ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 164, 2003

*Wiesław Wagner**, *Małgorzata Kobylińska***

NUMERICAL PRESENTATION OF THE SELECTED STATISTICAL NOTIONS BASED ON TUKEY'S CONCEPT OF OBSERVATION DEPTH IN THE SAMPLE

Abstract. The study provides presentation of selected statistical concepts based on data depth by Tukey. The concepts as: the rang of depth, the half-space convex, contour, simplicial depth breakdown points, position numerical measures, the trimmed mean depth, regression depth and the set of generally positive points were exemplified in a two-dimensional space of the dataset. There are also given numerical algorithms in some cases to indicate already mentioned concepts and to study their affined transformation. There are also given exemple for a one-dimensional space besides general description.

Key words: depth measure, contour of depth, simplicial depth, the cosine method, the convex combination method, the tree triangle area method, the angular transformation method.

1. INTRODUCTION

The development of the mathematical statistics theory is strongly based upon the limit theorems that incorporate appropriate sample statistics in the process of statistical inference. The following statistics are of particular importance: empirical distribution function, sample moments, sample quantiles, order statistics, empirical characteristic function, *U*-statistics and *L*-, *R*-, *M*-estimators (Serfling 1980).

In the recent years, thanks to Tukey's (1975) work, many new notions connected to numerical data explorative analyses have been introduced. One of the new notions is data depth used for visualisation of both one-dimensional or multivariate numerical data. Since their introduction, data exploration techniques have developed into different tool forms such as

* Professor, Department of Statistics, Academy of Physical Education, Poznań.
** Master of Science, University of Warmia and Mazury, Olsztyn.

- decision trees,
- neural networks,
- analysis time event,
- inductive search and recognition rule,
- data visualisation – detection of correlations in multivariate data,
- online analyses and ad hoc requests – data studying in different cross-sections and data hypercube dimensions, summing by particular dimensions and their subranges.

Many scientists (Liu 1990, Donoho and Gasko 1992, Rousseeuw and Ruts 1996, He and Wang 1997, Struyi and Rousseeuw 1999) have studied data depth in respect of its usefulness for one-dimensional and multivariate data statistical description. Observation depth in a sample has been proposed as a tool for the determination of multivariate order statistics, particularly for outliers data, encumbered with outliers observations.

2. GENERAL CONCEPT OF DEPTH MEASURE

Let the following set $P_n^d = \{x_1, x_2, ..., x_n\}$ be the system of the observable vectors expressing the d-dimensional n-sized sample, given from a certain d-dimensional distribution defined by F_d distribution function and let $\theta \in R^d$ be a certain point in the real space R^d. Point θ may in particular belong to the system of points from the P_n^d sample. The depth measure of θ point in the P_n^d sample is expressed by $D_d(\theta: x_1, x_2, ..., x_n) = D_d(\theta : P_n^d)$, which meets the following (He and Wang 1997):

(D1) Set $O_c^n = \{\theta : D_d(\theta : P_n^d) \geq c\}$ is convex and restricted to almost all n and c,

(D2) Occurs $\lim_{n \to \infty} D_d(\theta : P_n^d) = D(\theta)$ for almost all θ, and $D(\theta)$ are contours in the form of $\{\theta : e(\theta) = c\}$ for certain $e(\theta)$,

(D3) For certain compact sets $C \subset R^d$, occurs $\lim_{n \to \infty} |D_d(\theta : P_n^d) - D_d(\theta)| = 0$,

(D4) The $D(\theta)$ contour is a strictly monotonous function $e(\theta)$, which implies, that for certain $c > 0$, occurs the following probability $P(\theta : D(\theta) = c\} = 0$.

Intuitive θ point depth measure in the P_n^d sample is expressed by the smallest number of points from the P_n^d sample located at one side of the θ number. With the use of the depth measure we can sort P_n^d sample elements into the $x_{(1)}, x_{(2)}, ..., x_{(n)}$ array in which the following $D_d(x_{(1)} : P_n^d) \geq D_d(x_{(2)} : P_n^d) \geq ... \geq D_d(x_{(n)} : P_n^d)$ array occurs.

