Exact Bounds on Sample Variance of Interval Data

Scott Ferson\textsuperscript{1}, Lev Ginzburg\textsuperscript{1},
Vladik Kreinovich\textsuperscript{2}, and Monica Aviles\textsuperscript{2}

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

Abstract

We provide a feasible (quadratic time) algorithm for computing the
lower bound $V$ on the sample variance of interval data. The problem of
computing the upper bound $\overline{V}$ is, in general, NP-hard. We provide a
feasible algorithm that computes $\overline{V}$ for many reasonable situations.

Formulation of the problem. When we have $n$ results $x_1, \ldots, x_n$ of repeated
measurement of the same quantity, traditional statistical approach usually starts
with computing their sample average

$$E = \frac{x_1 + \ldots + x_n}{n}$$

and their sample variance

$$V = \frac{(x_1 - E)^2 + \ldots + (x_n - E)^2}{n - 1}$$

(or, equivalently, the sample standard deviation $\sigma = \sqrt{V}$); see, e.g., [1].

Sample variance is an unbiased estimator of the variance of the distribution
from which observations are assumed to be randomly sampled. For Gaussian
distribution, this estimator is a maximum likelihood estimator of the distribution
variance.

In some practical situations, we only have intervals $x_i = [\xi_i, \pi_i]$ 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 $\tilde{x}_i$ measured by an instrument
with a known upper bound $\Delta_i$ on the measurement error; then, the actual
(unknown) value is within the interval $x_i = [\tilde{x}_i - \Delta_i, \tilde{x}_i + \Delta_i]$. 
As a result, the sets of possible values of $E$ and $V$ are also intervals. The interval $E$ for the sample 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:

$$E = \frac{x_1 + \ldots + x_n}{n}.$$  

What is the interval $[V, \overline{V}]$ of possible values for sample variance $V$?

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

For this problem, straightforward interval computations sometimes overestimate: E.g., for $x_1 = x_2 = [0, 1]$, the actual $V = (x_1 - x_2)^2 / 2$ and hence, the actual range $V = [0, 0.5]$. On the other hand, $E = [0, 1]$, hence $(x_1 - E)^2 + (x_2 - E)^2 = [0, 2] \supset [0, 0.5]$. Three intervals $x_i$ equal to $[0, 1]$ show that a centered form also does not always lead to the exact range.

The problem reformulated in statistical terms. The traditional sample variance is an unbiased estimator for the following problem: observation points $x_i$ satisfy the equation $x_i = u - \varepsilon_i$, where $u$ is an unknown fixed constant and the $\varepsilon_i$ are independently and identically distributed random variables with zero expectation and unknown variance $\sigma^2$.

In our paper, we want to handle a situation in which each observation point $\tilde{x}_i$ satisfies the condition $\tilde{x}_i - u - \varepsilon_i \in \Delta_i \cdot [-1, 1]$, where the values $\Delta_i$ are assumed to be known. From this model, we can conclude that each $u + \varepsilon_i$ is contained in the corresponding interval $\tilde{x}_i + \Delta_i \cdot [-1, 1] = x_i$. As a solution to this problem, we take the interval consisting of all the results of applying the estimator $V$ to different values $x_1 \in x_1, \ldots, x_n \in x_n$.

Our first result: computing $V$. First, we design a feasible algorithm for computing the exact lower bound $\underline{V}$ of the sample variance. Specifically, our algorithm is quadratic-time, i.e., it requires $O(n^2)$ computational steps for $n$ interval data points $x_i = [x_i, \overline{x}_i]$. We have implemented this algorithm in C++, it works really fast. The algorithm is as follows (the proof that this algorithm is correct will be provided in the full paper):

- First, we sort all $2n$ values $x_i, \overline{x}_i$ into a sequence $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. This sorting requires $O(n \cdot \log(n))$ steps.
- Second, we compute $E$ and $\overline{E}$ and select all “small intervals” $[x_{(k)}, x_{(k+1)}]$ that intersect with $[E, \overline{E}]$.  


• For each of selected small intervals $[x(k), x(k+1)]$, we compute the ratio $r_k = S_k / N_k$, where

$$S_k \stackrel{\text{def}}{=} \sum_{i: x_i \geq x(k+1)} x_i + \sum_{j: x_j \leq x(k)} x_j,$$

and $N_k$ is the total number of such $i$’s and $j$’s. If $r_k \notin [x(k), x(k+1)]$, we go to the next small interval, else we compute

$$V'_k \stackrel{\text{def}}{=} \frac{1}{n-1} \left( \sum_{i: x_i > x(k+1)} (x_i - r)^2 + \sum_{j: x_j < x(k)} (x_j - r)^2 \right).$$

(if $N_k = 0$, we take $V'_k \stackrel{\text{def}}{=} 0$).

• Finally, we return the smallest of the values $V'_k$ as $V$.

**Second result:** computing $V$ is NP-hard. Our second result is that the general problem of computing $V$ from given intervals $x_i$ is NP-hard.

**Third result:** a feasible algorithm that computes $V$ 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 $V$) in reasonable time.

However, we show that there are algorithms for computing $V$ for many reasonable situations. For example, we propose an efficient algorithm $A$ that computes $V$ for the case when the “narrowed” intervals $[\bar{x}_i - \Delta_i / n, \bar{x}_i + \Delta_i / n]$ where $\bar{x}_i = (\bar{x}_i + \bar{x}_i) / 2$ is the interval’s midpoint and $\Delta_i = (\bar{x}_i - \bar{x}_i) / 2$ is its half-width do not intersect with each other. We also propose, for each positive integer $k$, an efficient algorithm $A_k$ that works whenever no more than $k$ “narrowed” intervals can have a common point.

**Acknowledgments.** This work was supported in part by NASA under cooperative agreement NCC5-209 and grant NCC 2-1232, by NSF grants CDA-9522207, ERA-0112968 and 9710940 Mexico/Conacyt, by the Air Force Office of Scientific Research grant F49620-00-1-0365, and by Grant No. W-00016 from the U.S.-Czech Science and Technology Joint Fund.

The authors are greatly thankful to the anonymous referees for very useful suggestions.

**References**
