Raport Badawczy

Research Report

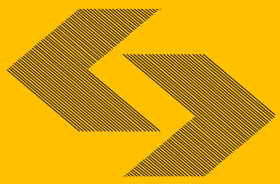
RB/67/2006

Tests for relation type - equivalence or tolerance - in a finite set of elementsns

L. Klukowski

Instytut Badań Systemowych
Polska Akademia Nauk

Systems Research Institute
Polish Academy of Sciences



Control and Cybernetics

vol. **35** (2006) No. 2

Tests for relation type - equivalence or tolerance - in a finite set of elements[1]

by

Leszek Klukowski

Ministry of Finance
12 Świętokrzyska Str., 00-916 Warsaw, Poland
e-mail: lkl@mofnet.gov.pl

Abstract: The statistical procedure for determination of the type of relation – equivalence or tolerance – in a finite set of elements, estimated on the basis of pairwise comparisons with random errors, is presented. The procedure consists of two tests based on Chebyshev's inequality for variance of a random variable; the test statistic is a mixture of some random variables. An example of application of the procedure – determination of relation type in the set of functions expressing profitability of treasury securities sold at auctions in Poland – is presented, too.

Keywords: tests for relation type, pairwise comparisons, nearest adjoining order method.

1. Introduction

The equivalence relation divides a set of elements into family of subsets with empty intersections, i.e. the relation is reflexive, symmetric and transitive. The tolerance relation also divides the set of elements into a family of subsets, but at least one nonempty intersection exists – the relation is not transitive. The methods of estimation of both relations, which rest on pairwise comparisons with random errors, are presented in Klukowski (1990, 2002). These methods are based on the idea of the nearest adjoining order (see Slater, 1961; David, 1988; and Klukowski, 1994, 2000). The methods of relation estimation presented in Klukowski (1990, 2002) are based on the assumption that the type of relation is known. In practice this may be often not true; therefore the method of determination of relation type is necessary in this case. A statistical procedure for this purpose is proposed in the paper (Section 3). The procedure is based on two statistical tests, which rest on Chebyshev's inequality for variance. The

[1]The investigations presented in the paper were partly sponsored by the Project MNiI no H02B 03828.

test statistic is a mixture of some random variables; two parameters of one component (random variable) of the mixture are determined: expected value and variance evaluation. The procedure may be effectively applied, if probability of error in each (pairwise) comparison is close to zero (it is assumed that comparison errors satisfy the assumptions formulated in Klukowski, 1990, 2002). The procedure is applied for examination of the "homogeneity" (similarity) of shapes of some functions (Section 4). Homogeneity of their shapes is verified with the use of three well-known statistical tests. The result of such examination can be used for forecasting purposes.

2. Basic definitions and notation

It is assumed that there exists (unknown) equivalence or tolerance relation in the finite set $\mathbf{X} = \{x_1, \ldots, x_m\}$ $(m \geqslant 3)$.

The equivalence relation (reflexive, symmetric, transitive) divides the set $\mathbf{X}$ into n_R $(n_R \geqslant 2)$ subsets χ_r^{*R} $(r = 1, \ldots, n_R)$ with empty intersections, i.e.:

$$\mathbf{X} = \bigcup_{r=1}^{n_R} \chi_r^{*R}, \quad \chi_r^{*R} \cap \chi_s^{*R} = \{\mathbf{0}\}, \quad \text{for} \quad r \neq s \tag{1}$$

where: $\{\mathbf{0}\}$ – empty set.

The tolerance relation is defined in similar way, but is not transitive, i.e. it satisfies the conditions:

$\mathbf{X} = \bigcup_{r=1}^{n_T} \chi_r^{*T}$ $(n_T \geqslant 2)$ and there exists at least one pair of subsets χ_r^{*T}, χ_s^{*T} $(r \neq s)$ with nonempty intersection: $\chi_r^{*T} \cap \chi_s^{*T} \neq \{\mathbf{0}\}$.

The equivalence relation can be characterized with the use of the function $T_1 : \mathbf{X} \times \mathbf{X} \to D$, $D = \{0, 1\}$, defined as follows:

$$T_1(x_i, x_j) = \begin{cases} 0 & \text{if there exists } q \text{ satisfying the condition} \\ & (x_i, x_j) \in \chi_q^{*R},\ i \neq j; \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

The tolerance relation can be characterized with use of the function $T_2 : \mathbf{X} \times \mathbf{X} \to D$, $D = \{0, 1\}$, defined as follows:

$$T_2(x_i, x_j) = \begin{cases} 0 & \text{if there exists } q \text{ and } s \ (q = s \text{ is not excluded}) \text{ such} \\ & \text{that } (x_i, x_j) \in \chi_q^{*T} \cap \chi_s^{*T},\ i \neq j; \\ 1 & \text{otherwise.} \end{cases} \tag{3}$$

It is assumed that the function $T_2(\cdot)$ characterizes completely the tolerance relation, i.e. there exists one-to-one relationship between the relation form and the set of values $T_2(x_i, x_j)$ for $(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$ (for example the relation in which

each subset χ_q^{*T} $(q = 1, \ldots, n_T)$ includes an element x_i, that is not included in any other subset ($x_i \in \chi_q^{*T}$ and $x_i \notin \chi_s^{*T}$ for $s \neq q$ satisfies this condition).

