Summary - The notion of Fuzzy Random Variable has been introduced to model random mechanisms generating imprecisely-valued data which can be properly described by means of fuzzy sets. Probabilistic aspects of these random elements have been deeply discussed in the literature. However, statistical analysis of fuzzy random variables has not received so much attention, in spite that implications of this analysis range over many fields, including Medicine, Sociology, Economics, and so on. A summary of the fundamentals of fuzzy random variables is presented. Then, some related “parameters” associated with the distribution of these variables are defined. Inferential procedures concerning these “parameters” are described. Some recent results related to linear models for fuzzy data are finally reviewed.

Key Words - Fuzzy data; Estimation; Hypothesis testing; Informational paradigm; Distances between fuzzy data; Regression analysis; ANOVA for fuzzy data.

1. INTRODUCTION

Statistical reasoning is a specific and relevant instance of approximate reasoning, under uncertainty. In particular, it refers to the analysis of collective phenomena, namely phenomena which are defined with reference to a collection of empirical observations. On each observation one or more variables are measured in various different ways. This constitutes the “empirical information” \( \mathcal{I}_E \) which is a basic component of the reasoning process in Statistics. The other component is the “theoretical information” \( \mathcal{I}_T \), namely the background theory, the assumptions and the models which allow us to process \( \mathcal{I}_E \) in order to draw the conclusions of the reasoning procedure, which represent the “informational gain” of this piece of knowledge acquisition based on the couple \( (\mathcal{I}_E, \mathcal{I}_T) \). This way of looking at statistical methodology has been called “Informational Paradigm” (Coppi, 2002). The advantage in using this perspective lies in the fact that it allows us to widen the spectrum of statistical theory and applications, considering different kinds of “informational ingredients” (either empirical or theoretical ones) as input of the statistical reasoning process.
In order to illustrate this point, we must recall that some kind of uncertainty generally affects the above mentioned reasoning process and, specifically, its informational ingredients (Coppi, 2007). As a matter of fact, the observed data may be affected by various types of uncertainty concerning:

(i) the measuring system;
(ii) the way of expressing the assessment of their measure (e.g., numerically, linguistically, by means of numerical intervals, etc.);
(iii) the way they are eventually selected from a larger “population”;
(iv) the possible “vagueness” in defining the underlying concepts through the use of observable variables (e.g., measuring an opinion by means of a visual analogue scale; see for instance, Coppi et al., 2006a).

Likewise, some theoretical informational ingredients can be affected by uncertainty. For example, we may doubt of the Gaussian assumption about the distribution of a given quantitative variable, utilized as a means for drawing inferences from an observed sample (the adoption of a Bayesian framework for coping with this type of uncertainty is often suggested in the literature; beside the use of robust procedures, nonparametric inference or resampling techniques).

A different type of uncertainty may concern, for instance, the assumption of a regression model linking the “response” variable with the “explanatory” ones. In classical inferential statistics, a “true”, albeit unknown, regression model is assumed, characterized by a vector of (unknown) precisely measured regression coefficients. This “crisp” assumption may be relaxed, by thinking of a “fuzzy” regression relationship, whereby the unknown coefficients are fuzzy numbers to be estimated by means of appropriate procedures. In this case, the uncertainty related to the assumption of a specific regression model is dealt with through the fuzziness of the regression system (see, e.g., Tanaka et al., 1982).

Fuzziness may also be adopted as a tool for coping with some types of uncertainty affecting statistical data (e.g., in case (iv) of the above mentioned list, or in case of linguistic data as mentioned in situation (ii) before). On the other hand, randomness managed by suitable probability models constitutes the usual tool for dealing with data uncertainty due to sampling from finite or infinite populations (type (iii) of the above list), or for processing subjective uncertainty concerning theoretical assumptions (such as the sampling distribution in a statistical set up, or the parameters of a given statistical model), in a Bayesian framework.

Randomness and fuzziness may act separately or jointly on the various informational ingredients of a statistical reasoning process (see Zadeh, 1995, for considerations on the simultaneous use of probability and fuzziness). From a mathematical viewpoint they are respectively managed by means of probability
Statistics with fuzzy random variables

and fuzzy sets theories. Due to the uncertainty of the ingredients, also the con-
cclusions of the process are uncertain. The latter uncertainty must be expressed
as a result of the propagation of the various types of involved uncertainties
through the given system of statistical reasoning. This opens new problems in
statistical methodology, as compared with traditional inferential statistics, where
the whole of uncertainty in the system was expressed in probabilistic terms (see
Coppi et al., 2006c and, for recent thorough discussions, the Special Issue of
the journal Computational Statistics and Data Analysis, 2006, Coppi, Gil and
Kiers, Eds., devoted to “The Fuzzy Approach to Statistical Analysis” and the
Special Issue of the journal Fuzzy Sets and Systems, 2006, Gil, López-Díaz and

In the present work, we will focus in particular on the case where the
observed data are jointly affected by two sources of uncertainty: fuzziness (due
to imprecision, vagueness, partial ignorance) and randomness (due to sampling
or measurement errors of stochastic nature). In this framework we will show
the theoretical and applicative potentialities of a specific tool developed for
coping with this type of complex uncertainty: the notion of “Fuzzy Random
Variable” in Puri and Ralescu’s sense (1986).

In collecting statistical data one can frequently come across an underlying
imprecision due to vagueness of the involved concepts/judgements/perceptions.
In many cases, such an imprecision can be modelled by means of fuzzy sets
in a more efficient way than considering only a single value or category.

In this way, judgements like “excellent”, “good”, “fair”, “bad”, and
“very bad”, or “risky”, “moderately safe”, and “quite safe”, or “low”,
“medium”, and “high”, etc. are data that can be frequently encountered in
many real-life situations in which randomness is involved in obtaining data.
Most of these labels, which are essentially imprecise, can be suitably modelled
by means of fuzzy sets of the space of real numbers. The methodology to
be illustrated in this paper will allow us to manage these values/categories by
exploiting all the information contained in their “meaning”, instead of only
considering whether these values are or not different or whether they occupy
different positions in a ranking (as it is usually done in traditional statistics
with categorical and ordinal data). The full exploitation of the information in
imprecise values is achieved in this approach through the use of convenient
distances between fuzzy sets.

When data are examined for statistical purposes, probabilistic models are
considered to formalize both the setting underlying the problem and the mech-
anism generating the data. For real/vectorial data, random variables/vectors
model such a mechanism, whereas for the case of interval/set-valued data, this
model is stated by random intervals/sets. The mechanisms generating essen-
tially fuzzy-valued data have been properly modelled through the concept of
fuzzy random variables (also referred to in the literature as random fuzzy sets).
It should be emphasized that there are other approaches to model fuzzy data in a random context. In fact, the concept of fuzzy random variable was first introduced by Kwakernaak (1978, 1979), and later formalized in a slightly different way by Kruse and Meyer (1987). Although mathematical conditions in the model stated by Puri and Ralescu and in that by Kwakernaak/Kruse and Meyer coincide for some relevant cases, the situations to be modelled essentially differ. Thus, fuzzy random variables in Kwakernaak/Kruse and Meyer’s sense formalize either fuzzy perceptions or fuzzy descriptions of existing real-valued data generated by a random mechanism (the so-called “original random variable”), and most of the statistical analysis refers to parameters and characteristics of the distribution of this original. Instead, fuzzy random variables in Puri and Ralescu’s sense were conceived to formalize random mechanisms which directly assign fuzzy judgements/labels, with no underlying original real-valued process behind. Of course, results and statistical methods using this last formal notion can be applied for the first one, but aim and scope frequently differ: the interest in Puri and Ralescu’s approach will be usually focused on the parameters/characteristics (sometimes fuzzy) of the distribution of the fuzzy random mechanism.

