11-2015

Why the Range of a Robust Statistic Under Interval Uncertainty Is Often Easier to Compute

Olga Kosheleva
The University of Texas at El Paso, olgak@utep.edu

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

Comments:
Technical Report: UTEP-CS-15-87
To appear in *Journal of Innovative Technology and Education*

Why the Range of a Robust Statistic Under Interval Uncertainty Is Often Easier to Compute

Olga Kosheleva and Vladik Kreinovich
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
olgak@utep.edu, vladik@utep.edu

Abstract

In statistical analysis, we usually use the observed sample values x_1, $\ldots$, x_n to compute the values of several statistics $v(x_1, \ldots, x_n)$ – such as sample mean, sample variance, etc. The usual formulas for these statistics implicitly assume that we know the exact values $x_1, \ldots, x_n$. In practice, the sample values $\widetilde{x}_1, \ldots, \widetilde{x}_n$ come from measurements and are, thus, only approximations to the actual (unknown) values $x_1, \ldots, x_n$ of the corresponding quantity. Often, the only information that we have about each measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$ is the upper bound Δ_i on the measurement error: $|\Delta x_i| \le \Delta_i$. In this case, the only information about each actual value x_i is that it belongs to the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$. It is therefore desirable to compute the range of each given statistic $v(x_1, \ldots, x_n)$ over these intervals. It is known that often, estimating the range of a robust statistic (e.g., median) is computationally easier than estimating the range of its traditional equivalent (e.g., mean). In this paper, we provide a qualitative explanation for this phenomenon.

1 Formulation of the Problem

Statistics: reminder. In statistical analysis, we often need to compute some values based on the given sample $x_1, \ldots, x_n$. For example, usually, we compute the sample mean $\mu = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$ and the sample variance $\sigma^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \mu)^2$. The sample mean is a good approximation to the mean of the corresponding probability distribution, and the sample variance is a good approximation to the variance of this distribution.

Alternatively, we can estimate other approximations to mean and variance or approximations to other characteristics of the probability distribution. In all these cases, we compute some value $v(x_1, \ldots, x_n)$ that depends on the sample

$x_1, \ldots, x_n$. The corresponding functions $v(x_1, \ldots, x_n)$ are known as *statistics*; see, e.g., [10].

Need to compute statistics under interval uncertainty. The usual formulas for computing the statistics are based on the implicit assumption that we know the exact values $x_1, \ldots, x_n$ from the corresponding sample. In real life, these values come from measurements, and measurements are never absolutely accurate. As a result, the measurement results $\widetilde{x}_1, \ldots, \widetilde{x}_n$ are, in general, different from the actual (unknown) values $x_1, \ldots, x_n$ of the corresponding quantities; see, e.g., [9].

Sometime, we know the probabilities of different values of the measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. However, in many practical situations, we do not know these probabilities, we only know the upper bound Δ_i on the (absolute value of the) measurement error: $|\Delta x_i| \le \Delta_i$ [9]. In such situations, once we know the measurement result $\widetilde{x}_i$, the only information that we have about the corresponding actual value x_i is that this value belongs to the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$.

For each statistic $v(x_1, \ldots, x_n)$, different combinations of inputs

$$x_i \in [\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$$

lead, in general, to different values v. To get a good understanding of the actual value of the corresponding characteristic of the probability distribution, it is desirable to find the range of all such possible values $v(x_1, \ldots, x_n)$. In other words, it is desirable to compute the range

$$\{v(x_1, \ldots, x_n) : x_1 \in [\widetilde{x}_1 - \Delta_1, \widetilde{x}_1 + \Delta_1], \ldots, x_n \in [\widetilde{x}_n - \Delta_n, \widetilde{x}_n + \Delta_n]\}.$$

The problem of computing such a range under interval uncertainty is known as the problem of *interval computations*; see, e.g., [5, 7].