It is assumed in the paper that the type and form of the relation (equivalence or tolerance) in the set $\mathbf{X}$ (i.e. the function $T_1(\cdot)$ or $T_2(\cdot)$) are not known and they have to be estimated on the basis of pairwise comparisons $g(x_i,\ x_j)$, $(x_i, x_j \in \mathbf{X} \times \mathbf{X})$, with random errors. The result of comparison $g(x_i, x_j)$ is the function:

$$g : \mathbf{X} \times \mathbf{X} \to D, \qquad D = \{0, 1\}, \tag{4}$$

which estimates the "true" value $T_1(x_i, x_j)$ or $T_2(x_i, x_j)$. In the case of the equivalence relation, $g(x_i, x_j) = 0$ $(i \neq j)$ if comparison indicates that there exist q, which satisfy: $x_i, x_j \in \chi_q^{*R}$ and $g(x_i, x_j) = 1$ if comparison indicates an opposite result. In the case of the tolerance relation, $g(x_i, x_j) = 0$ if comparison indicates that there exist q, s (the case $q = s$ is not excluded) such, that $x_i, x_j \in \chi_q^{*T} \cap \chi_s^{*T}$ and $g(x_i, x_j) = 1$ if comparison indicates an opposite result. The comparisons $g(x_i,\ x_j)$ do not determine directly the type of the relation; they are only the basis for inference.

It is assumed (see Klukowski, 1990, 2002), that probability of each comparison correctness satisfies the conditions:

$$P(g(x_i x_j) = T_f(x_i, x_j)) \geqslant 1 - \delta, \qquad \delta \in \left(0, \frac{1}{2}\right) \tag{5}$$

where f equals 1 or 2 – according to the actual relation in the set $\mathbf{X}$.

The comparisons, which satisfy the conditions (5) can be obtained as the result of application of the (two samples) statistical tests. If the result of test application indicates that both samples (namely x_i and x_j) are realizations of the random variables with the same type of distribution (e.g. exponential or symmetric), then $g(x_i, x_j) = 0$; in the opposite case $g(x_i, x_j) = 1$. The probabilities of errors in the tests have to satisfy the conditions (5).

Let us notice that any comparison $g(x_i, x_j)$, which satisfies the conditions (5), may be equal to $T_f(x_i, x_j)$ (f=1 or 2) or not, as a result of random error. In particular, the comparisons obtained for the equivalence relation may be not transitive (e.g.: $g(x_i, x_j) = 0$, $g(x_j, x_k) = 0$ and $g(x_i, x_k) = 1$), while comparisons for the tolerance relation may be transitive. Therefore, the type of actual relation is not directly indicated by the results of comparisons.

Under the assumption that the type of relation is known, the estimated form of the equivalence relation can be obtained as the optimal solution of the discrete mathematical programming problem (see Klukowski, 1990):

$$\min_{\chi_1^R, \ldots, \chi_v^R} \left[\sum_{\langle i,j \rangle \in I(\chi_1^R, \ldots, \chi_v^R)} g(x_i, x_j) + \sum_{\langle i,j \rangle \in J(\chi_1^R, \ldots, \chi_v^R)} (1 - g(x_i, x_j)) \right], \tag{6}$$

where:

– $\chi_1^R, \ldots, \chi_v^R$ – an element of feasible set (any form of the equivalence relation in the set $\mathbf{X}$),
– $I(\chi_1^R, \ldots, \chi_v^R)$ – the set of all index pairs $\langle i, j\rangle$ satisfying the conditions:

$$i, j \in \{1, \ldots, m\}, \quad j > i;$$
$$\langle i, j\rangle \in I(\chi_1^R, \ldots, \chi_v^R) \Leftrightarrow \exists\, q \quad \text{such, that:} (x_i, x_j) \in \chi_q^R,$$

– $J(\chi_1^R, \ldots, \chi_v^R)$ – the set of all index pairs $\langle i, j\rangle$ satisfying the conditions:

$$i, j \in \{1, \ldots, m\}, j > i;$$
$$\langle i, j\rangle \in J(\chi_1^R, \ldots, \chi_v^R) \Leftrightarrow \text{there does not exist } q \text{ such, that:}$$
$$(x_i, x_j) \in \chi_q^R.$$

The optimal solution of the task with the criterion function (6) (estimated form of the equivalence relation) will be denoted with the symbols $\hat{\chi}_1^R, \ldots, \hat{\chi}_{\hat{n}_R}^R$. The solution can be characterized with the function:

$$\hat{t}_1(x_i, x_j) = \begin{cases} 0 & \text{if there exists in (estimated) relation such } q \text{ that} \\ & \quad (x_i, x_j) \in \hat{\chi}_q^R, \quad i \neq j; \\ 1 & \text{otherwise.} \end{cases} \tag{7}$$

It should be noticed that the estimated form of the relation may be not unique, because the number of optimal solutions of discrete problem can exceed one. The minimal value of the function (6) equals zero; it is assumed in the case $g(x_i, x_j) = \hat{t}_1(x_i, x_j)$ for each $(x_i, x_j) \in \mathbf{X} \times \mathbf{X}$.

In case of the tolerance relation the optimization problem assumes the form:

$$\min_{\chi_1^T, \ldots, \chi_v^T} \left[\sum_{\langle i,j\rangle \in I(\chi_1^T, \ldots, \chi_v^T)} g(x_i, x_j) + \sum_{\langle i,j\rangle \in J(\chi_1^T, \ldots, \chi_v^T)} (1 - g(x_i, x_j)) \right], \tag{8}$$

where:
– $\chi_1^T, \ldots, \chi_v^T$ – an element of feasible set (any form of the tolerance relation in the set $\mathbf{X}$),
– $I(\chi_1^T, \ldots, \chi_v^T)$ – the set of all index pairs $\langle i, j\rangle$ satisfying the conditions:

$$i, j \in \{1, \ldots, m\}, \quad j > i;$$
$$\langle i, j\rangle \in I(\chi_1^T, \ldots, \chi_v^T) \Leftrightarrow \exists\, q, s \quad \text{such, that: } (x_i, x_j) \in \chi_q^T \cap \chi_s^T,$$
$$\text{there exists at least one nonempty intersection, i.e. } \chi_q^T \cap \chi_s^T \ (q \neq s);$$

– $J(\chi_1^T, \ldots, \chi_v^T)$ – the set of all index pairs $\langle i, j\rangle$ satisfying the conditions:

$$i, j \in \{1, \ldots, m\}, j > i;$$
$$\langle i, j\rangle \in J(\chi_1^T, \ldots, \chi_v^T) \Leftrightarrow \text{it does not exist such } q \text{ that: } (x_i, x_j) \in \chi_q^T.$$