This paper aims at reviewing the key statistical developments with fuzzy random variables, so far available in the literature, adopting in particular the Puri and Ralescu’s approach. In Section 2 fuzzy data to be considered along the paper are formalized, and some of their basic mathematical properties are described with specific reference to the metrics. The concept of fuzzy random variable is stated in Section 3. In Section 4 some of the most relevant (fuzzy- and real-valued) parameters of the distribution of fuzzy random variables (namely, the expected value, the variance and the inequality) are recalled. Section 5 gathers some inferential procedures (estimation and hypothesis testing) concerning these parameters. Section 6 is devoted to linear models. More concretely, the problems of analysis of variance and regression involving fuzzy data are addressed. Finally, some concluding remarks are made in Section 7.

2. Formalizing fuzzy data

Fuzzy-valued data we will deal with in this paper will be those (often called fuzzy numbers) belonging to the class

\[ \mathcal{F}_c(\mathbb{R}) = \{ U \mid \mathbb{R} \to [0, 1] \mid U_\alpha \text{ is a compact interval for all } \alpha \in [0, 1] \} , \]

where \( U_\alpha \) denotes the \( \alpha \)-level of fuzzy set \( U \), that is,

\[ U_\alpha = \{ x \in \mathbb{R} \mid U(x) \geq \alpha \} \]
if $\alpha \in (0, 1]$, and

$$U_0 = \text{cl}(\text{supp } U) = \text{cl} \{ x \in \mathbb{R} \mid U(x) > 0 \},$$

and $U(x)$ represents the degree of compatibility of $x$ with the property defining $U$ (also degree of membership of $x$ in $U$, or degree of possibility of $x$ being $U$).

In this way each fuzzy datum $U \in \mathcal{F}_c(\mathbb{R})$ to be considered can be characterized by means of the family of compact intervals

$$\{ [\inf U_\alpha, \sup U_\alpha] \}_{\alpha \in [0, 1]}.$$ 

Alternatively it can be characterized by the family of vectors

$$\{ (\text{mid } U_\alpha, \text{spr } U_\alpha) \}_{\alpha \in [0, 1]}$$

where $\text{mid } U_\alpha = (\inf U_\alpha + \sup U_\alpha)/2$ and $\text{spr } U_\alpha = (\sup U_\alpha - \inf U_\alpha)/2$ are the center and the radius of the $\alpha$-level, respectively.

In handling fuzzy data for probabilistic/statistical purposes the basic operations are the sum and the multiplication by scalars. These operations are usually assumed to be based on Zadeh's extension principle (1965, 1975). The application of this principle for elements in $\mathcal{F}_c(\mathbb{R})$ can be proven to satisfy, for all $\alpha \in [0, 1]:$

$$(U + V)_\alpha = \{ u + v \mid u \in U_\alpha, \ v \in V_\alpha \} = [\inf U_\alpha + \inf V_\alpha, \sup U_\alpha + \sup V_\alpha],$$

$$(\lambda \ U)_\alpha = \{ \lambda u \mid u \in U_\alpha \} = \begin{cases} [\lambda \inf U_\alpha, \lambda \sup U_\alpha] & \text{if } \lambda \geq 0 \\ [\lambda \sup U_\alpha, \lambda \inf U_\alpha] & \text{otherwise} \end{cases}$$

whatever $U, V \in \mathcal{F}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$, whence the simple arithmetic with fuzzy numbers coincides level-wise with the set-valued arithmetic.

With the above-mentioned operations $\mathcal{F}_c(\mathbb{R})$ is endowed with a semilinear structure, the absence of linearity being associated with the fact that there is no inverse for the fuzzy addition. When necessary, the Hukuhara difference will be considered. In this sense, $C$ is said to be the Hukuhara difference between $A$ and $B \in \mathcal{F}_c(\mathbb{R})$ if $A = B + C$ (that is, $C_\alpha = (A - H B)_\alpha = [\inf A_\alpha - \inf B_\alpha, \sup A_\alpha - \sup B_\alpha]$ provided that $\inf A_\alpha - \inf B_\alpha \leq \sup A_\alpha - \sup B_\alpha$ for all $\alpha \in [0, 1]$).

Since fuzzy data correspond to $[0, 1]$-valued mappings on $\mathbb{R}$, they are in fact functional data with a very intuitive interpretation. However, whereas functional data are often assumed to be Hilbert space-valued, the space of fuzzy data with the usual arithmetic above is not linear (although it can be “identified” with a cone of a Hilbert space by means of the support function, see Klement et al., 1986). As a consequence, in the development of many
providing probabilistic and statistical results one should take special care to guarantee that involved operations do not lead to elements out of $\mathcal{F}_c(\mathbb{R})$. For this reason, the case of fuzzy data often requires a specialized study.

Many distances between fuzzy data can be found in the literature (see, for instance, Puri and Ralescu, 1986, Klement et al., 1986 or Diamond and Kloeden, 1994). In order to analyze probabilistic and statistical aspects of fuzzy random variables, a metric which has been shown to be very convenient and easy to interpret (see, for instance, Lubiano et al., 2000) is the one introduced by Bertoluzza et al. (1995). For $U, V \in \mathcal{F}_c(\mathbb{R})$ the $D_W^\varphi$ distance between $U$ and $V$ is given by

$$D_W^\varphi(U, V) = \sqrt{\int_{[0,1]} \int_{[0,1]} \left( f_U(\alpha, \lambda) - f_V(\alpha, \lambda) \right)^2 dW(\lambda) d\varphi(\alpha)},$$

where $f_U(\alpha, \lambda) = \lambda \sup U_\alpha + (1 - \lambda) \inf U_\alpha$, and $W$ and $\varphi$ are weighting measures which can be identified with probability measures on the measurable space $([0,1], \mathcal{B}[0,1])$: $W$ being associated with a non-degenerate distribution, whereas $\varphi$ has a distribution function which is strictly increasing on $[0,1]$. Conditions for $W$ and $\varphi$ are imposed to guarantee that $D_W^\varphi$ is in fact a metric, but it should be noted that the associated weights have not a stochastic meaning.

In the preceding distance, for each $\alpha \in [0,1]$ a bijection is considered between the corresponding convex linear combinations of the extreme values of each $\alpha$-level, and $D_W^\varphi$ quantifies an average of the (Euclidean) distances between the corresponding points. The way in which (in each level) the choice of $W$ affects the metric $D_W^\varphi$ has been discussed in Gil et al. (2002). On the other hand, the choice of $\varphi$ allows us to weight the importance (“imprecision” or “consensus” in describing data) of each $\alpha$-level, and some statistical implications of some choices have been shown in González-Rodríguez et al. (2006a).

