UTILIZATION OF CERTAIN METHOD FROM THE FIELD OF DATA MINING

Zdeněk Půlpán
University of Hradec Králové, Pedagogical Faculty, Department of Mathematics

Abstract: *The paper describes the methodology of the so-called data mining on the example of a "healthy-ill" ensemble.*

Key words: data mining – answer tree methodology – statistical modelling and decision – data management

1 Introduction

Though the methodology of data mining is based on mathematical statistics, logic, and artificial intelligence, it also utilizes expert knowledge in the work with databases. The fundamental principle of the methodology is its systematic nature, both in the preparation and application of the procedure, and in its interpretation. The result is gradual obtaining of a foundation for making decisions on the basis of synthesized pieces of information from certain (usually large) collections of data. Another important characteristic of the methodology is heterogeneity and flexibility of available means. The researcher is not bound by a single "well-tried" methodology (e.g., a statistical one), but he or she can choose from a number of variants for a further, more detailed analysis. Software means are quite varied nowadays and they offer a number of levels of analysis of data ensembles consisting of various types of variables (both nominal and metric). Analyses are connected in a different way with statistical softwares (SPSS, NCSS, STATISTICA, etc.) and contain various kinds of procedures (Clementine, Data Miner, Neural Works, Answer Tree, Regression Tree, etc.). The present author will demonstrate one of the analyses elaborated with the use of the above-mentioned methodology. The results can be compared with its "classic" form published in [Půlpán 2003] and applied to the identical data.

Paper [Půlpán 2002, 2003, 2004] discussed the methodology of diagnosis determination on the basis of the construction of a multidimensional mathematical-statistical model containing four basic variables: LTH (lathosterol), SIT (sitosterol), CAM (camposterol), and TCH (total cholesterol). The diagnosis was formulated in the alternatives healthy-ill in connection with cholesterol metabolism. The decision-making was based on a basic sample of 101 subjects ("healthy" as regards with cholesterol metabolism) and samples of altogether 189 patients with various impairments of cholesterol metabolism. It has been shown that the data under study make it possible to establish diagnosis with the use of statistical methods with a degree of uncertainty not exceeding 30% of wrong diagnoses. In the present paper, an attempt will be made to establish the same diagnosis, but with the use of different means.

To obtain a set of measured values of the above-mentioned variables in healthy subjects is relatively expensive. The present author thus thinks that it is appropriate to present their more detailed processing, the results and possibilities of which can inspire further research.

1 Flexible algorithm

Let us begin with the premise that a disorder in cholesterol metabolism cannot be diagnosed from the measurement of only one of the variables under study, LTH, SIT, CAM, and TCH. At the same time, let us be aware of the fact that the weight of these variables for the above-mentioned diagnosis differs. For the time being, nevertheless, we will not estimate it and we will assume equivalence of the variables under study.

First, we will investigate the tetrads of the values of the variables LTH, SIT, and TCH in an ensemble of healthy subjects in order to determine the standard of nonpathological cholesterol metabolism. At the same time we will make an attempt to define "the prototype" of healthy subjects in order to be able to partially reduce the data ensemble of the healthy subjects (without losing essential information), if need be. As all variables under study are metrical and of continuous type, we will

attempt a certain reduction of information contained in it by means of monotonous transformation to discrete variables into five levels 0, 1, 2, 3, and 4 in such a way that the points of division of the original continuous scale into the discrete one will be the values

$x_1 = \bar{x}$ -0.4 . s; $x_2 = \bar{x}$ - 0.25 . s;

$x_3 = \bar{x}$ + 0.25 . *s*; $x_4 = \bar{x}$ + 0.84 . *s*,

Table 1. - *Basic statistical parameters of the variables under study.*

Variable *X*	Mean $\bar{x}$	Standard deviation *s*
LTH	7.769	4.896
SIT	5.044	2.553
CAM	10.244	4.249
TCH	4,921	1.094

