6-2011

Estimating Mean and Variance under Interval Uncertainty: Dynamic Case

Rafik Aliev
Azerbaijan State Oil Academy

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-11-29
Published in *Proceedings of the 2011 Sixth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control ICSCCW'2011*, Antalya, Turkey, September 1-2, 2011, pp. 85-93.

Estimating Mean and Variance under Interval Uncertainty: Dynamic Case

Rafik Aliev[1] and Vladik Kreinovich[2]
[1]Dept. of Computer Aided Control Systems
Azerbaijan State Oil Academy
Azadlig Ave. 20, AZ1010 Baki, Azerbaijan
raliev@asoa.edu.az
[2]Department of Computer Science, University of Texas at El Paso
El Paso, TX 79968, USA, vladik@utep.edu

Abstract

In many practical situations, it is important to estimate the mean E and the variance V from the sample values $x_1, \ldots, x_n$. Usually, in statistics, we consider the case when the parameters like E and V do not change with time and when the sample values x_i are known exactly. In practice, the values x_i come from measurements, and measurements are never 100% accurate. In many cases, we only know the upper bound Δ_i on the measurement error. In this case, once we know the measured value $\widetilde{x}_i$, we can conclude that the actual (unknown) value x_i belongs to the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. Different values x_i from these intervals lead, in general, to different values of E and V. It is therefore desirable to find the ranges $\mathbf{E}$ and $\mathbf{V}$ of all possible values of E and V. While this problem is, in general, NP-hard, in many practical situations, there exist efficient algorithms for computing such ranges.

In practice, processes are dynamic. As a result, reasonable estimates for E and V assign more weight to more recent measurements and less weight to the past ones. In this paper, we extend known algorithms for computing the ranges $\mathbf{E}$ and $\mathbf{V}$ to such dynamic estimates.

1 Introduction

Need for statistical estimates. In many practical situations, it is important to estimate statistical characteristics such as the mean E and the variance V from the sample values $x_1, \ldots, x_n$. There exist many methods for such estimation; see, e.g., [14].

Normal distribution and the standard estimates for E and V. Standard methods for estimating E and V are based on the assumption that the corresponding random quantity is normally distributed, with a probability density

$$\rho(x) = \frac{1}{\sqrt{2\pi \cdot V}} \cdot \exp\left(-\frac{(x-E)^2}{2V}\right).$$

This assumption is often empirically valid. The explanation for a frequent occurrence of normal distribution comes from the Central Limit Theorem, according to which if the random variable consists of several small independent components, then its distribution is close to normal – and it is often the case that the desired value x_i is influenced by a large number of different independent factors; see, e.g., [14].

It is usually assumed that different sample values are independent. In this case, for each pair of values E and V, the probability L that the observed sample $x_1, \ldots, x_n$ occurs for these particular values of E and V can be found as simply the product of the corresponding probabilities:

$$L = \prod_{i=1}^{n} \rho(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot V}} \cdot \exp\left(-\frac{(x_i - E)^2}{2V}\right).$$

It is reasonable to select the *most probable* values E and V, i.e., the values for which the above probability is the largest. This idea is known as the *Maximum Likelihood (ML) approach*.

We can find the corresponding maximum if we differentiate the expression L with respect to E and V and equate derivatives to 0. As a result, we get the following estimates:

$$E = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i; \quad V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2.$$

These are the estimates that are most frequently used to estimate the mean and variance.

Comment. Sometimes, statisticians use instead an un-biased estimate for the variance, i.e., an estimate for which the expected value is exactly the desired variance. This un-biased estimate $\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - E)^2$ differs from the ML estimate by a constant factor in front of the sum. Thus, from the computational viewpoint, we can easily reduce the computation of the un-biased estimate to the computation of the ML estimate. Namely, to compute the un-biased estimate, we can simply compute the ML estimate and multiply the result by $\frac{n}{n-1}$. Because of this reduction, in the following text, we will mainly talk about computing the ML estimate.