The metric above has been extended by Näther (2000), on the basis of the support function in $\mathcal{F}_c(\mathbb{R})$, by obtaining the generic $L_2$ distance in the class in which $\mathcal{F}_c(\mathbb{R})$ is embedded through the support function (see Klement et al., 1986, Diamond and Kloeden, 1994). This metric has been considered for fuzzy sets of finitely-dimensional Euclidean spaces.

3. **Formalizing the random mechanisms generating fuzzy data:**

Fuzzy random variables, as intended in this section (Puri and Ralescu, 1986) have been considered in the setting of a random experiment to model an essentially fuzzy-valued mechanism, that is, a mechanism associating a fuzzy value with each experimental outcome. The model is stated as follows:
Definition 1. Given a probability space \((\Omega, \mathcal{A}, P)\), a mapping \(X : \Omega \to \mathcal{F}_c(\mathbb{R})\) is said to be a fuzzy random variable (FRV for short) if it is \(D^*_W\)-Borel-measurable, that is it is a measurable mapping w.r.t. the Borel \(\sigma\)-field generated on \(\mathcal{F}_c(\mathbb{R})\) by the topology associated with \(D^*_W\).

Remark 1. The concept in the preceding definition can be proven (see Diamond and Kleoden, 1994, Colubi et al., 2001, Körner and Näther, 2002) to be equivalent to that introduced by Puri and Ralescu (1986) as an extension of the notion of random interval (and, consequently, as that of random variable). In accordance with Puri and Ralescu’s formalization, an FRV is a mapping from \(\Omega\) to \(\mathcal{F}_c(\mathbb{R})\) such that the set-valued \(\alpha\)-level mappings, \(X_\alpha\), defined so that 
\[
X_\alpha(\omega) = \left[ \chi(\omega) \right]_{\alpha},
\]
for all \(\omega \in \Omega\), are random sets. Puri and Ralescu developed the concept of FRV not only for fuzzy numbers but for spaces of more general fuzzy sets, so that FRVs also extend random sets and random vectors.

Remark 2. Mathematically the notion of FRV is also equivalent to the one given by Kruse and Meyer (1987), in accordance with which \(X : \Omega \to \mathcal{F}_c(\mathbb{R})\) is an FRV if and only if for all \(\alpha \in [0, 1]\) the real-valued mappings \(\inf X_\alpha\) and \(\sup X_\alpha\) are random variables. However, as it has been pointed out before, the situation which is modelled is conceptually different.

Remark 3. It should be emphasized that the formalization of an FRV as a Borel-measurable mapping, as stated in Definition 1, allows us immediately to refer to concepts like the induced probability distribution associated with the FRV, the independence of two or more FRVs, the identity of distribution of FRVs as the usual ones in Probability Theory for metric spaces (see, for instance, Billingsley, 1995). As a consequence, the notions of random sample from an FRV \(X\) (as a finite sequence of FRVs identically distributed as \(X\)) and simple random sample from \(X\) (as a finite sequence of FRVs being independent and identically distributed as \(X\)) can be directly considered.

4. Relevant “Parameters” of the Distribution of a Fuzzy Random Variable

Probabilistic and statistical studies for FRV’s usually concern fuzzy- and real-valued “parameters”. Parameters will be hereinafter applied to refer to “summary measures” of the (induced) distribution of the FRVs (instead of referring to those of the unknown real-valued original), and they extend either those from the set-valued case or other ones from the real-valued case.

The best known fuzzy parameter is the “fuzzy expected value” which is defined (Puri and Ralescu, 1986) as a fuzzy-valued measure of the “central tendency” of an FRV as follows:
Definition 2. Given a probability space \((\Omega, A, P)\), if \(X: \Omega \to \mathcal{F}_c(\mathbb{R})\) is an FRV such that the random variable \(\max\{|\inf X_0|, |\sup X_0|\}\) is integrable, then the (fuzzy) expected value of \(X\) corresponds to \(E(X) \in \mathcal{F}_c(\mathbb{R})\) such that

\[
(E(X))_\alpha = [E(\inf X_\alpha), E(\sup X_\alpha)]
\]

for all \(\alpha \in [0, 1]\).

Remark 4. The fuzzy expected value of \(X\), was introduced by Puri and Ralescu (in a more general space, as indicated in Remark 1) as the unique element in \(\mathcal{F}_c(\mathbb{R})\) such that for all \(\alpha \in [0, 1]\)

\[
(E(X))_\alpha = \text{Aumann's integral of the random set } X_\alpha
\]

\[
= \left\{ \int_{\Omega} X(\omega) \, dP(\omega) \mid X: \Omega \to \mathbb{R}, X \in L^1(\Omega, A, P), X \in X_\alpha \text{ a.s. } [P] \right\}.
\]

Properties of the fuzzy mean value can be found, for instance, in Puri and Ralescu (1986), Negoita and Ralescu (1987), Colubi et al. (1999). It should be remarked that the above definition of fuzzy expected value refers to an “abstract” population which is implicit in the random mechanism underlying the probability space. If the population is finite, the expectation operation is replaced by a simple ‘averaging’ over the observed fuzzy values. In fact, as an illustration of the above concept of fuzzy expected value, we will consider a sample of fuzzy observations as a finite population (notice that an analogous procedure will be sometimes adopted in the sequel, in order to practically illustrate some theoretical notions about FRVs. It should be underlined, in this connection, that the application of these procedures in this section is not meant as an estimation of the corresponding parameters of the underlying population).

Example 1. Clients of a Savings Bank are classified by the managers in accordance with their “degree of aversion to investment”. The labels assigned to clients, namely,

\[
\tilde{x}_1 = \text{‘very low degree of aversion’}, \quad \tilde{x}_2 = \text{‘low degree of aversion’}, \\
\tilde{x}_3 = \text{‘medium degree of aversion’}, \quad \tilde{x}_4 = \text{‘high degree of aversion’}, \\
\tilde{x}_5 = \text{‘almost total aversion’},
\]

have not been assigned on the basis of an underlying real-valued magnitude, but rather on the basis of subjective judgements/perceptions on the clients. As a consequence, the classification process can be viewed as a fuzzy random variable \(X\) which takes on five values, \(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{x}_4,\) and \(\tilde{x}_5\) which can be described, for instance, in terms of \(S\)-curves as those in Figure 1.
Figure 1. Values of the “degree of aversion to investment” of clients of a given Savings Bank.

\[ \tilde{x}_1 = \begin{cases} 
1 & \text{in } [0, 10] \\
1 - S(10, 20) & \text{in } [10, 15] \\
0 & \text{otherwise} 
\end{cases} \quad \tilde{x}_2 = \begin{cases} 
S(5, 20) & \text{in } [5, 20] \\
1 & \text{in } [20, 30] \\
1 - S(30, 40) & \text{in } [30, 40] \\
0 & \text{otherwise} 
\end{cases} 

