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

Grzegorz Kończak[*], *Janusz L. Wywiał*[**]

ON THE POWER OF THE CHI-SQUARE TEST FOR MULTIDIMENSIONAL NORMALITY

ABSTRACT. The one dimensional normality chi-square goodness-of-fit test is well known. Here we consider some generalization of this test into multidimensional case. Usually, the construction of the test is based on the contingency table in which the appropriate probabilities deals with events that multidimensional variable takes the value from rectangular cells. In this paper the cells are intersections of sets determined by appropriate ellipsoids and orthogonal planes. The main purpose of the paper is comparison of the test powers evaluated under the two systems of cells. The power is analyzed on the basis of computer simulation.

Key words: multivariate normal test, goodness-of-fit test, Monte Carlo.

I. INTRODUCTION

Our purpose is testing the hypothesis $H_0 : \mathbf{X} \sim N(\mathbf{m}, \mathbf{V})$ where $\mathbf{X} = [X_1, X_2, ..., X_k]$, $\mathbf{m} = E(\mathbf{X})$, $\mathbf{V}(\mathbf{X}) = E(\mathbf{X} - \mathbf{m})(\mathbf{X} - \mathbf{m})^T$. It is well known that there exists such affine transformation $\mathbf{Z} = \mathbf{GX} + \mathbf{g}$ that $E(\mathbf{Z}) = \mathbf{0}$, $\mathbf{V}(\mathbf{Z}) = \mathbf{I_k}$ if $\det(\mathbf{V}) \neq 0$. So, the hypothesis H_0 is equivalent to the following one $H_0 : \mathbf{Z} \sim N(\mathbf{0}, \mathbf{I_k})$.

Let us define the following sets in R^k (see Tallis G. M.,1963 and Tallis G. M. 1965)

- the ring $A_h = \{\mathbf{z} \in R^k : a_{h-1} \leq z^T z < a_h\}$ where $h = 1, 2, ..., r$ and $0 = a_0 < a_1 < ... < a_r = \infty$
- the set $B_{t_1 t_2 ... t_k} = \left\{ \mathbf{z} \in R^k : \begin{cases} z_i > 0, t_i = 1 \\ z_i \leq 0, t_i = 0 \end{cases} \quad i = 1, 2, ..., k \right\}$.
- the set $C(h, j) = \left\{ \mathbf{z} \in R^k : z_{h,j-1} < z_h \leq z_{h,j}, z_t \in R, t \neq h, t = 1, 2, ..., k \right\}$,

$h = 1, 2, ..., k$; $j = 1, 2, ..., e_h$, $-\infty = z_{h,0} < z_{h,1} < ... < z_{h,e_h} = \infty$.

[*] Ph.D., Department of Statistics, Katowice University of Economics.
[**] Professor, Department of Statistics, Katowice University of Economics.

The set $C(j_1, j_2, ..., j_k) = \bigcap_{h=1}^{k} C(h, j_h)$, where $j_h \in \{1, 2, ..., e_h\}$ for h = 1, 2, …, k, is a k-dimensional rectangular. Let us introduce following notations for some probabilities:

$$p_{C(h,j)} = P(\mathbf{z} \in C(h, j)),$$

$$p_{C(j_1,j_2,...,j_k)} = P(\mathbf{z} \in C(j_1, j_2, ..., j_k)) = \prod_{h=1}^{k} p_{C(h,j)},$$

$p_{A_h \cap B_{t_1 t_2 ... t_k}} = P(\mathbf{z} \in A_h \cap B_{t_1 t_2 ... t_k})$, where $A_h \cap B_{t_1 t_2 ... t_k}$ is a part of the ring .

The sample drawn from the distribution of the vector $\mathbf{X}$ is denoted by $\mathbf{X}=[X_{ij}]$, $i = 1, 2, ..., n$; $j=1, ..., k$. A value of the random variable X_{ij} is an i-th observation of the j-th random variable X_j. The sample vector and the sample variance are denoted by $\overline{\mathbf{X}} = [\overline{X}_1, \overline{X}_2, ..., \overline{X}_k]$ and $\overline{\mathbf{V}} = [\overline{V}_{ij}]$, respectively, where $\overline{V}_{js} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{is} - \overline{X}_s)$ where j, s = 1, 2, …, k.

