11-2004

Foundations of Statistical Processing of Set-Valued Data: Towards Efficient Algorithms

Hung T. Nguyen

Vladik Kreinovich
The University of Texas at El Paso, vladik@utep.edu

Gang Xiang

Comments:
UTEP-CS-04-35.
Published in *Proceedings of the Fifth International Conference on Intelligent Technologies InTech'04*, Houston, Texas, December 2-4, 2004.

Foundations of Statistical Processing of Set-Valued Data: Towards Efficient Algorithms

Hung T. Nguyen
Department of Mathematical Sciences
New Mexico State University
Las Cruces, NM 88003, USA
Email: hunguyen@nmsu.edu

Vladik Kreinovich and Gang Xiang
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968, USA
Emails: {vladik,gxiang}@utep.edu

***Abstract*— Due to measurement uncertainty, often, instead of the actual values x_i of the measured quantities, we only know the intervals $\mathbf{x_i} = [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$, where $\widetilde{x}_i$ is the measured value and Δ_i is the upper bound on the measurement error (provided, e.g., by the manufacturer of the measuring instrument). These intervals can be viewed as *random intervals*, i.e., as samples from the interval-valued random variable. In such situations, instead of the exact value of a sample statistic such as covariance $C_{x,y}$, we can only have an interval $\mathbf{C}_{x,y}$ of possible values of this statistic.**

In this paper, we extend the foundations of traditional statistics to statistics of such set-valued data, and describe how this foundation can lead to efficient algorithms for computing the corresponding set-valued statistics.

I. Statistical Estimation: An Important Real-Life Problem

In many real-life situations, we have a large population whose characteristics vary randomly. For example, the humankind is a group of people with different height, weight, etc.; a galaxy is a group of stars with different masses, brightnesses, etc. We would like to know certain characteristics of the corresponding probability distribution: we may be interested in the average height of humans, in the variance of the star mass, in the correlation between the star's mass and brightness, etc.

When the population is small, we can simply compute, e.g., the average height by simply adding all the heights and dividing the result by the total number of people. If the population is large, we cannot compute the average directly. Instead, we take a sample, and estimate the average (or any other desired characteristic) based on the values $x_1, \ldots, x_n$ corresponding to this sample.

Let us recall how this problem is usually formulated in precise mathematical terms; see, e.g., [38], [42].

II. Traditional Statistics: Brief Reminder

We have a sequence of independent identically distributed (i.i.d.) variables $x_1, \ldots, x_n, \ldots$ We are interested in a certain characteristic C of the corresponding probability distribution, e.g., in the expected value $E[f(x)]$ of a given function $f(x)$ of the corresponding random variable.

To estimate this characteristic, we select a *statistic*, i.e., a sequence of functions $s_n(x_1, \ldots, x_n)$ $(n = 1, 2, \ldots)$ for which, under reasonable assumptions,

$$s_n(x_1, \ldots, x_n) \to C;$$

(e.g., $s_n(x_1, \ldots, x_n) \to C$ with probability 1).

For example, to estimate the mean $C = E[x]$ of the distribution, we can take, as the corresponding statistic, the arithmetic average

$$s_n(x_1, \ldots, x_n) = \frac{x_1 + \ldots + x_n}{n}. \qquad (1)$$

For distributions with finite variance

$$\sigma^2 \stackrel{\text{def}}{=} E[(x - E[x])^2] < +\infty,$$

the difference

$$\frac{x_1 + \ldots + x_n}{n} - E[x]$$

is, in the limit $n \to \infty$, normally distributed with 0 average and variance

$$E\left[\left(\frac{x_1 + \ldots + x_n}{n} - E[x]\right)^2\right] = \frac{\sigma^2}{n} \to 0.$$

As a result, we can conclude. e.g., that for large n, we have

$$\left|\frac{x_1 + \ldots + x_n}{n} - E[x]\right| \le 2 \cdot \frac{\sigma}{\sqrt{n}}$$

with probability 95%.

III. Traditional Statistics: Precise Definitions

Notations.

- *In the following text, we will assume that we have fixed a set Ω and a σ-algebra $\mathcal{A}$ of subsets of the set Ω.*
- *By a* probability distribution *P, we will mean a probability measure on $(\Omega, \mathcal{A})$, i.e., a σ-additive function $P : \mathcal{A} \to [0, 1]$ for which $P(\Omega) = 1$.*
- *We will also assume that a class $\mathcal{P}$ of probability distributions is fixed.*

Definition 1. *By a* characteristic *C of a probability distribution, we mean a mapping $C : \mathcal{P} \to \mathbb{R}$.*

Definition 2. *By a* statistic, *we mean a sequence $\{s_n\}_n$ of functions $s_n : \Omega^n \to \mathbb{R}$.*

Definition 3. *We say that a statistic* $\{s_n\}_n$ approximates *a characteristic* C *if for every distribution* $P \in \mathcal{P}$*, we have*

$$s_n(x_1, \ldots, x_n) \to_{n\to\infty} C$$

with probability 1.