\]

\[ \tilde{x}_3 = \begin{cases} 
S(30, 40) & \text{in } [30, 40] \\
1 & \text{in } [40, 60] \\
1 - S(60, 70) & \text{in } [60, 70] \\
0 & \text{otherwise} 
\end{cases} \quad \tilde{x}_4 = \begin{cases} 
S(65, 75) & \text{in } [65, 75] \\
1 & \text{in } [75, 85] \\
1 - S(85, 100) & \text{in } [85, 100] \\
0 & \text{otherwise} 
\end{cases} 

\]

\[ \tilde{x}_5 = \begin{cases} 
S(80, 100) & \text{in } [80, 100] \\
0 & \text{otherwise} 
\end{cases} \]
with

\[ S(a, b)(t) = \begin{cases} 
0 & \text{if } t \leq a \\
2 \left( \frac{t - a}{b - a} \right)^2 & \text{if } t \in \left[ a, \frac{a + b}{2} \right] \\
1 - 2 \left( \frac{t - b}{b - a} \right)^2 & \text{if } t \in \left[ \frac{a + b}{2}, b \right] \\
1 & \text{otherwise,} 
\end{cases} \]

Bank managers are interested in the “mean degrees of aversion” in the three main offices of the Savings Bank in a given area, that will be denoted by \( \Omega_1 \), \( \Omega_2 \) and \( \Omega_3 \). For this purpose, they consider an overall sample of \( n = 122 \) clients, and observe fuzzy random variable \( X \) on the sub-samples of sizes \( n_1 = 33 \), \( n_2 = 44 \) and \( n_3 = 45 \). In this example, for illustrative purposes, the sample data will be considered as finite populations with the following distributions:

<table>
<thead>
<tr>
<th>( \tilde{x}_1 )</th>
<th>( \tilde{x}_2 )</th>
<th>( \tilde{x}_3 )</th>
<th>( \tilde{x}_4 )</th>
<th>( \tilde{x}_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Omega_1 )</td>
<td>.15</td>
<td>.24</td>
<td>.36</td>
<td>.15</td>
</tr>
<tr>
<td>( \Omega_2 )</td>
<td>.14</td>
<td>.25</td>
<td>.34</td>
<td>.18</td>
</tr>
<tr>
<td>( \Omega_3 )</td>
<td>.11</td>
<td>.20</td>
<td>.27</td>
<td>.29</td>
</tr>
</tbody>
</table>

As an illustration of the idea of fuzzy mean we can see now represented in Figure 2 those corresponding to the distributions in Table 1 (which can be understood as multinomial distributions with fuzzy values). In accordance with the graphics for these mean values in Figure 2, we can conclude that

1. the mean risk aversion to investment corresponding to the sample from \( \Omega_1 \) (represented by means of a continuous curve ———) could be interpreted as to be ‘RATHER LOW TO SLIGHTLY MODERATE’;
2. the mean risk aversion to investment corresponding to the sample from \( \Omega_2 \) (represented by means of a dash-dot curve −··−) could be interpreted as to be ‘SLIGHTLY LOW TO RATHER MODERATE’;
3. the mean risk aversion to investment corresponding to the sample from \( \Omega_3 \) (represented by means of a dashed curve −−−) could be interpreted as to be ‘MODERATE TO RATHER HIGH’.
Other useful parameters are those quantifying the absolute/relative variation of an FRV (see, for instance, Lubiano et al., 2000, Näther, 2000, 2001, Körner and Näther, 2002), when it is intended as a measure of how values differ among them in terms of distances/ratios (respectively). Since variation summary measures are usually employed to compare variables, populations, estimators, and so on, they are often stated to be real-valued.

A real-valued measure for the “absolute variation” of an FRV can be obtained, for instance, by expressing how much “in error” this number is expected to be as a description of variable values. This error could be quantified in a natural way as follows:

**Definition 3.** Let $\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$ be a fuzzy random variable so that $[\max\{|\inf \mathcal{X}_0|, |\sup \mathcal{X}_0|\}]^2$ is integrable, then, the **variance of $\mathcal{X}$** is given by the value

$$\text{Var}(\mathcal{X}) = E \left( \left[ D_w^\phi (\mathcal{X}, E(\mathcal{X})) \right]^2 \right) = \int_{\Omega} [D_w^\phi (\mathcal{X}(\omega), E(\mathcal{X}))]^2 dP(\omega).$$

This variation measure agrees with the Fréchet approach (Fréchet, 1948) and verifies the usual properties of the variance (see Lubiano et al., 2000, Näther, 2000, 2001, Körner and Näther, 2002).
Example 2. For the distributions in Example 1 the absolute variation of $X$ in the different finite populations can be quantified (by considering $W = \varphi = \text{Lebesgue measure}$). Thus, the variation is given by

$$\text{Var}(X|(\omega_{1,1}, \ldots, \omega_{1,33})) = 776.585,$$
$$\text{Var}(X|(\omega_{2,1}, \ldots, \omega_{2,44})) = 791.930,$$
$$\text{Var}(X|(\omega_{3,1}, \ldots, \omega_{3,45})) = 883.148.$$

Since the scales for the fuzzy mean values are similar, absolute variations are comparable, whence we can conclude that clients from $\Omega_1$ and $\Omega_2$ show close variations, whereas (in the absolute sense) $X$ is slightly more variable over the clients from $\Omega_3$ than in the previous two ones.

A real-valued measure for the "relative variation" of an FRV can be obtained, for instance, by considering the "ratio" of its values with respect to the mean one, so it distinguishes between values being above and below the mean. Another distinctive characteristic of the relative variation, opposite to the absolute one, is that the "contribution" of variable values to the inequality associated with this variable is not necessarily nonnegative (more precisely, the contribution for values above the mean one is usually assumed to be negative, for values below the mean it is positive, and for values coinciding with the mean the contribution is null). The key problem in the case of fuzzy data will be now that of dividing and ranking them. In Lubiano et al. (2000) we can find a detailed justification of the arguments leading to the following.

Definition 4. Assume that the FRV $X : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$ satisfies the condition that $X_0(\omega) \subset (0, +\infty)$ for all $\omega \in \Omega$. Let $f : (0, +\infty) \rightarrow \mathbb{R}$ be a strictly convex (intended as convex downward) and monotonic function verifying $f(u) + f(1/u) \geq 0$ for all $u \in (0, +\infty)$ and $f(1) = 0$. The $f$-inequality index associated with $X$ is the value given by

$$I_f(X) = \frac{1}{2} \int_{[0,1]} \mathbb{E} \left[ f \left( \frac{\inf X_\alpha}{\mathbb{E}(\sup X_\alpha)} \right) + f \left( \frac{\sup X_\alpha}{\mathbb{E}(\inf X_\alpha)} \right) \right] d\alpha,$$

whenever all the above involved expectations exist.

Properties of the above measures can be found, for instance, in Lubiano et al. (2000), Lubiano and Gil (2002). Inequality measures of FRVs have been also expressed in terms of fuzzy-valued indices (see, for instance, Gil et al., 1998) which become useful when the main goal is not that of comparing either populations or fuzzy attributes but rather to give an idea of the magnitude of the inequality.