There exists such an orthogonal matrix $\mathbf{G}$ that $\mathbf{G}\mathbf{G}^T=\mathbf{I}_k$, $\mathbf{G}^T \overline{\mathbf{V}} \mathbf{G} = \mathbf{D}$ where $\mathbf{I}_k$ is the unit matrix of degree k. Let $\mathbf{L} = \mathbf{G}\mathbf{D}^{-1/2}$. So, $\mathbf{L}^T \overline{\mathbf{V}} \mathbf{L} = \mathbf{I}_k$. The matrix $\mathbf{X}$ can be transformed as follows:

$$\mathbf{Y}=\mathbf{M}\mathbf{X},\ \mathbf{M} = \mathbf{I}_n - \frac{1}{n} \mathbf{J}_n \mathbf{J}_n^T \tag{1}$$

$$\mathbf{Z} = \mathbf{M}\mathbf{X}\mathbf{C}\mathbf{D}^{-1/2},\ \mathbf{Z}^T\mathbf{Z}=\mathbf{I}_k \tag{2}$$

where $\mathbf{J}_n$ is the unit column which all n-elements are equal to one. Hence, the columns of the matrix $\mathbf{Z}$ are not correlated. Let $\mathbf{Z}_{i*}$ be i-th row of the matrix $\mathbf{Z}$ and $\mathbf{z}_{i*}$ be the value of $\mathbf{Z}_{i*}$.

II. TEST STATISTICS

Transformations (1) and (2) lead to the evaluating the following test statistic of the hypothesis H_0.

$$U_C = \sum_{\{C(e_1,...,e_k)\}} \frac{(W_{C(e_1,...,e_k)} - p_{C(j_1,j_2,...,e_k)})^2}{p_{C(j_1,j_2,...,e_k)}} \tag{3}$$

$$U_{AB} = \sum_{\{A_h \cap B_{t_1 t_2 \ldots t_k}\}} \frac{\left(W_{A_h \cap B_{110\ldots01}} - p_{A_h \cap B_{t_1 t_2 \ldots t_k}}\right)^2}{p_{A_h \cap B_{t_1 t_2 \ldots t_k}}} \tag{4}$$

where

$$W_{C(e_1,\ldots,e_k)} = \frac{Card\left\{z_{i*} :\ z_{i*} \in C\left(j_1, j_2, \ldots, e_k\right)\right\}}{n}$$

and

$$W_{A_h \cap B_{110\ldots01}} = \frac{Card\left\{\mathbf{z}_{i*} :\ \mathbf{z}_{i*} \in A_h \cap B_{t_1 t_2 \ldots t_k}\right\}}{n}.$$

Particularly, if $k = 2$, $r = 5$ and $e_1 = e_2 = 4$ the systems of cells for the test statistic U_C and U_{AB} are showed by Figure 1 and Figure 2, respectively.

