Available online at www.sciencedirect.com

ScienceDirect

Information Sciences 177 (2007) 3228–3238

INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL

www.elsevier.com/locate/ins

# On-line algorithms for computing mean and variance of interval data, and their use in intelligent systems

Vladik Kreinovich [a,*], Hung T. Nguyen [b], Berlin Wu [c]

[a] *Computer Science Department, University of Texas, El Paso, TX 79968, USA*
[b] *Department of Mathematical Sciences, New Mexico State University, Las Cruces, NM 88003, USA*
[c] *Department of Mathematical Sciences, National Chengchi University, Taipei, Taiwan*

Received 10 November 2006; accepted 11 November 2006

**Abstract**

When we have only interval ranges $[\underline{x_i}, \overline{x_i}]$ of sample values $x_1, \ldots, x_n$, what is the interval $[\underline{V}, \overline{V}]$ of possible values for the variance $V$ of these values? There are quadratic time algorithms for computing the exact lower bound $\underline{V}$ on the variance of interval data, and for computing $\overline{V}$ under reasonable easily verifiable conditions. The problem is that in real life, we often make additional measurements. In traditional statistics, if we have a new measurement result, we can modify the value of variance in constant time. In contrast, previously known algorithms for processing interval data required that, once a new data point is added, we start from the very beginning. In this paper, we describe new algorithms for statistical processing of interval data, algorithms in which adding a data point requires only O($n$) computational steps.

*Keywords:* Interval data; Mean; Variance; On-line data processing

## 1. Introduction: data processing in intelligent systems – from probabilities to intervals

### *1.1. Let us start with a big picture*

Before we describe a specific problem that we solve in this paper, let us first describe how, in our view, this problem fits into a big picture of information processing in intelligent systems. Readers who are familiar with this big picture and/or who are only interested in our technical results can skip this subsection.

One of the main specific features of information processing in intelligent systems is that in such systems, we often have very limited knowledge. As a result, processing of *imprecise* information is necessary in intelligent systems.

A typical example is the processing of linguistic information, i.e., information represented by experts in terms of words from a natural language. This information can be modeled, e.g., by fuzzy sets (see, e.g.,

[*] Corresponding author. Tel.: +1 915 747 6951; fax: +1 915 747 5030.
*E-mail addresses:* vladik@cs.utep.edu (V. Kreinovich), hunguyen@nmsu.edu (H.T. Nguyen), berlin@math.nccu.edu.tw (B. Wu).

doi:10.1016/j.ins.2006.11.007

16,25). For such modeling, when an expert states that a value is, say, small but not very small, we describe this expert information in terms of an appropriate fuzzy set.

A particular case of such a statement is when an expert states that the actual value is between, say, 0.1 and 0.3. After such a statement, the only information about the actual (unknown) value of the desired quantity is that it belongs to the *interval* $[0.1, 0.3]$ – and each interval (and, more generally, each set) can be viewed as a particular example of a more general concept of a fuzzy set.

Since the knowledge about each quantity is represented in such a form, it is necessary to be able to develop *inference procedures* for such observations. Mathematical analysis of this problem is therefore crucial for designing intelligent systems. In this paper, we analyze an important particular case of this set-valued data. Specifically, in this paper, we investigate the computational aspects of processing interval-valued data. Let us now describe our problem and its motivation in more detail.

### 1.2. Why data processing?

In intelligent systems, there are at least two sources of information about physical quantities: measurements and expert estimates.

In many real-life situations, we are interested in the value of a physical quantity $y$ that is difficult or impossible to measure directly and difficult for experts to estimate. Examples of such quantities are the distance to a star and the amount of oil in a given well.

Since we cannot measure or estimate the value $y$ of the desired physical quantity directly, a natural idea is to measure or estimate $y$ *indirectly*. Specifically, we find some easier-to-measure or easier-to-estimate quantities $x_1, \ldots, x_n$ which are related to $y$ by a known relation $y = f(x_1, \ldots, x_n)$. For example, to find the resistance $R$, we measure or estimate current $I$ and voltage $V$, and then use the known relation $R = V/I$ to estimate resistance as $\widetilde{R} = \widetilde{V}/\widetilde{I}$. This relation may be a simple functional transformation, or a complex algorithm (e.g., for the amount of oil, a numerical solution to an inverse problem). It is worth mentioning that in the vast majority of these cases, the function $f(x_1, \ldots, x_n)$ that describes the dependence between physical quantities is continuous. In such cases, to estimate $y$, we first measure or estimate the values of the quantities $x_1, \ldots, x_n$, and then we use the results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of these measurements or estimates to compute an estimate $\tilde{y}$ for $y$ as $\tilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

*Comment.* In this paper, for simplicity, we consider the case when the relation between $x_i$ and $y$ is known exactly; in practical situations, we often only know an approximate relation between $x_i$ and $y$.

### 1.3. Why interval computations? From probabilities to intervals

Neither measurements nor estimates are 100% accurate, so in reality, the actual value $x_i$ of quantity $i$ can differ from the result $\widetilde{x}_i$ obtained by measurement or by estimation. Because of these *measurement* (*estimation*) *errors* $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$, the result $\tilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of data processing is, in general, different from the actual value $y = f(x_1, \ldots, x_n)$ of the desired quantity $y$ [29]. It is desirable to describe the error $\Delta y \stackrel{\text{def}}{=} \tilde{y} - y$ of the result of data processing. To do that, we must have some information about the errors of direct measurements and/or estimates.

What do we know about the errors $\Delta x_i$ related to expert estimation? Often, an expert can provide *bounds* $\underline{x}_i$ and $\overline{x_i}$ for the estimated quantity $x_i$. Then, the actual (unknown) value of $x_i$ belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x_i}]$. Often, these bounds come in the form of an unsigned error estimate $\Delta_i$ on the expert's estimation accuracy: for example, an expert may say that the actual fish population in a lake is 50,000 ± 20,000. In this case, $\widetilde{x}_i = 50{,}000$, $\Delta_i = 20{,}000$, so $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x_i} = \widetilde{x}_i + \Delta_i$.

*Comment.* For readers who may be interested in how the above description is related to fuzzy sets, here is an explanation. Often, in addition to (or instead of) the bounds, an expert can provide bounds that contain $x_i$ with a certain degree of confidence (not necessarily represented by a probability). Often, we know several such bounding intervals corresponding to different degrees of confidence. Such a nested family of intervals is also called a *fuzzy set*, because it turns out to be equivalent to a more traditional definition of fuzzy set [6,16,23–25]