Comment. Here, for a given probability measure P on Ω, we define the corresponding probability measure on the set Ω^ω of all infinite sequences $(x_1, \ldots, x_n, \ldots)$, $x_i \in \Omega$ in the usual way.

IV. Examples

- It is known that for $\Omega = \mathbb{R}$, under reasonable assumptions on $\mathcal{P}$, the population average (1) approximates the mean $C = E[x]$.
- It is also known that the population variance

$$s_n(x_1, \ldots, x_n) = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n},$$

where

$$\bar{x} \stackrel{\text{def}}{=} \frac{x_1 + \ldots + x_n}{n},$$

approximates the variance $C = E[(x - E[x])^2]$.
- Similarly, for the case when $\Omega = \mathbb{R}^2$, the population covariance

$$s_n((x_1, y_1) \ldots, (x_n, y_n)) = \frac{(x_1 - \bar{x}) \cdot (y_1 - \bar{y}) + \ldots + (x_n - \bar{x}) \cdot (y_n - \bar{y})}{n}$$

approximates the covariance

$$C = E[(x - E[x]) \cdot (y - E[y])].$$

V. Interval Uncertainty (and, More Generally, Set Uncertainty): A More Realistic Situation

In traditional statistics, we implicitly assume that the values x_i are directly observable. In real life, due to (inevitable) measurement uncertainty (see, e.g., [34]), often, what we actually observe is a *set* X_i that contains the actual (unknown) value of x_i. This phenomenon is called *coarsening*; see, e.g., [16]. Due to coarsening, instead of the actual values x_i, all we know is the sets $X_1, \ldots, X_n, \ldots$ that are known the contain the actual (un-observable) values x_i: $x_i \in X_i$.

In this case, inputs are sets X_i. The set uncertainty in the inputs lead, in general, to a similar set uncertainty in the value of the desired characteristic. As a result, in this case, the estimate for the desired characteristics – i.e., the value $S_n(X_1, \ldots, X_n)$ of the corresponding statistic – should also be a (non-empty) set.

It is reasonable to require that when the values x_i are known exactly, i.e., when $X_i = \{x_i\}$ for some values x_i, then the set corresponding set $S_n(\{x_1\}, \ldots, \{x_n\})$ should become a single-value set.

It is also reasonable to require that when we increase the uncertainty, i.e., replace the original sets X_i with larger sets $X_i' \supseteq X_i$, then the uncertainty in the resulting estimate should also increase:

$$S_n(X_1', \ldots, X_n') \supseteq S_n(X_1, \ldots, X_n).$$

As a result, we arrive at the following definitions.

VI. Set-Valued Statistics: Definitions

Notations. *In the following text, we will assume that we have fixed a class* $\mathcal{S}$ *of subsets of the set* X*. This class should include all one-element sets* $\{x\}$ *where* $x \in \Omega$*.*

Definition 4. *By a* set-valued statistic*, we mean a sequence* $\{S_n\}_n$ *of functions* $S_n : \mathcal{S}^n \to 2^{\mathbb{R}} - \{\emptyset\}$ *that satisfies the following two properties:*

- *when each of the* n *sets* $X_1, \ldots, x_n$ *is a one-element set, then the value of the statistic is also a 1-element set, i.e., for every* n *and for every* $x_1, \ldots, x_n$*,*

$$S_n(\{x_1\}, \ldots, \{x_n\}) = \{s_n(x_1, \ldots, x_n)\} \qquad (2)$$

for some real number $s_n(x_1, \ldots, x_n)$ *depending on* x_i*;*
- *if* $X_1 \subseteq X_1'$*,* $X_2 \subseteq X_2'$*, …, and* $X_n \subseteq X_n'$*, then*

$$S_n(X_1, \ldots, X_n) \subseteq S_n(X_1', \ldots, X_n'). \qquad (3)$$

Comment. According to Definition 4, for every set-valued statistic $\{S_n\}$, there exists a (normal) statistic $\{s_n\}$ that satisfies condition (2). We will say that:

- the statistic $\{s_n\}$ is a *restriction* of the set-valued statistic $\{S_n\}$; and
- the set-valued statistic $\{S_n\}$ is an *extension* of the statistic $\{s_n\}$.

