Exploratory Analysis of Random Variables Based on Fuzzifications

Colubi, A., González-Rodríguez. G., Lubiano, M.A. and Montenegro, M.[1]

Dpto. de Estadística e I.O. Universidad de Oviedo. 33007 Spain
{gil,colubi,lubiano,mmontenegro}@uniovi.es

In this paper we propose a new way of representing the distribution of a real random variable by means of the expected value of certain kinds of fuzzifications of the original variable. We will analyze the usefulness of this representation from a descriptive point of view. We will show that the graphical representation of the fuzzy expected value displays in a visible way relevant features of the original distribution, like the central tendency, the dispersion and the symmetry. The fuzzy representation is valuable for representing continuous or discrete distributions, thus, it can be employed both for representing population distributions and for exploratory data analysis.

1 Introduction

A family of fuzzy representations of real random variables has been proposed in [2]. Some of them were used to characterize the real distributions with inferential purposes. Some other ones capture visual information about the distributions by focusing mainly on the mean value and the variance, although these fuzzifications do not characterize the distribution and loose valuable information in descriptive analysis.

Actually, it seems quite complex to find a fuzzification in this family allowing to visualize properly any kind of distribution. However, on the basis on the same intuitive ideas, we find another family of fuzzification with valuable graphical properties. The fuzzy representation of a real random variable allows us to associate the distribution with the expected value of a fuzzy random variable. The one obtained in this paper will be referred to as *exploratory fuzzy expected value*.

The exploratory fuzzy expected value will allow to represent both continuous and discrete distribution, which leads to a double use. On one hand, population distributions will be graphically represented by displaying important features (mean value, variability, skewness, "density"). In this sense, it can be interpreted as a kind of "parametrical" density or distribution function.

On the other hand, it can be used with exploratory purposes. The aim of the exploratory and descriptive analysis is to gain understanding of data, which is one of the most important targets of the statistical analysis. Data visualization associated

A. Colubi et al.: *Exploratory Analysis of Random Variables Based on Fuzzifications*, Advances in Soft Computing **6**, 95–102 (2006)
www.springerlink.com

with the exploratory fuzzy expected value will allow to capture information about important features of the data, which will allow to formulate reasonable hypotheses that can later be checked using some of the inferential methods above-mentioned.

2 Preliminaries

Let $\mathscr{K}_c(\mathbb{R})$ be the class of the nonempty compact intervals of $\mathbb{R}$ and let $\mathscr{F}_c(\mathbb{R})$ be the class of the fuzzy subsets U of $\mathbb{R}$ such that the α-level sets $U_\alpha \in \mathscr{K}_c(\mathbb{R})$ for all $\alpha \in (0,1]$, where $U_\alpha = \{x \in \mathbb{R} \,|\, U(x) \geq \alpha\}$, and $U_0 = \mathrm{cl}\{x \in \mathbb{R} \,|\, U(x) > 0\}$. In this context, the *sendograph* of $U \in \mathscr{F}_c(\mathbb{R})$ is the region enclosed by U and the x-axis on U_0, and $\mathbf{A}(U)$ will denote the corresponding area.

The space $\mathscr{F}_c(\mathbb{R})$ can be endowed with a semilinear structure, induced by a sum and the product by a scalar, both based upon Zadeh's extension principle [4], in accordance with which the following properties can be derived $(U+V)_\alpha = U_\alpha + V_\alpha$ and $(\lambda U)_\alpha = \lambda U_\alpha$ for all $U,V \in \mathscr{F}_c(\mathbb{R})$, $\lambda \in \mathbb{R}$ and $\alpha \in [0,1]$.

Given a probability space $(\Omega, \mathscr{A}, P)$, a *fuzzy random variable* (FRV) associated with $(\Omega, \mathscr{A})$ is intended to be, in accordance with Puri and Ralescu [3], a mapping $\mathscr{X} : \Omega \to \mathscr{F}_c(\mathbb{R})$ such that for each $\alpha \in [0,1]$ the α-level mapping $\mathscr{X}_\alpha : \Omega \to \mathscr{K}_c(\mathbb{R})$, defined so that $\mathscr{X}_\alpha(\omega) = \big(\mathscr{X}(\omega)\big)_\alpha$ for all $\omega \in \Omega$, is a random set (that is, a Borel-measurable mapping w.r.t. the Borel σ-field generated by the topology associated with the well-known Hausdorff metric d_H on $\mathscr{K}(\mathbb{R})$). Alternatively, an FRV is an $\mathscr{F}_c(\mathbb{R})$-valued random element (i.e. a Borel-measurable mapping) when the Skorohod metric is considered on $\mathscr{F}_c(\mathbb{R})$ (see Colubi *et al.* [1]).

