Statistics with fuzzy data by using random fuzzy sets

Gil, María Ángeles
Universidad de Oviedo, Departamento de Estadística e I.O. y D.M.
C/ Calvo Sotelo, s/n
33007 Oviedo, Spain
E-mail: magil@uniovi.es

Many real-life random experiments involve variables which are associated with judgements, opinions, perceptions, ratings, and so on. 'Values' for these variables are usually non-numerical, but they correspond to imprecise values or categories. A well-known example of this type of experiments is the one corresponding to most of the usual questionnaires and surveys with a pre-specified response format, in which people are asked to respond to a series of questions and variable values are the different answers from respondents.

Most of these random experiments are designed for statistical analysis of the values. For this purpose, values are either treated as categorical ones or coded/ranked by using numbers (often, integer ones). The drawbacks for such treatments are mainly due to the fact that

- both the categorization and the integer coding discretize the variables so that the number of different values is usually a small one;
- the categorical scale can capture the imprecision underlying variables values, but subjectivity and diversity are usually not well reflected;
- the integer coding can capture neither the imprecision underlying variables values nor their subjectivity and diversity;
- the transition from a value to a 'consecutive' one is abrupt,
- and in case of the integer coding the distance between the coded values does not properly reflect the real deviation between the corresponding uncoded ones.

Furthermore, the methodologies to deal with data from such random experiments are rather limited in contrast to the methodology to deal with real- or vectorial-valued random variables.

In the last decades an alternative treatment has been suggested for these data. This treatment is based on modeling variable values by means of (normal, compact and convex) fuzzy sets of finite-dimensional Euclidean spaces, that is, on considering the statistical analysis of fuzzy data. In this respect, several studies have been carried out, many of them by analyzing data from a descriptive perspective, although inferential developments have been also considered.

To examine fuzzy data from a statistical perspective several concepts have been introduced in the literature to model the random mechanisms/processes leading to fuzzy data. Some of these concepts formalize either fuzzy perceptions of real-valued random variables (like, for instance, the fuzzy random variables as intended by Kwakernaak, 1978, Kruse and Meyer, 1987) or sets of random variables induced by a fuzzy relation describing an ill-defined probability (like the recent approach to fuzzy random variables by Couso and Dubois, 2009).

Another approach, which will be the one to be considered hereinafter, is that of modeling random mechanisms leading to intrinsically imprecise-valued random attributes. We will refer to this notion as random fuzzy sets (because they extend the concept of random sets and, hence, that of random variable), although Puri and Ralescu who introduced it in 1986 coined them as fuzzy random variables. Random fuzzy sets can be identified through the so-called support function with a special type of functional-valued random elements, and summary measures for these ones could be particularized to them, so the notion and related measures are well supported in the probabilistic setting.

Moreover, thanks to the use of random fuzzy sets, the usual fuzzy arithmetic, a versatile and intuitive metric between fuzzy values, and some limit results for generalized spaces, one can state a wide statistical methodology exploiting deeply the rich and expressive information contained in fuzzy data. Because of the way random fuzzy sets are introduced, and because of them to be identifiable with certain functional-valued random elements, it would be possible preserving almost all the ideas and goals orienting statistics with random variables. In this way, we can refer (for instance) to

- the unbiasedness or the consistency (in accordance with a metric sense) of a non-fuzzy or a fuzzy estimator,
- the p-value of a given test for a sample of fuzzy data,
- the power of a test based on a sample of fuzzy data, and so on.

This paper presents a review on some recent statistical developments concerning random fuzzy sets, and some instances of direct application to deal with fuzzy data as well as an implication to deal with real-valued data.

Preliminaries

In treating fuzzy-valued data from a statistical perspective several problems arise in contrast to treating real/vectorial-valued ones, namely,

- when the space of fuzzy values is endowed with the usual arithmetic it is not a linear but a semilinear space; in fact, it is not possible to state a difference between fuzzy values extending the one between real numbers/vectors and, simultaneously, being well-defined and preserving the meaning and properties of the difference between real numbers/vectors;
- there is no universally acceptable ranking between fuzzy values;
- there is not useful and widely applicable distribution models for random fuzzy sets; although there is a notion of normally distributed random fuzzy sets, the model is very restrictive and unrealistic so its practical applicability is rather questionable and weak;
- there are not limit results like Central Limit ones with immediate applicability to deal with random fuzzy sets for inferential purposes.

