APIC'95, El Paso, Extended Abstracts, A Supplement to the international journal of Reliable Computing

Interval Methods For Processing Statistical Characteristics

V. P. Kuznetsov

***Abstract*— Traditional applications of interval computations are based on the assumption that for all measured values, the only information that we have about the error is the error bound Δ that is guaranteed by the manufacturer of the measuring instrument. In such cases, the only information that we have about the actual value x of the measured quantity is that it belongs to an *interval* $[\tilde{x}-\Delta, \tilde{x}+\Delta]$, where $\tilde{x}$ is the result of the measurements.**

In this paper, we consider the case when in addition to the guaranteed error bound, we have an additional information about the probabilities of different values of errors. We will show that interval methods can be applied to this situation as well.

I. Where Do We Get Statistical Characteristics From?

The information about these characteristics can come from two possible sources:

- First, we may get it *theoretically*, if we know what components the measuring instrument consists of, and how exactly each component work (in particular, what are the probabilities of different possible deviations of each component from its desired behavior).
- In the majority of real-life cases, however, we do not have that theoretical information. In these cases, the only source of information about the error is *experiment*.

How can we get an experimental value of error? The error is defined as a difference $\tilde{x} - x$ between the measured and the actual values of x. So, if we want to get the experimental value of an error, we must somehow determine the actual value of the measured quantity x. We cannot measure x with an absolute precision, because every measurement is inevitably inaccurate. However, we can get a value that is (for our purposes) indistinguishable from the actual value x if we measure the same physical value x by another measuring instrument that is much more accurate than the one that we are testing. This second measuring instrument is much more accurate and therefore, much more expensive. For that reason, we do not use it for everyday measurements, we only use it as a *calibration standard* to *calibrate* the measuring instrument that we will be actually using.

The author is with the Dept. of Prognosis, Institute of Complex Systems, Moscow-Murmansk, Russia. The author is thankful to V. Kreinovich for his help in writing and editing this text.

So, we arrive at the following experimental method of determining the statistical characteristics of the measuring error: We simultaneously measure the same value of the physical quantity by two different measuring instrument:

- the one that we are testing (and that we will be later on using for measurements), and
- the *standard* measuring instrument that is known to be practically perfect.

Since we have chosen a (practically perfect) measuring instrument as the second measuring instrument in this *calibration* procedure, the value $\tilde{x}_{st}$ that is measured by this measurement instrument is practically indistinguishable from the actual value: $\tilde{x}_{st} \approx x$. Therefore, the difference $\tilde{x} - \tilde{x}_{st}$ between the two measurement results is practically indistinguishable from the error $\tilde{x} - x$ of this particular measurement. If we repeat this comparative measurement several times, then we get several different values of error.

In terms of mathematical statistics, these values from a *sample* of the unknown distribution. So, if we do not have any theoretical ideas about the error distribution, then we have to extract the information about the probabilities of different errors from this sample. Depending on the size of the sample, we have two different situations:

- If this sample is sufficiently large (contains $\approx 10^3$ error values), then we can check whether this sample is consistent with Gaussian or any other known family of probability distributions. If it is consistent with one of them, then we know the family of distributions to which this particular error distribution belongs, and the only thing that we still need to determine from the sample are the parameters of this distribution. This case will be considered in Section 1.3.

- If the sample is not so huge, then we do not have enough values to determine the shape of the distribution, However, for ≈ 100 elements, we can determine the statistical characteristics of this sample, such as the probability of the error to be in a certain interval, moments of this distribution, etc. This case will be considered in the current paper.

II. Typical Statistical Characteristics

Usually, as a first step of processing the sample $v_1, ..., v_N$ of error values, we estimate the values of one or several of the following statistical characteristics:

- *Moments*, i.e., characteristics of the type $E(v^a)$ for a given integer a.

  Here by E, we denoted *mathematical expectation* (i.e., $E(f(v)) = \int f(v)\,d\mu$, where μ is the desired probability measure).

  A moment is usually estimated simply as an arithmetic average over the sample: $(v_1^a + ... + v_N^a)/N$.

- *Generalized moments* of the type $E(f(v))$ for some function $f(v)$ (e.g., $f(v) = |v|$, or $f(v) = (v-a)^2$ for some a).

  Just like regular moments, a generalized moment is estimated by an arithmetic average $(f(v_1) + ... + f(v_N))/N$.

- *Probabilities* $p([a,b])$ that the error v belongs to the given interval $[a,b]$.

