

Fuzzy techniques in the analysis of distributions of real random variables

Gil González-Rodríguez



Statistical Methods with Imprecise Random Elements

European Centre for Soft Computing
Mieres, Spain

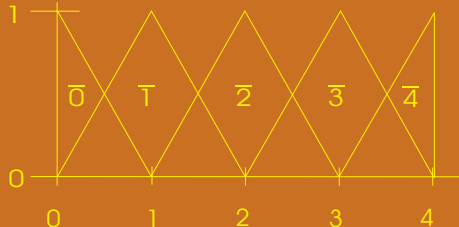
ERCIM WG Computing & Statistics 2008
Neuchâtel, June, 19-21 2008

Outline

- To **motivate** the use of fuzzy representations to handle random variables in certain cases and to analyze some interesting consequences
- To present some **fuzzy representations** of random variables (RVs) whose expected values **contains** different features of the original **distribution**
- To analyze **statistical/descriptive problems** concerning the distribution of RV **through the fuzzy expected value**
 - ✓ **Descriptive**: informative displays, easy identification of relevant parameter
 - ✓ **Probabilistic**: easy theoretical use
 - ✓ **Inferential**: sensitive and simple statistics for goodness-of-fit, equality of distributions and so on

A motivating example

- Forest fire risk index
 - ✓ Depending on certain variables (wind, temperature,...) a natural number from 0 (minimum risk) to 4 (maximum risk) is daily published
 - ✓ In practice, each natural number represents in some sense the “surrounding risks”
 - ⇒ values 0 and 4 are **different**
 - ✓ To represent these differences we can code the values as follows



Fuzzy representation of risk index

The **functions coding the values** represent membership functions of the **fuzzy sets** “the risk index is around i ” ($i = 0, \dots, 4$)

Usefulness?

- Some **inferences** on random variables like “forest fire risk index” can be developed by means of the preceding studies on fuzzy random variables after “fuzzifying” the original values
- This **triangular fuzzy representation** could be used to **code** other **finite variables** (say, Binomial ones)
- **Question:** is this a useful practice?

The space of fuzzy sets

$$\mathcal{F}_c(\mathbb{R}) = \{U : \mathbb{R} \rightarrow [0, 1] \mid U_\alpha \in \mathcal{K}_c(\mathbb{R}) \quad \forall \alpha \in [0, 1]\}$$

- $U_\alpha = \{x \in \mathbb{R} \mid U(x) \geq \alpha\}$ if $\alpha > 0$,
- $U_0 = \text{cl}\{x \in \mathbb{R} \mid U(x) > 0\}$
- $\mathcal{K}_c(\mathbb{R}) = \{\text{nonempty and bounded intervals of } \mathbb{R}\}$

Arithmetic on $\mathcal{F}_c(\mathbb{R})$

For all $U, V \in \mathcal{F}_c(\mathbb{R})$, $\lambda \in \mathbb{R}$ and $\alpha \in [0, 1]$

- $(U + V)_\alpha = \{u + v \mid u \in U_\alpha, v \in V_\alpha\}$
- $(\lambda \cdot U)_\alpha = \{\lambda u \mid u \in U_\alpha\}, \quad \forall \lambda \in \mathbb{R}$

Fuzzy Random Variables (FRV)

Given a probability space (Ω, \mathcal{A}, P) , an FRV $\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$ is a **Borel measurable** mapping (w.r.t. the usual metrics)

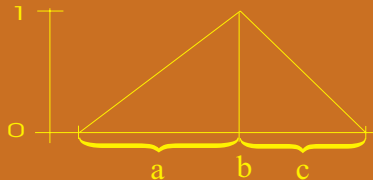
- \mathcal{X} is an FRV $\Leftrightarrow \inf \mathcal{X}_\alpha$ and $\sup \mathcal{X}_\alpha$ are random variables $\forall \alpha \in [0, 1]$

Mean value

- If $\sup_{x \in \mathcal{X}_0} \|x\| \in L^1$, the (fuzzy) **expected value** is the unique $\tilde{E}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R})$ s.t. for all $\alpha \in [0, 1]$
$$\left(\tilde{E}(\mathcal{X})\right)_\alpha = \text{Aumann's integral of } \mathcal{X}_\alpha = [E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)]$$

Triangular Fuzzy Representation

- Triangular fuzzy set: $Tri(a, b, c)$



- Let X be a finite random variable with mass function $\{(x_i, p_i)\}_{i=1}^k$.
- The **triangular fuzzy representation** of X will be the FRV $f(X)$ with mass function $\{(\tilde{x}_i, p_i)\}_{i=1}^k$, where
 - ✓ $\tilde{x}_1 = Tri(0, x_1, 1)$
 - ✓ $\tilde{x}_i = Tri(1, x_i, 1)$ if $1 < i < k$
 - ✓ $\tilde{x}_k = Tri(1, x_k, 0)$