Need for robust statistics. In addition to measurement errors, we can also have *outliers*, when the measuring instrument malfunctions, and the result is drastically different from the actual value.

For usual statistics, the presence of even a single outlier can be a disaster. For example, if we have 100 measurements results, all of which are close to 1.0, then the sample average is also close to 1.0. However, if instead of a single measurement result, we have an outlier, e.g., $\widetilde{x}_1 = 10000$, the the sample average becomes equal to $100 \gg 1.0$.

In the presence of outliers, it is desirable to consider *robust* statistics, i.e., statistics which are less vulnerable to the presence of outliers; see, e.g., [4].

Examples of robust statistics for estimating mean and standard deviation. For computing the mean of a symmetric distribution, the most robust estimate is *median*. For odd n, the median $\text{med}_i\, x_i$ can be defined as follows: when we order all the values in increasing order, into a sequence $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$, then the median is the middle value

$$\text{med}_i\, x_i \stackrel{\text{def}}{=} x_{((n+1)/2)}.$$

For even n, the median is defined as the arithmetic average of the two values which are the closest to the middle:

$$\mathrm{med}_i\, x_i \stackrel{\mathrm{def}}{=} \frac{x_{(n/2)} + x_{((n/2)+1)}}{2}.$$

A similar idea leads to a natural robust statistic for estimating the variance. The usual sample variance is defined, crudely speaking, as an arithmetic average of the square $(x_i - \mu)^2$ of the differences between each sample value x_i and the estimate μ for the mean. To have a robust statistic, it makes sense:

- to replace the sample mean μ with a robust estimate for the mean – e.g., with the median, and
- to replace the arithmetic average with a more robust operation of a median.

As a result, we get the following formula:

$$V_r(x_1, \ldots, x_n) = \mathrm{med}_i\ (x_i - \mathrm{med}_j\, x_j)^2\,.$$

Since the order between two non-negative numbers does not change when we take the squares of square roots of these numbers, this statistic can be written as

$$V_r(x_1, \ldots, x_n) = (\mathrm{MAD}(x_1, \ldots, x_n))^2,$$

where

$$\mathrm{MAD}(x_1, \ldots, x_n) = \mathrm{med}_i\ |x_i - \mathrm{med}_j\, x_j|$$

is called *mean absolute deviation.*

For robust statistics, interval computations are often easier. It is known that, in general, computing the range of sample variance under interval uncertainty is NP-hard; see, e.g., [2, 8]. In contrast, there exists a feasible (polynomial-time) algorithm for computing the range of the mean absolute deviations under interval uncertainty; see, e.g., [3]. So, computing the range of a robust statistic is, in general, much easier.

A similar comparison can be made for computing the mean and computing the median. Both mean and median are non-decreasing in terms of each of its variables. So, the smallest possible value of each statistic is attained when each of the inputs x_i takes its smallest possible value $x_i = \underline{x}_i$. Similarly, the largest possible value of each statistic is attained when each of the inputs x_i takes its largest possible value $x_i = \overline{x}_i$. Thus, to compute the range of each statistic, it is sufficient to compute two values of this statistic:

- the value corresponding to the lower endpoints $\underline{x}_1, \ldots, \underline{x}_n$, and
- the value corresponding to the upper endpoints $\overline{x}_1, \ldots, \overline{x}_n$.

Computing the arithmetic mean requires n arithmetic operations, and addition of b-bit numbers requires $O(b)$ bit operations. So, overall, to compute the two means, we need $O(n \cdot b)$ bit operations.

To compute the median of n numbers, we need $O(n)$ comparisons; see, e.g., [1]. Each comparison requires, on average, no more than 2 bit operations (irrespective of the number of bits b); see the explanation in the Appendix. Thus, overall, to compute the range of a robust statistic (median), we need $O(n)$ bit operations, which is smaller than the number $O(n \cdot b)$ of bit operations needed to compute the interval range of the traditional statistic (sample mean). So, in this case too computing the range of a robust statistic is much easier.

