Comparison of Fuzzy and Crisp Random Variables by Monte Carlo Simulations

Olgierd Hryniewicz

Systems Research Institute,
Newelska 6, 01-447 Warszawa, Poland
hryniewi@ibspan.waw.pl
http://www.ibspan.waw.pl/~hryniewi

Abstract. Fuzzy random variables are used when randomness is merged with imprecision described by fuzzy sets. When we need to use computer simulations for the comparison of a classical probabilistic approach with that based on fuzzy random variables we need to establish the method for the generation of crisp random variables compatible with existing fuzzy data. In the paper we consider this problem, and propose some practical solutions.

Keywords: random variable, comparison of fuzzy and crisp random variables, Monte Carlo simulation.

1 Introduction

Random phenomena are usually modeled by classical probabilistic models. These models are definitely appropriate when sample observations of random variables are precisely reported, even if their actual values are not precisely known or may substantially vary. However, real statistical data may be defined in imprecise way. Firstly, the observed data may be reported using imprecise linguistic terms like "about one hundred" etc. Moreover, in reality there is often no reason to assume that the unknown values of observations are governed by the same probability distribution. In contrast to the case of precisely known observations, there is no method for the statistical verification of this important hypothesis.

To overcome the problems with the analysis of imprecisely reported statistical data two general approaches are used. First approach, still mainly used, is entirely probabilistic. The supporters of this approach propose to use complex, often multi-level, probabilistic models with many assumptions that are hardly verifiable in practice. The second approach is based on the notion of fuzzy random variables. In this approach imprecise observations are described by fuzzy sets such as e.g. fuzzy numbers. Fuzzy random variables have been introduced in order to merge this imprecision with pure randomness.

When we are dealing with complex problems whose formal description involves random variables it is usually not possible to solve these problems in purely analytical way. Therefore, in such cases we use computer simulation methods, known

P. Grzegorzewski et al. (eds.), *Strengthening Links between Data Analysis & Soft Computing*,
Advances in Intelligent Systems and Computing 315, DOI: 10.1007/978-3-319-10765-3_2

as Monte Carlo methods. When fuzziness is additionally involved in the description of complex problems their solution requires the usage of simulation of fuzzy random variables. The paper by Colubi et al. [1] serves as a very good example how simulation techniques may be used for the analysis of the properties of fuzzy random variables. There also exist numerous papers whose authors propose different methods for simulating fuzzy random variables for the solution of practically oriented problems. However, only few of them provide more general information about the methodology of simulation. The most general formal model that can be used for the simulation of fuzzy random variable has been proposed by Gonzalez-Rodriguez et al. [3]. The methodology presented in this paper is based on the general definition of the fuzzy random variable proposed by Puri and Ralescu [7], and the concept of the simulation of random elements in the separable Hilbert space.

When fuzzy random variables are used for modeling imprecisely described random phenomena an important question often arises about the advantage of this methodology over the classical one. The adherents of purely probabilistic approach claim that it is always possible to describe imprecision using classical probabilistic methods. In this paper we claim that in general they are right if we define a fuzzy random variable according to the definition firstly introduced by Kwakernaak [6]. However, the purely probabilistic model of the fuzzy random variable may be extremely complicated. Fuzzy methodology, in our opinion, provides tools for good approximations. It is interesting, however, to compare these approximations with the results provided by restricted (simplified) purely probabilistic models. It seems hardly possible to do such comparisons analytically, but we could do it using Monte Carlo simulations. In order to do so we need methods for the simulation of crisp random variables whose observed values are *compatible* with existing imprecise information. The proposal of a useful method for doing this is the main goal of this paper.

The remaining part of the paper is organized as follows. In Section 2 we discuss some important problems related to the simulation of fuzzy random variables. In Section 3 we present main original results of this paper. We use the concept of the possibility distribution, understood according to the interpretation of Dubois and Prade [2], for the construction of a random mechanism that generates crisp random variables compatible with respective fuzzy ones. In the fourth section we illustrate our results with some examples of simulation experiments. The paper is concluded in the last section of the paper.

2 Monte Carlo Generation of Fuzzy Random Variables

The notion of a fuzzy random variable has been defined in several papers, starting from early works of Zadeh on the fuzzy probability. The first generally accepted definition was introduced in the paper by Kwakernaak [6]. Statistical methods based on Kwakernaak's proposal have been developed in the works of Kruse (see [5]), so nowadays this approach is often coined as Kwakernaak-Kruse approach. The definition, we present below, is taken from [4], and is consistent with the results of Kruse and Kwakernaak.