A fuzzy random variable $\mathscr{X} : \Omega \to \mathscr{F}_c(\mathbb{R})$ is said to be *integrably bounded* if and only if, $\max\{\,|\inf X_0|, |\sup X_0|\,\} \in L^1(\Omega, \mathscr{A}, P)$. If $\mathscr{X}$ is an integrably bounded fuzzy random variable, the *expected value (or mean)* of $\mathscr{X}$ is the unique $\widetilde{E}(\mathscr{X}) \in \mathscr{F}_c(\mathbb{R})$ such that $\big(\widetilde{E}(\mathscr{X})\big)_\alpha =$ Aumman's integral of the random set $\mathscr{X}_\alpha$ for all $\alpha \in [0,1]$ (see Puri and Ralescu [3]), that is,

$$\big(\widetilde{E}(\mathscr{X})\big)_\alpha = \big\{E(f) \,\big|\, f : \Omega \to \mathbb{R}, f \in L^1, f \in \mathscr{X}_\alpha \; a.s.\,[P]\big\}.$$

3 The Exploratory Fuzzy Representation

A *fuzzy representation of a random variable* transforms crisp data (variable values) into fuzzy sets (the associated FRV values). The representations in [2] are mappings $\gamma^C : \mathbb{R} \to \mathscr{F}_c(\mathbb{R})$ which transforms each value $x \in \mathbb{R}$ into the fuzzy number whose α-level sets are

$$\big(\gamma^C(x)\big)_\alpha = \Big[f_L(x) - (1-\alpha)^{1/h_L(x)}, f_R(x) + (1-\alpha)^{1/h_R(x)}\Big]$$

for all $\alpha \in [0,1]$, where $f_L : \mathbb{R} \to \mathbb{R}$, $f_R : \mathbb{R} \to \mathbb{R}$, $f_L(x) \leq f_R(x)$ for all $x \in \mathbb{R}$, and $h_L : \mathbb{R} \to (0,+\infty)$, $h_R : \mathbb{R} \to (0,+\infty)$ are continuous and bijective. By varying functions f_L, f_R, h_L and h_R it is possible to get representing fuzzy random variables whose expected value capture visual information about different parameters of the distribution, however it seems complex to show jointly the most important ones.

In order to overcome this inconveniency, we will consider a new family of fuzzifications based on the same idea, that is, in such a way that the fuzzy expected value of the transformed random element capture important information about the original distribution.

Let $f : [0,\infty) \to [0,1]$ be an injective function. We define the auxiliar functional $\gamma_f : \mathbb{R} \to \mathscr{F}_c(\mathbb{R})$ so that,

$$[\gamma_f(x)]_\alpha = \begin{cases} \left[0, x^2 + x^2\left(\dfrac{1-f(x)}{f(x)}\right)\left(\dfrac{f(x)-\alpha}{f(x)}\right)\right] & \text{if } 0 \leq \alpha \leq f(x) \\ \left[0, x^2\left(\dfrac{1-\alpha}{1-f(x)}\right)\right] & \text{if } f(x) < \alpha \leq 1 \end{cases} \tag{1}$$

for all $\alpha \in [0,1]$ and $x \in [0,\infty)$. Term $x^2(1-f(x))/f(x)$ has been defined to guarantee that the area of sendograph of $\gamma_f(x)$ is equal to x^2 (see Figure 1). This functional depends on the square values to make the variance visible in the exploratory fuzzy expected value.

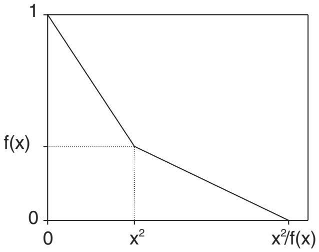

Fig. 1. Representation of the fuzzy set $\gamma_f(x)$

The family of *exploratory fuzzy representation* depends on a triple θ in a class

$$\Theta = \{(x_0, a, f)\,|\, x_0 \in \mathbb{R}, a \in \mathbb{R}^+, f : [0,\infty) \to [0,1] \text{ injective}\}$$

where x_0 will be a kind of ‘symmetry’ point, a a scale parameter and f the function above defined. Thus, if $\mathrm{sig}(x)$ is the sign of x, $\gamma^\theta : \mathbb{R} \to \mathscr{F}_c(\mathbb{R})$ is defined for $\theta = (x_0, a, f)$ so that

$$\gamma^\theta(x) = \mathbf{1}_{\{x\}} + \mathrm{sig}(x - x_0)\gamma_f\left(\left|\frac{x - x_0}{a}\right|\right)$$

for all $x \in \mathbb{R}$.

