A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 235, 2010

Andrzej Mantaj[*]

DETERMINATION AND ANALYSIS OF THE AREAS OF SCATTERED TWO-DIMENSIONAL SAMPLE OBSERVATIONS

Abstract: A base of the analysis of the two-dimensional sample is the correlational graph, where it can mark also the rectangle of the scattering erected on sample extreme positional statisticians of both features. One of manners of the analysis of the distribution of the observation of such sample in the rectangle of the scattering is the settlement of her number in different areas formed by straight parallels trippant by points being found on his sides or distant from his diagonal.

On the paper it present the construction of marking different areas contracted in the rectangle of the scattering and their numbers which it illustrated on empirical material concerning of chosen social-economic occurrences communes municipal of country and of country in Podkarpacie.

Key words: analysis of the two-dimensional sample, rectangle of the scattering.

I. INTRODUCTION

The basis of carrying out the statistical analysis of two variables, i.e. the two-dimensional analysis, is the two-dimensional sample (TDS). It is used in statistical description of each of the variables (edge analysis) and of both variables at the same time (associative analysis). In the first case we set classical and positional numerical characteristics of the sample allowing making the description of the empirical distribution of individual variables. In turn, the associative analysis includes such quantitative and graphical methods which enable characteristic of two-dimensional empirical distribution of pair of variables.

The starting point for TDS analysis is the correlation graph. It enables making assessment of the degree and direction of the dependence occurring between the variables, and the configuration of points on the plane facilitates stating possibly occurring concentrations of data, and, among them, two-dimensional diverging observations.

In the analysis of correlation graph there turns out to be useful the rectangle of dispersion of TDS built on sample extreme positional statistics of both variables (e.g. Wagner 2002). It is treated as generalization of range from the one-dimensional case, and within it there is located TDS.

[*] Ph.D., University of Information Technology and Management in Rzeszów.

One of the ways of analysis of arrangement of observations of TDS in the rectangle of dispersion is to state their quantities in its various regions. These regions can assume the form of triangles, squares, rectangles, polygons, circle, ellipse and ellipsoid of concentration. When creating polygonal regions we use cutting-off straight lines which in the special case are parallel to the main diagonals of the rectangle of dispersion. These straight lines always cross two neighbouring sides of this rectangle.

In literature (e.g. Barnett 1976, Everitt 1978, Mardia et al. 1979, Toit et al. 1986, Rousseeuw and Leroy 1987, Lee 1995, Ostasiewicz 1999, Manly 2005) we can meet many proposals of graphical presentation of TDS data. Among them we can distinguish the following graphs: correlation graphs, regression lines, confidence curves, concentration ellipsoids, convex hulls, plotting index variables and draftman's plot. Many graphs are available in computer statistical packages (e.g. STATISTICA, SPSS, Program R, EXCEL).

In the paper there is described the structure of determining of polygonal regions by cutting-off straight lines which was illustrated on the empirical material concerning municipal-rural and rural districts of Podkarpackie province.

II. TWO-DIMENSIONAL SAMPLE AND ITS RECTANGLE OF DISPERSE

Let (*X, Y*) be two quantitative continuous variables tested on the set of *n* statistical units, i.e. the sequence of their realization in the form of TDS $P_n^2 = \{(x_i, y_i) : i = 1,2,...,n\}$. Its geometrical image in the co-ordinate system is the correlation graph which expresses the configuration of points on the plane *OXY*. It assumes different shape in the form of compact set or disjoint subset of points.

We assume that the values of each feature of sample P_n^2 are ordered non-decreasingly, which is expressed by the inequalities $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ and $y_{(1)} \le y_{(2)} \le ... \le y_{(n)}$. Both sequences of inequalities represent sample positional statistics, respectively for the variables *X* and *Y*. The values of extreme positional statistics $x_{(1)}, x_{(n)}, y_{(1)}, y_{(n)}$ allow marking 4 points on the correlation graph: $A(x_{(1)}, y_{(1)})$, $B(x_{(n)}, y_{(1)})$, $C(x_{(n)}, y_{(n)})$ and $D(x_{(1)}, y_{(n)})$. Connecting these points with intervals *AB*, *BC*, *CD* and *AD* we obtain the rectangle *ABCD*, called rectangle of dispersion (fig. 1).