where the symbol *x* denotes the value of a random variable under study, $\bar{x}$, or *s*, its selective mean, or the standard deviation in the ensemble of healthy subjects (see Table 1). This transformation is employed for all variables under study (excepting the variable *LTH*) as at least approximately normal distribution is assumed in them. The variable *LTH* is, nevertheless, relatively well approximatable by log-normal distribution. The degree of agreement of the appropriate theoretical distribution with the experimental values can be assessed from Graphs 1, 2, 3, 4 (*p*-value of the pertinent χ^2 – test of good agreement is mostly greater than 0.01). For the sake of comparison, also Graphs 5 and 6 are presented, which show the degree of agreement of the experimental distribution of the variable *LTH* with the corresponding normal one and the variable *CAM* with the corresponding distribution of χ^2. Test χ^2 of good agreement in the first case is of a small *p*-value, so the above-mentioned approximation is out of the question, in the second case of the variable *CAM*, with regard to *p*-value, the approximation by division of χ^2 would be more suitable ($p = 0.13$). In the case of the variable *CAM*, we preferred normal distribution. It resulted from the belief that approximation by normal distribution is more acceptable for a directly measurable variable. Transformation divides the values of each variable under study with an approximately identical number (by 20 % of all values) into the individual intervals with the limit points according to Table 2.

As the values of some variables in the "healthy" subjects significantly differed from their mean, we considered it useful from the viewpoint of the establishment of the norm to exclude several respondents who in at least one of the variable showed values markedly different from the mean (e.g., by more than ± 2) from the ensemble of the "healthy" subjects.

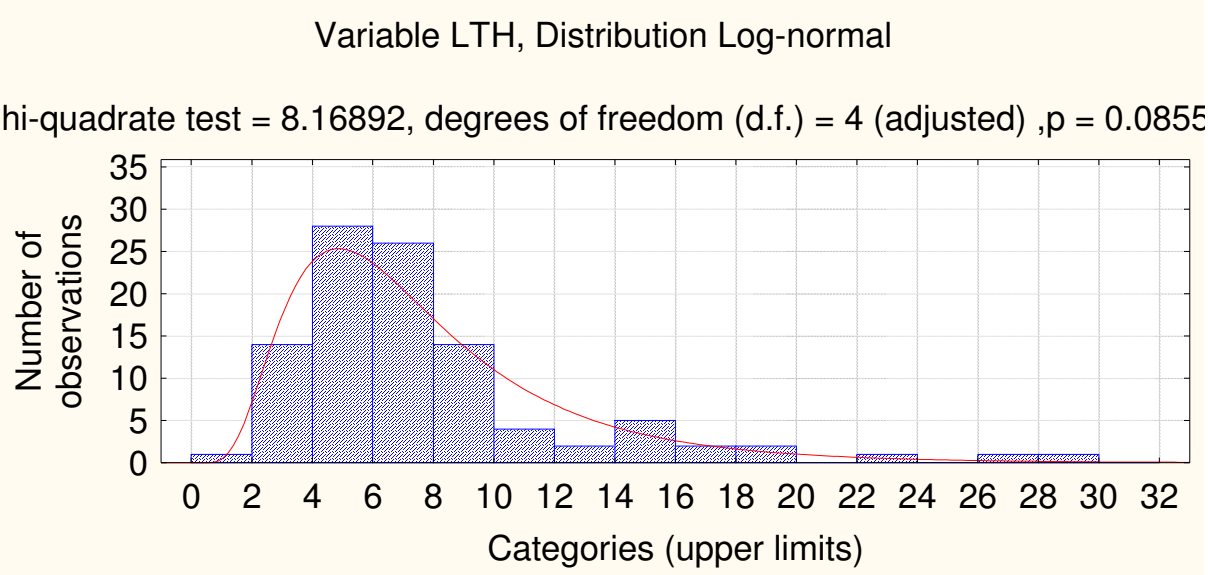
Nevertheless, it was medical evaluation that decided. (It considered possible variability of the values of the variables under study in healthy subjects.) The subjects in the ensemble of the "healthy" ones who in some variable did not show the measured values between the minimal and maximal ones as evaluated by expert determination were excluded from the representative ensemble of the "healthy" subjects.

Table 2 shows the "acceptable" maximal and minimal values which are considered possible for the representation of the "healthy" subjects from the medical viewpoint. Therefore subjects 82, 83, 85, 80, 22, and 62 were excluded from the ensemble of the "healthy" ones. (They are marked with an asterisk in Table 3a.)