General case when the distributions are not necessarily Gaussian. While these estimates are justified as optimal only for the normal distributions, they are used for other distributions as well. Their application to arbitrary distributions is justified by the fact that the mean can be equivalently defined as the limit of the arithmetic averages when the sample size n grows to infinity – similarly to how the probability can be defined as limit of the frequency when the sample size increases $n \to \infty$.

The variance is, by definition, the expected value of the square of the difference $(x - E)^2$: $V = E[(x - E)^2]$. It can be equivalently described as the difference $E[x^2] - (E[x])^2$. Similarly, the above formula can be described as $M - E^2$, where $M \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2$. The arithmetic average of x_i^2 tends to $E[x^2]$, the arithmetic average E tends to $E[x]$, so our estimate $M - E^2$ tends to the difference $E[x^2] - (E[x])^2$, i.e., to the actual variance.

The limits mean, in effect, that the estimates based on large n can serve as *good* estimates for the actual mean and variance; the larger the sample size n, the better these estimates.

(Although, for non-Gaussian distributions, these estimates are *not* necessarily *optimal* ones.)

Need to take interval uncertainty into account. Usually, in statistics, we consider the case when the sample values x_i are known exactly. In practice, the values x_i come from measurements, and measurements are never 100% accurate: the values $\widetilde{x}_i$ resulting from the measurement are, in general, different from the actual (unknown) values x_i of the corresponding quantities, and the corresponding measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ are non-zero.

Sometimes, we know the probabilities of different values of measurement errors. However, in many cases, we only know the upper bound Δ_i on the (absolute value of the) measurement error: $|\Delta x_i| \le \Delta_i$ [13]. In this case, once we know the measured value $\widetilde{x}_i$, we can conclude that the actual (unknown) value x_i belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i] \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

Different values x_i from these intervals lead, in general, to different estimates of $E(x_1, \ldots, x_n)$ and $V(x_1, \ldots, x_n)$. It is therefore desirable to find the ranges

$$\mathbf{E} = [\underline{E}, \overline{E}] = \{E(x_1, \ldots, x_n) | x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\} \text{ and}$$

$$\mathbf{V} = [\underline{V}, \overline{V}] = \{V(x_1, \ldots, x_n) | x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}$$

of all possible values of E and V.

Case of interval uncertainty: what is known. The general problem of estimating the range of a function under interval uncertainty is known as the *main problem of interval computations*; see, e.g., [7, 11].

The situation is the simplest with the mean $E(x_1, \ldots, x_n) = \dfrac{1}{n} \cdot \sum_{i=1}^{n} x_i$: since the mean is an increasing function of each of its variables $x_1, \ldots, x_n$, its smallest possible value $\underline{E}$ is attained when we take the smallest possible values $x_i = \underline{x}_i$ of all the inputs, and its largest possible value $\overline{E}$ is attained when we take the largest possible values $x_i = \overline{x}_i$ of all the inputs. Thus, the desired range has the form $[\underline{E}, \overline{E}] = \left[\dfrac{1}{n} \cdot \sum_{i=1}^{n} \underline{x}_i, \dfrac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i\right]$.

In contrast, the variance $V(x_1, \ldots, x_n)$ is not always monotonic, so for the variance, estimating the range is a more complex task. It is known that in general, the problem of computing this range is NP-hard [3, 4]. Specifically, the lower endpoint $\underline{V}$ can be computed in feasible time [3, 4, 15], but computing $\overline{V}$ is NP-hard. For some practically useful situations, there exist efficient algorithms for computing $\overline{V}$; see, e.g., [2, 5, 9, 8, 15].

Need to consider dynamic estimates. Usually, in statistics, we consider the case when the parameters like E and V do not change with time. In practice, processes are dynamic.

