Exact Bounds on Finite Populations of Interval Data

Scott Ferson[1], Lev Ginzburg[1],
Vladik Kreinovich[2], Luc Longpré, and Monica Aviles[2]

[1]Applied Biomathematics, 100 North Country Road,
Setauket, NY 11733, USA, {scott,lev}@ramas.com
[2]Computer Science Department, University of Texas at El Paso
El Paso, TX 79968, USA, {vladik,longpre}@cs.utep.edu

Abstract

In this paper, we start research into using intervals to bound the impact of bounded measurement errors on the computation of bounds on finite population parameters ("descriptive statistics"). Specifically, we provide a feasible (quadratic time) algorithm for computing the lower bound $\underline{\sigma^2}$ on the finite population variance function of interval data. We prove that the problem of computing the upper bound $\overline{\sigma^2}$ is, in general, NP-hard. We provide a feasible algorithm that computes $\overline{\sigma^2}$ under reasonable easily verifiable conditions, and provide preliminary results on computing other functions of finite populations.

1 Introduction

1.1 Formulation of the Problem

When we have n measurement results $x_1, \ldots, x_n$, traditional data processing techniques start with with computing such population parameters ("descriptive statistics") $f(x) \stackrel{\text{def}}{=} f(x_1, \ldots, x_n)$ as their finite population average

$$\mu \stackrel{\text{def}}{=} \frac{x_1 + \ldots + x_n}{n}$$

and their finite population variance

$$\sigma^2 \stackrel{\text{def}}{=} \frac{(x_1 - \mu)^2 + \ldots + (x_n - \mu)^2}{n} \tag{1.1}$$

(or, equivalently, the finite population standard deviation $\sigma = \sqrt{\sigma^2}$); see, e.g., [14].

In some practical situations, we only have intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values of x_i. This happens, for example, if instead of observing the actual value x_i of the random variable, we observe the value $\widetilde{x}_i$ measured by an instrument with a known upper bound Δ_i on the measurement error. In other words, we are assuming that

$$\mathbf{x}_i = \widetilde{x}_i + \Delta_i \cdot [-1, 1],$$

where the measurement error bounds $\Delta_i \cdot [-1, 1]$ are assumed to be known. Then, the actual (unknown) value of each measured quantity x_i is within the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

In these situations, for each population parameter $y = f(x_1, \ldots, x_n)$, we can only determine the *set* of possible values of y:

$$\mathbf{y} = \{f(x_1, \ldots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n.\}$$

For population parameters described by continuous functions $f(x_1, \ldots, x_n)$, this set is an interval. In particular, the sets of possible values of μ and σ^2 are also intervals. The interval $\boldsymbol{\mu}$ for the finite population average can be obtained by using straightforward interval computations, i.e., by replacing each elementary operation with numbers by the corresponding operation of interval arithmetic:

$$\boldsymbol{\mu} = \frac{\mathbf{x}_1 + \ldots + \mathbf{x}_n}{n}. \tag{1.2}$$

What is the interval $[\underline{\sigma^2}, \overline{\sigma^2}]$ of possible values for finite population variance σ^2?

When all the intervals $\mathbf{x}_i$ intersect, then it is possible that all the actual (unknown) values $x_i \in \mathbf{x}_i$ are the same and hence, that the finite population variance is 0. In other words, if the intervals have a non-empty intersection, then $\underline{\sigma^2} = 0$. Conversely, if the intersection of $\mathbf{x}_i$ is empty, then σ^2 cannot be 0, hence $\underline{\sigma^2} > 0$. The question is (see, e.g., [18]): What is the total set of possible values of σ^2 when the above intersection is empty?

The practical importance of this question was emphasized, e.g., in [10, 11] on the example of processing geophysical data.

A similar question can (and will) be asked not only about the finite population variance, but also about other finite population parameters.

1.2 For this Problem, Traditional Interval Methods Sometimes Lead to Excess Width

Let us show that for this problem, traditional interval methods sometimes lead to excess width.

1.2.1 Straightforward Interval Computations

Historically the first method for computing the enclosure for the range is the method which is sometimes called "straightforward" interval computations. This method is based on the fact that inside the computer, every algorithm consists of elementary operations (arithmetic operations, min, max, etc.). For each elementary operation $f(a, b)$, if we know the intervals $\mathbf{a}$ and $\mathbf{b}$ for a and b, we can compute the exact range $f(\mathbf{a}, \mathbf{b})$. The corresponding formulas form the so-called *interval arithmetic*. In straightforward interval computations, we repeat the computations forming the program f step-by-step, replacing each operation with real numbers by the corresponding operation of interval arithmetic. It is known that, as a result, we get an enclosure for the desired range.

For the problem of computing the range of finite population average, as we have mentioned, straightforward interval computations lead to exact bounds. The reason: in the above formula for μ, each interval variable only occurs once [6].

For the problem of computing the range of finite population variance, the situation is somewhat more difficult, because in the expression (1.1), each variable x_i occurs several times: explicitly, in $(x_i - \mu)^2$, and explicitly, in the expression for μ. In this cases, often, dependence between intermediate computation results leads to excess width of the results of straightforward interval computations. Not surprisingly, we do get excess width when applying straightforward interval computations to the formula (1.1).

For example, for $\mathbf{x}_1 = \mathbf{x}_2 = [0, 1]$, the actual $\sigma^2 = (x_1 - x_2)^2/4$ and hence, the actual range $\boldsymbol{\sigma^2} = [0, 0.25]$. On the other hand, $\boldsymbol{\mu} = [0, 1]$, hence

$$\frac{(\mathbf{x}_1 - \boldsymbol{\mu})^2 + (\mathbf{x}_2 - \boldsymbol{\mu})^2}{2} = [0, 1] \supset [0, 0.25].$$

It is worth mentioning that there are other formulas one can use to compute the variance of a finite population: e.g., the formula

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \mu^2.$$

In this formula too, each variable x_i occurs several times, as a result of which we get excess width: for $\mathbf{x}_1 = \mathbf{x}_2 = [0, 1]$, we get $\boldsymbol{\mu} = [0, 1]$ and

$$\frac{\mathbf{x}_1^2 + \mathbf{x}_2^2}{2} - \boldsymbol{\mu}^2 = [-1, 1] \supset [0, 0.25].$$

Unless there is a general formula for computing the variance of a finite population in which each interval variable only occurs once, then without using a numerical algorithm (as contrasted with am analytical expression), it is probably not possible to avoid excess interval width caused by dependence. The fact that we prove that the problem of computing of computing the exact bound for the finite population variance is computationally difficult (in precise terms, NP-hard) makes us believe that no such formula for finite population variance is possible.

1.2.2 Centered Form

A better range is often provided by a *centered form*, in which a range $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of a smooth function on a box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$ is estimated as

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) \subseteq f(\widetilde{x}_1, \ldots, \widetilde{x}_n) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(\mathbf{x}_1, \ldots, \mathbf{x}_n) \cdot [-\Delta_i, \Delta_i],$$

where $\widetilde{x}_i = (\underline{x}_i + \overline{x}_i)/2$ is the interval's midpoint and $\Delta_i = (\underline{x}_i - \overline{x}_i)/2$ is its half-width.

When all the intervals are the same, e.g., when $\mathbf{x}_i = [0,1]$, the centered form does not lead to the desired range. Indeed, the centered form always produced an interval centered in the point $f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$. In this case, all midpoints $\widetilde{x}_i$ are the same (e.g., equal to 0.5), hence the finite population variance $f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ is equal to 0 on these midpoints. Thus, as a result of applying the centered form, we get an interval centered at 0, i.e., the interval whose lower endpoint is negative. In reality, σ^2 is always non-negative, so negative values of σ^2 are impossible.

The upper endpoint produced by the centered form is also different from the upper endpoint of the actual range: e.g., for $\mathbf{x}_1 = \mathbf{x}_2 = [0,1]$, we have $\dfrac{\partial f}{\partial x_1}(x_1, x_2) = (x_1 - x_2)/2$, hence

$$\frac{\partial f}{\partial x_1}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\mathbf{x}_1 - \mathbf{x}_2}{2} = [-0.5, 0.5].$$

A similar formula holds for the derivative with respect to x_2. Since $\Delta_i = 0,5$, the centered form leads to:

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n) \subseteq 0 + [-0.5, 0.5] \cdot [-0.5, 0.5] + [-0.5, 0.5] \cdot [-0.5, 0.5] = [-0.5, 0.5]$$

– an excess width in comparison with the actual range $[0, 0.25]$.

1.3 For this Problem, Traditional Optimization Methods Sometimes Require Unreasonably Long Time

A natural way to solve the problem of computing the exact range $[\underline{\sigma^2}, \overline{\sigma^2}]$ of the finite population variance is to solve it as a constrained optimization problem. Specifically, to find $\underline{\sigma^2}$, we must find the minimum of the function (1.1) under the conditions $\underline{x}_1 \le x_1 \le \overline{x}_1$, …, $\underline{x}_n \le x_n \le \overline{x}_n$. Similarly, to find $\overline{\sigma^2}$, we must find the maximum of the function (1.1) under the same conditions.

