Fractional Degrees of Freedom in Statistics

Mikhail N. Mashkin
E-mail: mnmashkin@yandex.ru

The concept of observation and presentation of the count (reference) results in an interval form is considered. The transition to interval measurements is achieved by use of the total reduced number of measurements (number of degrees of freedom) as a sample parameter, which allows the use of non-integer (fractional) powers of freedom in the calculation of the estimates of static parameters and criteria values. The replacement of single measurements with interval measurements at their same quantities in all cases reduces the accuracy of statistical parameters estimates.

Introduction

Currently, there are known applications of fractional powers in statistics [1]. However, the use of different methods of data processing, in particular for small samples [2] and for processing with the use of methods similar to the method of group accounting arguments [3], allows to broaden their use in calculations.

The concept of observation

According to [4], observation is the experimental basis of scientific research. Observed results are most often recorded in the form of meanings of the measured values or their counts. For static methods of measurement, the result is a single number. With dynamic methods, it is possible to record the measured value in time as the implementation of a random (non-random) process. In the latter case, the results of measurements often are the evaluations of the process parameters. In both cases, statistical stability is a prerequisite, which in particular consists, in the approximation, with a sufficiently large number of observations* to the probability of a given value. In all cases, if the measurement of the value is repeated many times, the result is a statistical distribution series corresponding to any distribution law, which may be associated with the error of the measuring system or instrument.

Each single measurement (count), as well as their totality, gives an empirical distribution, which is described in the form of a histogram, statistical series, empirical distribution function, etc. In this case, along with the above, it is necessary to specify the number of measurements, i.e. empirical description requires specifying the number of experiments (sample size) on the basis of which it is obtained. We will refer to the number of measurements, on the basis of which the empirical description of the distribution law is obtained, as the number of degrees of freedom. However, there are measured values, which, by their nature, initially have a form corresponding to a certain distribution law [5]. In this case, the measured value is set not by a value, which is constant or changing in time, but by an area at each point of which it can be located with a certain probability. This allows each measurement to match the area of the measured value with the law of its distribution.

The area of determination of the value can be set with one or more than one interval, see Fig. 1. One dimension gives the area and the value of the parameters' estimates of the distribution law.

Interval measurements

Let us consider the basic prerequisites for using intervals as measurement results.

The possibility to express numerical values of quantities in the form of intervals is used in the theory of intervals [6]. The basic idea of interval analysis is that you can work with intervals as with plain numbers. Common operations such as addition, subtraction, multiplication and division, as well as set theory operations such as intersection and union, are quite applicable to them. Interval operations are described by a ratio:

$$A@B = \{\, x@y | x \in A,\ y \in B \,\}, \qquad (1)$$

where @ is one of the operations $\{+, -, *, /, \cup, \cap\}$, while A, B are intervals.

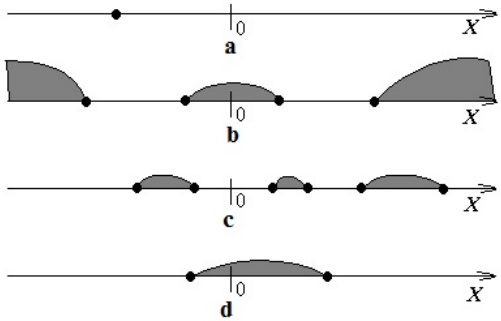

Fig. 1: Types of areas for determining the measurement value: a — one observation — a single numerical value; b — one observation — a set of intervals of numerical values, including those that are not limited to the left or to the right; c — one observation — a set of intervals strictly limited to the left and to the right; d — one observation — one interval of numerical values with one border to the left and one to the right.

*The ratio of the number of observations of a particular value to the total number of observations.

A single (real) number can be viewed as that an interval having a definition domain and the law of distribution in the form of a certain event probability:

$$P\,(a \in [a,\, a]) = 1,\ a = [a,\, a]\,, \tag{2}$$

i.e. just one numerical value is sufficient for the description of the measured value.

