BASIC STATISTICAL METHODS FOR INTERVAL DATA

Federica Gioia, Carlo N. Lauro

Dipartimento di Matematica e Statistica, Università degli Studi di Napoli "Federico II". e-mail: fgioia@unina.it

Abstract

Real world data analysis is often affected by different type of errors as: measurement errors, computation errors, imprecision related to the method adopted for estimating the data (parameters).

The uncertainty in the data, which is strictly connected to the above errors, may be treated by considering, rather than a single value for each data, the interval of values in which it may fall: the interval data. This kind of data representation imposes a new formulation of the classical statistical methods in the case that interval-valued variables are considered. Accordingly, purpose of the present work is to develop suitable statistical methods for: obtaining a synthesis of the data, analysing the variability in the data and the existing relations among interval-valued variables.

The proposed solutions are based on the following assessments:

– *The developed statistics for interval-valued variables are intervals.*
– *Statistical methods for interval-valued variables embrace classical statistical methods as special cases.*
– *The proposed interval solutions do not contain redundant elements with respect to a given criterion.*

In the present work particular interest is devoted to the proof of the properties of the proposed techniques and to the comparison of the obtained results with those already existing in the literature.

Keywords: interval-valued variable , interval algebra

1. INTRODUCTION

The statistical modelling of many problems must account in the majority of cases of "errors" both in the data and in the solution. These errors may be classified as:

- *measurement errors:* the measured value x_i of a physical quantity x (e.g.; the temperature) may be different of the exact value of the quantity.

- *computation errors:* due to the finite precision of computers the numerical results are distorted by roundoff errors;
- *errors due to uncertainty in estimating parameters:* frequently there is uncertainty associated with the estimate of the parameters used in the computation as far as their value cannot be set precisely.

Interval algebra provides a powerful tool for determining the effects of uncertainties or errors and for accounting them in the final solution.

Interval mathematics deals with numbers which are not single values but *sets of numbers* ranging between a maximum and a minimum value. Those sets of numbers are the sets all possible determinations of the errors.

A form of interval algebra appeared for the first time in the literature in [*Burkill 1924*], [*Young, 1931*]; then in [*Sunaga, 1958*]. Modern developments of such an algebra were started by R.E. Moore.

Beyond its main application in optimisation theory, interval algebra becomes more and more applied in domains like: statistics, economics, engineering etc..

Nowadays, many numerical aspects are solved thanks to the modern powerful computers and to innovative and efficient numerical algorithms. In the recent, we noticed an upsurge of scientific contributions published in specialized reviews and new software dedicated to the treatment of data in different application domains.

From a statistical point of view, interval data have a great importance. It is known that statistical methods have been primarily developed for single valued variables. However, in real life there are many situations in which the adoption of single valued variables cause a loss of information. For example the study of daily temperature if registered by its range of variation during all the day, surely will give much more information on weather conditions with respect to the case in which a single value of synthesis, as for example the daily mean, is considered. Furthermore, nowadays , the huge amount of data to be handled (transactional database, internet, etc.) and the necessity to create handy synthesis, sometimes precludes a direct applications of classical methods of statistical data analysis. This leads also the necessity of handling different kind of data as: interval data, histograms, probability distributions etc.

The above reasons have prompted the development of new methodologies of statistical analysis for treating *interval-valued variables*, that is variables that may assume not a single value on the individual on which have been measured, but an interval of values.

Statistical indexes for interval-valued variables have been defined in [*Canal-Pereira,1998*] as *scalar* statistical summaries. This scalar indexes, may cause loss of information inherent in the interval data

For preserving the information contained in the interval data many researchers

and in particular Diday and his school of *Symbolic Data Analysis (SDA)* have developed some methodologies for interval data which provide *interval index solutions* that sometimes appear oversized as they include unsuitable elements.

An alternative methodology, is proposed by [*Marino-Palumbo, 2003*] with an approach which is typical for handling *imprecise data*. Taking into account the centre and the radius of each considered interval and the relations between these two quantities.

The approach that we propose in this paper, having previously analysed the applicability of the interval algebra tools [*Alefeld-Herzberger, 1983*], [*Neumaier, 1990*], [*Kerarfott-Kreinovich, 1996*]is to develop some statistics for which it is possible to prove properties in accordance to those corresponding to single-valued variables.

The proposed solutions are based on the following assessments.

The developed statistics for interval-valued variables are intervals.

Statistical methods for interval-valued variables embrace classical statistical methods as special cases.

The proposed interval solutions do not contain redundant elements with respect to a given criterion.

In section 1,2,3 of the present work some statistical indexes for interval-valued variables are introduced and the properties, in accordance to those corresponding to single-valued variables, are proved. In section 4 a new method for fitting a simple linear regression equation for interval-valued variables is developed.

Section 5 is an appendix devoted to some basic results of interval algebra and to the description of the optimisation algorithms used in the present paper.

2. MEAN OF INTERVAL-VALUED VARIABLES

Let us consider an *interval-valued* variable $\boldsymbol{X} = \left(X_i = \left[\underline{x}_i , \overline{x}_i \right] \right), \quad i = 1,...,n$ i.e. a variable which assumes an *interval* of values on each of *n* considered individuals:

$$\left\{ \left[\underline{x}_1 , \overline{x}_1 \right], \left[\underline{x}_2 , \overline{x}_2 \right], \cdots, \left[\underline{x}_n , \overline{x}_n \right] \right\} \tag{1.1}$$

The aim is to define a *position index* for $\boldsymbol{X}$, that is an interval which may give a synthesis about the set of values assumed by the variable.

Let us consider the following interval $\boldsymbol{M}$ with lower and upper bound, equal to the arithmetic means of the lower and upper bounds respectively of the intervals (1.1):

$$\mathbf{M} = \left[\frac{1}{n} \sum_{i=1}^{n} \underline{x}_i \quad ; \quad \frac{1}{n} \sum_{i=1}^{n} \overline{x}_i \right] = \left[\underline{M} , \overline{M} \right] \tag{1.2}$$

Some requirements of ***M*** will be analysed in order to identify it as the *interval mean* of the *interval-valued variable* ***X***[1].

It is important to remark that characteristics elements of an interval mean are its *position* (or centre) and its *radius*; the same centre may refers to interval means having different radius and the same radius may refers to interval means having the same centre.

Internality criterion or Cauchy mean

It is known that a *mean* of a single-valued variable is a value, which is representative of the distribution function of the variable, with the only condition to be included between the minimum and the maximum value assumed by the variable. This requirement, known as *internality or Cauchy criterion*, is not an operative solution but a fundamental characteristic of the mean, in as much as that there are infinite numbers between the minimum and maximum value of ***X***.

In the case of an *interval-valued* variable, taking into account how ***M*** has been constructed, it is:

$$\min_i \underline{x}_i \le \underline{M} \le \max_i \underline{x}_i$$

$$\min_i \overline{x}_1 \le \overline{M} \le \max_i \overline{x}_n$$

so

$$\forall m \in M, \ \min_i \overline{x}_1 \le m \le \max_i \overline{x}_n$$

all the numbers belonging to ***M*** are included between the minimum lower bound and the maximum upper bound of the set (1.1); any element in ***M*** satisfies the internality criterion so the whole *M* satisfies the same criterion.