There exist optimization techniques that lead to computing "sharp" (exact) values of $\min(f(x))$ and $\max(f(x))$. For example, there is a method described in [7] (and effectively implemented). However, the behavior of such general constrained optimization algorithms is not easily predictable, and can, in general, be exponential in n.

For small n, this is quite doable, but for large n, the exponential computation time grows so fast that for reasonable n, it becomes unrealistically large: e.g., for $n \approx 300$, it becomes larger than the lifetime of the Universe.

1.4 We Need New Methods

Summarizing: the existing methods are either not always efficient, or do not always provide us with sharp estimates for $\underline{\sigma^2}$ and $\overline{\sigma^2}$. So, we need new methods.

In this paper, we describe several new methods for computing the variance of the finite population, and start analyzing the problem of computing other population parameters over interval data.

2 First Result: Computing $\underline{\sigma^2}$

First, we design a *feasible* algorithm for computing the exact lower bound $\underline{\sigma^2}$ of the finite population variance. Specifically, our algorithm is *quadratic-time*, i.e., it requires $O(n^2)$ computational steps (arithmetic operations or comparisons) for n interval data points $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.

The algorithm $\underline{\mathcal{A}}$ is as follows:

- First, we sort all $2n$ values $\underline{x}_i$, $\overline{x}_i$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}$.
- Second, we compute $\underline{\mu}$ and $\overline{\mu}$ and select all "small intervals" $[x_{(k)}, x_{(k+1)}]$ that intersect with $[\underline{\mu}, \overline{\mu}]$.
- For each of the selected small intervals $[x_{(k)}, x_{(k+1)}]$, we compute the ratio $r_k = S_k/N_k$, where

  $$S_k \stackrel{\text{def}}{=} \sum_{i:\underline{x}_i \ge x_{(k+1)}} \underline{x}_i + \sum_{j:\overline{x}_j \le x_{(k)}} \overline{x}_j,$$

  and N_k is the total number of such i's and j's. If $r_k \in [x_{(k)}, x_{(k+1)}]$, then we compute

  $$\sigma'^2_k \stackrel{\text{def}}{=} \frac{1}{n} \cdot \left(\sum_{i:\underline{x}_i \ge x_{(k+1)}} (\underline{x}_i - r_k)^2 + \sum_{j:\overline{x}_j \le x_{(k)}} (\overline{x}_j - r_k)^2 \right).$$

  If $N_k = 0$, we take $\sigma'^2_k \stackrel{\text{def}}{=} 0$.
- Finally, we return the smallest of the values σ'^2_k as $\underline{\sigma^2}$.

Theorem 2.1. *The algorithm $\underline{\mathcal{A}}$ always compute $\underline{\sigma^2}$ is quadratic time.*

(For readers' convenience, all the proofs are placed in the special Proofs section).

We have implemented this algorithm in C++, it works really fast.

Example. We start with 5 intervals: $\mathbf{x}_1 = [2.1, 2.6]$, $\mathbf{x}_2 = [2.0, 2.1]$, $\mathbf{x}_3 = [2.2, 2.9]$, $\mathbf{x}_4 = [2.5, 2.7]$, and $\mathbf{x}_5 = [2.4, 2.8]$. After sorting the bounds, we get the following "small intervals": $[x_{(1)}, x_{(2)}] = [2.0, 2.1]$, $[x_{(2)}, x_{(3)}] = [2.1, 2.1]$, $[x_{(3)}, x_{(4)}] = [2.1, 2.2]$, $[x_{(4)}, x_{(5)}] = [2.2, 2.4]$, $[x_{(5)}, x_{(6)}] = [2.4, 2.5]$,

$[x_{(6)}, x_{(7)}] = [2.5, 2.6]$, $[x_{(7)}, x_{(8)}] = [2.6, 2.7]$, $[x_{(8)}, x_{(9)}] = [2.7, 2.8]$, and $[x_{(9)}, x_{(10)}] = [2.8, 2.9]$.

The interval for finite population average is $\boldsymbol{\mu} = [2.24, 2.62]$, so we only keep the following four small intervals that have non-empty intersection with $\mathbf{E}$: $[x_{(4)}, x_{(5)}] = [2.2, 2.4]$, $[x_{(5)}, x_{(6)}] = [2.4, 2.5]$, $[x_{(6)}, x_{(7)}] = [2.5, 2.6]$, and $[x_{(7)}, x_{(8)}] = [2.6, 2.7]$. For these intervals:

- $S_4 = 7.0$, $N_4 = 3$, so $r_4 = 2.333\ldots$;
- $S_5 = 4.6$, $N_5 = 2$, so $r_5 = 2.3$;
- $S_6 = 2.1$, $N_6 = 1$, so $r_6 = 2.1$;
- $S_7 = 4.7$, $N_7 = 2$, so $r_7 = 2.35$.

Of the four values r_k, only r_4 lies within the corresponding small interval. For this small interval, ${\sigma'}_4^2 = 0.017333\ldots$, so $\underline{\sigma^2} = 0.017333\ldots$

3 Second Result: Computing $\overline{\sigma^2}$ is NP-Hard

Our second result is that the general problem of computing $\overline{\sigma^2}$ from given intervals $\mathbf{x}_i$ is computationally difficult, or, in precise terms, NP-hard (for exact definitions of NP-hardness, see, e.g., [5, 8, 13]).

Theorem 3.1. *Computing $\overline{\sigma^2}$ is NP-hard.*

Comment. This result was first announced in [3].

The very fact that computing the range of a quadratic function is NP-hard was first proven by Vavasis [15] (see also [8]). We have shown that this difficulty happens even for the very simple quadratic functions (1.1) frequently used in data processing.

A natural question is: maybe the difficulty comes from the requirement that the range be computed exactly? In practice, it is often sufficient to compute, in a reasonable amount of time, a usefully accurate estimate $\widetilde{\overline{\sigma^2}}$ for $\overline{\sigma^2}$, i.e., an estimate $\widetilde{\overline{\sigma^2}}$ which is accurate with a given accuracy $\varepsilon > 0$: $\left|\widetilde{\overline{\sigma^2}} - \overline{\sigma^2}\right| \le \varepsilon$. Alas, for any ε, such computations are also NP-hard:

Theorem 3.2. *For every $\varepsilon > 0$, the problem of computing $\overline{\sigma^2}$ with accuracy ε is NP-hard.*

It is worth mentioning that $\overline{\sigma^2}$ can be computed exactly in exponential time $O(2^n)$:

Theorem 3.3. *There exists an algorithm that computes $\overline{\sigma^2}$ in exponential time.*

4 Third Result: A Feasible Algorithm That Computes $\overline{\sigma^2}$ in Many Practical situations

NP-hard means, crudely speaking, that there are no general ways for solving all particular cases of this problem (i.e., computing $\overline{\sigma^2}$) in reasonable time.

However, we show that there are algorithms for computing $\overline{\sigma^2}$ for many reasonable situations. Namely, we propose an efficient algorithm $\mathcal{A}$ that computes $\overline{\sigma^2}$ for the case when all the interval midpoints ("measured values") $\widetilde{x}_i = (\underline{x}_i + \overline{x}_i)/2$ are definitely different from each other, in the sense that the "narrowed" intervals $[\widetilde{x}_i - \Delta_i/n, \widetilde{x}_i + \Delta_i/n]$ – where $\Delta_i = (\underline{x}_i - \overline{x}_i)/2$ is the interval's half-width – do not intersect with each other.

This algorithm $\overline{\mathcal{A}}$ is as follows:

- First, we sort all $2n$ endpoints of the narrowed intervals $\widetilde{x}_i - \Delta_i/n$ and $\widetilde{x}_i + \Delta_i/n$ into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(2n)}$. This enables us to divide the real line into $2n + 2$ segments ("small intervals") $[x_{(k)}, x_{(k+1)}]$, where we denoted $x_{(0)} \stackrel{\text{def}}{=} -\infty$ and $x_{(2n+1)} \stackrel{\text{def}}{=} +\infty$.
- Second, we compute $\underline{\mu}$ and $\overline{\mu}$ and pick all "small intervals" $[x_{(k)}, x_{(k+1)}]$ that intersect with $[\underline{\mu}, \overline{\mu}]$.
- For each of remaining small intervals $[x_{(k)}, x_{(k+1)}]$, for each i from 1 to n, we pick the following value of x_i:
  - if $x_{(k+1)} < \widetilde{x}_i - \Delta_i/n$, then we pick $x_i = \overline{x}_i$;
  - if $x_{(k)} > \widetilde{x}_i + \Delta_i/n$, then we pick $x_i = \underline{x}_i$;
  - for all other i, we consider both possible values $x_i = \overline{x}_i$ and $x_i = \underline{x}_i$.