Suppose that a random experiment is described as usual by a probability space $(\Omega, \mathcal{A}, \mathcal{P})$, where Ω is a set of all possible outcomes of the experiment, $\mathcal{A}$ is a $\sigma-$ algebra of subsets of Ω (the set of all possible events) and P is a probability measure

Definition 1. *A mapping $X : \Omega \to \mathcal{FN}$ is called a fuzzy random variable if it satisfies the following properties:*

(a) $\{X_\alpha(\omega) : \alpha \in [0,1]\}$ *is a set (α-cut) representation of $X(\omega)$ for all $\omega \in \Omega$,*
(b) for each $\alpha \in [0,1]$ both $X_\alpha^L = X_\alpha^L(\omega) = \inf X_\alpha(\omega)$ and $X_\alpha^U = X_\alpha^U(\omega) =$ $= \sup X_\alpha(\omega)$, are usual real-valued random variables on $(\Omega, \mathcal{A}, \mathcal{P})$.

According to Kruse [5] a fuzzy random variable X may be considered as an *imprecise perception* of an unknown usual random variable $V : \Omega \to \mathcal{R}$, called an *original* of X. There exists a more general definition proposed by Puri and Ralescu [7], but in this paper we restrict our attention to the case of the fuzzy random variable defined according to the Kwakernaak-Kruse approach.

Let us look at the definition of the fuzzy random variable from a point of view of computer simulations. It seems to be quite obvious that the ordinary random variables X_α^L and X_α^U must be dependent. Moreover, for all α-levels $0 \le \alpha \le 1$, and for all pairs of α-levels $0 \le \alpha' \le \alpha'' \le 1$ their joint probability distribution must fulfill the following requirements that assure the nested structure of the α-level subsets of the fuzzy observations.

$$P\left(x_\alpha^L < X_\alpha^L, X_\alpha^U \le x_\alpha^U\right) : \begin{cases} \ge 0 \ , x_\alpha^U > x_\alpha^L \\ = 0 \ , otherwise \end{cases} \tag{1}$$

$$P\left(x_{\alpha'}^L < X_{\alpha'}^L, X_{\alpha''}^L \le x_{\alpha''}^L\right) : \begin{cases} \ge 0 \ , x_{\alpha''}^L > x_{\alpha'}^L \\ = 0 \ , otherwise \end{cases} \tag{2}$$

$$P\left(x_{\alpha''}^U < X_{\alpha''}^U, X_{\alpha'}^U \le x_{\alpha'}^U\right) : \begin{cases} \ge 0 \ , x_{\alpha'}^U > x_{\alpha''}^U \\ = 0 \ , otherwise \end{cases} \tag{3}$$

Thus, we have the following proposition.

Proposition 1. *Let the fuzzy random variable $\tilde{X}$ be defined on a finite set of α levels $0 \le \alpha^{(1)} < \alpha^{(2)} < \cdots < \alpha^{(m)} \le 1$. Then, $\tilde{X}$ is fully described by a $2m$-dimensional vector $\left(X_{\alpha^{(1)}}^L, \ldots, X_{\alpha^{(m)}}^L, X_{\alpha^{(1)}}^U, \ldots, X_{\alpha^{(m)}}^U\right)$ of ordinary random variables whose joint probability distribution fulfills the conditions (1)-(3).*

When the values of α are not discretized the random vector mentioned in Proposition 1 becomes infinitely dimensional. Hence, any fuzzy random variable defined according to the Kwakernaak-Kruse approach can be represented by a fully probabilistic model described by dependent ordinary random variables whose marginal probability distributions must fulfill conditions(1)-(3). This property of the fuzzy random variables fully justifies the usage of Monte Carlo methods for the generation of fuzzy random samples.

3 Monte Carlo Simulation of Random Variables Compatible with Fuzzy Data

Let us assume that we have a sample of imprecise fuzzy data $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ observed in a random experiment (or simulated as the realizations of fuzzy random variable). If we want to compare fuzzy models and crisp probabilistic models using simulation methods we need to assume a certain probabilistic model in order to generate possible "origins" of the observed fuzzy data. The most frequently used approach consists in the transformation of the membership functions of the fuzzy numbers $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n$ into respective probability densities. For example, let us suppose that the fuzzy observation $\tilde{x_i}$ is described by a triangular normal fuzzy number (x_{i1}, x_{i2}, x_{i3}) (i.e. such that $\mu(x_{i2}) = 1$). Then, its membership function is easily transformed to the triangular probability density described by the triangle (x_{i1}, x_{i2}, x_{i3}) such that $f(x_{i2}) = 2/(x_{i3} - x_{i1})$.

