New Algorithms for Statistical Analysis of Interval Data

Gang Xiang, Scott A. Starks,
Vladik Kreinovich, and Luc Longpré
NASA Pan-American Center for Earth and
Environmental Studies (PACES)
University of Texas, El Paso, TX 79968, USA
vladik@utep.edu

Abstract

It is known that in general, statistical analysis of interval data is an NP-hard problem: even computing the variance of interval data is, in general, NP-hard. Until now, only one case was known for which a feasible algorithm can compute the variance of interval data: the case when all the measurements are accurate enough – so that even after the measurement, we can distinguish between different measured values $\widetilde{x}_i$. In this paper, we describe several new cases in which feasible algorithms are possible – e.g., the case when all the measurements are done by using the same (not necessarily very accurate) measurement instrument – or at least a limited number of different measuring instruments.

1 Introduction

Traditional statistical data processing. Once we have several results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ of measuring some physical quantity – e.g., the amount of pollution in a lake – traditional statistical data processing starts with computing the sample average $E = E(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, the sample median $M = M(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, and the sample variance $V = V(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ of these results. For example, $E(\widetilde{x}_1, \ldots, \widetilde{x}_n) = \dfrac{1}{n}\sum_{i=1}^{n} \widetilde{x}_i$, and $V(\widetilde{x}_1, \ldots, \widetilde{x}_n) = \dfrac{1}{n}\sum_{i=1}^{n} (\widetilde{x}_i - E)^2$.

To check whether the distribution is symmetric, we also compute its sample skewness $S(\widetilde{x}_1, \ldots, \widetilde{x}_n) = \dfrac{1}{n}\sum_{i=1}^{n} (\widetilde{x}_i - E)^3$.

Interval data. The values $\widetilde{x}_i$ come from measurements, and measurements are never 100% accurate. In many real-life situations, the only information about the corresponding measurement errors is the upper bound Δ_i on the absolute value of the measurement error. As a result, the only information we have about the actual value x_i of each measured quantity is that x_i belongs to the interval $\mathbf{x}_i \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

Statistical analysis of interval data: a problem. For interval data, instead of the exact values of E, M, and V, it is desirable to get the intervals $\mathbf{E}$, $\mathbf{M}$, and $\mathbf{V}$ of possible values, intervals formed by all possible values of E (correspondingly, M or V) when each x_i takes values from the interval $\mathbf{x}_i$.

Statistical analysis of interval data: what is known. Computing $\mathbf{E} = [\underline{E}, \overline{E}]$ and $\mathbf{M} = [\underline{M}, \overline{M}]$ is straightforward: indeed, both the sample average E and the sample median M are (non-strictly) increasing functions of the variables $x_1, \ldots, x_n$. So, the smallest possible value $\underline{E}$ (correspondingly, $\underline{M}$) is attained when we take the smallest possible values $\underline{x}_1, \ldots, \underline{x}_n$ from the corresponding intervals; similarly, the largest possible value $\overline{E}$ (correspondingly, $\overline{M}$) is attained when we take the largest possible values $\overline{x}_1, \ldots, \overline{x}_n$ from the corresponding intervals. Thus, $\underline{E} = E(\underline{x}_1, \ldots, \underline{x}_n)$, $\overline{E} = E(\overline{x}_1, \ldots, \overline{x}_n)$, $\underline{M} = M(\underline{x}_1, \ldots, \underline{x}_n)$, and $\overline{M} = M(\overline{x}_1, \ldots, \overline{x}_n)$.

On the other hand, computing the exact range $\mathbf{V} = [\underline{V}, \overline{V}]$ of V turns out to be an NP-hard problem; specifically, computing the upper endpoint $\overline{V}$ is NP-hard (see, e.g., [2]).

It is worth mentioning that computing the lower endpoint $\underline{V}$ is feasible; in [3], we show that it can done in time $O(n \cdot \log(n))$.

In the same paper [2] in which we prove that computing $\overline{V}$ is, in general, NP-hard, we also show that in the case when the measuring instruments are accurate enough – so that even after the measurements, we can distinguish between different measured values $\widetilde{x}_i$ (e.g, if the corresponding intervals $\mathbf{x}_i$ do not intersect) – we can compute $\overline{V}$ (hence, $\mathbf{V}$) in feasible time (actually, quadratic time).

In some practical examples, the measurement instruments are indeed very accurate, but in many other practical cases, their accuracy may be much lower – so the algorithm from [2] is not applicable.

What we are planning to do. In this paper, we describe new practically useful cases when we can compute the range $\mathbf{V}$ of the variance and the range $\mathbf{S}$ of the skewness by a feasible (polynomial-time) algorithm.

The first case is when all the measurements are made by the same measuring instrument or by similar measurement instruments. In this case, none of two input intervals $\mathbf{x}_i$ is a proper subset of one another, and as a result, we can find the exact range $\mathbf{V}$ in time $O(n \cdot \log(n))$, and the exact range $\mathbf{S}$ in time $O(n^2)$.

The second case is when instead of a single type of measuring instruments, we use a limited number ($m > 1$) of different types of measuring instruments. It turns out that in this case, we can compute $\mathbf{V}$ in polynomial time $O(n^m)$ and $\mathbf{S}$ in polynomial time $O(n^{2m})$.