Optimal solution of the task corresponding to the tolerance relation will be denoted $\hat{\chi}_1^T, \ldots, \hat{\chi}_{\hat{n}_T}^T$. The solution can be characterized with the use of the function $\hat{t}_2(x_i, x_j)$ defined as follows:

$$\hat{t}_2(x_i, x_j) = \begin{cases} 0 & \text{if there exist } q \text{ and } s \ (q = s \text{ not excluded}) \text{ such,} \\ & \quad \text{that } (x_i, x_j) \in \chi_q^{T*} \cap \chi_s^{T*}, \quad i \neq j; \\ 1 & \text{otherwise.} \end{cases} \tag{9}$$

The properties of the task (8) are similar to properties of the task (6).

3. Procedure of relation type testing

As it was mentioned above, both types of relation can be estimated on the basis of the same pairwise comparisons $g(x_i, x_j)$. In the case of unknown relation type the question arises which one is true. The statistical procedure proposed below allows to decide in this case.

The procedure rests on the differences between the estimated form of equivalence and tolerance relation, obtained as solutions of the optimization tasks with the criterion functions (6) and (8) – respectively. The procedure consists of two statistical tests; the test statistics is a function of inconsistencies between comparisons $g(x_i, x_j)$ and functions $\hat{t}_1(x_i, x_j)$ or $\hat{t}_2(x_i, x_j)$ for the pairs (x_i, x_j), which satisfy the condition $\hat{t}_1(x_i, x_j) \neq \hat{t}_2(x_i, x_j)$.

The basis for the tests proposed are the random variables S_{ij}, defined as follows:

$$S_{ij} = \left|\hat{t}_1(x_i, x_j) - g(x_i, x_j)\right| - \left|\hat{t}_2(x_i, x_j) - g(x_i, x_j)\right|, \ \ \langle i, j\rangle \in I_w \tag{10}$$

where:

I_w – the set of all pairs of indices $\langle i, j\rangle$, which satisfy the conditions:

$$\hat{t}_1(x_i, x_j) \neq \hat{t}_2(x_i, x_j);$$

($\hat{t}_1(x_i, x_j)$ and $\hat{t}_2(x_i, x_j)$ defined – respectively – in (7) and (9)).

The conditions, which define the set I_w, mean that:

- in the estimated form of the tolerance relation the elements x_i and x_j are included in an intersection of two subsets $\hat{\chi}_q^T \cap \hat{\chi}_s^T$ ($q = s$ not excluded), while in the (estimated) equivalence relation they are included in different subsets

or

- in the estimated form of the tolerance relation the elements x_i and x_j are not included in any intersection of subsets (also in the same subset), while in (estimated) equivalence relation they are included in the same subset.

The test statistic is the sum of random variables S_{ij} ($\langle i,j\rangle \in I_w$) divided by the number of elements of the set I_w :

$$S = \frac{1}{||I_w||} \sum_{\langle i,j\rangle \in I_w} S_{ij}\,, \tag{11}$$

where: $||I_w||$ – number of elements of the set I_w.

The properties of the statistics S depend on the "true" form of the relation in the set $\mathbf{X}$ under consideration. Let us consider first the case of the tolerance relation; the expected value and the evaluation of variance of the variable S are determined below.

For simplification it is assumed that probability of error in each comparison $g(x_i, x_j)$ ($j \neq i$) is equal to δ (see (5)). In the case, when some probabilities are less than δ the properties of the procedure proposed are not worse.

In the case when tolerance relation exists in the set $\mathbf{X}$, the estimated form of the relation is equivalent to the actual (errorless result of estimation), i.e. $\hat{\chi}_1^T, \ldots, \hat{\chi}_{\hat{n}T}^T \equiv \chi_1^{*T}, \ldots, \chi_n^{*T}$, with probability equal to or greater than $1 - 2\delta$ (see Klukowski, 2002). In this case the equalities $\hat{t}_2(x_i, x_i) = T_2(x_i, x_i)$, $\langle i,j\rangle \in I_w$, are valid. Moreover, each expression $\left|\hat{t}_1(x_i, x_j) - g(x_i, x_j)\right|$ and $\left|\hat{t}_2(x_i, x_j) - g(x_i, x_j)\right|$, $\langle i,j\rangle \in I_w$, is zero-one random variable; their distributions can be determined on the basis of the properties of the random variable (comparison) $g(x_i, x_i)$.

The probability function of each random variable $\left|\hat{t}_2(x_i, x_j) - g(x_i, x_j)\right|$, $\langle i,j\rangle \in I_w$, is determined as follows (assuming equality in (5)):

$$\left.\begin{aligned} &P(\,|\hat{t}_2(x_i, x_j) - g(x_i, x_j)| = 0\,|\hat{t}_2(\cdot) = T_2(\cdot)) \\ &\qquad = P(g(x_i, x_j) = \hat{t}_2(x_i, x_j)\,|\,\hat{t}_2(\cdot) = T_2(\cdot)) = 1 - \delta, \\ &P(\,|\,\hat{t}_2(x_i, x_j) - g(x_i, x_j)| \,= 1|\hat{t}_2(\cdot) = T_2(\cdot)) \\ &\qquad = P(g(x_i, x_j) \neq \hat{t}_2(x_i, x_j)|\hat{t}_2(\cdot) = T_2(\cdot)) = \delta. \end{aligned}\right\} \tag{12}$$