For set-valued statistics – i.e., in the presence of uncertainty – we may not get the exact values of the characteristic even in the limit, but we should be sure that the resulting bounds contains the actual value of the desired characteristic:

Definition 5. *We say that set-valued statistic* $\{S_n\}_n$ approximates *a characteristic* C *if for every distribution* $P \in \mathcal{P}$*, and for every sequence of sets* $X_i \ni x_i$*, we have*

$$C \in \left[\overline{\lim}_{n\to\infty} \underline{S}_n(X_1, \ldots, X_n), \underline{\lim}_{n\to\infty} \overline{S}_n(X_1, \ldots, X_n)\right]$$

with probability 1, where, for every set $S \subset \mathbb{R}$ *of real numbers,* $\underline{S}$ *means its infimum and* $\overline{S}$ *means its supremum.*

Comment. In particular, when the set X is an interval, i.e., $X = [\underline{x}, \overline{x}]$, then the values $\underline{X}$ and $\overline{X}$ coincide with the endpoints of this interval, i.e., $\underline{X} = \underline{x}$ and $\overline{X} = \overline{x}$.

VII. Important Comment: Resulting Definitions of Set-Valued Statistics Differ from the Definitions Motivated by Statistics of Shapes

Uncertainty is not the only reason why we may want to consider set-valued statistics; see, e.g., [14], [27]. For example, we may be interested in statistics of shapes. In this case, the

sets $X_1, \ldots, X_n \ldots$ describe different shapes: e.g., different skull shapes of different proto-humans. In such situation, we may be interested in finding out what is the average shape, what is the range of the deviation between the actual shape and the average shape, etc.; see, e.g., [1], [3], [8], [13], [24], [29], [36], [39], [40], [41], [45], [46].

Let us show that, in general, shape-related problems lead to different set-valued statistics than uncertainty-related problems. This difference can be illustrated on the simplest 1-D example, when all the sets are identical, e.g., $X_1 = X_2 = \ldots = X_n = \ldots = [0,1]$.

For *shapes*, this equality means that all the objects have exactly the same shape. So:

- the average shape is exactly the same shape $[0,1]$, and
- the actual shapes do not deviate from the average shape, so the variance – the measure of such deviation – should be identically equal to 0.

In case of *uncertainty*, however, we get a completely different result, because it is possible that the actual values are different: e.g., we could have $x_1 = x_3 = x_5 = \ldots = x_{2k+1} = \ldots = 0$ and $x_2 = x_4 = x_6 = \ldots = x_{2k} = \ldots = 1$ In this case, the average $\bar{x}$ is 0.5. The square $(x_i - \bar{x})^2$ of the difference between the actual value x_i and the average $\bar{x}$ is thus always equal to 1/4, so the variance can be equal to 1/4.

So, in case of uncertainty, we can have variance equal to 0 – if all the values $x_i \in X_i$ are identical – or we could have the variance equal to 1/4. Thus, in case of uncertainty, we have an entire interval of possible values of variance – in contrast to the shape-motivated definitions, where the variance is always equal to 0.

VIII. Additional Requirement for Set-Valued Statistics: That They Are the Least Uncertain

For a traditional statistic, the main request is that it should converge to the desired characteristic. The faster it converges, the better the statistic.

For a set-valued statistic, we also have a similar problem:

- just like the first estimate $s_1(x_1)$ for, e.g., average may be far away from the actual mean,
- similarly, the first interval-valued estimate $S_1(X_1)$ for the average may not include the actual mean.

It is therefore important to know how fast we get an intervals $S_n(X_1, \ldots, X_n)$ that actually contains the desired value C.

Here lies the difference between the traditional (number-valued) statistics and the set-valued statistics:

- in the traditional statistics, once the estimate $s_n(x_1, \ldots, x_n)$ is sufficiently close to the desired characteristic C, we have achieved our objective – of estimating C;
- on the other hand, for a set-valued statistic, if we can guarantee that C belongs to the interval $S_n(X_1, \ldots, X_n)$, it may mean that we have achieved our objective, but it may also mean that the estimate provided by this statistic is too wide.

If we simply artificially "expand" each interval $S_n(X_1, \ldots, X_n)$ by a factor of two, i.e.:

- represent each interval $S = [\underline{S}, \overline{S}]$ as $[\widetilde{S} - \Delta, \widetilde{S} + \Delta]$, where
$$\widetilde{S} = \frac{\underline{S} + \overline{S}}{2} \text{ and } \Delta = \frac{\overline{S} - \underline{S}}{2},$$
- and then replace it with a twice wider interval
$$[\widetilde{S} - 2\Delta, \widetilde{S} + 2\Delta] \supset S$$
with the same center,

then we will probably get the actual value C into the interval S_n faster – but the resulting estimate for C will be twice wider – hence, twice less accurate than before.

To avoid this artificial expansion, it is reasonable to require that our set-valued statistics be *the least uncertain* in the following precise sense.

Definition 6.