As a result, reasonable estimates for E and V should assign more weight to more recent measurements and less weight to the past ones. Specifically, if we sort the values x_i from the most recent one x_1 to the least recent one x_n, then, for each function $y(x)$, to estimate the mean value of y, instead of the arithmetic mean, we take the weighted mean

$$E[y] \approx \sum_{i=1}^{n} w_i \cdot x_i,$$

where $w_1 \geq w_2 \geq \ldots \geq w_n > 0$, and $\sum\limits_{i=1}^{n} w_i = 1$. In particular, for the mean E, we have the estimate

$$E = \sum_{i=1}^{n} w_i \cdot x_i.$$

Similarly, as an estimate for the actual variance $E[x^2] - (E[x])^2$, we take

$$V = \sum_{i=1}^{n} w_i \cdot x_i^2 - \left(\sum_{i=1}^{n} w_i \cdot x_i \right)^2.$$

One can show that this expression is equivalent to $V = \sum\limits_{i=1}^{n} w_i \cdot (x_i - E)^2$.

Comment. If $w_i = 0$, this simply means that we do not take into account the corresponding value x_i. Thus, if we restrict ourselves only to the inputs on which the characteristics actually depend, we conclude that $w_i > 0$.

What we do in this paper. In this paper, we extend known algorithms for computing the ranges **E** and **V** to such dynamic estimates.

2 Simplest Case: Estimates for the Mean

Let us first consider the simplest case: estimates for the mean. Since all the weights are non-negative, the function $E = \sum\limits_{i=1}^{n} w_i \cdot x_i$ is an increasing function of all its variables. Thus:

- the smallest possible value $\underline{E}$ is attained when we take the smallest possible values $x_i = \underline{x}_i$ of all the inputs, and
- the largest possible value $\overline{E}$ is attained when we take the largest possible values $x_i = \overline{x}_i$ of all the inputs.

Thus, the desired range of E has the form $[\underline{E}, \overline{E}] = \left[\sum\limits_{i=1}^{n} w_i \cdot \underline{x}_i, \sum\limits_{i=1}^{n} w_i \cdot \overline{x}_i \right]$.

3 Estimates for the Variance: Analysis of the Problem

When a function attains minimum and maximum on the interval: known facts from calculus. In this computation, we will use known facts from calculus.

A function $f(x)$ defined on an interval $[\underline{x}, \overline{x}]$ attains its minimum on this interval either at lone of its endpoints, or in some internal point of the interval. If it attains is minimum at a point $x \in (a, b)$, then its derivative at this point is 0: $\dfrac{df}{dx} = 0$.

If it attains its minimum at the point $x = \underline{x}$, then we cannot have $\dfrac{df}{dx} < 0$, because then, for some point $x + \Delta x \in [\underline{x}, \overline{x}]$, we would have a smaller value of $f(x)$. Thus, in this case, we must have $\dfrac{df}{dx} \geq 0$.

Similarly, if a function $f(x)$ attains its minimum at the point $x = \overline{x}$, then we must have $\dfrac{df}{dx} \le 0$.

For the maximum, a similar thing happens. If $f(x)$ attains is maximum at a point $x \in (a, b)$, then its derivative at this point is 0: $\dfrac{df}{dx} = 0$. If it attains its maximum at the point $x = \underline{x}$, then we must have $\dfrac{df}{dx} \le 0$. Finally, if a function $f(x)$ attains its maximum at the point $x = \overline{x}$, then we must have $\dfrac{df}{dx} \ge 0$.

Let us apply these known facts to our problem. We are interested in range of the expression $V = \sum\limits_{i=1}^{n} w_i \cdot x_i^2 - E^2$, where $E \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} w_i \cdot x_i$. For this estimate, $\dfrac{\partial E}{\partial x_i} = w_i$, hence

$$\frac{\partial V}{\partial x_i} = 2w_i \cdot x_i - 2E \cdot \frac{\partial E}{\partial x_i} = 2w_i \cdot (x_i - E).$$

To find this range, we must find the point where this expression attains its minimum, and the point where it attains its maximum.