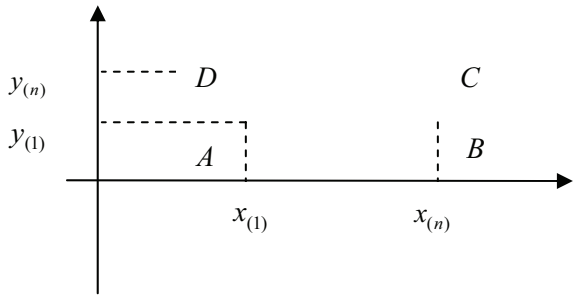

Fig. 1. Diagram of rectangle of dispersion of TDS

Source: Author's elaboration.

The points in this rectangle arrange most often around the diagonal *AC* or *BD*. As the points get closer to one of the diagonals, there increases the force of linear type relation between the tested variables which may be the relation:

- positive, when together with the increase of values of feature *X* there increase values of feature *Y* (points arrange around the diagonal *AC*),
- negative, when to increasing values of feature *X* there correspond decreasing values of feature *Y* (points arrange around the diagonal BD).

Each side of the rectangle of dispersion includes at least one point of TDS, and this rectangle is characterised by the following properties:

a) its surface is equal to the product of ranges of both variables,

b) lengths of its diagonals are equal to the Euclid's distance of their vertex points,

c) the diameter of the set of points of two-dimensional sample may in particular overlap the diagonal of the rectangle of dispersion..

III. DETERMINING POLYGONAL REGIONS

One of the problems of testing of arrangement of observations in the rectangle of dispersion is the assessment of distribution of their quantities in the regions:

a) triangular - lying above and below the diagonals *AC* or *BD*,

b) polygonal - created by two cutting-off straight lines, parallel to the diagonal *AC* or *BD* and equally distant from these diagonals,

c) created by the straight lines as in b) with vertexes *B*, *D* or *A*, *C* (fig. 2).

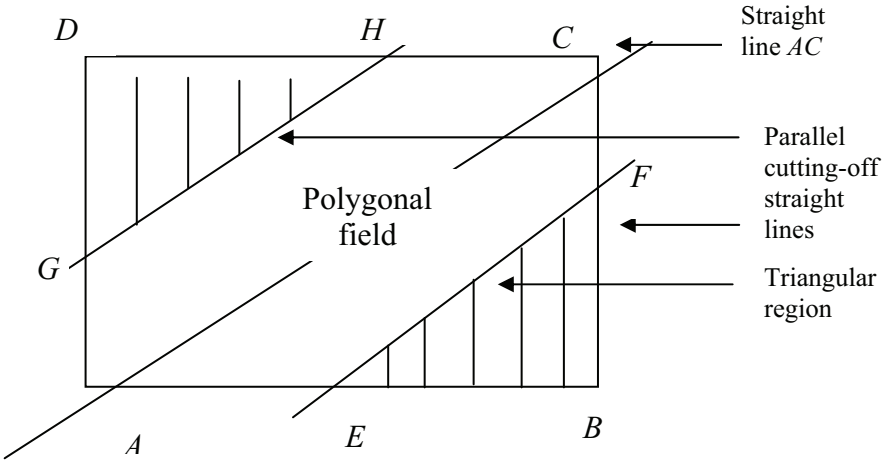

Fig. 2. Distinguished fields in rectangle of dispersion
Source: Author's elaboration

The adequate straight lines are determined in the following way:

A. Diagonal straight line *AC*.

The equation of diagonal straight line going through the points $A(x_{(1)}, y_{(1)})$ and $C(x_{(n)}, y_{(n)})$ in the general form expresses $y - y_{(1)} = \frac{y_{(n)} - y_{(1)}}{x_{(n)} - x_{(1)}}(x - x_{(1)})$, or $y - y_{(1)} = \frac{R_y}{R_x}(x - x_{(1)})$, where $R_y = y_{(n)} - y_{(1)}$ and $R_x = x_{(n)} - x_{(1)}$ are ranges of variables *Y* and *X* or lengths of sides *AD* = *BC* or *AB* = *CD*. We write this equation in the directional form as $y = ax + y_{(1)} - ax_{(1)}$, where $a = \frac{R_y}{R_x}$ is the coefficient of slope of straight line, i.e. the tangent of angle of slope of straight line towards the X-axis. This straight line allows classifying TDS observations (x_i, y_i) to the triangle *ABC* if $y_i \le y(x_i)$, otherwise to the triangle *ACD*.

