Absolute Bounds on the Mean of Sum, Product, etc.: A Probabilistic Extension of Interval Arithmetic

Scott Ferson[1], Lev Ginzburg[1],
Vladik Kreinovich[2], and Jorge Lopez[2]
[1]Applied Biomathematics, 100 North Country Road,
Setauket, NY 11733, USA, {scott,lev}@ramas.com
[2]Comp. Sci., U. Texas, El Paso, TX 79968, vladik@cs.utep.edu

Abstract

We extend the main formulas of interval arithmetic for different arithmetic operations $x_1 \oplus x_2$ to the case when, for each input x_i, in addition to the interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$ of possible values, we also know its mean E_i (or an interval $\mathbf{E}_i$ of possible values of the mean), and we want to find the corresponding bounds for $x_1 \oplus x_2$ and its mean.

Error estimation for indirect measurements: an important practical problem. A practically important class of statistical problems is related to data processing (indirect measurements). Some physical quantities y – such as the distance to a star or the amount of oil in a given well – are impossible or difficult to measure directly. To estimate these quantities, we use *indirect* measurements, i.e., we measure some easier-to-measure quantities $x_1, \ldots, x_n$ which are related to y by a known relation $y = f(x_1, \ldots, x_n)$, and then use the measurement results $\widetilde{x}_i$ $(1 \le i \le n)$ to compute an estimate $\widetilde{y}$ for y as $\widetilde{y} = f(\widetilde{x}_1, \ldots, \widetilde{x}_n)$. For example, to find the resistance R, we measure current I and voltage V, and then use the known relation $R = V/I$ to estimate resistance as $\widetilde{R} = \widetilde{V}/\widetilde{I}$.

Measurement are never 100% accurate, so in reality, the actual value x_i of i-th measured quantity can differ from the measurement result $\widetilde{x}_i$. In probabilistic terms, x_i is a random variable; its probability distribution describes the probabilities of different possible value of measurement error $\Delta x_i \stackrel{\text{def}}{=} \widetilde{x}_i - x_i$. It is desirable to describe the error $\widetilde{y} - y$ of the result of data processing.

Often, we know (or assume) that the measurement error Δx_i of each direct measurement is normally distributed with a known standard deviation σ_i, and that measurement errors corresponding to different measurements are independent. These assumptions – justified by the central limit theorem, according to which sums of independent identically distributed random variables with finite

moments tend quickly toward the Gaussian distribution – underly the traditional engineering approach to estimating measurement errors.

In some situations, the error distributions are not Gaussian, but we know their exact shape (e.g., lognormal). In many practical measurement situations, however, we only have *partial* information about the probability distributions.

The need for robust statistics. Traditional statistical techniques deal with the situations when we know the exact shape of the probability distributions. To deal with practical situations in which we only have a partial information about the distributions, special techniques have to be invented. Such techniques are called methods of *robust statistics*. They are called robust because they are usually designed to provide guaranteed estimates, i.e., estimates which are valid for all possible distributions from a given class.

Interval computations as a particular case of robust statistics. An important case of partial information about a random variable x is when we know (with probability 1) that x is within a given interval $\mathbf{x} = [\underline{x}, \overline{x}]$, but we have no information about the probability distribution within this interval. In other words, x may be uniformly distributed on this interval, it may be deterministic (i.e., distributed in a single value with probability 1), distributed according to a truncated Gaussian, bimodal distribution – we do not know.

So, we arrive at the following problem: for each of n random variables $x_1, \ldots, x_n$, we know that it is located (with probability 1) within a given interval $\mathbf{x}_i = [\underline{x}_i, \overline{x}_i]$. We do not know the distributions within the intervals, and we do not know whether the random variables x_i are independent or not. What can we then conclude about the probability distribution of $y = f(x_1, \ldots, x_n)$?

Since the only information we have about each variable x_i consists of its lower bound $\underline{x}_i$ and upper bound $\overline{x}_i$, it is natural to ask for similar bounds $\mathbf{y} = [\underline{y}, \overline{y}]$ for y. As a result, we arrive at the following problem:

GIVEN: an algorithm computing a function $f(x_1, \ldots, x_n)$ from R^n to R and n intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$,

TAKE: all possible joint probability distributions on R^n for which, for each i, $x_i \in \mathbf{x}_i$ with probability 1;

FIND: the set $\mathbf{Y}$ of all possible values of a random variable $y = f(x_1, \ldots, x_n)$ for all such distributions.