Testing about the mean: empirical conclusions

- It is **not equivalent** to test

$$\begin{array}{l} H_0 : E(X) = a \\ H_1 : E(X) \neq a \end{array} \quad \text{and} \quad \begin{array}{l} H_0 : \tilde{E}(f(X)) = \tilde{a} \\ H_1 : \tilde{E}(f(X)) \neq \tilde{a} \end{array}$$

(with $\tilde{a} = Tri(1, a, 1)$)

- Empirically the second test has been shown to be **more powerful** in many cases
- The reason is that the expected value of the fuzzy representation contains **more information** than the expected value of the original one

Justification

- If $k \leq 4 \Rightarrow \tilde{E}(f(X))$ characterizes completely the distribution of X

✓ The distribution of X is $\{(x_i, p_i)\}_{i=1}^4$

$$\Leftrightarrow \tilde{E}(f(X)) = \sum_{i=1}^4 p_i \tilde{x}_i$$

- The result can be extended to characterize any real-valued distribution by means of other fuzzy representations

Family of *characterizing fuzzy representations* γ^C

- Let $\gamma^C : \mathbb{R} \rightarrow \mathcal{F}_c(\mathbb{R})$ be such that $\forall \alpha \in [0, 1]$:

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

- ✓ $f_L, f_R : \mathbb{R} \rightarrow \mathbb{R}$, $f_L(x) \leq f_R(x)$ for all $x \in \mathbb{R}$
- ✓ $h_L, h_R : \mathbb{R} \rightarrow (0, +\infty)$ are continuous and bijective
- The FRV $\gamma^C \circ X$ is called the γ^C -fuzzy representation of X .

Characterizing fuzzy representations (González-Rodríguez et al., 2006)

Given a p.s. (Ω, \mathcal{A}, P) , and X and Y associated with this space

$$\tilde{E}(\gamma^C \circ X) = \tilde{E}(\gamma^C \circ Y) \Leftrightarrow X \text{ and } Y \text{ are ID}$$

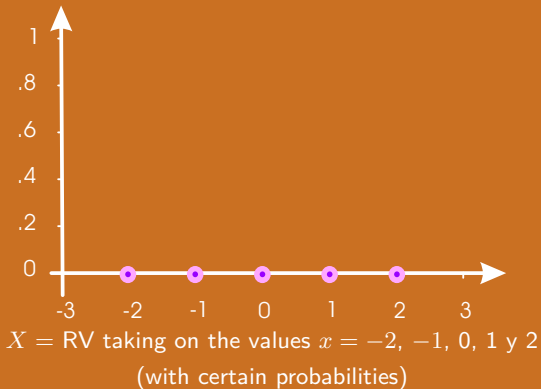
Remark

For each random variable X , the mapping $G_X^{\gamma^C} \in \mathcal{F}_c(\mathbb{R})$ given by

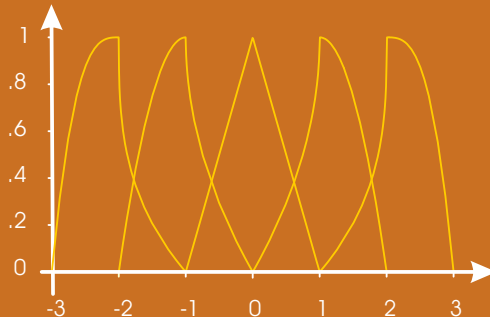
$$\begin{aligned} G_X^{\gamma^C} : \mathbb{R} &\rightarrow [0, 1] \\ t &\mapsto \tilde{E}(\gamma^C \circ X)(t) \end{aligned}$$

defines a **'characteristic' function** of the distribution of X .

A γ^C -fuzzy representation



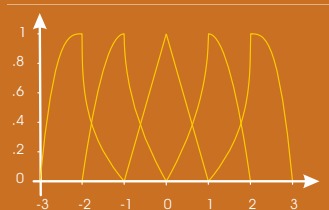
A γ^C -fuzzy representation



\mathcal{X} = FRV taking on the values $\gamma^C(x)$
(with the corresponding probabilities)

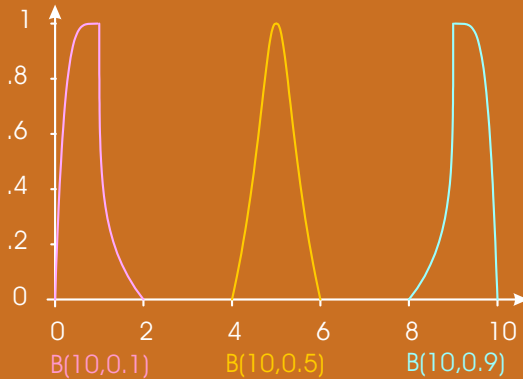
γ^C fuzzy representations

- **Modifications** of the triangular one in which different **degrees of curvature** for the infimum and supremum functions are allowed