Where is minimum attained: analysis. By considering the variance V as a function of x_i, for the point $(x_1, \ldots, x_n, y_1)$ at which V attains its minimum, we can make the following conclusions:

- if $x_i = \underline{x}_i$, then $x_i \ge E$;
- if $x_i = \overline{x}_i$, then $x_i \le E$;
- if $\underline{x}_i < x_i < \overline{x}_i$, then $x_i = E$.

So, if $\overline{x}_i < E$, this means that for the value $x_i \le \overline{x}_i$ also satisfies the inequality $x_i < E$. Thus, in this case:

- we cannot have $x_i = \underline{x}_i$ — because then we would have $x_i \ge E$; and
- we cannot have $\underline{x}_i < x_i < \overline{x}_i$ – because then, we would have $x_i = E$.

So, if $\overline{x}_i < E$, the only remaining option for x_i is $x_i = \overline{x}_i$.

Similarly, if $E < \underline{x}_i$, this means that the value $x_i \ge \underline{x}_i$ also satisfies the inequality $x_i > E$. Thus, in this case:

- we cannot have $x_i = \overline{x}_i$ — because then we would have $x_i \le E$; and
- we cannot have $\underline{x}_i < x_i < \overline{x}_i$ – because then, we would have $x_i = E$.

So, if $E < \underline{x}_i$, the only remaining option for x_i is $x_i = \underline{x}_i$.

What if $\underline{x}_i < E < \overline{x}_i$? In this case:

- the minimum cannot be attained for $x_i = \underline{x}_i$, because then we should have $x_i \ge E$, while we have $x_i < E$;

- the minimum cannot be attained for $x_i = \overline{x}_i$, because then we should have $x_i \le E$, while we have $x_i > E$.

Thus, the minimum has to be attained when $x_i \in (\underline{x}_i, \overline{x}_i)$. In this case, we have $x_i = E$.

Where is minimum attained: conclusion.

- If $\overline{x}_i \le E$, i.e., if the interval $\mathbf{x}_i$ is fully to the left of the mean E, the minimum is attained for $x_i = \overline{x}_i$.
- If $E \le \underline{x}_i$, i.e., if the interval $\mathbf{x}_i$ is fully to the right of the mean E, the minimum is attained for $x_i = \underline{x}_i$.
- If $\underline{x}_i < E < \overline{x}_i$, i.e., if the interval $\mathbf{x}_i$ contain the mean E, the minimum is attained for $x_i = E$.

In all three cases, once we know where the minimum is attained in relation to the endpoints $\underline{x}_i$ and $\overline{x}_i$, we can find out, for each i, where the minimum is attained – actually at the point which is the closest to E.

This conclusion is in good accordance with common sense: the variance is the smallest when all the values are the closest to the mean.

The value E must be found from the condition that it is the weighted mean of all corresponding minimal values, i.e., that

$$\sum_{i:\overline{x}_i \le E} w_i \cdot \overline{x}_i + \sum_{j:E \le \underline{x}_j} w_j \cdot \underline{x}_j + \sum_{k:\underline{x}_i < E < \overline{x}_i} w_k \cdot E = E.$$

By moving all the terms proportional to E to the right-hand side and dividing by the coefficient at E, we conclude that

$$E = \frac{\sum\limits_{i:\overline{x}_i \le E} w_i \cdot \overline{x}_i + \sum\limits_{j:E \le \underline{x}_j} w_j \cdot \underline{x}_j}{\sum\limits_{i:\overline{x}_i \le E} w_i + \sum\limits_{j:E \le \underline{x}_j} w_j}.$$

This conclusion will be used to design an efficient algorithm for computing $\underline{V}$.

Where is the maximum attained: analysis. The function $V(x_1, \ldots, x_n)$ is convex. Thus, its maximum is always attained at one of the endpoints of each intervals $[\underline{x}_i, \overline{x}_i]$. From our calculus-based analysis, we can now come up with the following conclusions:

- if the maximum is attained for $x_i = \underline{x}_i$, then we should have $x_i \le E$, i.e., $\underline{x}_i \le E$;
- if the maximum is attained for $x_i = \overline{x}_i$, then we should have $x_i \ge E$, i.e., $E \le \overline{x}_i$.