Transferability criterion of the character

Let us show now that the interval ***M*** satisfies the following proposition:

---

[1] *The definition of* ***M*** *is the same of the one adopted in [Piccolo, 1998].*

Proposition 1

It exists a function *t* for which it is:

$$t\left(\left[\underline{x}_1,\overline{x}_1\right],\left[\underline{x}_2,\overline{x}_2\right],...,\left[\underline{x}_n,\overline{x}_n\right]\right)=t\left(\left[\underline{M},\overline{M}\right],\left[\underline{M},\overline{M}\right],...,\left[\underline{M},\overline{M}\right]\right)$$

Dim:

let *t* be the following interval function:

$$t\left(\left[\underline{x}_1,\overline{x}_1\right],\left[\underline{x}_2,\overline{x}_2\right],...,\left[\underline{x}_n,\overline{x}_n\right]\right)=\sum_{i=1}^{n}\left[\underline{x}_i,\overline{x}_i\right] \tag{1.3}$$

Using the interval algebra instruments (as described in the appendix), it follows:

$$t\left(\left[\underline{M},\overline{M}\right],\left[\underline{M},\overline{M}\right],...,\left[\underline{M},\overline{M}\right]\right)=\sum_{i=1}^{n}\left[\underline{M},\overline{M}\right]=\left[n\underline{M},n\overline{M}\right]=$$

$$=\left[\sum_{i=1}^{n}\underline{x}_i,\sum_{i=1}^{n}\overline{x}_i\right]=t\left(\left[\underline{x}_1,\overline{x}_1\right],\left[\underline{x}_2,\overline{x}_2\right],...,\left[\underline{x}_n,\overline{x}_n\right]\right).$$

This property, known as *transferability of the character*, shows thatn***M*** may be substituted to the observations $\left[\underline{x}_1,\overline{x}_1\right],\left[\underline{x}_2,\overline{x}_2\right],\cdots,\left[\underline{x}_n,\overline{x}_n\right]$ without changing the value of the function *t* (1.3).

Satisfying ***M*** both the internality criterion and the transferability of the character, similarly to the mean of a single-valued variable [*Piccolo, 1998*], it is properly a *mean* of the interval-valued variable ***X***.

Let us now analyse some properties of ***M***.

Proposition 2

M is the set of—*all and only* the arithmetic means of *n* elements each of which is chosen in a different interval of the set (1.1).

Dim:

without loosing in generality, let us consider the case of only two intervals X_1 and X_2.

Let A be the set of all possible arithmetic means between an element of X_1 and an element X_2:

$$A=\left\{a=\frac{x_1+x_2}{2}\quad/\quad x_1\in X_1,\,x_2\in X_2\right\}$$

let us prove that $A \equiv M$.

"$A \subseteq M$"

taken an element in A, let us show that it is also in $\boldsymbol{M}$.

$$\forall a \in A\,, \quad \exists\, x_1 \in X_1\,, x_2 \in X_2 \Big/ a = \frac{x_1 + x_2}{2}$$

For the following inequalities and taking into account the expression of $\underline{M}, \overline{M}$, we have:

$$\frac{\underline{x}_1 + \underline{x}_2}{2} \le \frac{x_1 + x_2}{2} \le \frac{\overline{x}_1 + \overline{x}_2}{2} \quad \Rightarrow \quad a \in \left[\, \underline{M}, \overline{M} \,\right].$$

"$\boldsymbol{M} \subseteq A$"

taken an element in $\boldsymbol{M}$, let us show that it is also in A.

$$\forall m \in \mathbf{M} \quad \Rightarrow \quad \frac{\underline{x}_1 + \underline{x}_2}{2} \le m \le \frac{\overline{x}_1 + \overline{x}_2}{2}\,.$$

Let us define as r_1 and r_2 the radii of X_1 and X_2 respectively.
Thus

$$\exists \quad h,k \quad \textit{under the condition} \quad 0 \le h + k \le r_1 + r_2 \Big/$$

$$m = \frac{(\underline{x}_1 + h) + (\underline{x}_2 + k)}{2} \Rightarrow m \in A$$

where h and k are chosen in order to have $(\underline{x}_1 + h) \in X_1 \quad e \quad (\underline{x}_2 + k) \in X_2$.

The proposition is completely proved.

Linearity

Taken two real numbers a,b, the *interval-valued* variable $a\boldsymbol{X}+b$ will have mean equal to $a\boldsymbol{M}+b$.

Dim.:

Let us prove the proposition in the cases $a>0$ and $a<0$. The case $a=0$ is trivial.
Consider the variable $a\boldsymbol{X}+b$ which assumes the following values:

$$\left\{ a\left[\underline{x}_1, \overline{x}_1\right] + b \;,\; a\left[\underline{x}_2, \overline{x}_2\right] + b \;,\ldots,\; a\left[\underline{x}_n, \overline{x}_n\right] + b \right\} \tag{1.4}$$

case $a>0$:

by the definition of the product and the sum between an interval and a scalar (as described in the appendix), the set (1.4) will be:

$$\left\{\left[a\underline{x}_1+b, a\overline{x}_1+b\right],\ \left[a\underline{x}_2+b, a\overline{x}_2+b\right],\ldots,\ \left[a\underline{x}_n+b, a\overline{x}_n+b\right]\right\}$$

using the introduced definition (1.2), the interval mean of the variable $aX+b$ is:

$$\frac{1}{n}\left[\sum_{i=1}^{n}\left(a\underline{x}_i+b\right), \sum_{i=1}^{n}\left(a\overline{x}_i+b\right)\right]=\left[\ a\underline{M}+b\ , a\overline{M}+b\ \right]=a\left[\ \underline{M}\ , \overline{M}\ \right]+b=a\mathbf{M}+b$$

case a<0:

in this special case the set (1.4) will be:

$$\left\{\left[a\overline{x}_1+b, a\underline{x}_1+b\right],\ \left[a\overline{x}_2+b, a\underline{x}_2+b\right],\cdots,\ \left[a\overline{x}_n+b, a\underline{x}_n+b\right]\right\}$$

the interval mean of the variable $a\boldsymbol{X}+b$ is:

$$\frac{1}{n}\left[\sum_{i=1}^{n}\left(a\overline{x}_i+b\right), \sum_{i=1}^{n}\left(a\underline{x}_i+b\right)\right]=\left[\ a\overline{M}+b\ , a\underline{M}+b\ \right]=a\left[\ \underline{M}\ , \overline{M}\ \right]+b=a\mathbf{M}+b$$

so the proposition is completely proved.

Deviations from mean

Let us introduce the interval deviation of the *i-th* value X_i of $\boldsymbol{X}$ from the mean $\boldsymbol{M}$.

The *i-th interval deviation* may be regarded as the *set* of all possible deviations of an element in $\left[\underline{x}_i\ ,\ \overline{x}_i\right]$ from all possible arithmetic means between the considered element, and *n-1* elements each of them chosen in a different interval of the remaining *n-1*.

It is important to notice that, while using interval algebra instruments for calculating the i-th deviation, the analogy to the case of single-valued variables does not apply. In fact considering the interval difference:

$$\left[\underline{x}_i, \overline{x}_i\right]-\left[\underline{M}, \overline{M}\right]=\left[\underline{x}_i-\overline{M}\ , \overline{x}_i-\underline{M}\right],\qquad i=1,\ldots,n \tag{1.5}$$

the result would *contains* the searched *set* described before; it easy to see, for example, that $\overline{x}_i-\underline{M}$ is included in (1.5) but it is meaningless because $\underline{M}$ is a mean of a set of elements not including $\overline{x}_i$. Furthermore the interval (1.5) may contain elements which do not respect the definition of deviation from mean.

To avoid this drawback the *i-th deviation* $\boldsymbol{SC}(X_i)$ of X_i from $\boldsymbol{M}$ must be calculated as:

$$\boldsymbol{SC}(X_i)=\left[\underline{x}_i-\frac{1}{n}\left(\underline{x}_i+\sum_{\substack{j=1\\ j\neq i}}^{n}\overline{x}_j\right)\quad,\quad \overline{x}_i-\frac{1}{n}\left(\overline{x}_i+\sum_{\substack{j=1\\ j\neq i}}^{n}\underline{x}_j\right)\right],\quad i=1,\ldots n \tag{1.6}$$

In this way $\boldsymbol{SC}(X_i)$ is the set of *all and only* the deviations between an element $x_i \in \left[\underline{x}_i\,,\,\overline{x}_i\right]$ and the arithmetic mean of *n* elements in which x_i is included. We will indicate the vector of the deviations of $\boldsymbol{X}$ as:

$$\boldsymbol{SC}(\boldsymbol{X})=(\boldsymbol{SC}(X_i)),\qquad i=1,\ldots,n$$

Sum of deviations is nil

It is known that for a numerical variable the sum of the deviations from the mean is nil. The aim is to extend this property also to the case in which $\boldsymbol{X}$ is an interval-valued variable.