  As a result, we get one or several sequences of x_i. For each of these sequences, we check whether the average μ of the selected values $x_1, \ldots, x_n$ is indeed within this small interval, and if it is, compute the finite population variance by using the formula (1.1).
- Finally, we return the largest of the computed finite population variances as $\overline{\sigma^2}$.

Theorem 4.1. *The algorithm $\overline{\mathcal{A}}$ computes $\overline{\sigma^2}$ in quadratic time for all the cases in which the "narrowed" intervals do not intersect with each other.*

This algorithm also works when, for some fixed k, no more than k "narrowed" intervals can have a common point:

Theorem 4.2. *For every positive integer k, the algorithm $\overline{\mathcal{A}}$ computes $\overline{\sigma^2}$ in quadratic time for all the cases in which no more than k "narrowed" intervals can have a common point.*

This computation time is quadratic in n but it grows exponentially with k. So, when k grows, this algorithm requires more and more computation time;

as we will see from the proof, it requires $O(2^k \cdot n^2)$ steps. In the worst case, when our conditions are not satisfied and $k = O(n)$ narrowed intervals have a common point, this algorithm requires $O(2^n \cdot n^2)$ computational steps.

It is worth mentioning that the examples on which we prove NP-hardness (see proof of Theorem 3.1) correspond to the case when all n narrowed intervals have a common point.

5 Finite Population Mean, Finite Population Variance: What Next?

In the previous sections, we described conditions under which efficient ($O(n^2)$) algorithms exist for computing $\min(f(x))$ and $\max(f(x))$ for the finite population variance $f = \sigma^2$.

Average and variance are not the only population parameters used in data processing. A natural question is: when are efficient algorithms possible for other population parameters used in data processing?

5.1 Finite Population Covariance

When we have two sets of data $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, we normally compute *finite population covariance*

$$C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) \cdot (y_i - \mu_y),$$

where

$$\mu_x = \frac{1}{n} \sum_{i=1}^{n} x_i; \quad \mu_y = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

Finite population covariance is used to describe the correlation between x_i and y_i. If we take interval uncertainty into consideration, then, after each measurement, we do not get the exact values of $x_1, \ldots, x_n$, $y_1, \ldots, y_n$; instead, we only have *intervals* $[\underline{x}_1, \overline{x}_1], \ldots, [\underline{x}_n, \overline{x}_n]$, $[\underline{y}_1, \overline{y}_1], \ldots, [\underline{y}_n, \overline{y}_n]$. Depending on what are the actual values of $x_1, \ldots, x_n$, $y_1, \ldots, y_n$ within these intervals, we get different values of finite population covariance. To take the interval uncertainty into consideration, we need to be able to describe the interval $[\underline{C}, \overline{C}]$ of possible values of the finite population covariance C.

So, we arrive at the following problems: given the intervals $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$, compute the lower and upper bounds $\underline{C}$ and $\overline{C}$ for the interval of possible values of finite population covariance.

It turns out that these problems are also NP-hard:

Theorem 5.1. *The problem of computing $\overline{C}$ from the interval inputs $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$ is NP-hard.*

Theorem 5.2. *The problem of computing $\underline{C}$ from the interval inputs $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$ is NP-hard.*

Comment. These results were first announced in [12].

5.2 Finite Population Correlation

As we have mentioned, finite population covariance C between the data sets $x_1, \dots, x_n$ and $y_1, \dots, y_n$ is often used to compute finite population correlation

$$\rho = \frac{C}{\sigma_x \cdot \sigma_y}, \tag{5.1}$$

where $\sigma_x = \sqrt{\sigma_x^2}$ is the finite population standard deviation of the values $x_1, \dots, x_n$, and $\sigma_y = \sqrt{\sigma_y^2}$ is the finite population standard deviation of the values $y_1, \dots, y_n$.

When we only have *intervals* $[\underline{x}_1, \overline{x}_1], \dots, [\underline{x}_n, \overline{x}_n]$, $[\underline{y}_1, \overline{y}_1], \dots, [\underline{y}_n, \overline{y}_n]$, we have an interval $[\underline{\rho}, \overline{\rho}]$ of possible value of correlation. It turns out that, similar to finite population covariance, computation of the endpoints of this interval problems is also an NP-hard problem:

Theorem 5.3. *The problem of computing $\overline{\rho}$ from the interval inputs $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$ is NP-hard.*

Theorem 5.4. *The problem of computing $\underline{\rho}$ from the interval inputs $[\underline{x}_i, \overline{x}_i]$, $[\underline{y}_i, \overline{y}_i]$ is NP-hard.*

Comment. The fact that the problems of computing finite population covariance and finite population correlation are NP-hard means that, crudely speaking, that there is no feasible algorithm that would always compute the desired bounds for C and ρ. A similar NP-hardness result holds for finite population variance, but in that case, we were also able to produce a feasible algorithm that works in many practical cases. It is desirable to design similar algorithms for finite population covariance and finite population correlation.

5.3 Finite Population Median

Not all finite population parameters used in data processing are difficult to compute for interval data, some are easy. In addition to finite population mean, we can mention finite population *median.* Since the median is increasing in $x_1, \dots, x_n$, its smallest possible value is attained for $\underline{x}_1, \dots, \underline{x}_n$, and its largest possible value is attained for $\overline{x}_1, \dots, \overline{x}_n$.

So, to compute the exact bounds for the median, it is sufficient to apply the algorithm for computing the finite population median of n numbers twice:

- first, to the values $\underline{x}_1, \dots, \underline{x}_n$, to compute the lower endpoint for the finite population median;
- second, to the values $\overline{x}_1, \dots, \overline{x}_n$, to compute the upper endpoint for the finite population median.

To compute each median, we can sort the corresponding n values. It is known that one can sort n numbers in $O(n \cdot \log(n))$ steps; see, e.g., [1]. So, the above algorithm requires $O(n \cdot \log(n))$ steps – and is, therefore, quite feasible.

5.4 Other Population Parameters: Open Problem

In the previous sections, we described conditions under which efficient ($O(n^2)$) algorithms exist for computing $\min(f(x))$ and $\max(f(x))$ for the finite population variance $f = \sigma^2$. In this section, we analyzed the possibility of exactly computing a few more finite population characteristics under interval uncertainty.

The results from this section are mostly negative: that for the population parameters that we analyzed, in general, efficient algorithms for exactly computing the bounds are not possible. Since we cannot have efficient algorithms that work for all possible cases, it is desirable to find out under what conditions such efficient algorithms are possible.

It is desirable to analyze other finite population parameters from this viewpoint.

6 Proofs

Proof of Theorem 2.1

1°. Let us first show that the algorithm described in Section 2 is indeed correct.

1.1°. Indeed, let $x_1^{(0)} \in \mathbf{x}_1, \ldots, x_n^{(0)} \in \mathbf{x}_n$ be the values for which the finite population variance σ^2 attains minimum on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$.

Let us pick one of the n variables x_i, and let fix the values of all the other variables x_j ($j \neq i$) at $x_j = x_j^{(0)}$. When we substitute $x_j = x_j^{(0)}$ for all $j \neq i$ into the expression for finite population variance, σ^2 becomes a quadratic function of x_i.

This function of one variable should attain its minimum on the interval $\mathbf{x}_i$ at the value $x_i^{(0)}$.

1.2°. Let us start with the analysis of the quadratic function of one variable we described in Part 1.1 of this proof.

By definition, the finite population variance σ^2 is a sum of non-negative terms; thus, its value is always non-negative. Therefore, the corresponding quadratic function of one variable always has a global minimum. This function is decreasing before this global minimum, and increasing after it.

1.3°. Where is the global minimum of the quadratic function of one variable described in Part 1.1?

It is attained when $\partial(\sigma^2)/\partial x_i = 0$. Differentiating the formula (1.1) with respect to x_i, we conclude that

$$\frac{\partial(\sigma^2)}{\partial x_i} = \frac{1}{n} \cdot \left(2(x_i - \mu) + \sum_{j=1}^{n} 2(\mu - x_j) \cdot \frac{\partial \mu}{\partial x_j} \right). \tag{6.1}$$

Since $\partial\mu/\partial x_i = 1/n$, we conclude that

$$\frac{\partial(\sigma^2)}{\partial x_i} = \frac{2}{n} \cdot \left((x_i - \mu) + \sum_{j=1}^{n} (\mu - x_j) \cdot \frac{1}{n} \right). \tag{6.2}$$

Here,

$$\sum_{j=1}^{n} (\mu - x_j) = n \cdot \mu - \sum_{j=1}^{n} x_j. \tag{6.3}$$

By definition of the average μ, this difference is 0, hence the formula (6.2) takes the form

$$\frac{\partial(\sigma^2)}{\partial x_i} = \frac{2}{n} \cdot (x_i - \mu).$$

So, this function attains the minimum when $x_i - \mu = 0$, i.e., when $x_i = \mu$.