Thus, if $\overline{x}_i < E$, we cannot have $x_i = \overline{x}_i$, so the maximum is attained for $x_i = \underline{x}_i$. Similarly, if $E < \underline{x}_i$, then we cannot have $x_i = \underline{x}_i$, so the maximum is attained for $x_i = \overline{x}_i$. If $\underline{x}_i \le E \le \overline{x}_i$, then we can have both options $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$.

Where is maximum attained: conclusion.

- If $\overline{x}_i \le E$, i.e., if the interval $\mathbf{x}_i$ is fully to the left of the mean E, the maximum is attained for $x_i = \underline{x}_i$.
- If $E \le \underline{x}_i$, i.e., if the interval $\mathbf{x}_i$ is fully to the right of the mean E, the maximum is attained for $x_i = \overline{x}_i$.
- If $\underline{x}_i < E < \overline{x}_i$, i.e., if the interval $\mathbf{x}_i$ contain the mean E, the maximum can be attained at both values $\underline{x}_i$ and $\overline{x}_i$.

In all three cases, once we know where the maximum is attained in relation to the endpoints $\underline{x}_i$ and $\overline{x}_i$, we can find out, for each i, where the minimum is attained – actually at the point which is the farthest away from E. This conclusion is also in good accordance with common sense: the variance is the largest when all the values are the farthest way from the mean.

Now, we are ready to describe the corresponding algorithms.

4 Efficient Algorithm for Computing $\underline{V}$

Description of the algorithm. First, we sort all $2n$ endpoints $\underline{x}_i$ and $\overline{x}_i$ of the given intervals into a non-decreasing sequence $r_1 \le r_2 \le \ldots \le r_{2n-1} \le r_{2n}$. To cover the whole straight line, we add the points $r_0 = -\infty$ and $r_{2n+1} = +\infty$. As a result, the whole real line is divided into $2n+1$ zones $[r_k, r_{k+1}]$, with $k = 0, 1, \ldots, 2n$.

For each zone, we find the values x_i which minimize V under the condition that their weighted average E is contained in this zone. Namely, we compute $E_k = \dfrac{N_k}{D_k}$, where

$$N_k \stackrel{\text{def}}{=} \sum_{i:\overline{x}_i \le r_k} w_i \cdot \overline{x}_i + \sum_{j:r_{k+1} \le \underline{x}_j} w_j \cdot \underline{x}_j; \quad D_k = \sum_{i:\overline{x}_i \le r_k} w_i + \sum_{j:r_{k+1} \le \underline{x}_j} w_j.$$

If E_k is not within the zone $[r_k, r_{k+1}]$, we dismiss it and move to the next zone. If it is within the zone, we compute the corresponding value of the variance

$$V_k = \sum_{i:\overline{x}_i \le r_k} w_i \cdot (\overline{x}_i - E_k)^2 + \sum_{j:r_{k+1} \le \underline{x}_j} w_j \cdot (\underline{x}_j - E_k)^2.$$

This expression can be equivalently reformulated as $V_k = M_k - W_k \cdot E_k^2$, where we denoted

$$M_k = \sum_{i:\overline{x}_i \le r_k} w_i \cdot (\overline{x}_i)^2 + \sum_{j:r_{k+1} \le \underline{x}_j} w_j \cdot (\underline{x}_j)^2; \quad W_k = \sum_{i:\overline{x}_i \le r_k} w_i + \sum_{j:r_{k+1} \le \underline{x}_j} w_j.$$

Once the computations are performed for all $2n+1$ zones, we find the smallest of the corresponding values V_k as the desired smallest value $\underline{V}$.