Let us consider the measurement process of the diameter of a bearing ring as an example of a measurement that has a definition area of one interval. The measurement of the radius of the hole or the outer diameter (done with sufficiently accurate instruments) relative to the calculated center of the bearing ring gives the dependence of the radius to the point on the circumference surface of the hole or the outer diameter in the form of a realization of a random process that can be described by a random function like follows:

$$R = X\,(\alpha)\,, \tag{3}$$

where $0 \leqslant \alpha \leqslant 2\pi$ is the bearing ring angle of rotation.

Accurate lab instruments such as circular gauges allow us to fully record the kind of realization of a random process. Obviously, when such a record exists, it can be processed by well known methods of the theory of random processes. In production conditions, the use of precision instruments is impractical. The control devices used allow to quite precisely measure the diameter of a bearing ring. During the rotation of the bearing ring it is also possible to determine the maximum and minimum values of the diameter of the bearing. If we limit ourselves to only two of these values, then actually we come to a case of two independent observations. The information that there are other numerical values of the diameter, between these two values, becomes thus lost. For a more complete explanation of the essense of the observation, it is proposed to consider the considered measurement process as a single observation in the form of one interval, Fig. 1, d. The value of the measured diameter has a description in the form of a statistical series at a given interval:

$$\hat{P}\ (\,d : d \in [d_{\min},\, d_{\max}])\,, \tag{4}$$

where d is the value of the bearing ring diameter.

With an interval measurement, however, there are two degrees of freedom: the measurements of one and the other border of the interval. However, these two dimensions are considered together over the interval. For example, one dimension is a border, and the other is the interval value itself, that is, there is a relationship: for the first dimension, the entire numerical axis is available, and the second dimension describes the area of the finite length bound to the first measurement. The availability of the entire numerical axis here must be understood as a possibility to represent the first measurement only by selecting the initial value of the reference point by any number, including almost infinity. For the interval, whatever we choose as the reference point, its value remains constant. From this we can assume that the specified relationship as if reduces the number of degrees of freedom of choice of numerical values for the interval measurement. We can assume that it is less than two, but more than one. Interval measurement generally gives the values of the borders of intervals and parameters or their estimates of the distribution law. This can be described by displaying the interval in parameters' values:

$$G : [\,a_i\, b_i\,] \xrightarrow{P} \left\{\, \beta_j : j = 1, \ldots, k \,\right\}, \tag{5}$$

where G displays the set of numerical values of the interval measurement in the values of parameters or their estimates of the probability distribution law; a_i, b_i are borders of the i-th interval; P is the law of distribution of values of a random variable from the interval; β_j is the value or estimate of a parameter of the distribution law.

It should be noted that the borders of the interval can be displayed in the parameters of the distribution law explicitly (for example, the boundaries of the interval in the case of the law of equal probability density) or indirectly as the area of definition of this law.

One of the options for describing the distribution law P is the probability density. By the given probability density or histogram it is possible to calculate or to estimate the parameters of the distribution law. The previously declared commonality for the interval and for one number (2) allows these calculations to be applied for one number obtained during the measurement. Let us illustrate this by calculating the dispersion of a single observation.

Calculation of the dispersion estimation of one observation by known relations [1] can be performed by the formula:

$$\hat{\sigma}_x^2 = \frac{\sum\limits_{i=1}^{n} (x_i - m_x)^2}{n} = \frac{(x - m_x)^2}{1} \quad \text{if } m_x \text{ is known}, \tag{6}$$

where m_x is the mathematical expectation; x is the numerical value of the dimension.

For one number from the interval with coinciding borders, formula (6) is valid, because the mathematical expectation does not require an evaluation, but is equal to the number itself. The value of the dispersion estimate in this case is zero. This clearly indicates the non-randomness of the interval representation of the same number, i.e., the specific meaning of the measured value does not have a random component — it is a non-random value.

