Interval Approach to Preserving Privacy in Statistical Databases: Related Challenges and Algorithms of Computational Statistics

Luc Longpré, Gang Xiang,
Vladik Kreinovich, and Eric Freudenthal

Department of Computer Science
University of Texas, El Paso, Texas 79968, USA
contact email vladik@utep.edu

1. Need for Statistical Databases

- *Fact:* in many areas, statistics is gathered.
- *Why:* it is useful for many practical situations.
- *Example of gathering statistics:* a census.
- *Information gathered:* data about health, employment, and mortality in different regions.
- *Application:* so that resources can be allocated where they are needed the most.
- *Other applications:* industrial and medical fields.
- *Statistical databases:* databases whose intent is for outside users to compute statistics.

2. Need for Statistical Analysis, Need for Privacy

- *What we want to compute:* statistical characteristics such as
  - statistical moments, such as mean E, variance $V = M_2$, skewness $S = M_3$, and higher central moments M_m,
  - covariance C_{xy}, correlation ρ, etc.
- *Applications:* these characteristics provide valuable information on the distribution of the data.
- *Need for privacy:* a large part of this data is *sensitive*, such as salaries, medical information, etc.
- *Objective:*
  - outside users *should* be able to perform statistical analysis,
  - but outside users *should not* be able to get sensitive information about individuals.

3. Maintaining Privacy is Not Easy

- *Misconception:* anonymity, averaging protect privacy.
- *Main idea of anonymity:* delete the names from all the records.
  - *Toy example:* faculty data, with salary, department, education.
  - *Privacy violation:* ask for the data about a CS Dept. professor with PhD from Russia.
- *Main idea of averaging:* only return averages.
  - *Toy example:* same salaries database.
  - *Privacy violation:* ask for the average salary E_{all} and E_{nR} of all CS professors and all whose PhD is not from Russia:

$$E_{\text{all}} = \frac{1}{n} \cdot \sum_{i=1}^{n} s_i, \quad E_{\text{nR}} = \frac{1}{n-1} \cdot \sum_{i \neq i_R} s_i, \quad s_{i_R} = n \cdot E_{\text{all}} - (n-1) \cdot E_{\text{nR}}.$$

4. Maintaining Privacy: Interval Approach

- *Main idea:* instead of storing the actual values x_i, we only store *ranges* $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$.
- *Traditional approach:* we ask a person i for his or her age x_i.
- *Interval approach:* we only ask whether the age is between, say, 0 and 10, 10 and 20, 20 and 30, etc.
- *Example:* a 28 years-old person.
- *What we store:* we only store the interval value $[20, 30]$ years in the *age* field of this person's record.
- *Fact:* we do not store the actual data.
- *Result—privacy is preserved:* we cannot reconstruct the actual data, no matter how many queries we ask.

5. Interval Approach to Preserving Privacy: Computational Challenges

- *Reminder:* to preserve privacy, instead of the actual values x_i, we only store their ranges $\mathbf{x}_i$.
- *New problem:* what to return if a query asks for a statistical characteristic $C(x_1, \ldots, x_n)$ such as variance?
- *Difficulty:* different possible values $x_i \in \mathbf{x}_i$ lead, in general, to different values $C(x_1, \ldots, x_n)$.
- *Possible solution:* return the range of possible values of $C(x_1, \ldots, x_n)$:

$$\mathbf{C} = [\underline{C}, \overline{C}] \stackrel{\text{def}}{=} \{C(x_1, \ldots, x_n) : x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n\}.$$

- *Computational problem:* how to compute $\mathbf{C}$?