3. ONE-DIMENSIONAL SAMPLE DEPTH

Let now $P_n = \{x_1, x_2, ..., x_n\} = \{x_i, i = 1, 2, ..., n\}$ express one-dimensional n-sized sample from the population of distribution defined by F distribution function, $\theta \in P_n$ express the observation in this sample and let set measures: $i_-(\theta) = \#(i : x_i \leqslant \theta\}$, $i_+(\theta) = \#\{i : x_i \geqslant \theta\}$ express the observation number from P_n sample, not exceeding (not larger) or exceeding (not smaller) than any of the observations $\theta \in P_n$, where $\#\{.\}$ means the population (cardinal number) of the set in question.

Let now $P_n^{\leqslant}$ and $P_n^{\geqslant}$ express P_n sample sorted non-decreasingly or non-increasingly, respectively. Then the functions of $i_-(\theta)$ and $i_+(\theta)$ of the $\theta \in P_n$ variable define ranks the θ observation in the sample sorted non-decreasingly or non-increasingly.

Definition. The depth for the observation $\theta \in P_n$ is the number representing the lower rank from $P_n^{\leqslant}$ or $P_n^{\geqslant}$ samples, which is the smallest observation number from P_n sample at the same time not exceeding or exceeding the sample, which can be expressed as follows:

$$D_1(\theta) = \text{depth}_1(\theta; P_n) = \min\{i_-(\theta), i_+(\theta)\} = \min(\#\{i; x_i \leqslant \theta\}, \#\{i, x_i \geqslant \theta\}). \tag{1}$$

The above function allows for direct definition of the following order statistics:

– the smallest observation

$$x_{(1)} = \text{depth}_1(\theta; P_n) = 1 \text{ and } i_-(\theta) = 1,$$

– the largest observation

$$x_{(n)} = \text{depth}_1(\theta; P_n) = 1 \text{ and } i_+(\theta) = 1,$$

– lower quartile

$$Q_1 = \text{depth}_1(\theta; P_n) \approx n/4 \text{ and } i_-(\theta) \approx n/4,$$

– upper quartile

$$Q_3 = \text{depth}_1(\theta; P_n) \approx n/4 \text{ and } i_+(\theta) \approx n/4,$$

– median (middle quartile)

$$\text{Med} = \text{depth}_1(\theta; P_n) \approx n/2 \text{ and } i_-(x) \approx i_+(\theta) \approx n/2,$$

– a trimmed arithmetic average $(0 < a < 1)$

$$T_a(P_n) = \text{mean}\{x_i \in P_n : \text{depth}_1(x_i;\ P_n) \geqslant an\}.$$

Lemma. One-dimensional depth is invariant into linear transform $x_i \rightarrow ax_i + b$, which is $\text{depth}_1(a\theta + b;\ \{ax_i + b\}) = \text{depth}_1(\theta; P_n)$ for constants $a > 0$ and b.

Definition. Let D_k be the set of all $\theta \in P_n$, for which $\text{depth}_1(\theta;\ P_n) \geqslant k$. The D_k set is called contour of depth of the k-order.

The D_k can be identified with quasi-range $R_{k,n-k+1} = x_{(n-k+1)} - x_{(k)}$ for $k = 1, 2, \ldots, [n/2]$, and for $k = 1$, quasi-range changes into a range in simple respect, and $R = R_{1,n} = x_{(n)} - x_{(1)}$, and $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ is the array of order statistics from the P_n sample.

Directly related to the problem of depth is excessive outlyingness of the x observation in the P_n sample, whose depth is described as follows:

$$r_1(\theta;\ P_n) = \frac{|\theta - \text{Med}(P_n)|}{\text{MAD}(P_n)},$$

where $\text{Med}(P_n)$ is a median, and $\text{MAD}(P_n)$ median absolute deviation from the P_n sample in the form of

$$\text{MAD}(P_n) = \text{Med}\,|x_i - \text{Med}(P_n)|.$$

Simplex notion is one of the ways of defining depth for one-dimensional sample, which in the case of R^1 space is transformed into a section. Such conception was earlier reported by L i u (1990), in which the depth function has the following form:

$$D_1(\theta) = \binom{n}{2}^{-1} \sum_{1 \leqslant i < j \leqslant n} I[\theta \in \overline{x_i x_j}], \tag{2}$$

where $I(A)$ represents the indication function of the A event, where $I(A) = 1$, if it occurred and $I(A) = 0$ otherwise, a $\overline{x_i x_j}$ is a section of the x_i and x_j points.

