European Journal of Operational Research 241 (2015) 469–477

Contents lists available at ScienceDirect

European Journal of Operational Research

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

Decision Support

Spread measures and their relation to aggregation functions

Marek Gagolewski [a,b,*]

[a] Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warsaw, Poland
[b] Faculty of Mathematics and Information Science, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland

ARTICLE INFO

Article history:
Received 12 February 2014
Accepted 22 August 2014
Available online 19 September 2014

Keywords:
Group decisions and negotiations
Aggregation
Spread
Deviation
Variance

ABSTRACT

The theory of aggregation most often deals with measures of central tendency. However, sometimes a very different kind of a numeric vector's synthesis into a single number is required. In this paper we introduce a class of mathematical functions which aim to measure spread or scatter of one-dimensional quantitative data. The proposed definition serves as a common, abstract framework for measures of absolute spread known from statistics, exploratory data analysis and data mining, e.g. the sample variance, standard deviation, range, interquartile range (IQR), median absolute deviation (MAD), etc. Additionally, we develop new measures of experts' opinions diversity or consensus in group decision making problems. We investigate some properties of spread measures, show how are they related to aggregation functions, and indicate their new potentially fruitful application areas.

1. Introduction

Many introductory textbooks on applied statistics (or academic lectures on the subject) include a review of the so-called descriptive statistics, i.e. methods for summarizing quantitative data. Most often such methods are divided into at least two classes (cf. Aczel, 1996, Chap. 1 and e.g. Cramér, 1946):

1. *measures of central tendency* (also known as measures of location or centrality of observations); e.g. sample quantiles (including median, min, and max), arithmetic mean, mode, trimmed and winsorized mean etc.,
2. *measures of variability* (or data spread), e.g. range, interquartile range, variance, standard deviation.

At the most general level, the process of combining multiple numeric values into a single, representative number is called aggregation. The theory of aggregation became a genuine, rapidly developing research field in the 1980s (see e.g. Beliakov, Pradera, & Calvo, 2007; Calvo, Mayor, & Mesiar, 2002; Grabisch, Marichal, Mesiar, & Pap, 2009, 2011a,b). It may be observed, however, that the aggregation theory mainly focuses on the above-mentioned measures of central tendency, e.g. generalized means (OWA, OWMax operators, quasi-arithmetic means, etc.), averages, or "averaging functions". Such a broad class of tools is characterized by the following widely accepted definition of an aggregation function (see Grabisch et al., 2009, Def. 1.1).

Definition 1. Let $\mathbb{I} = [a, b]$. $\mathsf{A} : \mathbb{I}^n \to \mathbb{I}$ is an *aggregation function* if at least:

(a1) it is nondecreasing in each variable, i.e. for all $\mathbf{x}, \mathbf{x}' \in \mathbb{I}^n$ such that $\mathbf{x} \le_n \mathbf{x}'$, i.e. $(\forall i)\ x_i \le x_i'$, it holds $\mathsf{A}(\mathbf{x}) \le \mathsf{A}(\mathbf{x}')$,

and fulfills the boundary conditions:

(a2) $\inf_{\mathbf{x}\in\mathbb{I}^n} \mathsf{A}(\mathbf{x}) = \inf \mathbb{I}$,
(a3) $\sup_{\mathbf{x}\in\mathbb{I}^n} \mathsf{A}(\mathbf{x}) = \sup \mathbb{I}$.

It is true that these characteristic properties reflect somehow the concept of data synthesis: finding a value representative to the whole vector. Moreover, it is well-known that such functions are strongly connected to monotone (fuzzy) measures and integrals (cf. e.g. Greco, Mesiar, & Rindone, 2014).

Aggregation functions have many successful applications, for example in multicriteria or group decision making, statistics, quality management, engineering, approximate reasoning, fuzzy sets and fuzzy logic (cf. the notion of a t-norm and t-conorm, which are particular aggregation functions in $[0, 1]^2$), etc.