6. What Are the Statistical Properties of These Estimations?

- *We compute:* a statistical characteristic $C_n = C(x_1, \ldots, x_n)$.
- *Example:* population mean E.
- *We are interested in:* the *actual* value c of the actual distribution (e.g., mean).
- *Good news for point estimates:* usually, the difference $d(C_n, c) \stackrel{\text{def}}{=} |C_n - c|$ tends to 0 fast.
- *How good* are interval estimates $\mathbf{C}_n = C(\mathbf{x}_1, \ldots, \mathbf{x}_n)$?
- *Natural metric:* Hausdorff distance $d(c, \mathbf{C}_n) \stackrel{\text{def}}{=} \min\limits_{C \in \mathbf{C}_n} d(c, C)$.
- *Observation:* $x_i \in \mathbf{x}_i$, hence $C_n = C(x_1, \ldots, x_n) \in \mathbf{C}_n$, and $d(c, \mathbf{C}_n) = \min\limits_{C \in \mathbf{C}_n} d(c, C) \le d(c, C_n)$.
- *Conclusion:* the rate of convergence for interval estimates is the same (or better) as for point estimates.

7. Statistical Data Processing under Interval Uncertainty: General Formulation of the Problem

- *Given:* n intervals $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ and a function $C(x_1, \ldots, x_n)$.
- *Compute:* the range

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

- *Interval computations:* general case; we want

$$E = \frac{1}{n}\sum_{i=1}^{n} x_i,\ V = \frac{1}{n}\sum_{i=1}^{n}(x_i - E)^2,\ C_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - E)\cdot(y_i - E).$$

- *Known fact:* statistical data processing under interval uncertainty is NP-hard even for V.
- *Informal meaning:* no algorithm can efficiently solve all the instances of this problem.
- *Practical solution:* find practically useful subclasses for which an efficient algorithm is possible.

8. Estimating Mean under Interval Uncertainty: What Is Known

- *Fact:* the arithmetic average $E(x_1, \ldots, x_n)$ is an increasing function of $x_1, \ldots, x_n$.
- *Conclusions:*
  - the smallest possible value $\underline{E}$ of E is attained when each value x_i is the smallest possible ($x_i = \underline{x}_i$);
  - the largest possible value $\overline{E}$ of E is attained when $x_i = \overline{x}_i$ for all i.
- *Resulting formulas:* the range $\mathbf{E}$ of E is equal to

$$[E(\underline{x}_1, \ldots, x_n), E(\overline{x}_1, \ldots, \overline{x}_n)],$$

i.e., to

$$\mathbf{E} = [\underline{E}, \overline{E}] = \left[\frac{1}{n} \cdot (\underline{x}_1 + \ldots + \underline{x}_n), \frac{1}{n} \cdot (\overline{x}_1 + \ldots + \overline{x}_n)\right].$$

9. Estimating Variance under Interval Uncertainty: What is Known

- *Problem:* compute the range $\mathbf{V} = [\underline{V}, \overline{V}]$ of the variance V over interval data $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.
- *Known:* there is a polynomial-time algorithm for computing $\underline{V}$.
- *In general:* computing $\overline{V}$ is NP-hard.
- *In many practical situations:* there are efficient algorithms for computing $\overline{V}$.
- *Example:* consider narrowed intervals $[x_i^-, x_i^+]$, where $x_i^- \stackrel{\text{def}}{=} \widetilde{x}_i - \dfrac{\Delta_i}{n}$ and $x_i^+ \stackrel{\text{def}}{=} \widetilde{x}_i + \dfrac{\Delta_i}{n}$.
- *Case:* no two narrowed intervals are proper subsets of one another, i.e., $[x_i^-, x_i^+] \not\subseteq (x_j^-, x_j^+)$ for all i and j.
- *For this case:* there exists an $O(n \cdot \log(n))$ time algorithm for computing $\overline{V}$.