Calculation of the dispersion estimate for an interval measurement in the extreme case can be performed as that for two independent observations by formulas:

$$\hat{\sigma}_x^2 = \frac{(b - m_x)^2 + (a - m_x)^2}{1}\,, \quad \text{if } m_x \text{ is unknown}, \tag{7}$$

$$\hat{\sigma}_x^2 = \frac{(b - m_x)^2 + (a - m_x)^2}{2}\,, \quad \text{if } m_x \text{ is known}. \tag{8}$$

It can be assumed that the value of the dispersion estimate for the interval for each case, due to the lower value of the degrees of freedom, should exceed the values given by formulas (7) and (8). In addition, within the interval, the measured numerical values of the value are determined by its distribution law. If we choose as the basic one the law of equal probability density (EPD), then we lead the rest of the distributions to it by changing the value of the interval on the basis of equality of the entropy value.

Let us define the given number of measurements (degrees of freedom) for an interval measurement in the form of:

$$r_i = 1 + \Delta_i\,, \quad 1 \geqslant |\Delta_i| \geqslant 0; \tag{9}$$

where

$$\Delta_i = \begin{cases} +\Delta_{is}, \text{ boundaries are given from experience;} \\ -\Delta_{is}, \text{ one boundary is given by the researcher;} \\ -1, \text{ boundaries are given randomly.} \end{cases}$$

The value Δ_{is} can be determined by formula:

$$\Delta_{is} = \begin{cases} \frac{1}{1 + 1/h_{is}}, \text{ at } b_i,\ a_i \neq 0; \\ 0, \text{ at } b_i,\ a_i = 0; \end{cases} \tag{10}$$

where

$$h_{is} = \frac{b_i - a_i}{\frac{1}{2}\,|\,a_i + b_i\,|}$$

is choosen for the EPD law and

$$h_{is} = \frac{(b_i - a_i)\,H_x}{H_{EPD}\left|\,\mathrm{M}\left[X_{[a,b]}\right]\right|}$$

is choosen for any other law of the distribution of x along the interval $[a_i,\ b_i]$;

$$H_x = \mathrm{M}\left[\log P\left(X = x_j \in [a_i\, b_i]\right)\right] = -\sum_{j=1}^{n} P\left(X = x_j\right) \log P\left(X = x_j\right)$$

is chosen if the given measured value is discreet*;

$$H_x = \mathrm{M}\left[\log P\left(f\left(x\right)\right)\right] = -\int_{a_i}^{b_i} f\left(x\right) \log_c f\left(x\right) dx$$

is taken at $c < (b_i - a_i)$ if the measured value is continuous (relative entropy);

$$H_{EPD} = \log n_{[a,b]}$$

if the discrete measured value is distributed equally possible within the interval, where $n_{[a,b]}$ is the number of equally possible states in the interval;

$$H_{EPD} = \log_c\left(b_i - a_i\right)$$

if within the interval the measured value is distributed according to the EPD law;

$$\mathrm{M}\left[X_{[a,b]}\right]$$

is the mathematical expectation of the measured value in the interval $[a_i,\ b_i]$.

The total reduced number of measurements, the value for the calculation of statistical parameters for the sample, is equal to:

$$n_r = \sum_{i=1}^{n} r_i\ . \tag{11}$$

This assumes that, when creating a statistical series of distributions or histograms, each interval dimension must have its own share proportional to the value of r_i. If it is 0, this dimension is ignored. If it differs from zero, then this contribution, as the number of measurements (experiments), is equal to its value.