The third case is related to privacy in statistical databases; see details below.

2 First Case: Measurements by Same Measuring Instrument

When is it difficult to compute the variance of interval data? Mathematical description. In the proof that computing variance is NP-hard (given in [2]), we used interval data in which some intervals are proper subintervals of others: $[\underline{x}_i, \overline{x}_i] \subseteq (\underline{x}_j, \overline{x}_j)$.

When is it difficult to compute the variance of interval data? Practical interpretation. From the practical viewpoint, this situation makes perfect sense: the interval data may contain values measurement by more accurate measuring instruments – that produce narrower intervals $\mathbf{x}_i$ – and by less accurate measurement instruments – that produce wider intervals $\mathbf{x}_j$. When we measure the same value $x_i = x_j$, once with an accurate measurement instrument, and then with a less accurate instrument, then it is quite possible that the wider interval corresponding to the less accurate measurement properly contains the narrower interval corresponding to the more accurate instrument.

Similarly, if we measure close values $x_i \approx x_j$, it is quite possible that the wider interval coming from the less accurate instrument contains the narrower interval coming from the more accurate instrument.

Idea: how to avoid such difficult-to-compute situations. In view of the above analysis, a natural way to avoid such difficult-to-compute situations is to restrict ourselves to situations when all the measurement are done with the same measuring instrument.

Reformulating this idea is mathematical terms. For a single measuring instrument, it is not very probable that in two different measurements, we get two intervals in which one is a proper subinterval of the other: $[\underline{x}_i, \overline{x}_i] \subseteq (\underline{x}_j, \overline{x}_j)$.

Let us show that if this subset property is satisfied, then we have a feasible algorithm for computing $\overline{V}$. For each interval $\mathbf{x} = [\underline{x}, \overline{x}]$, we will denote its half-width $(\overline{x} - \underline{x})/2$ by Δ, and its midpoint $(\underline{x} + \overline{x})/2$ by $\widetilde{x}$.

Definition 1 *By an* interval data*, we mean a finite set of intervals* $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Definition 2 *We say that the interval data* $\mathbf{x}_1, \ldots, \mathbf{x}_n$ satisfies the subset property *if* $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$ *for every* i *and* j.

Definition 3 *For n real numbers $x_1, \ldots, x_n$, their* variance $V(x_1, \ldots, x_n)$ *is defined in the standard way – as* $V \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - E^2$, *where* $E \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$.

Definition 4 *By the* interval variance $\mathbf{V}$ *of the interval data, we mean the interval* $\mathbf{V} \stackrel{\text{def}}{=} \{V(x_1, \ldots, x_n) \,|\, x_i \in \mathbf{x}_i\}$ *filled by the values* $V(x_1, \ldots, x_n)$ *corresponding to different* $x_i \in \mathbf{x}_i$.

Theorem 1 *There exists an algorithm that computes the variance* $\mathbf{V}$ *of the interval data in time* $O(n \cdot \log(n))$ *for all the cases in which the interval data satisfies the subset property.*

Let us now show that in the situation when no two intervals are subsets of each other, we have a feasible algorithm for computing the range $\mathbf{S}$ of the skewness S.

For $n = 1$, we have $x_1 = E$, so the skewness is 0. For $n = 2$, we have $E = \frac{x_1 + x_2}{2}$, hence

$$S = \frac{1}{2} \cdot [(x_1 - E)^3 + (x_2 - E)^3] = \frac{1}{2} \cdot \left[\left(\frac{x_1 - x_2}{2} \right)^3 + \left(\frac{x_2 - x_1}{2} \right)^3 \right] = 0.$$

Thus, non-trivial skewness values are only possible for $n > 2$.

Definition 5 *By the* interval skewness $\mathbf{V}$ *of the interval data, we mean the interval* $\mathbf{S} \stackrel{\text{def}}{=} \{S(x_1, \ldots, x_n) \,|\, x_i \in \mathbf{x}_i\}$ *filled by the values* $S(x_1, \ldots, x_n)$ *corresponding to different* $x_i \in \mathbf{x}_i$.

Theorem 2 *There exists an algorithm that computes the skewness* $\mathbf{S}$ *of the interval data in time* $O(n^2)$ *for all the cases in which the interval data satisfies the subset property.*

Proof of Theorem 1. In order to compute the interval $\mathbf{V}$, we must compute both endpoints $\underline{V}$ and $\overline{V}$ of this interval.

The proof of the theorem consists of three parts:

- in Part A, we will mention that the algorithm for computing $\underline{V}$ in time $O(n \cdot \log(n))$ is already known;
- in Part B, we describe a new algorithm for computing $\overline{V}$;
- in Part C, we prove that the new algorithm is correct and has the desired complexity.

A. Algorithm for computing $\underline{V}$ with desired time complexity is already known. The algorithm for computing $\underline{V}$ in time $O(n \cdot \log(n))$ is described in [4].

B. New algorithm for computing $\overline{V}$: description. The proposed algorithm for computing $\overline{V}$ is as follows:

- First, we sort n intervals $\mathbf{x}_i$ in lexicographic order:

$$\mathbf{x}_1 \le_{\text{lex}} \mathbf{x}_2 \le_{\text{lex}} \ldots \le_{\text{lex}} \mathbf{x}_n,$$

  where $[\underline{a}, \overline{a}] \le_{\text{lex}} [\underline{b}, \overline{b}]$ if and only if either $\underline{a} < \underline{b}$, or $\underline{a} = \underline{b}$ and $\overline{a} \le \overline{b}$.

