# Towards Combining Probabilistic and Interval Uncertainty in Engineering Calculations

S. A. Starks, V. Kreinovich, L. Longpré
M. Ceberio, G. Xiang, R. Araiza, J. Beck
R. Kandathi, A. Nayak, R. Torres

NASA Pan-American Center for Earth
and Environmental Studies (PACES)
University of Texas at El Paso
El Paso, TX 79968, USA
contact email vladik@cs.utep.edu

# Statistical Analysis Is Important

- Many aspects of engineering involve statistical uncertainty: metallurgy, civil engineering (material, soil), environment.
- It is desirable to estimate statistical characteristics such as mean, variance, etc., i.e., compute statistics such as

$$E = \frac{1}{n}(x_1 + \ldots + x_n); \quad V = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - E)^2.$$

- In *non-destructive testing*, outliers are indications of faults; outliers are often detected as values outside $E \pm k_0 \cdot \sigma$ intervals.
- In *geophysics*, outliers indicate possible locations of minerals.
- In *biomedical systems*, statistical analysis often leads to improvements in medical recommendations.

# Interval Uncertainty

- *Traditional statistics:* we know the exact sample values $x_1, \ldots, x_n$.
- *In practice:* often, we only know $x_i$ with interval uncertainty: $x_i \in [\underline{x}_i, \overline{x}_i]$.
- *Measurements:* values $x_i$ come from measurements.
- We often only know the upper bounds $\Delta_i$ on the measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$.
- So, $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.
- *Detection limit:* if the sensor did not detect any $O^3$, this means that the ozone concentration is in $[0, DL]$.
- *Discretized data:* if a fish is alive on Day 5 but dead on Day 6, then its lifetime is $\in [5, 6]$.
- *Expert estimates:* often come as intervals.
- *Privacy in statistical databases:* e.g., age $\in [40, 50]$.

# Estimating Statistics under Interval Uncertainty: a Problem

- We want to estimate a statistic $C(x_1,\ldots,x_n)$.
- 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 $C$.
- It is desirable to find the range of such values:

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

- *Comment:* this problem is a specific problem related to a combination of interval and probabilistic uncertainty.

- Many other problems related to this combination have been (and are being) solved.

# Simple and Hard Cases

- *Mean* $E$ is monotonic, so $\mathbf{E} = [\underline{E}, \overline{E}]$, where

$$\underline{E} = \frac{1}{n}(\underline{x}_1 + \ldots + \underline{x}_n); \quad \overline{E} = \frac{1}{n}(\overline{x}_1 + \ldots + \overline{x}_n).$$

- *Variance:* in general, NP-hard.

- *Linearization:* $C \approx C_{\rm lin} = C_0 - \sum\limits_{i=1}^{n} C_i \cdot \Delta x_i$, where $C_0 \stackrel{\rm def}{=} C(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, $C_i \stackrel{\rm def}{=} \dfrac{\partial C}{\partial x_i}(\widetilde{x}_1, \ldots, \widetilde{x}_n)$, and $\Delta x_i \stackrel{\rm def}{=} \widetilde{x}_i - x_i$.

- *Linearized estimate:* $\mathbf{C} = [C_0 - \Delta, C_0 + \Delta]$, where $\Delta \stackrel{\rm def}{=} \sum\limits_{i=1}^{n} |C_i| \cdot \Delta_i$.

- *Linearization is not always acceptable.* Examples:
  - intervals are sometimes wide, so that quadratic terms cannot be ignored;
  - sometimes, we want to *guarantee* that, e.g., the variance $V$ is $\le V_0$.

# Classes of Problems

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):* $[\underline{x}_i, \overline{x}_i] \not\subseteq (\underline{x}_j, \overline{x}_j)$ (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:* 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 $DL_i$.

| class number | class description |
|---|---|
| 0 | general case |
| 1 | narrow intervals: no intersection |
| 2 | slightly wider intervals<br>$\leq K$ intervals intersect |
| 3 | single measuring instrument (MI):<br>subset property –<br>no interval is a "proper" subset of the other |
| 4 | same accuracy measurements:<br>all intervals have the same half-width |
| 5 | several ($m$) measuring instruments:<br>intervals form $m$ groups,<br>with subset property in each group |
| 6 | privacy case:<br>intervals same or non-intersecting |
| 7 | non-detects case: $[0, DL_i]$ |

# Main Statistics

- *Mean:*

$$E \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- *Variance:*

$$V \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} (x_i - E)^2.$$

- *Covariance:*

$$C_{xy} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} (x_i - E_x) \cdot (y_i - E_y).$$

- *Outlier-related characteristics:*

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

  largest value $k_0$ for which $x \notin [L, U]$:

$$R \stackrel{\text{def}}{=} \frac{|x - E|}{\sqrt{V}}.$$

- *Central moments:*

$$M_m \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} |x_i - E|^m.$$

# Results

| # | $E$ | $V$ | $C_{xy}$ | $L, U, R$ | $M_{2p}$ |
|---|---|---|---|---|---|
| 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 Intervals

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

*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

- *Weighted mean and weighted average:*

$$\sum_{i=1}^{n} \frac{(x_i - E)^2}{\sigma^2} \to \min_E .$$

  - *Formula:* $E_w = \sum_{i=1}^{n} p_i \cdot x_i$, where $p_i \stackrel{\text{def}}{=} \dfrac{\sigma_i^{-2}}{\sum_{j=1}^{n} \sigma_j^{-2}}$.
  - *Results:* mean monotonic, hence $O(n)$;

    weighted variance $O(n^2)$ for narrow intervals.