Example 1. In a group decision making problem, assume that n decision makers express as $x_1, \ldots, x_n \in [0, 1]$ the strength of preference toward an alternative. An aggregation function may be used to combine these assessments in order to obtain a global score $\mathsf{A}(x_1, \ldots, x_n)$. For example, let $n = 4$ and $\mathbf{x} = (1, 1, 1, 0)$. If all the experts have the same standing, one may use e.g. the arithmetic mean to combine their opinions; in such a case we get $\mathsf{A}(\mathbf{x}) = 0.75$. However, assume that the fourth decision maker is conceived as less competent (at least in a given matter) than the other ones, or his/her opinion has lower significance for some other reason (see e.g. Bernasconi, Choirat, & Seri, 2014; Saaty, 1994). If e.g. a weighting vector $\mathbf{w} = (2/7, 2/7, 2/7, 1/7)$

[*] Tel: +48 22 38 10 393; fax: +48 22 38 10 105.
E-mail address: gagolews@ibspan.waw.pl

http://dx.doi.org/10.1016/j.ejor.2014.08.034
0377-2217/© 2014 Elsevier B.V. All rights reserved.

describes the importance of the respective judges, then by calculating the weighted mean we get $\mathsf{A}'(\mathbf{x}) = 6/7 \simeq 0.86$.

It is evident that to understand the very nature of aggregation processes better, as well as to meet the practitioners' needs, we should explore new classes of methods for summarizing quantitative data. And so, the second group of measures from the above classification of descriptive statistics consists of single numbers that quantify the broadly-conceived "variability" of mathematical objects. Let us investigate it more deeply.

An important, yet not directly connected with our task, characterization of measures of entropy or uncertainty of discrete probability mass functions (represented by numeric vectors in $[0, 1]^n$ with elements summing up to 1) was proposed by Martín, Mayor, and Suñer (2001). Such a class includes e.g. the Shannon entropy and alike, cf. also (Kostal, Lansky, and Pokora (2013)). Other very loosely related measures include the notion of fuzziness of a fuzzy set, cf. (Sanchez & Trillas, 2012; Weber, 1984; Zeng & Li, 2006), multidiscances (Martin & Mayor, 2011), or a probability distribution's scale parameter estimates (non-negative, translation and ratio scale invariant functions discussed by Pitman, 1939).

Among the aggregation methods of our concern, on the other hand, we may find:

1. *Measures of absolute data spread*, e.g. standard deviation, IQR, MAD. In this case, an absolute spread measure V may accompany an aggregation function A in order to state that a numeric list $\mathbf{x}$ is concisely described as $\mathsf{A}(\mathbf{x}) \pm \mathsf{V}(\mathbf{x})$.
2. *Measures of relative data spread* (e.g. Gini coefficient, coefficient of variation), which are dependent on the order of magnitude of a numeric list's elements. For instance, imagine that we have two groups of people. The first group consists of $(1, 2, 3)$-year-olds and the second one of $(101, 102, 103)$-year-olds. Intuitively, the relative spread of age in the first group is greater than that of the second group.

Most importantly, to our best knowledge none of these has been discussed from the point of view of aggregation theory. In particular, it is still unknown what characteristic properties link the measures within both groups. Note that even in statistics there are many functions which aim – at least theoretically – to be used for the mentioned purposes. Also, diverse application areas require treatment with different suitable measures. We strongly believe that the measures of absolute and relative spread are worth of deeper, separate studies. Hence, this contribution will focus on the first subclass.

Example 1 (cont'd). If all the experts are of the same esteem, we may use e.g. the sample standard deviation to assess the consistency of decision makers' opinions, refer e.g. to Huang, Chang, and Lin (2013) for such an approach. However, if some form of weighting of the importance of opinions or their values is needed, then we should seek for a different kind of method for measuring the hetero/homogeneity. This, apart from measures of central tendency, could be an important, supplementary information on a numeric sequence, (cf. e.g. Ohki & Murofushi (2012)).

The paper is structured as follows. In Section 2 we propose a binary preorder which is further on used to determine whether a vector has no larger absolute spread than another one. Basing on this notion, in Section 3 we introduce the notion of a spread measure and indicate some additional properties that may be useful in particular application areas. In Section 4 we prove that the spread measures are naturally connected to aggregation functions. In Section 5 we show that the well-known descriptive statistics, like sample variance, standard deviation, interquartile range, range, median absolute deviation, and mean difference, are consistent with our definition and develop some new classes of functions which are of particular usefulness in DM tasks. Finally, Section 6 concludes the paper and indicates many ideas worth of deeper further studies.

