Interval Versions of Statistical Techniques, with Applications to Environmental Analysis, Bioinformatics, and Privacy in Statistical Databases

Vladik Kreinovich[1], Luc Longpré[1], Scott A. Starks[1],
Gang Xiang[1], Jan Beck[1], Raj Kandathi[1], Asis Nayak[1],
Scott Ferson[2], and Janos Hajagos[2,3]

[1]NASA Pan-American Center for Earth
and Environmental Studies (PACES)
University of Texas, El Paso, TX 79968, USA
contact email vladik@cs.utep.edu
[2]Applied Biomathematics
100 North Country Road, Setauket, NY 11733, USA
[3]Dept. of Ecology and Evolution
State University of New York
Stony Brook, NY 11794, USA

Abstract

Typical situation: we observe a pollution level $x(t)$ in a lake at different moments of time t, and we would like to estimate standard statistical characteristics such as mean, variance, covariance, etc.

In environmental measurements, we often know the values with interval uncertainty. Example: if we did not detect any pollution, the pollution value can be anywhere between 0 and the detection limit DL. Another example: to study the effect of a pollutant on the fish, we check on the fish daily; if a fish was alive on Day 5 but dead on Day 6, then the lifetime of this fish is $\in [5, 6]$. We must modify the existing statistical algorithms to process such interval data.

In general, the resulting problems are NP-hard; we overview cases when feasible algorithms exist: e.g., when measurements are very accurate, or when all the measurements are done with one (or few) instruments.

Other applications:

- In bioinformatics, we must solve systems of linear equations in which coefficients come from experts and are only known with interval uncertainty.

- To maintain privacy, we only keep a salary *range*; we must perform statistical analysis based on such interval data.

Keywords: intervals and probabilities, environmental analysis, bioinformatics, privacy, statistical databases

Statistical analysis is important. Many aspects of engineering and science involve statistical uncertainty. It is therefore desirable to estimate statistical characteristics such as mean, variance, covariance, etc., i.e., compute statistics such as $E(x) = \frac{1}{n}(x_1 + \ldots + x_n)$, $V(x) = \frac{1}{n} \cdot \sum\limits_{i=1}^{n}(x_i - E(x))^2$, and $C(x,y) = \frac{1}{n} \cdot \sum\limits_{i=1}^{n}(x_i - E(x)) \cdot (y_i - E(y))$. For example, in *non-destructive testing*, outliers are indications of faults; outliers are often detected as values outside the interval $[E(x) - k_0 \cdot \sqrt{V(x)}, E(x) + k_0 \cdot \sqrt{V(x)}]$ for $k_0 = 2$, 3, or 6. In *geophysics*, outliers indicate possible locations of minerals. In *biomedical systems*, statistical analysis often leads to improvements in medical recommendations.

Interval uncertainty. Traditional statistics assumes that we know the exact sample values $x_1, \ldots, x_n$. In practice, often, we only know x_i with interval uncertainty: $x_i \in [\underline{x}_i, \overline{x}_i]$.

For example, values x_i usually come from measurements, and we often only know the upper bounds Δ_i on the measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. So, the only information that we have about x_i is that $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

Another source of interval uncertainty is the existence of detection limits for different sensors: if a sensor, e.g., did not detect any ozone, this means that the ozone concentration is below its detection limit DL, i.e., in the interval $[0, DL]$.

Yet another source of interval uncertainty is discretized data: if we experiment on the fish and watch is daily, and a fish is alive on Day 5 but dead on Day 6, then all we know about its lifetime is that it is in the interval $[5, 6]$.

Expert estimates often come as intervals.

The need to keep privacy in statistical (e.g., medical) databases also often leads to the fact that instead of recording, e.g., exact age, we only record is in the interval $[40, 50]$.