Nevertheless, these drawbacks can be avoided in most of the studies by making use of some tools, like suitable metrics between fuzzy values (with a meaning similar to the distances between real/vectorial values, for instance, the one by Trutschnig *et al.*, 2009), the Strong Laws of Large Numbers for random fuzzy sets (see, for instance, Colubi *et al.*, 1999), and the bootstrapped Central Limit Theorem for generalized spaces by Giné and Zinn (1990).

This section aims to recall and summarize the basic preliminary notions supporting the statistical methodology we will briefly present in the next section.

The space of *fuzzy values* modeling data to be considered in the paper is formalized as the class $\mathcal{F}_c(\mathbb{R}^p)$ of mappings $\widetilde{U} : \mathbb{R}^p \to [0,1]$ so that for each $\alpha \in (0,1]$ the *α-level set*, $\widetilde{U}_\alpha = \{x \in \mathbb{R}^p : \widetilde{U}(x) \geq \alpha\}$, is a nonempty compact convex set of $\mathbb{R}^p$. When $p = 1$ we will refer to fuzzy values as *fuzzy numbers*.

A fuzzy value $\widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p)$ models an ill-defined property on (or subset of, or assertion on) $\mathbb{R}^p$, so that for each $x \in \mathbb{R}^p$ the value $\widetilde{U}(x)$ can be interpreted as the 'degree of compatibility' of x with the property 'defining' $\widetilde{U}$ (or 'degree of membership' of x to $\widetilde{U}$, or 'degree of truth' of the assertion x is $\widetilde{U}$). Equivalently, fuzzy values can be defined to be $[0,1]$-valued upper semicontinuous functions with convex bounded α-levels for all $\alpha \in (0,1]$ (i.e., compact and convex fuzzy sets), and attaining at least once the maximum value 1 (i.e., normal fuzzy sets). Real/vectorial/interval/set-valued data can be viewed as particular fuzzy data, by simply identifying them with the associated indicator functions.

The strength of the scale of fuzzy values in representing those taken on by variables involving judgements, opinions, ratings, and so on, lies in the fact that

- the fuzzy modeling is fully flexible, expressive and allows accurate description (especially when values are not constrained to be chosen from a prefixed set of properties or 'linguistic labels');
- the fuzzy scale can capture the imprecision, the subjectivity and the diversity of this type of variable values;
- the transition from a value to another (consecutiveness makes no real sense for fuzzy values) is gradual;
- the distance between fuzzy values to be considered later reflects appropriately the deviation between them.

The elementary arithmetic operations required to handle fuzzy data for statistical purposes will be the sum and the product by a scalar. These two operations can be approached either by applying directly Zadeh's extension principle (Zadeh, 1975) or, equivalently and based on the results by Nguyen (1978), as the level-wise extension of the usual set-valued arithmetic. In this way, given two fuzzy values $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c(\mathbb{R}^p)$ and a real number γ, the *sum of* $\widetilde{U}$ *and* $\widetilde{V}$ is defined as the fuzzy value $\widetilde{U} + \widetilde{V} \in \mathcal{F}_c(\mathbb{R}^p)$ such that for each $\alpha \in (0,1]$

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

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

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

The metric between fuzzy values to be considered is inspired on both, a metric between fuzzy numbers introduced by Bertoluzza *et al.* (1995) and the concept of *support function of a fuzzy value* by Puri and Ralescu (1985) which extends level-wise the well-known concept of support function of a compact convex set (see, for instance, Castaing and Valadier, 1977). On the basis of the support function one can define (see Trutschnig *et al.*, 2009) the *mid/spr characterization of a fuzzy value* which is given by $\operatorname{mid} s_{\widetilde{U}}(u,\alpha) =$ mid-point/center of $\Pi_u \widetilde{U}_\alpha$, $\operatorname{spr} s_{\widetilde{U}}(u,\alpha) =$ spread/radius of $\Pi_u \widetilde{U}_\alpha$, with $\Pi_u \widetilde{U}_\alpha$ denoting the projection of $\widetilde{U}_\alpha$ over a direction u in the unit sphere in $\mathbb{R}^p$, $\mathbb{S}^{p-1}$. This characterization takes into account the 'location' of the fuzzy number as well as the 'shape' which will be relevant also in the arithmetic and the distance to be now recalled.

