{n1},\ldots,\widetilde{x}_{nk})\Big\}.$$

Step 3. Obtain a sample of independent observations from each bootstrap population, say $\big((\widetilde{x}_{11},\ldots,\widetilde{x}_{1k})^*,\ldots,(\widetilde{x}_{n1},\ldots,\widetilde{x}_{nk})^*\big)$ and, for the sake of simplicity, denote $(\widetilde{x}^*_{i1},\ldots,\widetilde{x}^*_{ik}) = (\widetilde{x}_{i1},\ldots,\widetilde{x}_{ik})^*$ for $i = 1,\ldots,n$, $\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n\cdot k}} = (\widetilde{x}^*_{11},\ldots,\widetilde{x}^*_{n1}, \ldots,\widetilde{x}^*_{1k},\ldots,\widetilde{x}^*_{nk})$, $\widetilde{\boldsymbol{x}}^*_{\boldsymbol{j}} = (\widetilde{x}^*_{1j},\ \ldots,\widetilde{x}^*_{nj})$, $\overline{\widetilde{\boldsymbol{x}}^*_{\boldsymbol{j}}} = \frac{1}{n}\cdot(\widetilde{x}^*_{1j}+\ldots+\widetilde{x}^*_{nj})$ and $\overline{\widetilde{\boldsymbol{x}}^*_{\boldsymbol{i\cdot}}} = \frac{1}{k}\cdot(\widetilde{x}^*_{i1}+\ldots+\widetilde{x}^*_{ik})$.

Step 4. Compute the value of the bootstrap statistic

$$\mathsf{T}^*_{n\cdot k} = \frac{k\cdot C_k\big((\overline{\widetilde{\boldsymbol{x}}^*_{\boldsymbol{1}}},\ldots,\overline{\widetilde{\boldsymbol{x}}^*_{\boldsymbol{k}}}),(\overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{1}}},\ldots,\overline{\widetilde{\boldsymbol{x}}_{\boldsymbol{k}}})\big)}{\sum_{j=1}^{k} C_n\big((\widetilde{x}^*_{1j},\ldots,\widetilde{x}^*_{nj}),(\overline{\widetilde{\boldsymbol{x}}^*_{\boldsymbol{1\cdot}}},\ldots,\overline{\widetilde{\boldsymbol{x}}^*_{\boldsymbol{n\cdot}}})\big)}$$

where C_n and C_k use the notation in *Step 1* of Subsection 4.3.

Step 5. *Steps 3* and *4* should be repeated a large number B of times to get a set of B estimates, denoted by $\{\mathsf{t}^*_1,\ldots,\mathsf{t}^*_B\}$.

Step 6. Compute the bootstrap p-value as the proportion of values in $\{\mathsf{t}^*_1,\ldots,\mathsf{t}^*_B\}$ being greater than $\mathsf{T}_{n\cdot k}$.

To illustrate the application of this method we consider the following:

Example 4.6. Table 9 gathers the fuzzy rating scale-based responses to items *R*.3, *M*.3 and *S*.3 along with their mean values and the corresponding ANOVA's p-value.

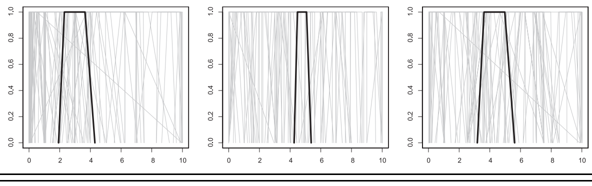
Testing equality of means	fuzzy rating scale data and means	p-value
R.3, *M*.3 and *S*.3		**.000*****

Table 9: p-Values in testing the equality of mean responses to items *R*.3, *M*.3 and *S*.3 ($^{***}p < .001$)

This p-value being equal to 0 indicates that differences are unequivocally significant at any significance level one can usually consider. By looking at the sample means (in black thick line), it can be concluded that: on the average and in contrast to other subjects, students consider reading as being less hard whereas, although math and science are viewed as harder, the comparison is more precise (narrower fuzzy mean) for math than for science. In case we wish to compare fuzzy rating scale with Likert conclusions (in this dependent samples case by using *Friedman test*) we would get also a p-value equal to 0.