- *We say that a set-valued statistic $\{S_n\}$ is* less uncertain *that the statistic $\{S'_n\}$ if the following two conditions are satisfied:*
  - *for every n and for all possible sets $X_1, \ldots, X_n \in \mathcal{S}$, we have*
  $$S_n(X_1, \ldots, X_n) \subseteq S'_n(X_1, \ldots, X_n);$$
  - *there exists an integer n and sets $X_1, \ldots, X_n \in \mathcal{S}$ for which*
  $$S_n(X_1, \ldots, X_n) \neq S'_n(X_1, \ldots, X_n).$$
- *We say that a set-valued statistics $\{S_n\}$ is* the least uncertain *if no other set-valued statistic is less uncertain than $\{S_n\}$.*

IX. Main Result: Formulation

Our main result is that we can always replace our original statistic with the least uncertain one and retain the property of approximating the desired characteristic:

Proposition. *Let $\{S_n\}$ be a set-valued statistic that approximates a characteristic C and that is not the least uncertain. Then, there exists a statistic $\{T_n\}$ that satisfies the following three properties:*

- *$\{T_n\}$ is less uncertain than $\{S_n\}$,*
- *$\{T_n\}$ is the least uncertain, and*
- *$\{T_n\}$ approximates C.*

X. Main Result: Proof

1°. We will show that the desired statistic is as follows:

$$T_n(X_1, \ldots, X_n) = \{s_n(x_1, \ldots, x_n) \,|\, x_1 \in X_1, \ldots, x_n \in X_n\} \qquad (4)$$

where $s_n(x_1, \ldots, x_n)$ is a restriction of the original set-valued statistic $\{S_n\}$.

It is easy to check that the expression (4) satisfies the properties required by Definition 4, i.e., it is indeed a set-valued statistic.

2°. Let us first prove that for every n and for all possible sets $X_1, \ldots, X_n \in \mathcal{S}$, we have

$$T_n(X_1, \ldots, X_n) \subseteq S_n(X_1, \ldots, X_n).$$

Indeed, let t be an arbitrary element of the set $T_n(X_1, \ldots, X_n)$; let us show that $t \in S_n(X_1, \ldots, X_n)$.

By definition of the set-valued statistic T_n (formula (4)), the fact that $t \in T_n(X_1, \ldots, X_n)$ means that there exist elements $x_1 \in X_1, \ldots, x_n \in X_n$ for which

$$t = s_n(x_1, \ldots, x_n). \tag{5}$$

Since

$$x_1 \in X_1, \ldots, x_n \in X_n,$$

we have

$$\{x_1\} \subseteq X_1, \ldots, \{x_n\} \subseteq X_n.$$

Then, due to Definition 4 of a set-valued statistic, formula (3) is true, according to which

$$S_n(\{x_1\}, \ldots, \{x_n\}) \subseteq S_n(X_1, \ldots, X_n). \tag{6}$$

By definition of a set-valued statistic (formula (2)), we have

$$S_n(\{x_1\}, \ldots, \{x_n\}) = \{s_n(x_1, \ldots, x_n)\},$$

so the formula (6) turns into

$$\{s_n(x_1, \ldots, x_n)\} \subseteq S_n(X_1, \ldots, X_n),$$

or, equivalently,

$$s_n(x_1, \ldots, x_n) \in S_n(X_1, \ldots, X_n). \tag{7}$$

By our choice of $x_1, \ldots, x_n$ (formula (5)), we have $s_n(x_1, \ldots, x_n) = t$, so (7) turns into $t \in S_n(X_1, \ldots, X_n)$.

The statement is proven.

3°. Because of the result proven in Part 2° of this proof, we can conclude that:

- either $\{T_n\}$ is less uncertain than $\{S_n\}$,
- or $\{T_n\}$ coincides with $\{S_n\}$.

We assumed that the original set-valued statistic $\{S_n\}$ is not the least uncertain. Thus, if we prove that $\{T_n\}$ *is* the least uncertain, then we will be able to exclude the second case and conclude that $\{T_n\}$ is less uncertain than $\{S_n\}$.

Let us prove it.

4°. Let us first prove that the set-valued statistic $\{T_n\}$ is the least uncertain.

We need to prove that no other statistic $\{U_n\}$ is less uncertain than $\{T_n\}$. Specifically, we will prove that if $\{U_n\}$ is a set-valued statistic for which, for every n and for all possible sets $X_1, \ldots, X_n \in \mathcal{S}$, we have

$$U_n(X_1, \ldots, X_n) \subseteq T_n(X_1, \ldots, X_n), \tag{8}$$

then $\{U_n\}$ is the same statistic as $\{U_n\}$, i.e.,

$$U_n(X_1, \ldots, X_n) = T_n(X_1, \ldots, X_n)$$

for all n and for all X_i.