Under the assumption $\hat{t}_1(x_i, x_j) \neq \hat{t}_2(x_i, x_j)$ (see (10)), the probability function of the random variable $|\hat{t}_1(x_i, x_j) - g(x_i, x_j)|$ assumes the form:

$$\left.\begin{aligned} &P(|\hat{t}_1(x_i, x_j) - g(x_i, x_j)| = 0|\hat{t}_2(\cdot) = T_2(\cdot)) \\ &\qquad = P(g(x_i, x_j) \neq \hat{t}_2(x_i, x_j)|\hat{t}_2(\cdot) = T_2(\cdot)) = \delta, \\ &P(|\hat{t}_1(x_i, x_j) - g(x_i, x_j)| = 1|\hat{t}_2(\cdot) = T_2(\cdot)) \\ &\qquad = P(g(x_i, x_j) = \hat{t}_2(x_i, x_j)|\hat{t}_2(\cdot) = T_2(\cdot)) = 1 - \delta. \end{aligned}\right\} \tag{13}$$

The probabilities (13) result from the fact that for $\langle i,j\rangle \in I_w$ inequalities $\hat{t}_1(\cdot) \neq \hat{t}_2(\cdot)$ and implications: $g(\cdot) = \hat{t}_1(\cdot) \Rightarrow g(\cdot) \neq \hat{t}_2(\cdot)$ and $g(\cdot) \neq \hat{t}_1(\cdot) \Rightarrow g(\cdot) = \hat{t}_2(\cdot)$ hold.

The equalities (12) and (13) indicate:

$$\begin{aligned} &P(S_{ij} = -1|\hat{t}_2(\cdot) = T_2(\cdot)) = \\ &P[(|\hat{t}_1(x_i, x_j) - g(x_i, x_j)| = 0) \\ &\qquad \cap(|\hat{t}_2(x_i, x_j) - g(x_i, x_j)| = 1)|\hat{t}_2(\cdot) = T_2(\cdot)] = \delta \end{aligned} \tag{14}$$

$$\begin{aligned} &P(S_{ij} = 1|\hat{t}_2(\cdot) = T_2(\cdot)) = \\ &P[(|\hat{t}_1(x_i, x_j) - g(x_i, x_j)| = 1) \\ &\qquad \cap(|\hat{t}_2(x_i, x_j) - g(x_i, x_j)| = 0)|\hat{t}_2(\cdot) = T_2(\cdot)] = 1 - \delta \end{aligned} \tag{15}$$

It follows from (14) and (15), that in the case of tolerance relation the expected value $E_2(S_{ij})$ and variance $Var_2(S_{ij})$ of each random variable S_{ij}, $\langle i, j\rangle \in I_w$, assume the form – respectively:

$$E_2(S_{ij}) = -\delta + 1 - \delta = 1 - 2\delta \tag{16}$$

$$Var_2(S_{ij}) = (-1 - (1 - 2\delta))^2\delta + (1 - (1 - 2\delta))^2(1 - \delta) = 4\delta(1 - \delta). \tag{17}$$

The random variable S is the sum of $||I_w||$ random variables S_{ij}; the expected value of each variable S_{ij} in the sum (11) is equal to $1 - 2\delta$ divided by $||I_w||$. Therefore, the expected value of the variable S equals:

$$E_2(S) = 1 - 2\delta. \tag{18}$$

The variance $Var(S)$ of the random variable S (see (11)) is evaluated under the assumption that any random variables S_{ij} andS_{kl}, which satisfy the conditions $i \neq k, l$ and $j \neq k, l$, are independent (i.e. their covariance equals to zero), while the remaining variables may be dependent. The number of covariances equal to zero is denoted $L(I_w)$; if the assumption does not hold, then $L(I_w) = 0$. The evaluation of variance of the variable S is based on the following facts: each variance of S_{ij} is equal to $4\delta(1 - \delta)$ and each non-zero covariance $C(S_{ij}, S_{kl})$ is not greater than $4\delta(1 - \delta)$. Moreover, the number of variances $Var(S_{ij})$ ($\langle i, j\rangle \in I_w$) is equal to $||I_w||$ and the number of covariances (in the set I_w) is equal to $2 * (||I_w|| * (||I_w|| - 1)/2 - L(I_w))$. As a result, $Var(S)$ satisfies the condition:

$$Var(S) \leq (1/||I_w||^2)(||I_w||^2 - 2L(I_w))4\delta(1 - \delta),$$

equivalent to:

$$Var(S) \leq 4(1 - 2L(I_w)/||I_w||^2)\delta(1 - \delta). \tag{19}$$

The right-hand side of the inequality (19) can significantly exceed the actual variance $Var(S)$, because covariances $C(S_{ij}, S_{kl})$ may be less than $Var(S_{ij})$, in particular – negative. More precise evaluation of the variance requires some additional knowledge about covariances $C(S_{ij}, S_{kl})$. Sometimes their values can

be evaluated, e.g. when the comparisons $g(x_i, x_j)$ are obtained from statistical test and covariance of test statistics is known. In the case, when test statistics is a function of difference of some random variables, namely X, Y, Z, with expected values μ_X, μ_Y, μ_Z respectively, the variance of the variable S can be significantly less than (19). It is so, because:

$$C[(X-Y),(Y-Z)] = E[(X-Y)-(\mu_x-\mu_y)][(Y-Z)-(\mu_y-\mu_z)] = \\ C(X,Y)+C(Y,Z)-\ Var(Y)-C(X,Z). \tag{20}$$

If values of covariances in (20) are similar, then covariance $C[(X-Y),(Y-Z)]$ is close to zero (or less). In such case the evaluation (19) can be replaced with the less restrictive formula:

$$Var(X+Y) \approx\ Var(X)+\ Var(Y)+\max\{\ Var(X)+\ Var(Y)\}, \tag{21}$$

which indicates:

$$Var(S) \leqslant 4(1/2+1/(2||I_w||)-L(I_w)/||I_w||^2)\delta(1-\delta). \tag{22}$$

The properties (18) and (19) of the random variable S are valid in the case of errorless estimation result of the tolerance relation $(\hat{\chi}_1^T,\ldots,\hat{\chi}_{\hat{n}_T}^T \equiv \chi_1^{*T},\ldots,\chi_n^{*T})$. If it is not true, then the properties mentioned do not hold. Moreover, it seems impossible to determine the probability of any non-errorless estimation result in an analytic way (the number of such results is quite large). Therefore, the realizations of the variable S obtained for any estimation result (errorless or not) can be treated as realizations of some mixture of distributions. However, the properties (expected value, evaluation of variance and probability of occurrence) of only one random variable from the mixture - corresponding to errorless estimation result - can be determined without difficulties. If the probability of comparison errors δ is close to zero, then the probability of this variable occurrence (equal to $1-2\delta$) is close to one. In other words, the realizations of the mixture is dominated by this component.