It should be commented that, although the two-sample testing for the equality of mean fuzzy rating scale responses to $M.3$ *vs* $S.3$ has not shown strong evidences against such an equality, the other pairwise comparisons do. Moreover, it should be taken into account that the sample of students is not exactly the same because the pairwise comparison $M.3$ *vs* $S.3$ involves one more student.

5. Concluding remarks

With this paper we have attempted to illustrate the potentiality of the already developed methodology for testing about means of fuzzy data, especially in case we deal with fuzzy rating scale-based ones.

All statistical analyses have been performed using R version 3.0.1 (The R Core Team, 2013), and the test procedures with independent samples have been run using the SAFD package (Trutschnig & Lubiano, 2012).

It should be remarked that there are many other studies to be developed, although they are beyond the length of this paper and also will depend in practice on the real interests users can have. Furthermore, there are many statistical methods to be developed yet. Among these ones, one can mention the developments of testing procedures about the equality of medians or other location measures, or even about the equality of distributions of random fuzzy numbers. These are some of the open problems we plan to tackle in a near future.

We would like finally indicate that the hypothesis testing methodology in this paper differs from some other ones also involving fuzzy data (see, among others, Filzmoser & Viertl, 2004, in which a fuzzy p-value is considered and Hryniewicz, 2006, in which a possibilistic approach is followed), but most of them refer to real-valued parameters of certain underlying random variables, whereas this paper involves fuzzy-valued ones and p-values should be probabilistically interpreted.

Acknowledgements

Authors are grateful to Colegio San Ignacio in Oviedo-Asturias (Spain), its staff and its fourth grade students and parents for the support and understanding shown when they allowed us to collect the data in the real-life example. Actually, the nine and ten-year old students in this school have quite properly caught the spirit and the mechanics behind fuzzy rating scale-based questionnaires. Authors also want to acknowledge their colleague Professor Carlos Carleos because of his design and implementation of the computerized format many students have used to answer the questionnaire. They also want to thank the reviewers and Editor in charge of the original of this paper for their valuable comments and suggestions.

The research in this paper has been partially supported by/benefited from Principality of Asturias Grants GRUPIN14-101, Research Contract E-33-2015-0040746 (Sinova) and Severo Ochoa BP12012 (De la Rosa de Sáa), and the Spanish Ministry of Economy and Competitiveness Grant MTM2013-44212-P. Their financial support is gratefully acknowledged.

References

Ballouard, J.M., Mullin, S.J., Ajtic, R., Brito, J.C., ElMouden, E.H., Erdogan, M., Feriche, M., Pleguezuelos, J.M., Prokop, P., Sánchez, A., Santos, X., Slimani, T., Sterijovski, B., Tomovic, L., Uşak, M., Zuffi, M., & Bonnet, X. (2015). Factors influencing schoolchildren's responses to a questionnaire in wildlife conservation education. *International Journal of Science Education, 37*, 469–483.

Ban, A., Coroianu, L., & Grzegorzewski, P. (2011). Trapezoidal approximation and aggregation. *Fuzzy Sets and Systems*, *177*, 45–59.

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

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. (2014). A distance-based statistical analysis of fuzzy number-valued data. *International Journal of Approximate Reasoning*, *55*, 1487–1501.

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. (2014). Rejoinder on "A distance-based statistical analysis of fuzzy number-valued data." *International Journal of Approximate Reasoning*, *55*, 1601–1605.

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

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

Corral-Blanco, N., Zurbano-Fernández, E., Blanco-Fernández, A., García-Honrado, I., & Ramos-Guajardo, A.B. (2013). Structure of the family educational environment: its influence on performance and differential performance. In *PIRLS-TIMSS 2011 International Study on Progress in Reading Comprehension, Mathematics and Sciences IEA. Volume II. Spanish Report. Secondary Analysis* (pp. 9-31). Ministerio de Educación, Cultura y Deporte, Instituto Nacional de Evaluación Educativa.

De la Rosa de Sáa, S., Gil, M.A., González-Rodríguez, G., López, M.T., & Lubiano, M.A. (2015). Fuzzy rating scale-based questionnaires and their statistical analysis. *IEEE Transactions on Fuzzy Systems*, *23*, 111–126.