This simple model has one serious disadvantage. As a matter of fact, it is a purely probabilistic model that fits the imprecise data to one specific probability distribution. We believe that this assumption is debatable. Consider, for example, two random fuzzy variables $\tilde{X}$ and $\tilde{Y}$ whose observations are described by intervals (i.e. by rectangular fuzzy numbers). According to the theory of fuzzy sets the observations of their sum should be also described by intervals. However, the probability distribution of the sum of their crisp "origins" simulated using the aforementioned method is not uniform.

In this paper we propose to interpret the membership functions of the observed fuzzy data as *possibility distributions*. The notion of the possibility distribution was introduced by Zadeh, and has many different interpretations. According to one of them, see [2], the possibility distribution can be understood as an upper envelope for all ordinary discrete probability distributions compatible with our imprecisely described value. Let $\mu_{[a,b]}(x)$ be the membership function of a fuzzy number $\tilde{x}$ with the support $[a, b]$. Consider now a representation of $[a, b]$ with a finite set of m real numbers $a \leq x_1 < x_2 < \ldots < x_m \leq b$. Now, let us define on this set the family $\mathcal{MN}$ of all *discrete* distributions $MN(p_1, p_2, \ldots, p_m)$ such that $p_j \leq \mu_{[a,b]}(x_j), j = 1, \ldots, m$, and $\sum_{j=1}^{m} p_i = 1$. The discrete distribution that belongs to the family $\mathcal{MN}$ we will call *compatible* with the possibility distribution $\mu_{[a,b]}(x)$. The value $x^\star$ randomly generated according to this distribution can be considered as a possible crisp "origin" of the fuzzy observation $\tilde{x}$.

For every fuzzy number $\tilde{x}$ defined on a non-degenerate interval $[a, b]$ there exist uncountably many distributions defined in the aforementioned way. However, for practical reasons we have to restrict the number of considered distributions. First, we should set the fixed number of points m. When $\mu_{[a,b]}(a) > 0$ and $\mu_{[a,b]}(b) > 0$ we set, respectively, $x_1 = a$ and $x_m = b$, and the remaining $m - 2$ points we may generate according to a certain probability distribution defined on (a, b). Otherwise, we generate all m points from this distribution. Note that any non-random generation of these m points (e.g. equidistant) can be considered as a special case of this general model.

Now, let us define some distributions on the set $\{x_1, \ldots, x_m\}$ that belong to the family $\mathcal{MN}$. Let us consider three such distributions: the *left concentrated* (LC), the *right concentrated* (RC), and the *random Dirichlet* (RD). The LC distribution is the distribution $MN(p_1, \ldots, p_k)$, where $p_j = \mu_{[a,b]}(x_j), j = 1, \ldots, k-1$ and $p_k = 1 - \sum_{j=1}^{k-1} p_j, 1 \geq k \leq m$. The respective RC distribution is the distribution $MN(p_l, \ldots, p_m)$, where $p_j = \mu_{[a,b]}(x_j), j = l+1, \ldots, m$ and $p_l = 1 - \sum_{j=l+1}^{m} p_j, 1 \geq l \leq m$. The interpretation of these distributions is simple when we are interested in the inference about the location parameter of the considered probability distribution or when the parameter of interest could be transformed to a location parameter by the appropriate transformation of the underlying random variable. For purely interval data the whole probability mass of the LC distribution is concentrated at the left limiting value of the considered interval. Similarly,for the RC distribution the whole probability mass is concentrated at the right limiting value of the considered interval.

Let us consider the LC distribution compatible with the triangular possibility distribution defined by the triangular fuzzy number (A, B, C). Let $s = |A, C|$ and $s_L = |A, B|$ be, respectively, the support and the left spread of this possibility distribution. We can now formulate the following proposition.

Proposition 2. *The expected value of the LC distribution compatible with triangular possibility distribution (A, B, C), and defined on the m evenly distributed points on the interval (A, C) is equal to A when m tends to infinity.*

Proof. Let X be a random variable defined on points $x_i = \frac{2i-1}{2}\frac{s}{s_L} + A$ evenly distributed on (A, C), and $p_i = \frac{2i-1}{2m}\frac{s}{s_L}$ be the corresponding probabilities of the LC distribution, such that $\sum_{i=1}^{k} p_i \leq 1 < \sum_{i=1}^{k+1} p_i$. One can prove that

$$Z_k = \sum_{i=1}^{k} \frac{2i-1}{2m}\frac{s}{s_L} = \frac{s}{2ms_L}k^2 \tag{4}$$