- Second, we use bisection to find the value k ($1 \le k \le n$) for which the following two inequalities hold:

$$\widetilde{x}_k + \frac{1}{n} \cdot \sum_{i=1}^{k-1} \Delta_i \le \frac{1}{n} \cdot \sum_{i=k+1}^{n} \Delta_i + \frac{1}{n} \cdot \sum_{i=1}^{n} \widetilde{x}_i; \tag{1}$$

$$\widetilde{x}_{k+1} + \frac{1}{n} \cdot \sum_{i=1}^{k} \Delta_i \ge \frac{1}{n} \cdot \sum_{i=k+2}^{n} \Delta_i + \frac{1}{n} \cdot \sum_{i=1}^{n} \widetilde{x}_i. \tag{2}$$

  At each iteration of this bisection, we have an interval $[k^-, k^+]$ that is guaranteed to contain k. In the beginning, $k^- = 1$ and $k^+ = n$. At each stage, we compute the midpoint $k_{\text{mid}} = \lfloor (k^- + k^+)/2 \rfloor$, and check both inequalities (1) and (2) for $k = k_{\text{mid}}$. Then:

  - If both inequalities (1) and (2) hold for his k, this means that we have found the desired k.
  - If (1) holds but (2) does not hold, this means that the desired value k is larger than k_{mid}, so we keep k^+ and replace k^- with $k_{\text{mid}} + 1$.
  - If (2) holds but (1) does not hold, this means that the desired value k is smaller than k_{mid}, so we keep k^- and replace k^+ with $k_{\text{mid}} - 1$.

- Once k is found, we compute

$$V_k \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{k} \underline{x}_i^2 + \frac{1}{n} \cdot \sum_{i=k+1}^{n} \overline{x}_i^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^{k} \underline{x}_i + \frac{1}{n} \cdot \sum_{i=k+1}^{n} \overline{x}_i \right)^2. \tag{3}$$

  This is the desired value $\overline{V}$.

C. Proof of correctness and complexity. Let us prove that this algorithm indeed produces the correct result and indeed requires time $O(n \cdot \log(n))$.

1°. Let us first prove that if the interval data satisfies the subset property, then, after we sort these elements in lexicographic order, both the lower endpoints $\underline{x}_i$

and the upper endpoints $\overline{x}_i$ are sorted in non-decreasing order: $\underline{x}_i \le \underline{x}_{i+1}$ and $\overline{x}_i \le \overline{x}_{i+1}$.

Indeed, by definition of a lexicographic order, we always have $\underline{x}_i \le \underline{x}_{i+1}$. If $\underline{x}_i = \underline{x}_{i+1}$, then, by definition of the lexicographic order, we have $\overline{x}_i \le \overline{x}_{i+1}$.

If $\underline{x}_i < \underline{x}_{i+1}$, then we cannot have $\overline{x}_i > \overline{x}_{i+1}$ – otherwise, we would have $(\underline{x}_{i+1}, \overline{x}_{i+1}) \subset \mathbf{x}_i$ which would contradict the subset property. Hence, if $\underline{x}_i < \underline{x}_{i+1}$, we also have $\overline{x}_i \le \overline{x}_{i+1}$. The statement is proven.

It is known that sorting requires time $O(n \cdot \log(n))$; see, e.g., [1].

In the following text, we will assume that the sequence of intervals has been sorted in this manner.

2°. Let us now prove that the desired maximum of the variance V is attained when each variable x_i is at one of the endpoints of the corresponding interval $\mathbf{x}_i$.

Indeed, if the maximum is attained in the interior point of this interval, this would means that in this point, $\partial V/\partial x_i = 0$ and $\partial^2 V/\partial x_i^2 \le 0$. For variance, $\partial V/\partial x_i = (2/n)\cdot(x_i - E)$, so $\partial^2 V/\partial x_i^2 = (2/n)\cdot(1-1/n) > 0$ – hence maximum cannot be inside.

3°. Let us show the maximum is attained at a vector

$$x = (\underline{x}_1, \ldots, \underline{x}_k, \overline{x}_{k+1}, \ldots, \overline{x}_n) \tag{4}$$

in which we first have lower endpoints and then upper endpoints.

What we need to prove is that there exists a maximizing vector in which, once we have an upper endpoint, what follows will also be an upper endpoint, i.e., in which we cannot have $x_k = \overline{x}_k > \underline{x}_k$ and $x_{k+1} = \underline{x}_{k+1} < \overline{x}_{k+1}$.

For that, let us start with a maximizing vector in which this property does not hold, i.e., in which $x_k = \overline{x}_k > \underline{x}_k$ and $x_{k+1} = \underline{x}_{k+1} < \overline{x}_{k+1}$ for some k. Based on this vector, we will now construct a different maximizing vector with the desired property. For that, let us consider two cases: $\Delta_k < \Delta_{k+1}$ and $\Delta_k \ge \Delta_{k+1}$, where $\Delta_i \stackrel{\text{def}}{=} (\overline{x}_i - \underline{x}_i)/2$ is the half-width of the interval $\mathbf{x}_i$.

