International Journal for Quality Research 14(1) 147–162
ISSN 1800-6450

Mikhail Lomakin
Alexey Buryi [1]
Alexander Dokukin
Anatoly Strekha
Julia Niyazova
Alexander Balvanovich

Article info:
Received 19.02.2019
Accepted 10.06.2019

UDC – 519.766.4
DOI – 10.24874/IJQR14.01-10

ESTIMATION OF QUALITY INDICATORS BASED ON SEQUENTIAL MEASUREMENTS ANALYSIS

Abstract: *The paper considers the task of evaluating product quality indicators based on the results of measurements obtained during the control testing. Data processing is proposed to be carried out in two stages. First stage comprises data compression via cluster analysis, and the second stage uses the procedure of non-parametric estimation of the observed measurements to evaluate the quality of small samples with an unknown distribution law. Quality scores are defined as guaranteed scores on a set of distributions with moments equal to sample points found from a small sample. A number of theoretical statements are formulated, and a model example is given.*

Keywords: *Quality indicator; Data compression; Cluster analysis; Sample; Distribution function; Guaranteed estimation; Probabilistic moments.*

1. Introduction

The shift of priorities in value to product quality has always been the main driver of economic development. Currently, information and intellectual technologies play a significant role in this direction.

At the same time, the main motion vector for improvement and increase of quality is traditionally set by theoretical studies, which are the foundation for the development of new models and methods implemented in production technologies in various fields and directions of application. An interesting approach in this regard is the integration of the specifics of information systems with general methodological issues of a quality management system that measures, controls and analyzes the processes necessary to achieve the required results, or when developing a strategy for integrated management systems that combines quality, environment and safety management (Barbosa et al., 2018), including by improving methods of quality management on the example of hard quality management (Abdullah & Tari, 2017).

It is the business and commercial practice that in most tasks of managing economic aspects determine the quality as a degree of consumer expectations (Lomakin, 2017).

Analysis of complex technical-organizational system related with the definition of a large number of characteristics, which leads to complex quality indicators (Mironov et al., 2017). For distributed information and control systems, the concept of quality is associated with function targets, see (Buryi, 2016).

Methods of obtaining indirect estimates of quality indicators of material flows of technological processes are of interest (Grebenyuk & Itskovich, 2017).

[1] Corresponding author: Alexey Buryi
Email: a.s.burij@gostinfo.ru

While the process of transition "from resources to product" aimed at obtaining the required guaranteed product quality is taking place, the stability of the technological processes, typical for production is provided by the standardization of technologies and compliance with the customer requirements (Figure 1). Technology management is carried out, on the one hand, based on measurements, and, on the other, the required processing of the measurement information enables the formation of the estimations of the quality indicators defining the state of a product as a whole.

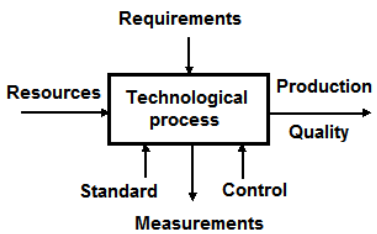

Figure 1. External and internal influences on the technological process of obtaining the required quality

A number of limitations on the amount of data lead researchers to the need to develop methods for obtaining the reliable results for the evaluation of the observed parameters for incomplete data, when receiving a full array of observations is either not available or economically unprofitable.

Tests of most products are carried out on pilot products batches. Based on the test results, a conclusion is given on the conformity or non-conformity of products to the accepted standards. The number of samples for testing depends on the required level of quality, as well as on the size of the controlled batch of products, that to a high extent determine the accuracy of the quality assessment task being solved, which can be determined by international standards organizations.

International standards are developed by international standards organizations. International standards are available for review and use worldwide. The commonly known organization is the International Organization for Standardization (ISO).

We assume sampled low when its volume is at least 25-30 observations, although in individual studies, for example, in medicine to the sample volume requirements are always unique (Vasileiou et al., 2018). Special criteria for certain distribution laws are known for small samples. This way, to test the normality hypothesis of a random variable distribution law, Shapiro-Wilk's test (Kapur & Lambson, 1977) is often used to describe the occurrence of failures caused by aging of materials, exponential distribution is used to specify sudden failures in reliability theory (Downton, 1970; Hahn & Shapiro, 1967), etc.

However, the establishment of the distribution law is not always a trivial task, and requires additional data or measurements, which are also not always available, especially at the stage of quality control of a new product or equipment.

It is proposed to divide the solution of the problem of reducing the initial set of observations into two successive stages. At the first stage, the initial data set will be structured based on the selected similarity measures into clusters containing objects with similar quality indicators. The resulting sample of the first stage will contain one object from each cluster.

At the final stage, we will consider the formed sample of the observed values as small with unknown distribution laws, which in the proposed approach can be any, and we will obtain estimates of the statistical parameters of the sample.

2. State of the Research Problem

Cluster analysis is widely used when comparing objects on many grounds, when it is necessary to classify them, to carry out statistical processing in order to identify trends, to solve a number of planning and forecasting problems.

Clustering refers to the process of splitting (distribution) of the set of observed objects

into clusters (subsets) of objects close by the selected ratio (similarity measure).

For the purpose of reducing the excess information obtained as a result of multivariate study of objects on a variety of features in the course of expert evaluation of the tests results it is proposed to use the methods of cluster analysis (Dalton et al., 2009; Duran and Odell, 1974).

A number of approaches are known. In particular, it is worth mentioning the algorithm of data reduction due to data mining based on the mimic algorithm of prototype selection from the class of evolutionary genetic algorithms (Derrac et al., 2010), as well as procedures for generating reduced data sets while maintaining the required level of representativeness using weighted average metrics (Kile & Uhlen, 2012). However, it should be noted that in practice of cluster analysis methods, the choice of the clustering criterion, the metrics themselves, the number of clusters, if they are not specified in advance, are determined by an expert, see e.g. (Shirkhorshidi et al., 2015).

Nonparametric statistical methods actively used in solving practical technical and economic problems are the basis for the evaluation of quality characteristics for small samples (Hahn & Shapiro, 1967).