Example 3. Consider the variable "annual income", $X$, when the income is not exactly quantified but rather classified in accordance with the classification
Statistics with fuzzy random variables

adopted in some credit assessment systems (intended as a kind of perception on the annual income situation). Following Cox (1994), this variable can be viewed as a variable whose (fuzzy) values are \( \tilde{x}_1 = \text{"SOMEWHAHT HIGH"}, \tilde{x}_2 = \text{"MODERATELY HIGH"}, \tilde{x}_3 = \text{"HIGH"} \) and \( \tilde{x}_4 = \text{"VERY HIGH"} \), where \( \tilde{x}_1, \tilde{x}_2, \tilde{x}_3 \) and \( \tilde{x}_4 \) are described (in US thousand dollars) by means of the following \( S \)-curves:

\[
\tilde{x}_1 = 1 - S(100, 125)
\]

\[
\tilde{x}_2 = \begin{cases} 
S(100, 125) & \text{in } [100, 125] \\
1 - S(125, 150) & \text{in } [125, 150] \\
0 & \text{otherwise}
\end{cases}
\]

\[
\tilde{x}_3 = \begin{cases} 
S(125, 147.5) & \text{in } [125, 147.5] \\
1 - S(147.5, 170) & \text{in } [147.5, 170] \\
0 & \text{otherwise}
\end{cases}
\]

\[
\tilde{x}_4 = S(147.5, 170)
\]

and \( (\tilde{x}_i)_0 \subset [90, 180] \), \( i = 1, 2, 3, 4 \), for all candidates for a credit in the considered system (see Figure 3).

Assume that a bank adopting the above system wishes to compare two different towns by means of the income inequality, and to this purpose we observe the values of \( X \) in the main offices of these two towns.

Suppose there are 125 candidates for a certain type of credit in the main office of the first town (\( \Omega_1 \)) during a certain period and 178 candidates for the same type of credit and period in the office (\( \Omega_2 \)). The distributions of the classification of candidates are shown in Table 2. Then, if we consider the
\( f \)-inequality index with \( f(x) = -\log x \), we obtain that

\[
\begin{align*}
I_f(x_{\Omega_1}) &= .01224 \\
I_f(x_{\Omega_2}) &= .009,
\end{align*}
\]

whence we can conclude that there is more inequality of annual income in the office from the first town than in the second one.

\begin{table}[h]
\centering
\begin{tabular}{cccc}
\hline
 & \( \tilde{x}_1 \) & \( \tilde{x}_2 \) & \( \tilde{x}_3 \) & \( \tilde{x}_4 \) \\
\hline \hline
\( \Omega_1 \) & .224 & .344 & .248 & .184 \\
\( \Omega_2 \) & .354 & .444 & .152 & .050 \\
\hline
\end{tabular}
\caption{Distributions of the classification of candidates for a credit in two Offices of a Bank.}
\end{table}

5. Estimation/testing on relevant parameters associated with fuzzy random variables

The first procedures of inferential analysis of fuzzy data developed in the context of FRVs concern the estimation and hypothesis testing problems about the parameters of the associated parent populations. We illustrate in the sequel some results obtained in this framework.

5.1. Inferences on the (fuzzy) mean value of an FRV

Consider an FRV \( \mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R}) \) associated with the probability space \((\Omega, \mathcal{A}, P)\) and such that \( \max \{ |\inf \mathcal{X}_0|, |\sup \mathcal{X}_0| \} \) is integrable. Let \( \mathcal{X}_1, \ldots, \mathcal{X}_n \) be FRVs which are identically distributed as \( \mathcal{X} \) (i.e., let \((\mathcal{X}_1, \ldots, \mathcal{X}_n)\) be a random sample from \( \mathcal{X} \), see Remark 3). Then (see Lubiano and Gil, 1999, Lubiano et al., 1999b),

**Theorem 1.** The sample fuzzy mean value

\[
\overline{\mathcal{X}}_n = \frac{1}{n} (\mathcal{X}_1 + \ldots + \mathcal{X}_n)
\]

is an unbiased and consistent estimator of the fuzzy parameter \( E(\mathcal{X}) \); that is, the (fuzzy) mean of the fuzzy-valued estimator \( \overline{\mathcal{X}}_n \) over the space of all random samples equals \( E(\mathcal{X}) \) and \( \overline{\mathcal{X}}_n \) converges almost-surely to \( E(\mathcal{X}) \).
Example 4. Since the distributions in Example 1 correspond in fact to a random sample of clients of each office, now we will utilize the previous data as usual samples from the unknown populations \( \Omega_1, \Omega_2, \Omega_3 \), that is,

<table>
<thead>
<tr>
<th>Table 3: Sample data on the classification of clients of three Main Offices of a Savings Bank.</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tilde{x}_1 )</td>
</tr>
<tr>
<td>( \Omega_1 )</td>
</tr>
<tr>
<td>( \Omega_2 )</td>
</tr>
<tr>
<td>( \Omega_3 )</td>
</tr>
</tbody>
</table>

Thus, the (sample) mean values in Figure 2 are fuzzy estimates of the population mean values.

On the other hand, consider an FRV \( \mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R}) \) associated with the probability space \( (\Omega, \mathcal{A}, P) \) and such that \( \max \{ |\inf \mathcal{X}_0|, |\sup \mathcal{X}_0| \}^2 \) is integrable. Let \( \mathcal{X}_1, \ldots, \mathcal{X}_n \) be FRVs which are independent and identically distributed as \( \mathcal{X} \) (i.e., let \( \mathcal{X}_1, \ldots, \mathcal{X}_n \) be a simple random sample from \( \mathcal{X} \), see Remark 3). Let \( \mathcal{X}_1^*, \ldots, \mathcal{X}_n^* \) be a bootstrap sample obtained from \( \mathcal{X}_1, \ldots, \mathcal{X}_n \). Then (see Montenegro et al., 2004b, González-Rodríguez et al., 2006b, and also Körner, 2000, for a previous asymptotic study which is too complex to apply in practice), we have that

**Theorem 2.** In testing the null hypothesis \( H_0 : E(\mathcal{X}) = U \in \mathcal{F}_c(\mathbb{R}) \) at the nominal significance level \( \alpha \in [0, 1] \), \( H_0 \) should be rejected whenever

\[
\frac{\left[ D_W^\varphi(\overline{\mathcal{X}}_n, U) \right]^2}{\widehat{S}_{(W, \varphi)}^2} > z_\alpha ,
\]

where \( z_\alpha \) is the 100(1 - \( \alpha \)) fractile of the bootstrap distribution of

\[
T_n = \left[ D_W^\varphi(\overline{\mathcal{X}}_n, \overline{\mathcal{X}}_n) \right]^2 / \widehat{S}_{(W, \varphi)}^2
\]

with

\[
\overline{\mathcal{X}}_n = \frac{1}{n} (\mathcal{X}_1 + \ldots + \mathcal{X}_n), \quad \widehat{S}_{(W, \varphi)}^2 = \sum_{i=1}^n [D_W^\varphi(\mathcal{X}_i, \overline{\mathcal{X}}_n)]^2 / (n - 1),
\]

\[
\overline{\mathcal{X}}_n^* = \frac{1}{n} (\mathcal{X}_1^* + \ldots + \mathcal{X}_n^*), \quad \widehat{S}_{(W, \varphi)}^2 = \sum_{i=1}^n [D_W^\varphi(\mathcal{X}_i^*, \overline{\mathcal{X}}_n^*)]^2 / (n - 1).
\]
Example 5. Bank managers in Example 1 are interested in checking whether or not the “mean degree of aversion” of each office is ‘MEDIUM/HIGH’, where this value is assumed to be described by means of the fuzzy set in Figure 4.