Since

$$\mu = \frac{x_i}{n} + \frac{\sum_i' x_j}{n},$$

where $\sum_i'$ means the sum over all $j \neq i$, the equality $x_i = \mu$ means that

$$x_i = \frac{x_i}{n} + \frac{\sum_i' x_j^{(0)}}{n}.$$

Moving terms containing x_i into the left-hand side and dividing by the coefficient at x_i, we conclude that the minimum is attained when

$$x_i = \mu_i' \stackrel{\text{def}}{=} \frac{\sum_i' x_j^{(0)}}{n-1},$$

i.e., when x_i is equal to the arithmetic average μ_i' of all other elements.

1.4°. Let us now use the knowledge of a global minimum to describe where the desired function attains its minimum on the interval $\mathbf{x}_i$.

In our general description of non-negative quadratic functions of one variable, we mentioned that each such function is decreasing before the global minimum and increasing after it. Thus, for $x_i < \mu_i'$, the function σ^2 is decreasing, for $x_i > \mu_i'$, this function in increasing. Therefore:

- If $\mu_i' \in \mathbf{x}_i$, the global minimum of the function σ^2 of one variable is attained within the interval $\mathbf{x}_i$, hence the minimum on the interval $\mathbf{x}_i$ is attained for $x_i = \mu_i'$.

- If $\mu_i' < \underline{x}_i$, the function σ^2 is increasing on the interval $\mathbf{x}_i$ and therefore, its minimum on this interval is attained when $x_i = \underline{x}_i$.
- Finally, if $\mu_i' > \overline{x}_i$, the function σ^2 is decreasing on the interval $\mathbf{x}_i$ and therefore, its minimum on this interval is attained when $x_i = \overline{x}_i$.

1.5°. Let us reformulate the above conditions in terms of the average

$$\mu = \frac{1}{n} \cdot x_i + \frac{n-1}{n} \cdot \mu_i'.$$

- In the first case, when $x_i = \mu_i'$, we have $x_i = \mu = \mu_i'$, so $\mu \in \mathbf{x}_i$.
- In the second case, we have $\mu_i' < \underline{x}_i$ and $x_i = \underline{x}_i$. Therefore, in this case, $\mu < \underline{x}_i$.
- In the third case, we have $\mu_i' > \overline{x}_i$ and $x_i = \overline{x}_i$. Therefore, in this case, $\mu > \overline{x}_i$.

Thus:

- If $\mu \in \mathbf{x}_i$, then we cannot be in the second or third cases. Thus, we are in the first case, hence $x_i = \mu$.
- If $\mu < \underline{x}_i$, then we cannot be in the first or the third cases. Thus, we are the second case, hence $x_i = \underline{x}_i$.
- If $\mu > \overline{x}_i$, then we cannot be in the first or the second cases. Thus, we are in the third case, hence $x_i = \overline{x}_i$.

1.6°. So, as soon as we determine the position of μ with respect to all the bounds $\underline{x}_i$ and $\overline{x}_i$, we will have a pretty good understanding of all the values x_i at which the minimum is attained.

Hence, to find the minimum, we will analyze how the endpoints $\underline{x}_i$ and $\overline{x}_i$ divide the real line, and consider all the resulting sub-intervals.

Let the corresponding subinterval $[x_{(k)}, x_{(k+1)}]$ by fixed. For the i's for which $\mu \notin \mathbf{x}_i$, the values x_i that correspond to the minimal finite population variance are uniquely determined by the above formulas.

For the i's for which $\mu \in \mathbf{x}_i$ the selected value x_i should be equal to μ. To determine this μ, we can use the fact that μ is equal to the average of all thus selected values x_i, in other words, that we should have

$$\mu = \frac{1}{n} \cdot \left(\sum_{i:\underline{x}_i \geq x_{(k+1)}} \underline{x}_i + (n - N_k) \cdot \mu + \sum_{j:\overline{x}_j \leq x_{(k)}} \overline{x}_j \right), \tag{6.4}$$

where $(n - N_k) \cdot \mu$ combines all the points for which $\mu \in \mathbf{x}_i$. Multiplying both sides of (6.4) by n and subtracting $n \cdot \mu$ from both sides, we conclude that (in notations of Section 2), we have $\mu = S_k/N_k$ – what we denoted, in the algorithm's description, by r_k. If thus defined r_k does not belong to the subinterval $[x_{(k)}, x_{(k+1)}]$, this contradiction with our initial assumption shows that there cannot be any minimum in this subinterval, so this subinterval can be easily dismissed.

The corresponding finite population variance is denoted by ${\sigma'}^2_k$. If $N_k = 0$, this means that μ belongs to all the intervals $\mathbf{x}_i$ and therefore, that the lower endpoint $\underline{\sigma^2}$ is exactly 0 – so we assign ${\sigma'}^2_k = 0$.

2°. To complete the proof of Theorem 2.1, we must show that this algorithm indeed requires quadratic time.

Indeed, sorting requires $O(n \cdot \log(n))$ steps (see, e.g., [1]), and the rest of the algorithm requires linear time ($O(n)$) for each of $2n$ subintervals, i.e., the total quadratic time.

The theorem is proven.

Proof of Theorem 3.1

1°. By definition, a problem is NP-hard if any problem from the class NP can be reduced to it. Therefore, to prove that a problem $\mathcal{P}$ is NP-hard, it is sufficient to reduce one of the known NP-hard problems $\mathcal{P}_0$ to $\mathcal{P}$.

In this case, since $\mathcal{P}_0$ is known to be NP-hard, this means that every problem from the class NP can be reduced to $\mathcal{P}_0$, and since $\mathcal{P}_0$ can be reduced to $\mathcal{P}$, thus, the original problem from the class NP is reducible to $\mathcal{P}$.

For our proof, as the known NP-hard problem $\mathcal{P}_0$, we take a *subset* problem: given n positive integers $s_1, \ldots, s_n$, to check whether there exist signs $\eta_i \in \{-1, +1\}$ for which the signed sum $\sum\limits_{i=1}^{n} \eta_i \cdot s_i$ equals 0.

We will show that this problem can be reduced to the problem of computing $\overline{\sigma^2}$, i.e., that to every instance $(s_1, \ldots, s_n)$ of the problem $\mathcal{P}_0$, we can put into correspondence such an instance of the $\overline{C}$-computing problem that based on its solution, we can easily check whether the desired signs exist.

As this instance, we take the instance corresponding to the intervals $[\underline{x}_i, \overline{x}_i] = [-s_i, s_i]$. We want to show that for the corresponding problem, $\overline{\sigma^2} = C_0$, where we denoted

$$C_0 \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2, \tag{6.5}$$

if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

2°. Let us first show that in all cases, $\overline{\sigma^2} \le C_0$.

Indeed, it is known that the formula for the finite population variance can be reformulated in the following equivalent form:

$$\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i^2 - \mu^2. \tag{6.6}$$

Since $x_i \in [-s_i, s_i]$, we can conclude that $x_i^2 \le s_i^2$ hence $\sum x_i^2 \le \sum s_i^2$. Since $\mu^2 \ge 0$, we thus conclude that

$$\sigma^2 \le \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 = C_0.$$

In other words, every possible value σ^2 of the finite population variance is smaller than or equal to C_0. Thus, the largest of these possible values, i.e., $\overline{\sigma^2}$, also cannot exceed C_0, i.e., $\overline{\sigma^2} \le C_0$.

3°. Let us now prove that if the desired signs η_i exist, then $\overline{\sigma^2} = C_0$.

Indeed, in this case, for $x_i = \eta_i \cdot s_i$, we have $\mu_x = 0$ and $x_i^2 = s_i^2$, hence

$$\sigma^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu_x)^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 = C_0.$$

So, the finite population variance σ^2 is always $\le C_0$, and it attains the value C_0 for some x_i. Therefore, $\overline{\sigma^2} = C_0$.

4°. To complete the proof of Theorem 3.1, we must show that, vice versa, if $\overline{\sigma^2} = C_0$, then the desired signs exist.

Indeed, let $\overline{\sigma^2} = C_0$. Finite population variance is a continuous function on a compact set $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$, hence its maximum on this compact set is attained for some values $x_1 \in \mathbf{x}_1 = [-s_1, s_1], \ldots, x_n \in \mathbf{x}_n = [-s_n, s_n]$. In other words, for the corresponding values of x_i, the finite population variance σ^2 is equal to C_0.

Since $x_i \in [-s_i, s_i]$, we can conclude that $x_i^2 \le s_i^2$; since $(\mu_x)^2 \ge 0$, we get $\sigma^2 \le C_0$. If $|x_i|^2 < s_i^2$ or $(\mu_x)^2 > 0$, then we would have $\sigma^2 < C_0$. Thus, the only way to have $\sigma^2 = C_0$ is to have $x_i^2 = s_i^2$ and $\mu_x = 0$. The first equality leads to $x_i = \pm s_i$, i.e., to $x_i = \eta_i \cdot s_i$ for some $\eta_i \in \{-1, +1\}$. Since μ_x is, by definition, the (arithmetic) average of the values x_i, the equality $\mu_x = 0$ then leads to $\sum\limits_{i=1}^{n} \eta_i \cdot s_i = 0$. So, if $\overline{\sigma^2} = C_0$, then the desired signs do exist.

