2011

A Case Study of Pre-Service Teachers Writing Mathematics for Teaching in a Second Language

Alberto Esquinca
University of Texas at El Paso, aesquinca@utep.edu

Olga Kosheleva
University of Texas at El Paso, olgak@utep.edu

Comments:
Wiest, L. R., & Lamberg, T. (Eds.). (2011). *Proceedings of the 33rd Annual Meeting of the North American Chapter of the International Group for the Psychology of Mathematics Education*. Reno, NV: University of Nevada, Reno.

International Journal of Approximate Reasoning 52 (2011) 1136–1146

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

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

Estimating sample mean under interval uncertainty and constraint on sample variance

Misha Koshelev [a,b,*], Ali Jalal-Kamali [c], Luc Longpré [c]

[a] Human Neuroimaging Lab, Division of Neuroscience, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA
[b] W. M. Keck Center for Interdisciplinary Bioscience Training, Houston, TX 77030, USA
[c] Department of Computer Science, University of Texas at El Paso, 500 W. University, El Paso, TX 79968, USA

ARTICLE INFO

Article history:
Received 14 September 2010
Revised 13 June 2011
Accepted 14 June 2011
Available online 26 June 2011

Keywords:
Statistics under interval uncertainty
Statistics under constraints
Variance
Mean

ABSTRACT

Traditionally, practitioners start a statistical analysis of a given sample $x_1, \ldots, x_n$ by computing the sample mean E and the sample variance V. The sample values x_i usually come from measurements. Measurements are never absolutely accurate and often, the only information that we have about the corresponding measurement errors are the upper bounds Δ_i on these errors. In such situations, after obtaining the measurement result $\widetilde{x}_i$, the only information that we have about the actual (unknown) value x_i of the i-th quantity is that x_i belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. Different values x_i from the corresponding intervals lead, in general, to different values of the sample mean and sample variance. It is therefore desirable to find the range of possible values of these characteristics when $x_i \in \mathbf{x}_i$.

Often, we know that the values x_i cannot differ too much from each other, i.e., we know the upper bound V_0 on the sample variance V: $V \le V_0$. It is therefore desirable to find the range of E under this constraint. This is the main problem that we solve in this paper.

1. Formulation of the problem

Traditional statistical data processing: computing sample mean and sample variance. In the traditional science and engineering practice (see, e.g., [11,12]), when we have a sample of values $x_1, \ldots, x_n$, we usually start by computing the sample mean

$$E = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \tag{1}$$

and the sample variance

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

Ubiquity of interval uncertainty. The above values E and V are easy to compute when we know the exact values of the characteristics $x_1, \ldots, x_n$. In practice, these values usually come from measurements, and measurements are never absolutely exact (see, e.g., [9,11]): the measurement results $\widetilde{x}_i$ are, in general, different from the actual (unknown) values x_i: $\widetilde{x}_i \neq x_i$.

* Corresponding author at: Human Neuroimaging Lab, Division of Neuroscience, Baylor College of Medicine, 1 Baylor Plaza, Houston, TX 77030, USA.
E-mail addresses: misha680hnl@gmail.com (M. Koshelev), ajalalkamali@miners.utep.edu (A. Jalal-Kamali), longpre@utep.edu (L. Longpré).

0888-613X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2011.06.002

Traditionally, it is assumed that we know the probability distribution of the measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. However, often, the only information we have is the upper bound Δ_i on the (absolute value of the) measurement error: $|\Delta x_i| \leq \Delta_i$.

In this case, based on the measurement result $\widetilde{x}_i$, the only information that we have about the actual (unknown) value x_i is that x_i belongs to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, where $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$.

Estimating sample mean under interval uncertainty: usual case of no additional constraints. In general, different values x_i from the corresponding intervals $\mathbf{x}_i$ lead to different values of the sample mean E. It is therefore desirable to describe the range of possible values of sample mean when x_i belong to the corresponding intervals:

$$\mathbf{E} = [\underline{E}, \overline{E}] \stackrel{\text{def}}{=} \{E(x_1, \ldots, x_n) \mid x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}. \tag{3}$$

Comment. The problem of computing the corresponding ranges is a particular case of a general problem of computing the range

$$\mathbf{y} = [\underline{y}, \overline{y}] \stackrel{\text{def}}{=} \{f(x_1, \ldots, x_n) \mid x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\} \tag{4}$$

of a given function $f(x_1, \ldots, x_n)$ when x_i are in known intervals. Computing such a range is called *interval computations*; see, e.g., [5,8].

Computing the range of the sample mean: resulting formula. When we pick any of the variables x_i and increase it to some value $x_i' > x_i$ (while leaving others intact, i.e., $x_j' = x_j$ for all $j \neq i$), the value E would increase as well. Thus, the smallest value $\underline{E}$ is attained when each of the variables x_i attains its smallest possible value $x_i = \underline{x}_i$, and its largest value $\overline{E}$ is attained when each of the variables x_i attains its largest possible value $x_i = \overline{x}_i$:

$$\underline{E} = \frac{1}{n} \cdot \sum_{i=1}^{n} \underline{x}_i; \quad \overline{E} = \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i. \tag{5}$$

Possibility of outliers. In some practical situations, the sample contains values which are drastically different from the others. Let us give an example. To estimate the average temperature in El Paso, a reasonable idea is to take the average E of the temperatures $x_1, \ldots, x_n$ measured at different locations within the city. Suppose that we have values $x_1 = 30.1$, $x_2 = 30.0$, $x_3 = 29.8$, and $x_4 = 27.0$. We know that the temperature within the city usually does not change much, the differences, on average, do not exceed 1 degree, so the value $x_4 = 27.0$ is an outlier.

In general, there may be several reasons why we get a different value. A usual reason is that the sample $x_1, \ldots, x_n$ contains *all* the measurement results, including both results with higher accuracy and results with lower accuracy. This is the main case that we consider in this paper.

In the above meteorological example, the sample contains both results made by high-accuracy thermometers placed at official meteorological sites *and* values measured by volunteers who simply place the reading of their low-accuracy thermometers on the web. For example, the values $x_1 = 30.1$, $x_2 = 30.0$, and $x_3 = 29.8$ come from the measurements with high accuracy $\Delta_1 = \Delta_2 = \Delta_3 = 0.5$, while the value $x_4 = 27.0$ comes from a thermometer with accuracy $\Delta_4 = 3.0$. Here, the corresponding intervals $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ have the form $\mathbf{x}_1 = [29.6, 30.6]$, $\mathbf{x}_2 = [29.5, 30.5]$, $\mathbf{x}_3 = [29.3, 30.3]$, and $\mathbf{x}_4 = [24.0, 30.0]$.