If $X : \Omega \to \mathbb{R}$ is a real-valued random variable so that $EX^2 < \infty$ and $f : [0,\infty) \to [0,1]$ is an injective function so that $(f(X))^{-1} \in L^1(\Omega, \mathscr{A}, P)$, then the *exploratory fuzzy expected value* is $\widetilde{E}(\gamma^\Theta \circ X)$. It should be noted that condition $(f(X))^{-1} \in$

$L^1(\Omega,\mathscr{A},P)$ is not restrictive, because functions like $f_p^\delta(x) = (p^x+\delta)/(1+\delta)$ with $p \in (0,1)$ and $\delta > 0$ for all $x \in \mathbb{R}$ verifies it irrespectively of X.

In this paper, we have considered $\theta_s = (EX, 1, f_{.6}^{.001}) \in \Theta$, which is a very simple and useful choice. Thus, the γ^{θ_s}-fuzzy representation of a random variable allows us to easily visualize features like the central tendency, variability, skewness, type of variable (discrete/continuous), and the existence of extreme values. More precisely, we can state that

If X is a random variable and

$$\gamma^{\theta_s} = \mathbf{1}_{\{x\}} + sig(x-EX)\gamma_{f_{.6}^{.001}}(|x-EX|)$$

for all $x \in \mathbb{R}$, where γ_f is defined as in (1) and

$$f_{.6}^{.001}(x) = \frac{.6^x+.001}{1.001},$$

then

i) $(\widetilde{E}(\gamma^{\theta_s} \circ X))_1 = \{EX\}$ *(that is, the* 1*-level set shows a* ***mean value*** *of* X*).*

ii) $\mathbf{A}(\widetilde{E}(\gamma^{\theta_s} \circ X)) = \mathrm{Var}(X)$ *(that is, the area of the sendograph shows the* ***variance*** *of* X*).*

iii) The symmetry of $\widetilde{E}(\gamma^{\theta_s} \circ X)$ *is connected with the symmetry of* X *around its mean value. The more skewness of* X *the more asymmetry of* $\widetilde{E}(\gamma^{\theta_s} \circ X)$*. Thus, the asymmetry of the exploratory fuzzy expected value shows the* ***skewness*** *of* X*.*

iv) If X *is a continuous variable, then* $\widetilde{E}(\gamma^{\theta_s} \circ X)$ *will be "smooth" (excepting at* EX*), whereas if it is discrete, the exploratory fuzzy expected value will show non-smooth changes of slope in each of the values* X *takes on (that is, the "smoothness" allows us to distinguish the* ***discrete*** *and* ***continuous*** *distributions).*

v) Large values of X *will be associated with large-spread* 0*-level sets (that is, thus the spread of the lower* α*-level sets can be useful to determine the presence of* ***extreme values****).*

In the following sections we will illustrate this properties by representing the exploratory fuzzy expected value of some relevant population/sample distributions.

4 Exploratory Analysis of Random Variables Through the Fuzzy Representation

In this Section the graphical representation of the exploratory fuzzification of different parametric distributions will be shown. Concretely, we will focus on the binomial, the poisson, the exponential, the normal and the χ^2 distribution. They have been chosen in order to show the different features of the exploratory fuzzy expected value that we have indicated in the preceding section. The distributions were approximated by Monte Carlo method on the basis of 100000 simulations.

In Figure 2 we show the exploratory fuzzy expected value of two random variables with binomial distributions. In both cases $n = 5$, but $p = .5$ at the left graphic

