Estimating Mean and Variance under Interval Uncertainty: Dynamic Case

Rafik Aliev[1] and Vladik Kreinovich[2]

[1]Dept. of Computer Aided Control Systems
Azerbaijan State Oil Academy
Azadlig Ave. 20, AZ1010 Baki, Azerbaijan
raliev@asoa.edu.az

[2]Department of Computer Science
University of Texas at El Paso
500 W. University, El Paso, TX 79968, USA
vladik@utep.edu

Statistical Analysis in . . .
Statistical Analysis: . . .
Need to Take Interval . . .
Case of Interval . . .
Need to Consider . . .
Simplest Case: . . .
Efficient Algorithm for . . .
Efficient Algorithm for . . .
Computing the Range . . .

1. Statistical Analysis in Gaussian Case: Reminder

- Standard methods for estimating the mean E and the variance V assume normal distribution:

$$\rho_N(x) = \frac{1}{\sqrt{2\pi \cdot V}} \cdot \exp\left(-\frac{(x-E)^2}{2V}\right).$$

- Normal distributions are ubiquitous, due to the Central Limit Theorem: sum of many small factors $\approx \rho_N(x)$.
- It is usually assumed that different sample values are independent, so

$$L = \prod_{i=1}^{n} \rho_N(x_i) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot V}} \cdot \exp\left(-\frac{(x_i-E)^2}{2V}\right).$$

- It is reasonable to select the *Maximum Likelihood* (*most probable*) values E and V s.t. $L \to \max$, then:

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

2. Statistical Analysis: General Case

- Often, distributions are non-Gaussian; Gaussian-generated estimated are used in the general case as well:

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

- *Justification:* the mean $E[x]$ is the limit of the expression $\frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ when $n \to \infty$.
- So, for large n, this expression is a good approximation for $E[x]$; the larger n, the better the approximation.
- Similarly, the Gaussian expression for V tends to the actual variance $V[x]$.
- *Caution:* for non-Gaussian distributions, the above estimates are *not* necessarily *optimal*.

3. Need to Take Interval Uncertainty into Account

- In practice, the values x_i come from measurements, and measurements are never 100% accurate: $\widetilde{x}_i \neq x_i$.
- Sometimes, we know the probabilities of different values of measurement errors $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$
- However, in many cases, we only know the upper bound Δ_i on the measurement error: $|\Delta x_i| \leq \Delta_i$.
- In this case, we know that $x_i \in \mathbf{x}_i \stackrel{\text{def}}{=} [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.
- Different values x_i from these intervals lead, in general, to different estimates of $E(x_1, \ldots, x_n)$ and $V(x_1, \ldots, x_n)$.
- It is therefore desirable to find the ranges

$$\mathbf{E} = [\underline{E}, \overline{E}] = \{E(x_1, \ldots, x_n) | x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\} \text{ and}$$

$$\mathbf{V} = [\underline{V}, \overline{V}] = \{V(x_1, \ldots, x_n) | x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

Statistical Analysis in . . .
Statistical Analysis: . . .
Need to Take Interval . . .
Case of Interval . . .
Need to Consider . . .
Simplest Case: . . .
Efficient Algorithm for . . .
Efficient Algorithm for . . .
Computing the Range . . .

4. Case of Interval Uncertainty: What Is Known

- Estimating the range of a function under interval uncertainty is known as *interval computations*.
- The mean $E(x_1, \ldots, x_n) = \dfrac{1}{n} \cdot \sum_{i=1}^{n} x_i$ is an increasing function of each of its variables $x_1, \ldots, x_n$, hence:

$$[\underline{E}, \overline{E}] = \left[\frac{1}{n} \cdot \sum_{i=1}^{n} \underline{x}_i, \frac{1}{n} \cdot \sum_{i=1}^{n} \overline{x}_i \right].$$

- For variance V, the situation is more complex:
  - the lower endpoint $\underline{V}$ can be computed in feasible time;
  - in general, computing $\overline{V}$ is NP-hard;
  - for some practically useful situations, there exist efficient algorithms for computing $\overline{V}$.