It is also possible that some of the measuring instruments simply malfunctioned, so the resulting value has nothing to do with the actual value of the observed quantity. In this case, we usually know the upper bound q on the number of malfunctioning sensors. This case will be analyzed in other sections.

A priori bounds on variability. As we have mentioned, one of the *a priori* constraints is that the "average difference" between different values of the sample is bounded. In the traditional statistical approach, the degree to which values vary is measured by the sample variance (2). Thus, a reasonable way to describe this constraint is to require that the sample variance cannot exceed a certain given value V_0: $V \leq V_0$.

This value V_0 may describe the upper bound on the variation of temperatures within a city, the variation of the values of biological characteristics (like height and weight) within a given species, etc.

Estimating sample mean under interval uncertainty and constraint on sample variance: a problem. In the presence of a constraint on sample variance, the problem of finding possible values of the sample mean E takes the following form:

- *given:* n intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ and a number $V_0 \geq 0$;
- *compute:* the range

$$[\underline{e}, \overline{e}] = \{E(x_1, \ldots, x_n) \mid x_i \in \mathbf{x}_i \,\&\, V(x_1, \ldots, x_n) \leq V_0\}. \tag{6}$$

This is the main problem that we will solve in this paper.

What is known: computing the range of the sample variance. Since we must take variance into account, it is reasonable to recall what is known about computing the range of the sample variance under interval uncertainty. The sample variance (2) is, in general, not monotonic; so, for the sample variance V, the problem of computing the range $[\underline{V}, \overline{V}]$ under interval uncertainty is more complex.

Specifically, it turns out that while the lower endpoint $\underline{V}$ can be computed in linear time [14], the problem of computing $\overline{V}$ is, in general, NP-hard [2,3].

A case when our problem is (relatively) easy to solve. Let us first consider the case when V_0 is larger than (or equal to) the largest possible value $\overline{V}$ of the sample variance corresponding to the given sample.

In this case, the constraint $V \le V_0$ is always satisfied. Thus, in this case, the desired range simply coincides with the range of all possible values of E, i.e., with the *arithmetic average* (5) of the corresponding intervals.

Another case when our problem is (relatively) easy to solve. Another such case is when $V_0 = 0$.

In this case, the constraint $V \le V_0$ means that the sample variance V should be equal to 0. In this case, all non-negative values $(x_i - E)^2$ should also be equal to 0 – otherwise, the average V of these values $(x_i - E)^2$ would be positive. So, we have $x_i = E$ for all i and thus, all the actual (unknown) values should coincide: $x_1 = \cdots = x_n$. In this case, we know that this common value x_i belongs to each of n intervals $\mathbf{x}_i$, so it belongs to their *intersection*

$$\mathbf{x}_1 \cap \cdots \cap \mathbf{x}_n. \tag{7}$$

A value E belongs to the interval $[\underline{x}_i, \overline{x}_i]$ if it is larger than or equal to its lower endpoint $\underline{x}_i$ and smaller than or equal to its upper endpoint $\overline{x}_i$. Thus, for a value E to belong to all n intervals, it has to be larger than or equal to all n lower endpoints $\underline{x}_1, \ldots, \underline{x}_n$, and it has to be smaller than or equal to all n upper endpoints $\overline{x}_1, \ldots, \overline{x}_n$.

A number E is larger than or equal to n given numbers $\underline{x}_1, \ldots, \underline{x}_n$ if and only if it is larger than or equal to the largest of these n numbers, i.e., if $\max(\underline{x}_1, \ldots, \underline{x}_n) \le E$. Similarly, a number E is smaller than or equal to n given numbers $\underline{x}_1, \ldots, \underline{x}_n$ if and only if it is smaller than or equal to the smallest of these n numbers, i.e., if $E \le \min(\overline{x}_1, \ldots, \overline{x}_n)$. So, the intersection consists of all the numbers which are located between these two bounds, i.e., the intersection coincides with the interval

$$[\underline{e}, \overline{e}] = [\max(\underline{x}_1, \ldots, \underline{x}_n), \min(\overline{x}_1, \ldots, \overline{x}_n)]. \tag{8}$$

Comment. In the above meteorological example, if we assume that $V_0 = 0$, then the intersection of the above intervals takes the form [29.6, 20.0]. It is easy to observe that this interval does not change if we slightly change the values $\overline{x}_1, \underline{x}_2, \overline{x}_2, \underline{x}_3, \overline{x}_3$, and $\underline{x}_4$. Informally, we can say that we *ignore* the numerical values of these quantities $\underline{x}_i$ and $\overline{x}_j$ – as long as they satisfy the corresponding inequalities $\underline{x}_i < \max(\underline{x}_1, \ldots, \underline{x}_n)$ and $\overline{x}_j > \min(\overline{x}_1, \ldots, \overline{x}_n)$. This "ignoring" is typical in *robust statistics*, when we find estimates in the presence of outliers; see, e.g., [4]. For example, one of the known robust estimates is a sample median. The median does not change if we slightly change the numerical values of all the x_i which are smaller or larger than the median value; so, in our informal terms, the median "ignores" all these values.

Similarly, in statistical clustering, it is often beneficial to "ignore" some of the input values; see, e.g., [6,13].

General case: analysis of the problem. Sample variance is a convex function of the variables $x_1, \ldots, x_n$, so the set $\{x_1, \ldots, x_n : V(x_1, \ldots, x_n) \le V_0\}$ is a convex set – specifically, an (infinite) ellipsoid. Thus, our main problem is the problem of minimizing and maximizing a linear function over a convex set – the intersection between the rectangular box $\mathbf{x}_1 \times \cdots \times \mathbf{x}_n$ and the ellipsoid.

Each linear function is both convex and concave, so to find the desired range, we can use known efficient algorithms for estimating the minimum of a convex function over a convex set and the maximum of a concave function over a convex set; see, e.g., [1] and references therein. [1]

General case: computational complexity of a known algorithm. A simplest way to optimize a linear objective function $\sum_{i=1}^n w_i \cdot x_i$ over an ellipsoid $\sum_{i=1}^n \sum_{j=1}^n a_{ij} \cdot (x_i - c_i) \cdot (x_j - c_j) \le w_0$ is to use the Lagrange multiplier technique, after which this constraint optimization problem reduces to an unconstrained optimization problem, with a quadratic objective function

$$J = \sum_{i=1}^n w_i \cdot x_i + \lambda \cdot \left(\sum_{i=1}^n \sum_{j=1}^n a_{ij} \cdot (x_i - c_i) \cdot (x_j - c_j) - w_0 \right).$$

This problem, in its turn, can be solved by equating all n partial derivatives $\dfrac{\partial J}{\partial x_i}$ to 0. A partial derivative of a quadratic function is a linear function, so, to find the optimizing values $x_1, \ldots, x_n$, we get a system of n linear equations with n unknowns. Such a system can be solved in time $O(n^{2.5})$ [1].