Indeed, let $\{U_n\}$ satisfy the property (8). In particular, for every $x_1, \ldots, x_n \in \Omega$, the formula (8) holds for $X_1 = \{x_1\}$, …, $X_n = \{x_n\}$:

$$U_n(\{x_1\}, \ldots, \{x_n\}) \subseteq T_n(\{x_1\}, \ldots, \{x_n\}). \tag{9}$$

For these one-element sets, according to the definition of a set-valued statistic, we have

$$U(\{x_1\}, \ldots, \{x_n\}) = \{u_n(x_1, \ldots, x_n)\},$$

where u_n is a restriction of $\{U_n\}$. By definition of $\{T_n\}$ (formula (4)), we know that

$$T(\{x_1\}, \ldots, \{x_n\}) = \{s_n(x_1, \ldots, x_n)\}.$$

Thus, the formula (9) means

$$u_n(x_1, \ldots, x_n) = s_n(x_1, \ldots, x_n)$$

for all $x_1, \ldots, x_n$ – i.e., that the set-valued statistic $\{U_n\}$ has the same restriction as the original set-valued statistic $\{S_n\}$ and as the newly constructed set-valued statistic $\{T_n\}$.

Similarly to Part 2° of the proof, we can now show that for every integer n and for every sequence of sets $X_1, \ldots, X_n$, we have $T_n(X_1, \ldots, X_n) \subseteq U_n(X_1, \ldots, X_n)$. Since we assumed that $U_n(X_1, \ldots, X_n) \subseteq T_n(X_1, \ldots, X_n)$ (formula (8)), we conclude that $T_n(X_1, \ldots, X_n) = U_n(X_1, \ldots, X_n)$ for all n and for all sets $X_1, \ldots, X_n$, i.e., that the statistics $\{T_n\}$ and $\{U_n\}$ indeed coincide: $\{U_n\} = \{U_n\}$.

Thus, we have proven that the set-valued statistic $\{T_n\}$ is the least uncertain. In view of Part 3° of this proof, we can thus conclude that $\{T_n\}$ is less uncertain than $\{S_n\}$.

5°. To complete the proof, it is sufficient to prove that the new set-valued statistic $\{T_n\}$ approximates the desired characteristic C. We will prove it in two steps:

- first, we will prove that the statistic $\{s_n\}$ – the restriction of the original set-valued statistic $\{S_n\}$ – approximates C;
- from this, we will conclude that the new set-valued statistic $\{T_n\}$ also approximates C.

6°. Let us first prove that the restriction $\{s_n\}$ of the original set-valued statistic $\{S_n\}$ approximates C.

By Definition 5, if we take $X_i = \{x_i\}$, then we conclude that with probability 1,

$$C \in \left[\overline{\lim}\ \underline{S}_n(\{x_1\}, \ldots, \{x_n\}), \underline{\lim}\ \overline{S}_n(\{x_1\}, \ldots, \{x_n\})\right]. \tag{10}$$

For one-element sets,

$$S_n(\{x_1\}, \ldots, \{x_n\}) = \{s_n(x_1, \ldots, x_n)\}.$$

Since the set $S_n(\{x_1\}, \ldots, \{x_n\})$ is a one-element set, its infimum and supremum both coincide with this same element:

$$\underline{S}_n(\{x_1\}, \ldots, \{x_n\}) = \overline{S}_n(\{x_1\}, \ldots, \{x_n\}) = s_n(x_1, \ldots, x_n).$$

Thus, the condition (10) turns into:

$$C \in \left[\overline{\lim}\ s_n(x_1,\ldots,x_n), \underline{\lim}\ s_n(x_1,\ldots,x_n)\right]. \qquad (11)$$

Since the interval is non-empty, we conclude that

$$\overline{\lim}\ s_n(x_1,\ldots,x_n) \le \underline{\lim}\ s_n(x_1,\ldots,x_n).$$

Since we always have

$$\underline{\lim}\ s_n(x_1,\ldots,x_n) \le \overline{\lim}\ s_n(x_1,\ldots,x_n),$$

we thus conclude that

$$\overline{\lim}\ s_n(x_1,\ldots,x_n) = \underline{\lim}\ s_n(x_1,\ldots,x_n),$$

and therefore, that the sequence $s_n(x_1,\ldots,x_n)$ converges:

$$\overline{\lim}\ s_n(x_1,\ldots,x_n) = \underline{\lim}\ s_n(x_1,\ldots,x_n) =$$

$$\lim\ s_n(x_1,\ldots,x_n). \qquad (12)$$

In view of the formula (12), the condition (11) takes the form

$$C \in [\lim\ s_n(x_1,\ldots,x_n), \lim\ s_n(x_1,\ldots,x_n)],$$

i.e., the form $C = \lim\ s_n(x_1,\ldots,x_n)$.

Thus, by Definition 3, the restriction $\{s_n\}$ of the original set-valued statistic $\{S_n\}$ approximates the characteristic C. The statement is proven.

7°. Let us now prove that the new set-valued statistic $\{T_n\}$ approximates the desired characteristic C.

Indeed, let x_i be a sequence of values and $x_i \in X_i$. Due to Part 6° of this proof, the statistic $\{s_n\}$ approximates C hence $s_n(x_1,\ldots,x_n) \to C$ with probability 1.