Formulas for calculation of the main estimates of statistical parameters for one, i-interval measurement, in the case of the EPD law for the measured value within the interval, have the form:

$$\hat{m}_{x_i} = \frac{b_i + a_i}{2}\ ; \tag{12}$$

$$\hat{\sigma}^2_{r_i} = \frac{\left(b_i - \hat{m}_{x_i}\right)^2 + \left(a_i - \hat{m}_{x_i}\right)^2}{r_i - 1} = \frac{\left(b_i - a_i\right)^2}{2\left(r_i - 1\right)}\,. \tag{13}$$

Example. With a rectangular contribution (EPD), let us define by formula (13) the estimate of the variance in the interval of an i-th observation for different ratios of the value of the interval and the values of its mathematical expectation, see Table 1.

Left border of the interval, a_i	Right border of the interval, b_i	Math. expectation estimate, $\hat{m}_{x_i}$	Reduced no. of measurements, r_i	Estim. variance, $\hat{\sigma}^2_{r_i}$ (13)
−4	4	0	2	32
−3	5	1	1.889	36
−2	6	2	1.8	40
−1	7	3	1.727	44
0	8	4	1.667	48
1	9	5	1.615	52
2	10	6	1.571	56
...	...	...	...	...
30	38	34	1.190	168

Table 1: Dispersion (variance) estimation via the given number of measurements.

Analysis of Table 1 shows that in the symmetric interval (the case when the estimate of the mathematical expectation is 0), the variance estimate coincides with the value calculated by formula (7) for two unit measurements. As the value

*The given relations for determination of $H_{\#}$ are similar to entropy formulas, and for the case of discrete measured values exactly coincide with them.

of mathematical expectation increases, the variance value increases due to the reduction of the reduced number of measurements, which can be taken as the number of degrees of freedom of the resulting measurement.

Taking into account the above, a single measurement can be considered as an interval measurement when the interval is equal to the rounding error of the instrument readings. In this case, a fairly small relative error gives the reduced number of measurements equal to 1.

Contributions method

To process the results of a small sample in the evaluation of the distribution laws, the contribution method is used [2, 6]. This approach allows us to obtain a paradoxical result: due to the empirical selection of the width of the interval of a rectangular or other contribution, the accuracy of the assessment increases. The paradox is that, by coarsening the measurement results (the numbers are replaced by fixed-width intervals), the accuracy of statistical parameters is allegedly improved.

When using the formalism published in the work [2], the proposed estimation formula for the method of contributions for the probability density is:

$$\tilde{f}(x) = \frac{\sum\limits_{i=1}^{n} r_i \cdot p_i(x,\, a_i,\, b_i)}{\sum\limits_{i=1}^{n} r_i}, \qquad (14)$$

where n is the number of observations; $p_i(x,\, a_i,\, b_i)$ is a generalized record of the empirical component of the distribution density associated with the interval of i-th observation (having all the properties of the distribution density), describes the law of distribution of measurements in the interval. Unlike the work [2], empiricism is limited by the choice of the distribution law in the interval. And there are two options:

1. The distribution law is the same for all intervals;
2. For each interval, its own law of distribution is picked.

For the case of the EPD law in the interval we have:

$$p_i(x,\, a_i,\, b_i) = \frac{1}{b_i - a_i}, \quad a_i \leqslant x \leqslant b_i. \qquad (15)$$

The work [7] presents a formula which uses the method of contributions for the empirical component of density estimation in the form of:

$$f_N^*(x) = C(\rho) \sum_{i=1}^{N} \mu_i \psi_i\,(\rho,\, x)\,, \qquad (16)$$

where the ρ parameter is equal to half of the contribution interval, $\rho = \frac{b_i - a_i}{2} = const$, that is, the interval in all dimensions is the same;

$$C(\rho) = \left(\int\limits_{-\rho}^{\rho} \psi_i(\rho,\, x)\, dx \right)^{-1}, \qquad (17)$$

the amplitude ensures the equality of each contribution 1; $\mu_i = 1/N$ is weight (the ratio for norming density estimation); and also

$$\psi_i(\rho,\, x) = \begin{cases} 1,\ x_i - \rho \leqslant x \leqslant x_i + \rho\,; \\ 0,\ \text{for others } x. \end{cases} \qquad (18)$$

Let us consider the use of formulas (14) and (16) for Example 2.1 from the work [7].