It is important to notice that the interval sum of $\boldsymbol{SC}(X_1)$, $\boldsymbol{SC}(X_2)$, …, $\boldsymbol{SC}(X_n)$ is ***not*** nil; in fact by the definition of interval sum:

$$\boldsymbol{SC}(X_1)+\boldsymbol{SC}(X_2)+\ldots+\boldsymbol{SC}(X_n)=0\Leftrightarrow$$
$$\forall\, s_i\in\boldsymbol{SC}(X_i)\,,\quad \forall\, s_j\in\boldsymbol{SC}(X_j)\,,\ j=1,\ldots,n\quad ,\ j\neq i\,/ \tag{1.7}$$
$$s_i+\sum_{\substack{j=1\\ j\neq i}}^{n}s_j=0$$

that is: chosen ***any*** element s_i in $\boldsymbol{SC}(X_i)$ and chosen ***any*** *n-1* elements in the remaining intervals, the sum of those elements is nil.

It is easy to see that the (1.7) would be satisfied only in the trivial case in which all the components of $\boldsymbol{SC}(\boldsymbol{X})$ are *thin* intervals[2]. Instead, according to the way $\boldsymbol{SC}(\boldsymbol{X})$ has been built (1.6), following statement holds true:

$$\forall\, s_i\in\boldsymbol{SC}(X_i)\,,\quad \exists\, s_j\in\boldsymbol{SC}(X_j)\,,\ j=1,\ldots,n\quad ,\ j\neq i\,/$$
$$s_i+\sum_{\substack{j=1\\ j\neq i}}^{n}s_j=0 \tag{1.8}$$

that is: chosen an element s_i in $\boldsymbol{SC}(X_i)$ there ***exist*** *n-1* elements in the remaining intervals which, summed to s_i, give a sum equal to zero.

In other words: chosen s_i in $\boldsymbol{SC}(X_i)$ the *n-1* elements s_1 , s_2 ,…, s_{i-1} , s_{i+1} ,…,

---

[2] *An interval is said to be thin if it has lower and upper bound equal one each other.*

s_n, cannot be picked up arbitrarily in their intervals as in (1.7), instead they must be *chosen* in a definite way in order to satisfy condition $s_i + \sum_{\substack{j=1 \\ j \neq i}}^{n} s_j = 0$.

With analogy to the case of numerical variables, condition (1.8) implies that the intervals $\boldsymbol{SC}(X_1), \boldsymbol{SC}(X_2), \ldots, \boldsymbol{SC}(X_n)$, are positioned around the zero; in particular for the case of only two intervals X_1, X_2, the relative deviations $\boldsymbol{SC}(X_1)$, $\boldsymbol{SC}(X_2)$ will have a *symmetric* position with respect to the origin.

Example

Let us consider an interval-valued variable $\boldsymbol{X}$ which assumes the following two values:

$$[1,4] \ , \ [11,12]$$

the interval mean will be:

$$\boldsymbol{M} = [6,8]$$

The relative deviations from M are:

$$\boldsymbol{SC}\left(\left[1,4\right]\right) = \left[1 - \frac{1+12}{2} \ , 4 - \frac{4+11}{2}\right] = \left[-5.5 \ , -3.5\right]$$

$$\boldsymbol{SC}\left(\left[11,12\right]\right) = \left[11 - \frac{11+4}{2} \ , 12 - \frac{12+1}{2}\right] = \left[3.5 \ , 5.5\right]$$

we can see that the two intervals are *symmetrically* positioned around the origin.

A numerical example

In this example we will consider an interval data set [*Billard-Diday 1*], in which the pulse rate, the blood systolic pressure and the diastolic pressure have been measured on 11 patients (taken from Raju(1997)).

The vector of the means and the deviations of each observation from the mean will be computed here below.

Pulse rate	*Systolic pressure*	*Diastolic pressure*
[44 , 68]	*[90 , 100]*	*[50 , 70]*
[62 , 72]	*[90 , 130]*	*[70 , 90]*
[56 , 90]	*[140 , 180]*	*[90 , 100]*
[70 , 112]	*[110 , 142]*	*[80 , 108]*
[54 , 72]	*[90 , 100]*	*[50 , 70]*
[70 , 100]	*[130 , 160]*	*[80 , 110]*
[63 , 75]	*[60 , 100]*	*[140 , 150]*
[72 , 100]	*[130 , 160]*	*[76 , 90]*
[76 , 98]	*[110 , 190]*	*[70 , 110]*
[86 , 96]	*[138 , 188]*	*[90 , 110]*
[86 , 100]	*[110 , 150]*	*[78 , 100]*

Vector of the Means

$$\left[[67.18\ ,89.36\ ] \quad [\ 108.90\ ,\ 145.45\ ] \quad [\ 79.45\ ,\ 100.72] \right]$$

Some deviations from mean

$$\begin{bmatrix} [-43.18\ ,-1.363\ ] & [\ -54.54\ ,-9.81] & [\ -48.90\ ,-11.27] \\ [-26.45,3.90\ ] & [\ -51.81\ ,17.45\ ] & [\ -28.90\ ,8.72\ ] \\ [\ -30.27\ ,19.72\ ] & [\ -1.81\ ,\ 67.45\ ] & [\ -9.81\ ,19.63\ ] \\ \vdots & \vdots & \vdots \\ [\ -11.36\ ,28.81\ ] & [\ -28.18\ ,\ 73.81\ ] & [\ -27.09\ ,26.90] \\ [\ -2.45\ ,27.90\ ] & [\ -2.90\ ,\ 74.54\ ] & [\ -8.90\ ,28.72\ ] \\ [\ -2.09\ ,31.54] & [\ -31.81\ ,37.45\ ] & [\ -20.72\ ,18.54\ ] \end{bmatrix}$$

2. VARIANCE AND VARIATION COEFFICIENT OF AN INTERVAL-VALUED VARIABLE

Let us consider a numerical variable:

$$X = \left(x_i \right) \quad i = 1,...,n$$

It is known that the variance of X may be computed as follow:

$$var(X) = \frac{1}{n} \sum_{1=1}^{n} (x_i - M)^2 \tag{2.1}$$

where M is the mean of the variable X.

The aim is to extend equation (2.1) to variables of type interval in order to have, also for this case, a measure of the variability of the phenomenon, by referring to the position of the statistical units with respect to the mean $\boldsymbol{M}$.

Let as consider now the following interval-valued variable :

$$\boldsymbol{X} = \left(X_i = \left[\underline{x}_i\ ,\ \overline{x}_i \right] \right), \qquad i = 1,...n$$

and the respective deviation interval valued variable:

$$\boldsymbol{SC}(\boldsymbol{X}) = \left(\boldsymbol{SC}(X_i) \right)_i , \quad i = 1,...n$$

In analogy with eq. (2.1), the first idea for computing the variance of $\boldsymbol{X}$ would be to use interval algebra for calculating the following interval sum:

$$\frac{1}{n}\sum_{1=1}^{n} \boldsymbol{SC}\left(X_i\right)^2 \tag{2.2}$$

We note that eq. (2.2) gives an interval which *contains* the *solution set* that we are searching for i.e., the set of *all and only* the variances that can be calculated when each component of the variable varies in its interval.

To avoid this drawback different instruments, with respect to those of interval algebra, should be used; as we will see in this section, the variance of an interval-valued variable may be reached solving a *minimization/maximization* problem of a multiple variable function *f*. The solution set, obtained by a numerical algorithm, will be the interval of all and only the values of function (2.1) when each variable ranges in its interval of values.

Let us write the (2.1) in the following alternative way:

$$var\left(X\right) = f\left(x_1,...x_n\right) = \frac{1}{n}\sum_{h=1}^{n}\left(x_h - \frac{1}{n}\sum_{k=1}^{n} x_k \right)^2 \tag{2.3}$$

When $\boldsymbol{X}$ is of type interval the idea is to compute its *interval variance* by minimizing/maximizing function (2.3), i.e. calculating the following set:

$$\boldsymbol{Var}(\mathbf{X}) = \left[\min_{\substack{x_i \in X_i \\ i=1,..n}} f(x_1,...,x_n), \quad \max_{\substack{x_i \in \boldsymbol{X}_i \\ i=1,..n}} f(x_1,...,x_n) \right] \tag{2.4}$$

where, for definition, X_i is the *i-th* component of the interval-valued variable $\boldsymbol{X}$.

