Journal of Engineering Physics and Thermophysics, Vol. 83, No. 3, 2010

UNIVERSAL ALGORITHM FOR TESTING STATISTICAL HYPOTHESES WITH RESPECT TO THE DISTRIBUTION

E. V. Chernukho

UDC 519.226

On the basis of the comparison principle, we have developed a universal algorithm for testing hypotheses, making this procedure simple and transparent. The algorithm is based on the formula for calculating the probability of filling the bucket of the experimental histogram by the probability distribution of the random process source. It has been established that the probability of filling the bucket is described by the binomial distribution. The relation of this formula to the order measure formula has been shown.

Keywords: *testing of hypotheses, comparison principle, repetitive experiment, order statistics.*

Introduction. The successful application of the comparison principle [1] to the problem of estimating the distribution parameters by the data of a repetitive experiment [2] permits supposing that the same principle can also be successfully applied to the problem of testing hypotheses.

Statistics as an applied discipline considers a large number of problems classified under the type of "testing hypotheses" problems and formulated in terms of the distribution. In stating these problems, at least two objects are used. One object of the hypothesis should be the distribution and the other one should be either also the distribution or the dataset, possibly, the statistics.

Formulations of the hypotheses are versatile and complicated for comparison and, evidently, they are classified on the principle of examples proceeding from the variants of their statement. The absence of classification of hypotheses by the character of the solution method used impedes both the formulation and the testing.

As we set it, a radical step for improving the situation is to split the criterion for testing a hypothesis into two stages: measurement of the validity probability of the hypothesis and making a decision by using an algorithm specially designed proceeding from the problem formulation. In the simplest case, this algorithm represents a threshold function.

For example, in testing the null hypothesis with respect to the complementing (simple) alternative, the standard criterion will make it possible to draw only one conclusion: the hypothesis is true or false. The first part of the testing algorithm determines the validity probability of the hypothesis, and the second part is a threshold function whose threshold value is the significance criterion. The recurrent result is Boolean.

The usefulness of the proposed approach is largely determined by the simplicity of calculating the validity probability of the hypothesis. Formally, an algorithm calculating the required probability by the given distribution density and the dataset is needed.

Source Testing Problem. The distribution and the dataset are given. It is required to measure the probability that the dataset can be generated by a process having exactly this distribution ($p(x)$, $\{d\} \to m$).

Then, having used the threshold function, we can state that with a certain probability q the hypothesis that the tested distribution is suitable for describing the source is valid if $h \geq q$ and false if vice versa. The value of the threshold probability is chosen from the context of the statistical problem similarly to the choice of the quantile parameter.

The dataset is always limited in the sense that $\{d\}.R < \infty$, i.e., the observed range is finite. The same holds for empirical distributions $p(\mathrm{x}).R < \infty$. The trivial algorithm for calculating the measure aTR:if($\{d\}.R \subseteq p(x).R$) $\to \{m = 1\}$else$\{m = 1\}$ is obvious. The algorithm can be interpreted as follows. If the range of values of the dataset is entirely in the domain of definition of the distribution density and approximately commensurate to it, then there is a

A. V. Luikov Heat and Mass Transfer Institute, National Academy of Sciences of Belarus, 15 P. Brovka Str., Minsk, 220072, Belarus. Translated from Inzhenerno-Fizicheskii Zhurnal, Vol. 83, No. 3, pp. 566–573, May–June, 2010. Original article submitted August 7, 2009.

1062-0125/10/8303-0602©2010 Springer Science+Business Media, Inc.