10. Hierarchical Case: Formulation of the Problem

- *Situation:* often,
  - we do not know the *individual* values of x_i;
  - we only have *average* values corresponding to several ($m < n$) groups $I_1, \ldots, I_m$ of observations;
  - example: statistics by counties or by states.
- *Typically:* for each group I_j, we know
  - the *frequency* p_j of this group (i.e., the probability that a randomly selected observation belongs to I_j),
  - the *average* E_j over this group, and
  - the *standard deviation* σ_j within j-th group.
- *Formulas:* $E = \sum_{j=1}^{m} p_j \cdot E_j$ and $V = V_E + V_\sigma$, where $V_E \stackrel{\text{def}}{=} M_E - E^2$, $M_E \stackrel{\text{def}}{=} \sum_{j=1}^{m} p_j \cdot E_j^2$, and $V_\sigma \stackrel{\text{def}}{=} \sum_{j=1}^{m} p_j \cdot \sigma_j^2$.

11. Hierarchical Case: Interval Uncertainty

- *Practical situation:* we only know the intervals $\mathbf{E}_j = [\underline{E}_j, \overline{E}_j]$ and $[\underline{\sigma}_j, \overline{\sigma}_j]$ that contain E_j and σ_j.
- *Mean* E is monotonic in E_j, hence

$$\mathbf{E} = [\underline{E}, \overline{E}] = \left[\sum_{j=1}^{m} p_j \cdot \underline{E}_j, \sum_{j=1}^{m} p_j \cdot \overline{E}_j\right].$$

- *Variance:* the terms V_E and V_σ in the expression for V depend on different variables.
- *Conclusion:* the range $\mathbf{V} = [\underline{V}, \overline{V}]$ of the population variance V is equal to the sum of the ranges:

$$\mathbf{V} = \mathbf{V}_E + \mathbf{V}_\sigma, \text{ where } \mathbf{V}_E = [\underline{V}_E, \overline{V}_E], \ \mathbf{V}_\sigma = [\underline{V}_\sigma, \overline{V}_\sigma].$$

- Due to monotonicity, $\mathbf{V}_\sigma = \left[\sum_{j=1}^{m} p_j \cdot (\underline{\sigma}_j)^2, \sum_{j=1}^{m} p_j \cdot (\overline{\sigma}_j)^2\right]$.
- *Conclusion:* it is sufficient to compute $\mathbf{V}_E$.

12. Formulation of the Problem in Precise Terms

GIVEN:

- an integer $m \geq 1$;
- m numbers $p_j > 0$ for which $\sum_{j=1}^{m} p_j = 1$; and
- m intervals $\mathbf{E}_j = [\underline{E}_j, \overline{E}_j]$.

COMPUTE the range

$$\mathbf{V}_E = \{V_E(E_1, \ldots, E_m) \,|\, E_1 \in \mathbf{E}_1, \ldots, E_m \in \mathbf{E}_m\},$$

where

$$V_E \stackrel{\text{def}}{=} \sum_{j=1}^{m} p_j \cdot E_j^2 - E^2; \quad E \stackrel{\text{def}}{=} \sum_{j=1}^{m} p_j \cdot E_j.$$

13. Analysis of the Problem

- *Fact:* the function V_E is convex.
- *Fact:* the box $\mathbf{E}_1 \times \ldots \times \mathbf{E}_m$ is convex.
- *Known:* a polynomial-time algorithm for computing minima of convex functions on convex sets.
- *Conclusion:* we can compute $\underline{V}_E$ in polynomial time.
- *Computing $\overline{V}_E$:* in general, NP-hard.
- *Proof of NP-hardness:*
  - for $p_1 = \ldots = p_m = \dfrac{1}{m}$, the expression V_E becomes a standard formula for the sample variance V;
  - so, in this case, $\overline{V}_E = \overline{V}$;
  - computing $\overline{V}$ under interval uncertainty is NP-hard;
  - thus, the more general problem of computing $\overline{V}_E$ is also NP-hard.