In the case, when the equivalence relation exists in the set $\mathbf{X}$ and the result of estimation is errorless (the probability of the event is equal or greater than $1-2\delta$, see Klukowski, 1990) the distribution of the random variable S (defined in (11)) can be obtained in a similar way. The distribution of each random variable S_{ij} ($\langle i,j\rangle \in I_w$) is the function of comparison results $g(x_i,x_j)$ (because $\hat{t}_1(\cdot) = T_1(\cdot)$ and $\hat{t}_1(\cdot) \neq \hat{t}_2(\cdot)$). Therefore, the distributions of the random variables $|\hat{t}_1(x_i,x_j)-g(x_i,x_j)|$ and $|\hat{t}_2(x_i,x_j)-g(x_i,x_j)|$ ($\langle i,j\rangle \in I_w$) are as follows (assuming equality in (5)):

$$P(|\hat{t}_1(x_i,x_j)-g(x_i,x_j)|=0|\hat{t}_1(\cdot)=T_1(\cdot)) \\ = P(g(x_i,x_j)=\hat{t}_1(x_i,x_j)|\hat{t}_1(\cdot)=T_1(\cdot))=1-\delta, \\ P(|\hat{t}_1(x_i,x_j)-g(x_i,x_j)|=1|\hat{t}_1(\cdot)=T_1(\cdot)) \\ = P(g(x_i,x_j)\neq\hat{t}_1(x_i,x_j)|\hat{t}_1(\cdot)=T_1(\cdot))=\delta, \tag{23}$$

and:

$$\begin{aligned}
&P(|\hat{t}_2(x_i,x_j) - g(x_i,x_j)| = 0|\hat{t}_1(\cdot) = T_1(\cdot)) \\
&= P(g(x_i,x_j) = \hat{t}_2(x_i,x_j)|\hat{t}_1(\cdot) = T_1(\cdot)) = \delta, \\
&P(|\hat{t}_2(x_i,x_j) - g(x_i,x_j)| = 1|\hat{t}_1(\cdot) = T_1(\cdot)) \\
&= P(g(x_i,x_j) \neq \hat{t}_2(x_i,x_j)|\hat{t}_1(\cdot) = T_1(\cdot)) = 1 - \delta. \qquad (24)
\end{aligned}$$

From (23) and (24) it follows that:

$$\begin{aligned}
&P(S_{ij} = -1|\hat{t}_1(x,x_j) = T_1(x_i,x_j)) = \\
&P[(|\hat{t}_1(x_i,x_j) - g(x_i,x_j)| = 0) \cap (|\hat{t}_2(x_i,x_j) - g(x_i,x_j)| = 1)| \\
&\qquad\qquad \hat{t}_1(x_i,x_j) = T_1(x_i,x_j)] = 1 - \delta \qquad (25) \\
&P(S_{ij} = 1|\hat{t}_1(x_i,x_j) = T_1(x_i,x_j)) = \\
&P[(|\hat{t}_1(x_i,x_j) - g(x_i,x_j)| = 1) \cap (|\hat{t}_2(x_i,x_j) - g(x_i,x_j)| = 0)| \\
&\qquad\qquad \hat{t}_1(x_i,x_j) = T_1(x_i,x_j)] = \delta\,. \qquad (26)
\end{aligned}$$

The formulas (25) and (26) are the basis for determination of the expected value and variance of each random variable S_{ij}:

$$E_1(S_{ij}) = -1 + \delta + \delta = 2\delta - 1, \qquad (27)$$

$$Var_1(S_{ij}) = 4\delta(1-\delta). \qquad (28)$$

The form of the above parameters allows to determine the expected value and evaluation of variance of the random variable S, when the equivalence relation exists in the set $\mathbf{X}$. The expected value can be expressed in the form:

$$E_1(S) = 2\delta - 1, \qquad (29)$$

while the variance satisfies the condition:

$$Var(S) \leq 4(1 - 2L(I_w)/||I_w||^2)\delta(1-\delta), \qquad (30)$$

the same as condition (19). The evaluations (20) and (22) can be also valid in case of the equivalence relation.

The properties (29) and (30) are valid for the equivalence relation, when errorless estimation result occurs. However, with some probability - less than 2δ - the result of estimation is different than the errorless one. Therefore, the distribution of the random variable S is a mixture of distributions, with similar properties, as in the case of tolerance relation.

On the basis of expected value and evaluation of variance of the random variable S, for both relation types, it is possible to determine some tests for distinction them. The Chebyshev's inequality can be used as the basis of tests:

$$P(|X - E(X)| > k\sigma_X) < 1/k^2,$$

where:

X – a random variable with expected value $E(X)$ and variance σ_X,

k – a positive constant.

The test for verification the tolerance relation in the set $\mathbf{X}$ rests on expected value (equal to $1-2\delta$) and evaluation (19) of variance of the random variable S. The null and the alternative hypotheses of the test can be formulated in the following way:

$$H_{T0} : E(S) = 1 - 2\delta,$$
$$H_{T1} : E(S) = 2\delta - 1,$$

with the critical region:

$$\Lambda_T = \{S|S < 1 - 2\delta - k\sigma_S\}, \tag{31}$$

where: σ_S – square root of the variance $Var(S)$ evaluation, i.e.:

$$\sigma_S = [4(1 - 2L(I_w)/||I_w||^2\delta(1-\delta)]^{1/2}.$$

The form of the test for the equivalence relation is "symmetric" :

$$H_{R0} : E(S) = 2\delta - 1,$$
$$H_{R1} : E(S) = 1 - 2\delta,$$

with the critical region:

$$\Lambda_R = \{S|S > 2\delta - 1 + k\sigma_S\} \tag{32}$$

(σ_S – the same, as in the formula (31)).