The theorem is proven.

Proof of Theorem 3.2

1°. Let $\varepsilon > 0$ be fixed. We will show that the subset problem can be reduced to the problem of computing $\overline{\sigma^2}$ with accuracy ε, i.e., that to every instance $(s_1, \ldots, s_n)$ of the subset problem $\mathcal{P}_0$, we can put into correspondence such an instance of the ε-approximate $\overline{C}$-computation problem that based on its solution, we can easily check whether the desired signs exist.

For this reduction, we will use two parameters. The first one – C_0 – is the same as in the proof of Theorem 3.1. We will also need a new real-valued parameter k; its value depend on ε and n. We could produce this value right away, but we believe that the proof will be much clearer if we keep it undetermined until it becomes clear what value k we need to choose for the proof to be valid.

As the desired instance, we take the instance corresponding to the intervals $[\underline{x}_i, \overline{x}_i] = [-k \cdot s_i, k \cdot s_i]$ for an appropriate value k. Let $\widetilde{\overline{\sigma^2}}$ be a number produced, for this problem, by a ε-accurate computation algorithm, i.e., a number for which $\left|\widetilde{\overline{\sigma^2}} - \overline{\sigma^2}\right| \le \varepsilon$. We want to to show that $\widetilde{\overline{\sigma^2}} \ge k^2 \cdot C_0 - \varepsilon$ if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

2°. When we multiply each value x_i by a constant k, the finite population variance is multiplied by k^2. As a result, the upper bound $\overline{\sigma^2}$ corresponding to $x_i \in [-k \cdot s_i, k \cdot s_i]$ is exactly k^2 times larger than the upper bound $\overline{v}$ corresponding to k times smaller values $z_i \in [-s_i, s_i]$: $\overline{v} = \overline{\sigma^2}/k^2$.

Hence, when $\widetilde{\overline{\sigma^2}}$ approximates $\overline{\sigma^2}$ with an accuracy ε, the corresponding value $\widetilde{\overline{v}} \stackrel{\text{def}}{=} \widetilde{\overline{\sigma^2}}/k^2$ approximates $\overline{v}$ $(= \overline{\sigma^2}/k^2)$ with the accuracy $\delta \stackrel{\text{def}}{=} \varepsilon/k^2$.

In terms of $\widetilde{\overline{v}}$, the above inequality $\widetilde{\overline{\sigma^2}} \ge k^2 \cdot C_0 - \varepsilon$ takes the following equivalent form: $\widetilde{\overline{v}} \ge C_0 - \delta$.

Thus, in terms of $\widetilde{\overline{v}}$, the desired property can be formulated as follows: $\widetilde{\overline{v}} \ge C_0 - \delta$ if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

3°. Let us first show that if the desired signs η_i exist, then $\widetilde{\overline{v}} \ge C_0 - \delta$.

Indeed, in this case, similarly to the proof of Theorem 3.1, we can conclude that $\overline{v} = C_0$. Since $\widetilde{\overline{v}}$ is a δ-approximation to the actual upper bound $\overline{v}$, we can therefore conclude that $\widetilde{\overline{v}} \ge \overline{v} - \delta = C_0 - \delta$. The statement is proven.

4°. Vice versa, let us assume that $\widetilde{\overline{v}} \ge C_0 - \delta$. Let us prove that in this case, the desired signs exist.

4.1°. Since $\widetilde{\overline{v}}$ is a δ-approximation to the upper bound $\overline{v}$, we thus conclude that $\overline{v} \ge \widetilde{\overline{v}} - \delta$ and therefore, $\overline{v} \ge C_0 - 2\delta$.

Similarly to the proof of Theorem 3.1, we can conclude that the maximum is attained for some values $z_i \in [-s_i, s_i]$ and therefore, there exist values $z_i \in [-s_i, s_i]$ for which the finite population variance v exceeds $C_0 - 2\delta$:

$$v \stackrel{\text{def}}{=} \frac{1}{n} \cdot \sum_{i=1}^{n} z_i^2 - (\mu_z)^2 \ge C_0 - 2\delta,$$

i.e., substituting the expression (6.5) for C_0, that

$$\frac{1}{n} \cdot \sum_{i=1}^{n} z_i^2 - (\mu_z)^2 \ge \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 - 2\delta. \tag{6.7}$$

4.2°. The following proof will be similar to the corresponding part of the proof of Theorem 3.1. The main difference is that we have approximate equalities instead of exact ones:

- In the proof of Theorem 3.1, we used the fact that $\sigma^2 = C_0$ to prove that the corresponding values x_i are equal to $\pm s_i$, and that their sum is equal to 0.

- Here, v is only approximately equal to C_0. As a result, we will only be able to show that the values z_i are *close* to $\pm s_i$, and that the sum of z_i is *close* to 0. From these closenesses, we will then be able to conclude (for sufficiently large k) that the sum of the corresponding terms $\pm s_i$ is exactly equal to 0.

4.3°. Let us first prove that for every i, the value z_i^2 is close to s_i^2. Specifically, we know that $z_i^2 \le s_i^2$; we will prove that

$$z_i^2 \ge s_i^2 - 2(n-1) \cdot \delta. \tag{6.8}$$

We will prove this inequality by reduction to a contradiction. Indeed, let us assume that for some i_0, this inequality is not true. This means that

$$z_{i_0}^2 < s_{i_0}^2 - 2(n-1) \cdot \delta. \tag{6.9}$$

Since $z_i \in [-s_i, s_i]$, for all i, in particular, for all $i \ne i_0$, we conclude, for all $i \ne i_0$, that

$$z_i^2 \le s_i^2. \tag{6.10}$$

Adding the inequality (6.9) and $(n-1)$ inequalities (6.10) corresponding to all values $i \ne i_0$, we get

$$\sum_{i=1}^n z_i^2 < \sum_{i=1}^n s_i^2 - 2(n-1) \cdot \delta. \tag{6.11}$$

Dividing both sides of this inequality by $n-1$, we get a contradiction with (6.7). This contradiction shows that (6.8) indeed holds for every i.

4.4°. The inequality (6.8) says, crudely speaking, that z_i^2 is close to s_i^2. According to our "action plan" (as outlined in Part 4.2 of this proof), we want to conclude that z_i is close to $\pm s_i$, i.e., that $|z_i|$ is close to s_i.

To be able to make a meaningful conclusion about z_i from the inequality (6.8), we must make sure that the right-hand side of the inequality (6.8) is positive: otherwise, this inequality is true simply because its left-hand side is non-negative, and the right-hand side is non-positive.

The value s_i is a positive integer, so $s_i^2 \ge 1$. Therefore, to guarantee that the right-hand side of (6.8) is positive, it is sufficient to select k for which, for the corresponding value $\delta = \varepsilon/k^2$, we have

$$2(n-1) \cdot \delta < 1. \tag{6.12}$$

In the following text, we will assume that this condition is indeed satisfied.

4.5°. Let us show that under the condition (6.12), the value $|z_i|$ is indeed close to s_i. To be more precise, we already know that $|z_i| \le s_i$; we are going to prove that

$$|z_i| \ge s_i - 2(n-1) \cdot \delta. \tag{6.13}$$

Indeed, since the right-hand side of the inequality (6.8) is supposed to be close to s_i, it makes sense to represent it as s_i^2 times a factor close to 1. To be more precise, we reformulate the inequality (6.8) in the following equivalent form:

$$z_i^2 \geq s_i^2 \cdot \left(1 - \frac{2(n-1)\cdot\delta}{s_i^2}\right). \tag{6.14}$$

Since both sides of this inequality are non-negative, we can extract the square root from both sides and get the following inequality:

$$|z_i| \geq s_i \cdot \sqrt{1 - \frac{2(n-1)\cdot\delta}{s_i^2}}. \tag{6.15}$$

The square root in the right-hand side of (6.15) is of the type $\sqrt{1-t}$, with $0 \leq t \leq 1$. It is known that for such t, we have $\sqrt{1-t} \geq 1-t$. Therefore, from (6.15), we can conclude that

$$|z_i| \geq s_i \cdot \sqrt{1 - \frac{2(n-1)\cdot\delta}{s_i^2}} \geq s_i \cdot \left(1 - \frac{2(n-1)\cdot\delta}{s_i^2}\right),$$

i.e., that

$$|z_i| \geq s_i - \frac{2(n-1)\cdot\delta}{s_i}.$$

Since $s_i \geq 1$, we have

$$\frac{2(n-1)\cdot\delta}{s_i} \leq 2(n-1)\cdot\delta,$$

hence

$$|z_i| \geq s_i - \frac{2(n-1)\cdot\delta}{s_i} \geq s_i - 2(n-1)\cdot\delta.$$

So, the inequality (6.13) is proven.