- *Robust estimates for the mean:*
  - *L-estimates:* $\sum_{i=1}^{n} w_i \cdot x_{(i)}$ (including median).
  - *M-estimates:* $\sum_{i=1}^{n} \psi(|x_i - a|) \to \max_a$.
  - *Algorithm:* monotonic so $O(n)$.

- *Robust estimates for the generalized central moments:* $M_\psi^{\text{rob}} = \min_E \left( \frac{1}{n} \cdot \sum_{i=1}^{n} \psi(x_i - E) \right)$.

- *Algorithm:* $O(n^2)$ for single MI, $O(n^{m+1})$ for $m$ MI.

- *Correlation:* we only know that it is NP-hard.

# Additional Issue: On-Line Data Processing

- *Traditional estimates* for mean and variance can be easily modified with the arrival of the new measurement result $x_{n+1}$:
$$E' = \frac{n \cdot E + x_{n+1}}{n+1}; \quad V' = M' - (E')^2,$$
where
$$M' = \frac{n \cdot M + x_{n+1}^2}{n+1}; \quad M = V + E^2.$$
- *Interval mean:* for **E**, we can have a similar adjustment.
- *Problem:* for other statistics, known algorithms for processing interval data require that we start computation from scratch.
- *What is known:* for variance, we need $O(n)$ steps to incorporate a new interval data point.

# Parallelization

- *Motivation:* often, computing the range **C** requires too much computation time.
- *Parallel* computers speed up computations.
- *Potentially unlimited number of processors:*
  - polynomial-time algorithms can be reduced to time $O(\log(n))$;
  - exponential-time algorithms can be, in principle, reduced to linear time.
- *Realistically:* for exponential-time algorithms:
  - *computation* time is linear, but
  - *communication* time grows exponentially.
- *Limited number of processors* $p \ll n$*:* ?
- *Quantum algorithms:* can also speed up computation of **C**.

# What If We Have Partial Information about Probabilities?

- We have considered the case when $x_i \in \mathbf{x}_i$ and we have no information about probabilities.
- In many real-life situations, we have a partial information about the corresponding probabilities.
- A natural way to describe probabilities is to use cdf $F(t) \stackrel{\text{def}}{=} \text{Prob}(\Delta x \le t)$.
- In practice, we store *quantiles*, i.e., values $t_i$ for which $F(t_i) = i/n$.
- Partial info means we do not know $F(t)$; we know an interval $\mathbf{F}(t) = [\underline{F}(t), \overline{F}(t)] \ni F(t)$ (*p-box*).
- Quantiles are then also known with interval uncertainty: $t_i \in [\underline{t}_i, \overline{t}_i]$ s.t. $\overline{F}(\underline{t}_i) = i/n$ and $\underline{F}(\overline{t}_i) = i/n$.

# Processing p-Boxes and How the Above Algorithms Can Help

- Statistical characteristics can be described in terms of quantiles: e.g., $V = \int (t(\alpha) - E)^2 \, d\alpha$.
- If we only know the quantiles $t_1 = t(1/n), \ldots, t_n = t(n/n)$, then we can use an integral sum:
$$V \approx V_{\text{approx}} = \frac{1}{n} \sum_{i=1}^{n} (t_i - E)^2.$$
- When $t_i \in \mathbf{t}_i$, we have a problem similar to the above estimates, with an extra constraint $t_i \le t_{i+1}$.
- This problem corresponds to single MI.
- For variance and single MI, both min and max are attained on monotonic $x_i$.
- So, the above algorithms apply for $V_{\text{approx}}$.
- To get guaranteed bounds (not just heuristic integral sum), we replace $\mathbf{t}_i$ with $\mathbf{t}'_i = [\underline{t}_{i-1}, \overline{t}_i]$.

# Multi-Dimensional Case

- *Traditional approach:*

$$F(t_1, \ldots, t_p) = \text{Prob}(x_1 \le t_1 \,\&\, \ldots \,\&\, x_p \le t_p).$$

- *Problem:*
  - often, multi-D data represent vectors;
  - the components depend on the coordinates;
  - so often:
    * the distribution is symmetric – e.g., a rotation-invariant Gaussian distribution,
    * but the description in terms of a multi-D cdf is *not* rotation-invariant.

- *Solution:* store, for each $\vec{e}$ and $t$, the probability

$$F(\vec{e}, t) \stackrel{\text{def}}{=} \text{Prob}(\vec{x} \cdot \vec{e} \le t),$$

where $\vec{x} \cdot \vec{e} = x_1 \cdot e_1 + \ldots + x_n \cdot e_n$ is a scalar (dot) product of the two vectors.

# p-Boxes: Problem

- *Known fact:* a p-box does not fully describe all kinds of possible partial information about the probability distribution.
- *Example:* the same p-box corresponds:
  - to the class of all distributions located on an interval $[0, 1]$ and
  - to the class of all distributions located at two points 0 and 1.
- *Multi-D case:* cdfs cannot distinguish between:
  - a set $S$ (= the class of all probability distributions localized on the set $S$ with probability 1) and
  - its convex hull.

# Beyond p-Boxes

- *Idea:*
  - in addition to the bounds on the probabilities

    $$\text{Prob}(f(x) \leq t)$$

    for all *linear* functions $f(x)$,
  - to also keep the bounds on the similar probabilities corresponding to all *quadratic* functions $f(x)$.
- *Result:* we can distinguish between different closed sets.
- *1-D case:* in addition to cdf, we also store the bounds on the probabilities of $x$ being within different intervals.
- *Comment:* this is exactly Berleant's approach.

# Acknowledgments

This work was supported in part:

- 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;
- by NIH grant 3T34GM008048-20S1;
- by Applied Biomathematics.