By applying bootstrap techniques to test such a hypothesis on the basis of the available sample data, we get the p-values for $\Omega_1$, $\Omega_2$, and $\Omega_3$, which are respectively given by .015, .013 and .474. This means that, at significance level .05, the ‘mean degree of aversion’ in office $\Omega_3$ can be accepted to be ‘MEDIUM/HIGH’, whereas this is not sustainable in Offices $\Omega_1$ and $\Omega_2$.

Testing hypothesis can also be developed for the two-sample and multi-sample cases (see Montenegro et al., 2001, and Montenegro et al., 2004a, Gil et al., 2006b, respectively). The multi-sample one will be later explained in the setting of linear models.

5.2. Inferences on the variance

Consider an FRV $\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$ associated with the probability space $(\Omega, \mathcal{A}, P)$ and such that $\left[\max\{|\inf\mathcal{X}_0|, |\sup\mathcal{X}_0|\}\right]^2$ is integrable. Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be FRVs which are independent and identically distributed as $\mathcal{X}$ (see Remark 3). Then (see Lubiano and Gil, 1999),

![Figure 4. Fuzzy value 'MEDIUM/HIGH'.](image-url)
Theorem 3. The corrected sample variance

\[ \hat{S}_{(W,\varphi)}^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} \left[ D_W^\varphi (X_i, \bar{X}_n) \right]^2 \]

is an unbiased and consistent estimator of \( \text{Var}(\mathcal{X}) \); that is, the mean of the real-valued estimator \( \hat{S}_{(W,\varphi)}^2 \) over the space of all random samples equals \( \text{Var}(\mathcal{X}) \) and \( \hat{S}_{(W,\varphi)}^2 \) converges almost-surely to \( \text{Var}(\mathcal{X}) \).

Example 6. Following the same reasoning as in Example 4, from the above results we conclude that unbiased estimates of the population variances of \( \mathcal{X} \) in \( \Omega_1, \Omega_2 \) and \( \Omega_3 \) are, respectively, given by 800.853, 810.347, and 903.216.

On the other hand, consider an FRV \( \mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R}) \) associated with the probability space \((\Omega, \mathcal{A}, P)\) and such that \( \left[ \max \{|\inf \mathcal{X}_0|, |\sup \mathcal{X}_0|\} \right]^4 \) is integrable. Let \( \mathcal{X}_1, \ldots, \mathcal{X}_n \) be FRVs which are independent and identically distributed as \( \mathcal{X} \), then by the asymptotic results in Lubiano et al. (1999a) involving \( D_W^\varphi \) metric, we have that

Theorem 4. In testing the null hypothesis \( H_0 : \text{Var}(\mathcal{X}) = \delta_0 \in \mathbb{R} \) at the nominal significance level \( \alpha \in [0, 1] \), \( H_0 \) should be rejected whenever

\[ \sqrt{n} \left( \frac{1}{n} \sum_{i=1}^{n} \left[ D_W^\varphi (X_i, \bar{X}_n) \right]^2 - \delta_0 \right) > z_\alpha, \]

where \( z_\alpha \) is the 100(1 - \( \alpha \)) fractile of an \( \mathcal{N}(0, 1) \) distribution.

By following the reasoning in Theorem 4, the corresponding one-sided tests are obtained as usual.

Example 7. Assume that Bank managers consider that an absolute variation in the “degree of aversion” lower than or equal to 800 can be admitted. Then, to test whether the variation in \( \Omega_1, \Omega_2 \), and \( \Omega_3 \) satisfies such a condition, the last test has been applied. The corresponding \( p \)-values are given by .57, .53, and .22, whence the hypothesis about the variability being admissible can be accepted for any of the three offices at the usual significance levels.
5.3. Inferences on the inequality

Consider an FRV $\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$ associated with the probability space $(\Omega, \mathcal{A}, P)$ and such that $\mathcal{X}_0(\omega) \subset (0, +\infty)$ for all $\omega \in \Omega$, and the real-valued function $[\max \{|\inf \mathcal{X}_0|, |\sup \mathcal{X}_0|\}]^2$ is integrable. Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be FRVs which are independent and identically distributed as $\mathcal{X}$. Then (see Lubiano and Gil, 2002),

**Theorem 5.** The corrected sample hyperbolic inequality index given by $\hat{I}_h(\mathcal{X})_n$ such that

$$\hat{I}_h(\mathcal{X})_n = \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} \int_{[0,1]} \left[ \frac{\inf(\mathcal{X}_j)}{\sup(\mathcal{X}_i)} + \frac{\sup(\mathcal{X}_j)}{\inf(\mathcal{X}_i)} \right] d\alpha$$

is an unbiased and consistent estimator of $I_h(\mathcal{X})$; that is, the mean of the real-valued estimator $\hat{I}_h(\mathcal{X})_n$ over the space of all random samples equals $I_h(\mathcal{X})$ and the sample hyperbolic inequality index converges almost-surely to $I_h(\mathcal{X})$.

On the other hand, consider an FRV $\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$ associated with the probability space $(\Omega, \mathcal{A}, P)$, such that $\mathcal{X}_0(\omega) \subset (0, +\infty)$ for all $\omega \in \Omega$, and satisfying conditions guaranteeing that $I_f(\mathcal{X})$ is well-defined. Let $\mathcal{X}_1, \ldots, \mathcal{X}_n$ be FRVs which are independent and identically distributed as $\mathcal{X}$, then by the asymptotic results in Lubiano et al. (1999a), we have that

**Theorem 6.** In testing the null hypothesis $H_0 : I_f(\mathcal{X}) = t_0 \in \mathbb{R}$ at the nominal significance level $\alpha \in [0, 1]$, $H_0$ should be rejected whenever

$$\sqrt{n} \left| \frac{1}{2n} \sum_{i=1}^{n} \int_{[0,1]} \left[ f \left( \frac{\inf(\mathcal{X}_i)}{\sup(\mathcal{X}_n)} \right) + f \left( \frac{\sup(\mathcal{X}_i)}{\inf(\mathcal{X}_n)} \right) \right] d\alpha - t_0 \right| > z_\alpha,$$

where $z_\alpha$ is the $100(1 - \alpha)$ fractile of an $N(0, 1)$ distribution, where

$$\hat{S}_{f,n}^2 = \frac{1}{4n} \sum_{i=1}^{n} \left( \int_{[0,1]} \left[ f \left( \frac{\inf(\mathcal{X}_i)}{\sup(\mathcal{X}_n)} \right) + f \left( \frac{\sup(\mathcal{X}_i)}{\inf(\mathcal{X}_n)} \right) \right] d\alpha \right)^2 - \frac{1}{n} \sum_{j=1}^{n} \int_{[0,1]} \left[ f \left( \frac{\inf(\mathcal{X}_j)}{\sup(\mathcal{X}_n)} \right) + f \left( \frac{\sup(\mathcal{X}_j)}{\inf(\mathcal{X}_n)} \right) \right] d\alpha.$$
6. Linear Models

As we have remarked above, the considered space of fuzzy sets is not linear. This fact makes the classical theory of Linear Models in Statistics rather complex or restrictive whenever the difference operator is needed. In spite of this, there are some studies in the literature involving this kind of models.