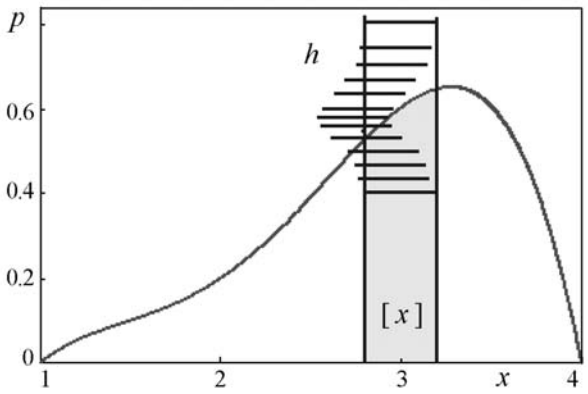

Fig. 1. Illustration of the bucket height distribution.

certain probability that the process was induced by precisely this distribution. The infinite distribution quantile can lead to the same interpretation with a supplement: "chance with a probability higher than $(1 - q)$."

Some hypotheses are perfectly well solvable by the trivial algorithm. But when the disagreement between the data and the tested distribution is not so apparent, then finer tools will be required. Below we will consider only those hypotheses that cannot be solved by the trivial algorithm.

It is easy to construct a precision measure if we note that in fact the distribution and the histogram of the dataset $p(x) \to \{d\}.H$ are compared. Indeed, the data value is influenced by the instrumental limitation because of the finiteness of the instrumentation resolution. Thus there appears a "natural" system of histogram buckets. Noteworthily, the method of partitioning the domain of definition of the histogram into buckets is of no fundamental significance. On the one hand, this is an instrumental system of buckets, and on the other hand, all these buckets can be combined arbitrarily or partitioned until a histogram in each bucket of which there is one dataset, or even some of the buckets will turn out to be empty. Thus, the histogram is a set of buckets $\{([x], H)\}$. Then the distribution adequately describes the random process only if the data are independent. What is more, the contribution of each bucket to the measure should be independent, and this independence is also predetermined by the method of calculating the measure with respect to the contribution of each bucket. If we denote by h the probability that a particular bucket has been generated by a process with the tested distribution, then the sought measure is $m = \prod_{\forall [x]_i} h_i(H_i)$, i.e., it is necessary to calculate the product of probabilities by all buckets of the histogram presented.

It remains to find a method for calculating the probability of obtaining the height of the histogram bucket by the given distribution of the process. The distribution arguments are apparent. So, we construct $h(n, [x], N)$, where we use n as an independent variable.

Distribution of the Histogram Bucket Height. *Key to the solution.* The key to the solution of problems for testing hypotheses is to calculate the probability of obtaining the bucket height. For any distribution and a specified base of the bucket, it is easy to calculate the bucket height distribution at a specified value of the dataset $(p(x), [x], N) \to h(n)$.

Figure 1 shows that as a result of the experiment, a random number of data can get into the bucket (shown by the line covering the bucket). After multiple repetitions it will turn out that the probability for various outcomes is different as to the number of hits, which is shown as the rise height of the closing line over the plane of the figure.

Below, for all illustrations we use the distributions described by a polynomial approximating four points on a limited definition interval:

$$p(x) = \mathrm{AP}\{(1, 0), (2, 1), (3, 3), (3.5, 3), (4, 0)\} \,\big|\, (x \in [1, 4]) .$$

Formulation of the problem. Let is be given the location of one bucket $[x]$ of the histogram $\{([x], H)\}$ constructed by a process having a distribution $p(x)$, and let it be even continuous. It is required to determine the distribution of the height of this bucket $h(n)$ generated by a dataset of size N. Apparently, $n \in [0, N]$ and the distribution is discrete.

Analysis of the problem. The bucket is formulated as the structure of data ([x], H), where the domain of definition of the bucket [x] represents the very "gates" that pick up data for computing the bucket height H. The histogram is constructed by a dataset $\{d\}_N$ of size N. Since the bucket-generating process is random, the computed bucket height is also random and can be obtained in each test. After a test series we find that the bucket height has its own distribution $h(n)$ depending on the distribution of the process being investigated, the location and width of the bucket, and the number of dataset elements. The calculation of this distribution is our goal and presents no difficulties.

