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 at El Paso
El Paso, TX 79968, USA, vladik@utep.edu

Abstract. In many practical situations, it is important to store large amounts of data and to be able to statistically process the data. A large part of the data is confidential, so while we welcome statistical data processing, we do not want to reveal sensitive individual data. If we allow researchers to ask all kinds of statistical queries, this can lead to violation of people's privacy. A sure-proof way to avoid these privacy violations is to store ranges of values (e.g., between 40 and 50 for age) instead of the actual values. This idea solves the privacy problem, but it leads to a computational challenge: traditional statistical algorithms need exact data, but now we only know data with interval uncertainty. In this paper, we describe new algorithms designed for processing such interval data.

1 Interval Approach to Preserving Privacy in Statistical Databases

Need for statistical databases. In many practical situations, it is very useful to collect large amounts of data.

For example, from the data that we collect during a census, we can extract a lot of information about health, mortality, employment in different regions – for different age ranges, and for people from different genders and of different ethnic groups. By analyzing this statistics, we can reveal troubling spots and allocate (usually limited) resources so that the help goes first to social groups that need it most.

Similarly, by gathering data about people's health in a large medical database, we can extract a lot of useful information on how the geographic location, age, and gender affect a person's health. Thus, we can make measures, which are aimed at improving public health, more focused.

Finally, a large statistical database of purchases can help find out what people are looking for, make shopping easier for customers and at the same time, decrease the stores' expenses related to storing unnecessary items.

Need for privacy. Privacy is an important issue in the statistical analysis of human-related data. For example, to check whether in a certain geographic area,

there is a gender-based discrimination, we can use the census data to check, e.g., whether for all people from this area who have the same level of education, there is a correlation between salary and gender. One can think of numerous possible questions of this type related to different sociological, political, medical, economic, and other questions. From this viewpoint, it is desirable to give researchers *ability to perform* whatever *statistical analysis* of this data that is reasonable for their specific research.

On the other hand, we do not want to give them direct access to the raw census data, because a large part of the census data is *confidential*. For example, for most people (those who work in private sector) salary information is confidential. Suppose that a corporation is deciding where to built a new plant and has not yet decided between two possible areas. This corporation would benefit from knowing the average salary of people of needed education level in these two areas, because this information would help them estimate how much it will cost to bring local people on board. However, since salary information is confidential, the company should not be able to know the exact salaries of different potential workers.

The need for privacy is also extremely important for *medical* experiments, where we should be able to make statistical conclusions about, e.g., the efficiency of a new medicine without disclosing any potentially embarrassing details from the individual medical records.

Such databases in which the outside users have cannot access individual records but can solicit statistical information are often called *statistical databases*.

Maintaining privacy is not easy. Maintaining privacy in statistical databases is not easy. Clerks who set up policies on access to statistical databases sometimes erroneously assume that once the records are made anonymous, we have achieved perfect privacy. Alas, the situation is not so easy: even when we keep all the records anonymous, we can still extract confidential information by asking appropriate questions.

Many examples of such extraction can be found in a book by D. Denning [1]. For example, suppose that we are interested in the salary of Dr. X who works for a local company. Dr. X's mailing address can be usually taken from the phone book; from the company's webpage, we can often get his photo and thus find out his race and approximate age. Then, to determine Dr. X's salary, all we need is to ask what is the average salary of all people with a Ph.D. of certain age brackets who live in a small geographical area around his actual home address – if the area is small enough, then Dr. X will be the only person falling under all these categories.

Even if only allow statistical information about salaries $s_1, \ldots, s_q$ when there are at least a certain amount n_0 people within a requested range, we will still be able to reconstruct the exact salaries of all these people. Indeed, for example, we can ask for the number and average salary of all the people who live on Robinson street at houses 1 through 1001, and then we can ask the same question about all the people who live in houses from 1 to 1002. By comparing the two numbers,

we get the average salary of the family living at 1002 Robinson – in other words, we gain the private information that we tried to protect.

In general, we can ask for the average $\frac{s_1 + \ldots + s_q}{q}$, and for several moments of salary (variance, third moment, etc): if we know the values v_j at least q different functions $f_j(s_1, \ldots, s_q)$ of s_i, then we can, in general, reconstruct all these values from the corresponding system of q equations with q unknowns: $f_1(s_1 \ldots, s_q) = v_1, \ldots, f_q(s_1, \ldots, s_q) = v_q$.

At first glance, moments are natural and legitimate statistical characteristics, so researchers would be able to request them, but on the other hand, we do not want them to be able to extract the exact up-to-cent salaries of all the folks leaving in a certain geographical area.

What restriction should we impose on possible statistical queries that would guarantee privacy but restrict research in the least possible way?

What is known. These are anecdotal examples of poorly designed privacy and security, but, as we have mentioned, the problem is indeed difficult: Many seemingly well-designed privacy schemes later turn out out to have unexpected privacy and security problem, and it is known that the problem of finding a privacy-preserving scheme is, in general, NP-hard [1].

Different aspects of the problem of privacy in statistical databases, different proposed solution to this problem, and their drawbacks, are described in [1, 7, 9] (see also references therein).

Interval approach to privacy protection. A sure-proof way to avoid these privacy violations is to store ranges (intervals) of values instead of the actual values. For example, instead of keeping the exact age, we only record whether the age is between 0 and 10, 10 and 20, 20 and 30, etc.

In this case, no matter what statistics we allow, the worst that can happen is that the corresponding ranges will be disclosed. However, in this situation, we do not disclose the original exact values – since these values are not stored in the database in the first place.

