International Journal of Approximate Reasoning 55 (2014) 1487–1501

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A distance-based statistical analysis of fuzzy number-valued data

A. Blanco-Fernández, M.R. Casals, A. Colubi, N. Corral, M. García-Bárzana, M.A. Gil, G. González-Rodríguez, M.T. López, M.A. Lubiano, M. Montenegro, A.B. Ramos-Guajardo, S. de la Rosa de Sáa, B. Sinova

SMIRE Research Group at the University of Oviedo *,[1]

Dpto. de Estadística e I.O. y D.M., Universidad de Oviedo, 33007 Oviedo, Spain

ARTICLE INFO

Article history:
Available online 16 October 2013

Keywords:
Fuzzy data
Random experiments
Distance between fuzzy data
Statistics

ABSTRACT

Real-life data associated with experimental outcomes are not always real-valued. In particular, opinions, perceptions, ratings, etc., are often assumed to be vague in nature, especially when they come from human valuations. Fuzzy numbers have extensively been considered to provide us with a convenient tool to express these vague data. In analyzing fuzzy data from a statistical perspective one finds two key obstacles, namely, the nonlinearity associated with the usual arithmetic with fuzzy data and the lack of suitable models and limit results for the distribution of fuzzy-valued statistics. These obstacles can be frequently bypassed by using an appropriate metric between fuzzy data, the notion of random fuzzy set and a bootstrapped central limit theorem for general space-valued random elements. This paper aims to review these ideas and a methodology for the statistical analysis of fuzzy number data which has been developed along the last years.

1. Introduction

Traditionally, most of the statistical developments refer to the analysis of real-valued data. However, in the real world other types of data often arise, so that Statistics is being enlarged by considering new ideas, concepts and results. In this respect, one can mention vector-valued, censored, set-valued, spatial, and functional data, among others, as well as their corresponding analyses, to cover a wide diversity of data that one can often find in real-life.

On the other hand, in the last decades fuzzy sets have been shown to be a powerful tool to represent quantitatively (more precisely, functionally) many imprecise real-life data. More concretely, we often find situations in which the values of some involved variables or attributes can only be observed imprecisely. For instance, when opinions, quality ratings, satisfaction valuations and many other surveys are conducted, responses that usually vary from respondent to respondent and from question to question (that is, they are subject to randomness) cannot be expected to be expressible in terms of values for a precise scale.

The approach followed in this paper is the so-called 'ontic' one (see, for instance, Dubois and Prade [13]), where ontic means to pertain to the essence of the object under study. Three main situations support this approach:

* Corresponding author. Tel.: +34 985103356; fax: +34 985103354.
[1] *E-mail address:* SMIRE@uniovi.es.

0888-613X/$ – see front matter © 2013 Published by Elsevier Inc.
http://dx.doi.org/10.1016/j.ijar.2013.09.020

- When the observer imprecisely measures a precisely defined attribute but we study the imprecise data he/she provides (fuzzy numbers are ontic w.r.t. the observer and they describe how much the observer is imprecise);
- When perceptions are collected from a human but there is no objective value behind (e.g., the taste of a dessert does not directly describe the objective ingredients of the dish); here the ontic view refers to the human tasting;
- When an intrinsically gradual representation subject to variability in the nature is described by means of fuzzy sets (e.g., regions with gradual boundaries in an image, flexible time intervals, a vector of ratings, etc.).

As we will later emphasize, the ontic approach is to be treated by considering Féron/Puri and Ralescu's fuzzy-valued random elements (see [14,46]). These random elements associate a fuzzy value with each experimental outcome and related statistical methods attempt to draw conclusions about them. In contrast, a different view is offered by Kruse and Meyer's fuzzy random variables [37] that formalize the fuzzy perception of an underlying (original) real-valued random variable (i.e., the 'epistemic' approach is followed, where epistemic means to pertain to the quality of the knowledge that can be obtained on the object under study). A fuzzy random variable in Kruse and Meyer's sense made reference to the set of possible originals and statistical methods attempt to draw conclusions about the unknown real-valued original.