Besides testing hypotheses, such distribution may also be needed for computing the a priori splitting of the range of data values in calculating a histogram of acceptable quality. For example, the spread of the bucket height values depends on the position in the distribution and on the width of the domain of definition of the bucket. The position and widths of buckets can be chosen so that the possible spread is the same.

Numerical solution of the problem. The simplest way of "brute force" is constructing an elementary stochastic model. For the problem under consideration, an inverse model will be better, i.e., let us begin by converting the bucket interval from the initial distribution to the stochastic generator distribution. The standard random number generator gives a uniform distribution on a unit interval (unif [0, 1]), and the corresponding transform is a cumulative distribution function: $[x] \overset{P(x)}{\rightarrow} [P(x_{\min}),\ P(x_{\max})]$. It remains to organize a cycle of the stochastic model in which a modeling dataset is generated and the result is divided and calculated:

$$\begin{array}{l} p\,(x),\,[x.\min,\,x.\max],\,N,\,\mathrm{ev}=0\,; \\ \left.\begin{array}{l} \left\{\mathrm{unif}\,([0,\,1]) \rightarrow \{d\}_N\,; \right. \\ \left. \{\mathrm{if}\,(d \in [P\,(x.\min),\,P\,(x.\max)]) \rightarrow \mathrm{ev}++\}_N\,; \right. \end{array}\right\}_{\rightarrow\infty}, \\ \{\mathrm{ev}\} \rightarrow h\,(n)\,; \end{array}$$

where the first line describes the initial data of the cycle and the cycle is carried out as long as possible, at least until a fairly stable result is obtained; the first line of the cycle generates a testing dataset, and the second line (embedded cycle) selects the event to be investigated; for example, it registers a hit; and the last line constructs a histogram of the selected event.

The algorithm is simple, it is relatively fast to execute, has no restrictions on the dataset size, and is very flexible as to the type of selected events. It is possible to calculate at once not only the number of hits, but also, singly, the contributions of different types, for example, the contribution of each order datum. The disadvantage of the algorithm is usual for a stochastic model — it requires many cycles for obtaining a result of required stability. But it has an important advantage — the accuracy is equal for all (N, n) and there is no factorial calculation problem.

Analytical solution. But if it is enough to obtain only the sought distribution, then we can propose a simple analytical formula obtained on the basis of distributions for order statistics:

$$h\,(n, N, [x]) = \frac{N!}{(N-n)!n!}\,(P\,[x])^n\,(1 - P\,[x])^{N-n}\,, \tag{1}$$

where

$$P\,[x] = \int\limits_{[x]} p\,(x)\,dx = \int\limits_{x.\min}^{x.\max} p\,(x)\,dx\,.$$

The derivation of the formula is simple and is based on combinatorial analysis of the contributions of ordered data to the final result. In fact, the above modeling algorithm is run analytically. The key to obtaining the formula is account of the combinations for each intermediate ordered datum of the dataset. The fact is that both the distribution and the histogram imply ordering of data and, therefore, the datum should necessarily also be sorted.

For the bucket under consideration, n of N data can get into the bucket (whence $(P[x])^n$ is the probability of this contribution), and the remaining N–n data will not get into it and the probability of this contribution is $(1 - P[x])^{N-n}$. The product of these probabilities will describe the contribution to (1) of the probability density of the

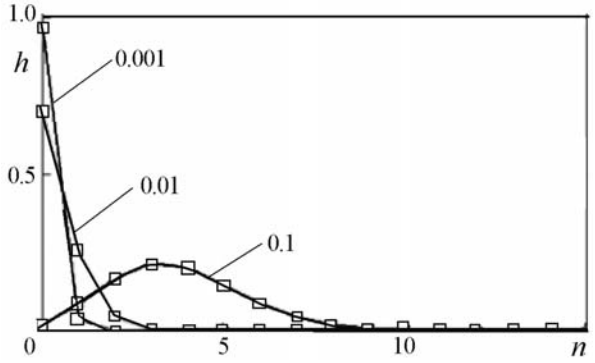

Fig. 2. Illustration of the narrowing of the bucket height prediction with decreasing width w of its base $[3, 3+w]$ by two orders of magnitude $w = 0.1$, 0.01; 0.001; $N = 60$.