The study of $\theta \in R^1$ point belonging to the $\overline{x_i x_j}$ section, as it was given in the depth criterion (2), which means that $\theta \in \overline{x_i x_j}$, could be replaced with the study, into whether the convex combination has been met

$$\theta \in (1 - \lambda)x_i + \lambda x_j, \text{ dla } \lambda \in \langle 0,\ 1\rangle. \tag{3}$$

If $\lambda \notin \langle 0, 1\rangle$, then the θ point does not belong to the section in question. Otherwise for non three numbers θ, x_i, x_j, the θ number belongs to the

$\langle x_i, x_j \rangle$ range, then and only then, when the convex combination has been met (3).

The algorithm of number belonging to none of the sets $\theta \in R^1$, includes the following steps:

(i) Sorting the P_n observation sample into the $P_n^{\leqslant}$ sample, which is a non-decreasing array $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$,

(ii) If the inequality $\theta < x_{(1)}$ lub $\theta > x_{(n)}$ occurs, then the θ number does not fit into the range of the sample variability (in range), when $D_1(\theta) = 0$,

(iii) If the $\theta \in \langle x_{(1)}, x_{(n)} \rangle$ condition has been met for all (i, j) indicator pairs, such as $1 \leqslant i < j \leqslant n$, then the following constant is to be calculated:

$$\lambda = \frac{\theta - x_i}{x_j - x_i}, \quad x_i \neq x_j, \tag{4}$$

as well as, exemplary k number of the $\lambda \in \langle 0, 1 \rangle$ condition satisfaction.

(iv) The depth measure for the given θ then equals $D_1(\theta) = \dfrac{2k}{n(n-1)}$,

(v) If in (4) occurs the following $x_i = x_j$ lub $|x_i - x_j| < \varepsilon$ (e.g. $\varepsilon = 10^{-4}$), then the section is reduced to a point which may be regarded as their ε – overlapping.

4. TWO-DIMENSIONAL SAMPLE DEPTH

Let $P_n^2 = \{x_1, x_2, \ldots, x_n\} = \{x_i, i = 1, 2, \ldots, n\}$ be a two-dimensional n-sized sample from a certain distribution described by F_2 two-dimensional distribution function. It means that each element from the P_n^2 sample is a point from the R^2 real space, and $x_i \in R^2$ occurs to $i = 1, 2, \ldots, n$. It is assumed that at least $h = [n/2] + 1$ number of points from the P_n^2 sample do not lie on one line, i.e. the system of $x_1, x_2, \ldots, x_n$ vectors meets the general position set conditions, according to the nomenclature introduced by Donoho and Gasko (1992). At the same time it is assumed that this set does not include nodes and no more than two points lie on one line.

The set of vectors from the P_n^2 sample forms on the R^2 plane a corelation chart, also called a points cloud. For three non-collinear random points: x_i, x_j, x_k from this chart, a closed triangle can be built from the points, which represent its vertex. Such a triangle is symbolically described as follows: $\Delta(x_i, x_j, x_k) \equiv \Delta_{ijk}$.

For the system of n points in R^2, the number of $\binom{n}{3} = \dfrac{1}{6} n(n-1)(n-2)$ different triangles can be built. Lets indicate that the Δ_{ijk} triangle transforms

into a section or becomes a single point when the vectors which form the triangle are collinear. This condition then means that there are such three constants $\lambda_1, \lambda_2, \lambda_3 \in R$, not all equalling zero at the same time, that the linear combination which is $\lambda_1 x_i + \lambda_2 x_j + \lambda_3 x_k = 0$. The mentioned collinearity may be of ε – order, so called ε – collinearity, when the area of the triangle $\Delta_{ijk} < \varepsilon$, where ε is a given sufficiently small constant (e.g. $\varepsilon = 10^{-4}$).

For a random $\theta \in R^2$ let $\theta \in \Delta_{ijk}$ be an event of θ point belonging to the interior of the triangle built on the x_i, x_j and x_k vertex. In other words, the θ vector is expressed by a convex linear combination: $\theta = \lambda_1 x_i + \lambda_2 x_j + \lambda x_k$ with the following conditions $\lambda_1, \lambda_2, \lambda_3 \geqslant 0$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. In case when the indicated convex combination does not occur then the point $\theta \notin \Delta_{ijk}$.