2. Vectors' spread

Fix $n \in \mathbb{N}$ and let $\mathbb{I} = [a, b]$, $b > a$. From now on for each $c \in \mathbb{I}$ we denote by $(n * c)$ a sequence $(c, c, \ldots, c) \in \mathbb{I}^n$. Additionally, we assume that $[k] = \{1, \ldots, k\}$ and that whenever at least one argument is a sequence, then all arithmetic operations are properly vectorized, e.g. we have $\mathbf{x} + \mathbf{x}' = (x_1 + x'_1, \ldots, x_n + x'_n)$ and $\mathbf{x} + c = \mathbf{x} + (n * c) = (x_1 + c, \ldots, x_n + c)$. In particular, Ind is a vectorized Boolean indicator function, i.e. $\mathrm{Ind}(c_1, \ldots, c_n) = (v_1, \ldots, v_1)$ with $v_i = 1$ iff logical condition c_i is true and 0 otherwise. What is more, let $x_{(i)}$, $i \in [n]$, denote the ith smallest element in $\mathbf{x} \in \mathbb{I}^n$, $\mathfrak{S}_{[n]}$ denote the set of all permutations of $[n]$, and for any $\sigma \in \mathfrak{S}_{[n]}$, $\mathbb{I}^n_\sigma = \{(x_1, \ldots, x_n) \in \mathbb{I}^n : x_{\sigma(1)} \le \cdots \le x_{\sigma(n)}\}$. Furthermore, if $\mathsf{F} : \mathbb{I}^n \to \mathbb{I}$, then let $\mathsf{F}|_\sigma$ denote the restriction of F to $\mathbb{I}^n_\sigma$, i.e. $\mathsf{F}|_\sigma : \mathbb{I}^n_\sigma \to \mathbb{I}$, $\mathsf{F}|_\sigma(\mathbf{x}) = \mathsf{F}(\mathbf{x})$ for any $\mathbf{x} \in \mathbb{I}^n_\sigma$.

2.1. Introductory remarks

Please note that the notion of $\le_n$ plays a central role in the definition of aggregation functions. It is because an aggregation function is a morphism between the partially ordered space $(\mathbb{I}^n, \le_n)$ and linearly ordered space $(\mathbb{I}, \le)$, cf. property (a1).

In other words, if $\mathbf{x} \le_n \mathbf{x}'$, then we are certain that each aggregation function ranks $\mathbf{x}$ no higher than $\mathbf{x}'$.

We shall introduce the class of absolute spread measures in a similar manner. Let us pose a question: In which case does a given vector in $\mathbb{I}^n$ *surely* have the same or not greater spread than another one in $\mathbb{I}^n$? Here is a list of the sine qua non postulates that seem reasonable for most applications.

- *Lowest possible spread.* Any constant vector, $(n * c)$, $c \in \mathbb{I}$ should have the lowest possible spread of all the vectors considered.
- *Invariance to translations.* Spread comparison results should not change when we translate all elements in at least one sequence considered, i.e. $\mathbf{x}$ and $\mathbf{x} + t$ are of the same spread for any $\mathbf{x}$, t. Note that such a condition would be inappropriate in case of measures of relative spread.
- *Non-symmetry.* In statistics and data analysis, perhaps we will not take into account the relative ordering of the elements in a sequence: for any $\sigma', \sigma'' \in \mathfrak{S}_{[n]}$ the vectors $(x_{\sigma'(1)}, \ldots, x_{\sigma'(n)})$ and $(x_{\sigma''(1)}, \ldots, x_{\sigma''(n)})$ have the same spread, as we treat all the observations as just "points in the real line"; however, here we should be interested in a more general setting in which the relative ordering may be important: for example, each element in a vector may have a corresponding weight which is determined by its position (the ith element may be more "important" than the jth, cf. the above example).

Moreover, how to modify a given vector $\mathbf{x}$ so that its spread surely does not decrease? A sensible answer may be given in terms of the notion of some kind of distance between all the pairs of elements. Namely, if the distance between each x_i and x_j does not decrease, then the spread also does not decrease. The most natural choice of the distance measure in $\mathbb{I}$ is of course an ℓ^p-norm generated one, $d(x_i, x_j) = |x_j - x_i|$. However, according to the non-symmetry postulate, we should rather insist on checking whether the *signed* distance between each pair of observations of the first vector is not greater than the distance between the corresponding pairs from the second one, cf. Rothschild & Stiglitz (1970) for a well-known approach concerning increasing a spread of a probability distribution.

2.2. Definition

The above intuitions are reflected by the following binary relation $\preccurlyeq_n$ on $\mathbb{I}^n$. Given $\mathbf{x}, \mathbf{x}' \in \mathbb{I}^n$, we write $\mathbf{x} \preccurlyeq_n \mathbf{x}'$ and say that $\mathbf{x}$ *has not*