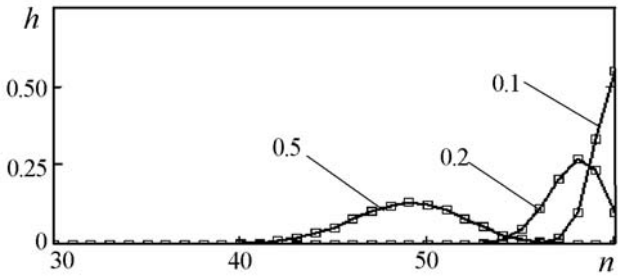

Fig. 3. Illustration of the prediction narrowing with a change in a wide bucket $[1, 4-w]$, $w = 0.5$, 0.2, 0.1; $N = 60$.

process source. It remains to take into account the number of all possible combinations. For each n the number of combinations is expressed by the binomial formula $\frac{N!}{(N-n)!n!}$. Taking into consideration all these factors, we will obtain formula (1).

As it turned out, (1) is just a binomial distribution probability bin (p, n, N), where $p = P[x]$, and the other parameters have the usual meaning. Apparently, the formula can be obtained directly proceeding from the abstraction of the Bernoulli formula. Indeed, hitting the bucket is a success, and missing it is a failure. The probability of success of the Bernoulli process is equal to $P[x]$. We obtained a binomial distribution where N acts as the number of degrees of freedom, and n — as the number of successes.

Check of the agreement between the formula and the numerical experiment has shown their coincidence to the accuracy of the calculation error.

Properties of the Solution. The distribution is discrete as to the independent variable and the limiting parameters. The bucket base can be any, including a discreet one. We assume the discreteness to be a serious advantage.

1. The probability o filling a very narrow bucket with anything is negligible. So, if $[x].R \to 0$, then $h(N >> 1, n \neq 0, [x]) \to 0$, a $h(N >> 1, n = 0, [x]) \to 1$. Further we will consider only the cases of $N >> 1$ illustrated in Fig. 2.

2. In the case of a wide base for the bucket $h(n = N, [x].R \to p(x).R) \to 1$; on the contrary, $h(n < N$ if $[x].R \to p(x).R) \to 0$. In particular, the bucket coinciding with the entire domain of definition of the distribution is always filled to overflowing. Formally, for the entire domain of definition of the tested distribution $h(n = N, p(x).R) = 1$ and 0 at other values of n. This case is illustrated in Fig. 3.

From this point of view, the quantile is a bucket with such a base that the probability of collecting all results reaches the normative value. Since the solution is not unique, the ambiguity can be decreased by additional restrictions, for example, by requiring connectedness of the bucket base and minimally of the base length.

3. At small values of N, the bucket can still be empty. Then, as N increases, the absolute value of the distribution width increases, and the relative value decreases (Fig. 4).

4. The shift variation of a fixed-width bucket follows, on average, the distribution form, as is shown in Fig. 5.

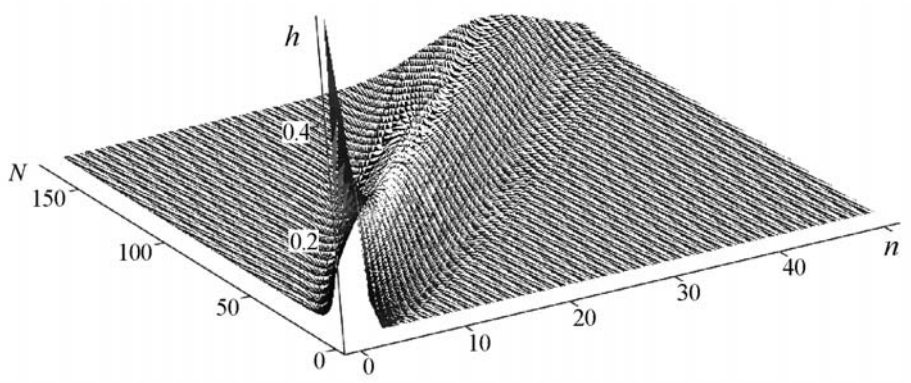

Fig. 4. Example of the distribution $h(n, N)$ for a fixed-base bucket.