Our problem is slightly more complex, since we want to optimize over the *intersection* between an ellipsoid and a box. So, the computation time may be slightly larger.

[1] The authors are thankful to the anonymous referee for this important observation.

Our new results. Our main new result is that the above problem can be solved much faster – in time $O(n \cdot \log(n))$.

We also analyze what happens if we want to estimate other characteristics (such as sample variance itself) under interval uncertainty and constraints on sample variance, and what happens if we also take into account the possibility of malfunctioning sensors.

2. Main result

Main problem (reminder):

- *given:* n intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ and a number $V_0 \geq 0$;
- *compute:* the range

$$[\underline{e}, \overline{e}] = \{E(x_1, \ldots, x_n) \mid x_i \in \mathbf{x}_i \,\&\, V(x_1, \ldots, x_n) \leq V_0\}.$$

Main result. *There exists an algorithm that solves the main problem in time $O(n \cdot \log(n))$.*

Algorithm. This algorithm is as follows:

- First, we compute the values

$$\underline{E} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} \underline{x}_i \quad \text{and} \quad V^- \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (\underline{x}_i - \underline{E})^2;$$

$$\overline{E} \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i \quad \text{and} \quad V^+ \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} (\overline{x}_i - \overline{E})^2.$$

- If $V^- \leq V_0$, then we return $\underline{e} = \underline{E}$.
- If $V^+ \leq V_0$, then we return $\overline{e} = \overline{E}$.
- If at least one these inequalities does not hold, i.e., if $V_0 < V^-$ or $V_0 < V^+$, then we sort the all $2n$ endpoints $\underline{x}_i$ and $\overline{x}_i$ into a non-decreasing sequence

$$z_1 \leq z_2 \leq \cdots \leq z_{2n}$$

and consider $2n - 1$ *zones* $[z_k, z_{k+1}]$.
- For each zone:
  - for every i for which $\overline{x}_i \leq z_k$, we take $x_i = \overline{x}_i$;
  - for every i for which $z_{k+1} \leq \underline{x}_i$, we take $x_i = \underline{x}_i$;
  - for every other i, we mark the corresponding values indicating that for all of them, we will select the same value $x_i = \alpha$, this common value α is determined later in this step; let us denote the number of such marked i's by n_k.

  The common value α is determined from the condition that for the resulting selected vector x, we have $V(x) = V_0$, i.e., from solving the following quadratic equation:

$$\frac{1}{n} \cdot \left(\sum_{i:\overline{x}_i \leq z_k} (\overline{x}_i)^2 + \sum_{i:z_{k+1} \leq \underline{x}_i} \underline{x}_i^2 + n_k \cdot \alpha^2 \right) - \frac{1}{n^2} \cdot \left(\sum_{i:\overline{x}_i \leq z_k} \overline{x}_i + \sum_{i:z_{k+1} \leq \underline{x}_i} \underline{x}_i + n_k \cdot \alpha \right)^2 = V_0. \tag{9}$$

  Then:
  - if neither of the two roots of the above quadratic equation belongs to the zone, this zone is dismissed;
  - if one or more roots belong to the zone, then for each of these roots, based on this α, we compute the value

$$E_k = \frac{1}{n} \cdot \left(\sum_{i:\overline{x}_i \leq z_k} \overline{x}_i + \sum_{i:z_{k+1} \leq \underline{x}_i} \underline{x}_i + n_k \cdot \alpha \right). \tag{10}$$

- After that:
  - if $V_0 < V^-$, we return the smallest of the values E_k as $\underline{e}$;
  - if $V_0 < V^+$, we return the largest of the values E_k as $\overline{e}$.

Comments.

- For readers' convenience, all the proofs, including the proof of correctness of this algorithm, are placed in the special Proofs section; the main ideas behind these proofs are similar to the ideas behind the proofs from [7].

- Which values are "ignored" by this algorithm? In this algorithm, we use all the variables for which the optimal E_k contains all the values $x_i = \underline{x}_j$ or $\overline{x}_j$ – in the sense that if we change these values, the resulting estimates $\underline{e}$ or $\overline{e}$ will change.

  On the other hand, for the intervals for which we take $x_i = \alpha$, the values are, in effect, ignored – because when we slightly change the corresponding interval, the resulting estimates $\underline{e}$ or $\overline{e}$ will not change.

Toy example. Let us illustrate the above algorithm on a simple example in which we have two intervals $\mathbf{x}_1 = [-1, 0]$ and $\mathbf{x}_2 = [0, 1]$, and the bound $V_0 \geq 0$.

In this case, according to the above algorithm, we compute the values

$$\underline{E} = \frac{1}{2} \cdot (-1 + 0) = -0.5; \quad V^- = \frac{1}{2} \cdot (((-1) - (-0.5))^2 + (0 - (-0.5))^2) = 0.25;$$

$$\overline{E} = \frac{1}{2} \cdot (0 + 1) = 0.5; \quad V^+ = \frac{1}{2} \cdot ((0 - 0.5)^2 + (1 - 0.5)^2) = 0.25.$$

Both conditions $V_0 \geq V^-$ and $V_0 \geq V^+$ are equivalent to $V_0 \geq 0.25$. Thus, when $V_0 \geq 0.25$, we take into account both endpoints of both intervals, and get

$$\underline{e} = \frac{1}{2} \cdot ((-1) + 0) = -0.5; \quad \overline{e} = \frac{1}{2} \cdot (0 + 1) = 0.5.$$

When $V_0 < 0.25$, then, for computing both bounds $\underline{e}$ and $\overline{e}$, we need to consider different zones.

By sorting the 4 endpoints -1, 0, 0, and 1, we get $z_1 = -1 \leq z_2 = 0 \leq z_3 = 0 \leq z_4 = 1$. Thus, here, we have three zones $[z_1, z_2] = [-1, 0]$, $[z_2, z_3] = [0, 0]$, and $[z_3, z_4] = [0, 1]$.