For nonparametric methods, a priori there are no assumptions about the nature of the distribution of the studied data, which is described in a model application by the authors (Hollander et al., 2014). For particular problems, the authors Corder and Foreman, (2014) used Wilcoxon criterion for samples of varying volume to check for normality. The most productive area of research is the creation of a model toolkit. So for regression models in nonparametric estimation of observations the degree of smoothing polynomial is being selected for the purpose of accuracy of a study, see (Efron, 2014). The model of interval data analysis for obtaining guaranteed estimates is obtained in the works by (Kieffer et al., 2002).

The ISO/TR 10017:2003 Standard provides practical procedures and recommendations for application of statistical methods and formation of sample for the required level of quality in the evaluation of measured data. The procedure for determining the level of quality for the studied product batches is given in the standard ISO 14560:2007, which show that the purpose of controlling the product batch is in a provision of a given limit level of quality – LQL. The main factors that explain the necessity for selective control are restrictions on financial resources, since the cost of controlled (tested) products automatically includes the cost of control operations and the cost of production, see the series of ISO 2859 standards.

A small sample is used by researchers in cases where organization of a continuous or complete control is impossible, which is typical for quality control of a number of products. It should be remembered that for small samples it is assumed that for a sample obtained from a normally distributed population, the distribution of the sample mean also possesses the normality characteristics.

Given this hypothesis, assuming that the small sample is also normal, even in asymptotic sense, the application of distribution quantiles is justified, see (Zielinski, 2006). However, most of the works are related to the study of the influence of sample size on the accuracy of the estimation of the observed parameters, e.g. see Shore H. (1998), where censored samples are used to reduce of errors, which is typical for cases of information loss, which also indicates ineffectiveness of clarification of distribution laws of the observed random parameters. A paper by (Dougherty et al., 2011) provides estimation of the influence of the reduction of data volume through elimination of classification errors in bio-information systems, and, for instance, a paper by (Yan et al., 2014) covers samples with less coordinated data in conditions of their noisiness or distortion.

It should be noted that the formation of the primary sample (measurement data array) can have both a probabilistic nature, typical for most measurements, and a subjective basis, typical for the processes of expert evaluation.

Table 1 presents the comparative characteristics of these two approaches taken into account (USEPA, 2002), as well as the features of the approaches depending on the size of the measurement sample.

Table 1. Probabilistic approach compared to subjective selection

	Probabilistic approach	Subjective selection	Sample size
Advantages	Provides ability to calculate uncertainty associated with estimates; Provides reproducible results within uncertainty limits; Can handle decision error criteria	Can be very efficient with knowledge of the site; Easy to implement	For complete control: confidence in the product Sampling of different volumes: save on the cost of testing; reducing the time of getting results
Disadvantages	Random locations may be difficult to locate; An optimal design depends on an accurate conceptual model	Depends upon expert knowledge; Cannot reliably evaluate precision of estimates; Depends on personal judgment to interpret data relative to study objectives	For complete control: expensive for the manufacturer; applicable for nondestructive testing only Sampling of different volumes: development of control methods for small samples are required; there is no complete confidence that there are no defective products in a small sample

Significant limitations on the sample size are set by cases related to the control of objects on telemetric measurements providing functional diagnosis by incomplete data (Buryi et al., 1998b). This is implemented via semantic compression of information and the functional relationships between the estimated parameters taken into account, semantic content and results of processing of measuring information in the evaluation of the quality of the studied products.

For the case where the data do correspond to non-normal processes, a number of robust methods are developed and a system of indicators is proposed to obtain qualitative estimates of production capabilities, see (Wooluru et al., 2016).

3. Methodology

Since the advent of information and measurement data in the sense in which we understand it today, there are at least two directions in the development of applications: one is dictated by the necessity to involve the most measurement information available into the process of evaluating the observed parameters (quality indicators); the other is determined by the requirement of the data reduction, based on a number of practical limitations caused by the capabilities of data transmission channels, computing facilities, the cost of the experiment and so on.

3.1. General directions to sample reduction by data compression methods

Data compression is in some cases an important stage in processing a measurement sample, both during preprocessing, for example, when rejecting anomalous measurements, during calibration or when censoring a sample, and during the main stage of processing, for example, when clustering, identification, or obtaining estimates of observed parameters.

With the development of the methods of forming and coding information messages the data compression algorithms also change. There is a large number of review articles on this topic, in particular, see e.g. (Buryi et al., 1998a; Uthayakumar et al., 2018). The variety of compression methods is due to the characteristics of data transmission channels (wired or wireless, carrier frequency range, etc.) from the source to the recipient. As applied to quality management systems, we will base our analysis on syntactic compression, which is implemented via structural and statistical redundancy of measurement data.

The structural approach includes coding methods and methods of signature analysis (Buryi & Lovtsov, 1988), methods of measurement compression see e.g. (Yang et al., 2013). Structural compression methods provide compression with Compression Ratio (CR), which is given in equation (1):

$$CR = 100\left(1 - \frac{\text{No. of bits in uncmprd}}{\text{No. of bits in cmprd}}\right), \% \quad (1)$$

where "cmprd" and "uncmprd" corresponds to the abbreviation for the phrase "compressed data" and "uncompressed data", and is usually CR≈5÷10% of the input data.

Statistical Compression (SC) methods include:

- adaptive discretization methods;
- interpolation by polynomial dependencies;
- algorithms based on the functional expansion of Kalman filter (resistant to information loss, provide reduced dimension of the state vector of objects and other modifications of filters elements);
- approximation methods based on orthogonal functions (discrete Chebyshev transform, Walsh-Hadamard transform, Haar transform).

At the same time the value of the compression ratio is $CR_{SC} \approx 50 \ldots 100$ for a separate parameter, depending on the number of elements combined in a cluster.

3.2. Reduction of measurement samples based on cluster analysis

The main task of clustering is the formation of subsets of objects according to a certain cluster feature or property, for which various measures (relations) of similarity or difference are often used. This approach is widely used in the tasks of searching for analogs (prototypes) – see (Iglesias & Kastner, 2013), pattern recognition – (Duran and Odell, 1974), diagnostics – (Buryi et al., 1998b).