Summarizing, often, instead of the actual values $x_1, \ldots, x_n$, we only know the intervals $\mathbf{x}_1 = [\underline{x}_1, \overline{x}_1], \ldots, \mathbf{x}_n = [\underline{x}_n, \overline{x}_n]$ that contain x_i. Different values $x_i \in \mathbf{x}_i$ lead to different values of the statistic $S(x_1, \ldots, x_n)$. It is desirable to find the range of such values:

$$S(\mathbf{x}_1, \ldots, \mathbf{x}_n) \stackrel{\text{def}}{=} \{S(x_1, \ldots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

Simple and hard cases. The mean $E(x)$ is monotonic, so $\mathbf{E}(x) = [\underline{E}(x), \overline{E}(x)]$, where $\underline{E}(x) = \dfrac{1}{n}(\underline{x}_1 + \ldots + \underline{x}_n)$ and $\overline{E}(x) = \dfrac{1}{n}(\overline{x}_1 + \ldots + \overline{x}_n)$.

For other statistics such as variance $V(x)$ or covariance $C(x, y)$, the problem is, in general, NP-hard [1, 3, 5]. In such cases, in general, we have to use approximate techniques.

Linearization and its limitations. One of the known approximate techniques is linearization, when we approximate the statistics S with the linear terms in its Taylor expansion: $S \approx S_{\text{lin}} = S_0 - \sum\limits_{i=1}^{n} S_i \cdot \Delta x_i$, where $S_0 \stackrel{\text{def}}{=} S(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, $S_i \stackrel{\text{def}}{=} \dfrac{\partial S}{\partial x_i}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, and $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. For the linear function, we get the exact formula for the range: $\mathbf{S} = [S_0 - \Delta_S, S_0 + \Delta_S]$, where $\Delta_S \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} |S_i| \cdot \Delta_i$.

However, linearization is not always acceptable. Sometimes, the intervals are wide, so that quadratic terms cannot be ignored. Sometimes – e.g., in cases of bioregulations – we want to *guarantee* that, e.g., the variance $V(x)$ is below a given threshold V_0. So, we need validated techniques.

Since we cannot provide efficient algorithms for the general case, we must find practically useful cases for which an efficient algorithm is possible.

Classes of problems for which efficient algorithms are known:

1. *Narrow intervals:* no two intervals $\mathbf{x}_i$ intersect.

2. *Slightly wider intervals:* for some integer K, no set of K intervals has a common intersection.

3. *Single measuring instrument (MI):* no two intervals are subsets of each other, i.e., $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$ (non-degenerate results are allowed).

4. *Same accuracy measurement:* $\Delta_1 = \ldots = \Delta_n$.

5. *Several MI:* intervals are divided into several subgroups each of which comes from a single MI.

6. *Privacy case:* intervals are formed from the given partition, e.g., 10 to 20, 20 to 30, etc.; in this case, every two non-degenerate intervals either coincide or do nor intersect.

7. *Non-detects:* every measurement result is either an exact value or a *non-detect*, i.e., an interval $[0, DL_i]$ for some real number DL_i.

In these cases, we have the following complexity results [4], where Class 0 means the general case (when almost all problems are NP-hard),

$$L \stackrel{\text{def}}{=} E(x) - k_0 \cdot \sqrt{V(x)}, \quad U \stackrel{\text{def}}{=} E(x) + k_0 \cdot \sqrt{V(x)},$$

R is the largest value k_0 for which $x \notin [L, U]$, i.e., $R \stackrel{\text{def}}{=} \dfrac{|x - E|}{\sqrt{V}}$, and M_m is m-th central moment: $M_m \stackrel{\text{def}}{=} \dfrac{1}{n} \sum\limits_{i=1}^{n} |x_i - E|^m$.

Class #	$E(x)$	$V(x)$	$C(x,y)$	L, U, R	M_{2p}
0	$O(n)$	NP-hard	NP-hard	NP-hard	NP-hard
1	$O(n)$	$O(n\log(n))$	$O(n^3)$	$O(n^2)$	$O(n^2)$
2	$O(n)$	$O(n^2)$	$O(n^3)$	$O(n^2)$	$O(n^2)$
3	$O(n)$	$O(n\log(n))$	?	$O(n^2)$	$O(n^2)$
4	$O(n)$	$O(n\log(n))$	$O(n^4)$	$O(n^2)$	$O(n^2)$
5	$O(n)$	$O(n^{m+1})$	?	$O(n^{m+1})$	$O(n^{m+1})$
6	$O(n)$	$O(n\log(n))$	$O(n^3)$	$O(n^2)$	$O(n^2)$
7	$O(n)$	$O(n\log(n))$	?	$O(n^2)$	$O(n^2)$

Comment: for M_{2p+1}, we have $O(n^3)$ for Classes 1 and 2, and ? (unknown) for all other classes.