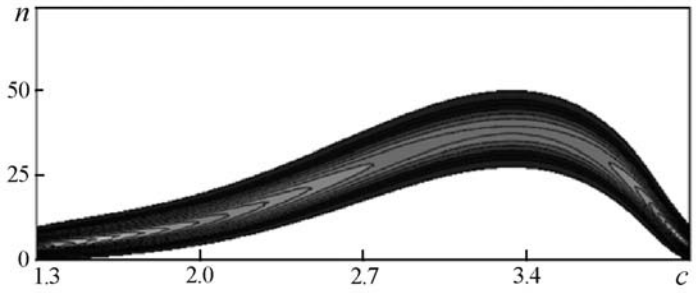

Fig. 5. Example of prediction of the height of a fixed-width bucket moving in the domain of definition of the distribution for $[c \mp 0.15]$, $c = 1.3 - 4.0$.

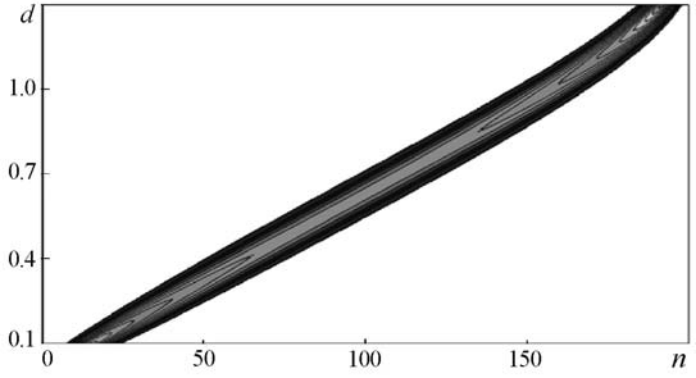

Fig. 6. Response of the bucket height distribution $[2.4 \mp 0.1 \mp d]$ to a widening of its base $d = 0 - 1.3$, $N = 200$.

5. Noteworthily, the dispersion of the bucket height prediction is wider at maximum values of the distribution density and decreases with decreasing density.

6. For a bucket beyond the domain of definition of the distribution, the height 0 is predicted with probability 1.

7. The response of the bucket height distribution to a widening of the bucket base is illustrated in Fig. 6.

8. The calculation for a divided-base bucket leads to a correct prediction. For instance, having given the bases of buckets $[x_\square]$ and $[x_\Delta]$, we can calculate two distributions of the heights for each of them $h(n, [x_\square])$ and $h(n, [x_\Delta])$, as well as the bucket height distribution on the common, though not connected, base $h(n, [x_\bigcirc]) = [x_\square] \cup [x_\Delta]$. This is admissible since the integral is additive over the integration domain. Figure 7 shows the above-mentioned distributions.

9. The function $h(n)$ can be interpreted as a distribution, since $\sum_{n=0}^{N} h(n) = 1$. The same holds also for a divided-base bucket.

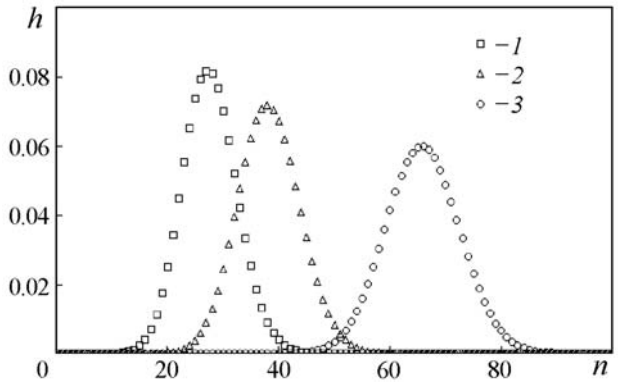

Fig. 7. Prediction of the height of the multibucket and initial buckets: 1) for a bucket with a base [2.6, 2.8]; 2) for a bucket [3.2, 3.4]; 3) for a multibucket consisting of buckets 1 and 2.