Let us describe set (2.4): *a) f* is a continuous function on a *conness* set: Bolzano's theorem assures that the set (2.4) is an *interval*. *b) f* is a continuous function on a *compact* set: Weierstrass's assures that *f* has *global extremes* on that set.

By properties *a), b)* assure us that the set (2.4) is a *closed interval*, in particular it is the interval of *all and only* the variances that may be computed when each component of the variables ranges in its interval of values.

Furthermore let us summarize here below some important requirements of $\boldsymbol{Var}(X)$ in order to assume it as the *variance* of an interval-valued variable.

1) *Non negative interval*

   Due to the expression of function (2.3), ***Var****(X)* will be a set of sums of squared elements, thus the elements belonging to ***Var****(X)* are all positive.

2) *The set of variances*

   As we had already seen, ***Var****(X)* is the interval of *all and only* the variances that may be calculated when each component of the variable varies in its interval of values.

3) *Values in [0,∞[*

   Var*(X)* is a set of elements that varies from a minimum of zero (when a single element is considered) trough a maximum which depends on the particular case considered and that may be infinitely large if there are elements infinitely distant from their mean.

Thus we may identify ***Var****(X)* as the *variance* of the interval-valued variable X^3.

The interval standard deviation ***Std*****(*****X*****)** will be computed as in (2.4) substituting $f^{1/2}$ to f.

It is known from classical theory that the variance is an absolute index so it depends from the scale of the analysed phenomenon. It is important to introduce also in the case of interval-valued variables, a relative normalized index of variation i.e. an index which is a pure number used for comparing the variability of a phenomenon which is observed in different conditions.

Let us write the classical *variation coefficient Cv* for numerical variables in the following way:

$$cv(X) = l(x_1,...x_n) = \frac{f(x_1,...x_n)}{\frac{1}{n}\cdot\sum_{k=1}^{n} x_k}$$

where $f(x_1,...x_n)$ is the "variance" function in (2.3) and where is supposed to be.

When ***X*** is of type interval, as in the case of interval variance, the *interval variation coefficient* ***Cv(X)*** may be computed as follow:

$$\frac{1}{n}\cdot\sum_{k=1}^{n} x_k > 0 \qquad (2.5)$$

function $l(x_1,...,x_n)$ verify the conditions *a)* and *b)* thus set (2.5) is the interval of

---

[3] *As in [Piccolo, 1998*

all and only the variation coefficients that may be computed when each component of the variables ranges in its interval of values.

It is important to remark that the *interval variation coefficient Cv(**X**)* it is defined only when the interval mean ***M*** of ***X*** is an interval with positive lower bound.

A numerical example

Let us consider the interval matrix introduced in section 1 and let us compute the interval *variance*, the interval *standard deviation* and the interval *variation coefficient* vectors with respect to the considered interval-valued variables.

Variance:

$$\left[[26.86 , 390.50] \quad [193.03 , 1632.06] \quad [298.18 , 737.71] \right]$$

Standard deviations

$$\left[[5.18 , 19.76] \; [13.89 , 40.39] \; [17.26 , 27.16] \right]$$

Variation coefficient

$$\left[[0.05 , 0.23] \quad [0.12 , 0.21] \quad [0.19 , 0.31] \right]$$

Rodriguez and Diday (2000) give an axiomatic definition of some interval indexes for interval-valued variables. It is worst to notice that in some cases the results obtained by Rodriguez and Diday (for example the variance) may be oversized with respect to a

proper interval solution relative to a chosen function, which is the interval containing all possible values assumed by the considered function when the observed values vary in their own interval of values.

Considering the results obtained for the standard deviation intervals, we can asses that the method that we have introduced in this paper is an improvement with respect to that used in [*Rodriguez, 2000*] in as much as the intervals calculated here are *narrower* than those calculated in [*Rodriguez, 2000*].

In analogy to the case of classical statistical variables, by the interval variation coefficients it is easier, with respect to the interval standard deviations, a comparison of two different interval-valued variables.

3. COVARIANCE AND CORRELATION OF INTERVAL-VALUED VARIABLES

Let us consider two numerical variables:

$$X_r = \left(x_{ir}\right), \quad X_s = \left(x_{is}\right), \quad i = 1,...,n$$

It is known that the covariance between X_r and X_s may be computed as follow:

$$\operatorname{cov}\left(X_r, X_s\right) = \frac{1}{n}\sum_{1=1}^{n}\left(x_{ir} - M_r\right)\cdot\left(x_{is} - M_s\right) \tag{3.1}$$

where M_r and M_s are respectively the mean of the *r-th* and the *s-th* variables.

Again the aim is to extend equation (3.1) to variables of type interval.

Let as consider the following interval-valued variables:

$$\mathbf{X}_r = \left(X_{ir} = \left[\underline{x}_{ir}, \overline{x}_{ir}\right]\right), \quad \mathbf{X}_s = \left(X_{is} = \left[\underline{x}_{is}, \overline{x}_{is}\right]\right)_i \quad i = 1,...n$$

The respective deviation interval vectors will be:

$$\boldsymbol{SC}(\boldsymbol{X}_r) = \left(\boldsymbol{SC}(X_{ir})\right)_i \quad , \quad \boldsymbol{SC}(\boldsymbol{X}_s) = \left(\boldsymbol{SC}(X_{is})\right)_i , \quad i = 1,...n$$

In analogy with (3.1), using interval algebra for calculating the covariance:

$$\frac{1}{n}\sum_{1=1}^{n}\mathbf{SC}\left(X_{ir}\right)\cdot\mathbf{SC}\left(X_{is}\right) \tag{3.2}$$

it is easy to understand that (3.2) *contains* the set of all and only the covariances that may be calculated when each component of the variables varies in its interval and further more elements.

As for the calculus of the variance, the covariance of an interval-valued variable may be reached insted solving a *minimization/maximization* problem of a function *g* of multiple variables, i.e. calculating the following set:

$$\boldsymbol{Cov}(\mathbf{X}_r, \mathbf{X}_s) = \left[\begin{array}{ll} \min\limits_{\substack{x_{ir}\in X_{ir}\\ x_{is}\in X_{is}\\ i=1,..,n}} g\left(x_{1,r},...x_{n,r}; x_{1,s},...x_{n,s}\right) & , \quad \max\limits_{\substack{x_{ir}\in X_{ir}\\ x_{is}\in X_{is}\\ i=1,..,n}} g\left(x_{1,r},...x_{n,r}; x_{1,s},...x_{n,s}\right) \end{array}\right] \tag{3.3}$$

where X_{ir} and X_{is} are the *i-th* components of the interval-valued variables X_r and X_s respectively, and the function *g* has the following mathematical expression:

$$g\left(x_{1,r},...x_{n,r}; x_{1,s},...x_{n,s}\right) = \frac{1}{n}\cdot\sum_{i=1}^{n}\left(\left(x_{ir} - \frac{1}{n}\cdot\sum_{k=1}^{n}x_{kr}\right)\cdot\left(x_{is} - \frac{1}{n}\cdot\sum_{k=1}^{n}x_{ks}\right)\right) \tag{3.4}$$

As for the function *f* in (2.3), also the function *g* in (3.4) satisfies the properties *a)* and *b)* of 2, i.e. *g* is a continuous function on a conness and compact set; therefore the set (3.3) is a closed *interval*, in particular it is the interval of *only and all* the covariances that we may compute when each component of the variables ranges in its interval of values.

Thus we may identify $\boldsymbol{Cov}(\boldsymbol{X}_r\,\boldsymbol{X}_s)$ as the *interval covariance* between $\boldsymbol{X}_r$ and $\boldsymbol{X}_s$.

It is important to remark that it is:

$$\begin{aligned} &\boldsymbol{Var}\left(\boldsymbol{X}\right) = \boldsymbol{Cov}\left(\boldsymbol{X},\boldsymbol{X}\right) \\ &\boldsymbol{Cov}\left(\boldsymbol{X}\right) \subset \left]-\infty,+\infty\right[\end{aligned} \tag{3.5}$$

perfectly in accordance to the case of single-valued variables.