5. Need to Consider Dynamic Estimates

- In practice, processes are dynamic: means and variances change with time.
- Reasonable estimates should assign more weight to more recent measurements $x_1, \ldots$ and less to the past ones.
- For each function $y(x)$, we thus take the weighted mean

$$E[y] \approx \sum_{i=1}^{n} w_i \cdot y(x_i); \quad w_i \geq 0 \quad \sum_{i=1}^{n} w_i = 1.$$

- In particular, for $E[x]$ and $V = E[(x - E)^2]$, we take

$$E = \sum_{i=1}^{n} w_i \cdot x_i; \quad V = \sum_{i=1}^{n} w_i \cdot (x_i - E)^2.$$

- What we do: we extend known algorithms for computing the ranges $\mathbf{E}$ and $\mathbf{V}$ to such dynamic estimates.

Statistical Analysis in . . .
Statistical Analysis: . . .
Need to Take Interval . . .
Case of Interval . . .
Need to Consider . . .
Simplest Case: . . .
Efficient Algorithm for . . .
Efficient Algorithm for . . .
Computing the Range . . .

6. Simplest Case: Estimates for the Mean

- Since all the weights are non-negative, the function $E = \sum_{i=1}^{n} w_i \cdot x_i$ is an increasing function of all x_i.
- Thus:
  - the smallest possible value $\underline{E}$ is attained when we take the smallest possible values $x_i = \underline{x}_i$, and
  - the largest possible value $\overline{E}$ is attained when we take the largest possible values $x_i = \overline{x}_i$.
- So, the desired range of E has the form

$$[\underline{E}, \overline{E}] = \left[\sum_{i=1}^{n} w_i \cdot \underline{x}_i, \sum_{i=1}^{n} w_i \cdot \overline{x}_i\right].$$

Statistical Analysis in . . .
Statistical Analysis: . . .
Need to Take Interval . . .
Case of Interval . . .
Need to Consider . . .
Simplest Case: . . .
Efficient Algorithm for . . .
Efficient Algorithm for . . .
Computing the Range . . .

7. Efficient Algorithm for Computing $\underline{V}$

- We sort all endpoints $\underline{x}_i$ and $\overline{x}_i$:

$$r_1 \le r_2 \le \ldots \le r_{2n-1} \le r_{2n}.$$

- Thus, the real line is divided into $2n+1$ zones $[r_k, r_{k+1}]$, with $k = 0, 1, \ldots, 2n$ ($r_0 = -\infty$ and $r_{2n+1} = +\infty$).
- For each zone, we compute $E_k = \dfrac{N_k}{D_k}$, where

$$N_k \stackrel{\text{def}}{=} \sum_{i:\overline{x}_i \le r_k} w_i \cdot \overline{x}_i + \sum_{j:r_{k+1} \le \underline{x}_j} w_j \cdot \underline{x}_j; \quad D_k = \sum_{i:\overline{x}_i \le r_k} w_i + \sum_{j:r_{k+1} \le \underline{x}_j} w_j.$$

- If $E_k \notin [r_k, r_{k+1}]$, we move to the next zone.
- If $E_k \in [r_k, r_{k+1}]$, we compute $V_k = M_k - D_k \cdot E_k^2$, where

$$M_k = \sum_{i:\overline{x}_i \le r_k} w_i \cdot (\overline{x}_i)^2 + \sum_{j:r_{k+1} \le \underline{x}_j} w_j \cdot (\underline{x}_j)^2.$$

- The smallest of the corresponding values V_k is the desired smallest value $\underline{V}$.