B. Lower cutting-off straight line *EF*.

To determine the lower cutting-off straight line one should assume the order of cut-off $p_1 \in (0,1)$ and to determine on the side *BC* the distance y_c (calculated parallel with the Y-axis) from the point *C*, which equals $y_c = y_{(n)} - p_1 \cdot R_y$. Designating by *E* and *F* the points of crossing of the straight line cutting off the sides *AB* and *BC* we obtain their coordinates: $E(x_e, y_{(1)})$ and $F(x_{(n)}, y_f)$, where

$y_f = y_c$, and we determine x_e from the condition of parallelism of the straight line *EF*, i.e. $y - y_{(1)} = \frac{y_f - y_{(1)}}{x_{(n)} - x_e}(x - x_e)$ to the straight line *AC*, i.e. $a = \frac{y_f - y_{(1)}}{x_{(n)} - x_e}$, and thus $x_e = \frac{ax_{(n)} - y_f + y_{(1)}}{a}$. It leads to the straight line *EF* of the directional form $y = ax + y_{(1)} - ax_{(n)}$. This straight line enables classification of TDS observations (x_i, y_i) to the triangle *EBF* if $y_i \le y(x_i)$, otherwise to the triangle *AEFCD*.

C. Upper cutting-off straight line *GH.*

We act analogically to B when determining the upper cutting-off straight line. We assume the order of cut-off $p_2 \in (0,1)$ and determine the length y_a of interval on the side *AD* with the beginning at the point A. It equals $y_a = y_{(1)} - p_2 \cdot R_y$. On the sides *AD* and *CD* of rectangle of dispersion we select the points *G* and *H*, in such a way that the straight line going through these points be parallel to the straight line *AC* and distant from it by the quantity y_a. Designating the coordinates of points $G(x_{(1)}, y_g)$ and $H(x_h, y_{(n)})$, we have for them $y_g = y_a$, and x_h is determined from the equation $y - y_a = \frac{y_{(n)} - y_a}{x_h - x_{(1)}}(x - x_{(1)})$. From the condition $AC \parallel GH$ we have $a = \frac{y_{(n)} - y_a}{x_h - x_{(1)}}$, and thus $x_h = \frac{ax_{(1)} - y_a + y_{(n)}}{a}$. Finally it leads to the equation of the straight line *GH* of the directional form $y = ax + y_a - ax_{(1)}$. This straight line enables classification of observations of two-dimensional sample (x_i, y_i) to the polygon *GABCH* if $y_i \le y(x_i)$, otherwise to the triangle *GHD*.

The selection of adequate regions is made on the basis of values of correlation coefficient. If its value is positive, then we use the fields around the diagonal *AC*, which is illustrated in fig. 2, and at the negative coefficient we create fields around the diagonal *BD*.

IV. NUMERICAL EXAMPLE

Assessment of position of observations in various rectangular and polygonal regions will be carried out on the basis of numerical data for two variables: *X* – population density and *Y* – employment indicator which was determined from

the quotient of actually employed people (Z) in relation to general number of people (L). The data for variables X, Z and L come from Statistical Yearbook of Podkarpackie Province 2004. They concern 144 rural and municipal-rural districts of Podkarpackie Province according to the state on 31.12.2002 (also Mantaj and Wagner 2007). The values of variables X and Y are presented on the correlation graph (fig. 3).