(1) For the first zone $[z_1, z_2] = [-1, 0]$, according to the above algorithm, we select $x_2 = 0$ and $x_1 = \alpha$. To determine the value α, we form the quadratic equation (9):

$$\frac{1}{2} \cdot (0^2 + \alpha^2) - \frac{1}{4} \cdot (0 + \alpha)^2 = V_0.$$

This equation is equivalent to

$$\frac{1}{2} \cdot \alpha^2 - \frac{1}{4} \cdot \alpha^2 = \frac{1}{4} \cdot \alpha^2 = V_0,$$

hence $\alpha^2 = 4 \cdot V_0$ and $\alpha = \pm 2 \cdot \sqrt{V_0}$. Of the two roots $\alpha = -2 \cdot \sqrt{V_0}$ and $\alpha = 2 \cdot \sqrt{V_0}$, only the first root belongs to the zone $[-1, 0]$. For this root, we compute the value (10):

$$E_1 = \frac{1}{2} \cdot (0 + \alpha) = \frac{1}{2} \cdot \left(0 + \left(-2 \cdot \sqrt{V_0}\right)\right) = -\sqrt{V_0}.$$

(2) For the second zone $[z_2, z_3] = [0, 0]$, according to the above algorithm, we select $x_1 = x_2 = 0$. In this case, there is no need to compute α, so we directly compute

$$E_2 = \frac{1}{2} \cdot (0 + 0) = 0.$$

(3) For the third zone $[z_3, z_4] = [0, 1]$, according to the above algorithm, we select $x_1 = 0$ and $x_2 = \alpha$. To determine the value α, we form the quadratic equation (9):

$$\frac{1}{2} \cdot (0^2 + \alpha^2) - \frac{1}{4} \cdot (0 + \alpha)^2 = V_0.$$

This equation is equivalent to

$$\frac{1}{2} \cdot \alpha^2 - \frac{1}{4} \cdot \alpha^2 = \frac{1}{4} \cdot \alpha^2 = V_0,$$

hence $\alpha^2 = 4 \cdot V_0$ and $\alpha = \pm 2 \cdot \sqrt{V_0}$. Of the two roots $\alpha = -2 \cdot \sqrt{V_0}$ and $\alpha = 2 \cdot \sqrt{V_0}$, only the second root belongs to the zone $[0, 1]$. For this root, we compute the value (10):

$$E_3 = \frac{1}{2} \cdot (0 + \alpha) = \frac{1}{2} \cdot \left(0 + 2 \cdot \sqrt{V_0}\right) = \sqrt{V_0}.$$

Here, we have a value E_k for all three zones, so we return

$$\underline{e} = \min(E_1, E_2, E_3) = -\sqrt{V_0}; \quad \overline{e} = \max(E_1, E_2, E_3) = \sqrt{V_0}.$$

Toy example: discussion. When $V_0 \geq 0.25$, in our (provably optimal) computations, we use all four endpoints $\underline{x}_i$ and $\overline{x}_i$ of the original intervals $[\underline{x}_i, \overline{x}_i]$ to estimate the values $\underline{e}$ and $\overline{e}$.

When $V_0 < 0.25$, we use only the value $\overline{x}_1$ (to compute $\overline{e}$) and the value $\underline{x}_2$ (to compute $\underline{e}$). The endpoints $\underline{x}_1$ and $\overline{x}_2$ are ignored in the optimal algorithm – in the sense that to find the bounds $\underline{e}$ and $\overline{e}$, we use the values $x_i = \alpha$ which do not change if we slightly change these endpoints.

In particular, when $V_0 = 0$, the resulting interval $[\underline{e}, \overline{e}]$ simply coincides with $[0, 0]$ – i.e., with the intersection $[-1, 0] \cap [0, 1]$ of the two given intervals. This is exactly one of the two above cases in which the problem has an easy solution.

3. Additional results

What if we consider other statistical characteristics? What if, in addition to the sample mean, we also consider other statistical characteristics such as sample variance? It turns out that already for sample variance, the corresponding problem is NP-hard.

First auxiliary problem:

- *given:* n intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ and a number $V_0 \geq 0$;
- *compute:* the range

$$[\underline{v}, \overline{v}] \stackrel{\text{def}}{=} \{V(x_1, \ldots, x_n) \mid x_i \in \mathbf{x}_i \,\&\, V(x_1, \ldots, x_n) \leq V_0\}.$$

First auxiliary result. *The first auxiliary problem is NP-hard.*

Second auxiliary problem: discussion. What if some sensors malfunction? We will show that in this case, the problem becomes NP-hard already for the sample mean – and even without interval uncertainty.

Second auxiliary problem: precise formulation.

- *given:* n values $x_1, \ldots, x_n$, a number $q \leq n$, and a number E;
- *check:* whether there exists a subset $x_{i_1}, \ldots, x_{i_k}$, $i_1 < i_2 < \cdots < i_k$, $k \geq n - q$, of the original sample for which

$$E = \frac{x_{i_1} + \cdots + x_{i_k}}{k}.$$

Second auxiliary result. *The second auxiliary problem is NP-hard.*

Comment. It is worth mentioning that if instead of checking which elements are possible we simply want to know the range of possible values, then the problem becomes feasible for interval uncertainty – for the sample mean and, more generally, for any increasing function. Let us describe this problem in precise terms.[2]

Third auxiliary problem: precise formulation.

- *given:* n intervals $\mathbf{x}_1 = [\underline{x}_1, \overline{x}_1], \ldots, \mathbf{x}_n = [\underline{x}_n, \overline{x}_n]$, a number $q \leq n$, and a family of feasible functions $f_k(s_1, \ldots, s_k)$ which is a (non-strictly) increasing function of each of its variables;
- *compute:* the smallest $\underline{y}$ and the largest $\overline{y}$ of the values $f(x_{i_1}, \ldots, x_{i_k})$ for all $k \geq n - q$ and for all $x_i \in \mathbf{x}_i$.

Comment. The case of the sample mean corresponds to the functions $f_k(s_1, \ldots, s_k) = \dfrac{s_1 + \cdots + s_k}{k}$.

Third auxiliary result. *There exists a feasible (polynomial-time) algorithm for solving the third auxiliary problem.*

Algorithm for solving the third auxiliary problem is as follows:

- to compute $\underline{y}$, we sort the lower endpoints $\underline{x}_i$ into an increasing sequence $\underline{x}_{(1)} \leq \underline{x}_{(2)} \leq \cdots \leq \underline{x}_{(n)}$, and then compute

$$\underline{y} = \min(f_{n-q}(\underline{x}_{(1)}, \ldots, \underline{x}_{(n-q)}), f_{n-q+1}(\underline{x}_{(1)}, \ldots, \underline{x}_{(n-q+1)}), \ldots, f_n(\underline{x}_{(1)}, \ldots, \underline{x}_{(n)}));$$

[2] The authors are thankful to the anonymous referee for this formulation.