Definition. The simplicial depth measure for the point of $\theta \in R^2$, built on triangles $\Delta(x_i, x_j, x_k)$ from the P_n^2 sample elements, is a ratio of the triangle number including the θ point to the number of all possible triangles, which could be illustrated as:

$$SD_2(\theta) = \text{depth}_2(\theta, P_n^2) = \binom{n}{3}^{-1} \sum_{1 \leqslant i \leqslant j \leqslant k \leqslant n} I[\theta \in \Delta(x_i, x_j, x_k)] \qquad (5)$$

where $I(A)$ is an indication function of the A event.

It was noticed that $SD_2(\theta)$ in the above definition may be regarded as an empirical distribution representing the following probability:

$$D(\theta) \equiv P_F(\theta \in \Delta(x_i, x_j, x_k,)),$$

if $x_1, x_2, \ldots, x_n$ are independent random variables from the two-dimensional distribution described by the F_2 distribution function.

There are several methods of studying whether a $\theta \in R^2$ point belongs to the given triangle $\Delta(x, y, z)$, on condition that the mentioned vectors are not collinear. These are: (a) the cosine method, (b) the linear convex combination method, (c) three triangle area method and (d) angular transformation method by Rousseeuw and Ruts (1996).

(a) The cosine method (Wagner and Kobylińska 2000)

For the given three vectors $x = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$, $y = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}$, $z = \begin{bmatrix} x_3 \\ y_3 \end{bmatrix}$ there will be a triangle built of vertex in A, B, C (Figure 1).

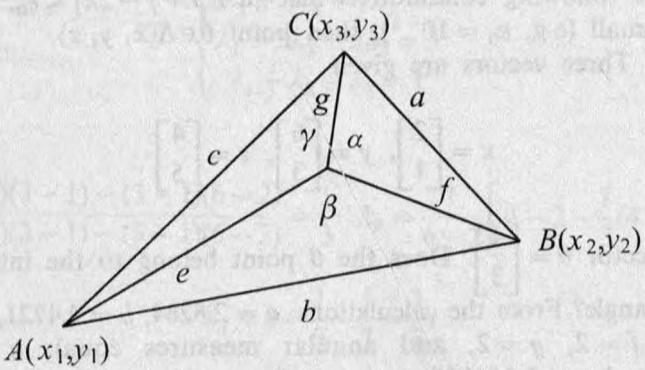

Fig. 1. Interior angles α, β and γ

The condition of the $\theta \in R^2$ point belonging to the triangle $\Delta(x, y, z)$ is expressed by $\alpha+\beta+\gamma=2\pi$, were α, β, γ are interior angles given in Figure 1. The α, β and γ angular measures are determined through the following steps:

(i) determination of triangle side lengths

$$a=\{(x_3-x_2)^2+(y_3-y_2)^2\}^{1/2},\quad b=\{(x_2-x_1)^2+(y_2-y_1)^2\}^{1/2},$$
$$c=\{(x_3-x_1)^2+(y_3-y_1)^2\}^{1/2},$$

(ii) determination of the distance between the vertex points A, B and C and the internal point $\theta=\begin{bmatrix} x_w \\ y_w \end{bmatrix}$, with the use of the following formulas:

$$e=\{(x_1-x_w)^2+(y_1-y_w)^2\}^{1/2},\quad f=\{(x_2-x_w)^2+(y_2-y_w)^2\}^{1/2},$$
$$g=\{(x_3-x_w)^2+(y_3-y_w)^2\}^{1/2},$$

(iii) application of the cosine theorem for each internal triangle

$$\Delta OBC:\ \alpha=\arccos\left[\frac{1}{2fg}(f^2+g^2-a^2)\right],$$
$$\Delta OAB:\ \beta=\arccos\left[\frac{1}{2ef}(e^2+f^2-b^2)\right],$$
$$\Delta OAC:\ \gamma=\arccos\left[\frac{1}{2eg}(e^2+g^2-c^2)\right],$$

(iv) if the following condition is met $|\alpha+\beta+\gamma-2\pi|<\varepsilon_0$, where ε_0 is sufficiently small (e.g. $\varepsilon_0 = 10^{-3}$), then point $\theta \in \Delta(x, y, z)$.