Hence, we have $p_{k+1} = 1 - Z_k$. If probabilities $p_1, p_2, \ldots, p_k, p_{k+1}$ describe the probability distribution defined on the set $\{x_1, x_2, \ldots, x_k, x_{k+1}\}$ the condition $Z_k + p_{k+1} = 1$ must be fulfilled. Note that $p_{k+1} \leq \frac{2k+1}{2m}\frac{s}{s_L}$, and for m tending to infinity this probability tends to zero. Thus, we have the following equation that defines the relationship between k and m

$$\frac{s}{2ms_L}k^2 = 1 \tag{5}$$

Now, let us calculate the expected value of X from the following formula

$$\begin{aligned} E(X) &= \textstyle\sum_{i=1}^{k} p_i x_i = \sum_{i=1}^{k}(\frac{2i-1}{2}\frac{s}{m} + A)(\frac{2i-1}{2m}\frac{s}{s_L}) \\ &= \frac{s^2}{s_L}\frac{1}{4m^2}\frac{1}{3}k(4k^2-1) + \frac{A}{2m}\frac{s}{s_L}k^2. \end{aligned} \tag{6}$$

From (5) we have $k = \sqrt{2ms_L/s}$, and hence

$$E(X) = \frac{s^{3/2}}{s_L^{1/2}}\frac{\sqrt{2}}{12}(\frac{8s_L}{s}m^{-1/2} - m^{-3/2}) + A. \tag{7}$$

Thus, from (7) we see that for $m \to \infty$ we have $E(X) \to A$, and this ends the proof. In the similar way we can prove a symmetric proposition concerning the RC distribution

Proposition 3. *The expected value of the RC distribution compatible with triangular possibility distribution* (A, B, C)*, and defined on the* m *evenly distributed points on the interval* (A, C) *is equal to* C *when* m *tends to infinity.*

In the general case, however, the values of the parameters $p_1, \ldots, p_m$ may be also chosen in a random way. The Dirichlet distribution is defined by the following density function

$$f(p_1, \ldots, p_m; \beta_1, \ldots, \beta_m) = \begin{cases} \frac{1}{B_m} \prod_{j=1}^{m} s_j^{\beta_j} & , (s_1, \ldots, p_m) \in S_m \\ 0 & , \quad otherwise \end{cases} , \tag{8}$$

where S_m is the closed $m-1$-dimensional simplex and B_m is the normalizing constant.

The Dirichlet distribution defined by (8) is very flexible, and allows to simulate very different "shapes" of the probability distribution compatible with a fuzzy number $\tilde{x}$ which can be used for the generation of the "origin" value representative for this fuzzy number. In the simulation algorithm the values of the parameters $\beta_1, \ldots, \beta_m$ can be chosen randomly, for example from a predefined interval $(\beta_{min}, \beta_{max})$. It gives an additional level of flexibility in the generation of probabilities $p_1, \ldots, p_m$. Then, the values of probabilities $p_1, \ldots, p_m$ can be generated from the Dirichlet distribution (8). Finally, the $MN(p_1, \ldots, p_m)$ distribution can be used for the generation of the "origin" of the observed value of the fuzzy random variable from among the set of (predefined or randomly generated) values $\{x_1, \ldots, x_m\}$. We have called this distribution the random Dirichlet (RD).

4 Properties of Probability Distributions Representing Fuzzy Random Variables – Results of Experiments

From the discussion presented in Section 2 we know that a fuzzy random observation $\tilde{z}$ can be represented as the sum of the unobserved crisp random "origin" y and a fuzzy number $\tilde{x}$ that represents our lack of knowledge about the "origin". In Monte Carlo experiments we can simulate "origins" from a given probability distribution. Then, we can use a certain predefined random mechanism for the generation of the membership function $\mu(x)$ of $\tilde{x}$. Thus, the simulated fuzzy observation is a fuzzy number $\tilde{z} = y + \tilde{x}$.

When we need to compare the approach based on random fuzzy numbers with a classical approach based on crisp random numbers we should simulate crisp random numbers that are compatible with our fuzzy observations. In this section we present the results of simulation experiments that have been performed in order to investigate the differences between different methods of the simulation of random variables compatible with given fuzzy observations. In this paper we describe only the results of experiments in which we assumed that the observed fuzzy numbers are described by triangular membership functions symmetric around zero with the randomly generated left and right spreads L and R. Moreover, we assumed that both are described by the same probability distribution characterized by its expected value w and a coefficient of variation v.