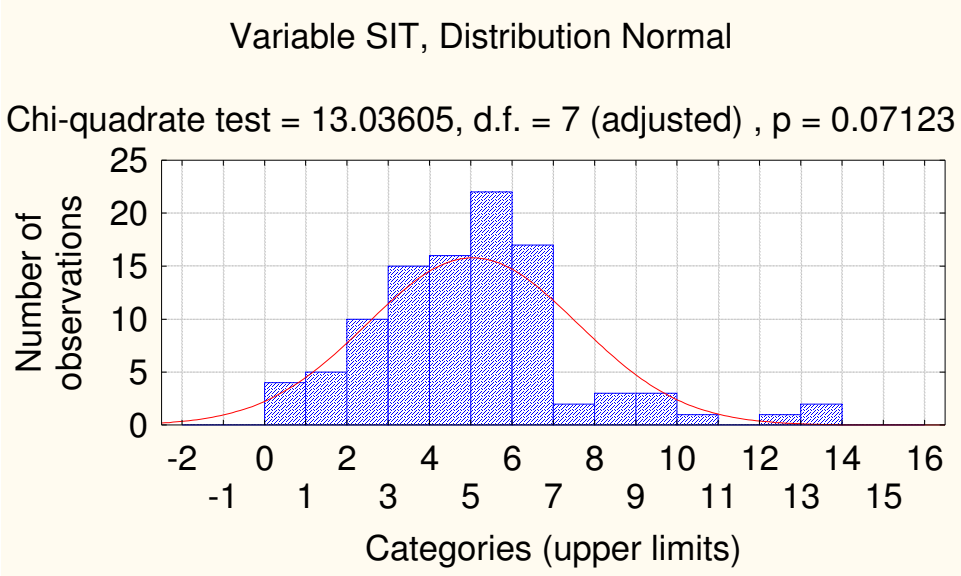
Table 2. - *Limits of the individual variables for transformation.*

	x_{min} -2	0	x_1	1	x_2	2	x_3	3	x_4	4	x_{max}	6
LTH	2.00		4.15		5.77		7.66		10.64		24.00	
SIT	0.01		2.90		4.41		5.68		7.19		13.77	
CAM	2.71		6.67		9.18		11.30		13.81		24.00	