- to compute $\overline{y}$, we sort the upper endpoints $\overline{x}_i$ into a decreasing sequence $\overline{x}_{(1)} \geq \overline{x}_{(2)} \geq \cdots \geq \overline{x}_{(n)}$, and then compute

$$\overline{y} = \max(f_{n-q}(\overline{x}_{(1)}, \ldots, \overline{x}_{(n-q)}), f_{n-q+1}(\overline{x}_{(1)}, \ldots, \overline{x}_{(n-q+1)}), \ldots, f_n(\overline{x}_{(1)}, \ldots, \overline{x}_{(n)})).$$

4. Proofs

4.1. Proof that the main algorithm is correct

1°. Let us first show that it is sufficient to prove correctness for the case of the upper endpoint $\overline{e}$.

Indeed, one can easily see that if we replace the original values x_i with the new values $x_i' = -x_i$, then the sample mean changes sign $E' = -E$ while the sample variance remains the same $V' = V$.

When each x_i is known with interval uncertainty $x_i \in \mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$, the corresponding interval for $x_i' = -x_i$ is equal to $\mathbf{x}_i' = [-\overline{x}_i, -\underline{x}_i]$. The resulting interval $\mathbf{e}' = [\underline{e}', \overline{e}']$ for E' is similarly equal to $[-\overline{e}, -\underline{e}]$, so $\overline{e}' = -\underline{e}$ and thus, $\underline{e} = -\overline{e}'$.

Thus, if we know how to compute the upper endpoint $\overline{e}$ for an arbitrary set of intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$, we can compute $\underline{e}$ for a given set of intervals $\mathbf{x}_1 = [\underline{x}_1, \overline{x}_1], \ldots, \mathbf{x}_n = [\underline{x}_n, \overline{x}_n]$ as follows:

- we compute n auxiliary intervals $\mathbf{x}_i' = [-\overline{x}_i, -\underline{x}_i]$, $i = 1, \ldots, n$;
- we use the known algorithm to find the upper endpoint $\overline{e}'$ for the range of the sample mean when $x_i' \in \mathbf{x}_i'$ and $V(x') \leq V_0$;
- we take $\underline{e} = -\overline{e}'$.

2°. Let us prove that the largest possible value $\overline{e}$ is attained for some values $x_i \in [\underline{x}_i, \overline{x}_i]$ for which $V(x) \leq V_0$.

Indeed, the sample variance function $V(x_1, \ldots, x_n)$ is continuous; thus, the set of all the values $x = (x_1, \ldots, x_n)$ for which $V(x_1, \ldots, x_n) \leq V_0$ is closed.

The box $\mathbf{x}_1 \times \cdots \times \mathbf{x}_n$ is closed and bounded and thus, compact. The set S of all the values $x \in \mathbf{x}_1 \times \cdots \times \mathbf{x}_n$ for which $V(x) \leq V_0$ is a closed subset of a compact set and therefore, compact itself. A continuous function attains its maximum on a compact set at some point. In particular, this means that the function $E(x)$ attains its maximum $\overline{e}$ at some point x, i.e., that there exist values $x = (x_1, \ldots, x_n)$ for which $E(x_1, \ldots, x_n) = \overline{e}$.

In the following text, we will consider these optimizing values.

3°. Let us prove that for the optimizing vector x, for all i for which we have $x_i < E$, we have $x_i = \overline{x}_i$.

Indeed, since $V = M - E^2$, where $M \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2$, we conclude that

$$\frac{\partial V}{\partial x_i} = \frac{\partial M}{\partial x_i} - \frac{\partial E^2}{\partial x_i} = \frac{\partial M}{\partial x_i} - 2 \cdot E \cdot \frac{\partial E}{\partial x_i}.$$

Here, $\frac{\partial E}{\partial x_i} = \frac{1}{n}$, $\frac{\partial M}{\partial x_i} = \frac{2x_i}{n}$, and therefore,

$$\frac{\partial V}{\partial x_i} = \frac{2 \cdot (x_i - E)}{n}. \tag{11}$$

If we change only one value x_i, by replacing it with $x_i + \Delta x_i$, with a small Δx_i, the value of V changes by

$$\Delta V = \frac{\partial V}{\partial x_i} \cdot \Delta x_i + o(\Delta x_i) = \frac{2}{n} \cdot (x_i - E) \cdot \Delta x_i + o(\Delta x_i). \tag{12}$$

When $x_i < E$, i.e., when $x_i - E < 0$, then for small $\Delta x_i > 0$, we have a negative ΔV, i.e., the sample variance decreases, while the sample mean E increases by $\frac{1}{n} \cdot \Delta x_i > 0$. Thus, if we had $x_i < E$ and $x_i \neq \overline{x}_i$ for some i, then we could, by slightly increasing x_i, further increase E while decreasing V (and thus, keeping the constraint $V \leq V_0$). So, in this case, the vector x cannot be the one that maximizes E under the constraint $V \leq V_0$.

This conclusion proves that for the optimizing vector, when $x_i < E$, we have $x_i = \overline{x}_i$.

4°. Let us assume that an optimizing vector has a component x_i which is strictly inside the corresponding interval $[\underline{x}_i, \overline{x}_i]$, i.e., for which $\underline{x}_i < x_i < \overline{x}_i$. Due to Part 3 of this proof, we cannot have $x_i < E$, so we must have $x_i \geq E$. Let us prove that in this case,

- for every j for which $E \leq x_j < x_i$, we have $x_j = \overline{x}_j$, and
- for every k for which $x_k > x_i$, we have $x_k = \underline{x}_k$.

4.1°. Let us first prove that if $x_i \in (\underline{x}_i, \overline{x}_i)$, and $E \le x_j < x_i$, then $x_j = \overline{x}_j$.

