Which Value $\bar{x}$ Best Represents a Sample $x_1, \ldots, x_n$: Utility-Based Approach Under Interval Uncertainty

Andrzej Pownuk and Vladik Kreinovich

Computational Science Program
University of Texas at El Paso
El Paso, TX 79968, USA
ampownuk@utep.edu, vladik@utep.edu

Abstract. In many practical situations, we have several estimates $x_1, \ldots, x_n$ of the same quantity $x$. In such situations, it is desirable to combine this information into a single estimate $\bar{x}$. Often, the estimates $x_i$ come with interval uncertainty, i.e., instead of the exact values $x_i$, we only know the intervals $[x_i, \bar{x}_i]$ containing these values. In this paper, we formalize the problem of finding the combined estimate $\bar{x}$ as the problem of maximizing the corresponding utility, and we provide an efficient (quadratic-time) algorithm for computing the resulting estimate.

1 Which Value $\bar{x}$ Best Represents a Sample $x_1, \ldots, x_n$: Case of Exact Estimates

Need to combine several estimates. In many practical situations, we have several estimates $x_1, \ldots, x_n$ of the same quantity $x$. In such situations, it is often desirable to combine this information into a single estimate $\bar{x}$; see, e.g., [6].

Probabilistic case. If we know the probability distribution of the corresponding estimation errors $x_i - x$, then we can use known statistical techniques to find $\bar{x}$, e.g., we can use the Maximum Likelihood Method; see, e.g., [8].

Need to go beyond the probabilistic case. In many cases, however, we do not have any information about the corresponding probability distribution [6]. How can we then find $\bar{x}$?

Utility-based approach. According to the general decision theory, decisions of a rational person are equivalent to maximizing his/her utility value $u$; see, e.g., [1, 4, 5, 7]. Let us thus find the estimate $\bar{x}$ for which the utility $u(\bar{x})$ is the largest.

Our objective is to use a single value $\bar{x}$ instead of all $n$ values $x_i$. For each $i$, the disutility $d = -u$ comes from the fact that if the actual estimate is $x_i$ and we use a different value $\bar{x} \neq x_i$ instead, we are not doing an optimal thing. For example, if the optimal speed at which the car needs the least amount of fuel is $x_i$, and we instead run it at a speed $\bar{x} \neq x_i$, we thus waste some fuel.
For each $i$, the disutility $d$ comes from the fact that the difference $\bar{x} - x_i$ is different from 0; there is no disutility if we use the actual value, so $d = d(\bar{x} - x_i)$ for an appropriate function $d(y)$, where $d(0) = 0$ and $d(y) > 0$ for $y \neq 0$.

The estimates are usually reasonably accurate, so the difference $x_i - \bar{x}$ is small, and we can expand the function $d(y)$ in Taylor series and keep only the first few terms in this expansion:

$$d(y) = d_0 + d_1 \cdot y + d_2 \cdot y^2 + \ldots$$

From $d(0) = 0$ we conclude that $d_0 = 0$. From $d(y) > 0$ for $y \neq 0$ we conclude that $d_1 = 0$ (else we would have $d(y) < 0$ for some small $y$) and $d_2 > 0$, so $d(y) = d_2 \cdot y^2 = d_2 \cdot (\bar{x} - x_i)^2$.

The overall disutility $d(\bar{x})$ of using $\bar{x}$ instead of each of the values $x_1, \ldots, x_n$ can be computed as the sum of the corresponding disutilities

$$d(\bar{x}) = \sum_{i=1}^{n} d(\bar{x} - x_i)^2 = d_2 \cdot \sum_{i=1}^{n} (\bar{x} - x_i)^2.$$  

Maximizing utility $u(\bar{x}) \overset{\text{def}}{=} -d(\bar{x})$ is equivalent to minimizing disutility.

The resulting combined value. Since $d_2 > 0$, minimizing the disutility function is equivalent to minimizing the re-scaled disutility function

$$D(\bar{x}) \overset{\text{def}}{=} \frac{d(\bar{x})}{d_2} = \sum_{i=1}^{n} (\bar{x} - x_i)^2.$$  

Differentiating this expression with respect to $\bar{x}$ and equating the derivative to 0, we get

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i.$$  

This is the well-known sample mean.

2 Case of Interval Uncertainty: Formulation of the Problem

Formulation of the practical problem. In many practical situations, instead of the exact estimates $x_i$, we only know the intervals $[\underline{x}_i, \overline{x}_i]$ that contain the unknown values $x_i$. How do we select the value $x$ in this case?