  This probability is estimated as the ratio M/N, where M is the number of values v_i that lie inside the interval $[a,b]$.

  In particular, for $a = -\infty$, this probability transforms into a *distribution function* $F(b) = p((-\infty, b])$.

- Another possibility is to measure the *inverse* to the distribution function, i.e., given a probability $p \in [0,1]$, try to find v for which $p((-\infty, v]) = p$. The values of this inverse characteristic are also called *quantiles*. If there are several such v, then we can take the smallest of them; if the distribution function is discontinuous, then we can define the quantile as $\sup\{x \,|\, p((-\infty, x] < x\}$.

  This characteristic is usually estimated by using the *ordinal* statistics, i.e., the value $v_{(k)}$ that is k–th in increasing order among the values $v_1, ..., v_N$, and $k \approx p \cdot N$. In particular, such characteristics include *maximum* $\max v_i$, *minimum* $\min v_i$, $\max |v_i|$, and the *median* $v_{(N/2)}$.

Comment. The probability $p([a,b])$ can also be represented as a generalized moment: namely, $p([a,b]) = E(\chi_{[a,b]}(v))$ (where $\chi_S(v)$ denotes a *characteristic function* of the set S, i.e., a function that is equal to 1 if $v \in S$, and to 0 else).

The characteristics obtained as a result of this processing are sometimes used to compute more complicated characteristics. For example:

- If we know the average $a = E(v)$, we can estimate the *mean square deviation* $E((v-a)^2)$.
- If we know the median $a = v_{(N/2)}$, then we can compute the *largest deviation* of the value from the median $\max |v_i - a|$.

These characteristics, in their turn, can be used to estimate further characteristics, etc.

III. For Each Statistical Characteristic, We Have An Interval Of Possible Values

A. Statistical characteristics cannot be precisely determined from experiments

The very idea of a random variable include *randomness, impossibility to predict exactly.* Because of that, it is impossible to precisely determine these characteristics based on finitely many observations v_i, because no matter how we process these observations, their random character will lead to an unpredictable fluctuations of the processing result.

B. The more observations, the better the estimate

Statistical characteristics are defined in terms of *probabilities* of different values, and the probability of each event is defined as a *limit* of the frequency M/N, where N is the total number of experiments, and M is the number of experiments in which this event occurred. Therefore, the more observations we have, the closer the resulting estimate for probability to the actual value of this probability. Therefore, the estimate of a statistical characteristics (that is based on these probability estimates) is getting closer to the actual value of this characteristic as $N \to \infty$. How closer?

Let us answer this question characteristic-by-characteristic.

C. The simplest case: estimating probabilities

Suppose that we want to determine the probability of a certain event. To determine this probability, we repeat the experiment N times. If in these N cases, the desired event occurred M times, then we take M/N as the estimate for the probability. However, it is not the actual probability. The larger N, the closer we are to the actual probability. So, the possible actual values of the probability form an *interval*. This interval is provided by mathematical statistics. (Actually, mathematical statistics provides us with *several* different confidence intervals, but for simplicity, we will assume that we have chosen one of them).

D. Moments and generalized moments

Mathematical statistics also provides us with *intervals* of possible values for moments and generalized moments.

E. Ordinal statistics

For the ordinal statistics, the situation is slightly different, because the inverse function to the probability distribution is not always well defined: for example, if we have a random variable that takes two values 0 and 1 with equal probability 0.5, then the median can be any number between 0 and 1, and therefore, its estimate $v_{(N/2)}$ either take the value 0, or the value 1. Therefore, no matter how many experiments we undertake, these estimates alternate between 0 and 1 and never converge to anything.

To express the fact that for the increasing number of experiments N, we do get a better estimate, we must consider the resulting estimate $v_{(k)}$ *not* as an estimate for the *inverse* function to a distribution function, but as an estimate for the *distribution function itself.* Namely, we say that the number p $(= k/N)$; the number that we started with) is an estimate for $p((-\infty, v_{(k)}))$.

F. Extremal ordinal characteristics

Extremal characteristics $v_{(1)}$ and $v_{(N)}$ (i.e., $\max v_i$ and $\min v_i$) have a different natural interpretation: they are approximations for the interval of possible values of the random variable. They can also be represented as estimates for generalized moments:

- If we knew *all* infinitely many values $v_1, ..., v_N, ...$, then we would be able to conclude that with probability 1, the values of v_i belong to the interval $[\min(v_i), \max(v_i)]$.
- In real life, we only know *finitely many* values $v_1, ..., v_N$. Therefore, we cannot guarantee that all the values belong to the interval $[\min(v_i), \max(v_i)]$. However, we *can* conclude that they belong to this interval with a certain probability that tends to 0 as $N \to \infty$.