We will prove this by contradiction. Indeed, let us assume that we have $E \le x_j < x_i$ and $x_j < \overline{x}_j$. In this case, we can, in principle, slightly increase x_j, to $x_j + \Delta x_j$ and slightly decrease x_i, to $x_i - \Delta x_i$, and still stay within the corresponding intervals $\mathbf{x}_i$ and $\mathbf{x}_j$. We select Δx_j and Δx_i in such a way that the resulting change ΔV in the sample variance V is non-negative. Here,

$$\Delta V = \frac{\partial V}{\partial x_j} \cdot \Delta x_j - \frac{\partial V}{\partial x_i} \cdot \Delta x_i + o(\Delta x_i) + o(\Delta x_j). \tag{13}$$

Substituting the formula (11) for the derivative $\frac{\partial V}{\partial x_j}$ into this formula, we conclude that

$$\Delta V = \frac{2}{n} \cdot ((x_j - E)\Delta x_j - (x_i - E) \cdot \Delta x_i) + o(\Delta x_i) + o(\Delta x_j). \tag{14}$$

Thus, for every Δx_j, to get $\Delta V = 0$, we select

$$\Delta x_i = \frac{x_j - E}{x_i - E} \cdot \Delta x_j + o(\Delta x_j). \tag{15}$$

For this selection, the sample variance does not change, but the sample mean E is changed by

$$\Delta E = \frac{1}{n} \cdot (\Delta x_j - \Delta x_i) = \left(1 - \frac{x_j - E}{x_i - E}\right) \cdot \Delta x_j + o(\Delta x_j) = \frac{x_i - x_j}{x_i - E} \cdot \Delta x_j + o(\Delta x_j). \tag{16}$$

Since $x_j < x_i$, for small Δx_j, we have $\Delta E > 0$. Thus, we can further increase the sample mean without violating the constraint $V \le V_0$. This contradicts to our assumption that x is the optimizing vector. Thus, when $E < x_j < x_i$, we cannot have $x_j < \overline{x}_j$ – so we must have $x_j = \overline{x}_j$.

4.2°. Let us first prove that if $x_i \in (\underline{x}_i, \overline{x}_i)$, $E \le x_i$, and $x_k > x_i$, then $x_k = \underline{x}_k$.

Let us assume that we have $x_k > x_i$ and $x_k > \underline{x}_k$. In this case, we can, in principle, slightly increase x_i, to $x_i + \Delta x_i$ and slightly decrease x_k, to $x_k - \Delta x_k$, and still stay within the corresponding intervals $\mathbf{x}_i$ and $\mathbf{x}_k$. We select Δx_i and Δx_k in such a way that the resulting change ΔV in the sample variance V is non-negative. Here,

$$\Delta V = \frac{\partial V}{\partial x_i} \cdot \Delta x_i - \frac{\partial V}{\partial x_k} \cdot \Delta x_k + o(\Delta x_i) + o(\Delta x_k) = \frac{2}{n} \cdot ((x_i - E)\Delta x_i - (x_k - E) \cdot \Delta x_k) + o(\Delta x_i) + o(\Delta x_k). \tag{17}$$

Thus, for every Δx_i, to get $\Delta V = 0$, we select

$$\Delta x_k = \frac{x_i - E}{x_k - E} \cdot \Delta x_i + o(\Delta x_i). \tag{18}$$

For this selection, the sample variance does not change, but the sample mean E is changed by

$$\Delta E = \frac{1}{n} \cdot (\Delta x_i - \Delta x_k) = \left(1 - \frac{x_i - E}{x_k - E}\right) \cdot \Delta x_i + o(\Delta x_i) = \frac{x_k - x_i}{x_k - E} \cdot \Delta x_i + o(\Delta x_i). \tag{19}$$

Since $x_k > x_i$, for small Δx_i, we have $\Delta E > 0$. Thus, we can further increase the sample mean without violating the constraint $V \le V_0$. This contradicts our assumption that x is the optimizing vector. Thus, when $x_i < x_k$, we cannot have $x_k > \underline{x}_k$ – so we must have $x_k = \underline{x}_k$.

5°. Let us now consider the case when for all the components $x_i \ge E$ of the optimizing vector x, we have either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$. Let us show that in this case, all the values x_i for which $x_i = \overline{x}_i$ are smaller than or equal to all the values x_j for which $x_j = \underline{x}_j$.

We will prove this statement by contradiction. Let us assume that there exist i and j for which $E \le x_j < x_i$, $x_j = \underline{x}_j$ and $x_i = \overline{x}_i$. In this case, we can slightly increase the value x_j, to $x_j + \Delta x_j$, and slightly decrease the value x_i, to $x_i - \Delta x_i$, and still stay within the corresponding intervals. Similarly to Part 4 of this proof, for every $\Delta x_j > 0$, to get $\Delta V = 0$, we must select

$$\Delta x_i = \frac{x_j - E}{x_i - E} \cdot \Delta x_j + o(\Delta x_j). \tag{20}$$

For this selection, the sample variance does not change, but the sample mean E is changed by

$$\Delta E = \frac{1}{n} \cdot (\Delta x_j - \Delta x_i) = \left(1 - \frac{x_j - E}{x_i - E}\right) \cdot \Delta x_j + o(\Delta x_j) = \frac{x_i - x_j}{x_i - E} \cdot \Delta x_j + o(\Delta x_j). \tag{21}$$

Since $x_j < x_i$, for small Δx_j, we have $\Delta E > 0$. Thus, we can further increase the sample mean without violating the constraint $V \le V_0$. This contradicts our assumption that x is the optimizing vector. So, when $E \le x_j < x_i$, we cannot have $x_j = \underline{x}_j$ and $x_i = \overline{x}_i$.

This contradiction proves that all the values x_i for which $x_i = \overline{x}_i$ are indeed smaller than or equal to all the values x_j for which $x_j = \underline{x}_j$.

6°. Due to Parts 3, 4, and 5 of this proof, there exists a threshold value α such that

- for all j for which $x_j < \alpha$, we have $x_j = \overline{x}_j$, and
- for all k for which $x_k > \alpha$, we have $x_k = \underline{x}_k$.

Indeed, in the case described in Part 4, at such α, we can take the value x_i that is strictly inside the corresponding interval $\mathbf{x}_i$. In the case described in Part 5, since all the upper endpoints from the optimizing vector are smaller than or equal to all the lower endpoints, we can take any value α between the largest of the optimal values $\overline{x}_j$ and smallest of the optimal values $\underline{x}_k$.

7°. Let us show that because of the property proven in Part 6, once we know to which zone α belongs, we can uniquely determine all the components x_j of the corresponding vector x – a candidate for the optimal vector.

7.1°. Indeed, if $\overline{x}_j < \alpha$, then, since we have $x_j < \overline{x}_j$, we get $x_j < \alpha$. Thus, due to Part 6, we have $x_j = \overline{x}_j$.

7.2°. If $\alpha < \underline{x}_j$, then, since we have $\underline{x}_j < x_j$, we get $\alpha < x_j$. Thus, due to Part 6, we have $x_j = \underline{x}_j$.

7.3°. Let us now consider the remaining case when neither of the above two conditions is satisfied and thus, we have $\underline{x}_j \le \alpha \le \overline{x}_j$.