Analogously it is known that the correlation between two numerical variables it may be written as follow:

$$corr(X_r, X_s) = h(x_{1,r},...x_{n,r}; x_{1,s},...x_{n,s}) = \frac{cov(X_r, X_s)}{\sqrt{var(X_r)}\sqrt{var(X_s)}}$$

In the case in which the two variables are of type interval the idea for calculating the correlation between them is the same as that one adopted for computing the relative variance and covariance i.e., solving a *minimization/ maximization* problem of a multiple variable function, in particular computing the following set:

$$\boldsymbol{Corr(X}_r\,,\,\boldsymbol{X}_s) = \left[\begin{array}{ll} \underset{\substack{x_{i\,r}\in X_{ir}\\ x_{is}\in X_{is}\\ i=1,...n}}{min}\; h(x_{1,r},...x_{n,r}; x_{1,s},...x_{n,s}) & , \quad \underset{\substack{x_{ir}\in X_{ir}\\ x_{is}\in X_{is}\\ i=1,...n}}{max}\; h(x_{1,r},...x_{n,r}; x_{1,s},...x_{n,s}) \end{array} \right] \tag{3.6}$$

Taking into account the expression and the characteristics of function *h* the set (3.6) satisfies the following requirements:

1) $\boldsymbol{Corr(X}_r\,,\,\boldsymbol{X}_s)$ is the interval of *all and only* the correlations that we can be calculated when the variables vary in their respective ranges of variations.
2) Any element belonging to $\boldsymbol{Corr(X}_r\,,\,\boldsymbol{X}_s)$ is between *–1* and *1*.
3) $\boldsymbol{Corr(X}\,,\,\boldsymbol{X})=1$.

We may assume $\boldsymbol{Corr(X}_r\,,\,\boldsymbol{X}_s)$ as the *correlation* between $\boldsymbol{X}_r$ and $\boldsymbol{X}_s$.

A numerical example

In this example we will consider the Oil's data set (Ichino 1988) largely used in SDA applications, whose characteristics are well-known to people working on the topic. The matrix presents eight different classes of oils described by five variables, we only refer to four quantitative variables: "Specific gravity", "Freezing point", "Iodine value" and "Saponification".

The covariance matrix, the correlation matrix and the vector of the standard deviations will be calculated in this section.

Tab. 1: Oil's matrix

	Spec.gravity	*Freezing point*	*Iodine value*	*Saponification*
Linseed	*[0.93 , 0.94]*	*[-27 , -18]*	*[170 , 204]*	*[118 , 196]*
Perilla	*[0.93 , 0.94]*	*[-5 , -4]*	*[192 , 208]*	*[188 , 197]*
Cotton	*[0.92 , 0.92]*	*[-6 , -1]*	*[99 , 113]*	*[189 , 198]*
Sesame	*[0.92 , 0.93]*	*[-6 , -4]*	*[104 , 116]*	*[187 , 193]*
Camellia	*[0.92 , 0.92]*	*[-21 , -15]*	*[80 , 82]*	*[189 , 193]*
Olive	*[0.91 , 0.92]*	*[0 , 6]*	*[79 , 90]*	*[187 , 196]*
Beef	*[0.86 , 0.87]*	*[30 , 38]*	*[40 , 48]*	*[190 , 199]*
Hog	*[0.86 , 0.86]*	*[22 , 32]*	*[53 , 77]*	*[190 , 202]*

The interval mean and standard deviation vectors, and the covariance and correlation matrices are presented here below:

Means

$$\left[\ [\,0.90\ ,\ 0.91\,]\quad [\,-1.62\ ,\ 4.25\,]\quad [\,102.12\ ,\ 117.25\,]\quad [\,179.75\ ,\ 196.75\,]\ \right]$$

Standard deviations

$$\left[\ [\,0.02\ ,\ 0.03\,]\quad [\,15.61\ ,\ 21.59\,]\quad [\,45.33\ ,\ 59.77\,]\quad [\,0\ ,\ 26.22\,]\ \right]$$

Covariance matrix

$$\begin{bmatrix} [\,0.00\ ,\ 0.00\,] & & & \\ [\,-0.60\ ,-0.35\,] & [\,243.96\ ,\ 466.48\,] & & \\ [\,0.76\ ,\ 1.49\,] & [\,-920.48\ ,-408.68\,] & [\,2055.10\ ,3573.23\,] & \\ [\,-0.43\ ,\ 0.06\,] & [\,-47.12\ ,363.06\,] & [\,-1093.45\ ,\ 176.26\,] & [\,4.520e-01\ ,\ 688\,] \end{bmatrix}$$

Correlation matrix

$$\begin{bmatrix} [\,1\,,\,1\,] & & & \\ [\,-0.99\,,-0.74\,] & [\,1\,,1\,] & & \\ [\,0.57\,,\,0.94\,] & [\,-0.86\,,-0.38\,] & [\,1\,,1\,] & \\ [\,-0.99\,,-0.02\,] & [\,0.02\,,\,0.99\,] & [\,-0.99\,,-0.03\,] & [\,1\,,1\,] \end{bmatrix}$$

1) Notice that all interval correlations are intervals which do not contain the zero.

2) The correlation matrix shows intervals which present a "small" absolute radius, as in the case of the interval correlation between *Specific gravity* and *Freezing point* variables, but also intervals with a "big" radius as in the case of the correlation of each variable with the *Saponification* one.

4. INTERVAL SIMPLE REGRESSION

In this section, we propose an extension of simple linear regression to the case of interval data. We assume that a single independent *interval-valued variable X* is used for predicting the dependent *interval-valued variable Y* according to a linear relationship.

Let us indicate with X^I and Y^I respectively the independent and the dependent *interval-valued variables*, which assume the following interval values on the *n* statistical units chosen for our experiment :

$$X^I = \left(X_i = \left[\underline{x}_i \,,\, \overline{x}_i \right] \right), \qquad i = 1,...n$$

$$Y^I = \left(Y_i = \left[\underline{y}_i \,,\, \overline{y}_i \right] \right), \qquad i = 1,...n$$

X^I and Y^I assume an *interval of values* on each statistical units, i.e. with difference to the case of single-valued variables, we don't know the exact value x_i or y_i but only the *range* in which this value falls.

In the proposed approach the idea is to contemplate *all possible values* of the components x_i, y_i each of which is in its interval of values $X_i = \left[\underline{x}_i \,,\, \overline{x}_i \right]$, $Y_i = \left[\underline{y}_i \,,\, \overline{y}_i \right]$ for $i=1,...n$.

For each different set of values $x_1,x_2,...,x_n$, $y_1,y_2,...,y_n$, chosen from each interval, a different cloud on points of the plane is univocally determined and the respective regression line is asked to be computed.

For example let us suppose for simplicity that $n=3$, which means that only *3*

observations are available for our experiment. Let us consider two different cases corresponding to two different set of values assumed by the components $x_1,x_2,...,x_3$, y_1,y_2,y_3,; these values will be indicated with $x_1^{(1)},x_2^{(1)},x_3^{(1)},y_1^{(1)},y_2^{(1)},y_3^{(1)}$ and $x_1^{(2)},x_2^{(2)},x_3^{(2)},y_1^{(2)},y_2^{(2)},y_3^{(2)}$:

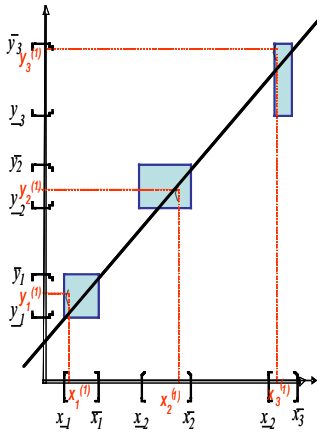

Fig. 4.1: cloud of points 1.