The probability that a random number v belongs to an interval $[a, b]$ can be also described as a generalized moment: namely, as $E(\chi_{[a,b]})$. So, the accuracy of the extremal ordinal characteristics can also be described by saying that the value of some generalized moment belongs to a certain interval.

G. Conclusion

In all cases, the estimates that we get from experiments are interval estimates of the generalized moments, of the type $a_i^- \leq E(f_i(v)) \leq a_i^+$.

IV. A General Description Of This Type Of Statistical Information

Definition 1. [1] *Let X be a set. By a (partial) statistical information (for a distribution defined on a set X), we mean a tuple $(m, f_1, ..., f_m, \mathbf{a}_1, ..., \mathbf{a}_m)$, where:*

- *m in a non-negative integer;*
- *f_i, $1 \leq i \leq m$, are functions from X to R; and*
- *$\mathbf{a}_i$, $1 \leq i \leq m$ are intervals.*

We say that a probability measure μ on X is consistent with this information if for all i from 1 to m, we have $E(f_i) \in \mathbf{a}_i$ (where $E(y)$ denotes $\int y\,d\mu$).

Comment 1. Since we are interested in probabilities of errors, we will only consider the case when X is either the set of possible values of a single error (i.e., $X = R$), or X is the set of all possible values of all directly measured quantities (i.e., $X = R^n$ for some $n \geq 1$).

Comment 2. In case of several variables, part of the information is about individual variables, and part of it may be global. The statistical information about the variable consists of the statements of the type $E(f_i) \in \mathbf{a}_i$, where f_i is a function of x_i. Each of these statements can be easily described as a statement about the entire vector $\vec{x}(x_1, ..., x_n)$: namely, as $E(\tilde{f}_i) \in \mathbf{a}_i$, where $\tilde{f}_i$ is defined as $\tilde{f}_i(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n) = f_i(x_i)$.

Comment 3. In addition to the information coming from measurements, we sometimes have an additional (a priori) information. In many cases, this additional (a priori) information can also be described in these terms. For example, suppose that we know beforehand that the value of the random variable v always belongs to an interval $[a, b]$. This information can be expressed as follows: the probability that $v \in [a, b]$ is equal to 1. If we express this probability in terms of a generalized moment (as we have done in the previous subsection), then we can express this condition as $E(f_i) \in \mathbf{a}_i$, where $f_i = \chi_{[a,b]}$ and $\mathbf{a}_i = [1, 1]$.

V. Basic Problems of Processing Statistical Information

A. Analysis

Before any data processing, the following question naturally arises [1]: Suppose that we have some partial statistical information, and we are interested in the value of some other statistical characteristics. For example, we know the second moment of the error, and we want to find the possible value of the probability that this error belongs to a certain interval $[a^-, a^+]$. Since the given information is *partial*, we do not expect a precise number, but we would like to know the *interval* of possible values of the desired characteristic.

Definition 2. *By the problem of estimating the value of a statistical characteristics, we mean the following problem:*
Given:

- *a set* X*;*
- *a partial statistical information* $(m, f_1, ..., f_m, \mathbf{a}_1, ..., \mathbf{a}_m)$ *on a set* X*; and*
- *a function* $g : X \to R$.

To compute: the interval $\mathbf{a} = [a^-, a^+]$ *of possible values of* $E(g(x))$ *for all distributions that are consistent with the given statistical information.*

B. Data processing

Informally, the problem of estimating the errors of the result of data processing can be described as follows:

- We have a partial statistical information about the inputs $\vec{x} = (x_1, ..., x_n)$.
- We know the function $y = f(x_1, ..., x_n)$ that is used for data processing.
- We want to know the possible values of a certain characteristics g of y.

The desired characteristics can be expressed as $E(y) = E(\tilde{g})$, where we defined a new function $\tilde{g} : R^n \to R$ as $\tilde{g}(x_1, ..., x_n) = g(f(x_1, ..., x_n))$; therefore, an arbitrary data processing problem can be represented as a particular case of the analysis problem.