Diamond, P., & Kloeden, P. (1990). Metric spaces of fuzzy sets. *Fuzzy Sets and Systems*, *35*, 241–249.

Filzmoser, P., & Viertl, R. (2004). Testing hypotheses with fuzzy data: the fuzzy p-value. *Metrika*, *59*, 21–29.

Gil, M.A., Lubiano, M.A., de la Rosa de Sáa, S., & Sinova, B. (2015). Analyzing data from a fuzzy rating scale-based questionnaire. A case study, *Psicothema*, *27*, 182–191.

Gil, M.A., Montenegro, M., González-Rodríguez, G., Colubi, A., & Casals, M.R. (2006). Bootstrap approach to the multi-sample test of means with imprecise data. *Computational Statistics* & *Data Analysis*, *51*, 148–162.

Giné, E., & Zinn, J. (1990). Bootstrapping general empirical measures. *Annals of Probability*, *18*, 851–869.

González-Rodríguez, G., Colubi, A. D'Urso, P., & Montenegro, M. (2009). Multi-sample test-based clustering for fuzzy random variables. *International Journal of Approximate Reasoning*, *50*, 721–731.

González-Rodríguez, G., Colubi, A., Gil, M.A., & D'Urso, P. (2006). An asymptotic two dependent samples test of equality of means of fuzzy random variables. In *Proceedings of the 17th Conference of IASC-ERS-COMPSTAT'2006* (pp. 689–695).

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

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

Grzegorzewski, P. (2008). Trapezoidal approximations of fuzzy numbers preserving the expected interval - algorithms and properties. *Fuzzy Sets and Systems*, *159*, 1354–1364.

Grzegorzewski, P., & Pasternak-Winiarska, K. (2011). Trapezoidal approximations of fuzzy numbers with restrictions on the support and core. In *Proceedings 7th Conference EUSFLAT-2011 and LFA-2011* (pp. 749–756). Paris: Atlantis Press.

Haelermans, C., & De Witte, K. (2012). The role of innovations in secondary school performance Evidence from a conditional efficiency model. *European Journal of Operational Research*, *223*, 541–549.

Herrera, F., Herrera-Viedma, E., & Martínez, L. (2008). A fuzzy linguistic methodology to deal with unbalanced linguistic term sets. *IEEE Transasctions on Fuzzy Systems*, *16*, 354–370.

Herrera, F., Herrera-Viedma, E., & Verdegay, J.L. (1998). Choice processes for non-homogeneous group decision making in linguistic setting. *Fuzzy Sets and Systems*, *94*, 287–308.

Hesketh, B., Griffin, B., & Loh, V. (2011). A future-oriented retirement transition adjustment framework. *Journal of Vocational Behavior*, *79*, 303–314.

Hesketh, T., & Hesketh, B. (1994). Computerized fuzzy ratings: the concept of a fuzzy class. *Behavior Research Methods, Instruments* & *Computers*, *26*, 272277.

Hesketh, B., Hesketh, T., Hansen, J.-I., & Goranson, D. (1995). Use of fuzzy variables in developing new scales from the strong interest inventory. *Journal of Counseling Psychology*, *42*, 85–99.

Hesketh, T., Pryor, R., & Hesketh, B. (1988). An application of a computerized fuzzy graphic rating scale to the psychological measurement of individual differences. *International Journal of Man-Machine Studies*, *29*, 21–35.

Hryniewicz, O. (2006). Possibilistic decisions and fuzzy statistical tests. *Fuzzy Sets and Systems*, *157*, 2665–2673.

Körner, R. (2000). An asymptotic α-test for the expectation of random fuzzy variables. *Journal of Statistical Planning and Inference*, *83*, 331–346.

Kräatschmer, V. (2002). Limit theorems for fuzzy-random variables. *Fuzzy Sets and Systems*, *126*, 253–263.

Lalla, M., Facchinetti, G., & Mastroleo, G. (2008). Vagueness evaluation of the crisp output in a fuzzy inference system. *Fuzzy Sets and Systems*, *159*, 3297–3312.

Lubiano, M.A., de la Rosa de Sáa, S., Sinova, B., & Gil, M.A. (2015). Empirical sensitivity analysis on the influence of the shape of fuzzy data on the estimation of some statistical measures. In P. Grzegorzewski, M. Gagolewski, O. Hryniewicz, & M.A. Gil (Eds.), *Strengthening Links Between Data Analysis and Soft Computing* (pp. 123–131). Heidelberg: Springer.