Computation time of this algorithm. Sorting takes time $O(n \log \log(n))$; see, e.g., [1]. Computing the sums D_0, N_0, M_0, and W_0 corresponding to the first zone take linear time $O(n)$. Each new sum is obtained from the previous one by changing a few terms which go

from $\underline{x}_i$ to $\overline{x}_i$; each value x_i changes only once, so we only need totally linear time to compute all these sums – and we also need linear time to perform all the auxiliary computations. Thus, the total computation time is $O(n \cdot \log(n)) + O(n) + O(n) = O(n \cdot \log(n))$.

It is possible to provide a linear-time algorithm. This computation time can be reduced to $O(n)$ of we use the ideas from [15], where instead of sorting, we used the known linear time algorithm for computing the median.

5 Efficient Algorithm for Computing $\overline{V}$ under a Reasonable Condition

Description of the condition. We assume that for some integer C, each set of more than C intervals has an empty intersection. For example, for $C = 1$, no two intervals have a common point. For $C = 2$, two intervals may have a common point, but no three intervals share a common point, etc.

Resulting algorithm. As with computing $\underline{V}$, we start by sorting all $2n$ endpoints $\underline{x}_i$ and $\overline{x}_i$ of the given intervals into a non-decreasing sequence $r_1 \le r_2 \le \ldots \le r_{2n-1} \le r_{2n}$. To cover the whole straight line, we then add the points $r_0 = -\infty$ and $r_{2n+1} = +\infty$. As a result, the whole real line is divided into $2n + 1$ zones $[r_k, r_{k+1}]$, with $k = 0, 1, \ldots, 2n$.

For each zone, we find the values x_i which maximize V under the condition that their weighted average E is contained in this zone:

- for those i for which $\overline{x}_i \le r_k$, we take $x_i = \underline{x}_i$;
- for those i for which $r_{k+1} \le \underline{x}_i$, we take $x_i = \overline{x}_i$;
- for all other indices i, for which $[r_k, r_{k+1}] \subseteq \mathbf{x}_i$, we consider both possibilities $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$.

Because of our condition, for each zone, there are no more than C indices i in the third category. Thus, for each zone, we have to consider $\le 2^C$ possible combinations of values $\underline{x}_i$ and $\overline{x}_i$. For each of these combinations, we compute the weighted average E and, if this weighter average is within the zone $[r_k, r_{k+1}]$, we compute the weighted variance V – e.g., $V = M - E^2$, where $M = \sum\limits_{i=1}^{n} w_i \cdot x_i^2$ is the weighted average of the squared values x_i^2.

The largest of all such computed values V is then returned as $\overline{V}$.

Computation time of this algorithm. Sorting takes time $O(n \cdot \log(n))$. Computing the original values of E and M requires linear time. Similarly to the case of computing $\underline{V}$, each new sum is obtained from the previous one by changing a few terms which go from $\underline{x}_i$ to $\overline{x}_i$; each value x_i changes only once, so we only need totally linear time to compute all these sums – and we also need linear time to perform all the auxiliary computations. Thus, the total computation time is also $O(n \cdot \log(n)) + O(n) + O(n) = O(n \cdot \log(n))$.

Comment. A similar modification of an algorithm presented in [6] can lead to a polynomial-time algorithm for computing the range of the weighted covariance

$$C = \sum_{i=1}^{n} w_i \cdot (x_i - E_x) \cdot (y_i - E_y) = \sum_{i=1}^{n} w_i \cdot x_i \cdot y_i,$$

where

$$E_x \stackrel{\text{def}}{=} \sum_{i=1}^{n} w_i \cdot x_i \text{ and } E_y \stackrel{\text{def}}{=} \sum_{i=1}^{n} w_i \cdot y_i,$$

under the condition that all x-intervals $\mathbf{x}_i$ are of the form $[t_0^{(x)}, t_1^{(x)}], [t_1^{(x)}, t_2^{(x)}], \ldots, [t_{N_x-1}^{(x)}, t_{N_x}^{(x)}]$ for some we have x-threshold values $t_0^{(x)} < t_1^{(x)} < \ldots < t_{N_x}^{(x)}$, and all y-intervals $\mathbf{y}_i$ are of the form $[t_0^{(y)}, t_1^{(y)}], [t_1^{(y)}, t_2^{(y)}], \ldots, [t_{N_y-1}^{(y)}, t_{N_y}^{(y)}]$ for some we have y-threshold values

$$t_0^{(y)} < t_1^{(y)} < \ldots < t_{N_y}^{(y)}.$$

This situation occurs in statistical data processing, when we use, as input, answers to threshold-related questions: e.g., whether the age is from 0 to 20, from 20 to 30, etc.