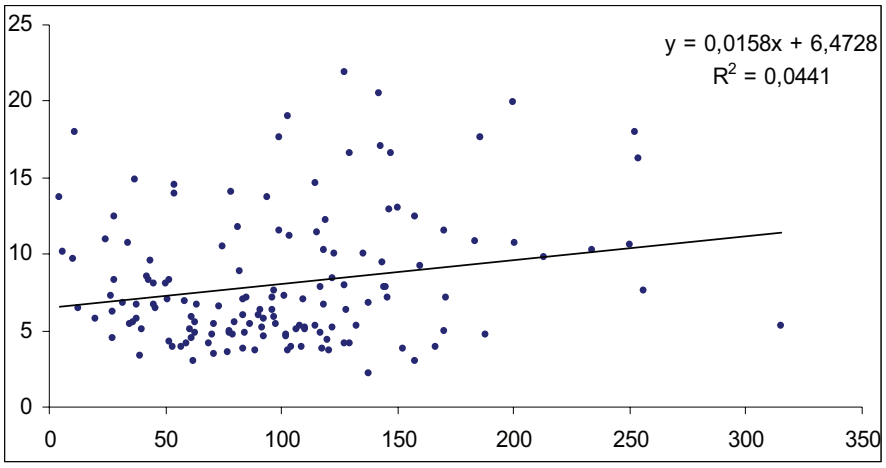

Fig. 3. Correlation graph for variables X and Y

Source: Author's elaboration.

In the figure there is also placed the regression equation for the tested variables and the value of coefficient of determination. General description of both tested variables is given in the statement:

Characteristics	X	Y
Min	4,702	2,193
Lower arithmetic mean	59,879	6,515
Arithmetic mean	99,710	8,053
Upper arithmetic mean	146,782	10,858
Lower quartile	59,176	4,976
Median	95,971	6,704
Upper quartile	127,693	10,255
Max	315,250	21,852

On the basis of given min and max values we create the rectangle of dispersion with the vertexes: $A(4{,}70; 2{,}19)$, $B(315{,}25; 2{,}19)$, $C(315{,}25, 21{,}85)$ and $D(4{,}70, 21{,}85)$. For the assigned rectangle we determine:

- lengths of sides, being ranges of variables $AB = CD = x_{(n)} - x_{(1)} = R_x = 310{,}55$ and $BC = AD = y_{(n)} - y_{(1)} = R_y = 19{,}66$, and since $R_x > R_y$ then the rectangle of dispersion is parallel with its longer side to the Y-axis,
- surface $P_{PR} = R_x \cdot R_y = 6105{,}413$, expressed by the quotient of ranges of both variables,
- length of diagonal $d_{PR} = \sqrt{R_x^2 + R_y^2} = 311{,}172$, equal to the root of sum of squares of ranges of both variables,
- tangent of angle $tg\varphi = \frac{R_y}{R_x} = \frac{19{,}66}{310{,}55} = 0{,}0633$ of slope of the diagonal AC to the side AB, and thus the measure of angle expressed in degrees $\varphi = \frac{180^0}{\pi} arctg(0{,}0633) = 3{,}62^0$,
- coefficient of linear correlation r between the indicated variables equals 0,2101, i.e. in the considered case there occurs very weak positive correlation

For the given rectangle of dispersion there will be illustrated several rectangular and polygonal regions which are described in chapter 3. Due to the positive value of correlation coefficient we determine the equations of diagonal straight line going through the vertex points A and C and of cutting-off straight lines parallel to the diagonal straight line.

The equation of diagonal straight line going through the points $A(4{,}70; 2{,}19)$ and $B(315{,}25; 21{,}85)$ expresses $y = 0{,}06331x + 1{,}89246$, which allowed determining the size of

sample $n = 144$ in the triangles: ABC – 84 (58,33 %) and ACD – 60 (41,67 %). When determining the cutting-off straight lines we assumed equal orders of cut-off $p_1 = p_2 = 0{,}1$, which leads to coordinates of points $E(35{,}75; 2{,}19)$ and $F(315{,}25; 19{,}88)$ and to the equation of lower cutting-off straight line in the directional form $y = 0{,}06331x - 0{,}07354$. The determined straight line EF led to the creation of two subsets of TDS observations, i.e. located in the triangle EBF and belonging to the remaining region of rectangle of dispersion in the number of 54 (37,5%) and 90 (62,5%) respectively.

When determining the upper cutting-off straight line we determined coordinates of points $G(4{,}7; 4{,}16)$ and $H(284{,}19; 21{,}85)$ and its equation $y = 0{,}06331x + 3{,}85846$. Basing on this straight line we classified to the triangle GHD 41 (28,47%), and to the remaining field 103 (71,52%) of observations.