In the first case, let us replace $\overline{x}_k = \underline{x}_k + 2\Delta_k$ with $\underline{x}_k$, and $\underline{x}_{k+1}$ with $\underline{x}_{k+1} + 2\Delta_k$ (since $\Delta_k < \Delta_{k+1}$, this new value is $< \overline{x}_{k+1}$). Here, the average E remains the same, so the only difference between the new value V' of the variance and its old value V comes from the change in terms x_k^2 and x_{k+1}^2. In other words,

$$V' - V = \frac{1}{n} \cdot ((\underline{x}_{k+1} + 2\Delta_k)^2 - \underline{x}_{k+1}^2) - \frac{1}{n} \cdot ((\underline{x}_k + 2\Delta_k)^2 - \underline{x}_k^2).$$

Opening parentheses and simplifying the resulting expression, we conclude that $V' - V = (4\Delta_k/n)\cdot(\underline{x}_{k+1} - \underline{x}_k)$. Since V is the maximum, we must have $V' - V \le$

0, hence $\underline{x}_{k+1} \le \underline{x}_k$. Due to our ordering, we thus have $\underline{x}_{k+1} = \underline{x}_k$. Since we assumed that $\Delta_k < \Delta_{k+1}$, we have $\overline{x}_k = \underline{x}_k + 2\Delta_k < \overline{x}_{k+1} = \underline{x}_{k+1} + 2\Delta_{k+1}$, hence the interval $\mathbf{x}_k$ is a proper subset of $\mathbf{x}_{k+1}$ – which is impossible.

In the second case, when $\Delta_k \ge \Delta_{k+1}$, let us replace $\overline{x}_k$ with $\overline{x}_k - 2\Delta_{k+1}$ (which is still $\ge \underline{x}_k$), and $\underline{x}_{k+1} = \overline{x}_{k+1} - 2\Delta_{k+1}$ with $\overline{x}_{k+1}$. Here, the average E remains the same, and the only difference between the new value V' of the variance and its old value V comes from the change in terms x_k^2 and x_{k+1}^2, hence

$$V' - V = \frac{1}{n} \cdot (\overline{x}_{k+1}^2 - (\overline{x}_{k+1} - 2\Delta_{k+1})^2) - \frac{1}{n} \cdot (\overline{x}_k^2 - (\overline{x}_k - 2\Delta_{k+1})^2),$$

i.e., $V' - V = (4\Delta_{k+1}/n) \cdot (\overline{x}_{k+1} - \overline{x}_k)$. Since V is the maximum, we must have $V' - V \le 0$, hence $\overline{x}_{k+1} \le \overline{x}_k$. Due to our ordering, we thus have $\overline{x}_{k+1} = \overline{x}_k$. Since we assumed that $\Delta_k \ge \Delta_{k+1}$, we have $\underline{x}_k = \overline{x}_k - 2\Delta_k \ge \underline{x}_{k+1} = \overline{x}_{k+1} - 2\Delta_{k+1}$, i.e., $\mathbf{x}_k \subseteq \mathbf{x}_{k+1}$. Since intervals cannot be proper subsets of each other, we thus have $\mathbf{x}_k = \mathbf{x}_{k+1}$. In this case, we can simply swap the values x_k and x_{k+1}, variance will not change.

If necessary, we can perform this swap for all needed k; as a result, we get the maximizing vector with the desired property.

4°. Due to Part 3 of this proof, the desired value $\overline{V} = \max V$ is the largest of $n+1$ values (3) corresponding to $k = 0, 1, \ldots, n$.

In principle, to compute $\overline{V}$, we can therefore compute each of these values and find the largest of them. Computing each value takes $O(n)$ times, so computing $n+1$ such values would require time $O(n^2)$. Let us show that we can compute $\overline{V}$ faster.

We must find the index k for which V_k is the largest. For the desired k, we have $V_k \ge V_{k-1}$ and $V_k \ge V_{k+1}$. Due to (3), we conclude that

$$V_k - V_{k-1} = \frac{1}{n} \cdot \left(\underline{x}_k^2 - \overline{x}_k^2\right)$$

$$- \left(\frac{1}{n} \cdot \sum_{i=1}^{k} \underline{x}_i + \frac{1}{n} \cdot \sum_{i=k+1}^{n} \overline{x}_i\right)^2 + \left(\frac{1}{n} \cdot \sum_{i=1}^{k-1} \underline{x}_i + \frac{1}{n} \cdot \sum_{i=k}^{n} \overline{x}_i\right)^2. \qquad (5)$$