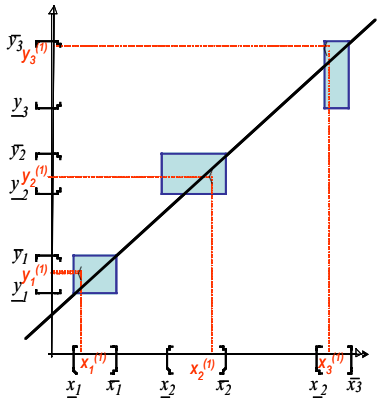

Fig. 4.2: Cloud of points 2.

Obviously we have two different regression lines (***Fig4.1***, ***Fig4.2***), with two different slopes and two different intercepts, corresponding to the represented clouds of points which are univocally determinate by the different values assumed by the components.

The task of the proposed methodology is to compute the *set* of slopes and the *set* of intercepts of all possible regression lines that we may have for each value of x_i, y_i each of which in its interval of values $X_i = \left[\underline{x}_i, \overline{x}_i\right]$, $Y_i = \left[\underline{y}_i, \overline{y}_i\right]$ for $i=1,...n$.

Thus *making regression* between two interval-valued variables means to compute the *set of regression lines* each of which realises the best fit, in the sense of Minimum Least Square, of a set of points in the plane. This set of points is univocally determined each time the components $x_1,x_2,...,x_n$, $y_1,y_2,...,y_n$, take a particular value in their own interval of variation.

Mathematically to compute the *interval regression line* between two interval-valued variables X^I and Y^I is equivalent to compute the following two sets:

$$\hat{\boldsymbol{\beta}}_1^I = \left\{ \hat{\beta}_1\left(x_1, x_2, ...x_n, y_1, y_2, ..., y_n\right) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \Big/ x_i \in \left[\underline{x}_i, \bar{x}_i\right], y_i \in \left[\underline{y}_i, \bar{y}_i\right] \quad , \quad i = 1,...,n \right\} \tag{4.5}$$

$$\hat{\boldsymbol{\beta}}_0^I = \left\{ \hat{\beta}_0\left(x_1, x_2, ...x_n, y_1, y_2, ..., y_n\right) = \bar{y} - \hat{\beta}_1 \bar{x}, \quad x_i \in \left[\underline{x}_i, \bar{x}_i\right], y_i \in \left[\underline{y}_i, \bar{y}_i\right] \quad , \quad i = 1,...,n \right\} \tag{4.6}$$

where $\bar{x}$ and $\bar{y}$, regarded as functions of $x_1, x_2, ..., x_n, y_1, y_2, ..., y_n$,, are given by (4.3).

Sets (4.5) and (4.6) are respectively the set described by the slopes and the set described by the intercepts of all regression lines varying x_i, y_i in their own interval of values $X_i = \left[\underline{x}_i, \bar{x}_i\right], Y_i = \left[\underline{y}_i, \bar{y}_i\right]$ for $i=1,...n$. These sets may be computed numerically by solving some optimization problems, i.e. searching for the minimum and for the maximum of functions $\hat{\beta}_1(x_1, x_2, ...x_n, y_1, y_2, ..., y_n)$ and $\hat{\beta}_0(x_1, x_2, ...x_n, y_1, y_2, ..., y_n)$ in (4.5) and (4.6).

These functions are both continuous[4] on a conness and compact set and this assures that sets (4.5) and (4.6) are the following closed intervals:

$$\hat{\boldsymbol{\beta}}_1^I = \left[\min_{\substack{x_i \in [\underline{x}_i, \bar{x}_i] \\ y_i \in [\underline{y}_i, \bar{y}_i] \\ i=1,...,n}} \hat{\beta}_1 \quad , \quad \max_{\substack{x_i \in [\underline{x}_i, \bar{x}_i] \\ y_i \in [\underline{y}_i, \bar{y}_i] \\ i=1,...,n}} \hat{\beta}_1 \right] \tag{4.7}$$

$$\hat{\boldsymbol{\beta}}_0^I = \left[\min_{\substack{x_i \in [\underline{x}_i, \bar{x}_i] \\ y_i \in [\underline{y}_i, \bar{y}_i] \\ i=1,...,n}} \hat{\beta}_0 \quad , \quad \max_{\substack{x_i \in [\underline{x}_i, \bar{x}_i] \\ y_i \in [\underline{y}_i, \bar{y}_i] \\ i=1,...,n}} \hat{\beta}_0 \right] \tag{4.8}$$

The *interval regression line* may be written as:

$$y = \hat{\boldsymbol{\beta}}_0^I + \hat{\boldsymbol{\beta}}_1^I \, x \tag{4.9}$$

and may be interpreted as follow:

chosen an intercept $\hat{\beta}_0$ in the interva $\hat{\boldsymbol{\beta}}_0^I$ it *exists* a slopen $\hat{\beta}_1$ in $\hat{\boldsymbol{\beta}}_1^I$ so that the regression line:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{4.10}$$

1) is the unique line that realises the best fit, in the sense of Minimum Least Square, of a given set of points (x_1 , y_1), (x_2 , y_2),…, (x_n , y_n) in the plane $\left(x_i \in \left[\underline{x}_i , \overline{x}_i\right], y_i \in \left[\underline{y}_i , \overline{y}_i\right], \quad i = 1,.....,n\right)$; i.e.

   indicating by: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. and $\hat{e}_i = y_i - \hat{y}_i$ the estimate value and the residual with respect to the observed value x_i respectively, equation (4.10) is the only line which *minimises* the sum of squared residuals.

$$\sum_{i=1}^{n} \hat{e}_i^2$$

2) the couple $(\overline{x} , \overline{y})$ of means:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i ; \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

   satisfies eq. (4.10).

It is important to notice that for each $\hat{\beta}_0 \in \hat{\boldsymbol{\beta}}_0^I$ it exists at least one $\hat{\beta}_0 \in \hat{\boldsymbol{\beta}}_1^I$ so that $y = \hat{\beta}_0 + \hat{\beta}_1 x$ verifies properties *1)* and *2)*, but this is *not* true for each chosen couple of parameters.

This means that, given an interval value $\left[\underline{x}_i , \overline{x}_i\right]$ of the independent variable X^I, the prevision $\left[\underline{y}_i , \overline{y}_i\right]$ by means of equation (4.10) of the interval value assumed by the dependent variable Y^I:

---

[4] *According to the definition of $\hat{\beta}_1$ in equation (4.5), the denominator $\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$ could be nil only in the case in which: $x_1 = x_2 = \cdots = x_n = \overline{x}$ but this would contradict the hypothesis that at least two different observations must be available in the experiment.*

$$\left[\underline{y}_i , \overline{y}_i\right] = \hat{\boldsymbol{\beta}}_0^I + \hat{\boldsymbol{\beta}}_1^I \left[\underline{x}_i , \overline{x}_i\right]$$

is oversized with respect to the set of all.

Numerical results

In this section some numerical results, relative to the interval regression, are presented.

In [*Billard-Diday, 2000*], [*Billard-Diday, 2002*] and [*Rodriguez, 2000*] the authors derive the results as a combination of two different regression equations for single-valued variables.

An alternative methodology, is proposed by [*Marino-Palumbo, 2003*] with an approach which is typical for handling *imprecise data*. Taking into account the centre and the radius of each considered interval and the relations between these two quantities.

Simulated data

Figures: "limit cases"

In order to show the good agreement between the proposed data and the result of our *interval regression model* we analyse two emblematic cases of ***Fig. 4.3*** and ***Fig4.4***.

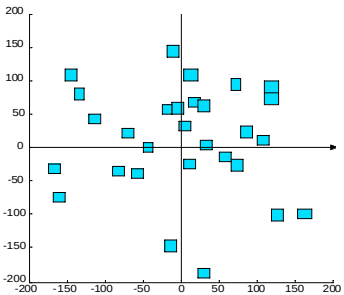

Fig. 4.3

BETA:
B0=[1.154,18.680]
B1= [-0.192,0.026]

CORR =[-0.243,0.034]

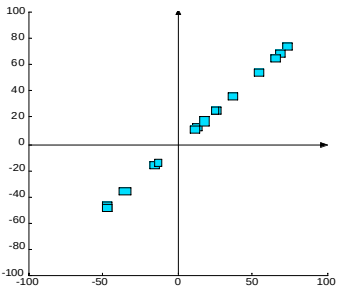

Fig. 4.4

BETA :
B0=[-6.000,5.999]
B1= [0.885,1.127]

CORR =[1,1]

Figures: "limit cases"

A cloud of rectangles rather dispersed and clearly not correlated is reported in ***Fig4.3***. The application of the proposed methodology produces a correlation range that clearly confirms the poor correlation of the data analysed.