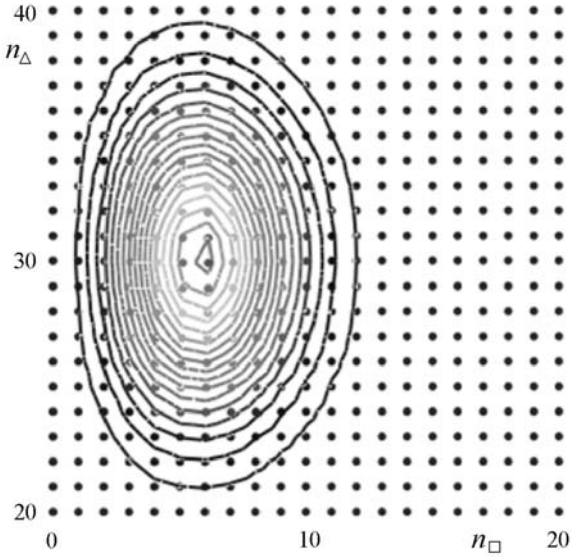

Fig. 8. Example of the contour diagram of the product of height distributions of two buckets $h(n_\Delta, n_\square)$ in bases $[x_\square] = [1, 2]$ and $[x_\Delta] = [3, 4]$, $N = 60$.

10. The product of the bucket height distributions $h(n_\Delta, n_\square) = h(n, [x_\Delta])h(n, [x_\square]$ whose contour diagram is given in Fig. 8, leads to a two-dimensional distribution, since $\sum_{n_\Delta=0}^{N} \sum_{n_\square=0}^{N} h(n_\Delta, n_\square) = 1$.

Apparently, generalization to a larger number of measurements can be performed. The obtained distribution can be interpreted as the probability that the presented histogram has been generated by a process having the distribution being tested. Hence we can immediately obtain the expression for the measure being constructed, since for an independent process the buckets will also be independent.

11. If $[x_\square]$ and $[x_\Delta]$ are mutually complementing (overlap the entire domain of definition of $p(x)$:$[x_\square] \cup [x_\Delta] = p(x).R$, as all as $[x_\square] = [p(x).R.\min, x_s]$ and $[x_\Delta] = [x_s, p(x).R.\max]$, then $h(n_\Delta, n_\square) = h(n, [x_\Delta])h(N-n, [x_\Delta]) = h(n, [x_\square])h(N-n, [x_\square])$, or $h(N-n, [x_\Delta]) = h(n, [x_\square])$, and, vice versa, $h(n, [x_\Delta]) = h(N-n, [x_\square])$.

Data Analysis. The input information for the data analysis is the experimental dataset and the a priori distribution model. The elementary goal of testing can be the measurement of the probability that the process that has generated the dataset has the distribution being tested. This elementary test is easy to carry out and can serve as the basis for formulating and solving more complex problems of testing hypotheses.

We first carry out the test by the trivial measure in order to exclude at once the obvious discrepancies. Actually, we check if there are buckets that must surely be empty but a datum has got into them, and vice versa. This check can be evaluated as a rough estimation of the shift and dispersion parameters.

The shift and dispersion parameters are interwoven into the notion of the partial reproducibility so deeply that in most cases exactly their values are investigated. The best variant of estimating these values is making use of the order measure for estimating the parameters designed for this purpose [2]. Knowing the shift and dispersion parameters, we can reduce the tested distribution to a standardized form; for example, we use the reduction of the domain of definition of the distribution to the interval [–1, 1]. Now the tested distribution is only characterized by its form, and the object of measurement is the estimation of the probability that the process has the same distribution form.