8. Computation Time of This Algorithm

- Sorting takes time $O(n \log \log(n))$.
- Computing the sums D_0, N_0, M_0 corresponding to the first zone take linear time $O(n)$.
- Each new sum is obtained from the previous one by changing a few terms which go from $\underline{x}_i$ to $\overline{x}_i$.
- Each value x_i changes only once, so we only need totally linear time to compute all these sums.
- We also need linear time to perform all the auxiliary computations.
- Thus, the total computation time is

$$O(n \cdot \log(n)) + O(n) + O(n) = O(n \cdot \log(n)).$$

- This time can be reduced to $O(n)$ if, instead of sorting, we use the $O(n)$ algorithm for computing the median.

9. Efficient Algorithm for Computing $\overline{V}$ under a Reasonable Condition

- We assume that for some integer C, each set of more than C intervals has an empty intersection.
- We sort $\underline{x}_i$ and $\overline{x}_i$: $r_1 \le r_2 \le \ldots \le r_{2n-1} \le r_{2n}$.
- For each zone $[r_k, r_{k+1}]$, we find optimal x_i under the condition that $E \in [r_k, r_{k+1}]$:
  - for those i for which $\overline{x}_i \le r_k$, we take $x_i = \underline{x}_i$;
  - for those i for which $r_{k+1} \le \underline{x}_i$, we take $x_i = \overline{x}_i$;
  - for all other i, we consider both $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$.
- For each of the resulting combinations, we compute the weighted average E.
- If $E \in [r_k, r_{k+1}]$, we compute the weighted variance V.
- The largest of all such computed values V is then returned as $\overline{V}$.

10. Computation Time of This Algorithm

- Sorting takes time $O(n \cdot \log(n))$.
- Computing the original values of E and M requires linear time.
- For each zone, we have $\leq C$ "other" indices, so we analyze $\leq 2^C = O(1)$ combinations.
- Each new sum is obtained from the previous one by changing a few terms – which go from $\underline{x}_i$ to $\overline{x}_i$.
- Each value x_i changes only once, so we only need totally linear time to compute all these sums.
- We also need linear time to perform all the auxiliary computations.
- Thus, the total computation time is also

$$O(n \cdot \log(n)) + O(n) + O(n) = O(n \cdot \log(n)).$$

Statistical Analysis in . . .
Statistical Analysis: . . .
Need to Take Interval . . .
Case of Interval . . .
Need to Consider . . .
Simplest Case: . . .
Efficient Algorithm for . . .
Efficient Algorithm for . . .
Computing the Range . . .

11. Computing the Range of Covariance

- In forming large statistical databases, we need to preserve privacy.
- One way is to only ask threshold-related questions: e.g., whether the age is from 0 to 20, from 20 to 30.
- In this case, all x-intervals are of the form $[t_i^{(x)}, t_{i+1}^{(x)}]$ for some we have x-threshold values $t_0^{(x)} < t_1^{(x)} < \ldots < t_{N_x}^{(x)}$.
- For these intervals, we want to compute the range of the weighted covariance

$$C = \sum_{i=1}^{n} w_i \cdot (x_i - E_x) \cdot (y_i - E_y) = \sum_{i=1}^{n} w_i \cdot x_i \cdot y_i,$$

$$\text{where } E_x \stackrel{\text{def}}{=} \sum_{i=1}^{n} w_i \cdot x_i \text{ and } E_y \stackrel{\text{def}}{=} \sum_{i=1}^{n} w_i \cdot y_i.$$

- For this computations, we also provide a similar feasible (polynomial-time) algorithm.

12. Acknowledgments

This work was supported in part:

- by the National Science Foundation grants HRD-0734825 and DUE-0926721, and
- by Grant 1 T36 GM078000-01 from the National Institutes of Health.

The author is greatly thankful to the conference organizers for the invitation.