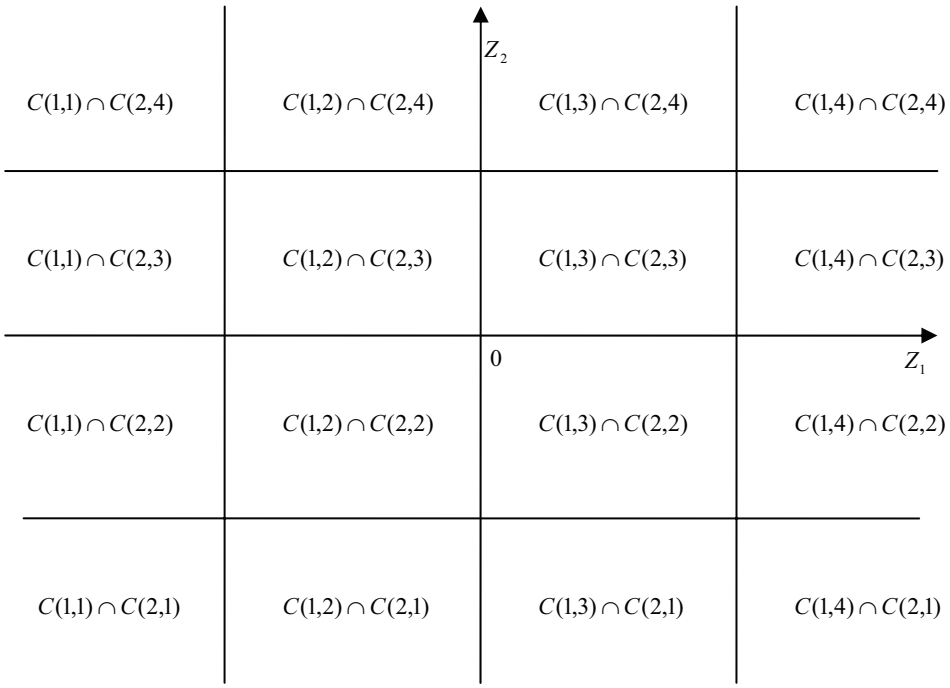

Figure 1. The cells for the statistic U_C

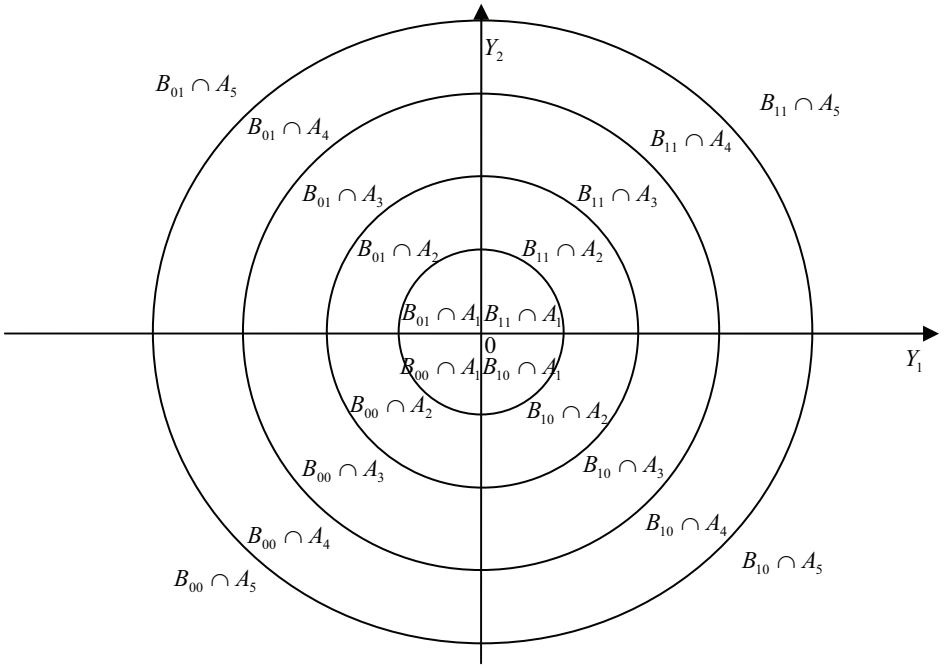

Figure 2. The cells for the statistic U_{AB}

Let $d_{AB} = Card\{A_k \cap B_{t_1 t_2 \ldots t_k}\}$ and $d_C = Card\{C(h_1, h_2, ..., h_k)\}$ be appropriate numbers of cells. When the hypothesis H_0 is true and sample size is sufficiently large then test statistics U_C and U_{AB} have asymptotically chi-square distribution with $d_{AB} - s - 1$ and $d_C - s - 1$ degrees of freedom respectively, where s is the number of estimated parameters.