Example. Three vectors are given

$$x=\begin{bmatrix}2\\1\end{bmatrix},\ y=\begin{bmatrix}6\\3\end{bmatrix},\ z=\begin{bmatrix}4\\5\end{bmatrix}$$

as well as vector $\theta=\begin{bmatrix}4\\3\end{bmatrix}$. Does the θ point belong to the interior of the $\Delta(x, y, z)$ triangle? From the calculations: $a = 2.8284$, $b = 4.4721$, $c = 4.4721$, $e = 2.8284$, $f = 2$, $g = 2$, and angular measures equal: $\alpha = 1.570796$, $\beta = 2.356195$ and $\gamma = 2.356195$, and $|\alpha+\beta+\gamma-2\pi = 6.283186 - 6.283185 = 0.000001 < 0.001$, which means that the given point θ belongs to the triangle interior.

(b) The convex combination method (W a g n e r and K o b y l i ń s k a 2000)

It could be assumed that a triangle is a set of all the convex combination points for the given three non-collinear points. The following are vectors: x, y, z and θ is as in (a). If $\theta \in \Delta(x, y, z)$, then $\theta = \lambda_1 x + \lambda_2 y + \lambda_3 z$, $\lambda_1, \lambda_2, \lambda_3 \geqslant 0$ and $\lambda_1 + \lambda_2 + \lambda_3 = 1$ occur, and the following system of linear equations is met:

$$\begin{cases} x_1\lambda_1 + x_2\lambda_2 + x_2\lambda_3 = x_w \\ y_1\lambda_1 + y_2\lambda_2 + y_3\lambda_3 = y_w \\ \lambda_1 + \lambda_2 + \lambda_3 = 1 \end{cases}$$

Here is the solution of the given system:

$$\lambda_3 = \frac{(x_w - x_1)(y_2 - y_1) - (y_w - y_1)(x_2 - x_1)}{(x_3 - x_1)(y_2 - y_1) - (y_3 - y_1)(x_2 - x_1)}$$

$$\lambda_2 = \frac{1}{x_2 - x_1}[x_w - x_1 - \lambda_3(x_3 - x_1)],\ x_2 - x_1 \neq 0,\ \lambda_1 = 1 - \lambda_2 - \lambda_3.$$

So, when the point $\theta \in \Delta(x, y, z)$, then $\lambda_1, \lambda_2, \lambda_3 \geqslant 0$, otherwise, when certain $\lambda_1, \lambda_2, \lambda_3$ are negative, then this point does not belong to the triangle in question.

Example. Let x, y, z and θ be as they are in the example in point (a). Calculations are made for λ_1, λ_2 i λ_3 according to the indicated formulas in the system:

$$\begin{cases} 2\lambda_1 + 6\lambda_2 + 4\lambda_3 = 4 \\ \lambda_1 + 3\lambda_2 + 5\lambda_3 = 3 \\ \lambda_1 + \lambda_2 + \lambda_3 = 1 \end{cases}$$

and

$$\lambda_1 = \frac{(4-2)(3-1)-(3-1)(6-2)}{(4-2)(3-1)-(5-1)(6-2)} = \frac{1}{3},\ \lambda_2 = \frac{1}{6-2}\left[4-2-\frac{1}{3}(4-2)\right] = \frac{1}{3},$$

$$\lambda_3 = 1 - \frac{1}{3} - \frac{1}{3} = \frac{1}{3},$$

then all $\lambda_1, \lambda_2, \lambda_3 \geqslant 0$, and the point in question θ belongs to the interior of the triangle built on the: x, y, z.

(c) The three triangle area method (W a g n e r and K o b y l i ń s k a 2000)

There are three non-collinear points given on a plane as in point (a), $x, y, z \in R^2$, and $x = (x_1, y_1)'$, $y = (x_2, y_2)$, $z = (x_3, y_3)'$. The given vectors, based on the assumption made, determine a certain closed triangle $\Delta = \Delta(x, y, z)$. Lets further assume that $\theta = (x_w, y_w)' \in R^2$ is an internal point of the triangle. This allows three internal triangles to be built and carry out the complete triangle division so they are disjointed and their area sum equals the area of the triangle Δ, as was shown in Figure 2.