While performing an expert assessment of the quality of complex objects, for example, software, the total number of measurements can reach several hundred. That is explained by the large number of indicators themselves and the number of experts involved (an average of 8-12 people). In accordance with the system of standards ISO / IEC 25000: 2011, which sets out the requirements and assessments of software quality, there is a number of integrated quality indicators: functional suitability, performance level, reliability, etc.

Given a positive assessment of the components of functional suitability (Figure 2) (completeness, correctness, expediency), there is sometimes no need for further analysis of sub-characteristics, accumulating them in an indicator of a higher hierarchical level.

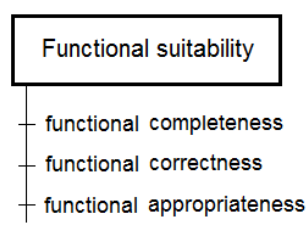

Figure 2. Sub-characteristics of functional suitability

The primary information obtained from the results of expert assessment is structured in the form of a table "object-feature", where the objects are the analyzed (tested) products, and the signs are the marks given by experts in accordance with a specific scale.

Let's denote the set of tested objects as $O_1, O_2, \dots, O_h \in \boldsymbol{O}$, each of which is described by a set of attributes (qualities) $Q_1, Q_2, \dots Q_q \in \boldsymbol{Q}$, which are estimated by experts, the total number of which is determined by the power of the set $\boldsymbol{E} = \{1,2, \dots, e\}$. Accordingly, for the measurement space $\boldsymbol{L}$ formed by the specified sets, i.e.

$$\boldsymbol{L} = \boldsymbol{O} \times \boldsymbol{Q} \times \boldsymbol{E},$$

the total measurement is $card(\boldsymbol{L}) = q \cdot h \cdot e$. Thus, each object is represented as a matrix $= \|l_{ij}\|_{q \times h}$, where l_{ij} is the value of attribute i for the j -th object, with ($i = \overline{1,q}$; $j = \overline{1,h}$; $e = 1$), i.e. an increase in the number of experts will lead to an increase in the rows of the matrix $\boldsymbol{L}$. On the other hand, the totality of the "object – feature" matrices represents a certain structure of relations.

At the same time, in view of the heterogeneity of individual signs, one can speak of the multidimensional scaling problem (MDS) – see also (De Leeuw & Mair, 2009). Usually, the measure of similarity between two objects a and a^s, where the index s corresponds to the image of some standard, with which the test sample is compared, is projected in the metric space over the distance between these objects or their properties (features). The distance is represented in absolute scale as a pairwise comparison of the properties of the object being measured with a standard image. In addition, the distance must satisfy metric axioms such as identity, symmetry, and triangle. Using the example of Euclidean space and manifolds of metrics (Choi et al., 2010), partially presented in Table 2 (for two data points x and y in n-dimentiions space), the matrix of objects that are closest in their properties to standard objects, i.e. $d_{ij}(L) \leq \Delta_{ij}$, where

$$d_{ij}(L) = \left(\sum_{i=1}^{q} \left(a_{ij} - a_{ij}^{s}\right)^2 \right)^{\frac{1}{2}}, \quad (2)$$

where Δ_{ij} are the requirement levels for each property of the corresponding j-th object.

Table 2. Definitions of the various measures

Distances	Formalization
Euclidean	$d_{euc}(x,y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}$
Average Distance	$d_{ad}(x,y) = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2 \right)^{\frac{1}{2}}$
Weighted Euclidean	$d_{we}(x,y) = \left(\sum_{i=1}^{n} w_i (x_i - y_i)^2 \right)^{\frac{1}{2}}$ where w_i is the weigth given to the i-th component
Manhattan	$d_{mh}(x,y) = \sum_{i=1}^{n} (x_i - y_i)$

To determine the object that is closest in its properties (features) to the standard sample, we use the minimization of a functional called stress (De Leeuw & Mair, 2009):

$$S(L) = \sum_{i<j} w_{ij} \left(\Delta_{ij} - d_{ij}(L)\right)^2 \rightarrow \min \quad (3)$$

Here, $W = \|w_{ij}\|$ is a known symmetric non-negative weight matrix, and the weights are selected based on the scaling goals. Usually weights are chosen from the physical sense. For example, $w_{ij} = \Delta_{ij}^{\gamma}$, a $\gamma < 0$, and $\gamma < 0$, if for an exact approximation smaller distances between objects are preferable and $\gamma > 0$ otherwise.

The stress $S(L)$ acquires the physical meaning of the potential energy at $\gamma = -2$ for a system of n connected points, then the equation (3) corresponds to the search for the equilibrium state of the system in which the potential energy is minimal.

The choice of clustering algorithm is based on factors such as the nature of the application, the characteristics of the analyzed objects, the expected number and shape of clusters, as well as the complexity of the task in

comparison with the available computing power. The sample reduced by cluster analysis be investigated below.

3.3 Probabilistic approach to the assessment of small measuring samples

Some cases of quality evaluation for technical, economic and other processes employ indicators like: $(\xi \geq (<)\varepsilon)$, where P – is a probability that a random value ξ is not less (less) than a set value ε. Such indicator is used to evaluate probability of a fault-free operation of systems, probability of company profits or project portfolio profitability to be not less than the target one or risk to be lower than a given level, etc.

Within probability theory (Gnedenko, 2005) it is common to use the following estimate for an arbitrary distribution $F(t)$ with a known mathematical expectation *m* and variance σ^2:

$$P(|\xi - m| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2}, \qquad (4)$$

where ξ – a random value with an unknown law of distribution $F(t)$; ε – an arbitrary positive value.

Despite being universal, relationship (4) provides rough estimates. Improvements and refinements of the estimate (4) a fairly large number of papers are devoted. The first paper solving the task of finding extreme or guaranteed (upper and lower) estimates of the probabilities $P(\xi \geq \varepsilon)$ for a known expectation and variance is solved is considered to be the work (Germeyer et al., 1966), the important direction of the guaranteed estimation, and also its comparison with a statistical estimation is considered in works of A.B. Kurzhansky and, in particular, Kurzhanski & Khapalov, (1986).