2 Related Challenges and Algorithms of Computational Statistics

Related challenges of computational statistics. This idea of storing intervals solves the privacy problem, but it leads to a computational challenge.

Indeed, suppose that we are interested in the value of a statistical characteristic $C(x_1, \ldots, x_n)$ such as population mean $E = \frac{x_1 + \ldots + x_n}{n}$, (biased) population variance $V = \frac{(x_1 - E)^2 + \ldots + (x_n - E)^2}{n}$, covariance, correlation, etc.

Traditional statistical algorithms for computing these characteristics assume that we know the exact values of the samples x_i, y_i, etc. However, in our case,

we do not know these actual values, we only know the *intervals* $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values of these characteristics. Since we do not know the actual values x_i, we cannot compute the exact range of the characteristic C, we can only find the *range* of this characteristic:

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

So, the challenge is: given the characteristic $C(x_1, \ldots, x_n)$ and the intervals $\mathbf{x}_i$, we must compute the corresponding range.

Important comment: what are the statistical properties of these estimations? What *really* interests the user is not a statistical characteristic like population mean E, but rather the *actual* mean of the actual distribution – of which the database contains only a sample. From this viewpoint, a population mean is interesting because it is a good approximation to the actual mean: when the sample size n increases, then with probability 1 the corresponding statistic $C_n \stackrel{\text{def}}{=} C(x_1, \ldots, x_n)$ converges to the actual value c of the desired characteristics, and the difference $d(C_n, c) \stackrel{\text{def}}{=} |C_n - c|$ tends to 0 fast.

In the case of privacy-related interval uncertainty, for every n, we get an *interval* $\mathbf{C}_n \stackrel{\text{def}}{=} C(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. The quality of this interval approximation can be naturally described by estimating the (Hausdorff) distance $d(c, \mathbf{C}_n) = \min\limits_{C \in \mathbf{C}_n} d(c, C)$ between the actual value c and the interval: e.g., this distance is 0 if and only if the interval $\mathbf{C}_n$ contains the desired value c.

Since each actual (hidden) value x_i belongs to the corresponding interval $\mathbf{x}_i$, we have $C_n = C(x_1, \ldots, x_n) \in \mathbf{C}_n$. The distance $d(c, \mathbf{C}_n)$ is defined as a minimum, hence we have $d(c, \mathbf{C}_n) \le d(c, C_n)$. We can therefore conclude that *the rate of convergence for interval estimates is the same (or better) than for the corresponding point estimates.*

Let us now go back to the computational problem.

The resulting computational problem is known – as interval computations. While privacy-related applications are reasonably novel, the problem of computing the range of a known function $f(x_1, \ldots, x_n)$ under interval uncertainty $x_i \in \mathbf{x}_i$ is a well-known and well-studied problem in applications, known as a problem of *interval computations*; see, e.g., [2] (see also [5]).

Indeed, in many real-life problems, we are interesting in the values of some quantity y which are difficult or impossible to measure directly; example include the amount of oil in a given well or a distance to a star. To estimate the value of this quantity y, we measure the values of easier-to-measure quantities $x_1, \ldots, x_n$ related to y in a known way $y = f(x_1, \ldots, x_n)$, and then use the measured values $\widetilde{x}_i$ of these quantities to estimate y as $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

Measurements are never 100% accurate. As a result, the result $\widetilde{x}$ of the measurement is, in general, different from the (unknown) actual value x of the desired quantity. The difference $\Delta x \stackrel{\text{def}}{=} \widetilde{x} - x$ between the measured and the actual values is usually called a *measurement error*.

The manufacturers of a measuring device usually provide us with an upper bound Δ for the (absolute value of) possible errors, i.e., with a bound Δ for which we guarantee that $|\Delta x| \le \Delta$. The need for such a bound comes from the very nature of a measurement process: if no such bound is provided, this means that the difference between the (unknown) actual value x and the observed value $\widetilde{x}$ can be as large as possible.

Since the (absolute value of the) measurement error $\Delta x = \tilde{x} - x$ is bounded by the given bound Δ, we can therefore guarantee that the actual (unknown) value of the desired quantity belongs to the interval $[\widetilde{x} - \Delta, \widetilde{x} + \Delta]$.

In many practical situations, we not only know the interval $[-\Delta, \Delta]$ of possible values of the measurement error; we also know the probability of different values Δx within this interval [6].

In practice, we can determine the desired probabilities of different values of Δx by comparing the results of measuring with this instrument with the results of measuring the same quantity by a standard (much more accurate) measuring instrument. Since the standard measuring instrument is much more accurate than the one use, the difference between these two measurement results is practically equal to the measurement error; thus, the empirical distribution of this difference is close to the desired probability distribution for measurement error.

There are two cases, however, when this determination is not done:

- First is the case of cutting-edge measurements, e.g., measurements in fundamental science. When a Hubble telescope detects the light from a distant galaxy, there is no "standard" (much more accurate) telescope floating nearby that we can use to calibrate the Hubble: the Hubble telescope is the best we have.
- The second case is the case of measurements on the shop floor. In this case, in principle, every sensor can be thoroughly calibrated, but sensor calibration is so costly – usually costing ten times more than the sensor itself – that manufacturers rarely do it.

In both cases, we have no information about the probabilities of Δx; the only information we have is the upper bound on the measurement error.

In this case, after performing a measurement and getting a measurement result $\widetilde{x}_i$, the only information that we have about the actual value x_i of the measured quantity is that it belongs to the interval $\mathbf{x}_i = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. In this situation, for each i, we know the interval $\mathbf{x}_i$ of possible values of x_i, and we need to find the range $\mathbf{y}$ of the function $f(x_1, \ldots, x_n)$ over all possible tuples $x_i \in \mathbf{x}_i$.