4.6°. Let us now prove that for the values z_i selected on Step 4.1, the average μ_z is close to 0. To be more precise, we will prove that

$$(\mu_z)^2 \leq 2\delta. \tag{6.16}$$

Similarly to Part 4.3 of this proof, we will prove this inequality by reduction to a contradiction. Indeed, assume that this inequality is not true, i.e., that

$$(\mu_z)^2 > 2\delta. \tag{6.17}$$

Since $z_i^2 \leq s_i^2$, we therefore conclude that

$$\sum_{i=1}^n z_i^2 \leq \sum_{i=1}^n s_i^2,$$

hence

$$\frac{1}{n} \cdot \sum_{i=1}^{n} z_i^2 \le \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2. \tag{6.18}$$

Adding, to both sides of the inequality (6.18), the inequality (6.17), we get an inequality

$$\frac{1}{n} \cdot \sum_{i=1}^{n} z_i^2 - (\mu_z)^2 < \frac{1}{n} \sum_{i=1}^{n} s_i^2 - 2\delta,$$

which contradicts to (6.7). This contradiction proves that that the inequality (6.16) is true.

4.7°. From the fact that the average μ_z is close to 0, we can now conclude that the sum $\sum z_i$ is also close to 0. Specifically, we will now prove that

$$\left| \sum_{i=1}^{n} z_i \right| \le n \cdot \sqrt{2\delta}. \tag{6.19}$$

Indeed, from (6.16), we conclude that $(\mu_z)^2 \le 2\delta$, hence $|\mu_z| \le \sqrt{2\delta}$. Multiplying both sides of this inequality by n, we get the desired inequality (6.19).

4.8°. Let us now show that for appropriately chosen k, we will be able to conclude that there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

From the inequalities (6.13) and $|z_i| \le s_i$, we conclude that

$$|s_i - |z_i|| \le 2(n-1) \cdot \delta. \tag{6.20}$$

Hence, $|z_i| \le s_i - 2(n-1) \cdot \delta$. Each value s_i is a positive integer, so $s_i \ge 1$. Due to the inequality (6.12), we have $2(n-1) \cdot \delta < 1$, so $|z_i| > 1 - 1 = 0$. Therefore, $z_i \ne 0$, hence each value z_i has a sign. Let us take, as η_i, the sign of the value z_i. Then, the inequality (6.20) takes the form

$$|\eta_i \cdot s_i - z_i| \le 2(n-1) \cdot \delta. \tag{6.21}$$

Since the absolute value of the sum cannot exceed the sum of absolute values, we therefore conclude that

$$\left| \sum_{i=1}^{n} \eta_i \cdot s_i - \sum_{i=1}^{n} z_i \right| = \left| \sum_{i=1}^{n} (\eta_i \cdot s_i - z_i) \right| \le \sum_{i=1}^{n} |\eta_i \cdot s_i - z_i| \le$$

$$\sum_{i=1}^{n} 2(n-1) \cdot \delta = 2n \cdot (n-1) \cdot \delta. \tag{6.22}$$

From (6.22) and (6.19), we conclude that

$$\left| \sum_{i=1}^{n} \eta_i \cdot s_i \right| \le \left| \sum_{i=1}^{n} z_i \right| + \left| \sum_{i=1}^{n} \eta_i \cdot s_i - \sum_{i=1}^{n} z_i \right| = n \cdot \sqrt{2\delta} + 2n \cdot (n-1) \cdot \delta. \tag{6.23}$$

All values s_i are integers, hence, the sum $\sum \eta_i \cdot s_i$ is also an integer, and so is its absolute value $|\sum \eta_i \cdot s_i|$. Thus, if we select k for which the right-hand side of the inequality (6.23) is less than 1, i.e., for which

$$n \cdot \sqrt{2\delta} + 2n \cdot (n-1) \cdot \delta < 1, \tag{6.24}$$

we therefore conclude that the absolute value of an integer $\sum \eta_i \cdot s_i$ is smaller than 1, so it must be equal to 0: $\sum \eta_i \cdot s_i = 0$.

Thus, to complete the proof, it is sufficient to find k for which, for the corresponding value $\delta = \varepsilon/k^2$, both the inequalities (6.12) and (6.24) hold. To guarantee the inequality (6.24), it is sufficient to have

$$n \cdot \sqrt{2\delta} \le \frac{1}{3} \tag{6.25}$$

and

$$2n \cdot (n-1) \cdot \delta \le \frac{1}{3}. \tag{6.26}$$

The inequality (6.25) is equivalent to

$$\delta \le \frac{1}{18n^2};$$

the inequality (6.26) is equivalent to

$$\delta \le \frac{1}{6n \cdot (n-1)};$$

and the inequality (6.12) is equivalent to

$$\delta \le \frac{1}{2(n-1)}.$$

Thus, to satisfy all three inequalities, we must choose δ for which $\delta = \varepsilon/k^2 = \delta_0$, where we denoted

$$\delta_0 \stackrel{\text{def}}{=} \min\left(\frac{1}{18n^2}, \frac{1}{6n \cdot (n-1)}, \frac{1}{2(n-1)}\right).$$

The original expression (1.1) for the finite population variance only works for $n \ge 2$. For such n, $18n^2 > 6n \cdot (n-1)$ and $18n^2 > 2(n-1)$, hence the above formula can be simplified into

$$\delta_0 = \frac{1}{18n^2}.$$

To get this δ as $\delta_0 = \varepsilon/k^2$, we must take $k = \sqrt{\varepsilon/\delta_0} = 3n \cdot \sqrt{2\varepsilon}$. For this k, as we have shown before, the reduction holds, so the theorem is proven.

Proof of Theorem 3.3

Let $x_1^{(0)} \in \mathbf{x}_1, \ldots, x_n^{(0)} \in \mathbf{x}_n$ be the values for which the finite population variance σ^2 attains maximum on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$.

Let us pick one of the n variables x_i, and let fix the values of all the other variables x_j ($j \neq i$) at $x_j = x_j^{(0)}$. When we substitute $x_j = x_j^{(0)}$ for all $j \neq i$ into the expression for finite population variance, σ^2 becomes a quadratic function of x_i.

This function of one variable should attain its maximum on the interval $\mathbf{x}_i$ at the value $x_i^{(0)}$.

As we have mentioned in the proof of Theorem 2.1, by definition, the finite population variance σ^2 is a sum of non-negative terms; thus, its value is always non-negative. Therefore, the corresponding quadratic function of one variable always has a global minimum. This function is decreasing before this global minimum, and increasing after it. Thus, its maximum on the interval $\mathbf{x}_i$ is attained at one of the endpoints of this interval.

In other words, for each variable x_i, the maximum is attained either for $x_i = \underline{x}_i$, or for $x_i = \overline{x}_i$. Thus, to find $\overline{\sigma^2}$, it is sufficient to compute σ^2 for 2^n possible combinations $(x_1^{\pm}, \ldots, x_n^{\pm})$, where $x_i^- \stackrel{\text{def}}{=} \underline{x}_i$ and $x_i^+ \stackrel{\text{def}}{=} \overline{x}_i$, and find the largest of the resulting 2^n numbers.

Proof of Theorems 4.1 and 4.2

1°. Similarly to the proof of Theorem 2.1, let us first show that the algorithm described in Section 4 is indeed correct.

2°. Similarly to the proof of Theorem 2.1, let $x_1, \ldots, x_n$ be the values at which the finite population variance attain its maximum on the box $\mathbf{x}_1 \times \ldots \times \mathbf{x}_n$. If we fix the values of all the variables but one x_i, then σ^2 becomes a quadratic function of x_i. When the function σ^2 attains maximum over $x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n$, then this quadratic function of one variable will attain its maximum on the interval $\mathbf{x}_i$ at the point x_i.

We have already shown, in the proof of Theorem 2.1, that this quadratic function has a (global) minimum at $x_i = \mu_i'$, where μ_i' is the average of all the values $x_1, \ldots, x_n$ except for x_i. Since this quadratic function of one variable is always non-negative, it cannot have a global maximum. Therefore, its maximum on the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ is attained at one of the endpoints of this interval.

An arbitrary quadratic function of one variable is symmetric with respect to the location of its global minimum, so its maximum on any interval is attained at the point which is the farthest from the minimum. There is exactly one point which is equally close to both endpoints of the interval $\mathbf{x}_i$: its midpoint $\widetilde{x}_i$. Depending on whether the global minimum is to the left, to the right, or exactly at the midpoint, we get the following three possible cases:

1. If the global minimum μ_i' is to the left of the midpoint $\widetilde{x}_i$, i.e., if $\mu_i' < \widetilde{x}_i$, then the upper endpoint is the farthest from μ_i'. In this case, the maximum of the quadratic function is attained at its upper endpoint, i.e., $x_i = \overline{x}_i$.