Acknowledgments. This work was supported in part by the National Science Foundation grants HRD-0734825 and DUE-0926721 and by Grant 1 T36 GM078000-01 from the National Institutes of Health.

References

[1] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Introduction to Algorithms, MIT Press, Cambridge, Massachusetts, 2009.

[2] E. Dantsin, V. Kreinovich, A. Wolpert, and G. Xiang, Population Variance under Interval Uncertainty: A New Algorithm, Reliable Computing, 12(4) (2006) 273–280.

[3] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Computing Variance for Interval Data is NP-Hard, ACM SIGACT News, 33(2) (2002) 108–118.

[4] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, Exact Bounds on Finite Populations of Interval Data, Reliable Computing, 11(3) (2005) 207–233.

[5] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg, Experimental Uncertainty Estimation and Statistics for Data Having Interval Uncertainty, Sandia National Laboratories, Report SAND2007-0939, May 2007.

[6] A. Jalal-Kamali, V. Kreinovich, and L. Longpré, Estimating covariance for privacy case under interval (and fuzzy) uncertainty, In: R. R. Yager, M. Z. Reformat, S. N. Shahbazova, and S. Ovchinnikov (eds.), Proceedings of the World Conference on Soft Computing, San Francisco, CA, May 23–26, 2011.

[7] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics, Springer-Verlag, London, 2001.

[8] V. Kreinovich, L. Longpré, S. A. Starks, G. Xiang, J. Beck, R. Kandathi, A. Nayak, S. Ferson, and J. Hajagos, Interval versions of statistical techniques, with applications to environmental analysis, bioinformatics, and privacy in statistical databases, Journal of Computational and Applied Mathematics, 199(2) (2007) 418–423.

[9] V. Kreinovich, G. Xiang, S. A. Starks, L. Longpré, M. Ceberio, R. Araiza, J. Beck, R. Kandathi, A. Nayak, R. Torres, and J. Hajagos, Towards combining probabilistic and interval uncertainty in engineering calculations: algorithms for computing statistics under interval uncertainty, and their computational complexity, Reliable Computing, 12(6) (2006) 471–501.

[10] V. Kreinovich and G. Xiang, "Fast algorithms for computing statistics under interval uncertainty: an overview", In: Van-Nam Huynh, Y. Nakamori, H. Ono, J. Lawry, V. Kreinovich, and H. T. Nguyen (eds.), Interval/Probabilistic Uncertainty and Non-Classical Logics, Springer-Verlag, Berlin-Heidelberg-New York, 2008, pp. 19–31.

[11] R. E. Moore, R. B. Kearfott, and M. J. Cloud, Introduction to Interval Analysis, SIAM Press, Philadelphia, Pennsylvania, 2009.

[12] R. Osegueda, V. Kreinovich, L. Potluri, and R. Al'o, Non-destructive testing of aerospace structures: granularity and data mining approach, Proc. FUZZ-IEEE'2002, Honolulu, Hawaii, May 12–17, 2002, vol. 1, pp. 685–689.

[13] S. Rabinovich, Measurement Errors and Uncertainties: Theory and Practice, Springer Verlag, New York, 2005.

[14] D. J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, Boca Raton, Florida, 2007.

[15] G. Xiang, M. Ceberio, and V. Kreinovich, Computing population variance and entropy under interval uncertainty: linear-time algorithms, Reliable Computing, 2007, 13(6) (2007) 467–488.