14. Efficient Algorithm for Computing $\overline{V}_E$

- *Notations:* $\widetilde{E}_j \stackrel{\text{def}}{=} \dfrac{\underline{E}_j + \overline{E}_j}{2}$, $\Delta_j \stackrel{\text{def}}{=} \dfrac{\overline{E}_j - \underline{E}_j}{2}$.
- *Narrowed intervals* $[E_j^-, E_j^+]$, where $E_j^- \stackrel{\text{def}}{=} \widetilde{E}_j - p_j \cdot \Delta_j$ and $E_j^+ \stackrel{\text{def}}{=} \widetilde{E}_j + p_j \cdot \Delta_j$.
- *Case:* no two narrowed intervals are proper subsets of each other, i.e., $[E_j^-, E_j^+] \not\subseteq (E_k^-, E_k^+)$ for all j and k.
- *Efficient $O(m \cdot \log(m))$ algorithm for this case:*
  - First, sort the $\widetilde{E}_1, \ldots, \widetilde{E}_m$ into an increasing sequence $\widetilde{E}_1 \le \widetilde{E}_2 \le \ldots \le \widetilde{E}_m$.
  - Then, for every k from 0 to m, compute the value $V_E^{(k)} = M^{(k)} - (E^{(k)})^2$ of V_E for the vector $\vec{E}^{(k)} = (\underline{E}_1, \ldots, \underline{E}_k, \overline{E}_{k+1}, \ldots, \overline{E}_m)$.
  - Finally, compute $\overline{V}_E$ as the largest of $m+1$ values $V_E^{(0)}, \ldots, V_E^{(m)}$.

15. Number of Computation Steps

- *Known:* sorting requires $O(m \cdot \log(m))$ steps.
- Computing the initial values $M^{(0)}$, $E^{(0)}$, and $V_E^{(0)}$ requires linear time $O(m)$.
- *Reminder:* $V_E^{(k)} = M^{(k)} - (E^{(k)})^2$ is the value for the vector $\vec{E}^{(k)} = (\underline{E}_1, \ldots, \underline{E}_k, \overline{E}_{k+1}, \ldots, \overline{E}_m)$.
- *Transition:* once we have $M^{(k)} = \sum_{j=1}^{m} p_j \cdot \left(E_j^{(k)}\right)^2$ and $E^{(k)} = \sum_{j=1}^{m} p_j \cdot E_j^{(k)}$, we compute, in $O(1)$ steps,

$$M^{(k+1)} = M^{(k)} + p_{k+1} \cdot (\underline{E}_{k+1})^2 - p_{k+1} \cdot (\overline{E}_{k+1})^2,$$

$$E^{(k+1)} = E^{(k)} + p_{k+1} \cdot \underline{E}_{k+1} - p_{k+1} \cdot \overline{E}_{k+1}.$$

- Finding the largest of $V_E^{(0)}, \ldots, V_E^{(m)}$ requires $O(m)$ steps.
- Thus, overall, we need $O(m \cdot \log(m)) + O(m) + m \cdot O(1) + O(m) = O(m \cdot \log(m))$ steps.

16. Conclusion

- *In medicine, in social studies:* it is important to perform statistical data analysis (age $\leftrightarrow$ income, etc.)
- *Problem:* it is possible to extract confidential data.
- *Solution:* replace confidential values x_i with ranges $\mathbf{x}_i$.
- *Example:* $[20, 30]$ instead of the actual age of 26.
- *Computational challenge:* compute the range $\mathbf{C}$ of possible values of the statistic $C(x_1, \ldots, x_n)$ when $x_i \in \mathbf{x}_i$.
- *What was known:* algorithms efficiently computing this range based on the intervals $\mathbf{x}_i$.
- *New challenge:* combine results E_j, V_j $(1 \le j \le m)$ of statistical analysis of different data sets $I_1, \ldots, I_m$.
- *Algorithms known:* for exact data.
- *New result:* combination is also possible when we only have privacy-related interval ranges $\mathbf{E}_j$ and $\mathbf{V}_j$.

17. Acknowledgments

This work was supported in part:

- by NSF grants EAR-0225670 and EIA-0080940,
- by the Texas Department of Transportation grant No. 0-5453, and
- by the Max Planck Institut für Mathematik.