VI. The General Methodology of Solving the Basic Problem Of Statistical Data Processing: Main Idea

In [1], a methodology (based on linear programming) is proposed that solves this problem. Linear programming is a conditional optimization problem in which the objective function is linear in the variables, and the conditions that restrict the values of the variables are linear inequalities. A standard representation of a linear programming problem is in the following form:

$$\sum_{j=1}^{n} c_j x_j \to \max$$

under the conditions that $x_j \geq 0$ and for each j from 1 to n, either

$$\sum_{j=1}^{m} a_{ij} x_j \leq b_i,$$

or

$$\sum_{j=1}^{m} a_{ij} x_j = b_i.$$

So, to apply linear programming techniques, we must reformulate our problem in this form. This will be done in three steps:

- First, we describe our problem as an optimization problem. We need to find the interval of possible values of $E(g)$. This means that we must find two values:
  - the *largest* possible value of $E(g)$ for all statistical distributions consistent with the given partial information, and
  - the *smallest* possible value of $E(g)$ for all statistical distributions consistent with the given partial information.

  In other words, we must solve *two* optimization problems:
  - $E(g) \to \max$ under the condition that $a_i^- \leq E(f_i) \leq a_i^+$ for all i, and
  - $E(g) \to \min$ under the same conditions.
- If X is a finite set $X = \{a_1, ..., a_n\}$, then, to define a probability measure, it is sufficient to define n non-negative numbers $p_i = p(\{a_i\})$, $1 \leq i \leq n$. These numbers must satisfy the condition that the total probability is 1: $p_1 + ... + p_n = 1$. In terms of p_i, the value of the desired characteristic $E(g) = \int g\, d\mu$ can be represented as $\sum f(a_i) p_i$, i.e., a linear combination of the variables p_i. Similarly, the conditions $a_i^- \leq E(f_i) \leq a_i^+$ can be represented a linear inequalities.
  To get a similar representation for the general case (when the set X may be infinite), we can use the continuous analogue of the probability at a point: namely, a probability density $\rho(x)$ with respect to some standard measure (e.g., with respect to Lebesgue measure for $X = R^d$). Therefore, instead of *finitely many* variables p_i, we will have *infinitely many* variables $\rho(x)$ that correspond to all possible values $x \in X$. Another complication is that for some probability measures (e.g., for a measure for which $x = 0$ with probability 1), $\rho(x)$ if a *generalized* function (i.e., a *distribution* in the sense of functional analysis). However, we can still, *formally*, write down our problem in

a form that is linear in the "variables" $\rho(x)$: $\int g(x)\rho(x)\,dx \to \max$ under the conditions that $\rho(x) \geq 0$ for all $x \in X$, $\int \rho(x)\,dx = 1$, and $a_i^- \leq \int f_i(x)\rho(x)\,dx \leq a_i^+$, $1 \leq i \leq m$.

- This expression is almost in the desired form. The only remaining step is to express the conditions in the standard form. Namely, we need conditions of the type $\sum a_{ij}x_j \leq b_i$ or $\sum a_{ij}x_j = b_i$, while some of our conditions are inequalities of the opposite sign: $\sum a_{ij}x_j \geq b_i$. In linear programming, it is well known how to reformulate these conditions in a standard form: namely, we reformulate each of these opposite-sign conditions as $\sum(-a_{ij})x_j \leq (-b_i)$.

As a result, we get the following representation of our original problem: $\int g(x)\rho(x)\,dx \to \max$ under the conditions that $\rho(x) \geq 0$, $\int \rho(x)\,dx = 1$, $\int f_i(x)\rho(x)\,dx \leq a_i^+$, and $\int(-f_i(x))\rho(x)\,dx \leq (-a_i^-)$. Here, we have:

- $x \in X$ instead of $j \in \{1, ..., n\}$;
- $\rho(x)$ instead of x_j;
- $g(x)$ instead of c_j;
- 1, $f_i(x)$, and $-f_i(x)$ instead of a_{ij};
- 1, a_i^+, and $-a_i^-$ instead of b_i.

Since we have a linear programming problem with *infinitely many variables*, we cannot expect to apply known algorithms here. However, we can apply the general tricks that helps to solve these problems. One of the main tools in linear programming is the notion of *duality*. Namely, according to duality theorem, the maximum attained in the above-described linear programming problem coincides with the minimum attained in the so-called *dual* problem. Namely, to each of the conditions of the original problem, we assign the new variable y_i, and define the dual problem as

$$\sum_{i=1}^{n} b_i y_i \to \min$$

under the conditions that $y_i \geq 0$ for those conditions that were inequalities, and

$$\sum_{i=1}^{m} a_{ij} y_i \geq c_j, \ 1 \leq j \leq m.$$

Let us express this formula in our case. Each of the new variables y_i corresponds to one of the conditions of the original linear programming problem. So, to make out notations clearer, let us use the following denotations:

- By y_0, we denote the variable that correspond to the equality $\int \rho(x)\,dx = 1$.
- By y_i^+, we denote the variable that corresponds to the inequality $\int f_i(x)\rho(x)\,dx \leq a_i^+$.
- By y_i^-, we denote the variable that corresponds to the inequality $\int(-f_i(x))\rho(x)\,dx \leq (-a_i^-)$.