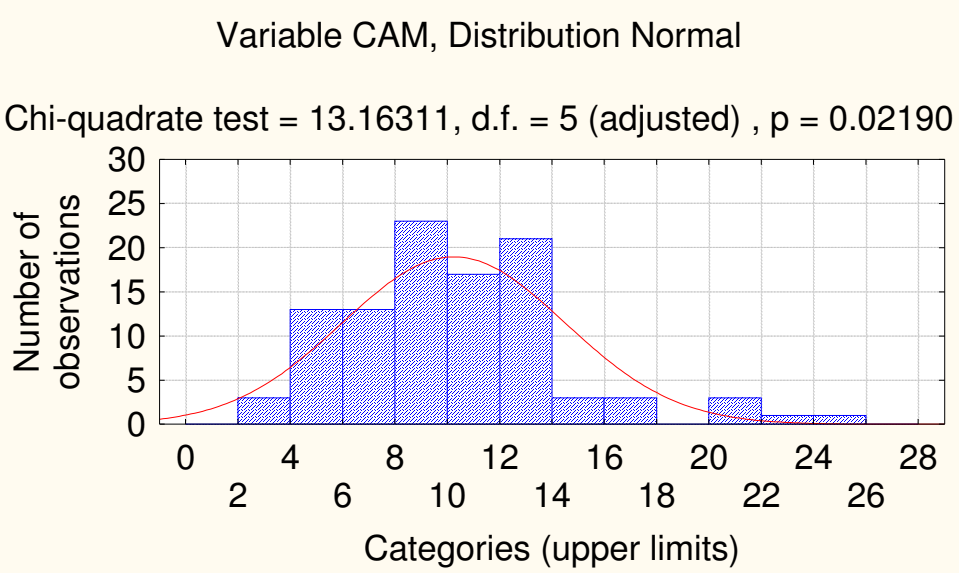
TCH	3.00	4.00	4.65	5.19	5.84	8.00



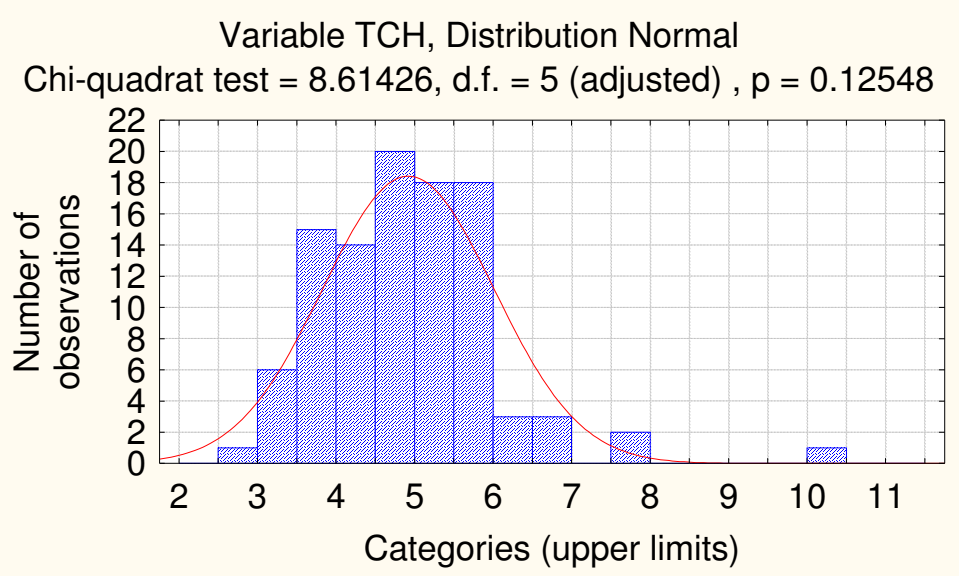
Graph 1 - *Histogram of distribution of values of variable LTH (p ~ 0.09)*



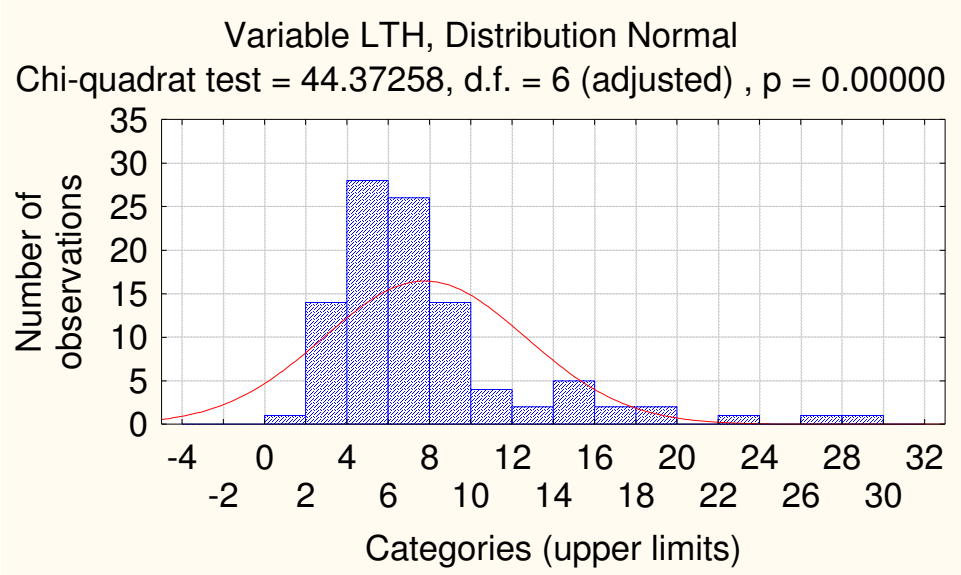
Graph 2 - *Histogram of distribution of values of variable SIT (p ~ 0.07)*



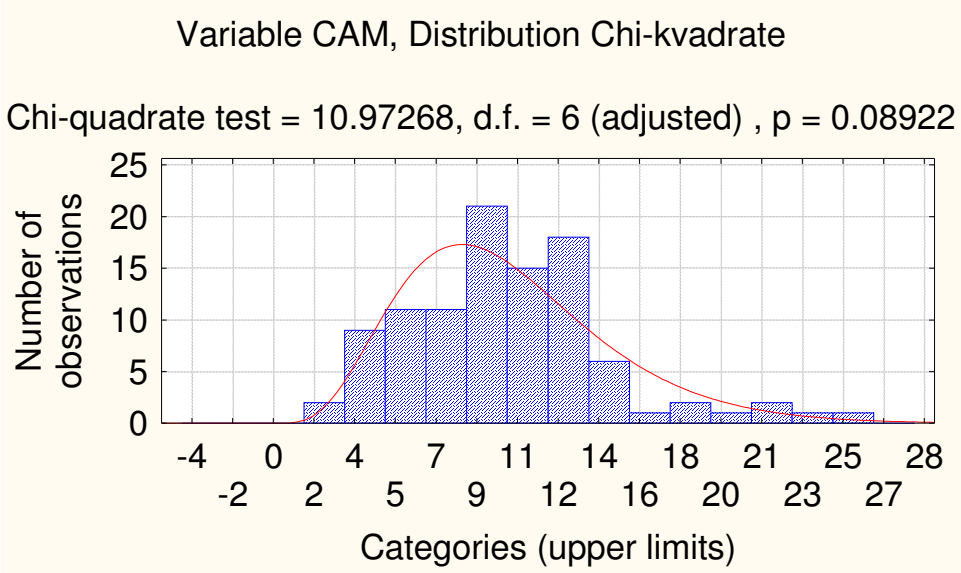
Graph 3 - *Histogram of distribution of values of variable CAM (p ~ 0.02).*