For data analysis, we can propose two techniques. For the simple problem being considered, this does not seem to be especially useful, but their importance increases for more complex problems. The first technique consists of the possibility to divide the bucket geometrically, leaving it single analytically; thus we sum the content of the arbitrarily arranged buckets. For two buckets, this technique has been illustrated above (property 8). By the formula

$$P\{[x]\} = \sum_{\{[x]\}} \int_{[x]_i} p(x)dx$$

we give the structure used as a single multibucket.

We can predict by this multibucket the distribution of its height $h(n) \mid P\{[x]\}, N$. If the H value has been obtained experimentally by calculating the number of times the value of the measured quantity hit the set of intervals $\{[x]\}$, we can, by comparing the prediction and the experimental results, give the value of the probability that the generating process distribution coincides with the theoretical one (holds exactly for this multibucket). Formally, if $H \in h(n).Q$, then the hypothesis can be accepted, and if $H \in h(n)$, then it has to be rejected.

The problem is that will have to carry out this procedure for each multibucket. Let the domain of definition of the distribution be arbitrarily split by a system of buckets $\{B\}$. Each bucket can be simple, i.e., its base is one interval, or a multibucket. It is reasonable to require that the bucket bases do not overlap; this will considerably simplify the calculations. The requirement is not too burdensome. Taking into account the overlap of buckets is also a simple problem, but there is no use of this, and overlaps create confusions.

For this system of buckets, we obtain an experimental histogram $H(B)$. For each bucket, we construct a prediction $\{h(n), B\}$. Comparing the experiment and the prediction, we obtain a system of estimates $\{h(H)\}$. All of them will have a different value. If we test the hypothesis separately for each bucket, we will obtain a set of contradicting results. To avoid this, let us make use of $\Pi\{h(H)\}$ for making the final decision. Here a difficulty of calculating the value of the criterion threshold arises.

There is an elegant way to avoid all problems at one stroke. The key is that we have to compare the whole distribution with all data. The obstacle which we have to overcome is the specific feature of predicting the bucket height, covering the entire domain of definition of the distribution, which is the tending to the delta function (property 2).

The second technique is as follows. We divide the domain of definition by any method only into two multibuckets. If the sums of their bases are comparable in value, then no problems connected with the prediction arise. Now note that $H_1 = N - H_2$ (property 11), from which we will combine the two predictions, having transformed one to the other, no matter in what direction. Let it be the transformation of the second prediction to the first one resulting in $h(n_1) = h_1(n_1)h_2(N - n_2)$. Now it suffices to make a decision as to $n(H_1)$ in order to obtain a response of the same quality as for the entire histogram. The reverse transformation is symmetric.

We can obviously generalize the method of transformations to several buckets, to a multi-dimensional model, and derive a recursive algorithm for the purpose, but we discovered nothing useful on this way and preferred a simple way of combining, although here there is no scope for mathematical search.

Relation between the Solution Methods for Problems. A small dataset will lead to the fact that almost all empty buckets will be predicted, which would seemingly complicate the situation, but this fact opens the transition to an order measure of estimating the distribution parameters. As important step is to order the dataset, equate the number of buckets to the number of their data, and bring the boundaries of buckets and the coordinates of ordered data into coincidence. Then $P[x] = P(d_i) - P(d_{i-1})$. To transform to the cumulative distribution, all these integrals should begin from $-\infty$. A simple replacement of the lower bound in all integrals is equivalent to the integration of the whole

expression $h(n, N, [x])$ predicting the bucket height. To compensate for this effect, it is necessary to differentiate the result. As a result, we will obtain precisely the kernel of the order measure [2]. Reversing the line of reasoning, we will discover a relation between the order measure and the binomial distribution.