The lower and upper of estimates probability for approach (1) are equal, respectively:

$$P(\xi \geq \varepsilon) = \frac{(m-\varepsilon)^2}{(m-\varepsilon)^2+\sigma^2};\ \varepsilon \leq m, \qquad (5)$$

$$P(\xi \geq \varepsilon) = min\left\{1,\ \frac{\mathrm{m}}{\varepsilon},\ \frac{\sigma^2}{(\mathrm{m}\text{-}\varepsilon)^2+\sigma^2}\right\}. \qquad (6)$$

However, the results obtained in the paper (Germeyer et al., 1966), cannot be generalized onto arbitrary number of moments of distribution. Such generalization of the results onto any number of moments was obtained in the paper (Lomakin, 1991).

Let us define a set of distribution functions with given moments $m = m_1, m_2, \dots, m_k$ in the following form:

$$F_0 = \left\{F(t): \int_{-\infty}^{\infty} t^i dF(t) = m_i; i = \overline{1,k}\right\}. (7)$$

The following statement may then be formulated (Lomakin, 1991).

Statement 1. The upper (lower) value of an integral

$$J(F) = \int_0^{\tau+0} c(t)\, dF(t) \qquad (8)$$

with continuous sub-integral function $c(t)$ having $k+1$ non-negative derivative and $F(t) \in F_0$ is achieved with single step-function distribution, which has a point τ among the points of growth $t_1, t_2, \dots, t_v$; with uneven k the number of growth points v in the distribution function $F(t)$ is determined with a relationship $v = (k+3)/2$, which has $t_0 = 0 < t_1 < t_2 < \cdots < t_v < \infty$; with even k the number of growth points v in the distribution function $F(t)$ is determined with a relationship $v = \frac{k}{2} + 1$, which has $0 < t_1 < t_2 < \cdots < t_v < \infty$; numbers $p_j > 0$, t_j, $j = 1,2, \dots, v$, comply with the following

$$m_i = \sum_{j=1}^{v} t_j^i\, p_j;\ \ i = \overline{0,k};\ \ p_0 = 0. \quad (9)$$

The relationship (9) provided a way to directly find upper (lower) (or guaranteed) probability estimates for a random variable (quality characteristic values, etc.) to be higher than a certain given level ε, i.e. the value of probability $P(r \geq \varepsilon)$.

Let us consider a case when first two moments distribution of a random variable m_1, m_2 are known (and, certainly, $m_0 = 1$. Let us write down the equations for the moments

with a condition that the guaranteed probability estimate $P(r \geq \varepsilon)$ or $F(\varepsilon) = 1 - P(r \geq \varepsilon)$ is obtained with step-function distribution $F(t)$ having $v = 2$ growth points. This way we have:

$$\begin{cases} p_1 + p_2 = 1, \\ p_1 t_1 + p_2 t_2 = m_1, \\ p_1 t_1^2 + p_2 t_2^2 = m_2, \end{cases} \quad (10)$$

a set of three equations with four unknowns. It is apparent that

$$F^x(\varepsilon) = \begin{cases} 0 \ when \ \ 0 \leq \varepsilon \leq t_1, \\ p_1 when \ \ t_1 \leq \varepsilon \leq t_2, \\ 1 \ \ when \ \ \ \varepsilon \geq t_2. \end{cases} \quad (11)$$

The last relationship enables us to drop one unknown from the set of equations (10) and this way obtain a guaranteed estimate $F^x(\varepsilon)$. Since $F(\varepsilon) = F(\varepsilon - 0)$, i.e. the distribution function is a left-tail continuous function, the upper value of the integral (8) takes the form:

$$J(F) = F(\tau) \int_0^{\tau+0} dF(t), \quad (12)$$

given $F(t) \in F_0$ is obtained with step-function distribution $F(t)$, which has a point ε among the growth points t_1 , t_2. The relationship (11) demonstrates that the case where $F^x(\varepsilon) = p_1$ is the most interesting one. Then we suppose that $t_1 = \varepsilon$ and solve a set of three equations with three unknowns:

$$\begin{cases} p_1 + p_2 = 1, \\ p_1 \varepsilon + p_2 t_2 = m_1, \\ p_1 \varepsilon_1^2 + p_2 t_2^2 = m_2, \end{cases} \quad (13)$$

from which we derive the following, provided a series of plain transformations:

$$F^x(\varepsilon) = \begin{cases} 0 & when \ \ \varepsilon = 0, \\ \dfrac{m_2 - m_1^2}{m_2 - 2m_1\varepsilon + \varepsilon^2} & when \ \ 0 \leq \varepsilon \leq m_1, \\ 1 & when \ \ \ \varepsilon \geq m_1. \end{cases} \quad (14)$$

$$P(\varepsilon) = \begin{cases} 1 & when \ \ \varepsilon = 0, \\ \dfrac{(m_1 - \varepsilon)^2}{m_2 - 2m_1\varepsilon + \varepsilon^2} & when \ \ 0 \leq \varepsilon \leq m_1, \\ 0 & when \ \ \ \varepsilon \geq m_1. \end{cases} \quad (15)$$

The latter result was first obtained using a special method (Germeyer et al., 1966), which does not allow further generalizations and does not work for three or more moments.

Let $k = 3$, i.e. three moments are known – m_1, m_2, m_3. Following the same judgement as with two moments it is not difficult to obtain the following result:

$$P(\varepsilon) = \begin{cases} 1 - p_1 \ \ for & \varepsilon = 0, \\ 1 - p_1 - p_2 \ \ for \ \ 0 \leq \varepsilon \leq min\left(\dfrac{m_2}{m_1}; \dfrac{m_3}{m_2}; \sqrt{\dfrac{m_3}{m_1}}\right), \\ 0 \ \ for & \varepsilon \geq min\left(\dfrac{m_2}{m_1}; \dfrac{m_3}{m_2}; \sqrt{\dfrac{m_3}{m_1}}\right). \end{cases} \quad (16)$$

The most cunning case is when $P(\varepsilon) = 1 - p_1 - p_2$. Skipping intermediate calculations, we obtain the final relationship for the guaranteed estimate $P(\varepsilon)$:

$$P(\varepsilon) = \frac{3m_2 m_1^2 \varepsilon^2 - 3m_1 m_2^2 \varepsilon - m_1^3 \varepsilon^3 + m_2^3}{2m_2^2 \varepsilon^2 + m_3^2 - m_1 m_3 \, \varepsilon^2 - 3m_2 m_3 \varepsilon - m_1 m_3 \varepsilon^3}. \quad (17)$$

For any number of fixed moments of distribution $k > 3$ the guaranteed estimates may be obtained in the same way as in the case when two or three distribution moments are known. In practice, finding such estimates requires numerical methods (Lomakin, 1991).