Interval computations are sometimes easy. In some cases, it is easy to estimate the desired range. For example, the arithmetic average E is a monotonically increasing function of each of its n variables $x_1, \ldots, x_n$, so its smallest possible value $\underline{E}$ is attained when each value x_i is the smallest possible ($x_i = \underline{x}_i$) and its largest possible value is attained when $x_i = \overline{x}_i$ for all i. In other words, the range

E of E is equal to $[E(\underline{x}_1, \ldots, x_n), E(\overline{x}_1, \ldots, \overline{x}_n)]$, where, $\underline{E} = \dfrac{1}{n} \cdot (\underline{x}_1 + \ldots + \underline{x}_n)$ and $\overline{E} = \dfrac{1}{n} \cdot (\overline{x}_1 + \ldots + \overline{x}_n)$.

Interval computations are, in general, computationally difficult. For more complex functions $C(x_1, \ldots, x_n)$, the problem of computing the range is often more computationally difficult.

For example, it is known that the problem of computing the exact range $\mathbf{V} = [\underline{V}, \overline{V}]$ for the variance V over interval data $x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ is, in general, NP-hard; see, e.g., [3, 4]. Specifically, there is a polynomial-time algorithm for computing $\underline{V}$, but computing $\overline{V}$ is, in general, NP-hard.

Efficient algorithms exist for several practically useful situations. In many practical situations, there are efficient algorithms for computing $\overline{V}$; see, e.g., [3, 4].

For example, an $O(n \cdot \log(n))$ time algorithm exists when no two narrowed intervals $[x_i^-, x_i^+]$, where $x_i^- \stackrel{\text{def}}{=} \widetilde{x}_i - \dfrac{\Delta_i}{n}$ and $x_i^+ \stackrel{\text{def}}{=} \widetilde{x}_i + \dfrac{\Delta_i}{n}$, are proper subsets of one another, i.e., when $[x_i^-, x_i^+] \not\subseteq (x_j^-, x_j^+)$ for all i and j.

Computational problem are usually easier in the privacy case. In the privacy case, intervals correspond to the fixed subdivision of the real line, For such situations, efficient algorithms exist for computing most statistical characteristics; see, e.g., [3, 4].

Important comment: there is a strong need to implement interval-related algorithms in Oracle and SQL. In [3, 4]:

- we have *theoretically* proven that our new algorithms produce correct results in *reasonable* time (usually linear or quadratic), and
- we have shown, by implementing these algorithms in standard programming languages, that the corresponding computation time is also *practically* reasonable.

It is worth mentioning that in most real-world applications of statistical databases, practitioners do not write new code in high-level programming languages, they use systems like Oracle and SQL. So, to promote the use of interval methods, it is important to implement our algorithms in systems such Oracle and SQL.

We hope that the privacy-enhancing character of interval-related algorithms and their efficiency will inspire database designers to incorporate such algorithms in the future versions of database management systems.

3 New Problem: Hierarchical Statistical Analysis under Privacy-Related Interval Uncertainty

Need for hierarchical statistical analysis. In the above description, we assumed that we have all the data in one large database, and we process this large statistical database to estimate the desired statistical characteristics.

In reality, the data is often stored hierarchically. For example, it makes sense to store the census results by states, get averages and standard deviations per state, and then combine these characteristics to get nation-wide statistics.

Formulas behind hierarchical statistical analysis. Let the data values $x_1 \ldots, x_n$ be divided into $m < n$ groups $I_1, \ldots, I_m$. For each group j, we know the frequency p_j of this group (i.e., the number n_j of elements of this group divided by the overall number of records), the average E_j over this group, and the population variance V_j within j-th group.

One can show that in this case, $E = \sum\limits_{j=1}^{m} p_j \cdot E_j$ and $V = V_E + V_\sigma$, where $V_E = \sum\limits_{j=1}^{m} p_j \cdot E_j^2 - E^2$ and $V_\sigma = \sum\limits_{j=1}^{m} p_j \cdot V_j$.

Hierarchical case: situation with interval uncertainty. When we start with values x_i which are only known with interval uncertainty, we end up knowing E_j and V_j also with interval uncertainty. In other words, we only know the intervals $\mathbf{E}_j = [\underline{E}_j, \overline{E}_j]$ and $[\underline{V}_j, \overline{V}_j]$ that contain the actual (unknown) values of E_j and V_j. In such situations, we must find the ranges of the possible values for the population mean E and for the population variance V.

Analysis of the interval problem. The formula that describes the dependence of E on E_j is monotonic in E_j. Thus, we get an explicit formula for the range $[\underline{E}, \overline{E}]$ of the population mean E: $\underline{E} = \sum\limits_{j-1}^{m} p_j \cdot \underline{E}_j$ and $\overline{E} = \sum\limits_{j-1}^{m} p_j \cdot \overline{E}_j$.