Graph 4 - *Histogram of distribution of values of variable TCH (p ~ 0.13)*



Graph 5 - *Histogram of distribution of values of variable LTH – approximation by normal distribution.*



Graph 6 - *Histogram of distribution of values of variable CAM – approximations by distribition chi- quadrate.*

Table 3a presents the values for 95 healthy subjects transformed from the original ones for all variables under study into a five-degree scale (according to Table 2). (Original data are presented in [1], Table 6). Now let us introduce discrete metric in the set of all arranged tetrads of transformed values of the five-point scale

$$d_1(r, t) = max_i\,(\,|r_i - t_i|\,), \qquad (1)$$

where $r = (r_1, r_2, r_3, r_4)$, $t = (t_1, t_2, t_3, t_4)$, $r_i, t_i \in \{0,1,...,4\}$, $I = 1, 2, 3, 4$.

If we mark the symbols r1, r2, … , r95 of the arranged tetrad of transformed values of variables LTH, SIT, CAM, and TCH gradually for all healthy subjects, we define on the set {r1, r2, … , r95 } matrix D1 by means of (1) thus:

$$\boldsymbol{D}_1 = (d_1(r^i, r^j)),\ i, j = 1, 2, \ldots, 95. \qquad (2)$$

Matrix D1 is symmetrical and with zeros in the diagonal; the following expression holds true for its elements

$$0 \le d_1(r^i, r^j) \le 4. \quad (3)$$

The space of all possible tetrads of scores r = (r1, … , r 4) possesses 54 = 625 elements and it is expected that not all of the possibilities will appear in the sample under investigation. If in some of the more extensive sets of data of the "healthy" subjects nearly all tetrads of reduced data under study would be covered by their experimental values, it would not be possible inside the above-mentioned set of scale values to separate the set of the "healthy" subjects from patients. The cause can be then either incorrect selection of the variables under study, the width of the interval of possible values, or too rough discretisation of the selected continuous variables. Here only 95 subjects are available in whose group, in addition, there can be two subjects equivalent from the viewpoint of metric d1. For reasons of economy, we will therefore exclude from the table of healthy subjects the subjects with the result rj, who with the fixed i ≠ j have the value d1(ri, rj) = 0; i = 1, 2, … , 95; j = i +1. Of the subjects with the mutual value of metric equal to zero, only one has thus remained.

Let us now compare the reduced set of all healthy subjects Z = { i1, i2, … , im }with the set of all patients N. Prior to it, we will extend the five-level scale of the transformed values of variables to further levels. It will be carried out (again only with regard to the ensemble of healthy subjects) in such a way that when the measured values in the ensemble of patients will be smaller than the minimal value of this variable in the ensemble of healthy subjects, we will assign to this value transformed –2 and in case the value of the pertinent variable will be greater than the maximum of the value of this variable in the set of the healthy subjects, we will assign to it the transformed value equal to 6 (see Table 2). We will thus obtain a set of patients N, newly represented by 189 tetrads of values transformed to a seven-degree scale:

$$T_N = \left\{t^1, t^2, ..., t^{189}\right\}, \; t^i = (t^i_1, t^i_2, t^i_3, t^i_4,) \quad (4)$$