Each pair of terms in the right-hand side of (5) can be simplified if we use the fact that $a^2 - b^2 = (a-b) \cdot (a+b)$ and use the notations Δ_k and $\widetilde{x}_k \stackrel{\text{def}}{=} (\underline{x}_k + \overline{x}_k)/2$. First, we get $\underline{x}_k^2 - \overline{x}_k^2 = (\underline{x}_k - \overline{x}_k) \cdot (\underline{x}_k + \overline{x}_k) = -4\Delta_k \cdot \widetilde{x}_k$. Second, we get

$$\left(\frac{1}{n} \cdot \sum_{i=1}^{k-1} \underline{x}_i + \frac{1}{n} \cdot \sum_{i=k}^{n} \overline{x}_i\right)^2 - \left(\frac{1}{n} \cdot \sum_{i=1}^{k} \underline{x}_i + \frac{1}{n} \cdot \sum_{i=k+1}^{n} \overline{x}_i\right)^2 =$$

$$\frac{2}{n} \cdot (\overline{x}_k - \underline{x}_k) \cdot \left(\frac{1}{n} \cdot \sum_{i=1}^{k-1} \underline{x}_i + \frac{1}{n} \cdot \widetilde{x}_k + \frac{1}{n} \cdot \sum_{i=k+1}^{n} \overline{x}_i\right).$$

Here, $\overline{x}_k - \underline{x}_k = 2\Delta_k$, hence the formula (5) takes the following form:

$$V_k - V_{k-1} = \frac{4}{n} \cdot \Delta_k \cdot \left(-\widetilde{x}_k + \frac{1}{n} \cdot \sum_{i=1}^{k-1} \underline{x}_i + \frac{1}{n} \cdot \widetilde{x}_k + \frac{1}{n} \cdot \sum_{i=k+1}^{n} \overline{x}_i \right).$$

Since $V_k \geq V_{k-1}$ and $\Delta_k > 0$, we conclude that

$$-\widetilde{x}_k + \frac{1}{n} \cdot \sum_{i=1}^{k-1} \underline{x}_i + \frac{1}{n} \cdot \widetilde{x}_k + \frac{1}{n} \cdot \sum_{i=k+1}^{n} \overline{x}_i \geq 0. \tag{6}$$

Substituting the expressions $\underline{x}_i = \widetilde{x}_i - \Delta_i$ and $\overline{x}_i = \widetilde{x}_i + \Delta_i$ into the formula (6) and moving all the negative terms to the other side of the inequality, we get the inequality (1). Similarly, the inequality $V_{k+1} \leq V_k$ leads to (2).

When k increases, the left-hand side of the inequality (1) increases – because $\widetilde{x}_k$ increases as the average of the two increasing values $\underline{x}_k$ and $\overline{x}_k$, and the sum is increasing. Similarly, the right-hand side of this inequality decreases with k. Thus, if this inequality holds for k, it should also hold for all smaller values, i.e., for $k-1$, $k-2$, etc.

Similarly, in the second desired inequality (2), when k increases, the left-hand side of this inequality increases, while the right-hand side decreases. Thus, if this inequality is true for k, it is also true for $k+1$, $k+2$, ...

If both inequalities (1) and (2) are true for two different values $k < k'$, then they should both be true for all the values intermediate between k and k', i.e., for $k+1, k+2, \ldots, k'-1$. If (1) and and (2) are both true for k and $k+1$, this means that in both cases, we have equality, thus $V_k = V_{k+1}$, so it does not matter which of these values k we take.

Thus, modulo this equality case, there is, in effect, only one k for which both inequalities are true, and this k can be found by the bisection method as described in the above algorithm.

How long does this algorithm take? In the beginning, we only know that k belongs to the interval $[1, n]$ of width $O(n)$. At each stage of the bisection step, we divide the interval (containing k) in half. After I iterations, we decrease the width of this interval by a factor of 2^I. Thus, to find the exact value of k, we must have I for which $O(n)/2^I = 1$, i.e., we need $I = O(\log(n))$ iterations. On each iteration, we need $O(n)$ steps, so we need a total of $O(n \cdot \log(n))$ steps. With $O(n \cdot \log(n))$ steps for sorting, and $O(n)$ for computing the variance, we get a $O(n \cdot \log(n))$ algorithm.

The theorem is proven.

Proof of Theorem 2.
1°. In order to compute the interval $\mathbf{S}$, we must compute both endpoints $\underline{S}$ and $\overline{S}$ of this interval.

Let us first show that if we can compute $\overline{S}$, then we can easily compute $\underline{S}$ as well. Indeed, skewness is an odd function: $S(-x_1, \ldots, -x_n) = -S(x_1, \ldots, x_n)$;

thus, for intervals $-\mathbf{x}_i = -[\underline{x}_i, \overline{x}_i] = [-\overline{x}_i, -\underline{x}_i]$, we have $\mathbf{S}(-\mathbf{x}_1, \ldots, -\mathbf{x}_n) = -\mathbf{S}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. From this relation between the skewness intervals, we can conclude that $\overline{S}(-\mathbf{x}_1, \ldots, -\mathbf{x}_n) = -\underline{S}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Thus, if we know how to compute $\overline{S}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ for an arbitrary collection of intervals $\mathbf{x}_i$, we can thus compute $\underline{S}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ as $-\overline{S}(-\mathbf{x}_1, \ldots, -\mathbf{x}_n)$

In view of this comment, in the remaining part of the proof, we will only consider an algorithm for computing $\overline{S}$.

2°. As we have shown while proving Theorem 1, since the interval data satisfies the subset property, after we sort these elements in lexicographic order, both the lower endpoints $\underline{x}_i$ and the upper endpoints $\overline{x}_i$ are sorted in non-decreasing order: $\underline{x}_i \le \underline{x}_{i+1}$ and $\overline{x}_i \le \overline{x}_{i+1}$.