Since the terms V_E and V_σ in the expression for V depend on different variables, the range $[\underline{V}, \overline{V}]$ of the population variance V is equal to the sum of the ranges $[\underline{V}_E, \overline{V}_E]$ and $[\underline{V}_\sigma, \overline{V}_\sigma]$ of the corresponding terms: $\underline{V} = \underline{V}_E + \underline{V}_\sigma$ and $\overline{V} = \overline{V}_E + \overline{V}_\sigma$. Due to similar monotonicity, we can find an explicit expression for the range $[\underline{V}_\sigma, \overline{V}_\sigma]$ for V_σ: $\underline{V}_\sigma = \sum\limits_{j=1}^{m} p_j \cdot \underline{V}_j$ and $\overline{V}_\sigma = \sum\limits_{j=1}^{m} p_j \cdot \overline{V}_j$. Thus, to find the range of the population variance V, it is sufficient to find the range of the term V_E. So, we arrive at the following problem:

4 Formulation of the Problem in Precise Terms and Main Result

GIVEN: an integer $m \geq 1$, m numbers $p_j > 0$ for which $\sum\limits_{j=1}^{m} p_j = 1$, 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

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

Main result. Since the function V_E is convex, we can compute its minimum $\underline{V}_E$ on the box $\mathbf{E}_1 \times \ldots \times \mathbf{E}_m$ by using known polynomial-time algorithms for minimizing convex functions over interval domains; see, e.g., [8].

For computing maximum $\overline{V}_E$, even the particular case when all the values p_j are equal $p_1 = \ldots = p_m = 1/m$, is known to be NP-hard. Thus, the more general problem of computing $\overline{V}_E$ is also NP-hard. Let us show that in a reasonable class of cases, there exists a feasible algorithm for computing $\overline{V}_E$.

For each interval $\mathbf{E}_j$, let us denote its midpoint by $\widetilde{E}_j \stackrel{\text{def}}{=} \dfrac{\underline{E}_j + \overline{E}_j}{2}$, and its half-width by $\Delta_j \stackrel{\text{def}}{=} \dfrac{\overline{E}_j - \underline{E}_j}{2}$. In these terms, the j-th interval $\mathbf{E}_j$ takes the form $[\widetilde{E}_j - \Delta_j, \widetilde{E}_j + \Delta_j]$.

In this paper, we consider narrowed intervals $[E_j^-, E_j^+]$, where $E_j^- \stackrel{\text{def}}{=} \widetilde{E}_j - p_j \cdot \Delta_j$ and $E_j^+ \stackrel{\text{def}}{=} \widetilde{E}_j + p_j \cdot \Delta_j$. We show that there exists an efficient $O(m \cdot \log(m))$ algorithm for computing $\overline{V}_E$ for the case when no two narrowed intervals are proper subsets of each other, i.e., when $[E_j^-, E_j^+] \not\subseteq (E_k^-, E_k^+)$ for all j and k.

Algorithm.

- First, we sort the midpoints $\widetilde{E}_1, \ldots, \widetilde{E}_m$ into an increasing sequence. Without losing generality, we can assume that $\widetilde{E}_1 \le \widetilde{E}_2 \le \ldots \le \widetilde{E}_m$.
- Then, for every k from 0 to m, we compute the value $V_E^{(k)} = M^{(k)} - (E^{(k)})^2$ of the quantity V_E for the vector $E^{(k)} = (\underline{E}_1, \ldots, \underline{E}_k, \overline{E}_{k+1}, \ldots, \overline{E}_m)$.
- Finally, we compute $\overline{V}_E$ as the largest of $m+1$ values $V_E^{(0)}, \ldots, V_E^{(m)}$.

To compute the values $V_E^{(k)}$, first, we explicitly compute $M^{(0)}$, $E^{(0)}$, and $V_E^{(0)} = M^{(0)} - E^{(0)}$. Once we computed the values $M^{(k)}$ and $E^{(k)}$, we can compute

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

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

5 Proof

Number of computation steps.

- It is well known that sorting requires $O(m \cdot \log(m))$ steps.
- Computing the initial values $M^{(0)}$, $E^{(0)}$, and $V_E^{(0)}$ requires linear time $O(m)$.
- For each k from 0 to $m-1$, we need a constant number $O(1)$ of steps to compute the next values $M^{(k+1)}$, $E^{(k+1)}$, and $V_E^{(k+1)}$.
- Finally, finding the largest of $m+1$ values $V_E^{(k)}$ also requires $O(m)$ steps.

Thus, overall, we need

$$O(m \cdot \log(m)) + O(m) + m \cdot O(1) + O(m) = O(m \cdot \log(m)) \text{ steps.}$$

Proof of correctness. The function V_E is convex. Thus, its maximum $\overline{V}_E$ on the box $\mathbf{E}_1 \times \ldots \times \mathbf{E}_m$ is attained at one of the vertices of this box, i.e., at a vector $(E_1, \ldots, E_m)$ in which each value E_j is equal to either $\underline{E}_j$ or to $\overline{E}_j$.

To justify our algorithm, we need to prove that this maximum is attained at one of the vectors $E^{(k)}$ in which all the lower bounds $\underline{E}_j$ precede all the upper bounds $\overline{E}_j$. We will prove this by reduction to a contradiction. Indeed, let us assume that the maximum is attained at a vector in which one of the lower bounds follows one of the upper bounds. In each such vector, let i be the largest upper bound index followed by the lower bound; then, in the optimal vector $(E_1, \ldots, E_m)$, we have $E_i = \overline{E}_i$ and $E_{i+1} = \underline{E}_{i+1}$.