As a result, we get the following problem: $J^+ \to \min$, where we denoted

$$J^+ = y_0 + \sum_{i=1}^{m} y_i^+ a_i^+ - \sum_{i=1}^{m} y_i^- a_i^-, \tag{1}$$

under the conditions that

$$y_i^+ \geq 0, y_i^- \geq 0, \tag{2a}$$

and

$$g(x) \leq y_0 + \sum_{i=1}^{m} y_i^+ f_i(x) - \sum_{i=1}^{m} y_i^- f_i(x) \text{ for all } x \in X. \tag{2b}$$

Similarly, to estimate the lower bound a^- for $E(g)$, we must solve the following problem: $J^- \to \max$, where

$$J = y_0 + \sum_{i=1}^{m} y_i^+ a_i^- - \sum_{i=1}^{m} y_i^- a_i^+ \tag{3}$$

under the conditions that

$$y_i^+ \geq 0, y_i^- \geq 0, \tag{4a}$$

and

$$g(x) \geq y_0 + \sum_{i=1}^{m} y_i^+ f_i(x) - \sum_{i=1}^{m} y_i^- f_i(x) \text{ for all } x \in X. \tag{4b}$$

So, we arrive at the following methodology:

VII. The General Methodology of Solving the Problems of Statistical Data Processing: Description

Definition 3. *Suppose that a problem of estimating the value of a statistical characteristic is given. By its dual, we mean the following two problems:*

- *the problem of minimizing the function (1) under the conditions (2), and*
- *the problem of maximizing the function (3) under the conditions (4).*

The solutions of these problems will be denoted correspondingly by m^- and m^+. The corresponding interval $\mathbf{m} = [m^-, m^+]$ is called the solution of a dual problem.

Definition 4. *We say that for a given problem of estimating the value of a statistical characteristics, the duality theorem holds if the solution $\mathbf{a}$ to this problem coincides with the solution $\mathbf{m}$ to its dual problem.*

Comment 1. Because of this definition, if the duality theorem holds, then one way to find the bounds $a^\pm$ for $E(f)$ is to find the following two numbers:

- the minimum m^+ of the function (1) under the conditions (2), and
- the maximum m^- of the function (3) under the conditions (4).

In the cases when duality theorem is true, these values $m^\pm$ coincide with the desired values $a^\pm$.

Comment 2. If the duality theorem is not true, or if we do not know whether it is true or not, computing $m^\pm$ still leads to reasonable estimates for $a^\pm$:

PROPOSITION. [1] *For every particular case of the problem of estimating the value of a statistical characteristic, its desired solution* $\mathbf{a}$ *is a subset of the solution* $\mathbf{m} = [m^-, m^+]$ *to a dual problem.*

So, we arrive at the following methodology:

METHODOLOGY 1. (solving the basic problem of statistical data processing)

Input:

- `a set` X;
- `a partial statistical information` $(m, f_1, ..., f_m, \mathbf{a}_1, ..., \mathbf{a}_m)$ `on a set` X; `and`
- `a function` $g : X \to R$.

Objective: `To estimate the interval` $\mathbf{a} = [a^-, a^+]$ `of possible values of` $E(g(x))$ `for all distributions that are consistent with the given statistical information.`

Objective in precise mathematical terms: `To estimate the interval of possible values of the mathematical expectation` $E(g(x))$ `for all probability distributions for which` $E(f_i(x)) \in \mathbf{a}_i$ `for` $i = 1, ..., m$.

Methodology:

- `Solve the problem of minimizing the function (1) under the condition (2); the resulting minimum will be denoted by` m^-.
- `Solve the problem of maximizing the function (3) under the condition (4); the resulting minimum will be denoted by` m^+.
- `Return the interval` $[m^-, m^+]$ `as the desired estimate.`

Comment 1. If the duality theorem holds, then the resulting estimate is precise.