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

$$D_\theta^\varphi(\widetilde{U}, \widetilde{V}) = \sqrt{\left(\|\operatorname{mid} \widetilde{U} - \operatorname{mid} \widetilde{V}\|^\varphi\right)^2 + \theta\left(\|\operatorname{spr} \widetilde{U} - \operatorname{spr} \widetilde{V}\|^\varphi\right)^2}$$

$\|f\|^\varphi = [f,f]^\varphi = \int_{(0,1]} \int_{\mathbb{S}^{p-1}} (f(u,\alpha))^2 \, du \, d\varphi(\alpha)$. The choice of θ allows us to weight the effect of the deviation between spreads (difference in 'shape' or 'imprecision'), in contrast to the effect of the deviation between mid's (which can be intuitively translated into the difference in 'location'), and it is usually constrained to belong to $(0,1]$. On the other hand, the choice of φ enables to weight the relevance of different levels (i.e., the degree of 'compatibility'), and this measure has no stochastic but weighting mission.

It should be remarked that the support function states an isometrical embedding of $\mathcal{F}_c(\mathbb{R}^p)$ (with the fuzzy arithmetic and D_θ^φ) onto a closed convex cone of the L^2-type real-valued functions defined on $\mathbb{S}^{p-1} \times (0,1]$ (with the usual functional arithmetic and a certain metric). Based on this embedding, one can establish a one-to-one identification of fuzzy data with functional data in a closed convex cone of a certain Hilbert space. As a consequence, the concept of functional-valued random element and associated summary measures can be particularized to deal with fuzzy-valued random elements as we will see now. This also means that some Functional Data Analysis techniques could be particularized to deal with fuzzy data, although care should be taken in ensuring that in the functional treatment of fuzzy data we move always within the cone. This forces sometimes either to use tools with such a guarantee or to develop *ad-hoc* methods to handle fuzzy data.

Random fuzzy sets were introduced by Puri and Ralescu (1986) as a mathematical model for mechanisms associating a fuzzy value with each experimental outcome and extending level-wise the concept of random set. Given a probability space $(\Omega, \mathcal{A}, P)$ modeling a random experiment, a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R}^p)$ is said to be a *random fuzzy set* (for short RFS) if and only if it is a Borel measurable mapping w.r.t. $\mathcal{A}$ and the Borel σ-field generated by the topology induced by D_θ^φ on $\mathcal{F}_c(\mathbb{R}^p)$. Equivalently, the notion of RFS could be either induced directly from that of Hilbert space-valued random element, or induced directionally- and level-wise from that of real-valued random variable by using either the mid/spr characterization. In González-Rodríguez *et al.*, 2011, one can find some other equivalences for this concept. The Borel measurability of RFS's assures that one can immediately refer in this setting to notions like the distribution induced by an RFS, the stochastic independence of RFS's, and so on.

In analyzing fuzzy data, summary measures (also called parameters) for RFS's can be induced or stated, the most popular one being the mean value. Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated RFS $\mathcal{X}$, its (Aumann type) *mean value* is the fuzzy value $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R}^p)$ such that for all $\alpha \in (0, 1]$

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

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

Due to the properties of the support function and the Hilbertian random elements, the mean value of an RFS satisfies the usual properties of linearity and it is coherent with the fuzzy arithmetic. Furthermore, it is the Fréchet expectation w.r.t. D_θ^φ, which corroborates the fact that it is a central tendency measure. The (fuzzy) mean value of an RFS is supported by Strong Laws of Large Numbers for RFS's (cf. Colubi *et al.*, 1999), so that it is the almost sure limit (w.r.t. different metrics) of the 'sample fuzzy mean'.

The Fréchet variance of an RFS could be also defined (see Lubiano *et al.*, 2000, Körner and Näther, 2002, and González-Rodríguez *et al.*, 2011), as well as the covariance of two RFS's. They preserve most of the valuable properties from the corresponding parameters in the non-fuzzy case, those not being preserved being mainly related and due to the lack of linearity of the space of fuzzy values.