In this case, we cannot have $x_j < \alpha$, because then, due to Part 6, we would have $x_j = \overline{x}_j$ and thus, $\overline{x}_j < \alpha$, which contradicts to the inequality $\alpha \le \overline{x}_j$.

Similarly, we cannot have $\alpha < x_j$, because then, due to Part 6, we would have $x_j = \underline{x}_j$ and thus, $\alpha < \underline{x}_j$, which contradicts to the inequality $\underline{x}_j \le \alpha$.

Thus, the only possible value here is $x_j = \alpha$.

7.3°. Overall, we conclude that for each α, we get exactly the arrangement formulated in our algorithm.

8°. Let us prove that when $V_0 < V^+$, then the maximum is attained when $V = V_0$.

Let us prove this by contradiction. Let us assume that $V_0 < V^+$ and that the maximum of E is attained for some vector $x = (x_1, \ldots, x_n)$, with $x_i \in [\underline{x}_i, \overline{x}_i]$, for which $V(x) < V_0$.

Since $V < V_0 < V^+$, we have $V(x) < V^+ = V(\overline{x}_1, \ldots, \overline{x}_n)$. Thus, $x = (x_1, \ldots, x_n) \ne \overline{x} \stackrel{\text{def}}{=} (\overline{x}_1, \ldots, \overline{x}_n)$ – otherwise, we would get $V(x) = V(\overline{x}) = V^+$. So, there exists an index i for which $x_i \ne \overline{x}_i$. Since $x_i \in [\underline{x}_i, \overline{x}_i]$, this means that $x_i < \overline{x}_i$. Thus, we can increase x_i by a small positive value $\varepsilon > 0$, to a new value $x_i' = x_i + \varepsilon > x_i$, and still remain inside the interval $[\underline{x}_i, \overline{x}_i]$.

The function $V(x_1, \ldots, x_n)$ describing sample variance continually depends on x_i. Since $V(x) < V_0$, for sufficiently small ε, we will have

$$V(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) < V_0.$$

Thus, the new vector still satisfies the constraint – but for this new vector, the sample mean is larger (by $\varepsilon/n > 0$) than for the original vector x.

This contradicts our assumption that the sample mean $E(x)$ of the vector x is the largest possible under the given constraint $V \le V_0$.

The above contradiction shows that when $V_0 < V^+$, then for the optimizing vector x, we have $V(x) = V_0$. This fact enables us to determine α – as do in the algorithm – by solving the equation $V(x(\alpha)) = V_0$, where $x(\alpha)$ is a vector corresponding to the given α.

Correctness is proven.

4.2. Proof that the main algorithm takes time $O(n \cdot \log(n))$

Sorting $2n$ numbers takes time $O(n \cdot \log(n))$; see, e.g., [1].

Once the values are sorted, we can then go zone-by-zone, and perform the corresponding computations. A straightforward implementation of the above algorithm would require time $O(n^2)$ – for each of $2n$ zones, we need linear time to compute several sums of n numbers.

However, in reality, only the sum for the first zone requires linear time. Once we have the sums for each zone, computing the sum for the next zone requires changing a few terms – values x_j which changed status. Each value x_j changes once, so overall, to compute all these sums, we still need linear time.

Thus, after sorting, the algorithm requires only linear computations time $O(n)$. So, if the endpoints are already given to us as sorted, we only take linear time.

If we still need to sort, then we need time

$$O(n \cdot \log(n)) + O(n) = O(n \cdot \log(n)).$$

4.3. Proof that the first auxiliary problem is NP-hard

By definition, a problem $\mathcal{P}$ is NP-hard if every problem from a certain class NP can be reduced to $\mathcal{P}$; see, e.g., [1,10]. A usual way to prove NP-hardness of a problem $\mathcal{P}$ is to show that some known NP-hard problem $\mathcal{P}_0$ can be reduced to $\mathcal{P}$. Indeed, in this case, by definition of NP-hardness, every problem from the class NP can be reduced to $\mathcal{P}_0$, and since $\mathcal{P}_0$ can be reduced to the problem $\mathcal{P}$, every problem from the class NP can be reduced to $\mathcal{P}$ as well. By definition, this means that $\mathcal{P}$ is NP-hard.

In our proof, as a known NP-hard problem, we take the above-mentioned problem of computing the upper endpoint $\overline{V}$ of the sample variance under interval uncertainty. Let us show that this problem can be reduced to the new problem of computing the range $[\underline{v}(V_0), \overline{v}(V_0)]$. Indeed, one can easily check that:

- if $V_0 \le \overline{V}$, then $\overline{v}(V_0) = V_0$; and
- if $V_0 > \overline{V}$, then $\overline{v}(V_0) = \overline{V} < V_0$.

Thus, $\overline{v}(V_0) < V_0$ if and only if $\overline{V} < V_0$. So, if we could compute $\overline{v}(V_0)$ for a given value V_0, we could then compare this value with V_0 and check whether, for a given number V_0, we have $\overline{V} < V_0$. Thus, by using bisection, we could locate $\overline{V}$ with a given accuracy quickly. In other words, if we can solve our problem, then we can solve the problem of computing $\overline{V}$ with a given accuracy as well. Reduction is proven, so our problem is indeed NP-hard.

4.4. Proof that the second auxiliary problem is NP-hard

This proof is similar to the proof of the previous result, except that as the known NP-hard problem $\mathcal{P}_0$, we take the following *subset sum* problem [1,10]:

- *given:* $m + 1$ positive integers $s_1, \ldots, s_m, S$;
- *check:* whether it is possible to find a subset $s_{i_1}, \ldots, s_{i_k}, i_1 < i_2 < \cdots < i_k$ for which $s_{i_1} + \cdots + s_{i_k} = S$.

We reduce each instance of the subset sum problem to our problem as follows: we take $n = m + 1$, $x_i = s_i$ for $i \le m$, $x_{m+1} = -S, E = 0$, and $q = n-1$. In this case, if there is a subset for which $s_{i_1} + \cdots + s_{i_k} = S$, then $x_{i_1} + \cdots + x_{i_k} + x_{m+1} = 0$ and thus, $E = 0$.

Vice versa, $E = 0$ means that the sum of the selected values x_i is 0. Since all the values x_i except for the last one are positive, the only way for the sum to be 0 is to have the only negative term $x_{m+1} = -S < 0$ in this sum. In this case, the fact that $x_{i_1} + \cdots + x_{i_k} + x_{m+1} = 0$ means that $s_{i_1} + \cdots + s_{i_k} = S$.

The reduction is proven, so the second auxiliary problem is indeed NP-hard.