Example 2.1 [7]. As a result of measurement of parameter X of three products after adjustment of the equipment, the following results were obtained: 6.0; 6.4; 6.6. Let us estimate the empirical density that characterizes the quality of the equipment setup.

Some assumptions must be made to calculate by (16). Let us suppose that. Let us suppose that $\mu_i = 1/N = 1/3 = 0.3$.

Then by formula (17)

$$C(\rho) = \left(\int\limits_{-0,3}^{0,3} \psi_i(\rho,\, x)\, dx \right)^{-1} = \frac{1}{0.6} \approx 1.67.$$

Summing the kernels (contributions) $\psi_i(\rho,\, x)$ for all $i = 1, 2, 3$ with amplitudes of 1.67 and weights 1/3, we obtain

$$f_N^*(x) = \begin{cases} 0.56, & 5.7 \leqslant x < 6.1; \\ 1.11, & 6.1 \leqslant x < 6.7; \\ 0.56, & 6.7 \leqslant x \leqslant 6.9; \end{cases} \qquad (19)$$

(see Fig. 2):

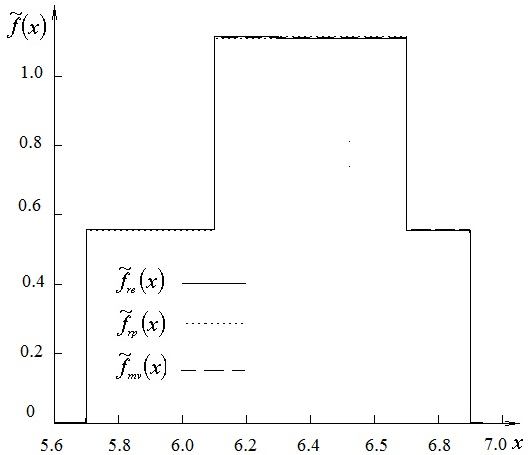

Fig. 2: Empirical estimates of the probability density, Example 2.1.

Using formula (14), the example data can be interpreted as follows: three intervals are used as input: [5.7; 6.3], [6.1; 6.7], [6.3; 6.9]. The length of each interval is equal to 0.6. The distribution law within the interval is EPD. The distribution density is equal to:

$$p_i(x,\, a_i,\, b_i) = 1/0.6; \quad a_i \leqslant x \leqslant b_i. \qquad (20)$$

N/N i/o	Left border of the interv., a_i	Right border of the interv., b_i	Math. expectated estim., $\hat{m}_{x_i}$	Reduced number, r_i	
				Both borders are experim., r_{ie}	One border is set by the researcher, r_{ip}
1	5.7	6.3	6	1.091	0.909
2	6.1	6.7	6.4	1.086	0.914
3	6.3	6.9	6.6	1.083	0.917
			Total:	3.26	2.74

Table 2: The reduced number of measurements by intervals, Example 2.1.

The calculated numerical values according to formula (9) of the given numbers of measurements for each the interval are shown in Table 2.

The total number of measurements calculated by formula (11) is equal to:

$$n_{re}\sum_{i=1}^{n} r_{ie} = 3.26$$

if all parameters of the interval are obtained experimentally (experimental data);

$$n_{rp}\sum_{i=1}^{n} r_{ip} = 2.74$$

if one of the interval's limits is specified by the researcher (a priori data).