Statistical developments concerning random fuzzy sets

Random fuzzy sets fit appropriately many real-life classification/qualification processes associated with valuations, ratings, judgements leading to imprecise data. As we have commented, they mean a well-formalized model within the probabilistic setting and, consequently, ideas, concepts and methods in the statistical analysis of real/vectorial-valued data make (at least theoretically) sense in handling experimental fuzzy data. In this respect, we have tried to preserve as many as possible of these ideas, concepts and methods.

More precisely, the aim of Statistics with fuzzy data obtained from random fuzzy sets will be to draw conclusions about the distribution of RFS's over populations, on the basis of the information supplied by samples of observations from these RFS's. Several inferential procedures based on RFS's have been already developed. Among them, one can mention

- distance-based inferential statistics about parameters of the distribution of RFS's, namely,
  - 'point' and 'region' (fuzzy) estimation of the mean of an RFS;
  - testing 'two-sided' and 'approximative/comparative' hypotheses about means of RFS's;
  - testing hypotheses about Fréchet-variances of RFS's, and so on.
- least squares regression/correlation analysis involving RFS's, etc.

In connection with the *'point' estimation of the population mean of an RFS*, the objective has been to approximate this unknown population fuzzy mean by means of the sample fuzzy mean. This sample mean, which is defined similarly to the non-fuzzy case in terms of the usual fuzzy arithmetic, satisfies convenient properties like unbiasedness and strong consistency (in the D_θ^φ-metric case, and even in many other metric senses). Furthermore, one can quantify a mean squared error in the fuzzy estimation by considering D_θ^φ metric (cf. Lubiano, 1999).

Concerning the *'region' estimation of the population mean of an RFS*, the objective has been to approximate this unknown population fuzzy mean by means of a confidence region defined as a ball w.r.t. a given metric, which is centered in the sample fuzzy mean and for which the radius is determined via bootstrapping (cf. González-Rodríguez *et al.*, 2009).

The *'two-sided' hypotheses about the fuzzy means* that have been considered for testing purposes refer to the equality of the population (fuzzy) mean(s) of RFSs. The global objective has been to test, at a given significance level, whether a null hypothesis formalized as the equality of two fuzzy values (the population fuzzy mean and a given one in the one-sample case, and population fuzzy means in the multi-sample case) could be accepted or should be rejected on the basis of the available sample of fuzzy data. Through the D_θ^φ metric these null hypotheses have been identified with 'two-sided' hypotheses of real values and different procedures have been carried out.

For the *one-sample case* some of the developed tests have been: an exact test for 'normal' RFS's (cf. Montenegro *et al.*, 2004); asymptotic tests for general RFS's (cf. Körner, 2000, Montenegro *et al.*, 2004); a bootstrap test for general RFS's (cf. González-Rodríguez *et al.*, 2006a). For the *two-sample* and *multi-sample cases* similar tests have been stated (see, for instance, Montenegro *et al.*, 2001, Gil *et al.*, 2006, González-Rodríguez *et al.*, 2011). It should be pointed out that the exact method for normal RFS's (in Puri and Ralescu's sense, 1985) is exact and easy-to-apply, but this assumption of normality is quite restrictive and unrealistic. Regarding the asymptotic procedures, they are based on some Central Limit Theorems for generalized space-valued random elements, with the practical inconvenient of the need for large samples and the fact that sometimes the asymptotic results involve unknown parameter or lead to elements out of the cone the space of fuzzy values is embedded into. The Generalized Bootstrapped Central Limit Theorem by Giné and Zinn (1990) has enabled to incorporate bootstrap techniques, so that for small/medium samples the bootstrap test usually outperforms the asymptotic one and for large sample sizes the improvement is not that remarkable, but the bootstrap approach still provides the best approximation to the nominal significance level. Furthermore, the bootstrap tests have been proved to be consistent.

The *'approximative/comparative' hypotheses about the fuzzy means* have been recently examined (see Ramos-Guajardo, 2011) and their test will be useful to formalize certain hypotheses concerning the fuzzy means which cannot be expressed in terms of the equality of some fuzzy values. The problem of *testing the equality of Fréchet-variances* of RFS's have been also recently discussed (see Ramos-Guajardo and Lubiano, 2011).