Since the maximum is attained for $E_i = \overline{E}_i$, replacing it with $\underline{E}_i = \overline{E}_i - 2\Delta_i$ will either decrease the value of V_E or keep it unchanged. Let us describe how V_E changes under this replacement. Since V_E is defined in terms of M and E, let us first describe how E and M change under this replacement. In the sum for M, we replace $(\overline{E}_i)^2$ with

$$(\underline{E}_i)^2 = (\overline{E}_i - 2\Delta_i)^2 = (\overline{E}_i)^2 - 4 \cdot \Delta_i \cdot \overline{E}_i + 4 \cdot \Delta_i^2.$$

Thus, the value M changes into $M + \Delta_i M$, where

$$\Delta_i M = -4 \cdot p_i \cdot \Delta_i \cdot \overline{E}_i + 4 \cdot p_i \cdot \Delta_i^2.$$

The population mean E changes into $E + \Delta_i E$, where $\Delta_i E = -2 \cdot p_i \cdot \Delta_i$. Thus, the value E^2 changes into $(E + \Delta_i E)^2 = E^2 + \Delta_i(E^2)$, where

$$\Delta_i(E^2) = 2 \cdot E \cdot \Delta_i E + (\Delta_i E)^2 = -4 \cdot p_i \cdot E \cdot \Delta_i + 4 \cdot p_i^2 \cdot \Delta_i^2.$$

So, the variance V changes into $V + \Delta_i V$, where

$$\Delta_i V = \Delta_i M - \Delta_i(E^2) = -4 \cdot p_i \cdot \Delta_i \cdot \overline{E}_i + 4 \cdot p_i \cdot \Delta_i^2 + 4 \cdot p_i \cdot E \cdot \Delta_i - 4 \cdot p_i^2 \cdot \Delta_i^2 =$$

$$4 \cdot p_i \cdot \Delta_i \cdot (-\overline{E}_i + \Delta_i + E - p_i \cdot \Delta_i).$$

By definition, $\overline{E}_i = \widetilde{E}_i + \Delta_i$, hence $-\overline{E}_i + \Delta_i = -\widetilde{E}_i$. Thus, we conclude that $\Delta_i V = 4 \cdot p_i \cdot \Delta_i \cdot (-\widetilde{E}_i + E - p_i \cdot \Delta_i)$. So, the fact that $\Delta_i V \le 0$ means that $E \le \widetilde{E}_i + p_i \cdot \Delta_i = E_i^+$.

Similarly, since the maximum of V_E is attained for $E_{i+1} = \underline{E}_{i+1}$, replacing it with $\overline{E}_{i+1} = \underline{E}_{i+1} + 2\Delta_{i+1}$ will either decrease the value of V_E or keep it unchanged. In the sum for M, we replace $(\underline{E}_{i+1})^2$ with

$$(\overline{E}_{i+1})^2 = (\underline{E}_{i+1} + 2\Delta_{i+1})^2 = (\underline{E}_{i+1})^2 + 4 \cdot \Delta_{i+1} \cdot \underline{E}_{i+1} + 4 \cdot \Delta_{i+1}^2.$$

Thus, the value M changes into $M + \Delta_{i+1} M$, where

$$\Delta_{i+1} M = 4 \cdot p_{i+1} \cdot \Delta_{i+1} \cdot \underline{E}_{i+1} + 4 \cdot p_{i+1} \cdot \Delta_{i+1}^2.$$

The population mean E changes into $E + \Delta_{i+1} E$, where $\Delta_{i+1} E = 2 \cdot p_{i+1} \cdot \Delta_{i+1}$. Thus, the value E^2 changes into $E^2 + \Delta_{i+1}(E^2)$, where

$$\Delta_{i+1}(E^2) = 2 \cdot E \cdot \Delta_{i+1} E + (\Delta_{i+1} E)^2 = 4 \cdot p_{i+1} \cdot E \cdot \Delta_{i+1} + 4 \cdot p_{i+1}^2 \cdot \Delta_{i+1}^2.$$

So, the term V_E changes into $V_E + \Delta_{i+1} V$, where

$$\Delta_{i+1} V = \Delta_{i+1} M - \Delta_{i+1}(E^2) =$$

$$4 \cdot p_{i+1} \cdot \Delta_{i+1} \cdot \underline{E}_{i+1} + 4 \cdot p_{i+1} \cdot \Delta_{i+1}^2 - 4 \cdot p_{i+1} \cdot E \cdot \Delta_{i+1} - 4 \cdot p_{i+1}^2 \cdot \Delta_{i+1}^2 =$$

$$4 \cdot p_{i+1} \cdot \Delta_{i+1} \cdot (\underline{E}_{i+1} + \Delta_{i+1} - E - p_{i+1} \cdot \Delta_{i+1}).$$

By definition, $\underline{E}_{i+1} = \widetilde{E}_{i+1} - \Delta_{i+1}$, hence $\underline{E}_{i+1} + \Delta_{i+1} = \widetilde{E}_{i+1}$. Thus, we conclude that

$$\Delta_{i+1} V = 4 \cdot p_{i+1} \cdot (\widetilde{E}_{i+1} - E - p_{i+1} \cdot \Delta_{i+1}).$$

Since V_E attains maximum at $(E_1, \ldots, E_i, E_{i+1}, \ldots, E_m)$, we have $\Delta_{i+1} V \le 0$, hence $E \ge \widetilde{E}_{i+1} - p_{i+1} \cdot \Delta_{i+1} = E_{i+1}^-$.

We can also change both E_i and E_{i+1} at the same time. In this case, from the fact that V_E attains maximum, we conclude that $\Delta V_E \le 0$.