III. STUDY OF THE TEST POWER

Description of random variables in Monte Carlo study

The size and the power of analyzed tests was estimated in Monte Carlo study. The computer simulations were prepared using R Cran software (www.r-project.org). There are presented the results in the paper for the two dimensional case. The size of tests was obtained for generated two-dimensional normal random variable $\mathbf{X} \sim N(\mathbf{m}, \mathbf{V})$, where $\mathbf{m} = (0,0)$ and $\mathbf{V} = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}$. The power of the described tests were analyzed in Monte Carlo study for following two dimensional distributions:

- LN – Log-normal distribution where both of two variables have log-normal distribution with parameters 0 and 1.
- B(1, 6), B(2, 2) – beta distribution with the shape and the scale parameters
- G(1), G(40) – gamma distribution with the shape parameter.
- U – uniform distribution on the two-dimensional sphere.

The symbolic notation, base description and the asymmetry (third central moment) of these random variables are presented in Table 1.

Table 1. The random variables details in Monte Carlo study

Distribution	Symbolic notation of the random variable	Description of $\mathbf{X} = (X_1, X_2)$	Asymmetry
Normal	N(**m**,**V**)	$\mathbf{m} = (0,0)$ $\mathbf{V} = \begin{bmatrix} 10 & 8 \\ 8 & 10 \end{bmatrix}$	$\gamma_3 = 0$
Log-normal	LN	$X_i = e^X$, i=1, 2. X~N(0, 1)	$\gamma_3 \approx 4$
Gamma	G(1)	X_i has gamma distribution (i=1, 2), with the shape parameter equal to 1.	$\gamma_3 \approx 2$
	G(40)	X_i has gamma distribution (i=1, 2), with the shape parameter equal to 40.	$\gamma_3 \approx 0.4$
Beta	B(1, 6)	X_i has beta distribution (i=1, 2), with the shape parameters equal to 1 and 6.	$\gamma_3 \approx 1.4$
	B(2, 2)	X_i has beta distribution (i=1, 2) with the shape parameters equal to 2 and 2.	$\gamma_3 = 0$
Uniform	U	The density function $f(x,y) = \begin{cases} 1 & x^2 + y^2 \le 1 \\ 0 & x^2 + y^2 > 1 \end{cases}$	$\gamma_3 = 0$

The Monte Carlo study was prepared for h = 3, 4, 5 and 6 zones for each variable in the rectangles cells case and for h = 2, 4, 6 and 8 rings in the parts of rings cells case. The two dimensional case was analyzed. It leads to 9, 16, 25 and 36 cells in the first case and 8, 16, 24 and 32 cells in the second case. The sample size was 45, 80, 120 and 200. The cells were constructed such the expected numbers were equal for each cell. It was taken into account that the minimum expected numbers for each cell should be c.a. 5 observations. The

description of sample sizes for each considered number of cells in two analyzed cases are presented in table 2.

Table 2. Sizes of samples in Monte Carlo study for rectangles cells and parts of rings cells

Rectangles cells		Parts of rings cells	
Number of cells	Sample sizes	Number of cells	Sample sizes
9	45, 80, 120, 200	8	45, 80, 120, 200
16	80, 120, 200	16	80, 120, 200
25	120, 200	24	120, 200
36	200	32	200

The results of Monte Carlo study

The Monte Carlo study was prepared for random variables described in table 1 and for sample sizes and number of cells presented in table 2.

1. The sample of size n (see table 2) from distribution D (see table 1) was generated.
2. The values of the statistics U_{AB} and U_C was calculated and were compared to appropriate critical values.
3. Steps 1–2 were repeated $N = 10\,000$ times.
4. The empirical probabilities of rejection H_0 for the parts of rings (U_{AB}) and the rectangle cells (U_C) were calculated.

There was assumed the significance level a = 0.05 in the Monte Carlo study. The estimated probabilities of rejection of the null hypothesis in the multivariate goodness-of-fit tests was tabled for each case (see tables 3–6).