Hence, the estimated values for the probability density (14) with account of contributions (18) look like these:

$$\tilde{f}_{re}(x) = \Delta^{re}_{f1} + \Delta^{re}_{f2} + \Delta^{re}_{f3} \quad \text{for experimental data,} \tag{21}$$

$$\tilde{f}_{rp}(x) = \Delta^{rp}_{f1} + \Delta^{rp}_{f2} + \Delta^{rp}_{f3} \quad \text{for a priori data,} \tag{22}$$

$$\tilde{f}_{mv}(x) = \Delta^{mv}_{f1} + \Delta^{mv}_{f2} + \Delta^{mv}_{f3} \quad \text{for a small sampling,} \tag{23}$$

where

$$\Delta^{re}_{f1} = 0.558, \quad \Delta^{re}_{f2} = 0.555, \quad \Delta^{re}_{f3} = 0.554;$$

$$\Delta^{rp}_{f1} = 0.553, \quad \Delta^{rp}_{f2} = 0.556, \quad \Delta^{rp}_{f3} = 0.558;$$

$$\Delta^{mv}_{f1} = 0.556, \quad \Delta^{mv}_{f2} = 0.556, \quad \Delta^{mv}_{f3} = 0.556;$$

are contribution of the intervals, while i is the interval number,

$$\Delta^{\#}_{fi} = \begin{cases} \Delta_{hi}\,, & a_i \leqslant x \leqslant b_i; \\ 0, & a_i > x > b_i; \end{cases}$$

is a contribution of the i-th interval under # (here re means "experimental", rp means "a priori", mv means "calculated by data method" [7]);

$$\Delta_{hi} = \frac{r_{i\#}}{n_{\#}\cdot(b_i - a_i)}$$

N/N i/o	Left border of the interv., a_i	Right border of the interv., b_i	Math. expectated estim., $\hat{m}_{x_i}$	Reduced number, r_i	
				Both borders are experim., r_{ie}	One border is set by the researcher, r_{ip}
1	5.56	6.44	6	1.128	0.872
2	5.96	6.84	6.4	1.121	0.879
3	6.16	7.04	6.6	1.118	0.882
4	5.0	7.2	6.1	1.265	0.735
			Total:	4.631	3.369

Table 3: The reduced number of measurements by intervals of Example 2.2.

Interv. no.	Contribution height:		
	Experim. data	A priori data	Small sampl.
1	0.277	0.294	0.284
2	0.275	0.297	0.284
3	0.274	0.298	0.284
4	0.124	0.099	0.114

Table 4: Height of contributions for Example 2.2.

is the height of the i-th contribution; $n_{mv} = 3$ is number of intervals; $r_{imv} = 1$ is the value of the method contribution [7].

The graphs of probability density estimation for dependencies (21–23) are shown in Fig. 2.

Let us also consider Example 2.2 [7], in which, along with the intervals of Example 2.1, an interval different from the others by length is included.

Example 2.2 [7]. Let us assume that in the conditions of Example 2.1 there is a priori information in the form of an interval [5.0; 7.2]. Let us calculate the estimates of the probability density. The length of the interval for readings 6.0; 6.4 and 6.6 is calculated [7] equal to 0.88, i.e. $\rho = 0.44$.

The given numbers of measurements (9) for each the interval are shown in Table 3.

The estimated probability density values in this case is:

$$\tilde{f}_{re}(x) = \sum_{i=1}^{4} \Delta^{re}_{fi} \quad \text{for experimental data,} \tag{24}$$

$$\tilde{f}_{rp}(x) = \sum_{i=1}^{4} \Delta^{rp}_{fi} \quad \text{for a priori data,} \tag{25}$$

$$\tilde{f}_{mv}(x) = \sum_{i=1}^{4} \Delta^{mv}_{fi} \quad \text{for small sample contributions.} \tag{26}$$

The heights of contributions for the intervals are shown in Table 4.

Probability density estimates for dependencies (24–26) are shown in Fig. 3.