Here, the change ΔM in M is simply the sum of the changes coming from E_i and E_{i+1}: $\Delta M = \Delta_i M + \Delta_{i+1} M$, and the change in E is also the sum of the corresponding changes: $\Delta E = \Delta_i E + \Delta_{i+1} E$. So, for $\Delta V = \Delta M - \Delta(E^2)$, we get

$$\Delta V = \Delta_i M + \Delta_{i+1} M - 2 \cdot E \cdot \Delta_i E - 2 \cdot E \cdot \Delta_{i+1} E - (\Delta_i E)^2 - (\Delta_{i+1} E)^2 -$$

$$2 \cdot \Delta_i E \cdot \Delta_{i+1} E.$$

Hence,

$$\Delta V = (\Delta_i M - 2 \cdot E_i \cdot \Delta_i E - (\Delta_i E)^2) + (\Delta_{i+1} M - 2 \cdot E_{i+1} \cdot \Delta_{i+1} E - (\Delta_{i+1} E)^2) -$$

$$2 \cdot \Delta E_i \cdot \Delta E_{i+1},$$

i.e., $\Delta V = \Delta_i V + \Delta_{i+1} V - 2 \cdot \Delta_i E \cdot \Delta_{i+1} E$.

We already have expressions for $\Delta_i V$, $\Delta_{i+1} V$, $\Delta_i E$, and $\Delta_{i+1} E$, and we already know that $E_{i+1}^- \le E \le E_i^+$. Thus, we have $D(E) \le 0$ for some $E \in [E_{i+1}^-, E_i^+]$, where

$$D(E) \stackrel{\text{def}}{=} 4 \cdot p_i \cdot \Delta_i \cdot (-E_i^+ + E) + 4 \cdot p_{i+1} \cdot \Delta_{i+1} \cdot (E_{i+1}^- - E) + 8 \cdot p_i \cdot \Delta_i \cdot p_{i+1} \cdot \Delta_{i+1}.$$

Since the narrowed intervals are not subsets of each other, we can sort them in lexicographic order; in which order, midpoints are sorted, left endpoints are sorted, and right endpoints are sorted, hence $E_i^- \le E_{i+1}^-$ and $E_i^+ \le E_{i+1}^+$.

For $E = E_{i+1}^-$, we get

$$D(E_{i+1}^-) = 4 \cdot p_i \cdot \Delta_i \cdot (-E_i^+ + E_{i+1}^-) + 8 \cdot p_i \cdot \Delta_i \cdot p_{i+1} \cdot \Delta_{i+1} =$$

$$4 \cdot p_i \cdot \Delta_i \cdot (-E_i^+ + E_{i+1}^- + 2 \cdot p_{i+1} \cdot \Delta_{i+1}).$$

By definition, $E_{i+1}^- = E_{i+1} - p_{i+1} \cdot \Delta_{i+1}$, hence $E_{i+1}^- + 2 \cdot p_{i+1} \cdot \Delta_{i+1} = E_{i+1}^+$, and $D(E_{i+1}^-) = 4 \cdot p_i \cdot \Delta_i \cdot (E_{i+1}^+ - E_i^+) \ge 0$. Similarly,

$$D(E_i^+) = 4 \cdot p_{i+1} \cdot \Delta_{i+1} \cdot (E_{i+1}^- - E_i^+) \ge 0.$$

The only possibility for both values to be 0 is when interval coincide; in this case, we can easily swap them. In all other cases, all intermediate values $D(E)$ are positive, which contradicts to our conclusion that $D(E) \le 0$. The statement is proven.

6 Auxiliary Result: What if the Frequencies Are Also Only Known with Interval Uncertainty?

Reminder: hierarchical statistical data processing. If we know the frequency of the group j, the mean E_j of the group j, and its second moment $M_j = V_j + E_j^2 = \frac{1}{p_j \cdot n} \cdot \sum_{i \in I_j} x_i^2$, then we can compute the overall mean E and the overall variance as $E = \sum_{j=1}^{m} p_j \cdot E_j$ and $V = \sum_{j=1}^{m} p_j \cdot M_j - E^2$.

Reminder: hierarchical statistical data processing under interval uncertainty. In the above text, we considered the case when the statistical characteristics E_j and V_j corresponding to different groups are known with interval uncertainty, but the frequencies p_j are known exactly.

New situation. In practice, the frequencies p_j may also only be known with interval uncertainty. This may happen, e.g., if instead of the full census we extrapolate data – or if we have a full census and try to take into account that no matter how thorough the census, a certain portion of the population will be missed.

In practice, the values x_i (age, salary, etc.) are usually non-negative. In this case, $E_j \geq 0$. In this section, we will only consider this non-negative case. Thus, we arrive at the new formulation of the problem:

GIVEN: an integer $m \geq 1$, and for every j from 1 to m, intervals $[\underline{p}_j, \overline{p}_j]$, $[\underline{E}_j, \overline{E}_j]$, and $[\underline{M}_j, \overline{M}_j]$ for which $\underline{p}_j \geq 0$, $\underline{E}_j \geq 0$, and $\underline{M}_j \geq 0$.

COMPUTE the range $\mathbf{E} = [\underline{E}, \overline{E}]$ of $E = \sum_{j=1}^{m} p_j \cdot E_j$ and the range $\mathbf{M} = [\underline{M}, \overline{M}]$ of $M = \sum_{j=1}^{m} p_j \cdot M_j - E^2$ under the conditions that $p_j \in [\underline{p}_j, \overline{p}_j]$, $E_j \in [\underline{E}_j, \overline{E}_j]$, $M_j \in [\underline{M}_j, \overline{M}_j]$, and $\sum_{j=1}^{m} p_j = 1$.

Derivation of an algorithm for computing $\underline{E}$. When the frequencies p_j are known, we can easily compute the bounds for E. In the case when p_j are also known with interval uncertainty, it is no longer easy to compute these bounds.