One of the open questions left unanswered while using *statement* 1 is the quantity of moments required for finding the estimates of quality indicators.

4. Main results

Let $\rho = (r_1, r_2, \dots, r_n) \in R^n$, be a sample of values of a certain parameter characterizing the quality of a certain process, for instance, mean time to failure of a system. The elements of the sample r_i>0 are independent variables with the same distribution from a certain unknown distribution $F(t)$. Let's define a set F_1 as a set of all possible distribution functions c, from which a sample ρ may be derived, i.e. we define the set of distribution functions F_1, as follows:

$$F_1 = \{F(t): F^{-1}(\xi_i) = r_i\}. \quad (18)$$

The notation $F^{-1}(\xi_i) = r_i$ should be understood as the solution of an equation $(r_i) = \xi_i$, which has ξ_i as an implementation of evenly distributed random value r within a range of $[0,1]$.

Suppose it is required to obtain the upper and the lower estimate (boundary) of a distribution function $F(t)$ for a given $t = const$ within a set F_1 or to find guaranteed estimates for a distribution function within a set F_1, i.e. find such $F_x(t)$ and $F^x(t)$ that:

$$\begin{aligned} F_x(t) &= \min_{F(t)\in F_1} F(t); \\ F^x(t) &= \max_{F(t)\in F_1} F(t). \end{aligned} \quad (19)$$

Basing on the sample ρ let us define n sample moments of distribution $F(t)$ in the following relationships:

$$m_i = \frac{1}{n}\sum_{j=1}^{n} r_j^i. \quad (20)$$

Let us define a set of distribution functions F_0, which have distribution moments equal to sample moments obtained based on the sample ρ using relationships (20), i.e.:

$$F_0 = \left\{F(t): \int_0^{\infty} t^i \, dF(t) = m_i; i = \overline{1,n}\right\}.$$

Let's consider a task: within a set F_0 find the lower and upper estimate (boundary) for a distribution function $F(t)$ with a given t, i.e. find such $F_x(t)$ and $F^x(t)$, that:

$$\begin{aligned} F_x(t) &= \min_{F(t)\in F_0} F(t); \\ F^x(t) &= \max_{F(t)\in F_0} F(t). \end{aligned} \quad (21)$$

The following statement may be formulated.

Statement 2. Tasks defined by relationships (19) and (21) are equivalent to each other.

Proof. In order to prove the statement, it is required to demonstrate that the sets of distributions F_0 and F_1 are equal to each other. The equality of two sets F_0 and F_1 is understood as an identical equality, i.e. it means that every element of one set belongs to the other set and vice versa.

Let $F_1(t)$ be a certain function such that $F_1(t) \in F_1$, i.e. the sample ρ could be obtained from the distribution $F_1(t)$. Each sample value $r_i \in \rho$ can be considered as a solution to the equation $(r_i) = \xi_i$, in which ξ_i is the realization of a uniformly distributed quantity within the interval $[0,1]$. Let's prove that $F_1(t) \in F_0$.

Since $\boldsymbol{F_0}$ is a set of distribution functions, in which the first $\boldsymbol{n}$ moments are equal to the sample moments determined from the sample $\boldsymbol{\rho}$ using relationship (20), the statement that $\boldsymbol{F_1(t)} \in \boldsymbol{F_0}$ is then obvious.

Let's assume $F_0(t) \in F_0$ and $F_0(t)$ to be the distribution function for which the first n moments are equal to the sample moments. Let's prove that $F_0(t) \in F_1$, i.e. it may be the distribution function from which the sample ρ was obtained. In order to do that, it is sufficient to prove that the sample moments $m = (m_1, m_2, \dots, m_n)$ uniquely determine the sample ρ. Let's prove the following statement.

Statement 3. Let $X^* = \{x_1^*, x_2^*, \dots, x_n^*\}$ be one of the (possibly complex) solutions of a set of equations:

$$\begin{cases} \boldsymbol{x_1 + x_2 + \dots + x_n = b_1}; \\ \boldsymbol{x_1^2 + x_2^2 + \dots + x_n^2 = b_2}; \\ \dots \dots \dots \dots \dots \dots \\ \boldsymbol{x_1^n + x_2^n + \dots + x_n^n = b_n}, \end{cases} \quad (\mathbf{22})$$

then the system of equations (19) does not have other solutions except for those obtained through a number of rearrangements used to impact the set $\boldsymbol{X^*}$.

Proof of Statement 3. Let's introduce the following notation:

$$s_k = x_1^k + x_2^k + \dots + x_n^k ; k = 1,2, \dots, n .$$

For any k the polynomial s_k is a symmetric polynomial, i.e. such polynomial that does not change under any permutation of variables and therefore, as stated by the main theorem on symmetric polynomials (Okunev, 1966), can be uniquely represented as a polynomial from elementary symmetric functions:

$$\begin{cases} \boldsymbol{\sigma_1 = x_1 + x_2 + \dots + x_n}; \\ \boldsymbol{\sigma_2 = x_1x_2 + x_1x_3 + \dots + x_{n-1}x_n}; \\ \dots \dots \dots \dots \dots \dots \\ \boldsymbol{\sigma_{n-1} = x_1x_2 \dots x_{n-1} + x_1x_2 \dots x_{n-2}x_n + \dots + x_2x_3 \dots x_n}; \\ \boldsymbol{\sigma_n = x_1x_2 \dots x_n} . \end{cases}$$