6.1. Analysis of variance with fuzzy data

Consider a factor which can act at $k$ possible different levels and having fixed effects, and a fuzzy random response variable determining $k$ independent populations. In this way, let $(\Omega_1, A_1, P_1), \ldots, (\Omega_k, A_k, P_k)$ be probability spaces and $\mathcal{X}_1 : \Omega_1 \to \mathcal{F}_c(\mathbb{R})$, $\ldots$, $\mathcal{X}_k : \Omega_k \to \mathcal{F}_c(\mathbb{R})$ fuzzy random variables associated with these spaces. From the $i$-th population, $\mathcal{X}_i$, we can generate a simple random sample, $\mathcal{X}_{i1}, \ldots, \mathcal{X}_{in_i}$ ($i = 1, \ldots, k$). Consider the linear model

$$\mathcal{X}_{ij} = A_i + \epsilon_{ij}$$

where $A_i \in \mathcal{F}_c(\mathbb{R})$ and the expected value of the fuzzy random errors $\epsilon_{ij}$ is equal to $B \in \mathcal{F}_c(\mathbb{R})$ for all $i = 1 \ldots k$ and $j = 1 \ldots n_i$.

The goal for this section is verifying whether the effect of the factor at the $k$ levels is the same, which is equivalent to testing the equality of the fuzzy means across the $k$ populations. The technique of analysis of variance (ratio between-group variation/within-group variation) will be considered.

Whatever $i \in \{1, \ldots, k\}$ may be, consider a simple random sample from $\mathcal{X}_i$: $\mathcal{X}_{i1}, \ldots, \mathcal{X}_{in_i}$. The sample fuzzy mean in the $i$-th group is given by

$$\overline{\mathcal{X}}_i = \frac{1}{n_i} (\mathcal{X}_{i1} + \ldots + \mathcal{X}_{in_i}) ,$$

and the total sample fuzzy mean is given by

$$\overline{\mathcal{X}} = \frac{1}{n} (\mathcal{X}_{11} + \ldots + \mathcal{X}_{1n_1} + \ldots + \mathcal{X}_{k1} + \ldots + \mathcal{X}_{kn_k}) = \frac{n_1}{n} \overline{\mathcal{X}}_1 + \ldots + \frac{n_k}{n} \overline{\mathcal{X}}_k$$

(with $n = n_1 + \ldots + n_k$).

To get bootstrap populations with a common fuzzy mean from the available sample information in this case, we can add to each sample the mean of the other samples. In other words, for each $i = 1, \ldots, k$ we will consider

$$\mathcal{X}^{*i} = \mathcal{X}_i + \overline{\mathcal{X}}_1 + \ldots + \overline{\mathcal{X}}_{(i-1)} + \overline{\mathcal{X}}_{(i+1)} + \ldots + \overline{\mathcal{X}}_k .$$
Then, we will resample from these new populations, that is, for any \( i \in \{1, \ldots, k\} \) we draw a large number of samples of \( n_i \) independent observations \( X_{i1}^*, \ldots, X_{in_i}^* \), from population \( X_i^* \). The test statistic will be given by

\[
T^*_{(n_1, \ldots, n_k)} = \frac{\sum_{i=1}^{k} n_i \left[ D_w^q(X_i^*, \bar{X}^*) \right]^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[ D_w^q(X_{ij}^*, \bar{X}_i^*) \right]^2},
\]

where for all \( i \in \{1, \ldots, k\} \)

\[
\bar{X}_i^* = \frac{1}{n_i} \left( X_{i1}^* + \ldots + X_{in_i}^* \right), \quad \bar{X}^* = \frac{n_1}{n} \bar{X}_1^* + \ldots + \frac{n_k}{n} \bar{X}_k^*.
\]

Under regularity conditions (see Gil et al., 2006b) it can be shown that

**Theorem 7.** To test at the nominal significance level \( \alpha \in [0, 1] \) the null hypothesis

\[
H_0 : E(X_1) = \ldots = E(X_k),
\]

**H_0 should be rejected if**

\[
T_{(n_1, \ldots, n_k)} = \frac{\sum_{j=1}^{k} n_i \left[ D_w^q(X_i^*, \bar{X}) \right]^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[ D_w^q(X_{ij}^*, \bar{X}_i) \right]^2} > z_\alpha,
\]

where \( z_\alpha \) is the 100(1 − \( \alpha \)) fractile of the distribution of \( T^*_{(n_1, \ldots, n_k)} \) (this distribution can be approximated by the MonteCarlo method).

Moreover, the probability of rejecting \( H_0 \) under the alternative hypothesis \( H_a \) (which assumes that there is not coincidence of all the fuzzy population means) converges to 1 as \( n_i \to \infty \) for all \( i = 1, \ldots, k \).

In practice, and especially when samples have quite different sizes, the replacement of the statistic in Theorem 7 and its bootstrap version by

\[
\frac{\sum_{i=1}^{k} n_i \left[ D_w^q(X_i^*, \bar{X}) \right]^2}{\sum_{i=1}^{k} \frac{1}{n_i^2} \sum_{j=1}^{n_i} \left[ D_w^q(X_{ij}^*, \bar{X}_i) \right]^2}
\]

usually leads to more stable results.
Example 8. Bank managers in Example 1 are interested in checking whether the “mean degree of aversion” is the same in the three offices. By applying the method in this section, a p-value of .302 is obtained. Thus, at the usual significance level, this hypothesis is accepted.

6.2. Regression analysis involving fuzzy data

The regression problem involving input or/and output fuzzy data has been widely studied in the literature. The viewpoint of many of these studies regards a fitting problem (see, for instance, Celminš, 1987, Diamond, 1988, Bandemer and Näther, 1992, Coppi and D’Urso, 2003, D’Urso, 2003, D’Urso and Santoro, 2006, Coppi et al., 2006b) and only few works assume an underlying stochastic model expressed through the fuzzy random variables (see Körner and Näther, 1998, Näther, 2000, Wünsche and Näther, 2002, Kratschmer, 2004 and Näther, 2006, Flórez et al., 2006, González-Rodríguez et al., 2007).

In a Least Squares setting, there are two fundamental approaches with a fuzzy response variable. The first approach regards crisp explanatory variables, whereas the second approach refers to fuzzy explanatory variables.

Concerning the first approach, FRVs $\gamma_1, \ldots, \gamma_n \in \mathcal{F}_c(\mathbb{R})$ are observed at a set of $m$-dimensional crisp design points $\{x_1, \ldots, x_n\} \subset \mathbb{R}^m$. The usual choices of weighting measures $W$ and $\varphi$ in linear models have been the discrete uniform in $\{0, 1\}$ and the continuous uniform in $[0, 1]$ (i.e., the Lebesgue measure), respectively.