Table 3. The estimated probabilities of rejection of H_0 for rectangles cells (9 cells) and parts of rings cells (8 cells)

Distribution of random variable	Rectangles 3 zones – 9 cells				The parts of rings 2 rings – 8 cells			
	$n = 45$	$n = 80$	$n = 120$	$n = 200$	$n = 45$	$n = 80$	$n = 120$	$n = 200$
N(**m**,**V**)	0.1787	0.1674	0.1798	0.1777	0.2386	0.2325	0.2283	0.2259
LN	1.0000	1.0000	1.0000	1.0000	0.9989	1.0000	1.0000	1.0000
G(1)	0.9705	0.9991	1.0000	1.0000	0.9790	1.0000	1.0000	1.0000
G(40)	0.2023	0.2205	0.2369	0.2868	0.2720	0.2787	0.2938	0.3352
B(1,6)	0.8780	0.9837	1.0000	1.0000	0.8788	0.9729	1.0000	1.0000
B(2,2)	0.2117	0.252	0.2785	0.3666	0.3250	0.4255	0.5403	0.7451
U	0.3326	0.4832	0.6486	0.8639	0.6535	0.8904	0.9784	0.9996

Source: Monte Carlo study.

Table 4. The estimated probabilities of rejection of H_0 for rectangles cells (16 cells) and parts of rings cells (16 cells)

Distribution of random variable	Rectangles 4 zones – 16 cells			The parts of rings 4 rings – 16 cells		
	$n = 80$	$n = 120$	$n = 200$	$n = 80$	$n = 120$	$n = 200$
N(**m**,**V**)	0.0940	0.0928	0.0879	0.1062	0.1017	0.1001
LN	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
G(1)	0.9991	1.0000	1.0000	1.0000	1.0000	1.0000
G(40)	0.1272	0.1442	0.1749	0.1372	0.1628	0.2085
B(1,6)	0.9874	1.0000	1.0000	0.9961	1.0000	1.0000
B(2,2)	0.1925	0.2435	0.3602	0.2010	0.2673	0.4018
U	0.3449	0.5295	0.8249	0.8097	0.9357	0.9971

Source: Monte Carlo study.

Table 5. The estimated probabilities of rejection of H_0 for rectangles cells (25 cells) and parts of rings cells (24 cells)

Distribution of random variable	Rectangles 5 zones – 25 cells		The parts of rings 6 rings – 24 cells	
	$n = 120$	$n = 200$	$n = 120$	$n = 200$
N(**m**,**V**)	0.0715	0.0756	0.0823	0.0841
LN	1.0000	1.0000	1.0000	1.0000
G(1)	1.0000	1.0000	1.0000	1.0000
G(40)	0.1009	0.1383	0.1290	0.1777
B(1,6)	1.0000	1.0000	1.0000	1.0000
B(2,2)	0.2291	0.3845	0.2896	0.4515
U	0.4789	0.7926	0.9770	1.0000

Source: Monte Carlo study.

Table 6. The estimated probabilities of rejection of H_0 for rectangles cells (36 cells) and parts of rings cells (32 cells)

Distribution of random variable	Rectangles 6 zones – 36 cells	The parts of rings 8 rings – 32 cells
	$n = 200$	$n = 200$
N(**m**,**V**)	0.0610	0.0754
LN	1.0000	1.0000
G(1)	1.0000	1.0000
G(40)	0.1238	0.1476
B(1,6)	1.0000	1.0000
B(2,2)	0.3866	0.4981
U	0.7780	0.9999

Source: Monte Carlo study.

The Monte Carlo studies have shown that the test in which the parts of rings cells are used is most powerful. It can be seen in Uniform, Beta (2,2) and Gamma (40) distribution cases (see table 3–6). The results of Monte Carlo study are presented in the fig. 3 and 4.