The tests may be used together or separately (one of them only). In the first case, their critical regions have to be non-overlapping; the value of parameter k may be different in each test (leading to different probabilities of errors in the tests). The evaluations of the probabilities of errors are determined below. In the case, when $2\delta - 1 + k\sigma_S < 1 - 2\delta - k\sigma_S$, there exists some non-decision region of the procedure.

Application of one test only allows to reject the hypothesis H_{T0} or H_{R0} (significance test); the alternative hypothesis can assume the form $H_{T1} : E(S) < 1 - 2\delta$ or $H_{R1} : E(S) > 2\delta - 1$.

Let us notice that if the exact values or evaluations of covariances $C(S_{ij}, S_{kl})$ are known, then the critical regions can be determined more precisely, which would improve the properties of the tests.

The critical regions of both tests are based on Chebyshev inequality. Such evaluations of probabilities of test errors are not precise - typically overestimated (e.g. if distribution of test statistics is symmetric, then the expression $1/k^2$ can be replaced with the expression $1/2k^2$). However, it is not easy to examine the asymmetry (or other "useful" features) of the distribution of S statistics.

The properties (18), (19), (29), (30) of the statistics S are valid in the case of errorless estimation result of the relation form, when its type is known. The probability of the errorless estimation result (optimal solution of the task (6) or (8)) is equal or greater than $1-2\delta$ and therefore, the evaluation of the first and the second type error in the tests (31) and (32) has to be corrected with the use of this factor. Denoting the significance level of the tests with the symbol α (its value results from the term $1/k^2$ in Tshebyshev's inequality) the corrected significance level can be expressed in the form:

$$1-(1-\alpha)(1-2\delta)=\alpha+2\delta(1-\alpha). \tag{33}$$

The formula (33) results from the fact that the probability of correct decision (the event opposite to the first type error) in the test is equal to $1-\alpha$, but it is valid in the case of errorless estimation result (probability of this event is equal to $1-2\delta$). Therefore, the probability of correct result of the test, multiplied by the factor $1-2\delta$, equals $(1-\alpha)(1-2\delta)$ and finally - corrected first type error equals (33). The corrected significance level is higher than α; the component $2\delta(1-\alpha)$ determines the increase of the probability resulting from the fact that test statistics is the mixture of distributions and only one component of the mixture, with known parameters, is taken into account. If δ is close to $1/2$, then the corrected probability of the error is close to one.

The evaluation of the probability of the second type error is obtained under the assumption that the value of parameter k is the same in both tests. The probability can be evaluated for both tests in the following way.

In the case of the tolerance relation, the second type error occurs, when H_{T0} is tested and accepted (i.e. $S \geqslant 1-2\delta-k\sigma_S$), while the equivalence relation is true (i.e. $E(S|H_{R0})=2\delta-1$). The probability of such event can be evaluated in the following way:

$$\begin{aligned}
&P(S \geqslant 1-2\delta-k\sigma_S|H_{R0}) = \\
&P(S-(2\delta-1) \geqslant 1-2\delta-(2\delta-1)-k\sigma_S|H_{R0}) = \\
&P(S-(2\delta-1) \geqslant 2(1-2\delta)-k\sigma_S|H_{R0}) = \\
&P(S-(2\delta-1) \geqslant k_{R0}\sigma_S|H_{R0}) \geqslant P(|S-(2\delta-1)| \geqslant |k_{R0}\sigma_S|) \leqslant 1/k_{R0}^2,
\end{aligned} \tag{34}$$

where the value of k_{R0} is determined in the following way (the expression $k_{R0}\sigma_S$ is positive under assumptions made):

$$2(1-2\delta)-k\sigma_S = k_{R0}\sigma_S \Rightarrow k_{R0}=(2(1-2\delta)-k\sigma_S)/\sigma_S. \tag{35}$$

The probability of the second type error in the case of the equivalence relation

$(E(S|H_{T0}) = 1 - 2\delta)$ is obtained in a similar way:

$$\begin{aligned}
&P(S \leq 2\delta - 1 + k\sigma_S|H_{T0}) = \\
&P(S - (1-2\delta) \leq 2\delta - 1 - (1-2\delta) + k\sigma_S|H_{T0}) = \\
&P(S - (1-2\delta) \leq 2(2\delta - 1) + k\sigma_S|H_{T0}) = \\
&P(S - (1-2\delta) \leq k_{T0}\sigma_S|H_{T0}) \leq P(|S - (1-2\delta)| \leq |k_{R0}\sigma_S| \leq 1/k_{T0}^2,
\end{aligned} \tag{36}$$

where:

$$k_{T0} = (2(2\delta - 1) + k\sigma_S)/\sigma_S = (k\sigma_S - 2(1-2\delta))/\sigma_S. \tag{37}$$

Let us notice that the values k_{T0}^2 and k_{R0}^2 are equal for the same value of the parameter k in both tests; therefore the evaluations of the second type error probabilities are also the same.

Evaluations (34) and (36) correspond to the case of errorless estimation result, while the realizations of the random variable S are obtained from the mixture of distributions. Therefore, these evaluations have to be corrected – similarly as in (33). Denoting the probability of the second type error resulting from inequalities (34) and (36) with the symbol β, the corrected probability of this error occurrence can be expressed in the form:

$$1 - (1-\beta)(1-2\delta) = \beta + 2\delta(1-\beta). \tag{38}$$

Let us notice that if the probability $\beta \to 0$, then the probability $\beta + 2\delta(1 - \beta) \to 2\delta$; which means that the tests are not consistent.