With respect to the *linear regression and correlation involving RFS's* different approaches have been considered depending on different models, distances, etc. (see, for instance, Näther, 2006, González-Rodríguez *et al.*, 2009, Ferraro *et al.*, 2010).

Lubiano and Trustchnig (2011) have developed an R-package called SAFD (Statistical Analysis of Fuzzy Data) to perform statistical analysis of RFS's. For more details readers are referred to `http://bellman.ciencias.uniovi.es/SMIRE/SAFDpackage.html`.

Applications

One of the most immediate real-life applications for the preceding developments are those related to the *design and analysis of questionnaires, surveys, and so on*, mainly when a *free-fuzzy response format* is allowed and driven. The concept of fuzzy values, and especially the particular and most common one of fuzzy numbers, is easy to understand and quite friendly to employ. So, it would be possibly either unfeasible or not useful to set up the use of questionnaires or surveys involving such a free-fuzzy response format in case of polls carried out by phone, at the street and, in general whenever the final goal is not extremely relevant. However, in case the analysis have a deep scientific, economical, political, interest and implication (say, for educational studies, medical diagnoses concerning the seriousness of an illness, the valuation on the loyalty of a customer to a bank, the human resource information on hiring employees, and so on) it would be really valuable in capturing expressiveness, imprecision, subjectivity and diversity, and the statistical analysis of the responses will exploit a lot of relevant information. An example for such a type of questionnaire can be found in Gil and González-Rodríguez (2011).

Another useful implication is not a direct one, but it is based on the combination of the statistical analysis of fuzzy data recalled in the preceding section and the so-called *characterizing fuzzy representation of a real-valued random variable* (see González-Rodríguez *et al.*, 2006b). The fuzzy mean value of such a representation of a real-valued random variable is a functional characterization of its distribution. This characterization have several key features, namely, its functional 'dominium' is bounded and included in $\mathbb{R}$, its functional image is the unit interval $[0, 1]$, it has an intuitive interpretation as a 'mean value' of a functionally-valued random element, and probabilistic/statistical results for mean values of generalized space-valued random elements can be applied.

Moreover, it can be used to develop consistent statistical inferences on the distribution of random variables. In this way: the 'point' estimator of the mean value of the characterizing fuzzy representation of a random variable determines an estimator of its distribution; the one-sample test about such a fuzzy mean value can be used to test the goodness-of-fit of a random variable to a specified distribution; the multi-sample test about mean values corresponding to characterizing fuzzy representations of random variables can be applied to develop an ANOVA test for distributions; and son on. Consequently, based on this representation, an integral methodology to develop statistical inferences on the distributions of real-valued random variables has been derived. This methodology shows convenient properties (like strong consistency and others) because of being based on mean values.

Acknowledgements

This research has been partially supported by/benefited from the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-01 and MTM2009-09440-C02-02, and the COST Action IC0702. Their financial support is gratefully acknowledged.

REFERENCES

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

Castaing, C., Valadier, M. (1977). *Convex Analysis and Measurable Multifunctions.* Lect. Notes Math. 580. Springer-Verlag, Berlin.

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

Couso, I., Dubois, D. (2009). On the variability of the concept of variance for fuzzy random variables. *IEEE Trans. Fuzzy Syst.* **17**, 1070–1080.

Ferraro, M.B., Coppi, R., Gonzlez Rodrguez, G., Colubi, A. (2010). A linear regression model for imprecise response. *Int. J. Approx. Reason.* **51**, 759–770.

Gil, M.A., Gonzlez-Rodrguez, G. (2011). Fuzzy *vs* Likert scale in Statistics. In: Trillas, E., Bonissone, P., Magdalena, L., Kacprzyk, J. (Eds.) *Combining Experimentation and Theory (A Hommage to Abe Mamdani).* Springer, Heidelberg (in press).

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. *Comp. Statist. Data Anal.* **51**, 148–162.

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

Gonzlez-Rodrguez, G., Blanco, A., Colubi, A., Lubiano, M.A. (2009). Estimation of a simple linear regression model for fuzzy random variables. *Fuzzy Sets and Systems* **160**, 357–370.

Gonzlez-Rodrguez, G., Colubi, A., Gil, M.A. (2006b). A fuzzy representation of random variables: an operational tool in exploratory analysis and hypothesis testing. *Comp. Statist. Data Anal.* **51**, 163–176.