3°. The maximum of a differentiable function $S(x_1, \ldots, x_n)$ on an interval $[\underline{x}_i, \overline{x}_i]$ can be attained either in an internal point of this interval, or at one of the endpoints.

If the maximum is attained at an internal point, then the first derivative is 0 $\left(\dfrac{\partial S}{\partial x_i} = 0\right)$ and the second derivative should be non-positive $\left(\dfrac{\partial^2 S}{\partial x_i^2} \le 0\right)$.

If the maximum is attained at the left endpoint, the function S cannot be increasing at this point, so we must have $\dfrac{\partial S}{\partial x_i} \le 0$. Similarly, if the maximum is attained at the right endpoint, the function S cannot be decreasing at this point, so we must have $\dfrac{\partial S}{\partial x_i} \ge 0$.

For skewness,

$$\frac{\partial S}{\partial x_i} = \frac{3}{n} \cdot (x_i - E)^2 - \frac{3}{n} \cdot \sum_{j=1}^{n} (x_j - E)^2 \cdot \frac{\partial E}{\partial x_i}.$$

Since $\dfrac{\partial E}{\partial x_i} = \dfrac{1}{n}$, we thus get $\dfrac{\partial S}{\partial x_i} = \dfrac{3}{n} \cdot ((x_i - E)^2 - V)$. So, the first derivative of S has the same sign as the expression $(x_i - E)^2 - V$.

To compute the second derivative of S, we must take into account that $\dfrac{\partial V}{\partial x_i} = \dfrac{2}{n} \cdot (x_i - E)$, hence

$$\frac{\partial^2 S}{\partial x_i^2} = \frac{3}{n} \cdot \left(2(x_i - E) - 2(x_i - E) \cdot \frac{1}{n} - \frac{2}{n} \cdot (x_i - E) + \frac{2}{n} \cdot \sum_{j=1}^{n} (x_j - E) \cdot \frac{1}{n}\right).$$

Since $\sum\limits_{j=1}^{n} (x_j - E) = 0$, we conclude that

$$\frac{\partial^2 S}{\partial x_i^2} = \frac{3}{n} \cdot 2 \cdot \left(1 - \frac{2}{n}\right) \cdot (x_i - E).$$

We have already mentioned that the problem of computing skewness only makes sense for $n > 2$, because for $n \le 2$, the skewness is identically 0. For $n > 2$, the second derivative $\dfrac{\partial^2 S}{\partial x_i^2}$ has the same sign as the expression $x_i - E$.

Thus, for skewness, we value x_i at which the maximum is attained satisfies one of the following three conditions:

- either $\underline{x}_i < x_i < \overline{x}_i$, $(x_i - E)^2 - V = 0$, and $x_i - E \le 0$,
- or $x_i = \underline{x}_i$ and $(x_i - E)^2 - V \le 0$,
- or $x_i = \overline{x}_i$ and $(x_i - E)^2 - V \ge 0$.

In the first case, $(x_i - E)^2 = V = \sigma^2$, hence $x_i - E = \pm\sigma$. Since $x_i - E \le 0$, we cannot have $x_i - E = \sigma$, so in this case, $x_i = E - \sigma$. In the second case, $(x_i - E)^2 \le V = \sigma^2$, hence $E - \sigma \le x_i \le E + \sigma$. In the third case, $(x_i - E)^2 \ge V = \sigma^2$, so either $x_i \le E - \sigma$ or $x_i \ge E + \sigma$. So:

- either $\underline{x}_i < x_i = E - \sigma < \overline{x}_i$,
- or $x_i = \underline{x}_i$ and $E - \sigma \le \underline{x}_i \le E + \sigma$,
- or $x_i = \overline{x}_i$ and either $\overline{x}_i \le E - \sigma$ or $\overline{x}_i \ge E + \sigma$.

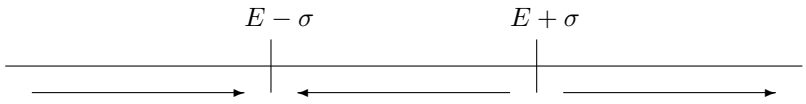
In all three cases, the desired maximum of the skewness S is attained when x_i is either at one of the endpoints of the corresponding interval $\mathbf{x}_i$, or has the value $\mu \stackrel{\text{def}}{=} E - \sigma$.

4°. Let us now deduce a more specific information about the values x_i at which the maximum is attained.

Based on the above description of possible cases, once we know how the intervals are located in relation to $E - \sigma$ and $E + \sigma$, we can sometimes uniquely determine the value x_i at which the maximum is attained. Namely,

- If $\overline{x}_i \le E - \sigma$, then the maximum cannot be attained at an internal point and it cannot be attained at the value $\underline{x}_i$, so it is attained when $x_i = \overline{x}_i$.
- If $\underline{x}_i \le E - \sigma \le \overline{x}_i \le E + \sigma$, then the maximum can only be attained when $x_i = E - \sigma$.
- If $E - \sigma \le \underline{x}_i \le E + \sigma$, then the maximum is attained at $x_i = \underline{x}_i$.
- Finally, if $E + \sigma \le \underline{x}_i$, then the maximum is attained at $x_i = \overline{x}_i$.

