Computational Statistics and Data Analysis 56 (2012) 943–955

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Fuzzy data treated as functional data: A one-way ANOVA test approach

Gil González-Rodríguez [a,*], Ana Colubi [b], María Ángeles Gil [b]

[a] *European Centre for Soft Computing, Edificio Científico-Tecnológico, 33600 Mieres, Asturias, Spain*
[b] *Departamento de Estadística, I.O. y D.M., Universidad de Oviedo, 33071 Oviedo, Spain*

A R T I C L E   I N F O

Article history:
Received 14 January 2010
Received in revised form 13 June 2010
Accepted 13 June 2010
Available online 22 June 2010

Keywords:
Functional data
Fuzzy data
k-samples test
ANOVA statistic
Hilbert space
Convex cone
Bootstrap
Local alternatives

A B S T R A C T

The use of the fuzzy scale of measurement to describe an important number of observations from real-life attributes or variables is first explored. In contrast to other well-known scales (like nominal or ordinal), a wide class of statistical measures and techniques can be properly applied to analyze fuzzy data. This fact is connected with the possibility of identifying the scale with a special subset of a functional Hilbert space. The identification can be used to develop methods for the statistical analysis of fuzzy data by considering techniques in functional data analysis and *vice versa*. In this respect, an approach to the FANOVA test is presented and analyzed, and it is later particularized to deal with fuzzy data. The proposed approaches are illustrated by means of a real-life case study.

1. Introduction and motivation

Data which cannot be exactly described by means of numerical values, such as evaluations, medical diagnosis or quality ratings, to name but a few, are frequently classified as either nominal or ordinal. A well-known example is the so-called Likert scales (cf. Likert, 1932; Allen and Seaman, 2007) in which categories are labeled with numerical values. Using these scales, the statistical analysis is limited. Many parameters and techniques cannot be directly used or, when they can, the interpretation and reliability of the conclusions are considerably reduced (see, for instance, Stevens, 1946, or Chimka and Wolfe, 2009). Additionally, the transition from one category to another is rather abrupt (see, for instance Ammar and Wright, 2000). A third concern is that categories are not perceived in the same manner by different observers, so that the variability and accuracy cannot always be well captured.

A new easy-to-use representation of such data through fuzzy values is to be considered. The measurement scale of fuzzy values includes, in particular, real vectors and set values as special elements. It is more expressive than ordinal scales and more accurate than rounding or using real or vectorial-valued codes. The arithmetic and the metric to be used make it possible to extend naturally many of the usual statistical measures and techniques. The transition between closely different values can be made gradually, and the variability, accuracy and possible subjectiveness can be well reflected in describing data. Although fuzzy data could be directly viewed as special functional data, the arithmetic and the metric being coherent with their meaning do not coincide with the usual ones for functional data. Nevertheless, the so-called support function establishes a useful embedding of the space of fuzzy values into a cone of a functional Hilbert space (see Puri and Ralescu, 1983, 1985; Körner and Näther, 2002).

* Corresponding author. Tel.: +34 985456545; fax: +34 985456699.
E-mail address: gil.gonzalez@softcomputing.es (G. González-Rodríguez).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.06.013

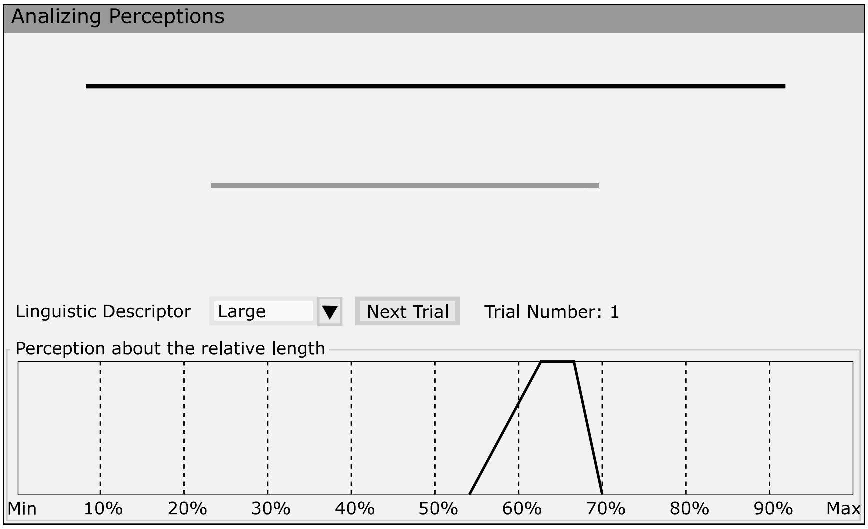

Fig. 1. Application to express the perception on the relative length of segments.