For power sums of s_k there exists a unique inverse representation expressing elementary symmetric polynomials in terms of power sums. This representation is given by Waring's second formula (Okunev, 1966):

$$\sigma_k = \sum \frac{(-1)^{j_1+\dots+j_k+k}}{1^{j_1}2^{j_2} \dots k^{j_k} j_1! j_2! \dots j_k!} s_1^{j_1} \times \dots \times s_k^{j_k} ; k = 1,2, \dots, n ,$$

where the summation applies to all sets of non-negative integers $j_1, j_2, \dots, j_n$ with the property: $j_1 + 2j_2 + \dots + nj_n = k$

Thus, the system of equations (22) is equivalent to the following system

$$\begin{cases} \boldsymbol{c_1 = x_1 + x_2 + \dots + x_n}; \\ \boldsymbol{c_2 = x_1x_2 + x_1x_3 + \dots + x_{n-1}x_n}; \\ \dots \dots \dots \dots \dots \dots \\ \boldsymbol{c_{n-1} = x_1x_2 \dots x_{n-1} + x_1x_2 \dots x_{n-2}x_n + \dots + x_2x_3 \dots x_n}; \\ \boldsymbol{c_n = x_1x_2 \dots x_n} . \end{cases} \quad (\mathbf{23})$$

where

$$c_k = \sum \frac{(-1)^{j_1+\dots+j_k+k}}{1^{j_1}2^{j_2} \dots k^{j_k} j! j_2! \dots j_k!} b_1^{j_1} \dots b_k^{j_k}; k = 1,2, \dots, n . \quad (24)$$

Let's consider the equation:

$$x^n + a_1 x^{n-1} + \dots + a_{n-1}x + a_n = 0, \quad (25)$$

where $a_k = (-1)^k c_k; k = 1,2, \dots, n.$

Since the set X^* of possible solutions satisfies the system of equations (22), and, consequently, the system of equations (25), then, as it follows from the inverse theorem of Viet, all elements of the set X^* are the roots of the equation (25).

Let us suppose further that the system of equations (23) in addition to the set of roots X^* has a set of roots $X_* \neq X^*$. Since, as proved above, all the elements of a set X^* are the roots of equation (25), we thereby conclude that all elements of the set

$$X_*^* = X^* \cup X_*$$

are the roots of equation (22), and due to the inequality $X_* \neq X^*$ the cardinal number of the set X_*, i.e. $card\ X_* > n$.

However, this contradicts the main theorem of algebra, which states that equation (25) has exactly n roots.

This contradiction proves Statement 3.
Statement 3 proves that Statement 2 is justified.
Consequently, instead of the original problem defined by relationship (19), we can solve the problem defined by relationships (21), i.e. we could consider the problem of finding the upper and lower bounds of the distribution function on a set of distributions, in which n moments are equal to n sample moments found from the relationship (22) for sample ρ with the volume of n dimensions.
Let's consider an algorithm for the numerical solution of the problem of finding guaranteed estimates of the distribution function, defined by relationships (21).

To stay distinct, we will consider the problem of finding the upper estimate of the distribution function. The upper bound for the distribution function $F(t)$ for a given $t = \tau$ in accordance with the above result is achieved on a discrete distribution. Let $t_1, t_2, \dots, t_v$ be the growth points of the distribution function; $p_1, p_2, \dots, p_v$ are the values of the growth (jump) of the distribution function at the corresponding points. Then the task may be rewritten in the form:

$$F^x(\tau) = \max_{p_j} F(\tau) = \sum_{j=1}^{d} p_j \qquad (26)$$

under conditions

$$\sum_{i=1}^{v} t_i^j p_i = m_j; j = \overline{0,k}, \ m_0 = 1. \qquad (27)$$

The relationship (26) p_d is the value of growth (jump) of the distribution function $F(t)$in the point where $t = \tau$.

The problem defined by relationships (26) and (27) is a nonlinear multidimensional programming problem. We solve this problem of multidimensional programming by the following iterative method.

Let's introduce some set of growth points of the distribution function $T_s = (t_1, t_2, \dots, t_{vs})$, e.g. in the following way: $t_g = \delta g$. In the latter relation, the index g takes values 1, 2, ..., vs; δ – is a constant value – a step of discretization. Let $\pi_1, \pi_2, \dots, \pi_{vs}$ be the growth (jump) values of the distribution function at these points, and for some growth points from the set T_s, the growth may be zero.
Let us consider the problem: find the upper boundary of the distribution function

$$F^x(\tau) = \max_{p_j} F(\tau) = \sum_{j=1}^{d} p_j \qquad (28)$$

under conditions

$$\sum_{i=1}^{vs} t_i^j \pi_i = m_j; j = \overline{0,k}, \ m_0 = 1. \qquad (29)$$

In relation (28), π_{ds} is the growth rate of the distribution function at the point t_{ds}, moreover, $t_{ds} \le \tau \le t_{ds+1}$. The latter problem, defined by relations (28) and (29), is a linear programming problem and can be solved using standard software packages for solving linear programming problems.

Then we reduce the discretization step, for example, though half-division and solve the linear programming problem each time until the difference between the value of the distribution function $F^x(\tau) \approx F_h^x(t_{ds})$ in the previous and subsequent $F^x(\tau) \approx F_{h+1}^x(t_{ds})$ steps is material, i.e. until module of the difference meets the following condition

$$|F_{h+1}^x(t_{ds}) - F_h^x(t_{ds})| > \gamma, \qquad (30)$$

where $\gamma > 0$ is a small value determining the precision of estimation of the upper bound of a distribution function; $h, h+1$ – previous and next step in the solution of linear programming problems.

Let's consider an example of usage of the proposed method for determining the upper estimate of the distribution function. Let's assume that we have data on a certain process represented by four dimensions. Basing on this sample of four dimensions, estimations of four moments m_1 , m_2, m_3, m_4 have been found. The source data for the estimates of the moments are presented in table 3.

Table 3. Data on estimates of moments

m_1	m_2	m_3	m_4
2,85	9,45	34,05	129.45

Let us define the set of points of growth for the distribution function $T_s = (1,2, \dots ,5)$, then for δ=1 we will assume $t_g = g$, and $g = \overline{1,5}$. Let $\pi_1, \pi_2, \dots, \pi_5$ be the growth (jump) of the distribution function at these points.