Towards precise formulation of the problem. For different values $x_i$ from the corresponding intervals $[\underline{x}_i, \overline{x}_i]$, we get, in general, different values of utility

$$U(\bar{x}, x_1, \ldots, x_n) = -D(\bar{x}, x_1, \ldots, x_n),$$
Which Value $x$ Best Represents a Sample $x_1, \ldots, x_n$

where $D(\bar{x}, x_1, \ldots, x_n) = \sum_{i=1}^{n} (\bar{x} - x_i)^2$. Thus, all we know is that the actual (unknown) value of the utility belongs to the interval $[U(\bar{x}), \overline{U}(\bar{x})] = [-D(\bar{x}), -\overline{D}(\bar{x})]$, where

$$D(\bar{x}) = \min D(\bar{x}, x_1, \ldots, x_n),$$

$$\overline{D}(\bar{x}) = \max D(\bar{x}, x_1, \ldots, x_n),$$

and min and max are taken over all possible combinations of values $x_i \in [x_i, \bar{x}_i]$.

In such situations of interval uncertainty, decision making theory recommends using Hurwicz optimism-pessimism criterion [2–4], i.e., maximize the value

$$U(\bar{x}) \overset{\text{def}}{=} \alpha \cdot U(\bar{x}) + (1 - \alpha) \cdot \overline{U}(\bar{x}),$$

where the parameter $\alpha \in [0, 1]$ describes the decision maker’s degree of optimism. For $U = -D$, this is equivalent to minimizing the expression

$$D(\bar{x}) = -U(\bar{x}) = \alpha \cdot D(\bar{x}) + (1 - \alpha) \cdot \overline{D}(\bar{x}).$$

**What we do in this paper.** In this paper, we describe an efficient algorithm for computing the value $\bar{x}$ that minimizes the resulting objective function $D(\bar{x})$.

### 3 Analysis of the Problem

**Let us simplify the expressions for $D(\bar{x})$, $\overline{D}(\bar{x})$, and $D(\bar{x})$.** Each term $(\bar{x} - x_i)^2$ in the sum $D(\bar{x}, x_1, \ldots, x_n)$ depends only on its own variable $x_i$. Thus, with respect to $x_i$:

- the sum is the smallest when each of these terms is the smallest, and
- the sum is the largest when each term is the largest.

One can easily see that when $x_i$ is in the $[x_i, \bar{x}_i]$, the maximum of a term $(\bar{x} - x_i)^2$ is always attained at one of the interval’s endpoints:

- at $x_i = \underline{x}_i$ when $\bar{x} \geq \bar{x}_i \overset{\text{def}}{=} \underline{x}_i + \frac{\overline{x}_i - \underline{x}_i}{2}$ and
- at $x_i = \bar{x}_i$ when $\bar{x} < \bar{x}_i$.

Thus,

$$\overline{D}(\bar{x}) = \sum_{i: \bar{x} < \bar{x}_i} (\bar{x} - \underline{x}_i)^2 + \sum_{i: \bar{x} \geq \bar{x}_i} (\bar{x} - \bar{x}_i)^2.$$

Similarly, the minimum of the term $(\bar{x} - x_i)^2$ is attained:

- for $x_i = \bar{x}$ when $\bar{x} \in [\underline{x}_i, \overline{x}_i]$ (in this case, the minimum is 0);
- for $x_i = \underline{x}_i$ when $\bar{x} < \underline{x}_i$; and
- for $x_i = \overline{x}_i$ when $\bar{x} > \overline{x}_i$. 

Thus,

\[ D(\bar{x}) = \sum_{i: \bar{x} > \bar{x}_i} (\bar{x} - \bar{x}_i)^2 + \sum_{i: \bar{x} < \bar{x}_i} (\bar{x} - \bar{x}_i)^2. \]

So, for \( D(\bar{x}) = \alpha \cdot D(\bar{x}) + (1 - \alpha) \cdot D(\bar{x}) \), we get

\[
D(\bar{x}) = \alpha \cdot \sum_{i: \bar{x} > \bar{x}_i} (\bar{x} - \bar{x}_i)^2 + \alpha \cdot \sum_{i: \bar{x} < \bar{x}_i} (\bar{x} - \bar{x}_i)^2 + (1 - \alpha) \cdot \sum_{i: \bar{x} > \bar{x}_i} (\bar{x} - \bar{x}_i)^2.
\]

Towards an algorithm. The presence or absence of different values in the above expression depends on the relation of \( \bar{x} \) with respect to the values \( \bar{x}_i, \bar{x}_i \), and \( \bar{x}_i \). Thus, if we sort these \( 3n \) values into a sequence \( s_1 \leq s_2 \leq \ldots \leq s_{3n} \), then on each interval \([s_j, s_{j+1}]\), the function \( D(\bar{x}) \) is simply a quadratic function of \( \bar{x} \).