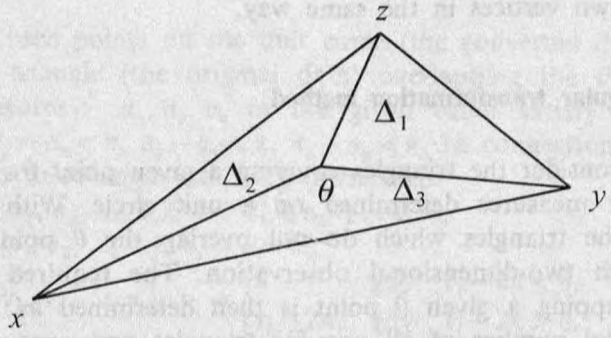

Fig. 2. Complete division of the triangle $\Delta(x, y, z)$

There are three triangles such as $\Delta_1 = \Delta_1(\theta, y, z)$, $\Delta_2 = \Delta_2(\theta, x, z)$, $\Delta_3 = \Delta_3(\theta, x, y)$, which meet the condition of the complete division: $\Delta = \Delta_1 \cup \Delta_2 \cup \Delta_3$ and $\Delta_i \cap \Delta_j = \varnothing$, $i \neq j$ and $i, j = 1, 2, 3$. Let $S(x, y, z)$ be the Δ triangle area, which is a determinant function in the following form:

$$S(x, y, z) = \operatorname{abs}\left\{\frac{1}{2}\begin{vmatrix} 1 & 1 & 1 \\ x & y & z \end{vmatrix}\right\} = \operatorname{abs}\left\{\frac{1}{2}\begin{vmatrix} 1 & 1 & 1 \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix}\right\} =$$

$$= \frac{1}{2}\operatorname{abs}\{(y_1 - x_1)(z_2 - x_2) - (z_1 - x_1)(y_2 - x_2)\},$$

where the symbol abs(.) is an absolute value. In the same way, the areas of $S_1(\theta, y, z)$, $S(\theta, x, z)$ and $S(\theta, x, y)$ for the following triangles Δ_1, Δ_2 i Δ_3 were determined. If the condition of $|S - S_1 - S_2 - S_3| < \varepsilon$ is satisfied when ε is sufficiently small (e.g. equalling 10^{-4}), then point θ is an inner point of the triangle $\Delta(z, y, z)$.

We need to consider two cases in the following issue: when point θ belongs to the triangle (a) circumference, (b) vertex.

Case (a). In this situation the θ point should satisfy the equation of the triangle side line to which it belongs. The lines determining the Δ sides according to the symbols in Figure 2, are the following: $L_1 = L_1(y, z)$, $L_2(x, z)$ i $L_3(x, y)$. So, if $\theta \in L_1(y, z)$, the following equality $(y_2 - y_1)(\theta_2 - z_2) = (z_2 - z_1)(\theta_1 - y_1)$ is satisfied. It means that if the $|(y_2 - y_1)(\theta_2 - z_2) - (z_2 - z_1)(\theta_1 - y_1)| < \varepsilon$ inequality is satisfied, the θ point belongs to the L_1 line. Similarly, the same conditions are checked for the other two L_2 and L_3 lines, for the constant, sufficiently small value of ε.

Case (b). If the θ in a vertex point of the Δ triangle, and overlaps with the x, y, or z, then the following conditions should be satisfied e.g. for the x vertex x: $|x_1 - \theta_1| < \varepsilon$ and $|x_2 - \theta_2| < \varepsilon$. These conditions are checked for the other two vertices in the same way.

(d) The angular transformation method

Now let's consider the triangles covering a given point $\theta \in R^2$ with the use of angular measures determined on a unit circle. With the use of these angles, the triangles which do not overlap the θ point are determined for each two-dimensional observation. The required number of triangles overlapping a given θ point is then determined as the subtraction of the total number of all possible triangles and ones that do not overlap. The following steps illustrate the operations leading to obtaining this number:

Step 1. Transferring vectors from the bivariate sample $P_n^2 = \{x_1, x_2, ..., x_n\}$ into centralised vectors: $y_i = x_i - \theta$.