By definition of T_n (formula (4)), we have $s_n(x_1,\ldots,x_n) \in T_n(X_1,\ldots,X_n)$; this means that

$$\underline{T}_n(X_1,\ldots,X_n) \le s_n(x_1,\ldots,x_n) \le$$

$$\overline{T}_n(X_1,\ldots,X_n). \qquad (13)$$

From $\underline{T}_n(X_1,\ldots,X_n) \le s_n(x_1,\ldots,x_n)$, we conclude that

$$\overline{\lim}\ \underline{T}_n(X_1,\ldots,X_n) \le$$

$$\overline{\lim}\ s_n(x_1,\ldots,x_n) = \lim\ s_n(x_1,\ldots,x_n) = C. \qquad (14)$$

Similarly, from (13), we conclude that

$$s_n(x_1,\ldots,x_n) \le \overline{T}_n(X_1,\ldots,X_n)$$

and therefore, that

$$\underline{\lim}\ s_n(x_1,\ldots,x_n) = C \le \underline{\lim}\ \overline{T}_n(X_1,\ldots,X_n). \qquad (15)$$

The inequalities (14) and (15) mean that

$$C \in \left[\overline{\lim}\ \underline{T}_n(X_1,\ldots,X_n), \underline{\lim}\ \overline{T}_n(X_1,\ldots,X_n)\right],$$

i.e., according to Definition 5, that the set-valued statistic $\{T_n\}$ approximates C.

The statement is proven, and so is our main result.

XI. Resulting Computational Problem and How to Solve It

According to our result, the optimal (least uncertain) set-valued statistic (4) is uniquely determined by the corresponding traditional statistic $s_n(x_1,\ldots,x_n)$ – its restriction to the case when we know the exact values x_i (i.e., when $X_i = \{x_i\}$).

Specifically, once the statistic $s_n(x_1,\ldots,x_n)$ is known, we can describe the corresponding optimal statistic $\{T_n\}$ as follows: it assigns, to every sets $X_1,\ldots,X_n$, the range of

$$\mathbf{s}_n(X_1,\ldots,X_n) \stackrel{\text{def}}{=}$$

$$\{s_n(x_1,\ldots,x_n) \,|\, x_1 \in X_1,\ldots,x_n \in X_n\}$$

of the statistic $s_n(x_1,\ldots,x_n)$ when $x_i \in X_i$.

So, from the computational viewpoint, the problem of computing the value of the set-valued statistic $S_n(X_1,\ldots,X_n)$ can be reformulated as the problem as the problem of computing the range of a given function $s_n(x_1,\ldots,x_n)$ when $x_i \in X_i$.

For a practically important case when the sets X_i are intervals, the problem of computing the range is one of the main problems solved by *interval computations* [17], [18], [19], [30]. It is known that in general, the problem is NP-hard (see, e.g., [22]).

Comment. NP-hard means, crudely speaking, that no feasible algorithm can compute the exact range of $s_n(x_1,\ldots,x_n)$ for all possible functions $s_n(x_1,\ldots,x_n)$ (even for all possible polynomials $s_n(x_1,\ldots,x_n)$) and for all possible intervals $X_1,\ldots,X_n$.

For a specific case when s_n is a statistic, this problem been described, in a general context, in the monographs [26], [43]; for further developments, see, e.g., [2], [4], [5], [6], [7], [9], [10], [11], [12], [15], [20], [21], [23], [25], [28], [31], [32], [33], [35], [37], [44] and references therein.

For example, for the population average

$$E(x_1,\ldots,x_n) = \frac{x_1+\ldots+x_n}{n},$$

this function s_n is an increasing function of each of its variables, hence its range is equal to $[\underline{E}, \overline{E}]$, where

$$\underline{E} = \frac{\underline{x}_1+\ldots+\underline{x}_n}{n}, \text{ and } \overline{E}_x = \frac{\overline{x}_1+\ldots+\overline{x}_n}{n}.$$

For the population variance V, all known algorithms lead to an excess width. Specifically, there exist feasible algorithms for computing the lower endpoint $\underline{V}$ (see, e.g., [10]), but in general, the problem of computing $\overline{V}$ is NP-hard [10].

It is also known that in some practically important cases, feasible algorithms for computing $\overline{V}$ are possible. One such practically useful case is when the measurement accuracy is good enough so that we can tell that the different measured values $\widetilde{x}_i$ are indeed different – e.g., the corresponding intervals $\mathbf{x}_i$ do not intersect. In this case, there exists a quadratic-time algorithm for computing $\overline{V}$; see, e.g., [10].

There exist other practically useful cases when efficient aslgorithms are possible; see above references.

XII. Acknowledgments

This work was supported in part by NASA under cooperative agreement NCC5-209, by the Future Aerospace Science and Technology Program (FAST) Center for Structural Integrity of Aerospace Systems, effort sponsored by the Air Force Office of Scientific Research, Air Force Materiel Command, USAF, under grant F49620-00-1-0365, by NSF grants EAR-0112968, EAR-0225670, and EIA-0321328, by the Army Research Laboratories grant DATM-05-02-C-0046, and by the NIH grant 3T34GM008048-20S1.