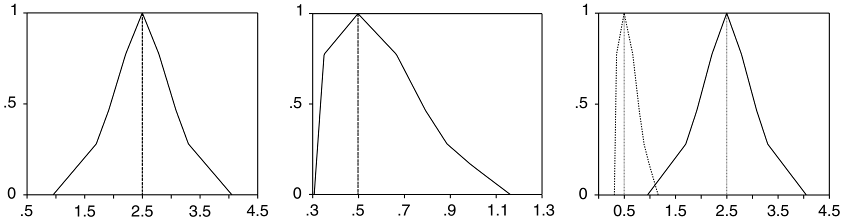

Fig. 2. Exploratory fuzzy expected value associated with $\mathscr{B}(5,.5)$ (left) and $\mathscr{B}(5,.1)$ (center) distributions. Comparison (right)

and $p = .1$ at the center one. We can see the respective mean values at the 1-level sets. The symmetry of the $\mathscr{B}(5,.5)$ and the skewness of the $\mathscr{B}(5,.1)$ is evident. It should be noted that the area of the sendograph shows the variance, although to make comparisons we have to take into account the range of the supports. The graphic on the right shows both fuzzy representations in the same scale. The difference in the areas, associated to the variabilities, is clear. If the aim were to compare the two distributions irrespectively of the variance, we could make use of the scale parameter a. The discrete character of the binomial distribution is connected with the lack of smoothness of the fuzzy sets and the right spread of the 0-level of the binomial $\mathscr{B}(5,.1)$ shows the presence of values far away from the mean.

In Figure 3, random variables with Poisson and exponential distributions, both with expected value equal to 4, are represented. The most remarkable difference is the large left-spreads with respect to the mean value of the exponential distribution, which indicates that in the exponential distribution the values lower than the mean have a greater density than in the Poisson distribution. In this case, since the Poisson is discrete but not finite, the lack of smoothness is less evident than for the binomial. We can also observe than the exponential distribution is considerably more asymmetric and variable than the Poisson.

The χ^2 distributed random variables were chosen with 1 and 2 degrees of freedom (see Figure 4). The left-spreads with respect to the mean values are more homogeneous than those for the Poisson and the exponential distributions, which indicates that the low values w.r.t. the corresponding expected value are relatively less fre-

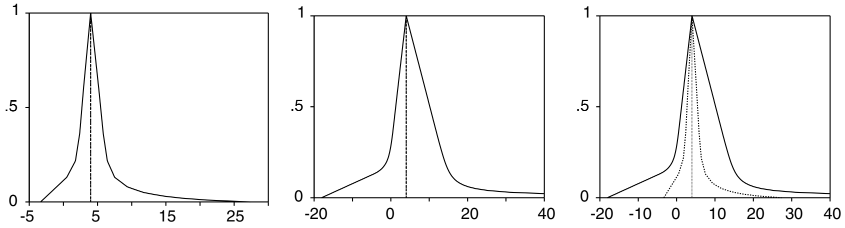

Fig. 3. Exploratory fuzzy expected value associated with $\mathscr{P}(4)$ (left) and Exp(.25) (center) distributions. Comparison (right)

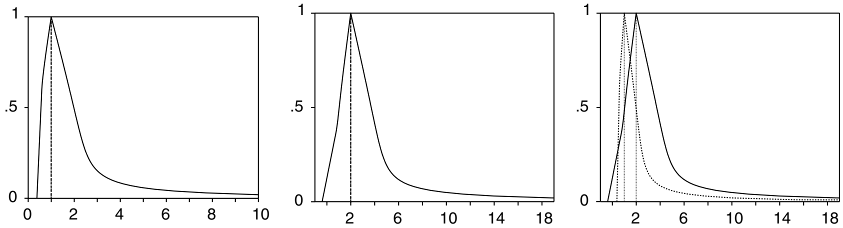

Fig. 4. Exploratory fuzzy expected value associated with χ_1^2 (left) and χ_2^2 (center) distributions. Comparison (right)

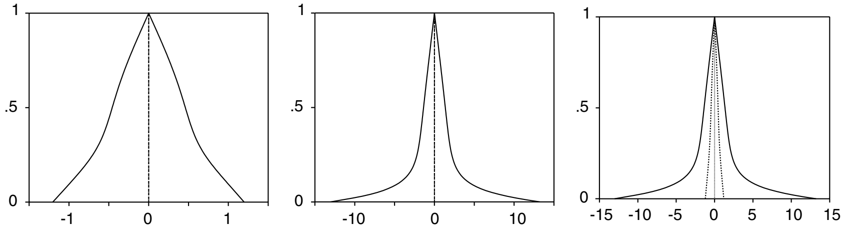

Fig. 5. Exploratory fuzzy expected value associated with $\mathcal{N}(0,1)$ (left) and $\mathcal{N}(0,2)$ (center) distributions. Comparison (right)

quent, mainly for the χ_1^2. The asymmetry of both distributions is evidenced and the greater variability of the χ_2^2 is easily noticed.