Since E monotonically depends on E_j, the smallest value $\underline{E}$ of E is attained when $E_j = \underline{E}_j$ for all j, so the only problem is to find the corresponding probabilities p_j. Suppose that $p_1, \ldots, p_n$ are minimizing probabilities, and for two indices j and k, we change p_j to $p_j + \Delta p$ (for some small Δp) and p_k to $p_k - \Delta p$. In this manner, the condition $\sum_{j=1}^{m} p_j$ is preserved. After this change, E changes to $E + \Delta E$, where $\Delta E = \Delta p \cdot (\underline{E}_j - \underline{E}_k)$.

Since we start with the values at which E attains its minimum, we must have $\Delta E \geq 0$ for all Δp. If both p_j and p_k are strictly inside the corresponding

intervals, then we can have Δp of all signs hence we should have $\underline{E}_j = \underline{E}_k$. So, excluding this degenerate case, we should have at most one value p_i strictly inside – others are at one of the endpoints.

If $p_j = \underline{p}_j$ and $p_k = \overline{p}_k$, then we can have $\Delta p > 0$, so $\Delta E \geq 0$ implies $\underline{E}_j \geq \underline{E}_k$. So, the values $\underline{E}_j$ for all j for which $p_j = \underline{p}_j$ should be $\leq$ than all the values $\underline{E}_k$ for which $p_k = \overline{p}_k$. This conclusion can be reformulated as follows: if we sort the groups in the increasing order of $\underline{E}_j$, we should get first $\overline{p}_j$ then all $\underline{p}_k$. Thus, it is sufficient to consider only such arrangements of probabilities for which we have $\overline{p}_1, \ldots \overline{p}_{l_0-1}, p_{l_o}, \underline{p}_{l_0+1}, \cdots \underline{p}_m$. The value l_0 can be uniquely determined from the condition that $\sum\limits_{j=1}^{m} p_j = 1$. Thus, we arrive at the following algorithm:

Algorithm for computing $\underline{E}$. To compute $\underline{E}$, we first sort the values $\underline{E}_j$ in increasing order. Let us assume that the groups are already sorted in this order, i.e., that $\underline{E}_1 \leq \underline{E}_2 \leq \ldots \leq \underline{E}_m$.

For every l from 0 to k, we then compute $P_l = \overline{p}_1 + \ldots + \overline{p}_l + \underline{p}_{l+1} + \ldots + \underline{p}_n$ as follows: we explicitly compute the sum P_0, and then consequently compute P_{l+1} as $P_l + (\overline{p}_{l+1} - \underline{p}_{l+1})$. This sequence is increasing. Then, we find the value l_0 for which $P_{l_0} \leq 1 \leq P_{l_0+1}$, and take $\underline{E} = \sum\limits_{j=1}^{l_0-1} \overline{p}_j \cdot \underline{E}_j + p_{l_0} \cdot \underline{E}_{l_0} + \sum\limits_{j=l_0+1}^{m} \underline{p}_j \cdot \underline{E}_j$, where $p_{l_0} = 1 - \sum\limits_{j=1}^{l_0-1} \overline{p}_j - \sum\limits_{j=l_0+1}^{m} \underline{p}_j$.

Computation time. We need $O(m \cdot \log(m))$ time to sort, $O(m)$ time to compute P_0, $O(m)$ time to compute all P_l and hence, to find l_0, and $O(m)$ time to compute $\underline{E}$ – to the total of $O(m \cdot \log(m))$.

Algorithm for computing $\overline{E}$. Similarly, we can compute $\overline{E}$ in time $O(m \cdot \log(m))$. We first sort the values $\overline{E}_j$ in increasing order. Let us assume that the groups are already sorted in this order, i.e., that $\overline{E}_1 \leq \overline{E}_2 \leq \ldots \leq \overline{E}_m$.

For every l from 0 to k, we then compute $P_l = \underline{p}_1 + \ldots + \underline{p}_l + \overline{p}_{l+1} + \ldots + \overline{p}_n$ as follows: we explicitly compute the sum P_0, and then consequently compute P_{l+1} as $P_l - (\overline{p}_{l+1} - \underline{p}_{l+1})$. This sequence is decreasing. Then, we find the value l_0 for which $P_{l_0} \geq 1 \geq P_{l_0+1}$, and take $\overline{E} = \sum\limits_{j=1}^{l_0-1} \underline{p}_j \cdot \overline{E}_j + p_{l_0} \cdot \overline{E}_{l_0} + \sum\limits_{j=l_0+1}^{m} \overline{p}_j \cdot \overline{E}_j$, where $p_{l_0} = 1 - \sum\limits_{j=1}^{l_0-1} \underline{p}_j - \sum\limits_{j=l_0+1}^{m} \overline{p}_j$.

Derivation of an algorithm for computing $\underline{M}$. First, we notice that the minimum is attained when M_j are the smallest ($M_j = \underline{M}_j$) and E_j are the largest ($E_j = \overline{E}_j$). So, the only problem is to find the optimal values of p_j.

Similarly to the case of $\underline{E}$, we add Δp to p_j and subtract Δp from p_k. Since we started with the values at which the minimum is attained we must have $\Delta M \leq 0$,

i.e., $\Delta \cdot [\underline{M}_j - \underline{M}_k - 2E \cdot (\overline{E}_j - \overline{E}_k)] \le 0$. So, at most one value p_j is strictly inside, and if $p_j = \underline{p}_j$ and $p_k = \overline{p}_k$, we must have $\underline{M}_j - \underline{M}_k - 2E \cdot (\overline{E}_j - \overline{E}_k) \le 0$, i.e., $\underline{M}_j - 2E \cdot \overline{E}_j \le \underline{M}_k - 2E \cdot \overline{E}_j$.