In our experiments we have considered four types of probability distributions compatible with the given fuzzy observation: triangular, left(right) concentrated, and random Dirichlet. Because the spreads L and R have been simulated from the same distribution the average behavior of the LC and RC distributions was the same. Therefore, we present here only the results for the RC distribution. In the experiments we generated samples of n random fuzzy variables $\tilde{X}$, Then, for each generated sample item we generated its "origin" from its compatible probability distribution. In the next step we calculated the sample average of the generated "origins". The procedure has been repeated 100 000 times in order to evaluate the properties of the simulated distributions of sample averages, such as the expected value and the standard deviation.

In the first group of simulation experiments we investigated the dependence of the expected value of the RC (LC) distribution on the number of discretization points m when the triangular membership functions were generated from different probability distributions. The convergence to the limiting values defined by Proposition 3 (or 2) was rather slow. For example, when the spreads were generated from the uniform distribution defined on the interval $(0, 4)$ the expected value for the RC distribution for $m = 500$ was equal to $1,88$. Note however, that according to Proposition 3 for $m \to \infty$ this expected value should be equal to 2, i.e. to the expected length of the right spread. The results obtained in similar experiments have shown that for a realistic discretization of the possibility distribution described by a fuzzy number the observed average values of the left and right concentrated distributions are not so far from their theoretical, but rather improbable, values.

The important question may arise about the difference between the random variables generated from the triangular (Tr) distribution and the random variables generated with the usage of the proposed random Dirichlet (RD) distribution. Because of the way the triangular membership functions are generated (maximum at zero, the same distribution of the both, left and right, spreads) the expected value of the sample average for this distribution must be equal to zero. However, in the case of the random Dirichlet distribution the similar behavior of the sample average is somewhat unexpected. Only in the case of small values of m the estimated average is slightly different than zero (e.g. for $m = 10$ it is equal to -0.046 while σ is equal to 0.265). The situation is different when we consider the standard deviation of $\bar{x}$.

In all considered cases the variability of sample means generated from the random Dirichlet distribution was greater than the similar variability in the case of the triangular distribution. This difference becomes significant for moderate and large values of m (e.g. larger than 200). This means that the random Dirichlet distribution is less "informative", and represents fuzziness in a better way.

Moreover, for the large values of m the standard deviation of $\bar{x}$ practically does not depend upon the value of this parameter. Therefore, the maximal variability of the generated crisp observations that are compatible with the given fuzzy number can be obtained even for a moderate number of discretization points. This property tells us that in real experiments the value of m need not be too large, and thus, simulation experiments need not be time consuming. One should also remember that in practice the variability of the distribution compatible with fuzzy observations is equal to the sum of the variability of an unobserved “origin” and the variability of the distribution representing observed fuzziness. When the former is much larger than the latter the difference between the triangular and the random Dirichlet distributions may be neglected.

5 Conclusions

The widely used methods of the generation of fuzzy random variables are fully compatible with the Kwakernaak-Kruse definition of the fuzzy random variable. The concept of the probability distribution compatible with a fuzzy observation introduced in this paper provides a simple methodology for the comparison of classical (non-fuzzy) and fuzzy approaches for dealing with imprecise data. In the classical approach the lack of knowledge is modeled by a predefined and difficult to identify probability distribution. When the fuzzy approach is used this lack of knowledge may be modeled by several probability distributions that are compatible, in the sense introduced in the paper, with imprecise observations. Therefore, this approach provides more flexibility in the description of imprecisely observed random phenomena.

Acknowledgments. The author thanks Dr. Piotr Nowak for his helpful comments regarding the interpretation of fuzzy random variables.

References

1. Colubi, A., Fernandez-Garcia, C., Gil, M.A.: An Empirical Approach to Statistical/Probabilistic Studies With Fuzzy Experimental Data. IEEE Transactions on Fuzzy Systems 10, 384–390 (2002)
2. Dubois, D., Prade, H.: Possibility Theory. Plenum Press, New York (1988)
3. Gonzalez-Rodriguez, G., Colubi, A., Trutschnig, W.: Simulation of fuzzy random variables. Information Sciences 179, 642–653 (2008)
4. Grzegorzewski, P., Hryniewicz, O.: Computing with words and life data. Int. Journ. of Appl. Math. & Comp. Science 12, 337–345 (2002)
5. Kruse, R., Meyer, K.D.: Statistics with Vague Data. Riedel, Dordrecht (1987)
6. Kwakernaak, H.: Fuzzy random variables, part I: definitions and theorems. Information Sciences 15, 1–15 (1978)
7. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. Journ. of Math. Anal. and Appl. 114, 409–422 (1986)