It is necessary to find the upper estimate of the distribution function $F(\tau)$ with $\tau = 2$. The accuracy of the estimate of the distribution function is set equal to $\gamma = 0{,}01$.

Let us describe the solution of the problem in the form of a sequence of steps.

Step 1. Let's write down the problem of finding the upper estimate (boundary) of the distribution function in expanded form: find

$$F^x(2) = \max(\pi_1 + \pi_2) \qquad (31)$$

under conditions:

$$\pi_1 + 2\pi_2 + 3\pi_3 + 4\pi_4 + 5\pi_5 = 2{,}85 \qquad (32)$$

$$\pi_1 + 4\pi_2 + 9\pi_3 + 16\pi_4 + 25\pi_5 = 9{,}45 \qquad (33)$$

$$\pi_1 + 8\pi_2 + 27\pi_3 + 64\pi_4 + 125\pi_5 = 34{,}05 \qquad (34)$$

$$\pi_1 + 16\pi_2 + 81\pi_3 + 256\pi_4 + 625\pi_5 = 129{,}45 \qquad (35)$$

As a result of solving this linear programming problem we obtain $F^x(2) = 0{,}30$.

Step 2. The expanded form of the task of finding the upper estimate of the distribution function is not given here due to its «bulkiness»; it is determined by relationships similar to (14) – (18).

We reduce the discretization step $\delta = 0{,}5$. The set of points of growth of the distribution function $T_s = (0{,}5;\ 1;\ 1{,}5; \dots ; 4{,}5;\ 5)$.

As a result of the repeat solution of the linear programming problem with the changed initial data, the estimat is obtained $F^x(2) = 0{,}475$.

Step 3. We reduce the discretization step to $\delta = 0{,}25$ on the interval [1,5] and obtain, respectively the set of growth points of the distribution function $T_s = (0{,}25; 0{,}5; \dots ; 5)$.

As a result of the solution of the linear programming problem the estimate is obtained $F^x(2) = 0{,}514$.

Step 4. We reduce the discretization step to $\delta = 0{,}125$ on the interval [1,5] and obtain, respectively the set of growth points of the distribution function $T_s = (0{,}125; 0{,}25; \dots ; 4{,}75; 5)$. As a result of the solution of the linear programming problem the estimate is obtained $F^x(2) = 0{,}528$.

Step 5. We reduce the discretization step $\delta = 0{,}0625$ on the interval [1,5] and obtain, respectively the set of growth points of the distribution function $T_s = (0{,}0625; 0{,}125; \dots ; 49375; 5)$. As a result of the solution of the linear programming problem the estimate is obtained $F^x(2) = 0{,}534$.

For this step we find $\gamma = 0{,}006$. With this the solution of the task is complete. In order to obtain a more precise estimate a smaller discretization step may be used.

5. Conclusion

A comprehensive two-step data mining approach for product quality control is proposed. It is shown that at the first stage of processing the measurement information, by applying cluster analysis methods and attracting well-known proximity measures, the amount of data can be reduced to the level when it may be identified as a small sample in the accepted terminology. To determine the quality indicators, a method for determining guaranteed estimates on a set of distributions with given moments equal to sample moments is proposed. It is shown that in order to obtain a guaranteed assessment of the quality indicator in a case of small samples, a number of distribution points should be used equal to the sample size. It is also proposed to use an algorithm for determining quality indicators in the context of incomplete information, which is based on the numerical solution of a sequence of linear programming problems. The algorithm was practically tested for finding the upper estimate of the distribution function at four known points.

The upper estimate of the distribution function is found with the required accuracy in the numerical solution of a sequence of linear programming problems until a given level of accuracy of quality estimates is achieved.

References:

Abdullah, M. M. B., & Tari, J. J. (2017). Hard quality management and performance: The moderating role of soft quality management. *International Journal for Quality Research*, *11*(3), 587-602. doi: 10.18421/IJQR11.03-07

Barbosa, L. C. F. M., Oliveira, O., Santos, G. (2018). Proposition for the alignment of the integrated management system (quality, environmental and safety) with the business strategy, *International Journal for Quality Research*, *12*(4), 925-940. DOI: 10.18421/IJQR12.04-09

Buryi, A. S. (2016). *Fault-tolerant distributed systems of the processing of information*. Moscow: Hot line – Telecom.

Buryi, A. S., & Lovtsov, D. A. (1988). Telemetry system with information compression. *Patent RF*, no. 1425754.

Buryi, A. S., Loban, A. V., & Lovtsov, D.A. (1998a). Compression models for arrays of measurement data in an automatic control systems. *Automation and Remote Control*, *59*(5), Pt 1, 613-631.

Buryi, A. S., Polous, A. I., & Shlyakonov, V. A. (1998b). A functional diagnostic method for program-controlled objects, *Automation and Remote Control*, *59*(4), Pt 2, 599-602.

Choi, S-S., Cha, S-H., Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, *8*(1), 43-48.

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: a step-by-step approach*, 2nd edition. John Wiley & Sons.

Dalton, L., Ballarin, V., & Brun, M. (2009). Clustering algorithms: on learning, validation, performance, and applications to genomics. *Current Genomics*, *10*(6), 430-445. doi: 10.2174/138920209789177601

De Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, *31*(3), 1-30. doi: 10.18637/jss.v031.i03

Derrac, J., García, S., & Herrera, F. (2010). Stratified prototype selection based on a steady-state memetic algorithm: a study of scalability. *Memetic Computing*, *2*(3), 183-199. doi: 10.1007/s12293-010-0048

Dougherty, E. R., Zollanvari, A., & Braga-Neto, U. M. (2011). The illusion of distribution-free small-sample classification in genomics. *Current Genomics*, *12*(5), 333-341, doi: 10.2174/138920211796429763

Downton, F. (1970). Bivariate exponential distributions in reliability theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, *32*(3), 408-417.

Duran, B. S., & Odell, P. L. (1974). *Cluster analysis: A survey*. New York: Springer-Verlag.