4.5. Proof of correctness of the algorithm for solving the third auxiliary problem

For each set of indices $i_1, \ldots, i_k$, since the function $f_k(s_1, \ldots, s_k)$ is monotonic, the smallest possible value of $f_k(x_{i_1}, \ldots, x_{i_k})$ when $x_i \in \mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ is attained when each of the values x_{i_j} attains its smallest possible value $\underline{x}_{i_j}$. In this case, the value of the desired quantity $y = f_k(x_{i_1}, \ldots, x_{i_k})$ is equal to $y = f_k(\underline{x}_{i_1}, \ldots, \underline{x}_{i_k})$.

Similarly, for a given k, the smallest possible value of $f_k(\underline{x}_{i_1}, \ldots, \underline{x}_{i_k})$ is attained when $\underline{x}_{i_1}, \ldots, \underline{x}_{i_k}$ take the k smallest values, i.e., when $\underline{x}_{i_1} = \underline{x}_{(1)}, \ldots, \underline{x}_{i_k} = \underline{x}_{(k)}$. In this case, the value of the desired quantity $y = f_k(x_{i_1}, \ldots, x_{i_k})$ is equal to $f_k(\underline{x}_{(1)}, \ldots, \underline{x}_{(k)})$.

The overall minimum is attained for some k, so the absolute minimum $\underline{y}$ can be computed as the minimum of the above expressions over all possible values $k \ge n - q$. This leads us exactly to the formula given in the above algorithm.

Similarly, we can justify the formula for $\overline{y}$. The correctness is proven.

5. Conclusions

Main problems: reminder. In many practical situations, for a sample consisting of n observations, we do not know the exact values $x_1, \ldots, x_n$ of the corresponding characteristic. Instead, from measurements, we know the lower and upper bounds $\underline{x}_i$ and $\overline{x}_i$ on these values ($\underline{x}_i \le x_i \le \overline{x}_i$), and also, we know that the values x_i cannot differ too much from each other, i.e., we know the upper bound V_0 on the sample variance V: $V \le V_0$. Under these assumptions, we need to find the find the ranges of possible values of different statistical characteristics such as the sample mean E and the sample variance V. These are the main problems analyzed in our paper.

Main problems: results. The problem of computing the range $[\underline{E}, \overline{E}]$ for the mean E can be, in principle, solved by using known feasible algorithms for convex optimization; however, these general-purpose convex optimization algorithms require computation time $O(n^{2.5})$ which is much longer than the usual linear-time ($O(n)$) algorithm for computing the sample mean E. In this paper, we show that the desired range can be computed much faster than in the general convex case – namely, in time $O(n \cdot \log(n))$, almost as fast as in the absence of interval uncertainty.

We also prove that the problem of computing the range of the sample variance is, in general, NP-hard.

Additional problems. In the formulation of the main problems, we implicitly assumed that the measuring instruments function correctly – or at least that we know when these instruments malfunction. In reality, sometimes, sensors malfunction without any indication of the malfunctioning. In such situations, some of the resulting ranges $[\underline{x}_i, \overline{x}_i]$ come from these malfunctioning cases and thus, do not contain the actual values x_i. Usually, we know the probability of malfunctioning, so we know the percentage of measurements which can be erroneous. In this case, we want to find estimates the bounds on the mean (and variance) of the actual values x_i – estimates that would not be affected by the erroneous sensor readings.

In this paper, we show that for the sample mean E, the problem of computing its range $[\underline{E}, \overline{E}]$ is still feasible, even when some of the ranges $[\underline{x}_i, \overline{x}_i]$ may be caused by malfunctioning sensors. However, if we are interested in knowing which values E from the range $[\underline{E}, \overline{E}]$ are possible values of the sample mean and which are not, the problem becomes NP-hard – even in the absence of interval uncertainty, when all correct (=not malfunction-caused) measurement results are exact.

Remaining open problems. Our results about estimating the mean in situations when some sensors malfunction can be extended to arbitrary monotonic statistical characteristics. It is desirable to provide a similar analysis for other statistical characteristics which are not necessarily monotonic, such as higher moments, covariance, correlation, etc.

Acknowledgments

This work was partially supported by a training fellowship from the National Library of Medicine (NLM) to the Keck Center for the Interdisciplinary Bioscience Training of the Gulf Coast Consortia (NLM Grant No. T15 LM0070931).

The authors are thankful to the anonymous referees for valuable suggestions.

References

[1] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stein, Introduction to Algorithms, MIT Press, Cambridge, MA, 2009.
[2] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, M. Aviles, Computing variance for interval data is NP-hard, ACM SIGACT News 33 (2) (2002) 108–118.
[3] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, M. Aviles, Exact bounds on finite populations of interval data, Reliable Computing 11 (3) (2005) 207–233.
[4] P.J. Huber, E. Ronchetti, Robust Statistics, Wiley, Hoboken, NJ, 2009.
[5] L. Jaulin, M. Kieffer, O. Didrit, E. Walter, Applied Interval Analysis, with Examples in Parameter and State Estimation, Robust Control and Robotics, Springer-Verlag, London, 2001.
[6] S. Kim, M.G. Tadesses, M. Vannucci, Variable selection in clustering via Dirichlet process mixture models, Biometrika 93 (4) (2006) 877–893.
[7] V. Kreinovich, G. Xiang, S. Ferson, Computing mean and variance under Dempster–Shafer uncertainty: towards faster algorithms, International Journal of Approximate Reasoning 42 (3) (2006) 212–227.
[8] R.E. Moore, R.B. Kearfott, M.J. Cloud, Introduction to Interval Analysis, SIAM Press, Philadelphia, PA, 2009.
[9] H.T. Nguyen, O. Kosheleva, V. Kreinovich, S. Ferson, Trade-off between sample size and accuracy: case of measurements under interval uncertainty, International Journal of Approximate Reasoning 50 (8) (2009) 1164–1176.
[10] C. Papadimitriou, Computational Complexity, Addison Wesley, Reading, MA, 1994.
[11] S. Rabinovich, Measurement Errors and Uncertainties: Theory and Practice, Springer-Verlag, New York, 2005.
[12] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, Boca Raton, FL, 2004.
[13] M.G. Tadesse, N. Sha, M. Vannucci, Bayesian variable selection in clustering high-dimensional data, Journal of the American Statistical Association 100 (470) (2005) 602–617.
[14] G. Xiang, M. Ceberio, V. Kreinovich, Computing population variance and entropy under interval uncertainty: linear-time algorithms, Reliable Computing 13 (6) (2007) 467–488.