As it was mentioned above, the determination of properties of the proposed procedure (except for evaluations of the probabilities of errors in the tests) is not easy; simulation approach can be applied for this purpose.

The tests are based on "weak" probabilistic inequality. Therefore the results of their application can be also of rough type; it is a *cost* of non-restricted assumptions about comparison errors. However, such approach provides some progress in comparison with an arbitrary decision.

4. Example of application of the procedure

The procedure presented above is applied to the problem of determination of relation type in the set comprising seven elements - some functions with values from the range $(0, 1]$. They are approximations of empirical functions, expressing profitability of treasury securities sold at auctions in Poland. The application of the procedure is aimed at selecting functions with similar shapes. The comparison of shapes was made for each pair with the use of three statistical tests (correlation, regression and goodness-of-fit); the resultant comparison (from three tests) was determined using the majority rule. The results of comparisons are presented in Table 1, shapes of functions - in Chart 1. The probability δ (upper limit of probability of error in pairwise comparisons) equals 0.01.

The optimal solution of the optimization task for the equivalence relation indicates the following form of estimated relation $\hat{\chi}_1^R = \{x_1, x_3, x_6\}$, $\hat{\chi}_2^R = \{x_2, x_5, x_7\}$, $\hat{\chi}_3^R = \{x_4\}$; the value of the criterion function (6) equals three. Optimal solution corresponding to the tolerance relation has multiple variants. Therefore, the variant with the biggest fraction of elements included in the intersections of different subsets is assumed as the basis for testing relation type. The optimal solution of the task for equivalence relation indicates the following form of relation: $\hat{\chi}_1^T = \{x_1, x_3, x_6, x_7\}$, $\hat{\chi}_2^T = \{x_2, x_3, x_5, x_7\}$, $\hat{\chi}_3^T = \{x_4\}$; the value of the criterion function (8) equals two.

The set I_w, comprising pairs of elements defined in (10), assumes the form: $I_w = \{\langle 1,7\rangle, \langle 3,7\rangle, \langle 6,7\rangle, \langle 2,3\rangle, \langle 3,5\rangle\}$. The number of elements of this set is equal to five; the number of pairs with different indices equals four ($\langle 1,7\rangle$ and $\langle 2,3\rangle$, $\langle 1,7\rangle$ and $\langle 3,5\rangle$, $\langle 6,7\rangle$ and $\langle 2,3\rangle$, $\langle 6,7\rangle$ and $\langle 3,5\rangle$). The test statistic assumes the form:

$$\begin{aligned} S =& [(|t_1(x_{1,7}) - g(x_{1,7})| - |t_2(x_{1,7}) - g(x_{1,7})|) + \\ &+ (|t_1(x_{3,7}) - g(x_{3,7})| - |t_2(x_{3,7}) - g(x_{3,7})|) + \\ &+ (|t_1(x_{6,7}) - g(x_{6,7})| - |t_2(x_{6,7}) - g(x_{6,7})|) + \\ &+ (|t_1(x_{2,3}) - g(x_{2,3})| - |t_2(_{2,3})| - g(x_{2,3})|) + \\ &+ (|t_1(x_{3,5}) - g(x_{3,5})| - |t_2(_{3,5})| - g(x_{3,5})|)]/5 = \\ =& [|1-0| - |0-0| + |1-0| - |0-0| + |1-1| - |0-1| + \\ &+ |1-0| - |0-0| + |1-1| - |0-1|]/5 = 1/5. \end{aligned}$$

The critical region of the tests is based on the variance $Var(S)$ evaluation (see (19)); the evaluation assumes the form:

$$Var(S) \leqslant 4 * (1 - 2 * (4/25)) * 0.01 * (1 - 0.01) = 0.027;$$

the square root of its value (σ_S in (31)) is equal 0.164.

The hypothesis for the tolerance relation is verified first, because the test statistic is positive. The critical region for the null hypothesis (see (31)), using k=5, is of the form:

$$\Lambda_T = \{S | S < 1 - 0.02 - 5 * 0.164 = 0.160\}.$$

The value of the test statistics S (equal 0.2) is greater than the critical value of 0.160 and therefore it is not included in the critical region; the null hypothesis must be accepted. The significance level for k=5 is not greater than $(1/5)^2 = 0.04$ and corrected significance level (see (33)) is equal or less than 0.04+2*0.01(1-0.04)=0.059.

The critical region for the equivalence relation test (see (32)) assumes the form:

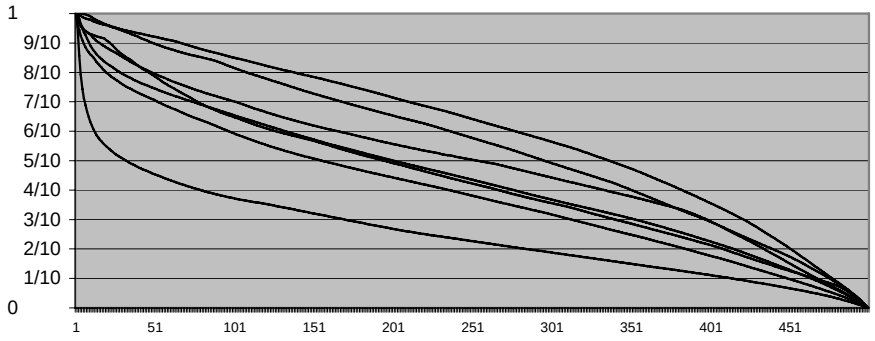
$$\Lambda_R = \{S|S > 0.02 - 1 + 5 * 0.164 = -0.160\}.$$

Therefore, the null hypothesis must be rejected; results of both tests are not contradictory.