The authors are thankful to the anonymous referees for valuable suggestions.

References

[1] A. J. Baddeley and I. Molchanov, "Averaging of random sets based on their distance functions", *Journal of Mathematical Imaging and Vision*, 1996.

[2] J. B. Beck, V. Kreinovich, and B. Wu, "Interval-Valued and Fuzzy-Valued Random Variables: From Computing Sample Variances to Computing Sample Covariances", In: M. Lopez, M. A. Gil, P. Grzegorzewski, O. Hrynewicz, and J. Lawry (eds.), *Soft Methodology and Random Information Systems*, Springer-Verlag, 2004, pp. 85-92.

[3] J. K. Beddow and T. Meloy, *Testing and characterization of powder and fine particles*, Heyden & Sons, London, 1980.

[4] D. Berleant, "Automatically verified arithmetic with both intervals and probability density functions", *Interval Computations*, 1993, No. 2, pp. 48–70.

[5] D. Berleant, "Automatically verified arithmetic on probability distributions and intervals", In: R. B. Kearfott and V. Kreinovich, eds., *Applications of Interval Computations*, Kluwer, Dordrecht, 1996.

[6] D. Berleant and C. Goodman-Strauss, "Bounding the results of arithmetic operations on random variables of unknown dependency using intervals", *Reliable Computing*, 1998, Vol. 4. No. 2, pp. 147–165.

[7] D. Berleant, L. Xie, and J. Zhang, "Statool: A Tool for Distribution Envelope Determination (DEnv), an Interval-Based Algorithm for Arithmetic on Random Variables", *Reliable Computing*, 2003, Vol. 9, No. 2, pp. 91–108.

[8] F. L. Bookstein, *Morphometric tools for landmark data: geometry and biology*, Cambridge University Press, Cambridge, 1991.

[9] S. Ferson, *RAMAS Risk Calc 4.0: Risk Assessment with Uncertain Numbers*, CRC Press, Boca Raton, Florida, 2002.

[10] S. Ferson, L. Ginzburg, V. Kreinovich, and J. Lopez, "Absolute Bounds on the Mean of Sum, Product, etc.: A Probabilistic Extension of Interval Arithmetic", *Extended Abstracts of the 2002 SIAM Workshop on Validated Computing*, Toronto, Canada, May 23–25, 2002, pp. 70–72.

[11] S. Ferson, L. Ginzburg, V. Kreinovich, H. T. Nguyen, and S. A. Starks, "Uncertainty in Risk Analysis: Towards a General Second-Order Approach Combining Interval, Probabilistic, and Fuzzy Techniques", *Proceedings of FUZZ-IEEE'2002*, Honolulu, Hawaii, May 12–17, 2002, Vol. 2, pp. 1342–1347.

[12] S. Ferson, D. Myers, and D. Berleant, *Distribution-free risk analysis: I. Range, mean, and variance*, Applied Biomathematics, Technical Report, 2001.

[13] L. A. Galway, *Statistical analysis of star-shaped sets*, Ph.D. Thesis, Carnegie-Mellon University, 1987.

[14] J. Goutsias, R. P. S. Mahler, and H. T. Nguyen (eds.), *Random Sets: Theory and Applications*, Springer-Verlag, N.Y., 1997.

[15] L. Granvilliers, V. Kreinovich, and N. Mueller, "Novel Approaches to Numerical Software with Result Verification", In: R. Alt, A. Frommer, R. B. Kearfott, and W. Luther (eds.), *Numerical Software with Result Verification*, International Dagstuhl Seminar, Dagstuhl Castle, Germany, January 19–24, 2003, Revised Papers, Springer Lectures Notes in Computer Science, 2004, Vol. 2991, pp. 274–305.

[16] D. F. Heitjan and D. B. Rubin, "Ignorability and coarse data", *Ann. Stat.*, 1991, Vol. 19, No. 4, pp. 2244–2253.

[17] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied Interval Analysis*, Springer-Verlag, Berlin, 2001.

[18] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Kluwer, Dordrecht, 1996.

[19] R. B. Kearfott and V. Kreinovich (eds.), *Applications of Interval Computations*, Kluwer, Dordrecht, 1996.

[20] V. Kreinovich, "Probabilities, Intervals, What Next? Optimization Problems Related to Extension of Interval Computations to Situations with Partial Information about Probabilities", *Journal of Global Optimization*, 2004, Vol. 29, No. 3, pp. 265–280.

[21] V. Kreinovich, S. Ferson, and L. Ginzburg, "Exact Upper Bound on the Mean of the Product of Many Random Variables With Known Expectations", *Reliable Computing*, 2003, Vol. 9, No. 6, pp. 441–463.