González-Rodríguez, G., Colubi, A., Gil, M.A. (2011). Fuzzy data treated as functional data. A one-way ANOVA test approach. *Comp. Statist. Data Anal.* (in press, doi:10.1016/j.csda.2010.06.013).

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

González-Rodríguez, G., Trutschnig, W., Colubi, A. (2009). Confidence regions for the mean of a fuzzy random variable. In *Abst. IFSA-EUSFLAT 2009*, Lisbon, Portugal (`http://www.eusflat.org/publications/proceedings/IFSA-EUSFLAT_2009/pdf/tema_1433.pdf`).

Körner, R. (2000). An asymptotic α-test for the expectation of random fuzzy variables. *J. Statist. Plann. Infer.* **83**, 331–346.

Körner, R., Näther, W. (2002). On the variance of random fuzzy variables. In: Bertoluzza, C., Gil, M.A., Ralescu, D.A. (Eds.) *Statistical Modeling, Analysis and Management of Fuzzy Data.* Physica-Verlag, Heidelberg, 22–39.

Kruse, R., Meyer, K. D. (1987). *Statistics with Vague Data.* Reidel Publ. Co., Dordrecht.

Kwakernaak, H. (1978). Fuzzy random variables. Part I: Definitions and Theorems. *Inform. Sci.* **15**, 1–29.

Lubiano, M.A., Gil, M.A. (1999). Estimating the expected value of fuzzy random variables in random samplings from finite populations. *Stat. Pap.* **40**, 277–295.

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

Lubiano, M.A., Trutschnig, W. (2010). ANOVA for Fuzzy Random Variables Using the R-package SAFD. In Borgelt, C., Gonzlez-Rodrguez, G., Trutschnig, W., Lubiano, M.A., Gil, M.A., Grzegorzewski, P., Hryniewicz, O. (Eds.) *Combining Soft Computing and Statistical Methods in Data Analysis.* Springer, Heidelberg, 449–456.

Montenegro, M., Casals, M. R., Lubiano, M. A., Gil, M. A. (2001). Two-sample hypothesis tests of means of a fuzzy random variable. *Inform. Sci.* **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.

Näther, W. (2006). Regression with fuzzy random data. *Comp. Statist. Data Anal.* **51**, 235–252.

Nguyen, H.T. (1978). A note on the extension principle for fuzzy sets. *J. Math. Anal. Appl.* **64**, 369–380.

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

Puri, M. L., Ralescu, D. A. (1986). Fuzzy random variables. *J. Math. Anal. Appl.* **114**, 409–422.

Ramos-Guajardo, A.B. (2011). Contrastes de hipótesis: tratamiento de la variabilidad y la imprecisión. PhD Thesis. University of Oviedo.

Ramos-Guajardo, A.B., Lubiano, M.A. (2011). K-sample tests for equality of variances of random fuzzy sets. *Comp. Statist. Data Anal.* (in press, doi:10.1016/j.csda.2010.11.025)

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 *Inform. Sci.* **179**, 3964–3972.

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

ABSTRACT

Data obtained from the sampling of a random experiment are often supposed to be either numerical or vectorial and exactly known/perceived. However, this assumption does not fit some practical situations, especially those concerning judgments, perceptions or ratings involving imprecision and subjectivity.

Data of this type are usually modeled and treated as categorical/qualitative ones, and statistical techniques dealing with them are in some senses rather limited. Many of these data could be alternatively modeled and handled in a suitable way as fuzzy-valued ones. This alternative approach enables on one hand capturing the underlying imprecision, subjectivity and diversity/variability and, on the other hand, stating distances between data with a meaning similar to that for numerical/vectorial ones.

By using the notion of random fuzzy sets (or fuzzy random variables in Puri and Ralescu's sense) a distance-based statistical methodology can be developed; actually, several techniques have been already formalized mainly in connection with the fuzzy-valued mean values of random fuzzy sets. The added value of this methodology lies in the fact that it exploits the information associated with the imprecision, subjectivity and diversity/variability captured by the use of the fuzzy scale. Furthermore, combining this methodology with the so-called characterizing fuzzy representation of a real-valued random variable will provide us with an appealing approach to develop inferential statistics on the distribution of random variables.