- $\tilde{E}(\gamma^C \circ X)$ characterizes the distribution of X
 - ✓ the intuitive meaning is lost, but the fuzzification remains interesting as a tool
 - ✓ easy-to-use, good inferential results, general **but** non-informative enough display

Fuzzy expected value of γ^C for some distributions



D_W^φ -metric (Bertoluzza et al., 1995)

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

where $f_U(\alpha, \lambda) = \lambda \sup U_\alpha + (1 - \lambda) \inf U_\alpha$

- φ and W are certain normalized measures on $[0, 1]$
- φ allows to weight the importance of “distances concerning the X-axis, concerning the shape...”

Metric between distributions

The D_W^φ -metric becomes a **versatile distance between distributions** through the fuzzy representation

$$D_W^\varphi \left(\tilde{E}(\gamma^\theta \circ X), \tilde{E}(\gamma^\theta \circ Y) \right)$$

Applications

If X is an RV and $\gamma^C \circ X$ is a characterizing fuzzy representation, then

- An estimator of $\tilde{E}(\gamma^C \circ X)$ becomes an **estimator of the distribution** of X
- A one-sample test about $\tilde{E}(\gamma^C \circ X)$ becomes a **goodness-of-fit test** about the distribution of X
- A k -sample test about $\tilde{E}(\gamma^C \circ X_1), \dots, \tilde{E}(\gamma^C \circ X_k)$ becomes an **equality of distribution test** (*ANOVA of distributions*) about X_1, \dots, X_k

Remark

The preceding **bootstrap** hypothesis testing procedures become **tests about distributions based on tests about means** (which are usually quite **powerful**)

Hypothesis testing: empirical comparison of the power functions

- Quite good power results in comparison with
 - ✓ Kolmogorov-Smirnov, Cramer-Von Mises, χ^2 and likelihood ratio
- There is not a “uniformly best test”, but
 - ✓ if the mean values are relevant in order to detect differences between the hypothetical distribution and the simulated one
 - ★ D_W^φ should be more focused on ‘locations’
 $\varphi = \text{Lebesgue measure}$ is quite convenient
 - ✓ if the variability of the simulated distribution is much greater than that of the hypothetical one (i.e., the mean value is less relevant)
 - ★ D_W^φ should be more focused on ‘shapes’
 φ weighting more lower α -level sets, c.f. **Beta(1, 10)**

Situation

- The **forest fires** were especially severe in Asturias (Spain) in the nineties
- The policy concerning the protection against fire have changed since 2000
- A deep descriptive study **to compare** the forest fires before and after 2000 was carried out in order to verify whether or not the changes have been appropriate
- One of the variables in the study has been the **erosion risk** caused by the fire
- The data collected by the fire brigades from 1988 to 2004 were

	Before 2000	After 2000
Low	5519	6836
Medium	2646	4752
High	523	1221

Empirical conclusions

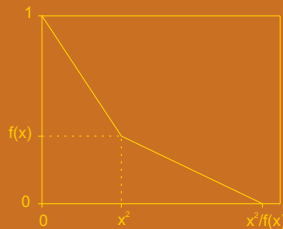
- The **aim** is to compare the erosion risk before and after 2000 from an inferential point of view
- The **values** of the erosion risk can be identified with its “order”, namely
 - ✓ 1=“low”
 - ✓ 2=“medium”
 - ✓ 3=“high”
- These categories could be represented by the **triangular fuzzy numbers** of the first representation

Empirical conclusions

- The **bootstrap two-sample test** for the equality of fuzzy means leads to a p -value=.000
- This can be interpreted in two ways
 - ✓ since the fuzzified data seem a suitable representation of the original categories, the test of equality of means allows to conclude that the **fuzzy mean** risk was different before and after 2000
 - ✓ since the number of values of the variable is $k = 3$, the preceding test is also a test of the *equality of distributions*, and from this point of view, it can be concluded that the **distribution** of the risk was different before and after 2000
- Thus, from both viewpoints a **significant difference** has been detected

Family of *exploratory fuzzy representations* γ^θ

- $\Theta = \{(x_0, a, f) \mid x_0 \in \mathbb{R}, a \in \mathbb{R}^+, f : [0, \infty) \rightarrow [0, 1] \text{ injective}\}$
 - ✓ f allows us to embed the positive part of X into $[0, 1]$ to define the auxiliary fuzzifying mechanism γ_f