Step 2. Calculating the length of the centralised vectors $d_i = \sqrt{y_i' y_i}$, which are their Euclidean norms.

Step 3. Determining the normalised vectors $u_i = y_i/d_i$ of the length of 1, that means that they are vectors lying on a unit circle.

Step 4. For each $u_i = (u_{i1}, u_{i2})$ vector, inner angles in rd are determined according to the principle given in a specification below including formulas in two versions:

Quarter	Signs	φ = arc sin	φ = arctan
I	$u_1 > 0,\ u_2 > 0,$	$\varphi = \text{arc sin}(u_2)$	$\varphi = \text{arc tg}(u_2/u_1)$
II	$u_1 < 0,\ u_2 > 0,$	$\varphi = \pi/2 - \text{arc sin}(u_1)$	$\varphi = \pi/2 - \text{arc tg}(u_1/u_2)$
III	$u_1 < 0,\ u_2 < 0$	$\varphi = \pi - \text{arc sin}(u_2)$	$\varphi = \pi - \text{arc tg}(u_2/u_1)$
IV	$u_1 > 0,\ u_2 < 0$	$\varphi = 1.5\pi + \text{arc sin}(u_1)$	$\varphi = 1.5\pi - \text{arc tg}(u_1/u_2)$

Step 5. Sequencing the φ_i angles into a non-decreasing array – $a_1, a_2, \ldots, a_n$, and then calculating the subtractions $\delta = a_i - a_{i-1}$, $i = 2, 3, \ldots, n$. If there is $\delta_i > \pi$, then the θ point lies outside the P_n^2 sample and the depth measure equals $D_2(\theta) = \text{zan}_2(\theta, P_n^2) = 0$, which is the end of the algorithm.

Step 6. Rotating by the a_j angle which means transforming $a_i \rightarrow a_i - a_j$, $i = 1, 2, \ldots, n$, which in consequence leads to the following array: $0 = a_1 \leqslant a_2 \leqslant \ldots \leqslant a_n$.

Step 7. Checking the number of angles located in the I and II quarter, by calculating the number of angles, from the array given in the step 6, that satisfy the inequality $a_i < \pi - \varepsilon$, where ε is a given small number (e.g. $\varepsilon = 10^{-4}$). Let m be the number of such angles. If $n = m$, then all angles lie in the upper semi-circle, which at the same time means the lack of triangles overlapping the given θ point and that the depth measure equals zero.

Step 8. Three points on the unit circle (the converted data) determine a $\Delta(x, y, z)$ triangle (the original data) overlapping the θ point, if the angular measures – a_x, a_y, a_z in the given order satisfy the following conditions: $a_y - a_x < \pi$, $a_z - a_y < \pi$, $a_z - a_x < \pi$. In connection to the above determination three angle sets were introduced:

(a) $A_n = \{a_1, a_2, \ldots, a_n\}$, $0 = a_1 \leqslant a_2 \leqslant \ldots \leqslant a_n$,

(b) $T_n = \{\tau_j, \tau_2, \ldots, \tau_n\}$, $\tau_i = a_i + \pi$,

(c) $V_n = \{v_1, v_2, \ldots, v_n\}$, $v_i = \begin{cases} 0 & \text{gdy} \quad \tau_i - 2\pi < 0, \\ \tau_i - 2\pi, & \text{gdy} \quad \tau_i - 2\pi > 0. \end{cases}$

Step 9. Determination of the $h = \{h_1, h_2, \ldots, h_n\}$ angle number table from the A_n, T_n and V_n, sets which exclude the triangles that do not overlap the θ point, with the elements $h_i = f_i - i$. where

$$f_i = \underset{a_j \in A_n}{\#} \{a_i : a_j < \tau_i\} + \underset{a_j \in A_n}{\#} \{a_j : a_j < v_i\}.$$

Step 10. Determination of the "bad" Δ triangles number for each point on a unit circle, which do not overlap the θ point, so $m_i = \binom{h_i}{2}$ and if $h_i < 2$, then $m_i = 0$.

Step 11. The total number q of the triangles overlapping the point θ equals $q = k - \sum_{i=1}^{n} m_i$, where $k = \binom{n}{3} = \frac{1}{6}n(n-1)(n-2)$, and the depth measure for the given θ point equals $D^2(\theta, P_n^2)q/k$.