Why? We have shown that in several cases, computing the interval range of a robust statistic is computationally easier than computing the range of the corresponding traditional statistics. A natural question is: why?

In this paper, we provide a qualitative explanation for this empirical fact.

2 Our Explanation

In general, the narrower the interval range, the easier its estimation. Let us first show that, in general, the narrower the resulting interval range, the easier it is to compute this range.

Indeed, in general, the problem of computing the range of a given function $v(x_1, \ldots, x_n)$ on given intervals is NP-hard – as we have mentioned earlier, this problem is NP-hard even for the sample variance. Let us show that, in general, when the desired range is narrow, this range is easier to compute – namely, this range can be computed feasibly, in polynomial time.

Indeed, each value x_i from the interval $[\widetilde{x}_i - \Delta_i, \widetilde{x}_i + \Delta_i]$ can be represented as $x_i = \widetilde{x}_i + \Delta x_i$, where $|\Delta x_i| \le \Delta_i$. For smooth functions $v(x_1, \ldots, x_n)$, the corresponding value $v(x_1, \ldots, x_n) = v(\widetilde{x}_1 + \Delta x_1, \ldots, \widetilde{x}_n + \Delta x_n)$ can be expanded in Taylor series:

$$v(\widetilde{x}_1 + \Delta x_1, \ldots, \widetilde{x}_n + \Delta x_n) = \widetilde{v} + \sum_{i=1}^{n} c_i \cdot \Delta x_i + \ldots,$$

where $\widetilde{v} \stackrel{\text{def}}{=} v(\widetilde{x}_1, \ldots, \widetilde{x}_n)$ and each c_i denotes the value of the partial derivative $\dfrac{\partial f}{\partial x_i}$ at the point $(\widetilde{x}_1, \ldots, \widetilde{x}_n)$.

When the ranges Δ_i are small – and when, therefore, the resulting range of $v(x_1, \ldots, x_n)$ is narrow – we can, with good accuracy, ignore terms which are quadratic (and of higher order) in terms of Δx_i and use an approximate formula

$$v(\widetilde{x}_1 + \Delta x_1, \ldots, \widetilde{x}_n + \Delta x_n) = \widetilde{v} + \sum_{i=1}^{n} c_i \cdot \Delta x_i.$$

This is a linear function. Its largest value is attained when each of the terms in the sum is the largest. For each i, the largest value of the term $c_i \cdot \Delta x_i$ on the

interval $\Delta x_i \in [-\Delta_i, \Delta_i]$ is attained either when $\Delta x_i = \Delta_i$ (for $c_i \geq 0$) or for $\Delta x_i = -\Delta_i$ (for $c_i \leq 0$). In both cases, this largest value is equal to $|c_i| \cdot \Delta_i$ and thus, the largest value of $v(x_1, \ldots, x_n)$ is equal to $\widetilde{v} + \Delta$, where $\Delta \stackrel{\text{def}}{=} \sum_{i=1}^{n} |c_i| \cdot \Delta_i$.

Similarly, we can show that the smallest possible value of $v(x_1, \ldots, x_n)$ is equal to $\widetilde{v} - \Delta$. Thus, the range of possible value of the statistic $v(x_1, \ldots, x_n)$ is, in this approximation, equal to $[\widetilde{v} - \Delta, \widetilde{v} + \Delta]$. The above formulas for $\widetilde{v}$ and Δ enable us to compute this range feasibly – in polynomial time for a feasibly computable statistic $v(x_1, \ldots, x_n)$; see, e.g., [6].

Resulting explanation. By definition, a robust statistic $v(x_1, \ldots, x_n)$ is the one whose value changes less when we change the inputs $x_1, \ldots, x_n$. In particular, the values of the robust statistic change less when we replace each input $\widetilde{x}_i$ with a modified input $x_i = \widetilde{x}_i + \Delta x_i$. Thus, for the robust statistic, the range of all such values is narrower than for the corresponding traditional statistic.