One can easily prove that $\mathbf{Y}$ is equal to the range $f(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of the given function f on given intervals, i.e., to $\{f(x_1, \ldots, x_n) \,|\, x_1 \in \mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

This is exactly the problem solved by interval computations. The main interval computations approach to solving this problem is to take into consideration that inside the computer, every algorithm consists of elementary operations (arithmetic operations, min, max, etc.). For each elementary operation $f(x, y)$, if we know the intervals $\mathbf{x}$ and $\mathbf{y}$ for x and y, we can compute the exact range $f(\mathbf{x}, \mathbf{y})$; the corresponding formulas form the so-called *interval arithmetic*. We

can therefore repeat the computations forming the program f step-by-step, replacing each operation with real numbers by the corresponding operation of interval arithmetic. It is known that, as a result, we get an enclosure for the desired range.

Comment. In the above text, we considered the case when we have no information about the correlation between the random variables. We have proven that in the above problem, if we assume independence, we still get the same range.

For functions of two variables, we can consider two additional cases: when x_1 and x_2 are highly positively correlated (i.e., crudely speaking, that x_1 is (non-strictly) increasing in x_2, and when x_i is highly negatively correlated (i.e., when x_1 is decreasing in x_2). In both cases, we get the same range $\mathbf{Y}$ as in the above case of no information about the correlation.

New problem. In some practical situations, in addition to the lower and upper bounds on each random variable x_i, we know the bounds $\mathbf{E}_i = [\underline{E}_i, \overline{E}_i]$ on its mean E_i. In such situations, we arrive at the following problem:

GIVEN: an algorithm computing a function $f(x_1, \ldots, x_n)$ from R^n to R; n intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$, and n intervals $\mathbf{E}_1, \ldots, \mathbf{E}_n$,

TAKE: all possible joint probability distributions on R^n for which, for each i, $x_i \in \mathbf{x}_i$ with probability 1 and the mean E_i belongs to $\mathbf{E}_i$;

FIND: the set $\mathbf{Y}$ of all possible values of a random variable $y = f(x_1, \ldots, x_n)$ and the set $\mathbf{E}$ of all possible values of $E[y]$ for all such distributions.

A similar problem can be formulated for the case when x_i are known to be independent, and for the cases when $n = 2$ and the values x_i are highly positively or highly negatively correlated.

If we can find the range for degenerate intervals $\mathbf{E}_i = [E_i, E_i]$, then we can use interval computation to extend these formulas to arbitrary intervals $\mathbf{E}_i$.

Similarly to interval computations, our main idea is to find the corresponding formulas for the cases when $n = 2$ and $f = \oplus$ is one of the basic arithmetic operations ($+$, $-$, $\cdot$, min, max). For example, if we know two "triples" $(\underline{x}_i, E_i, \overline{x}_i)$, $(i = 1, 2)$, what are the possible triples $(\underline{y}, E, \overline{y})$ for $y = x_1 \cdot x_2$?

Main results. For all basic operations, the interval part $(\underline{y}, \overline{y})$ of the result is the same as for interval arithmetic.

We provide explicit formulas for the interval $\mathbf{E}$ of possible values of $E = E[y]$. For example, for multiplication, when we know nothing about the correlation,

$$\overline{E} = \min(p_1, p_2) \cdot \overline{x}_1 \cdot \overline{x}_2 + \max(p_1 - p_2, 0) \cdot \overline{x}_1 \cdot \underline{x}_2 + \max(p_2 - p_1, 0) \cdot \underline{x}_1 \cdot \overline{x}_2 + \min(1 - p_1, 1 - p_2) \cdot \underline{x}_1 \cdot \underline{x}_2,$$

where $p_i \stackrel{\text{def}}{=} (E_i - \underline{x}_i)/(\overline{x}_i - \underline{x}_i)$.