A guideline suggesting how to collect/describe fuzzy data associated with random experiments will be presented in Section 2. In Section 3, the natural arithmetic, the support function and a family of metrics on the space of fuzzy values will be recalled. Random fuzzy sets and their connection with Hilbert space-valued random elements will be described in Section 4. In Section 5, a one-way ANOVA test for functional data will be introduced, and it will be particularized to deal with fuzzy data. Specifically, an asymptotic test and its behaviour under local alternatives, as well as a bootstrap procedure, will be analyzed. A case study introduced in Section 2 will be considered in Section 6 to illustrate the approach.

2. Collecting fuzzy data from random experiments

The space $\mathcal{F}_c(\mathbb{R}^p)$ of *fuzzy values* to be considered contains the mappings $\widetilde{U} : \mathbb{R}^p \to [0, 1]$ so that for each $\alpha \in (0, 1]$ the *α-level set* $\widetilde{U}_\alpha = \{x \in \mathbb{R}^p : \widetilde{U}(x) \geq \alpha\}$ is a nonempty compact convex set of $\mathbb{R}^p$. When $p = 1$ the fuzzy values are referred to as fuzzy numbers. Formally, fuzzy values are $[0, 1]$-valued upper semicontinuous functions with nonempty convex bounded α-levels. Real, vectorial, interval and set-valued data can be viewed as particular fuzzy data, by identifying them with the associated indicator functions.

A fuzzy value $\widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p)$ models an ill-defined subset of $\mathbb{R}^p$, so that for each $x \in \mathbb{R}^p$ the value $\widetilde{U}(x)$ can be interpreted as 'degree of membership' of x to $\widetilde{U}$. Alternatively, $\widetilde{U}$ may be interpreted as the 'degree of compatibility' of x with an ill-defined property $\widetilde{U}$. In practice, fuzzy data usually come from either a pre-established classification, such as the danger of forest fires (see Colubi and González-Rodríguez, 2007), or from a designed experiment. This is the case of the expert evaluation of the trees in a reforestation analyzed in Colubi (2009), where the ill-defined characteristic 'quality' is individually described through a fuzzy set. Obviously, accuracy and variability of data are much better captured by using individual fuzzy assessments than by considering a pre-fixed list of fuzzy values.

The main concepts and methods are to be illustrated by means of a case study which is now introduced along with some guidelines to describe fuzzy data. The case study regards an experiment in which people have been asked for their perception of the relative length of different line segments with respect to a fixed longer segment that is used as a standard for comparison. Fig. 1 displays the screen of the application. On the center top of the screen the pattern (longest line segment) is drawn in black. At each trial a gray shorter line segment is generated and placed below the pattern one, parallelly and without considering a concrete location (i.e., indenting or centering).

After an explanation of the fuzzy values, participants are asked by their judgment of relative length for each of several line segments in two ways. First, to choose a label from a Likert-like list, {VERY SMALL, SMALL, MEDIUM, LARGE, VERY LARGE}. Second, to describe the perception through a trapezoidal fuzzy number with support included in [0, 100] (0% indicating the minimum relative length and 100% maximum the one). The support is to be chosen as the set of all values that the participant subjectively considers to be compatible with the relative length of the generated segment to a greater or lesser extent. The 1-level has to be the set of all values that the participant considers to be completely compatible with his/her perception about the relative length of the generated segment. The trapezoidal fuzzy set is formed by the linear interpolation of both intervals, although it is possible to change the shape. In other words, out of the support are the values that the participant is not willing to accept as possible values for the relative length at all. The membership degree is linearly increasing from the minimum of the support to the first value for which the participant would say that it is the relative length of the line (see Fig. 1). However, since the participant may have doubts, often there is not a unique value in these conditions, but an