2. Similarly, if the global minimum μ'_i is to the right of the midpoint $\widetilde{x}_i$, i.e., if $\mu'_i > \widetilde{x}_i$, then the lower endpoint is the farthest from μ'_i. In this case, the maximum of the quadratic function is attained at its lower endpoint, i.e., $x_i = \underline{x}_i$.

3. If $\mu'_i = \widetilde{x}_i$, then the maximum of σ^2 is attained at both endpoints of the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.

3°. In the third case, we have either $x_i = \underline{x}_i$ or $x_i = \overline{x}_i$. Depending on whether x_i is equal to the lower or to the upper endpoints, we can "combine" the corresponding situations with Cases 1 and 2. As a result, we arrive at the conclusion that one of the following two situations happen:

1. either $\mu'_i \le \widetilde{x}_i$ and $x_i = \overline{x}_i$;

2. either $\mu'_i \ge \widetilde{x}_i$ and $x_i = \underline{x}_i$.

4°. Similarly to the proof of Theorem 2.1, let us reformulate these conclusions in terms of the average μ of the maximizing values $x_1, \ldots, x_n$.

The average μ'_i can be described as

$$\frac{\sum'_i x_j}{n-1},$$

where $\sum'_i$ means the sum over all $j \ne i$. By definition, $\sum'_j x_j = \sum_j x_j - x_i$, where $\sum_j x_j$ means the sum over all possible j. By definition of μ, we have

$$\mu = \frac{\sum_j x_j}{n},$$

hence $\sum_j x_j = n \cdot \mu$. Therefore,

$$\mu'_i = \frac{n \cdot \mu - x_i}{n-1}.$$

Let us apply this formula to the above three cases.

4.1°. In the first case, we have $\widetilde{x}_i \ge \mu'_i$. So, in terms of μ, we get the inequality

$$\widetilde{x}_i \ge \frac{n \cdot \mu - x_i}{n-1}.$$

Multiplying both sides of this inequality by $n-1$, and using the fact that in this case, $x_i = \overline{x}_i = \widetilde{x}_i + \Delta_i$, we conclude that

$$(n-1) \cdot \widetilde{x}_i \ge n \cdot \mu - \widetilde{x}_i - \Delta_i.$$

Moving all the terms but $n \cdot \mu$ to the left-hand side and dividing by n, we get the following inequality:

$$\mu \le \widetilde{x}_i + \frac{\Delta_i}{n}.$$

4.2°. In the second case, we have $\widetilde{x}_i \le \mu'_i$. So, in terms of μ, we get the inequality

$$\widetilde{x}_i \le \frac{n \cdot \mu - x_i}{n-1}.$$

Multiplying both sides of this inequality by $n-1$, and using the fact that in this case, $x_i = \underline{x}_i = \widetilde{x}_i - \Delta_i$, we conclude that

$$(n-1) \cdot \widetilde{x}_i \le n \cdot \mu - \widetilde{x}_i + \Delta_i.$$

Moving all the terms but $n \cdot \mu$ to the left-hand side and dividing by n, we get the following inequality:

$$\mu \ge \widetilde{x}_i - \frac{\Delta_i}{n}.$$

5°. Parts 4.1 and 4.2 of this proof can be summarized as follows:

- In Case 1, we have $\mu \le \widetilde{x}_i + \Delta_i/n$ and $x_i = \overline{x}_i$.
- In Case 2, we have $\mu \ge \widetilde{x}_i - \Delta_i/n$ and $x_i = \underline{x}_i$.

Therefore:

- If $\mu < \widetilde{x}_i - \Delta_i/n$, this means that we cannot be in Case 2. So we must be in Case 1 and therefore, we must have $x_i = \overline{x}_i$.
- If $\mu > \widetilde{x}_i + \Delta_i/n$, this means that we cannot be in Case 1. So, we must be in Case 2 and therefore, we must have $x_i = \underline{x}_i$.

The only case when we do not know which endpoint for x_i we should choose is the case when μ belongs to the narrowed interval $[\widetilde{x}_i - \Delta/n, \widetilde{x}_i + \Delta_i]$.

6°. Hence, once we know where μ is with respect to the endpoints of all narrowed intervals, we can determine the values of all optimal x_i – except for those that are within this narrowed interval. Since we consider the case when no more than k narrowed intervals can have a common point, we have no more than k undecided values x_i. Trying all possible combinations of lower and upper endpoints for these $\le k$ values requires $\le 2^k$ steps.

Thus, the overall number of steps is $O(2^k \cdot n^2)$. Since k is a constant, the overall number of steps is thus $O(n^2)$.

The theorem is proven.

Proof of Theorem 5.1

1°. Similarly to the proof of Theorem 3.1, we reduce a subset problem to the problem of computing $\overline{C}$.

Each instance of the subset problem is as follows: given n positive integers $s_1, \ldots, s_n$, to check whether there exist signs $\eta_i \in \{-1, +1\}$ for which the signed sum $\sum_{i=1}^{n} \eta_i \cdot s_i$ equals 0.

We will show that this problem can be reduced to the problem of computing $\overline{C}$, i.e., that to every instance $(s_1, \ldots, s_n)$ of the subset problem $\mathcal{P}_0$, we can put into correspondence such an instance of the $\overline{C}$-computing problem that based on its solution, we can easily check whether the desired signs exist.

As this instance, we take the instance corresponding to the intervals $[\underline{x}_i, \overline{x}_i] = [\underline{y}_i, \overline{y}_i] = [-s_i, s_i]$. We want to to show that for the corresponding problem, $\overline{C} = C_0$ (where C_0 is the same as in the proof of Theorem 3.1) if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

2°. Let us first show that in all cases, $\overline{C} \le C_0$.

Indeed, it is known that the finite population covariance C is bounded by the product $\sigma_x \cdot \sigma_y$ of finite population standard deviations $\sigma_x = \sqrt{\sigma_x^2}$ and $\sigma_y = \sqrt{\sigma_y^2}$ of x and y. In the proof of Theorem 3.1, we have already proven that the finite population variance σ_x^2 of the values $x_1, \ldots, x_n$ satisfies the inequality $\sigma_x^2 \le C_0$; similarly, the finite population variance σ_y^2 of the values $y_1, \ldots, y_n$ satisfies the inequality $\sigma_y^2 \le C_0$. Hence, $C \le \sigma_x \cdot \sigma_y \le \sqrt{C_0} \cdot \sqrt{C_0} = C_0$. In other words, every possible value C of the finite population covariance is smaller than or equal to C_0. Thus, the largest of these possible values, i.e., $\overline{C}$, also cannot exceed C_0, i.e., $\overline{C} \le C_0$.

3°. Let us now show that if $\overline{C} = C_0$, then the desired signs exist.

Indeed, if $\overline{C} = C$, this means that for the corresponding values of x_i and y_i, the finite population covariance C is equal to C_0, i.e.,

$$C = C_0 = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2.$$

On the other hand, we have shown that in all cases (and in this case in particular), $C \le \sigma_x \cdot \sigma_y \le \sqrt{C_0} \cdot \sqrt{C_0} = C_0$. If $\sigma_x < \sqrt{C_0}$, then we would have $C < C_0$. So, if $C = C_0$, we have $\sigma_x = \sigma_y = \sqrt{C_0}$, i.e., $\sigma_x^2 = \sigma_y^2 = C_0$. We have already shown, in the proof of Theorem 3.1, that in this case the desired signs exist.

4°. To complete the proof of Theorem 5.1, we must show that, vice versa, if the desired signs η_i exist, then $\overline{C} = C_0$.

Indeed, in this case, for $x_i = y_i = \eta_i \cdot s_i$, we have $\mu_x = \mu_y = 0$ and $x_i \cdot y_i = s_i^2$, hence

$$C = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu_x) \cdot (y_i - \mu_y) = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i^2 = C_0.$$

The theorem is proven.

Proof of Theorem 5.2

This proof is similar to the proof of Theorem 5.1, with the only difference that in this case, we use the other part of the inequality $|C| \le \sigma_x \cdot \sigma_y$, namely, that $C \ge -\sigma_x \cdot \sigma_y$, and in the last part of the proof, we take $y_i = -x_i$.

Proof of Theorem 5.3

1°. Similarly to the proof of Theorems 3.1 and 5.1, we reduce a subset problem to the problem of computing $\overline{\sigma^2}$.

Each instance of the subset problem is as follows: given m positive integers $s_1, \ldots, s_m$, to check whether there exist signs $\eta_i \in \{-1, +1\}$ for which the signed sum $\sum_{i=1}^{m} \eta_i \cdot s_i$ equals 0.

We will show that this problem can be reduced to the problem of computing $\overline{\rho}$, i.e., that to every instance $(s_1, \ldots, s_m)$ of the subset problem $\mathcal{P}_0$, we can put into correspondence such an instance of the $\overline{\rho}$-computing problem that based on its solution, we can easily check whether the desired signs exist.