To treat fuzzy number data with statistical purposes some peculiarities should be pointed out:

- If one tries to perform a descriptive study summarizing the information contained in a sample of fuzzy data, then one should basically focus on the arithmetic, the ranking and the distances between these data. At this point, two main drawbacks can be stressed in developing descriptive statistics with fuzzy number data, namely,
  - The usual and most natural arithmetic between fuzzy numbers extends the one between real values and the usual one between intervals. It involves special features in contrast to the real-valued case, due to the lack of linearity associated with it. As a consequence, there is no definition for the difference of fuzzy numbers which is simultaneously well-defined and preserves some essential properties of the difference in the real-valued case, whence the use of the difference should be mostly avoided;
  - A universally acceptable complete order between fuzzy numbers cannot be established, and therefore the notion of empirical cumulative distribution function and related summary measures cannot be formalized in the usual way.

  To overcome these drawbacks, one possible/common solution is to define a proper distance between fuzzy numbers. A suitable metric between fuzzy numbers can be employed to quantify the average variability of data, the error associated with the approximation of some measures or relationships, etc. In this way, one can cope with the formalization and computation of some summary measures of the fuzzy datasets, and with the development of the descriptive fitting of some functional relationships (i.e., regression analyses) when two- or multi-dimensional fuzzy data are analyzed.
- If one tries to perform an inferential study about a random mechanism generating fuzzy number-valued data (in which fuzziness affects data values, whereas randomness affects their generation) on the basis of a sample from it, then one should mainly focus on defining suitable statistics and obtaining their sampling distributions. At this point, in addition to the drawbacks already mentioned for the descriptive analysis, two new concerns can be noticed in developing inferential statistics with fuzzy number data, namely,
  - There are not yet probability distribution models which have been shown and proved to be convenient and realistic enough to deal with the random mechanisms producing fuzzy number data;
  - There are not Central Limit Theorems for random mechanisms producing fuzzy data which could be directly applied in inferential developments because it ensures that the limit element takes values on the space of fuzzy numbers.

  To overcome these drawbacks, three tools will be crucial: the notion of random fuzzy number and its induced distribution; the real-valued distance-based measures or statistics (quantifying either the error in the fuzzy number-valued estimation or the deviation between the reality of the sample fuzzy information and a hypothetical model in testing hypothesis); the existence of a bootstrapped Central Limit Theorem for generalized space-valued random elements allowing us to apply re-sampling techniques which guarantee the well-definition of the involved elements as fuzzy numbers. In this way, one can cope with establishing procedures either to estimate some 'parameters' of the distribution of random mechanisms or to test some statistical hypotheses about them, by preserving the essential ideas, concepts and approaches used to solve these problems in the real-valued case.

This paper aims to describe the main ideas and results on a distance-based methodology for the statistical analysis of fuzzy number data which has been developed along the last years. In Section 2, the considered metric and some metric properties will be recalled; some equivalent expressions will also be given, due to the practical and theoretical implications derived from such equivalences. Section 3 will be devoted to the formalization of the random mechanisms producing fuzzy number data (random fuzzy numbers), their connection with the distance and some relevant distance-based summary measures. Section 4 will examine the distance-based methodology to analyze fuzzy number data. The paper will conclude with some comments related to an application of the preceding methodology to develop inferential analysis on the distributions of real-valued random variables.

The concepts and methods that appear in this review, along with other ones which are not based on the same metric, are described in detail in the original manuscripts that introduce them (references will be given in due course) as well as, to some extent, in some recent reviews prepared by invitation (see Blanco-Fernández et al. [2], Colubi et al. [3] and Gil et al. [19]). These reviews refer to the more general case of fuzzy-valued data and random fuzzy sets indeed. The review in