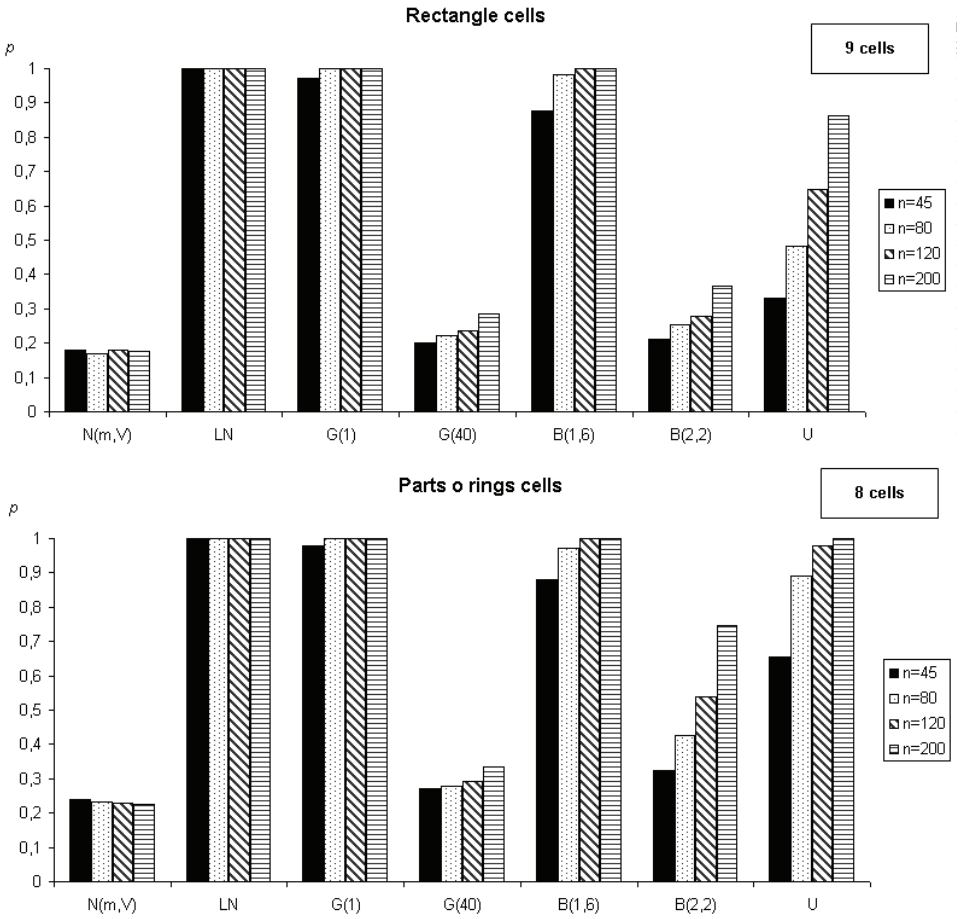

Figure 3. The estimated probabilities for rejection of H_0 for the rectangle cells (9 cells) and the parts of rings cells (8 cells).

Source: Table 3.

The comparison of chi-square goodness-of-fit tests and Shapiro-Wilk's normality test

The power of analyzed normality tests based on the chi-square statistic was compared to the power of the multivariate extension (Royston P., 1982, Domański Cz., 1998) well known Shapiro-Wilk's normality test. The case of

sample of size n = 200 were analyzed in Monte Carlo study. The estimated probabilities of rejection of the null hypothesis for these tests were obtained in N = 10 000 simulations. The results are presented in table 7.