$t^i_j \in \{-2,0,1,...,6\}$, $i = 1, 2, \ldots, 189, j = 1, 2, 3, 4.$

Let us calculate all d(r, t), where r = (r1, … , r4) goes through all tetrads of values of the pertinent variables of the healthy subjects from Z and t = (t1, … , t4) goes through all tetrads of the values of the variables in patients from in N. Let us now discard from the already reduced group of healthy subjects Z all subjects who will have for some j (where j goes through all patients from N) the value of the metric d1 in the interval

$$0 \le d_1(r^i, t^j) \le 1, \; i \in Z. \quad (5)$$

If there is a new, hitherto not included subject A with the tetrad of the originally found values xA = (xA1, xA2, xA3, xA4) = (LTHA, SITA, CAMA, TCHA), we transform his or her tetrad of measurements xA according to Table 2 to the transformed values rA = (rA1 , …, rA4). Then we take the reduced repertory of healthy subjects and step by step determine the values d1(rA, ri), I ∈ Z. If there holds true for some i that 0 ≤ d1(rA, ri) ≤ 1, we include the subject into the healthy ones. If the subject is included into the healthy ones, we examine whether in the patients from N there is at least one, e.g., the j-th, with results represented by the tetrad tj of the character that

$$0 \le d_1(r^A, t^j) \le 1, \; j = 1, 2, \ldots, 189.$$

If it is so, we include the subject into the group of patients. Otherwise, the patient is unclassified.

As d1 is metric, for each healthy subject of the reduced group Z with the values of the tetrad of transformed measurements rZ, for each patient of N with the above-mentioned tetrad rN, and for each of the unclassified subjects A with the tetrad rA , the following inequalities hold true (with regard to (5))

$$2 \le d_1(r^Z, r^N) \le d_1(r^A, r^Z) + d_1(r^A, r^N). \quad (6)$$

If then $0 \le d_1(r^A, r^Z) \le 1$, $d_1(r^A, r^N)$ must be equal to at least 1.

When tightening up condition (5) to the form

$$0 \le d_1(r^i, t^j) \le 2, \; i \in Z, \; j \in N \tag{7}$$

then (6) would be changed into the form

$$3 \le d_1(r^Z, r^N) \le d_1(r^A, r^Z) + d_1(r^A, r^N) \tag{8}$$

and if $0 \le d_1(r^A, r^Z) \le 1$, then $d_1(r^A, r^N)$ must be at least 2.

This consideration means that in the first case some subject can fulfil the condition for the inclusion both in the healthy subjects and in patients. But in the second case every subject classified as a healthy one cannot fulfil the condition of inclusion in patients.

The decision about the classification of the patient as healthy in the first case would therefore deserve an evaluation using the quantitative index of certainty of correct classification.

Inclusion of the subject A into the "healthy" subjects under the condition

$$0 \le d_1(r^A, r^z) \le 1, z \in Z, \tag{9}$$

can be evaluated by the measure of certainty under the presumption that the set of the "healthy" subjects in the original values of variables can be represented as a defined region (characteristic of the healthy subjects) and each subject is the "healthier", the "further" his or her data are from the limit values of the "healthy" subjects towards the "centre" of this region; e.g., a subject is the "healthier", the larger the radius of the "circle" with the centre given by the coordinates of the healthy subject, containing only the values of the healthy ones, is. (The "circle" $K_{a,\rho}$ with the radius ρ and the centre r^A in the metric space (M, d_1) is

$$K_{a,\rho} = \{r \in M; d_1(r^A, r) \le \rho\}.$$

To each "healthy" subject of z ∈ Z, represented by the pertinent tetrad of the values of the transformed scale rz, the weight v(r2) is assigned in the form

$$v(r^z) = \frac{card(I_Z)}{card(Z)}, \; v(r^Z) \in \left\langle \frac{1}{card(Z)}; 1 \right\rangle, \tag{10}$$

where $I_z = \{j \in Z; d_1(r^z, r^j \le 1\}$; the symbol *card(M)* means the number of the elements of the set *M*. Then by the *measure of certainty* of the inclusion of the subject A into the "healthy" ones we understand the number $J_A \in \langle 0;1 \rangle$, which is derived from (11):

$$J_A = \begin{cases} 0, \text{ when } z_0 \in Z \text{ does not exist, so that } d_1(r^{z0}, r^A) \le 1, \\ max_{z \in Z} \{v(r^z); d(r^Z, r^A) \le 1\}; \end{cases} \tag{11}$$

the greater J_A, the higher the certainty of the inclusion of subject A into the healthy ones.

The values of indices (10) and (11) depend on the reduced sample of the "healthy" ones. The more and in greater detail (i.e., when the transformed scale has more levels in each variable) this sample "covers" possible variability of the healthy population, the more reliable the derived measure of certainty of the pertinent decision is.

In Table 3b, each healthy subject z ∈ Z with the diagnosis rz is assigned the weight v (rz) according to (10).