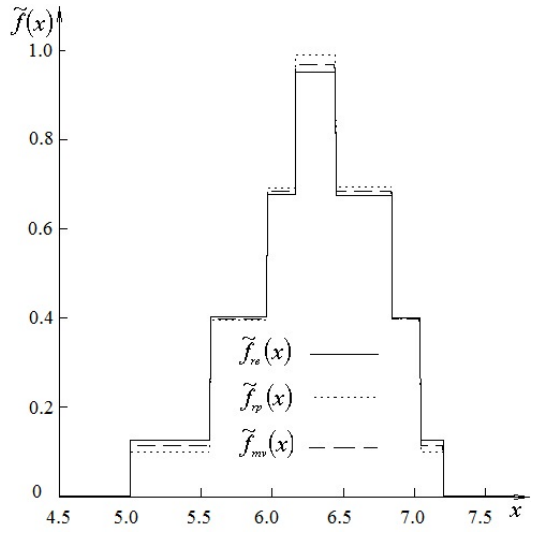

Fig. 3: Empirical estimates of probability density of Example 2.2.

Types of data	Measurement characteristics	Estimates	Examples	
			2.1	2.2
Discrete	Borders of the interval	n	—	5
		ME	—	6.24
		D	—	0.668
Discrete	Average values of the intervals	n	3	4
		ME	6.333	6.275
		D	0.093	0.076
Interval	Experimental	n_r	3.26	4.631
		ME	6.333	6.269
		D	0.133	0.272
Interval	A priori	n_r	2.74	3.369
		ME	6.334	6.283
		D	0.145	0.279
	Small samples	n	3	4
		ME	6.333	6.275
		D	0.138	0.275

Table 5: The reduced number of measurements by intervals of Example 2.2. The following designations are used here: ME — the mathematical expectation, D — the dispersion (variance), n — the number of experiments or intervals (for a single measurement, when the borders of the interval coincide, the number of intervals is equal to 1), n_r — the total given number of measurements.

Results

The reduced estimates of probability densities, Fig. 2 and Fig. 3, can be used in practical applications only when specifying for each of them the number of observations (experiments), which can be considered as the number of degrees of freedom, see formula (11) for the reduced number of measurements. In work [7] the number of experiments is equal to the number of intervals. The results of mathematical expectation and variance estimates for Examples 2.1 and 2.2, with taking different approaches into account (determination of the number of experiments as the number of intervals, or the use of the reduced number of measurements instead) are given in Table 5.

Analysis of the results displayed in Table 5 allows us to make the main conclusion: replacement of single measurements with interval measurements at the same numbers in all cases reduces the accuracy of estimates of statistical parameters. This follows from the fact that single measurements, rather than interval measurements, have the lowest variance. The application of interval measurements allows to expand the possibilities of statistical processing of measuring information. It is essential to use as a sample parameter the total reduced number of measurements (number of degrees of freedom), which allows the use of non-integer (fractional) degrees of freedom in the calculation of estimates of static parameters and criteria values.

Submitted on May 17, 2018

References

1. Bomas V.V., Bulygin V.S., Mashkin M.N. Probability Theory and Mathematical Statistics: Lectures. Moscow, Potok Publ., 2000 (*in Russian*).
2. Hahulin G.F. Principles of Designing Simulation Models: A Tutorial. Moscow, Potok Publ., 2001 (*in Russian*).
3. Ivakhnenko A.G. Heuristic Self-Organization Systems in Technical Cybernetics. Kazan, Technika, 1971 (*in Russian*).
4. Mitropolsky A.K. Technique of Statistical Calculations. Moscow, Nauka, 1971 (*in Russian*).
5. Gurney R.W. Elementary Quantum Mechanics, Cambridge University Press, 1934.
6. Knyazev G.N., Korzenev G.N. Graphic implementation in the architecture of the basic laws of management. In: *Effectiveness of Scientific Research RosZITLP*, Vol. 1, Moscow, Roszitlp Publ., 1997, pp. 41–45 (*in Russian*).
7. Gaskarov D.V., Shapovalov V.I. Small Samples. Moscow, Statistica, 1978, pp. 31–35 (*in Russian*).