A quadratic function attains its minimum on an interval either at one of its midpoints, or at a point when the derivative is equal to 0 (if this point is inside the given interval). Differentiating the above expression for \( D(\bar{x}) \), equating the derivative to 0, dividing both sides by 0, and moving terms proportional not containing \( \bar{x} \) to the right-hand side, we conclude that

\[
(\alpha \cdot \#\{i: \bar{x} < \bar{x}_i \text{ or } \bar{x} > \bar{x}_i\} + 1 - \alpha) \cdot \bar{x} = \alpha \cdot \sum_{i: \bar{x} > \bar{x}_i} \bar{x}_i + \alpha \cdot \sum_{i: \bar{x} < \bar{x}_i} \bar{x}_i + (1 - \alpha) \cdot \sum_{i: \bar{x} \geq \bar{x}_i} \bar{x}_i.
\]

Since \( s_j \) is a listing of all thresholds values \( \bar{x}_i, \bar{x}_i \), and \( \bar{x}_i \), then for \( \bar{x} \in (s_j, s_{j+1}) \), the inequality \( \bar{x} < \bar{x}_i \) is equivalent to \( s_{j+1} \leq \bar{x}_i \). Similarly, the inequality \( \bar{x} > \bar{x}_i \) is equivalent to \( s_j \geq \bar{x}_i \). In general, for values \( \bar{x} \in (s_j, s_{j+1}) \), the above equation gets the form

\[
(\alpha \cdot \#\{i: \bar{x} < \bar{x}_i \text{ or } \bar{x} > \bar{x}_i\} + 1 - \alpha) \cdot \bar{x} = \alpha \cdot \sum_{i: \bar{x}_j \leq \bar{x}_i} \bar{x}_i + \alpha \cdot \sum_{i: \bar{x}_{j+1} \leq \bar{x}_i} \bar{x}_i + (1 - \alpha) \cdot \sum_{i: \bar{x}_j \leq \bar{x}_i} \bar{x}_i.
\]

From this equation, we can easily find the desired expression for the value \( \bar{x} \) at which the derivative is 0.

Thus, we arrive at the following algorithm.

4 Resulting Algorithm

First, for each interval \([\bar{x}_i, \bar{x}_i]\), we compute its midpoint \( \bar{x}_i = \frac{\bar{x}_i + \bar{x}_i}{2} \). Then, we sort the \( 3n \) values \( \bar{x}_i, \bar{x}_i \), and \( \bar{x}_i \) into an increasing sequence \( s_1 \leq s_2 \leq \ldots \leq s_{3n} \). To cover the whole real line, to these values, we add \( s_0 = -\infty \) and \( s_{3n+1} = +\infty \).
We compute the value of the objective function (1) on each of the endpoints $s_1, \ldots, s_{3n}$. Then, for each interval $(s_i, s_{j+1})$, we compute the value

$$
\tilde{x} = \frac{\alpha \cdot \sum_{i:s_i \leq \tau_i} \tau_i + \alpha \cdot \sum_{i:s_{j+1} \leq \tau_i} x_i + (1 - \alpha) \cdot \sum_{i:s_{j+1} \leq \tau_i} x_i + (1 - \alpha) \cdot \sum_{i:s_j \geq \tau_i} x_i}{\alpha \cdot \# \{ i : \tilde{x} < x_i \text{ or } \tilde{x} > \tau_i \} + 1 - \alpha}.
$$

If the resulting value $\tilde{x}$ is within the interval $(s_i, s_{j+1})$, we compute the value of the objective function (1) corresponding to this $\tilde{x}$.

After that, out of all the values $\tilde{x}$ for which we have computed the value of the objective function (1), we return the value $\tilde{x}$ for which objective function $D(\tilde{x})$ was the smallest.

**What is the computational complexity of this algorithm.** Sorting $3n = O(n)$ values $x_i, \tau_i$, ad $\tilde{x}$, takes time $O(n \cdot \ln(n))$.

Computing each value $d(\tilde{x})$ of the objective function requires $O(n)$ computational steps. We compute $d(\tilde{x})$ for $3n$ endpoints and for $\leq 3n + 1$ values at which the derivative is 0 at each of the intervals $(s_j, s_{j+1})$ – for the total of $O(n)$ values.

Thus, overall, we need $O(n \cdot \ln(n)) + O(n) \cdot O(n) = O(n^2)$ computation steps. Hence, our algorithm runs in quadratic time.

**Acknowledgments.** This work was supported in part by the National Science Foundation grants HRD-0734825 and HRD-1242122 (Cyber-ShARE Center of Excellence) and DUE-0926721, and by an award “UTEP and Prudential Actuarial Science Academy and Pipeline Initiative” from Prudential Foundation.

**References**