On the contrary rectangles which are visually strongly correlated are reported in ***Fig4.4***. The application of the method produces a correlation which confirms this strong correlation and a range of the regression coefficient which are well in agreement with the position of the rectangles.

Figures "shift of a rectangle"

Some numerical results are proposed in order to analyse the interval regression coefficients and the interval correlation of some clouds of rectangles in which one rectangle is changing its position with respect to the second axis.

Also in this case the regression coefficient and the correlation between the variables are well in agreement with the position of the rectangles presented in ***Fig4.5***.

Furthermore changing the position as in ***Fig:4.6-4.7-4.8-4.9-4.10***, the regression coefficient and the correlation, if compared with the solutions of ***Fig4.5***, are intervals with a greater radius; this is perfectly in agreement with the increasing "variability" of the considered cloud of the rectangles.

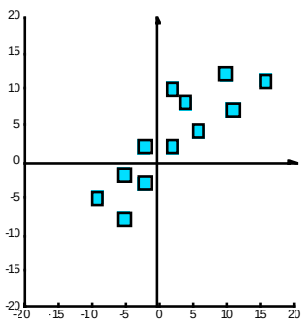

Fig. 4.5.

BETA: B0=[-1.139 , 2.019]
B1=[0.294 , 0.720]

CORR =[0.291 , 0.645]

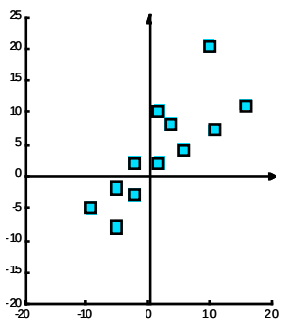

Fig. 4.6.

BETA :B0=[-0.863 , 2.386]
B1=[0.381 , 0.808]

CORR =[0.440 , 0.772]

Figures: "shift of a rectangle"

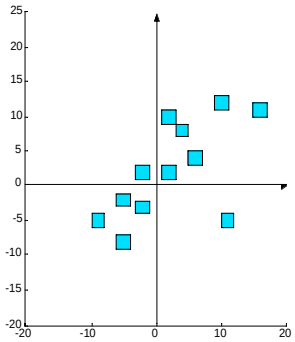

Fig. 4.7.

BETA :B0=[-0.863 , 2.386]
B1=[0.381 , 0.808]

CORR =[0.440 , 0.772]

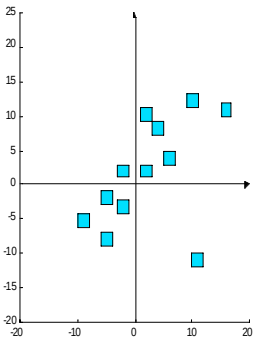

Fig. 4.8.

BETA :B0=[-1.139 , 2.019]
B1=[0.294 , 0.720]

CORR =[0.291 , 0.645]

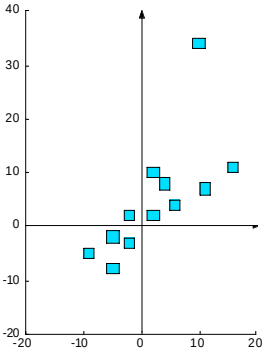

Fig. 4.9.

BETA :B0=[0.500 , 4.624]
B1=[0.773 , 1.327]

CORR =[0.571 , 0.811]

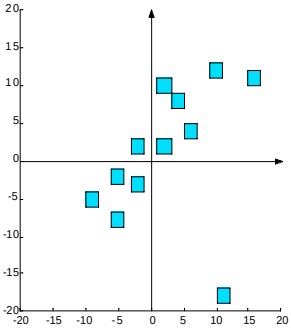

Fig. 4.10.

BETA :B0=[-1.460 , 1.606]
B1=[0.177 , 0.625]

CORR =[0.143 , 0.499]

Figures "different forms"

The regression coefficients and the interval correlation are computed for two clouds of rectangles having different *forms* (***Fig:4.11-4.12***) with respect to those presented in ***Fig4.5***.

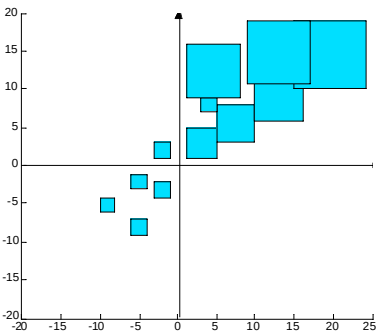

Fig. 4.5.

BETA :B0=[-1.174 , 5.00]
B1=[0.399 ,1.37]

CORR =[0.552 , 0.995]

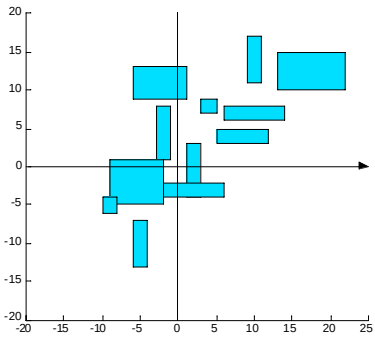

Fig. 4.6.

BETA :B0=[-1.91 , 5.5]
B1=[0.23 , 1.38]

CORR =[0.27 , 0.95]

Figures:"different forms"

In both situations the computed intervals, relative to the regression coefficients and to the correlation between the variables, are intervals which present a bigger radius according to the visible higher variability of the problem.

Real data

In the following example [*Marino-Palumbo, 2003*] we will take into account the relationship between the weaving of ground and the water retention. We will compute a simple regression involving the water retention as dependent variable and the bulk density (for the weaving of ground) as independent one. Let us consider the following (*49◊2)* interval data matrix:

Water retention min	*Water retention max*	*Bulk density min*	*Bulk density max*
0.5885	0.6127	0.9609	0.9632
0.6261	0.6261	0.9350	0.9350
0.6120	0.6222	0.9081	0.9799
0.5371	0.5717	1.0461	1.0885
.	.	.	.
.	.	.	.
.	.	.	.
0.4617	0.4652	1.2292	1.2479
0.4310	0.4410	1.3194	1.3667
0.4482	0.4656	1.2156	1.2418
0.4931	0.4931	1.1345	1.1345

By applying the proposed approach for computing the interval regression line, we find the following interval intercepts and interval slope: β_0= *[0.79, 1.05]*, β_1 = *[– 0.46, – 0.25]*, which are narrower than those calculated in [*Marino-Palumbo, 2003*].

CONCLUSIONS

Interval data analysis has been treated in the literature by many authors [*Billard-Diday, 2000*], [*Billard-Diday, 2002*], [*Bock-Diday, 2000*], [*Canal-Marques Pereira, 1998*], [*Gioia, 2001*], [*Lauro-Palumbo, 2000*], [*Marino-Palumbo, 2003*], [*Rodriguez, 2000*].

Some proposed approaches for studying interval-valued variables produce numbers rather than intervals; in other works some statistics for interval-valued variables are obtained ex-post using interval bounds or centre and radius of each interval. In general, classical statistical methods applied to the characteristic elements of the considered intervals do not produce satisfactory results as the used criteria do not consider intervals as a whole structure or special kind of data. Aiming at reconstructing interval solution only ex post.é

Statistical indexes for location and variability for interval-valued variables have been defined in this work. A new method for fitting a*simple regression* equation for linear dependent interval-valued variables has been also proposed.

Interval algebra in many situations has presented some drawback, in particular some basic formula when directly computed with interval operations may produce oversized interval solutions. This drawback has been reduced solving some *optimisation problems* by means of numerical algorithms implemented in Matlab.

The presented numerical results show that the methods adopted in this paper

features the input data fairly well; furthermore some computed interval indexes are narrower with respect to some already existing in literature.