Now let us have the subject A with the diagnosis, e.g., rA = (4,2,2,3). Let us calculate all d1(rA, rz), z ∈ Z. We see that, e.g., d1(rA, r15) = 0. The subject A is then included into the healthy ones with the weight v(r15) = 0.099 (see Table 3b).

Besides metric (1), which is shown here as "very strict", in the space of all tetrads of transformed values we can introduce also metric (12):

$$d^p{}_2(r, t) = (\sum_i |r_i - t_i|^p)^{\frac{1}{p}}, \qquad (12)$$

$r = (r_1, r_2, r_3, r_4)$, $t = (t_1, t_2, t_3, t_4)$, where $r_i, t_i \in \{-2,0,1,\ldots,6\}$, $i = 1,2,3,4$, p is a random number $p \geq 1$. (A suitable selection of p makes it possible to change the metric, for $p = 2$ the Euclidean metric is selected.)

We can proceed in a similar way, i.e., we will first reduce the number of the "healthy" ones Z (represented by the tetrads of discrete transformed values according to Table 1) in such a way that only those subjects will remain who differ from each other in metric (12). Then, by comparing with the ensemble of patients N , we will discard from the ensemble of the "healthy" ones Z those subjects with diagnosis rZ, for whom there exists in the ensemble of patients at least one with the diagnosis t of such a character that

$$0 \leq d^p{}_2(r^Z, t) \leq 1. \qquad (13)$$

Then we declare as "healthy" the subject A with the diagnosis rA, which for some z ∈ Z possesses the value of metric dp2, fulfilling the condition

$$0 \leq d^p{}_2(r^z, r^A) \leq 1. \qquad (14)$$

The measure of certainty JA of this decision is again determined according (11), where v (rz) for z ∈ Z is determined from (10), where, however, IZ = $\{j \in Z; d_2{}^p(r^z, r^j) = 1\}$.

Note: For the reference set of patients A the measure of certainty can be constructed similarly by including the subject into patients (when dp2(rz ,rA) >1 for all z ∈ Z).

Table 3a. - *Transformed values of the group of the "healthy" subjects.*

Sample	*LTH*	*SIT*	*CAM*	*TCH*		Sample	*LTH*	*SIT*	*CAM*	*TCH*
1	1	3	2	0		31	4	1	1	2
2	3	3	1	3		32	4	4	4	3
3	3	4	4	3		33	4	2	3	2
4	2	4	4	2		34	1	2	2	0
5	0	4	3	0		35	2	1	0	1
6	2	4	4	3		36	1	2	3	1
7	2	4	2	4		37	1	0	0	1
8	4	3	2	3		38	0	1	0	0
9	2	3	3	0		39	3	1	2	3
10	1	3	2	0		40	0	1	0	0
11	1	3	4	1		41	2	1	1	2
12	2	4	4	2		42	2	3	3	3
13	1	1	2	0		43	3	2	1	0

14	3	3	3	1		44	1	3	3	2
15	4	2	2	3		45	2	4	4	3
16	3	0	0	0		46	1	1	2	1
17	1	4	4	1		47	2	3	3	3
18	2	1	2	2		48	2	0	0	1
19	4	2	2	3		49	3	1	0	3
20	4	4	3	3		50	2	1	1	2
21	3	3	3	1		51	1	2	2	4
22*	1	1	1	0		52	2	1	2	4
23	4	3	2	1		53	2	1	0	3
24	2	2	2	0		54	1	3	1	4
25	3	3	4	0		55	1	1	0	2
26	3	3	3	1		56	0	2	1	4
27	2	2	3	0		57	0	0	0	2
28	1	1	3	0		58	1	2	1	3
29	4	0	0	0		59	1	1	0	2
30	2	3	3	1		60	3	1	0	2

Sample	*LTH*	*SIT*	*CAM*	*TCH*		Sample	*LTH*	*SIT*	*CAM*	*TCH*
61	0	1	0	1		81	3	0	0	2
62*	0	3	2	4		82*	0	0	2	2
63	2	1	1	2		83*	4	0	2	0
64	3	3	3	4		84	4	2	2	1
65	2	2	3	4		85*	4	3	3	3
66	1	2	3	3		86	2	1	2	3
67	3	2	4	2		87	0	2	3	3
68	1	2	2	2		88	2	0	1	3
69	1	3	3	2		89	4	0	0	4
70	2	0	0	2		90	4	0	0	1
71	0	2	1	3		91	2	2	4	2
72	0	3	2	4		92	3	0	0	0
73	2	3	3	4		93	1	2	3	0
74	2	0	0	0		94	0	0	1	0
75	0	2	2	4		95	2	3	4	3
76	2	1	0	1		96	0	3	4	3
77	1	0	1	0		97	1	0	3	1
78	0	1	2	1		98	1	0	1	3
79	3	3	3	3		99	4	0	1	2
80*	4	4	4	4		100	3	1	1	3
						101	1	3	3	3