[22] V. Kreinovich, A. Lakeyev, J. Rohn, and P. Kahl, *Computational Complexity and Feasibility of Data Processing and Interval Computations*, Kluwer, Dordrecht, 1997.

[23] V. Kreinovich and L. Longpré, "Computational complexity and feasibility of data processing and interval computations, with extension to cases when we have partial information about probabilities", In: V. Brattka, M. Schröder, K. Weihrauch, and N. Zhong, *Proceedings of the Conference on Computability and Complexity in Analysis CCA'2003*, Cincinnati, Ohio, USA, August 28–30, 2003, pp. 19–54.

[24] V. Kreinovich and I. Molchanov, "How to define an average of several sets?", *Geombinatorics*, 1998, Part I, Vol. 7, No. 4, pp. 123–131; Part II, 1998, Vol. 8, No. 1, pp. 160–165.

[25] V. Kreinovich, G. N. Solopchenko, S. Ferson, L. Ginzburg, and R. Aló, "Probabilities, intervals, what next? Extension of interval computations to situations with partial information about probabilities", *Proceedings of the 10th IMEKO TC7 International Symposium on Advances of Measurement Science*, St. Petersburg, Russia, June 30–July 2, 2004, Vol. 1, pp. 137–142.

[26] V. P. Kuznetsov, *Interval Statistical Models*, Radio i Svyaz, Moscow, 1991 (in Russian).

[27] S. Li, Y. Ogura, and V. Kreinovich, *Limit Theorems and Applications of Set Valued and Fuzzy Valued Random Variables*, Kluwer, Dordrecht, 2002.

[28] W. A. Lodwick and K. D. Jamison, "Estimating and Validating the Cumulative Distribution of a Function of Random Variables: Toward the Development of Distribution Arithmetic", *Reliable Computing*, 2003, Vol. 9, No. 2, pp. 127–141.

[29] I. Molchanov, "Statistical problems for random sets", in: [14].

[30] R. E. Moore, *Methods and Applications of Interval Analysis*, SIAM, Philadelphia, 1979.

[31] P. Nivlet, F. Fournier, and J. Royer, "A new methodology to account for uncertainties in 4-D seismic interpretation", *Proc. 71st Annual Int'l Meeting of Soc. of Exploratory Geophysics SEG'2001*, San Antonio, TX, September 9–14, 2001, pp. 1644–1647.

[32] P. Nivlet, F. Fournier, and J. Royer, "Propagating interval uncertainties in supervised pattern recognition for reservoir characterization", Proc. 2001 Society of Petroleum Engineers Annual Conf. SPE'2001, New Orleans, LA, September 30–October 3, 2001, paper SPE-71327.

[33] R. Osegueda, V. Kreinovich, L. Potluri, and R. Aló, "Non-Destructive Testing of Aerospace Structures: Granularity and Data Mining Approach", In: *Proc. FUZZ-IEEE'2002*, Honolulu, HI, May 12–17, 2002, Vol. 1, pp. 685–689.

[34] S. Rabinovich, *Measurement Errors: Theory and Practice*, American Institute of Physics, New York, 1993.

[35] H. Regan, S. Ferson, and D. Berleant, "Equivalence of five methods for bounding uncertainty", *International Journal of Approximate Reasoning*, 2004, Vol. 36, No. 1, pp. 1–30.

[36] A. Rosenfeld and J. L. Pfaltz, "Distance functions on digital pictures", *Pattern Recognition*, 1968, Vol. 1, pp. 33–61.

[37] N. C. Rowe, "Absolute bounds on the mean and standard deviation of transformed data for constant-sign-derivative transformations", *SIAM Journal of Scientific Statistical Computing*, 1988, Vol. 9, pp. 1098–1113.

[38] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2004.

[39] D. Stoyan and I. Molchanov, "Set-valued means of random particles", *Journal of Mathematical Imaging and Vision*, 1996.

[40] D. Stoyan and H. Stoyan, *Fractals, random shapes, and point fields*, Wiley, Chichester, 1994.

[41] O. Yu. Vorob'ev, O. Yu. *Mean-measure modeling*, Nauka Publ., Moscow, 1984 (in Russian).

[42] H. M. Wadsworth Jr., *Handbook of statistical methods for engineers and scientists*, McGraw-Hill, N.Y., 1990.

[43] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman & Hall, N.Y., 1991.

[44] R. Williamson and T. Downs, “Probabilistic arithmetic I: numerical methods for calculating convolutions and dependency bounds”, *International Journal of Approximate Reasoning*, 1990, Vol. 4, pp. 89–158.

[45] H. Ziezold, H. “On expected figures in the plane”, In: *Geobild’89, Math, Research Series* Vol. 51, Akademie-Verlag, Berlin, 1989, pp. 105–110.

[46] H. Ziezold, “Mean figures and mean shapes applied to biological figures and shape distributions in the plane”, *Biom. J.*, 1994, Vol. 36, pp. 491–510.