These conclusions can be described in the following graphical manner, in which the arrows indicate the direction towards the corresponding maximum:



The only case when we cannot exactly determine the optimal value x_i is when the interval $\mathbf{x}_i$ contains the value $E+\sigma$: it this case, we may have $x_i = \overline{x}_i$, and we may also have $x_i = \max(E-\sigma, \underline{x}_i)$.

5°. Let us show that the maximum of skewness is always attained at a vector $x = (x_1, \ldots, x_n)$ which can be divided into three consequent fragments (some of which may be empty):

- first, we have values $\overline{x}_i$ which are smaller than $E-\sigma$;
- then, we have the values $\max(E-\sigma, \underline{x}_i)$;
- finally, we have the values $\overline{x}_i$ which are larger than $E+\sigma$.

All the intervals $\mathbf{x}_i$ that do not contain $E+\sigma$ inside naturally fall into this scheme. The only intervals that we do need to consider to prove this result are the intervals that do contain $E+\sigma$. For each of these intervals, the corresponding values x_i are either $\max(E-\sigma, \underline{x}_i)$ or $\overline{x}_i$. What we claim is that after we sort the intervals in lexicographic order, we will first have the values equal to $\max(E-\sigma, \underline{x}_i)$, and then the values equal to $\overline{x}_i$. In other words, once we have a value $x_i = \overline{x}_i$, all the following values will also be of the same type.

We will show that if there is an optimizing vector at which this condition is not satisfied, then we can rearrange it into a new vector with the same optimal value of S for which this condition holds.

Indeed, let us start with a vector for which, for some i, for two consequent intervals $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$, a value $x_i = \overline{x}_i \geq E+\sigma$ is followed by a value $x_{i+1} = \max(E-\sigma, \underline{x}_{i+1}) \leq E+\sigma$. If there are several such indices i, we take the smallest i with this property.

According to Part 2 of this proof, we have $\underline{x}_i \leq \underline{x}_{i+1} \leq E+\sigma$ and $E+\sigma \leq \overline{x}_i \leq \overline{x}_{i+1}$. Thus, $x_i = \overline{x}_i \leq \overline{x}_{i+1}$ and $x_i = \overline{x}_i \geq E+\sigma \geq \underline{x}_{i+1}$; hence, $x_i \in \mathbf{x}_{i+1}$. Similarly, $x_{i+1} \in \mathbf{x}_i$. Thus, we can "swap" the values x_i and x_{i+1}: as a new value of x_i, we take the old value of x_{i+1}, and vice versa. The swap does not change the average E and does not change the sample skewness S, so the function S attains the maximum at the new values as well.

As a result of this swap, if there is now a value i' for which $\overline{x}_{i'}$ is followed by $\max(E-\sigma, \underline{x}_{i'+1})$, this value i' has to be equal to at least $i+1$. If there still is such an index i', we apply a new swap again and thus again increase the smallest problematic value i. After $\leq n$ such swaps, there will be no problematic cases anymore, so we will get a sequence which has the desired property.

6°. To determine the optimal vector x, we must thus select a zone $[x_{(p)}, x_{(p+1)}]$ that contains $\mu = E - \sigma$, and an index k at which the optimal value x_i switches from $\max(\mu, \underline{x}_i)$ to $\overline{x}_i$.

Once p and k are fixed, we can uniquely determine each of the optimal values x_i – some as known numbers, some as equal to the (unknown) value μ:

- when $\overline{x}_i \le x_{(p)}$, we have $x_i = \overline{x}_i$;
- when $\underline{x}_i < x_{(p)} < x_{(p+1)} \le \overline{x}_i$ and $i < k$, we have $x_i = \mu$;
- when $x_{(p+1)} \le \underline{x}_i$ and $i < k$, we have $x_i = \underline{x}_i$;
- finally, when $i \ge k$, we have $x_i = \overline{x}_i$.

To find μ, we must use the fact that $\mu = E - \sigma$. Specifically, the average E can be determined as

$$\frac{1}{n} \cdot \sum_{i \in N'} x_i + \frac{n - n'}{n} \cdot (E - \sigma) = E,$$

where the sum is taken over the set N' of all the indices for which x_i is known, and n' is the total number of such indices. Similarly, the sample second moment $E^2 + \sigma^2$ can be determined as

$$\frac{1}{n} \cdot \sum_{i \in N'} x_i^2 + \frac{n - n'}{n} \cdot (E - \sigma)^2 = E^2 + \sigma^2.$$

From the first of these equations, we can determine σ as a linear function of E. Substituting this expression into the second equation, we get a quadratic equation with the only unknown σ, from which we can determine σ. Then, we can use the first equation to find E – and hence find $\mu = E - \sigma$.

If the resulting value of μ is indeed within the zone $[x_{(p)}, x_{(p+1)}]$, then we compute the sample skewness for the corresponding values x_i. Specifically, the skewness can be computed as

$$\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^3 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^3 - \frac{3 \cdot E}{n} \cdot \sum_{i=1}^{n} x_i^2 + \frac{3 \cdot E^2}{n} \cdot \sum_{i=1}^{n} x_i - E^3 =$$

$$\frac{1}{n} \cdot \sum_{i=1}^{n} x_i^3 - \frac{3 \cdot E}{n} \cdot \sum_{i=1}^{n} x_i^2 + 2 \cdot E^3 =$$

$$\frac{1}{n} \cdot \sum_{i \in N'} x_i^3 + \frac{n - n'}{n} \cdot \mu^3 - \frac{3 \cdot E}{n} \cdot \left(\sum_{i \in N'} x_i^2 + (n - n') \cdot \mu^2 \right) + 2 \cdot E^3.$$

The largest of these skewnesses is the desired value $\overline{S}$.