Example. For a given sample of $P_5^2 = \{(10, 13), (13, 9), (4, 7), (15, 6), (5, 9)\}$ consisting of five ($n = 5$) two-dimensional observations, the depth measure for $\theta = (6, 1; 8, 1)$ point should be calculated. The table below includes angular measures obtained from the above calculations:

i	x_i	y_i	u_i	v_i	φ_i
1	10	13	0.623	0.782	0.899
2	13	9	0.992	0.129	0.130
3	4	7	–0.886	–0.464	3.624
4	15	6	0.973	–0.230	6.051
5	5	9	–0.774	0.633	2.456

Three angle sets are then created according to the given algorithm: $A_5 = \{0, 0.739, 2.326, 3.494, 5.921\}$, $T_5 = \{3.142, 3.910, 5.467, 6.636, 9.063\}$ and $V_5 = \{0, 0, 0, 0.353, 2.779\}$. Elements f_i, are determined from the given sets in the following way:

$$f_1 = \#\{a_j : a_j < 3.142\} + \#\{a_j : a_j < 0\} = 3 + 0 = 3,$$

$$f_2 = \#\{a_j : a_j < 3.910\} + \#\{a_j : a_j < 0\} = 4 + 0 = 4,$$

$$f_3 = \#\{a_j : a_j < 5.467\} + \#\{a_j : a_j < 0\} = 4 + 0 = 4,$$

$$f_4 = \#\{a_j : a_j < 6.636\} + \#\{a_j : a_j < 0.353\} = 5 + 1 = 6,$$

$$f_5 = \#\{a_j : a_j < 9.063\} + \#\{a_j : a_j < 2.779\} = 5 + 3 = 8,$$

then we can determine as follows h_i: $h_1 = 2$, $h_2 = 2$, $h_3 = 1$, $h_4 = 2$, $h_5 = 3$, and from these numbers the following numbers are obtained m_i: $m_1 = 1$, $m_2 = 1$, $m_3 = 0$, $m_4 = 1$ and $m_5 = 3$, and in total we have 6 bad triangles, which do not overlap any of the $\theta = (6, 1, 8, 1)$ points. In total there are $k = 10$ possible triangles for the given sample, the number of the triangles overlapping the given θ point equals 4, and for this point the depth measure equals 0.4.

REFERENCES

Donoho D. L., Gasko M. (1992), *Breakdown Properties of Location Estimates Based on Half-space Depth and Projected Outlyingness*, „The Annals of Statistics", **20**, 1803–1827.

He X., Wang G. (1997), *Convergence of Depth Contours for Multivariate Datasets*, „The Annals of Statistics", **25**, 495–504.

Liu R. Y. (1990), *On a Notion of Data Depth Based on Random Simplices*, „The Annals of Statistics", **18**, 405–414.

Rousseeuw P. J., Ruts I. (1996), *Bivariate Location Depth*, „Applied Statistics", **45**, 516–526.

Serfling R. J. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley & Sons, New York (tłum. pol. *Twierdzenia graniczne statystyki matematycznej*, PWN, Warszawa 1991).

Struyf A., Rousseeuw P. J. (1999), *Halfspace Depth and Regression Depth Characterize the Empirical Distribution*, „Journal of Multivariate Analysis", **69**, 135–153.

Tukey J. W. (1975), *Mathematical and Picturing Data*, Proceedings of International Congress of Mathematics, Vancouver, **2**, 523–531.

Wagner W., Kobylińska M. (2000), *Miary i kontury zanurzania w opisie statystycznym prób dwuwymiarowych*, „Wyzwania i Dylematy Statystyki XXI Wieku", 201–216.

Wiesław Wagner, Małgorzata Kobylińska

WYBRANE POJĘCIA STATYSTYCZNE W ŚWIETLE KONCEPCJI ZANURZENIA PUNKTU W PRÓBIE – UJĘCIE NUMERYCZNE

(Streszczenie)

W pracy omówiono definicję zanurzenia punktu w próbie oraz wywodzące się z tej koncepcji pewne inne pojęcie statystyczne. Przedstawiono między innymi wskaźniki określające stopień zanurzenia oraz zaproponowano metody numerycznego ich wyznaczania.