Lubiano, M.A., Gil, M.A., López-Díaz, M., & López, M.T. (2000). The lambda-mean squared dispersion associated with a fuzzy random variable. *Fuzzy Sets and Systems*, *111*, 307–317.

Montenegro, M., Casals, M.R., Lubiano, M.A., & Gil, M.A. (2001). Two-sample hypothesis tests of means of a fuzzy random variable. *Information Sciences*, *133*, 89–100.

Montenegro, M., Colubi, A., Casals, M.R., & Gil, M.A. (2004). Asymptotic and Bootstrap techniques for testing the expected value of a fuzzy random variable. *Metrika*, *59*, 31–49.

Montenegro, M., López-García, M.T., Lubiano, M.A., & González-Rodríguez, G. (2009). A dependent multi-sample test for fuzzy means. In *Abstracts of Second Workshop of the ERCIM Working Group on Computing* & *Statistics-ERCIM'09* (p. 102).

Nikolaidis, Y., & Dimitriadis, S.G. (2014). On the student evaluation of university courses and faculty members' teaching performance. *European Journal of Operational Research*, *238*, 199–207.

Pedrycz, W. (1989). A fuzzy cognitive structure for pattern recognition. *Pattern Recognition Letters*, *9*, 305–313.

Pedrycz, W. (1994). Why triangular membership functions? *Fuzzy Sets and Systems*, *64*, 21–30.

Peró, M., Soriano, P.P., Capilla, R., Guàrdia i Olmos, J., & Hervás, A. (2015). Questionnaire for the assessment of factors related to university degree choice in Spanish public system: A psychometric study. *Computers in Human Behavior*, *47*, 128–138.

Puri, M.L., & Ralescu, D.A. (1985). The concept of normality for fuzzy random variables. *Annals of Probability*, *11*, 1373–1379.

Puri, M.L., & Ralescu, D.A. (1986). Fuzzy random variables. *Journal of Mathematical Analysis and Applications*, *114*, 409–422.

Roszkowska, E., & Wachowicz, T. (2015). Application of fuzzy TOPSIS to scoring the negotiation offers in ill-structured negotiation problems. *European Journal of Operational Research*, *242*, 920–932.

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds). (2014). *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis.* Boca Raton: Chapman and Hall/CRC.

Rutkowski, L., & Rutkowski, D. (2010). Getting it 'better': The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies*, 42, 411–430.

Takemura, K. (1999). A fuzzy linear regression analysis for fuzzy input-output data using the least squares method under linear constraints and its application to fuzzy rating data. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *3*, 36–41.

Takemura, K. (2007). Ambiguous comparative judgment: fuzzy set model and data analysis. *Japanese Psychological Research*, *49*, 148–156.

The R Core Team (2013). R: A language and environment for statistical computing. *R Foundation for Statistical Computing.* Vienna, Austria. http://www.r-project.org/.

Tong, R.M., & Bonissone, P.P. (1980). A linguistic approach to decision making with fuzzy sets. *IEEE Transactions on Systems Man and Cybernetics*, *10*, 716–723.

Trutschnig, W., González-Rodríguez, G., Colubi, A., & Gil, M.A. (2009). A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Information Sciences*, *179*, 3964–3972.

Trutschnig, W., & Lubiano, M.A. (2012). SAFD: Statistical Analysis of Fuzzy Data. R package version 0.4. (http://cran.r-project.org/web/packages/SAFD/index.html).

Wu, H.-C. (1999). The central limit theorems for fuzzy random variables. *Information Sciences*, *120*, 239–256.

Yan, H.-B., & Ma, T. (2015). A group decision-making approach to uncertain quality function deployment based on fuzzy preference relation and fuzzy majority. *European Journal of Operational Research*, *241*, 815–829.

Zadeh, L.A. (1975). The concept of a linguistic variable and its application to approximate reasoning. Part 1. *Information Sciences*, *8*, 199–249; Part 2. *Information Sciences*, *8*, 301–353; Part 3. *Information Sciences*, *9*, 43–80.