Comment 2. This Proposition reduces the solution of a problem to the solution of its dual. In general, we do not know how to solve the dual problem either, but in many important cases, the dual problem is solvable. In these cases, we get a reasonable estimate for the solution of the original problem. Several such cases are described in [1]:

- The case when we know the second moment, and we want to estimate the probability of belonging to a given interval $[a, b]$. In this case, $g(x) = \chi_{[a,b]}(x)$, $m = 1$, $f_1(x) = x^2$, and we are looking for functions $y_0 + y_1 x^2$ with the property that $y_0 + y_1 x^2 \geq \chi_{[a,b]}(x)$ for all x. This property can be explicitly expressed in terms of y_i, and therefore, we get a simple and solvable optimization problem.
- The case when we know the moment $E(|x|^\alpha)$ for some $\alpha > 0$, and we want to estimate $E(|x|^\beta)$ for some $\beta \neq \alpha$.
- The case when we know the generalized moments $E(\exp(-k|x|)$, and we want to estimate probabilities $p([a, b])$.
- The case when we know moments, and we want to estimate Fourier transforms $E(\sin(\omega x))$ and $E(\cos(\omega x))$.

3. In some cases, we do not know how to solve the dual problem either. In these case, we can still use this Proposition to get some estimates for $\mathbf{a}$. This possibility is based on the following facts:

- according to our Proposition, $a^- \geq m^-$;
- m^- is defined as the maximum of the function J^- (defined by the formula (3)) for all $y_i^\pm$ that satisfy the condition (4); therefore, if for some $y_i^\pm$, we get the conditions (4), then the corresponding value of J^- is $\leq m^-$ and therefore, $\leq a^-$.

A similar estimate can be done for a^+, so, we get the following corollary:

COROLLARY. *Assume that a problem of estimating the value of a statistical characteristics is given. Then, the solution* $\mathbf{a} = [a^-, a^+]$ *of this problem satisfies the following properties:*

- *If the numbers* y_0 *and* $y_i^\pm$ *satisfy the condition (2), then* $J^+(y_0, \{y_i^\pm\}) \geq a^+$.
- *If the numbers* y_0 *and* $y_i^\pm$ *satisfy the condition (4), then* $J^-(y_0, \{y_i^\pm\}) \leq a^-$.

METHODOLOGY 2. (solving the basic problem of statistical data processing)

Input:

- `a set` X;
- `a partial statistical information` $(m, f_1, ..., f_m, \mathbf{a}_1, ..., \mathbf{a}_m)$ `on a set` X; `and`
- `a function` $g : X \to R$.

Objective: `To estimate the interval` $\mathbf{a} = [a^-, a^+]$ `of possible values of` $E(g(x))$ `for all distributions that are consistent with the given statistical information.`

Objective in precise mathematical terms:
`To estimate the interval of possible values of the mathematical expectation` $E(g(x))$ `for all probability distributions for which` $E(f_i(x)) \in \mathbf{a}_i$ `for` $i = 1, ..., m$.

Methodology:

- `Find one or several tuples` $(y_0, y_1^{\pm}, ...)$ `for which the inequalities (2) are true; for each of these tuples, compute the value` J^+ `by using the formula (1). Denote the smallest of the resulting values by` m^-.
- `Find one or several tuples` $(y_0, y_1^{\pm}, ...)$ `for which the inequalities (4) are true; for each of these tuples, compute the value` J^- `by using the formula (3). Denote the largest of the resulting values by` m^+.
- `Return the interval` $[m^-, m^+]$ `as the desired estimate.`

VIII. Proof of the Proposition

The proof is reasonably simple: Suppose that the conditions (2) are true. Then, the difference

$$d(x) = y_0 + \sum y_i^+ f_i(x) - \sum y_i^- f_i(x) - g(x)$$

is non-negative. Therefore, its mathematical expectation $E(d)$ is also non-negative. Hence,

$$E(g) \le E(y_0 + \sum y_i^+ f_i(x) - \sum y_i^- f_i(x)) =$$

$$y_0 + \sum y_i^+ E(f_i) - \sum y_i^- E(f_i).$$

From $a_i^- \le E(f_i) \le a_i^+$, it follows that $E(g) \le J^+$. Hence, $E(g) \le \inf J^+ = m^+$. Since this is true for every measure that is consistent with the given information, we conclude that $a^+ = \sup E(g) \le m^+$. Similarly, one can prove that $a^- \ge m^-$. Q.E.D.

References

[1] V. P. Kuznetsov, *Interval statistical models*, Moscow, Radio i Svyaz Publ., 1991 (in Russian).