The presented straight lines have been used for determining the sizes of polygonal regions in the rectangle of dispersion presented in fig. 2, assuming various orders of cut-off. Adequate results for the indicated fields are presented in the statement:

Order of cut-off $p_1 = p_2$	Size of cut-off	*EBF*	*AEFC*	*GHD*	*ACHG*	*EFHG*
0,1	1,966	54	30	41	19	49
	%	37,50	20,83	28,47	13,19	34,03
0,15	2,949	40	44	34	26	70
	%	27,78	30,56	23,61	18,06	48,61
0,2	3,932	24	60	26	34	94
	%	16,67	41,67	18,06	23,61	65,28
0,25	4,915	18	66	21	39	105
	%	12,50	45,83	14,58	27,08	72,92
0,3	5,898	11	73	17	43	116
	%	7,64	50,69	11,81	29,86	80,56

Interpretation:

- low percentage (34,03%) of observations in the field *EFHG*, formulated at 10% cut-off of range of feature *Y* indicates a considerable distance of TDS observation from the diagonal *AC*,
- at higher cut-off, i.e. 30% of range of feature *Y*, over 80% of observations gather around the diagonal *AC*, and in the triangles *EBF* and *GHD* there are respectively 11 (7,64%) and 17 (11,81%) of TDS observations.

SUMMARY

The presented analysis of various polygonal regions in the rectangle of dispersion of two-dimensional sample (TDS) enables better learning of configurations of arrangement of observations on the correlation chart. The structure of these regions is determined by the researcher through giving the parameters of orders of cut-off. The idea of this procedure was illustrated by the numerical example, in which, for preserving the symmetry of created polygonal regions there were assumed equal values of both parameters of cut-off.

It should be noted that we can use cutting-off parameters p_1, p_2 at determining of abscissas of rectangle of TDS dispersion. In this case these parameters can be determined as corresponding quantiles of marginal distributions.

The proposed method of analysis of configuration of points on the correlation chart may constitute a supplement of the correlation table.

BIBLIOGRAPHY

Barnett V. (1976), *The Ordering of Multivariate Data (with discussion)*. J. Roy. Statist. Soc, A139, 318–354.

Everitt B. (1978), *Graphical Techniques for Multivariate Data.* North-Holland, New York.

Lee Y.S. (1995), *Graphical Demonstration of an Optimality property of the Median.* The American Statistician 49, No 4, 369–372.

Mardia K.V., Kent J.T., Bibby J.M. (1979), *Multivariate Analysis.* Academic Press, London.

Manly B.F.J. (2005), *Multivariate Statistical Methods, A primer.* Chapman&Hall/CRC, New York.

Mantaj A., Wagner W. (2007), *Comparative analysis of number characteristics of selected social-economic characteristics of communes of Podkarpackie province*, ACTA UNIVERSITATIS LODZIENSIS, Folia Oeconomica 206, s. 445–461

Ostasiewicz W. (1999), *Statystyczne metody analizy danych*, *(Statistical Methods of Data Analysis)* Wyd. AE Wrocław

Rousseeuw P.J., Leroy A.M. (1987), *Robust Regression and Outlier Detection.* Wiley, New York.

Toit S.H.C., Steyn A.G.W., Stumf R.H. (1986), *Graphical Exploratory Data Analysis.* Springer-Verlag, New York.

Wagner W. (2002), *Podstawy metod statystycznych w turystyce i rekreacji*, *(Basics of Statistical Methods in Tourism and Recreation)*, AWF Poznań.

Andrzej Mantaj

WYZNACZANIE I ANALIZA OBSZARÓW ROZMIESZCZENIA OBSERWACJI PRÓBY DWUWYMIAROWEJ

Podstawą analizy próby dwuwymiarowej jest wykres korelacyjny, na którym zaznaczyć można również prostokąt rozrzutu zbudowany na próbkowych skrajnych statystykach pozycyjnych obu cech. Jednym ze sposobów analizy rozmieszczenia obserwacji takiej próby w prostokącie rozrzutu jest ustalenie jej liczebności w różnych obszarach utworzonych przez równoległe proste przechodzące przez punkty znajdujące się na jego bokach lub oddalone od jego przekątnej.

W pracy została opisana konstrukcja wyznaczania różnych obszarów zawartych w prostokącie rozrzutu oraz ich liczebności, którą zilustrowano na materiale empirycznym dotyczącym wybranych zjawisk społeczno-gospodarczych gmin miejsko-wiejskich i wiejskich woj. podkarpackiego.