Table 1. Results of comparisons $g(x_i, x_j)$

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	×	1	0	1	1	0	0
x_2		×	0	1	0	1	0
x_3			×	1	1	0	0
x_4				×	1	1	1
x_5					×	1	0
x_6						×	1
x_7							×

Chart 1. The set $\mathbf{X}$ – graphs of the functions

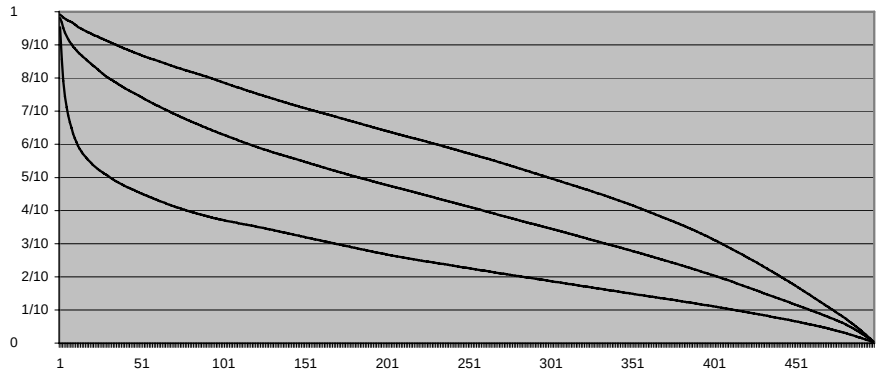


The evaluation of probability of the second type error for the tolerance relation is determined in (36) – (38); it amounts to $1/k^2_{T0} = 1/k^2_{R0} = 0.021$ and the corrected probability level is equal or less than 0.040.

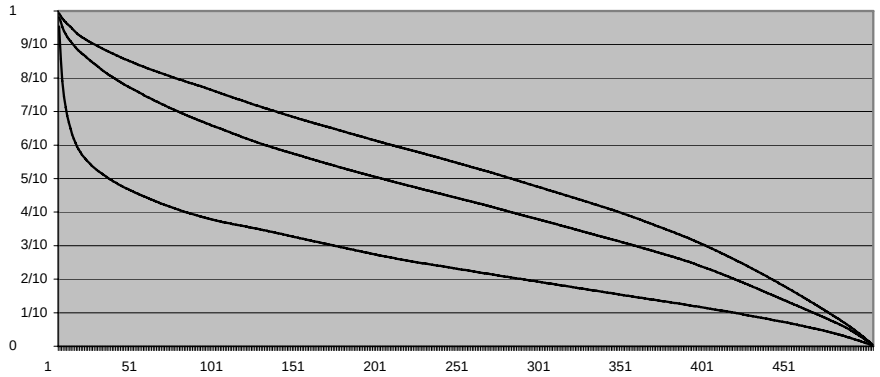
The results of the test application are depicted in Chart 2: (a) and (b). Chart 2(a) presents the functions averaged on the basis of (estimated) equivalence relation (i.e. the average of the functions from each subset $\hat{\chi}^R_1 = \{x_1, x_3, x_6\}$, $\hat{\chi}^R_2 = \{x_2, x_5, x_7\}$, $\hat{\chi}^R_3 = \{x_4\}$. Chart 2(b) presents the results averaged in the same way - corresponding to the tolerance relation. It can be noticed that the shapes of functions averaged on the basis of subsets $\hat{\chi}^R_1$ and $\hat{\chi}^R_2$ (equivalence

Chart 2. Results of estimation of the equivalence relation and tolerance relation

a) the functions averaged according to the equivalence relation



b) the functions averaged according to the tolerance relation



relation) are more dissimilar, than those averaged on the basis $\hat{\chi}_1^T$ and $\hat{\chi}_2^T$ (tolerance relation). The toleration relation generates more "fuzzy" result, because the functions denoted with symbols x_3 and x_7 are included in both sets $\hat{\chi}_1^T$ and $\hat{\chi}_2^T$. The results of the procedure application (both tests) indicate acceptance of the tolerance relation; it suggests that the set $\mathbf{X}$ comprises some functions (elements: x_3 and x_7) with non-homogenous features.

The parameters of the procedure, especially k in Tshebyshev's inequality, indicate existence of no-decision region – the interval [-0.16; 0.16]. The interval may be narrowed down; such modification changes the probability of the errors - increases the probability of the second type error and decreases the probability of the first type error.

5. Summary

The procedure presented in the paper is the tool for determination of the relation type (equivalence or toleration) in a finite set of elements. It is based on the assumption that both relations are estimated with the use of the idea of the nearest adjoining order; the basis for estimation are the pairwise comparisons with random errors. Procedure consists of two tests resting on Chebyshev's inequality; the variance of random variable necessary in the inequality is replaced with its evaluation. The test statistic is the mixture of distributions; the expected value and evaluation of variance are determined for one component of the mixture. Therefore the results of the procedure are of rough type; in consequence it is effective, when the probabilities of comparison errors are close to zero. It seems rational to examine the properties of the procedure with the use of simulation.

References

David, H.A. (1988) *The Method of Paired Comparisons*, 2^{nd} ed. Ch. Griffin, London.

Klukowski, L. (1990) Algorithm for classification of samples in the case of unknown number of random variables generating them. *Przegląd Statystyczny* **XXXVII** (3) (in Polish), 167-177.

Klukowski, L. (1994) Some probabilistic properties of the nearest adjoining order method and its extensions. *Annals of Operations Research* **51**, 241-261.

Klukowski, L. (2000) The nearest adjoining order method for pairwise comparisons in the form of difference of ranks. *Annals of Operations Research* **97**, 357-378.

Klukowski, L. (2002) Estimation of tolerance relation on the basis of pairwise comparisons with random errors. In: Z. Bubnicki, O. Hryniewicz, R. Kulikowski, eds., *Methods and techniques of data analysis and decision support*. Akademicka Oficyna Wydawnicza EXIT, Warszawa 2002 (in Polish), V-21–V-35.

Slater, P. (1961) Inconsistencies in a schedule of paired comparisons. *Biometrika* **48**, 303-312.