The methods for interval-valued variables, which have been proposed for treating errors in the data, may be as well applied to different kind of data that in the real life are of interval type. For example:

- Financial data; e. g., (opening value and closing value in a session, or maxima an minima observed in a specified period).
- Customer satisfaction data (expected or perceived characteristic of the quality of a product).
- Tolerance limits in quality control.
- Confidence intervals of estimates from sample surveys.
- Query on a database.

With the prospective of analysing interval data in many different fields it could be interesting to extend the proposed statistical methods for univariate and bivariate interval-valued variables also to the case of multivariate interval-valued variables. Naturally it will not be a simple extension of classical statistical methods. Some solutions based on the imprecise data theory have *[Lauro-Palumbo 2000]* or interval algebra"*[Gioia, 2000]*. It must be noticed that these approaches require very restrictive conditions. As an alternative the optimal interval of solutions criterion here introduced for computing different interval statistical indexes (i.e. interval variance, interval regression coefficients and correlations) could offer a valid answer in these area. New conceptual approaches must probably be introduced for treating ill-defined problems; for example an open problem is the *ortogonality* of the interval eigenvectors in the *Interval Principal Component Analysis*

5. APPENDIX

Interval Algebra

Extensions of number systems involving ordered pairs of numbers from the given system are commonplace. The rational numbers are essentially ordered pairs of integers; complex numbers are ordered pairs of real numbers; in each case arithmetic operations are defined with rules for computing the components of a pair resulting from an arithmetic operation on a *pair* of pairs.

An interval number *[a,b]* with $a \leq b$, is defined as the set of real numbers between *a* and *b*:

$$\left[a,b\right] = \left\{x \,/\, a \leq x \leq b\right\}$$

Degenerate intervals of the form *[a,a]*, also named *thin* intervals, are equivalent to real numbers.

The symbols $\in, \subset, \cup, \cap,$ will be used in the common sense of set theory. For example by $[a,b] \subset [c,d]$ we mean that interval *[a,b]* is included as a *set* in the interval *[c,d]*.

Furthermore it is $[a,b]=[c,d] \Leftrightarrow a=b\ ,\ c=d$.

Let $\Im$ be the set of interval numbers. Thus $I \in \Im$ then $I=[a,b]$ for some $a \leq b$. Let us introduce an arithmetic on the elements of $\Im$. The arithmetic will be an extension of real arithmetic.

If • is one of the symbols +, -, , /, we define arithmetic operations on intervals by:

$$\left[a,b\right] \bullet \left[c,d\right] = \left\{x \bullet y \mid \quad a \leq x \leq b, c \leq y \leq d\right\} \tag{5.1}$$

except that we do not define *[a,b]/[c,d]* if $0 \in [c,d]$.

We can say that the sum, the difference, the product, and the ratio (when defined) between two intervals is the set of the sums, the differences, the products, and the ratios between any two numbers from the first and the second interval respectively.

Let us write an equivalent set of definitions in terms of formulas for the endpoints of resultant intervals.

Let *[a,b] , [c,d]* be elements of $\Im$, it is:

$$\begin{gathered} [a,b] + [c,d] = [a + c, b + d] \\ [a,b] - [c,d] = [a - d, b - c] \\ [a,b] \cdot [c,d] = [\min(ac,ad,bc,bd), \max(ac,ad,bc,bd)] \end{gathered} \tag{5.2}$$

if $0 \notin [c,d]$, then:

$$[a,b] \ / \ [c,d] = [a,b] \cdot [1 \ / \ d, 1 \ / \ c]$$

It can be easily proved that the addition and the product (5.2) are associative and commutative, i.e. taken $I,J,K \in \Im$ the following relations hold:

$$I+(J+K)=(I+J)+K$$

$$I \cdot (J \cdot K)=(I \cdot J)\ K$$

$$I+J=J+I$$

$$I \cdot J=J \cdot I$$

Real numbers *0* and *1* can be both regarded as units for addition and for

product respectively. In other words if $I \in \mathfrak{I}$ then:

$$0+I=I+0=I$$

$$1 \cdot I=I \cdot 1=I$$

When an interval number is raised to a power the algebraic sign of the endpoints involved determine the formulas to be applied. Thus:

$$\begin{aligned}
&\left[a,b\right]^n=\left[a^n,b^n\right] \quad if \quad a>0 \quad or \quad n \quad is \quad odd \\
&\left[a,b\right]^n=\left[b^n,a^n\right] \quad if \quad b<0 \quad and \quad n \quad is \quad even \\
&\left[0,\left|\left[a,b\right]\right|^n\right] \quad if \quad 0 \in \left[a,b\right] \quad and \quad n \quad is \quad even
\end{aligned}$$

Other properties may be found in [*Moore 1966*], [*Kerarfott-Kreinovich, 1996*].

The min/max optimisation algorithm for interval-data analysis

The algorithms, for computing the statistical index just introduced, have been implemented in MATLAB.

In particular the *minimization/maximization* problem, of a multiple variable function, is solved here numerically using the optimisation function *fmincon*.

The function *fmincon* finds the minimum(maximum) of a constrained nonlinear multivariable function.

$$\min_x f(x) \qquad \text{subiect t} \qquad \begin{aligned} c(x) &\le 0 \\ ceq(x) &= 0 \\ A \cdot x &\le b \\ Aeq \cdot x &\le beq \\ lb \le x &\le ub \end{aligned}$$

where *x*, *b*, *beq*, *lb*, and *ub* are vectors, *A* and *Aeq* are matrices, *c(x)* and *ceq(x)* are functions that return vectors, and *f(x)* is a function that returns a scalar. *f(x)*, *c(x)*, and *ceq(x)* can be nonlinear functions.

It is important to notice that the solution of all proposed optimization problems always exists because all object functions respect the pertinent hypothesis.

ACKNOWLEDGMENTS

We should like to express appreciation to Prof. Aversa for his critical appraisals and for his encouraging interest throughout this research project.

BIBLIOGRAPHY

ALEFELD G., HERZBERGER J., *Introduction to Interval Computations*, Computer Science and Applied Mathematics, 1983.

BILLARD L., DIDAY E., *Regression Analysis for Interval-Valued Data*. In: Data Analysis,Classification and Related Methods (eds.H.-H. Bock and E. Diday), Springer,103-124, (2000).

BILLARD L., DIDAY E., *Symbolic regression Analysis*, Proceedings IFCS. In KRZYSZTOF JAJUGA et al (Eds.) (2002), Data Analysis, Classification and Clustering Methods Heidelberg, Springer-Verlag.

BOCK H.-H.- DIDAY E. (2000), *Analysis of Symbolic Data*, Springer.

BURKILL J. C., *Functions of Intervals*, Proceedings of the London Mathematical Society, 22:375-446.

CANAL L., MARQUES PEREIRA R.A., *Towards statistical indices for numeroid data*, Pre-Proceedings NTTS, Sorrento, Italy, 4/6 November 1998, p.97.

GIOIA F. (2001), *Metodi Statistici per Variabili di Intervallo*, P.h.D. Thesis.

KERARFOTT R. B. KREINOVICH V. (Eds.) (1996), *Applications Of Interval Computations*, Kluwer Academic Publishers.

LAURO C.N, PALUMBO F. (2000), *Principal component analysis of interval data: A symbolic data analysis approach*, Computational Statistics 15, 1, 73-87.

MARINO M., PALUMBO F. (2003), *Interval Arithmetic for the evaluation of imprecise data effects in least squares linear regression*, Statistica Applicata, 3.

MOORE R.E. (1966), *Interval Analysis*, Prentice-Hall, Series in Automatic Computation,

NEUMAIER A. (1990), *Interval methods for Systems of Equations*, Cambridge University Press.

PICCOLO D. (1998), *Statistica*, Il Mulino

RODRIGUEZ O. (2000), *Classification et Modeles Lineaires en Analyse des Donnes Symboliques.* Doctoral Thesis, Universite de Paris Dauphine IX.

SUNAGA T. (1958), *Theory of an Interval Algebra and its Application to Numerical Analysis*, Gaukutsu Bunken Fukeyu-kai, Tokyo.

YOUNG R.C. (1931), *The algebra of many-valued quanties*, Math. Ann. 104:260-290.

Inserire il Titolo e riassunto in italiano