It turns out that the estimation of parameters and the testing of hypotheses with respect to the distribution are based on the same principle of analysis of the combinations of data distributions. The independence of the data is realized as an arbitrary order of their analysis, which leads to the formulas of combinations.

The similarity of the methods will be even greater if we take into account that in practice we will have to work with buckets whose bases were predetermined a priori, mainly by hardware techniques, and, besides, their number is usually small (measuring techniques are chosen with a small resolution to spare). Their sorting is easy and rough, and often determine exactly the discreteness of the estimate.

Conclusions. The foregoing can be given as a simple linear algorithm $(p(x), \{d\}_N) \overset{\text{aTR}}{\to} (0 \,|\, x_s) \to h(H) \overset{Q(q)}{\to} (0 \,|\, 1)$, where by the data (a priori or experimental), by means of a trivial algorithm, a decision is made (to reject the hypothesis or continue the investigation having chosen the division point of the domain of definition). Then the probability distribution of a bucket with a base size twice smaller than the distribution domain for the number of experimental data that have hit this bucket is calculated. By the given significance threshold the final decision about the validity of the hypothesis is made.

The proposed algorithm of testing a simple hypothesis can obviously be generalized to a simple hypothesis with an alternative, for example, as a differential problem of estimating the difference between two standardized distributions (differing only in form). The main means for solving this problem is the splitting of the common domain of definition of the distribution into two multibuckets taking the best account of the differences between the distribution forms. Accordingly, the differential decision rule usually formulated in terms of the probabilities of taking a wrong decision when the hypothesis is true and vice versa is used.

Generalization to complex hypotheses can rely, for example, on the transformation of the set of distributions to a set of measured estimates and the construction of the decision rule by these estimates. The final result of these operations is the ordering or even parameterization of the set of tested distributions.

Since the algorithms for testing hypothesis and estimating parameters are actually isomorphous, we suppose that the advantage of the algorithms for estimating parameters is more universal and easier to interpret, the more so since the parameterization of a set of distributions is often a simple procedure and easy to carry out.

As an alternative, we can consider an algorithm using an adaptive system of buckets that is constructed at each comparison of distributions. By the metric of alternative hypotheses the matrix for all variants of distributions is constructed. As a point estimate, the most often accepted distribution is chosen. The confidence interval will consist of indices of all those hypotheses that are accepted at least once. Although the approach is cumbersome and fraught with internal conflicts, it may turn out to be useful for some specific problems.

NOTATION

AP, polynomial approximation algorithm; aTR, title of the trivial algorithm; B, abstract bucket, connected or not connected; c, variable describing the procedure of bucket widening; d, datum: an element of measuring instrument readings; $\{d\}_n$, dataset: a collection of n data processed as a single structure of data; ev, internal variable of the algorithm; h, probability of filling the bucket; H, bucket height or the number of data that have hit it; $.H$, representation of the object in the form of a histogram; i, index of a set of objects; m, probability measure; N, total number of data used to construct the whole histogram; n, number of data that have hit the bucket; $P(x)$, cumulative distribution function; $p(x)$, distribution density function; Q, distribution quantile; q, quantile parameter; $.R$, method of calculating the range of the object having such a property; the result is an interval number; x, variable denoting the measurand; x_s, point of division of the domain of definition of the distribution into two buckets; $[x]$, interval describing the bucket base; $\{[x]\}$, multibucket, set of buckets used as a single bucket; w, variable describing the variation of the bucket width. Subscripts: s, statistics.

REFERENCES

1. E. V. Chernukho, Comparison principle as an alternative of the substitution principle in metrology and statistics, in: *Heat and Mass Transfer-2008*, A. V. Luikov Institute of Heat and Mass Transfer of the National Academy of Sciences of Belarus, Minsk (2009), pp. 418–423.
2. E. V. Chernukho, Solution of the problem of estimating arbitrary distribution parameters by the data of a repetitive experiment, *Inzh.-Fiz. Zh.*, **83**, No. 2, 403–409 (2010).