As this instance, we take the instance corresponding to the following intervals:

- $n = m+2$ (note the difference between this reduction and reductions from the proofs of Theorems 3.1 and 5.1, where we have $n = m$);
- $[\underline{x}_i, \overline{x}_i] = [-s_i, s_i]$ and $\mathbf{y}_i = [0, 0]$ for $i = 1, \ldots, m$;
- $\mathbf{x}_{m+1} = \mathbf{y}_{m+2} = [1, 1]$; $\mathbf{x}_{m+2} = \mathbf{y}_{m+1} = [-1, -1]$.

Like in the proof of Theorem 3.1, we define C_1 as

$$C_1 = \sum_{i=1}^{m} s_i^2. \tag{6.27}$$

We will prove that for the corresponding problem, $\overline{\rho} = -\sqrt{\dfrac{2}{C_1 + 2}}$ if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

2°. The correlation coefficient is defined as $\rho = C/\sqrt{\sigma_x^2} \cdot \sqrt{\sigma_y^2}$. To find the range for ρ, it is therefore reasonable to first find ranges for C, σ_x^2, and σ_y^2.

3°. Of these three, the variance σ_y^2 is the easiest to compute because there is no interval uncertainty in y_i at all. For y_i, we have $\mu_y = 0$ and therefore,

$$\sigma_y^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} y_i^2 - (\mu_y)^2 = \frac{2}{n} = \frac{2}{m+2}. \tag{6.28}$$

4°. To find the range for the covariance, we will use the known equivalent formula

$$C = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \cdot y_i - \mu_x \cdot \mu_y. \tag{6.29}$$

Since $\mu_y = 0$, the second sum in this formula is 0, so C is equal to the first sum. In this first sum, the first m terms are 0's because for $i = 1, \ldots, m$, we have $y_i = 0$. The only non-zero terms correspond to $i = m+1$ and $i = m+2$, so

$$C = -\frac{2}{n} = -\frac{2}{m+2}. \tag{6.30}$$

5°. Substituting the formulas (6.28) and (6.30) into the definition (5.1) of finite population correlation, we conclude that

$$\rho = -\frac{\dfrac{2}{m+2}}{\sqrt{\dfrac{2}{m+2}} \cdot \sqrt{\sigma_x^2}} = -\sqrt{\frac{2}{(m+2)\cdot\sigma_x^2}}. \tag{6.31}$$

Therefore, the finite population correlation ρ attains its maximum $\overline{\rho}$ if and only if the finite population variance σ_x^2 takes the largest possible value $\overline{\sigma}_x^2$:

$$\overline{\rho} = -\sqrt{\frac{2}{(m+2)\cdot\overline{\sigma}_x^2}}. \tag{6.32}$$

Thus, if we can know $\overline{\rho}$, we can reconstruct $\overline{\sigma}_x^2$ as

$$\overline{\sigma}_x^2 = \frac{2}{(m+2)\cdot(\overline{\rho})^2}. \tag{6.33}$$

In particular, the desired value $\overline{\rho} = -\sqrt{\dfrac{2}{C_1+2}}$ corresponds to $\overline{\sigma}_x^2 = \dfrac{C_1+2}{m+2}$. Therefore, to complete our proof, we must show that $\overline{\sigma}_x^2 = \dfrac{C_1+2}{m+2}$ if and only if there exist signs η_i for which $\sum \eta_i \cdot s_i = 0$.

6°. Similarly to the proof of Theorem 3.1, we will use the equivalent expression (6.6) for the finite population variance σ_x^2; we will slightly reformulate this expression by substituting the definition of μ_x into it:

$$\sigma_x^2 = \frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2. \tag{6.34}$$

We can (somewhat) simplify this expression by substituting the values $n = m+2$, $x_{m+1} = 1$, and $x_{m+2} = -1$. We have

$$\sum_{i=1}^n x_i = \sum_{i=1}^m x_i + x_{m+1} + x_{m+2} = \sum_{i=1}^m x_i$$

and

$$\sum_{i=1}^n x_i^2 = \sum_{i=1}^m x_i^2 + x_{m+1}^2 + x_{m+2}^2 = \sum_{i=1}^m x_i + 2.$$

Therefore,

$$\sigma_x^2 = \frac{1}{m+2} \cdot \sum_{i=1}^m x_i^2 + \frac{2}{m+2} - \frac{1}{(m+2)^2} \cdot \left(\sum_{i=1}^m x_i\right)^2. \tag{6.35}$$

Similarly to the proof of Theorem 3.1, we can show that always $\sigma_x^2 \le \dfrac{C_1+2}{m+2}$, and that $\overline{\sigma}_x^2 = \dfrac{C_1+2}{m+2}$ if and only if there exist the signs η_i for which $\sum \eta_i \cdot s_i = 0$.

The theorem is proven.

Proof of Theorem 5.4

This proof is similar to the proof of Theorem 5.3, with the only difference that we take $y_{m+1} = 1$ and $y_{m+2} = -1$. In this case,

$$C = \frac{2}{m+2},$$

hence

$$\rho = \sqrt{\frac{2}{(m+2)\cdot\sigma_x^2}},$$

and so the largest possible value of σ_x^2 corresponds to the smallest possible value of ρ.

Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209 and grant NCC 2-1232, NSF grants CDA-9522207, EAR-0112968, EAR-0225670, and 9710940 Mexico/Conacyt, by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grants numbers F49620-95-1-0518 and F49620-00-1-0365, by grant No. W-00016 from the U.S.-Czech Science and Technology Joint Fund, by IEEE/ACM SC2001 Minority Serving Institutions Participation Grant, and by Small Business Innovation Research grant 9R44CA81741 to Applied Biomathematics from the National Cancer Institute (NCI), a component of the National Institutes of Health (NIH). The opinions expressed herein are those of the author(s) and not necessarily those of NASA, NSF, AFOSR, NCI, or the NIH.

The authors are greatly thankful to Daniel E. Cooke, Michael Gelfond, Marek W. Gutowski, R. Baker Kearfott, Arnold Neumaier, Philippe Nivlet, Renata Maria C. R. de Souza, G. William Walster for valuable discussions, and to the anonymous referees for very useful suggestions.

References

[1] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*, MIT Press, Cambridge, MA, and Mc-Graw Hill Co., N.Y., 2001.

[2] S. Ferson, *RAMAS Risk Calc: Risk Assessment with Uncertain Numbers*, Applied Biomathematics, Setauket, NY, 1999.

[3] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Computing Variance for Interval Data is NP-Hard", *ACM SIGACT News*, 2002, Vol. 33, No. 2, pp. 108–118.

[4] W. A. Fuller, *Measurement error models*, J. Wiley & Sons, New York, 1987.

[5] M. R. Garey and S. D. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*, W.H. Freeman and Company, New York, 1979.

[6] E. Hansen, "Sharpness in interval computations", *Reliable Computing*, 1997, Vol. 3, pp. 7–29.

[7] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Kluwer, Dordrecht, 1996.

[8] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational complexity and feasibility of data processing and interval computations*, Kluwer, Dordrecht, 1997.

[9] V. P. Kuznetsov, *Interval statistical models*, Moscow, Radio i Svyaz Publ., 1991 (in Russian).

[10] P. Nivlet, F. Fournier, and J. Royer, "A new methodology to account for uncertainties in 4-D seismic interpretation", *Proceedings of the 71st Annual International Meeting of the Society of Exploratory Geophysics SEG'2001*, San Antonio, Texas, September 9–14, 2001, pp. 1644–1647.

[11] P. Nivlet, F. Fournier, and J. Royer, "Propagating interval uncertainties in supervised pattern recognition for reservoir characterization", *Proceedings of the 2001 Society of Petroleum Engineers Annual Conference SPE'2001*, New Orleans, Louisiana, September 30-October 3, 2001, paper SPE-71327.

[12] R. Osegueda, V. Kreinovich, L. Potluri, and R. Aló, "Non-Destructive Testing of Aerospace Structures: Granularity and Data Mining Approach", *Proceedings of FUZZ-IEEE'2002*, Honolulu, Hawaii, May 12–17, 2002, pp. 685–689.

[13] C. H. Papadimitriou, *Computational Complexity*, Addison Wesley, San Diego, 1994.

[14] S. Rabinovich, *Measurement Errors: Theory and Practice*, American Institute of Physics, New York, 1993.

[15] S. A. Vavasis, *Nonlinear optimization: complexity issues*, Oxford University Press, N.Y., 1991.

[16] H. M. Wadsworth, Jr. (eds.), *Handbook of statistical methods for engineers and scientists*, McGraw-Hill Publishing Co., New York, 1990.

[17] P. Walley, *Statistical reasoning with imprecise probabilities*, Chapman and Hall, N.Y., 1991.

[18] G. W. Walster, "Philosophy and practicalities of interval arithmetic", In: R. E. Moore (ed.), *Reliability in Computing*, 1988, pp. 307–323.