Case when only d out of n data points are non-degenerate intervals. In this case, we have the following complexity results:

Class #	$E(x)$	$V(x)$	$C(x,y)$	L, U, R	M_{2p}
0	$O(n)$	NP-hard	NP-hard	NP-hard	NP-hard
1	$O(n)$	$O(n\log(d))$	$O(n \cdot d^2)$	$O(n \cdot d)$	$O(n \cdot d)$
2	$O(n)$	$O(nd)$	$O(n \cdot d^2)$	$O(n \cdot d)$	$O(n \cdot d)$
3	$O(n)$	$O(n\log(d))$	?	$O(n \cdot d)$	$O(n \cdot d)$
4	$O(n)$	$O(n\log(d))$	$O(n \cdot d^3)$	$O(n \cdot d)$	$O(n \cdot d)$
5	$O(n)$	$O(nd^m)$	?	$O(n \cdot d^m)$	$O(n \cdot d^m)$
6	$O(n)$	$O(n\log(d))$	$O(n \cdot d^2)$	$O(n \cdot d)$	$O(n \cdot d)$
7	$O(n)$	$O(n\log(d))$	?	$O(n \cdot d)$	$O(n \cdot d)$

Comment: for M_{2p+1}, we have $O(n \cdot d^2)$ for Classes 1 and 2, and ? (unknown) for all other classes.

Other statistics. Other methods for estimating mean include [6]:

- *weighted mean* that is defined by the condition $\sum\limits_{i=1}^{n} \frac{(x_i - E)^2}{\sigma^2} \to \min\limits_{E}$, so $E_w = \sum\limits_{i=1}^{n} p_i \cdot x_i$, where $p_i \stackrel{\text{def}}{=} \frac{\sigma_i^{-2}}{\sum\limits_{j=1}^{n} \sigma_j^{-2}}$;
- *L-estimates:* $\sum\limits_{i=1}^{n} w_i \cdot x_{(i)}$, where $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ are the results of ordering x_i in increasing order; median is a particular case of an L-estimate;
- *M-estimates:* $\sum\limits_{i=1}^{n} \psi(|x_i - a|) \to \max\limits_{a}$ for some function $\psi(x)$; average is a particular case of an M-estimate, corresponding to $\psi(x) = x^2$.

They are all monotonic functions of x_i, so their ranges can be computed in time $O(n)$.

Case study: bioinformatics. In cancer research, it is important to find out the genetic difference between the cancer cells and the healthy cells. In the ideal world, we should be able to have a sample of cancer cells, and a sample of healthy cells, and thus directly measure the concentrations c and h of a given gene in cancer and in healthy cells. In reality, it is very difficult to separate the cells, so we have to deal with samples that contain both cancer and normal cells. Let y_i denote the result of measuring the concentration of the gene in i-th sample, and let x_i denote the percentage of cancer cells in i-th sample. Then, we should have $x_i \cdot c + (1 - x_i) \cdot h \approx y_i$ (approximately equal because there are measurement errors in measuring y_i).

If we knew the exact percentages x_i, then, to find the desired values c and n, we would have to solve a system of equations $x_i \cdot c + (1 - x_i) \cdot h \approx y_i$, or, equivalently, $a \cdot x_i + h \approx y_i$, where we denoted $a \stackrel{\text{def}}{=} c - h$. The error of measuring y_i are normally i.i.d. random variables, so to estimate a and h, we can use the Least Squares Method (LSM) $\sum\limits_{i=1}^{n} (a \cdot x_i + h - y_i)^2 \to \min\limits_{a,h}$, according to which $a = \frac{C(x,y)}{V(x)}$ and $h = E(y) - a \cdot E(x)$. Once we know $a = c - h$ and h, we can then estimate c as $a + h$.

The problem is that the concentrations x_i comes from experts who manually count different cells, and experts can only provide interval bounds on the values x_i such as $x_i \in [0.7, 0.8]$. Different values of x_i in the corresponding intervals lead to different values of a and h. It is therefore desirable to find the range of a and h corresponding to all possible values $x_i \in [\underline{x}_i, \overline{x}_i]$.