Efron, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association*, *109*(507), 991-1007. doi: 10.1080/01621459.2013.823775

Germeyer, Y. B., Inger, D. S., & Kalabukhova, E. P. (1966). On guaranteed estimates of system reliability with incomplete information about the reliability of the elements. *Computational mathematics and mathematical physics*, *6*(4), 733-747.

Gnedenko, B. V. (2005). *Course of probability theory*, Eighth edition, Editorial URSS, Moscow.

Grebenyuk, E. A., & Itskovich, E. L. (2017). Indirect estimations of current quality indicators of the process unit material flows and their correlation. *Tenth International Conference Management of Large-Scale System Development (MLSD)*, Moscow, 1-4. doi: 10.1109/MLSD.2017.8109630

Hahn, G. J., & Shapiro, S. S. (1967). *Statistical models in engineering*. New York: John Wiley & Sons.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). *Nonparametric statistical methods*, 3rd edition. New York: John Wiley & Sons.

Iglesias, F., & Kastner, W. (2013). Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, *6*, 579-597. doi: 10.3390/en6020579

ISO/TR 10017:2003. *Guidance on statistical techniques for ISO 9001:2000*.

ISO 14560:2004. *Acceptance sampling procedures by attributes - Specified quality levels in nonconforming items per million*.

Kapur, K. C., & Lamberson, L. R. (1977). *Reliability in engineering design*. New York: John Wiley & Sons.

Kieffer, M., Jaulin, L., & Walter, E. (2002). Guaranteed recursive non-linear state bounding using interval analysis. *International Journal of Adaptive Control and Signal Processing*, *16*(3), 193-218. doi: 10.1002/acs.680

Kile, H., & Uhlen, K. (2012). Data reduction via clustering and averaging for contingency and reliability analysis. *International Journal of Electrical Power & Energy Systems*, *43*(1), 1435-1442. doi: 10.1016/j.ijepes.2012.07.011

Kurzhanski, A. B., & Hapalov, A. Yu. (1986). On the state estimation problem for distributed systems. *Lecture notes in control and information sciences*, *83*, 102-113. doi: 10.1007/BFb0007551

Lomakin, M. I. (1991). Guaranteed bounds on failfree operation probability in the class of distributions with fixed moments. *Automation and remote control*, *52*(1), 126-131.

Lomakin, M. I. (2017). Identification of impacting factors of foreign-economic activity of Russian Federation on economic growth using econometrics modeling. *Economy and socium*, *3*(34), 35-39. (in Russian)

Mironov, A. N., Shestopalova, O. L., Niyazova, Yu. M., & Sidorov, D. A. (2017). Evaluation of the quality of transportation on the basis of together probabilistic and statistical and fuzzy expert estimations of indicators consumer properties transport services. *Information and economic aspects of standardization and technical regulation*, *6* (40), 13.

Okunev, L. Ya. (1966). *Higher algebra*. Moscow: Prosveschenie.

Shirkhorshidi, A. S., Aghabozorgi, S., & Wah, T. Y. (2015). A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data. *PloS one*, *10*(12), e0144059. doi: 10.137/journal.pone.0144059

Shore, H. (1998). Approximating an unknown distribution when distribution information is extremely limited, *Communication in Statistics - Simulation and Computation*, *27*(2), 501-523. doi: 10.1080/03610919808813492

USEPA. (2002). *Guidance for choosing a sampling design for environmental data collection. EPA/240/R-02/005*. US Environmental Protection Agency, Washington.

Uthayakumar, J, Vengattaraman, T., & Dhavachelvan, P. (2018). A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications. *Journal of King Saud University - Computer and Information Sciences*. doi: 10.1016/j.jksuci.2018.05.006 (Accessed 25 January 2019).

Vasileiou, K., Barnett, J., Thorpe, S., & Young, T. (2018). Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period, *BMC medical research methodology*, *18*(1), 148. doi: 10.1186/s12874-018-0594-7

Wooluru, Y., Swamy, D. R., & Nadesh, P. (2016). Process capability estimation for non-normally distributed data using robust methods - a comparative study. *International Journal for Quality Research*, *10*(2), 407-420. doi: 10.18421/IJQR10.02-11

Yan, Y., Chen, L. J., & Zhang, Z. (2014). Error-bounded sampling for analytics on big sparse data. *Proceedings of the VLDB Endowment*, *7*(13), 1508-1519. doi: 10.14778/2733004.2733022

Yang, F., Chia, N., White, B. A., & Schook, L. B. (2013). Compression-based distance (CBD): a simple, rapid, and accurate method for microbiota composition comparison. *BioMedCentral Bioinformatics*, *14*:136. doi: 10.1186/1471-2105-14-136

Zielinski, R. (2006). Small-sample quantile estimators in a large nonparametric model, *Communications in Statistics: Theory and Methods*, *35*(7), 1223-1241. doi: 10.1080/03610920600692656

Mikhail Lomakin
Russian scientific and technical centre of information on standardization, metrology and conformity assessment,
31, Bldg 2, Nahimov Avenue,
Moscow, 117418
Russia
m.i.lomakin@ yandex.ru

Alexey Buryi
Russian scientific and technical centre of information on standardization, metrology and conformity assessment,
31, Bldg 2, Nahimov Avenue,
Moscow, 117418
Russia
a.s.burij@gostinfo.ru

Alexander Dokukin
Russian scientific and technical centre of information on standardization, metrology and conformity assessment,
31, Bldg 2, Nahimov Avenue,
Moscow, 117418
Russia
a.v.dokukin@gostinfo.ru

Anatoly Strekha
Russian scientific and technical centre of information on standardization, metrology and conformity assessment,
31, Bldg 2, Nahimov Avenue,
Moscow, 117418
Russia
a.a.streha@gostinfo.ru

Julia Niyazova
Moscow state University of geodesy and cartography
4, Gorokhovsky Lane,
Moscow, 105064
Russia
julia_lomakina@mail.ru

Alexander Balvanovich
Russian scientific and technical centre of information on standardization, metrology and conformity assessment,
31, Bldg 2, Nahimov Avenue,
Moscow, 117418
Russia
a.v.balvanovich@gostinfo.ru