13. Estimates for the Variance: Analysis of the Problem

- In designing our algorithms, we used known facts from calculus.
- A function $f(x)$ defined on an interval $[\underline{x}, \overline{x}]$ attains its minimum on this interval
  - either at one of its endpoints,
  - or in some internal point of the interval.
- If it attains is minimum at a point $x \in (a, b)$, then its derivative at this point is 0: $\dfrac{df}{dx} = 0$.
- If it attains its minimum at the point $x = \underline{x}$, then we cannot have $\dfrac{df}{dx} < 0$, so $\dfrac{df}{dx} \geq 0$.
- Similarly, if a function $f(x)$ attains its minimum at the point $x = \overline{x}$, then we must have $\dfrac{df}{dx} \leq 0$.

14. Where Is The Minimum Attained: Analysis

- *For the weighted variance:* $\dfrac{\partial V}{\partial x_i} = 2w_i \cdot (x_i - E)$; so:

$$x_i = \underline{x}_i \Rightarrow x_i \geq E; \quad x_i = \overline{x}_i \Rightarrow x_i \leq E; \quad \underline{x}_i < x_i < \overline{x}_i \Rightarrow x_i = E.$$

- If $\overline{x}_i < E$, this means that for the value $x_i \leq \overline{x}_i$ also satisfies the inequality $x_i < E$; thus, in this case:
  - we cannot have $x_i = \underline{x}_i$ — because then we would have $x_i \geq E$; and
  - we cannot have $\underline{x}_i < x_i < \overline{x}_i$ – because then, we would have $x_i = E$.
- So, if $\overline{x}_i < E$, the only remaining option is $x_i = \overline{x}_i$.
- Likewise, if $E < \underline{x}_i$, the only remaining option for x_i is $x_i = \underline{x}_i$.

15. Where Is The Minimum Attained (cont-d)

- When $\underline{x}_i < E < \overline{x}_i$, then:
  - the minimum cannot be attained for $x_i = \underline{x}_i$, because then $x_i \geq E$, while we have $x_i < E$;
  - the minimum cannot be attained for $x_i = \overline{x}_i$, because then $x_i \leq E$, while we have $x_i > E$.
- Thus, the minimum has to be attained when $x_i \in (\underline{x}_i, \overline{x}_i)$. In this case, we have $x_i = E$; So:

$$\overline{x}_i \leq E \to x_i = \overline{x}_i; \;\; E \leq \underline{x}_i \Rightarrow x_i = \underline{x}_i; \;\; \underline{x}_i < E < \overline{x}_i \Rightarrow x_i = E.$$

- In all 3 cases, once we know where E is relative to $\underline{x}_i$ and $\overline{x}_i$, we can find, for each i, the minimizing x_i.
- The value E must be found from the condition that it is the weighted mean of all minimizing x_i.
- This leads to the above algorithm for computing $\underline{V}$.

16. Justification of the Algorithm for Computing $\overline{V}$

- The function $V(x_1, \ldots, x_n)$ is convex, so its maximum is always attained at one of the endpoints of $[\underline{x}_i, \overline{x}_i]$.
- From a calculus-based analysis, we can now come up with the following conclusions:
  - if the maximum is attained for $x_i = \underline{x}_i$, then we should have $x_i \le E$, i.e., $\underline{x}_i \le E$;
  - if the maximum is attained for $x_i = \overline{x}_i$, then we should have $x_i \ge E$, i.e., $E \le \overline{x}_i$.
- Thus, if $\overline{x}_i < E$, we cannot have $x_i = \overline{x}_i$, so the maximum is attained for $x_i = \underline{x}_i$.
- Similarly, if $E < \underline{x}_i$, then we cannot have $x_i = \underline{x}_i$, so the maximum is attained for $x_i = \overline{x}_i$.
- If $\underline{x}_i \le E \le \overline{x}_i$, then we can have both options $x_i = \underline{x}_i$ and $x_i = \overline{x}_i$.

Statistical Analysis in . . .
Statistical Analysis: . . .
Need to Take Interval . . .
Case of Interval . . .
Need to Consider . . .
Simplest Case: . . .
Efficient Algorithm for . . .
Efficient Algorithm for . . .
Computing the Range . . .