7°. How much times does this algorithm require? Sorting takes time $O(n \cdot \log(n))$.

For n interval data points, we have $2n$ possible zone and n possible indices k – totally, $O(n^2)$ possible pairs (p, k). For the first pair, computing the corresponding values n', $\sum\limits_{i \in N'} x_i$, $\sum\limits_{i \in N'} x_i^2$, and $\sum\limits_{i \in N'} x_i^3$ requires linear time. For each next pair, we, in general, change one value in comparison with the previous pair, so each new computation requires a constant number of steps. Thus, for $O(n^2)$ pairs, we need $O(n^2)$ time. (In some cases, we change more than one value, but still, each value changes only once, so we still need $O(n^2)$ times.)

So, overall, we need time $O(n \cdot \log(n)) + O(n) + O(n^2) = O(n^2)$.

The theorem is proven.

3 Second Case: Using a Limited Number of Different Types of Measuring Instruments

In this case, the interval data consists of m families of intervals such that within each family, no two intervals are proper subsets of each other.

Similarly to the proof of Theorem 1, we can conclude that if we sort each family in lexicographic order, then, within each family, the maximum of V is attained on one of the sequences (4). Thus, to find the desired maximum $\overline{V}$, it is sufficient to know the value $k_\alpha \le n$ corresponding to each of m families. Overall, there are $\le n^m$ combinations of such values. For the first combination, computing the corresponding value of the variance requires $O(n)$ steps. Each next combination differs from the previous one by a single term, so overall, we need $O(n^m)$ steps to compute all the values of the variance – and thus, to find the largest of them, which is $\overline{V}$.

For $\overline{S}$, we need to consider n^2 options for each of the m subsequences corresponding to a single MI; thus, overall, we must consider $(O(n^2))^m = O(n^{2m})$ possible combinations – hence we need time $O(n^{2m})$.

4 Third Case: Privacy in Statistical Databases

Why privacy leads to intervals. When the measurements $\widetilde{x}_i$ correspond to data that we want to keep private, e.g., health parameters of different patients, we do not want statistical programs to have full access to the data – because otherwise, by computing sufficiently many different statistics, we would be able to uniquely reconstruct the actual values $\widetilde{x}_i$. One way to prevent this from happening is to supply the statistical data processing programs not with the exact data, but only with intervals of possible values of this data, intervals corresponding to a fixed partition; see, e.g., [4]. For example, instead of the exact age, we tell the program that a person's age is between 30 and 40.

Privacy-related intervals: formal description. To implement the above idea, we need to fix a partition, i.e., to fix the values $t_1 < t_2 < \ldots < t_n$. In this case, instead of the actual value of the quantity, we return the partition-related interval $[t_i, t_{i+1}]$ that contains this value.

Relation with the first case. From the *practical* viewpoint, the first case – when intervals come from measurement uncertainty – is very different from this case, in which intervals are formed by us, to maintain privacy. However, from the purely *mathematical* viewpoint, privacy-related intervals $[t_i, t_{i+1}]$ satisfy the same property as intervals from the first case: none of them is a proper subset of the other.

This relation leads to a feasible algorithm for privacy-related intervals. Because of the above relation, apply the algorithm described in Section 2 to privacy-related case intervals as well. In other words, for privacy-related interval data, it is also possible to and compute the exact range $\mathbf{V}$ and $\mathbf{S}$ in polynomial time – namely, we can compute $\mathbf{V}$ in time $O(n \cdot \log(n))$ and $\mathbf{S}$ in time $O(n^2)$.

Acknowledgments. This work was supported in part by NASA grant NCC5-209, by the AFOSR grant F49620-00-1-0365, by NSF grants EAR-0112968, EAR-0225670, and EIA-0321328, by the Army Research Laboratories grant DATM-05-02-C-0046, and by NIH grant 3T34GM008048-20S1.

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

References

[1] Cormen Th. H., Leiserson C. E., Rivest R. L., and Stein C.: Introduction to Algorithms, MIT Press, Cambridge, MA, 2001.

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

[3] Granvilliers, L., Kreinovich, V., Müller, L.: Novel Approaches to Numerical Software with Result Verification", In: Alt, R., Frommer, A., Kearfott, R. B., Luther, W. (eds.), Numerical software with result verification, Springer Lectures Notes in Computer Science, 2004, Vol. 2991, pp. 274–305.

[4] Kreinovich, V., Longpré, L.: Computational complexity and feasibility of data processing and interval computations, with extension to cases when we have partial information about probabilities, In: Brattka, V., Schroeder,

M., Weihrauch, K., Zhong, N.: Proc. Conf. on Computability and Complexity in Analysis CCA'2003, Cincinnati, Ohio, USA, August 28–30, 2003, pp. 19–54.