Table 3b - *Each healthy subject from $z \in Z$ is assigned the weight $v(r^z)$.*

z	1	2	3	4	5	6	7	8	9	10
$v(r^z)$	0.089	0.04	0.129	0.158	0.04	0.109	0.079	0.079	0.149	0.089
$card(I_{Z_j})$	9	4	13	16	4	11	8	8	15	9
z	11	12	13	14	15	16	17	18	19	20
$v(r^z)$	0.119	0.149	0.119	0.139	0.099	0.069	0.079	0.139	0.099	0.059
$card(I_{Z_j})$	12	15	12	14	10	7	8	14	10	6
z	21	22*	23	24	25	26	27	28	29	30
$v(r^z)$	0.139	0.149	0.059	0.158	0.069	0.139	0.158	0.089	0.003	0.208
$card(I_{Z_j})$	14	15	6	16	7	14	16	9	3	21
z	31	32	33	34	35	36	37	38	39	40
$v(r^z)$	0.099	0.005	0.119	0.119	0.168	0.168	0.168	0.059	0.139	0.059
$card(I_{Z_j})$	10	5	12	12	17	17	17	6	14	6
z	41	42	43	44	45	46	47	48	49	50
$v(r^z)$	0.218	0.188	0.05	0.168	0.119	0.158	0.188	0.158	0.119	0.218
$card(I_{Z_j})$	22	19	5	17	12	16	19	16	12	22

z	51	52	53	54	55	56	57	58	59	60
$v(r^z)$	0.158	0.089	0.129	0.079	0.158	0.069	0.05	0.168	0.158	0.149
$card(I_{Zj})$	16	9	13	8	16	7	5	17	16	15
z	61	62*	63	64	65	66	67	68	69	70
$v(r^z)$	0.089	0.099	0.218	0.168	0.119	0.188	0.109	0.178	0.168	0.158
$card(I_{Zj})$	9	10	22	17	12	19	11	18	17	16
z	71	72	73	74	75	76	77	78	79	80*
$v(r^z)$	0.099	0.099	0.129	0.079	0.099	0.168	0.129	0.109	0.198	0.059
$card(I_{Zj})$	10	10	13	8	10	17	13	11	20	6
z	81	82*	83*	84	85*	86	87	88	89	90
$v(r^z)$	0.149	0.04	0	0.069	0.099	0.139	0.099	0.158	0.02	0.069
$card(I_{Zj})$	15	4	0	7	10	14	10	16	2	7
z	91	92	93	94	95	96	97	98	99	100
$v(r^z)$	0.139	0.069	0.119	0.089	0.149	0.05	0.05	0.129	0.069	0.178
$card(I_{Zj})$	14	7	12	9	15	5	5	13	7	18
z	101									
$v(r^z)$	0.198									
$card(I_{Zj})$	20									

2 Conclusion

We have demonstrated mathematical model (and therefore necessarily formalized method) which under different conditions analyzed a set of measurements of several variables simultaneously. The analysis aimed to find the relation between the variables under study and the variable categorial Y characterizing the condition of a statistical unit (a respondent). The variable Y was categorized into two levels – "healthy" ($Y = 1$) or "ill" ($Y = 0$).

Reference:

[1] PŮLPÁN, Z. Shluková analýza a její aplikace, Acta medica (Hradec Králové), Suppl. 2002, 45 (1), 25 – 43

[2] PŮLPÁN, Z. K formální definici nemoci. Acta medica (Hradec Králové) SUPPL 2003; 46 (1 – 2), 79 – 99

[3] PŮLPÁN, Z. K problematice zpracování empirických šetření v humanitních vědách, Academia, Praha: 2004, 182 s. ISBN 80-200-1221-4

Contact:

Prof. RNDr. PhDr. Zdeněk Půlpán, CSc.
Katedra matematiky, PdF UHK
Rokitanského 62
500 03 Hradec Králové 3
e-mail: zdenek.pulpan@uhk.cz, tel. č.: 493 331 153