Bioinformatics problem: linear approximation. Let $\widetilde{x}_i = (\underline{x}_i + \overline{x}_i)/2$ be the midpoint of i-th intervals, and let $\Delta_i = (\overline{x}_i - \underline{x}_i)/2$ be its half-width. For a, we have

$$\frac{\partial a}{\partial x_i} = \frac{1}{n \cdot V(x)} \cdot (y_i - E(y) - 2a \cdot x_i + 2a \cdot E(x)).$$

We can use the formula $E(y) = a \cdot E(x) + h$ to simplify this expression, resulting in $\Delta_a = \dfrac{1}{n \cdot V(x)} \sum\limits_{i=1}^{n} |\Delta y_i - a \cdot \Delta x_i| \cdot \Delta_i$, where we denoted $\Delta y_i \overset{\text{def}}{=} y_i - a \cdot x_i - h$ and $\Delta x_i \overset{\text{def}}{=} x_i - E(x)$.

Since $h = E(y) - a \cdot E(x)$, we have $\dfrac{\partial h}{\partial x_i} = -\dfrac{\partial a}{\partial x_i} \cdot E(x) - \dfrac{1}{n} \cdot a$, so $\Delta_h = \sum\limits_{i=1}^{n} \left| \dfrac{\partial h}{\partial x_i} \right| \cdot \Delta_i$.

Prior estimation of the resulting accuracy. The above formulas provides us with the accuracy *after* the data has been processed. It is often desirable to have an estimate *prior* to measurements, to make sure that we will get c and h with desired accuracy.

The difference Δy_i is a measurement errors, so it is normally distributed with 0 mean and standard deviation $\sigma(y)$ corresponding to the accuracy of measuring y_i. The difference Δx_i is distributed with 0 mean and standard deviation $\sqrt{V(x)}$. For estimation purposes, it is reasonable to assume that the values Δx_i are also normally distributed. It is also reasonable to assume that the errors in x_i and y_i are uncorrelated, so the linear combination $\Delta y_i - a \cdot \Delta x_i$ is also normally distributed, with 0 mean and variance $\sigma_y^2 + a^2 \cdot V(x)$. It is also reasonable to assume that all the values Δ_i are approximately the same: $\Delta_i \approx \Delta$.

For normal distribution ξ with 0 mean and standard deviation σ, the mean value of $|\xi|$ is equal to $\sqrt{2/\pi} \cdot \sigma$. Thus, the absolute value $|\Delta y_i - a \cdot \Delta x_i|$ of the above combination has a mean value $\sqrt{2/\pi} \cdot \sqrt{\sigma_y^2 + a^2 \cdot V(x)}$. Hence, the expected value of Δ_a is equal to $\dfrac{2}{\pi} \cdot \dfrac{\sqrt{\sigma_y^2 + a^2 \cdot V(x)} \cdot \Delta}{V(x)}$.

Since measurements are usually more accurate than expert estimates, we have $\sigma_y^2 \ll V(x)$, hence $\Delta_a \approx \dfrac{2}{\pi} \cdot a \cdot \Delta$.

Similar estimates can be given for Δ_h.

Bioinformatics: in general, finding the exact range is NP-hard. Let us show that in general, finding the exact range for the ratio $C(x,y)/V(x)$ is an NP-hard problem.

The proof is similar to the proof that computing the range for the variance is NP-hard [1, 3, 5]: namely, we reduce a partition problem (known to be NP-hard) to our problem. In the partition problem, we are given m positive integers $s_1,\ldots,s_m$, and we must check whether there exist values $\varepsilon_i \in \{-1,1\}$ for which $\sum\limits_{i=1}^m \varepsilon_i \cdot s_i = 0$. We will reduce this problem to the following problem: $n = m+2$, $y_1 = \ldots = y_m = 0$, $y_{m+1} = 1$, $y_{m+2} = -1$, $x_i = [-s_i, s_i]$ for $i \le m$, $x_{m+1} = 1$, and $y_{m+2} = -1$. In this case, $E(y) = 0$, so $C(x,y) = \dfrac{1}{n} \sum\limits_{i=1}^n x_i \cdot y_i - E(x) \cdot E(y) = \dfrac{2}{m+2}$. Therefore, $C(x,y)/V(x) \to \min$ if and only if $V(x) \to \max$.