We have already shown that, in general, the narrower the range, the easier it is to compute this range. Thus, it is reasonable to expect that for a robust statistic, computation of the range will be easier – which is exactly what we observe.

Acknowledgments

This work was supported in part by the US National Science Foundation grants HRD-0734825, HRD-1242122, and DUE-0926721.

References

[1] Th. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, MIT Press, Cambridge, Massachusetts, 2009.

[2] S. Ferson, L. Ginzburg, V. Kreinovich, L. Longpré, and M. Aviles, "Computing Variance for Interval Data is NP-Hard", *ACM SIGACT News*, 2002, Vol. 33, No. 2, pp. 108–118.

[3] B. Gladysz and A. Kasperski, "Computing mean absolute deviation under uncertainty", *Applied Soft Computing*, 2010, Vol. 10, No. 2, pp. 361–366.

[4] P. J. Huber and E. M. Ronchetti, *Robust Statistics*, Wiley, Hoboken, New Jersey, 2009.

[5] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied Interval Analysis*, Springer, London, 2001.

[6] V. Kreinovich, "Interval Computations and Interval-Related Statistical Techniques: Tools for Estimating Uncertainty of the Results of Data Processing and Indirect Measurements", In: F. Pavese and A. B. Forbes (eds.), *Data*

Modeling for Metrology and Testing in Measurement Science, Birkhauser-Springer, Boston, 2009, pp. 117–145.

[7] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM Press, Philadelphia, Pennsylvania, 2009.

[8] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer, Berlin, Heidelberg, 2012.

[9] S. G. Rabinovich, *Measurement Errors and Uncertainty. Theory and Practice*, Springer Verlag, Berlin, 2005.

[10] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall/CRC, Boca Raton, Florida, 2011.

A Appendix: An Explanation of Why Comparing Two Numbers Requires, on Average, Less than 2 Bit Operations

To compare two binary numbers, we can compare them bit-by-bit, starting with the largest bit. If all the bits until this one coincided, and this bit is different, this means that the number with bit 1 is larger.

For a random sequence of bits, the probability that the largest bits are different is 1/2. In this case, we need 1 bit operation. With probability 1/2, the largest bits are equal, then with conditional probability 1/2 the second largest bits are different. So, with probability 1/4, we need 2 bit operations. Similarly, with probability 2^{-k}, we need k bit operations. So, the average number of bit operations is equal to $2^{-1} \cdot 1 + 2^{-2} \cdot 2 + \ldots + k \cdot 2^{-k} + \ldots + b \cdot 2^{-b}$. This finite sum does not exceed the corresponding infinite sum $s \stackrel{\text{def}}{=} \sum\limits_{k=1}^{\infty} k \cdot 2^{-k}$.

This infinite sum can be computed if we take into account that

$$2s = \sum_{k=1}^{\infty} k \cdot 2^{-(k-1)}.$$

Introducing a new variable $j = k - 1$ for which $k = j + 1$, we get

$$2s = \sum_{j=0}^{\infty} (j+1) \cdot 2^{-j}.$$

This sum can be represented as the sum of two terms:

$$2s = \sum_{j=0}^{\infty} j \cdot 2^{-j} + \sum_{j=0}^{\infty} 2^{-j}.$$

The first sum differs from the original sum s only by the term for $j = 0$ which is equal to 0 anyway, so the first sum is equal to s. The second sum is a geometric

progression whose sum is equal to 2. Thus, $2s = s + 2$, hence $s = 2$. Thus, on average, we need

$$2^{-1} \cdot 1 + 2^{-2} \cdot 2 + \ldots + k \cdot 2^{-k} + \ldots + b \cdot 2^{-b} \leq \sum_{k=1}^{\infty} k \cdot 2^{-k} = 2$$

bit operations.