The following result assures the existence and uniqueness of the least squares solution when a linear model is considered (see, for instance, Körner and Näther, 2002).

**Theorem 8.** There exists a unique $(\hat{B}_1, \ldots, \hat{B}_m) \in (\mathcal{F}_c(\mathbb{R}))^m$ so that

$$\sum_{i=1}^n \left[ D^\varphi_W \left( \gamma_i, \sum_{j=1}^m x_{ij} \hat{B}_j \right) \right]^2 = \inf_{(B_1, \ldots, B_m) \in (\mathcal{F}_c(\mathbb{R}))^m} \sum_{i=1}^n \left[ D^\varphi_W \left( \gamma_i, \sum_{j=1}^m x_{ij} B_j \right) \right]^2 .$$

In Körner and Näther (2002) explicit solutions are obtained when values from a wide class of fuzzy numbers (the so-called LR-fuzzy numbers) are considered. A wider discussion on this issue can be found in Näther (2006), where it is proved that under a stochastic linear regression model

$$E(Y_i) = \sum_{j=1}^m x_{ij} B_j, \quad i = 1, \ldots, n \quad \text{and} \quad (B_1, \ldots, B_m) \in (\mathcal{F}_c(\mathbb{R}))^m,$$

it is not possible to find in general a BLUE of the fuzzy parameters if $m \geq 2$. It should be noted that in the simple regression case, $m = 1$, for instance, the sample fuzzy mean is a BLUE.
As to the second approach, consider \((X, Y)\) to be a bidimensional FRV associated with a probability space \((\Omega, A, P)\) so that \(\text{Var}(X) < \infty\) and \(\text{Var}(Y) < \infty\). Wünsche and Näther (2002) proved that \(E(Y|X)\) is the best approximation of \(Y\) by a measurable function of \(X\) w.r.t. \(D_W^0\), which justifies calling \(E(Y|X)\) the regression function of \(Y\) w.r.t. \(X\).

Regarding the problem of finding the best approximation of \(Y\) by a linear function of \(X\), that is, search for \(\hat{a} \in \mathbb{R}\) and \(\hat{B} \in \mathcal{F}_c(\mathbb{R})\) so that\n
\[
E\left(\left[D_W^0(Y, \hat{a}X + \hat{B})\right]^2\right) = \inf_{a \in \mathbb{R}, B \in \mathcal{F}_c(\mathbb{R})} E\left(\left[D_W^0(Y, aX + B)\right]^2\right),
\]

partial solutions, provided that some involved Hukuhara differences exist, were obtained. In this respect, it should be noted that the lack of linear structure of the space of fuzzy sets makes it difficult to handle linear regression models.

Recently (see González-Rodríguez et al., 2007) solutions for the general regression problem above have been found. These solutions depend on different moments of the FRVs, and the expressions for the solutions often become quite complex.

To avoid the difficulties which can arise in handling Linear Models, in Flórez et al. (2006) a nonparametric regression model has been considered. This model is given by \(Y = g(X) + \varepsilon_X\), where \(g : \mathcal{F}_c(\mathbb{R}) \to \mathcal{F}_c(\mathbb{R})\) is any function and \(\varepsilon_X\) is a fuzzy random variable defined on \((\Omega, A, P)\). It should be noted that the arithmetic of fuzzy sets makes a regression model for which \(E(\varepsilon_X) = 1_{\{0\}}\) very restrictive. By means of the support function it is possible to take advantage of some recent results concerning regression models developed for functional data in this context. More concretely, in Flórez et al. (2006) a kernel estimator of \(m = g + B\) has been proposed, the estimator being based on the random sample \(\{(X_i, Y_i)\}_{i=1}^n\) obtained from \((X, Y)\), given by

\[
\hat{m}(\tilde{x}) = \frac{\sum_{i=1}^n Y_i K(h^{-1}D_W^0(\tilde{x}, X_i))}{\sum_{i=1}^n K(h^{-1}D_W^0(\tilde{x}, X_i))}
\]

where \(h\) is the bandwidth, and the kernel \(K\) satisfies the usual regularity conditions (see, for instance, Ferraty and Vieu, 2004). It should be noted that the choice of the kernel function is not very relevant whereas the choice of the bandwidth is decisive.
7. **Concluding remarks**

It must be underlined that the notion of fuzzy random variable (in the Puri and Ralescu’s sense) is not the only one capable to model and process fuzzy data. Among the alternative notions and approaches we can mention:

- the approach introduced by Tanaka *et al.* (1979), followed by different authors (like Gil *et al.*, 1998ab, 1989);
- the one based on the formalization of FRVs as fuzzy perceptions of real-valued ones (see, Kruse and Meyer, 1987, Gebhardt *et al.* 1998);
- the approach by Viertl and collaborators, mainly devoted to extend Bayesian methodology to deal with non-precise data (see, for instance, Viertl, 1996, 2006);
- the perspective based on the use of random fuzzy closed sets, generalizing the theory of random closed sets (see, e.g., Nguyen and Wu, 2006);
- the approaches by Hryniewicz and Grzegorzewski, mainly devoted to develop techniques of association, reliability, quality control and hypothesis testing with fuzzy data (see, for instance, Hryniewicz, 2006, Grzegorzewski, 2006);
- the one by Denoeux and others based on belief functions (see, for instance, Petit-Renaud and Denoeux, 2004).

On the other hand, reasons supporting the interest of fuzzy random variables are mainly as follows:

- they model essentially fuzzy-valued classification processes, mechanisms, and so on, which are usually associated with perceptions, value judgments, opinions, leading to imprecise assessments;
- they are well-formalized in the probabilistic setting, which allows that concepts and properties in analyzing real-valued data (like, for instance, *p*-values, power functions, etc.) make sense in this wider context, and several statistical techniques inspired by those for real-valued data can be extended/adapted to the fuzzy-valued case;
- the use of the scale of measurement $F_c(\mathbb{R})$ allows us to consider distances between data, whence many classical concepts and procedures (especially those based on the squared error) can be extended and applied for statistical purposes.
ACKNOWLEDGMENTS

The authors are grateful to the referees of the original version for their helpful suggestions and comments. The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grant MTM2006-07501 and the Italian Ministry of University and Research, Grant PRIN 2005. This financial support is gratefully acknowledged.

REFERENCES


ANA COLUBI
Departamento de Estadística
I.O. y D.M.
Universidad de Oviedo
33071 Oviedo, Spain
colubi@uniovi.es

RENATO COPPI
Dipartimento di Statistica
Probabilità e Statistiche Applicate
Sapienza Università di Roma
00185 Roma, Italy
renato.coppi@uniroma1.it

PIERPAOLO D’URSO
Dipartimento di Scienze Economiche,
Gestionali e Sociali
Università degli Studi del Molise
86100 Campobasso, Italy
durso@unimol.it
pierpaolo.durso@uniroma1.it

MARÍA ÁNGELES GIL
Departamento de Estadística
I.O. y D.M.
Universidad de Oviedo
33071 Oviedo, Spain
magil@uniovi.es