Here, $V(x) = \dfrac{1}{m+2} \cdot \left(\sum\limits_{i=1}^m x_i^2 + 2\right) - \left(\dfrac{1}{m+2} \cdot \sum\limits_{i=1}^m x_i\right)^2$. Since $|x_i| \le s_i$, we always have $V(x) \le V_0 \stackrel{\text{def}}{=} \dfrac{1}{m+2} \cdot \left(\sum\limits_{i=1}^m s_i^2 + 2\right)$, and the only possibility to have $V(x) = V_0$ is when $x_i = \pm s_i$ for all i and $\sum x_i = 0$. Thus, $V(x) = V_0$ if and only if the original partition problem has a solution. Hence, $C(x,y)/V(x) = \dfrac{2}{\sum s_i^2 + 2}$ if and only if the original instance of the partition problem has a solution.

The reduction is proven, so our problem is indeed NP-hard.

Comment. In this proof, we consider the case when the values x_i can be negative and larger than 1, while in bioinformatics, x_i is always between 0 and 1. However, we can easily modify this proof: First, we can shift all the values x_i by the same constant to make them positive; shift does not change neither $C(x,y)$ nor $V(x)$. Second, to make the positive values ≤ 1, we can then rescale the values x_i ($x_i \to \lambda \cdot x_i$), thus multiplying $C(x,y)/V(x)$ by a known constant.

As a result, we get new values $x'_i = \dfrac{1}{2} \cdot (1 + x_i/K)$, where $K \stackrel{\text{def}}{=} \max s_i$, for which $x'_i \in [0,1]$ and the problem of computing $C(x,y)/V(x)$ is still NP-hard.

Bioinformatics: what can we do? One possibility is to known use algorithms to find the ranges for $C(x,y)$ and for $V(x)$, and then use the division operation from interval arithmetic to get the interval that is guaranteed to contain $C(x,y)/V(x)$.

Acknowledgments.

This work was supported by NASA under cooperative agreement NCC5-209, by NSF grants EAR-0112968, EAR-0225670, and EIA-0321328, by Army Research Laboratories grant DATM-05-02-C-0046, and by NIH grant 3T34GM008048-20S1.

The authors are very thankful to Ilya Shmulevich from the University of Texas M. D. Anderson Cancer Center for the formulation of the case study problem and valuable discussions.

References

[1] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Exact Bounds on Finite Populations of Interval Data", *Reliable Computing* (to appear).

[2] L. Jaulin, M. Keiffer, O. Didrit, and E. Walter, *Applied Interval Analysis*, Springer-Verlag, Berlin, 2001.

[3] V. Kreinovich, "Probabilities, Intervals, What Next? Optimization Problems Related to Extension of Interval Computations to Situations with Partial Information about Probabilities", *J. Global Optim.*, 2004, Vol. 29, No. 3, pp. 265–280.

[4] V. Kreinovich, J. Beck, C. Ferregut, A. Sanchez, G. R. Keller, M. Averill, and S. A. Starks, "Monte-Carlo-type techniques for processing interval uncertainty, and their engineering applications", *Proceedings of the Workshop on Reliable Engineering Computing*, Savannah, Georgia, September 15–17, 2004, pp. 139–160.

[5] V. Kreinovich and L. Longpré, "Computational complexity and feasibility of data processing and interval computations, with extension to cases when we have partial information about probabilities", In: V. Brattka, M. Schröder, K. Weihrauch, and N. Zhong, *Proc. Conf. on Computability and Complexity in Analysis CCA'2003*, Cincinnati, Ohio, USA, August 28–30, 2003, pp. 19–54.

[6] H. M. Wadsworth Jr., *Handbook of statistical methods for engineers and scientists*, McGraw-Hill, N.Y., 1990.

[7] W. Zhang, I. Shmulevich, and J. Astola, *Microarray Quality Control*, Wiley, Hoboken, New Jersey, 2004.