Finally, the exploratory fuzzy expected values corresponding to centered normal distributions with variances 1 and 4 are shown in Figure 5. The difference with the preceding distributions is obvious. As expected, the most similar shape to the standard normal distribution is the $\mathscr{B}(5,0.5)$, although we can see the difference in the smoothness of the curve. We observe the greater variability, the greater area and, in this case, the greater spreads for the 0-level.

5 Exploratory Data Analysis Through the Fuzzy Representation

When only data are available and the aim is to gain understanding of them, we can also make use of the graphical representation of the fuzzy mean. To illustrate it, we have simulated 4 samples with different sample sizes.

The exploratory fuzzy expected value associated with the first simulated samples are presented in Figure 6. We can observe the same features that we have commented in the preceding section. The distribution of sample 1 seems to be more skewed than the one in sample 2. The right spread of the 0-level in sample 1 seems to point out the presence of values quite greater than the mean. On the contrary, sample 2 seems to be quite symmetric around its mean. The sample sizes are quite low, although the clear lack of smoothness points out that there are repeated values, which indicates that they could come from discrete population distributions.

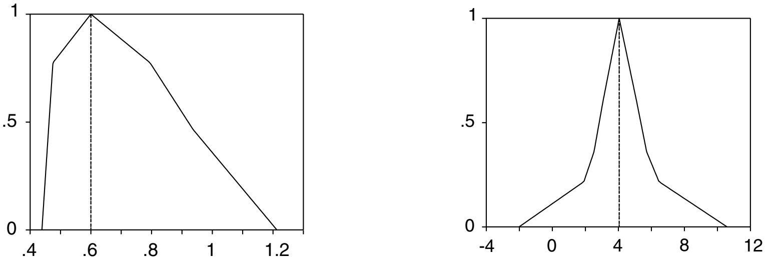

Fig. 6. Exploratory fuzzy expected value associated with sample 1 $n = 10$ (left) and sample 2 $n = 20$ (right)

In Figure 7 we present the graphical representation corresponding to the other simulated samples. We observe that the sample 4 is strongly asymmetric, with extreme values much greater than the sample mean, while sample 3 seems to be slightly asymmetric. The range of the supports suggests that the sample 3 is quite less variable than the sample 4. In this case the sample sizes are larger than in the preceding case, and the curves seems to be quite smooth, which suggests that the population distributions could be continuous.

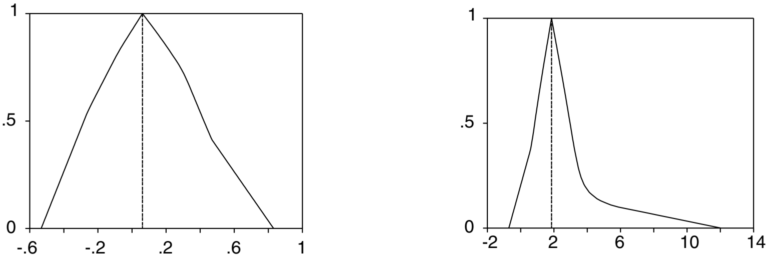

Fig. 7. Exploratory fuzzy expected value associated with sample 3 $n = 30$ (left) and sample 4 $n = 50$ (right)

Actually, sample 1 have been simulated form a $\mathscr{B}(5,.1)$, sample 2 from a $\mathscr{P}(4)$, sample 3 from a $\mathscr{N}(0,1)$ and sample 4 from a χ^2_2. If we compare the population distributions with the sample ones, we can note the similarities.

Acknowledgement

The research in this paper has been partially supported by the Spanish Ministry of Education and Science Grant MTM2005-00045. Its financial support is gratefully acknowledged.

References

[1] A. Colubi, J. S. Domínguez-Menchero, M. López-Díaz, and D. A. Ralescu. A $d_e[0,1]$-representation of random upper semicontinuous functions. *Proc. Amer. Math. Soc.*, 130:3237–3242, 2002.

[2] G. González-Rodríguez, A. Colubi, and M.A. Gil. A fuzzy representation of random variables: an operational tool in exploratory analysis and hypothesis testing. *Comput. Statist. Data Anal.*, 2006. (accepted, in press).

[3] M. L. Puri and D. A. Ralescu. Fuzzy random variables. *J. Math. Anal. Appl.*, 114:409–422, 1986.

[4] L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning, II. *Inform. Sci.*, 8:301–353, 1975.