Once we know E, we can similarly sort and get the optimal p_j. The problem is that we do not know the value E, and for different values E, we have different orders. The solution to this problem comes from the fact that the above inequality is equivalent to comparing $2E$ with the ratio $\dfrac{\underline{M}_j - \underline{M}_k}{\overline{E}_j - \overline{E}_k}$. Thus, if we compute all n^2 such ratios, sort them, then within each zone between the consequent values, the sorting will be the same. Thus, we arrive at the following algorithm.

Algorithm for computing $\underline{M}$. To compute $\underline{M}$, we first compute all the ratios $\dfrac{\underline{M}_j - \underline{M}_k}{\overline{E}_j - \overline{E}_k}$, sort them, and take Es between two consecutive values in this sorting.

For each such E, we sort the groups according to the value of the expression $\underline{M}_j - 2E \cdot \overline{E}_j$. In this sorting, we select the values $p_j = (\overline{p}_1, \ldots, \overline{p}_{l_0-1}, p_{l_0}, \underline{p}_{l_0+1}, \ldots, \underline{p}_m)$ and pick l_0 in the same manner as when we computed $\underline{E}$. For the resulting p_j, we then compute $\underline{M} = \sum\limits_{j=1}^{m} p_j \cdot \underline{M}_j - \left(\sum\limits_{j=1}^{m} p_j \cdot \overline{E}_j \right)^2$.

Computation time. We need $O(m \cdot \log(m))$ steps for each of m^2 zones, to the (still polynomial) total time $O(m^3 \cdot \log(m)))$.

Algorithm for computing $\overline{M}$. A similar polynomial-time algorithm can be used to compute $\overline{M}$. We first compute all the ratios $\dfrac{\overline{M}_j - \overline{M}_k}{\underline{E}_j - \underline{E}_k}$, sort them, and take Es between two consecutive values in this sorting.

For each such E, we sort the groups according to the value of the expression $\overline{M}_j - 2E \cdot \underline{E}_j$. In this sorting, we select the values $p_j = (\underline{p}_1, \ldots, \underline{p}_{l_0-1}, p_{l_0}, \overline{p}_{l_0+1}, \ldots, \overline{p}_m)$ and pick l_0 in the same manner as when we computed $\overline{E}$. For the resulting p_j, we then compute $\overline{M} = \sum\limits_{j=1}^{m} p_j \cdot \overline{M}_j - \left(\sum\limits_{j=1}^{m} p_j \cdot \underline{E}_j \right)^2$.

7 Conclusion

In medicine, in social studies, etc., it is important to perform statistical data analysis. By performing such an analysis, we can find, e.g., the correlation between the age and income, between the gender and side effects of a medicine, etc. People are often willing to supply the needed confidential data for research purposes. However, many of them are worried that it may be possible to extract their confidential data from the results of statistical data processing – and indeed such privacy violations have occurred in the past.

One way to prevent such privacy violations is to replace the original confidential values with ranges. For example, we divide the set of all possible ages into ranges $[0, 10]$, $[10, 20]$, $[20, 30]$, etc. Then, instead of storing the actual age of 26, we only store the range $[20, 30]$ which contains the actual age value.

This approach successfully protects privacy, but it leads to a computational challenge. For example, if we want to estimate the variance, we can no longer simply compute the statistic such as population variance $V = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2$; since we only know the intervals $[\underline{x}_i, \overline{x}_i]$ of possible values of x_i, we can only compute the range $\mathbf{V}$ of possible values of this statistic when $x_i \in \mathbf{x}_i$. In our previous papers, we designed algorithms efficiently computing this range based on the intervals $\mathbf{x}_i$.

In many real-life situations, several research groups independently perform statistical analysis of different data sets. The more data we use for statistical analysis, the better the estimates. So, it is desirable to get estimates based on the data from all the data sets. In principle, we can combine the data sets and reprocess the combined data. However, this would require a large amount of data processing. It is known that for many statistics (e.g., for population variance), we can avoid these lengthy computations: the statistic for the combined data can be computed based on the results of processing individual data sets.

In this paper, we show that a similar computational simplification is possible when instead of processing the exact values, we process privacy-related interval ranges for these values.

Acknowledgments. This work was supported in part by NSF grants EAR-0225670 and EIA-0080940, by Texas Department of Transportation grant No. 0-5453, and by the Max Planck Institut für Mathematik. The authors are thankful to the editors and to the anonymous referees for valuable suggestions.

References

1. Denning, D.: *Cryptography and Data Security*, Addison-Wesley, Reading, MA, 1982
2. Jaulin, L., et al.: *Applied Interval Analysis*, Springer-Verlag, London, 2001
3. Kreinovich, V., et al.: Interval versions of statistical techniques, with applications to environmental analysis, bioinformatics, and privacy in statistical databases. Journal of Computational and Applied Mathematics **199** (2) (2007) 418–423
4. Kreinovich, V., et al.: Towards combining probabilistic and interval uncertainty in engineering calculations. Reliable Computing **12**(6) (2006) 471–501
5. Nguyen, H.T., Kreinovich, V., Gorodetski, V.I., et al.: Applications of interval-valued degrees of belief. In: *Information Technologies and Intellectual Methods*, Vol. 3, Inst. for Information and Automation, 1999, 6–61 (in Russian)
6. Rabinovich, S.: *Measurement Errors and Uncertainties*, Springer, N.Y., 2005
7. Sweeney, L.: Weaving technology and policy together to maintain confidentiality. Journal of Law, Medicine and Ethics **25** (1997) 98–110
8. Vavasis, S.A.: *Nonlinear Optimization*, Oxford University Press, N.Y., 1991
9. Willenborg, L., De Waal, T.: *Statistical disclosure control in practice*, Springer Verlag, New York, 1996