- ✓ x_0 is a kind of ‘symmetry’ point
- ✓ a is a scale parameter

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

The γ^{θ_s} -fuzzy representation of a random variable

$$\theta_s = (EX, 1, f_{.6}^{.001}) \in \Theta.$$

If X is a random variable and

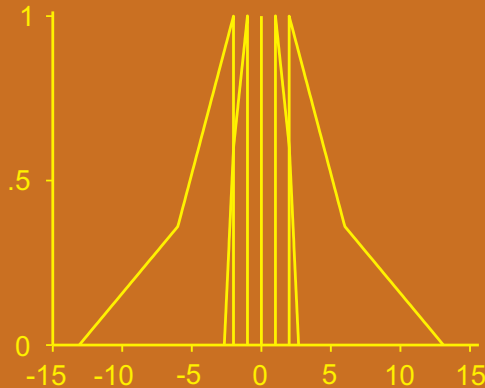
$$\gamma^{\theta_s}(x) = \mathbf{1}_{\{x\}} + \text{sig}(x - EX)\gamma_{f_{.6}^{.001}}(|x - EX|)$$

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

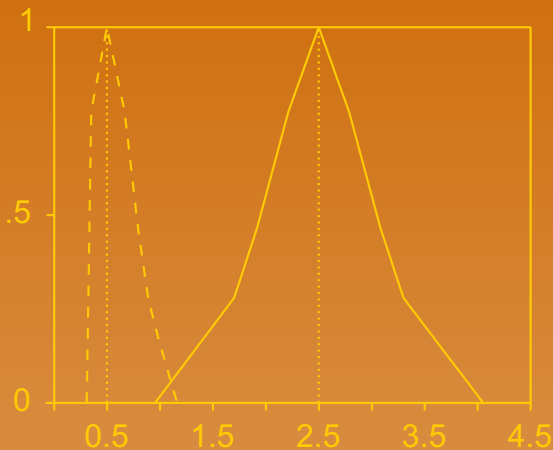
$$f_p^q(x) = \frac{p^x + q}{1 + q} \text{ with } p, q \in (0, 1)$$

The γ^{θ_s} -fuzzy representation of a random variable

$\theta_s = (0, 1, f_{.6}^{.001}) \in \Theta$ where $f_p^q(x) = \frac{p^x + q}{1+q}$ with $p, q \in (0, 1)$

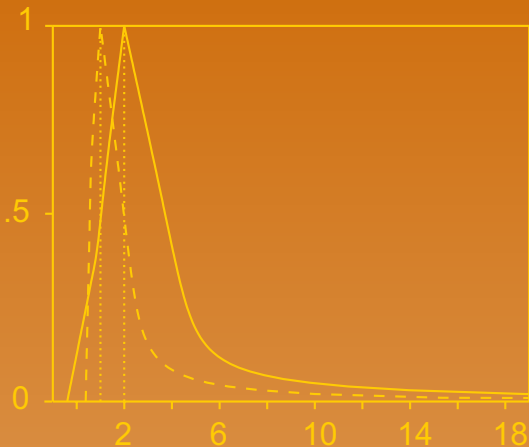


The exploratory fuzzy expected value associated with binomial distribution



Exploratory fuzzy expected value associated with $\mathcal{B}(5, 0.5)$ - $\mathcal{B}(5, 0.1)$. Comparison

The exploratory fuzzy expected value associated with χ^2 distribution



Exploratory fuzzy expected value associated with χ_1^2 and χ_2^2 distributions. Comparison

Exploratory analysis

- **Central tendency:** $(\tilde{E}(\gamma^{\theta_s} \circ X))_1 = \{EX\}$
- **Variability:** the area of the sendograph of $\tilde{E}(\gamma^{\theta_s} \circ X)$ is equal to $Var(X)$
- **Asymmetry:** the more skewness of X the more asymmetry of $\tilde{E}(\gamma^{\theta_s} \circ X)$
- **Continuity:**
 - ✓ if X is a continuous variable, then $\tilde{E}(\gamma^{\theta_s} \circ X)$ will be “smooth” (excepting at EX),
 - ✓ if X is discrete, $\tilde{E}(\gamma^{\theta_s} \circ X)$ will show non-smooth slope changes in each of the values $f(X)$ takes on
- **Extreme values:** large values of X will be associated with large-spread 0-level sets

Concluding remarks

- The fuzzy representations of random variables lead to an **integral methodology** to represent and estimating/testing about **distributions** of RVs which is quite easy to use, with good average empirical inferential results (specially in some cases) and with informative graphical displays
- The main **difference** with other ways of characterizing the distribution is that it is strongly focused on relevant parameters
- Further theoretical/empirical comparisons to other relevant techniques depending on concrete problems are being developed

More information...

- Contact:
 - ✓ Gil González-Rodríguez
European Centre for Soft Computing
Mieres. Spain.
 - ✓ gil.gonzalez@softcomputing.es

Thank you!