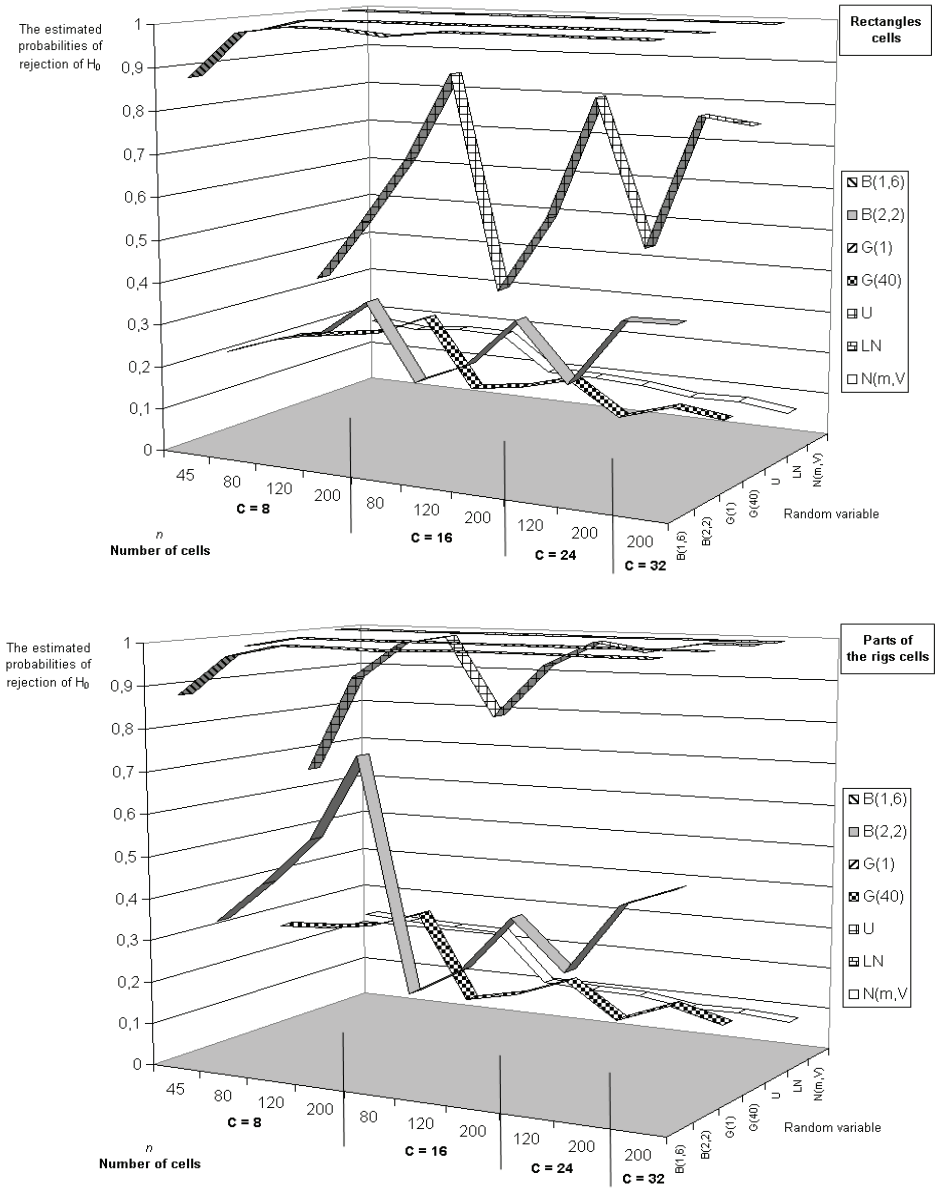

Figure 4. The estimated probabilities of rejection of H_0 for the rectangle cells (upper part) and the part of rings (bottom part) cells

Source: Table 3–6.

Table 7. The estimated probabilities of rejection of H_0 for rectangles cells (25 cells), parts of rings cells (24 cells) and multivariate Shapiro-Wilk's test (n = 200)

Distribution of random variable	Rectangles 6 zones – 36 cells	The parts of rings 8 rings – 32 cells	Shapiro-Wilk's test
N(**m**,**V**)	0.0610	0.0754	0.0969
LN	1.0000	1.0000	1.0000
G(1)	1.0000	1.0000	1.0000
G(40)	0.1238	0.1476	0.4154
B(1,6)	1.0000	1.0000	1.0000
B(2,2)	0.3866	0.4981	0.0629
U	0.7780	0.9999	0.9530

Source: Monte Carlo study.

The results of comparison the power of analyzed chi-square goodness-of-fit tests and multivariate extension of Shapiro-Wilk's test are presented in the figure 7.

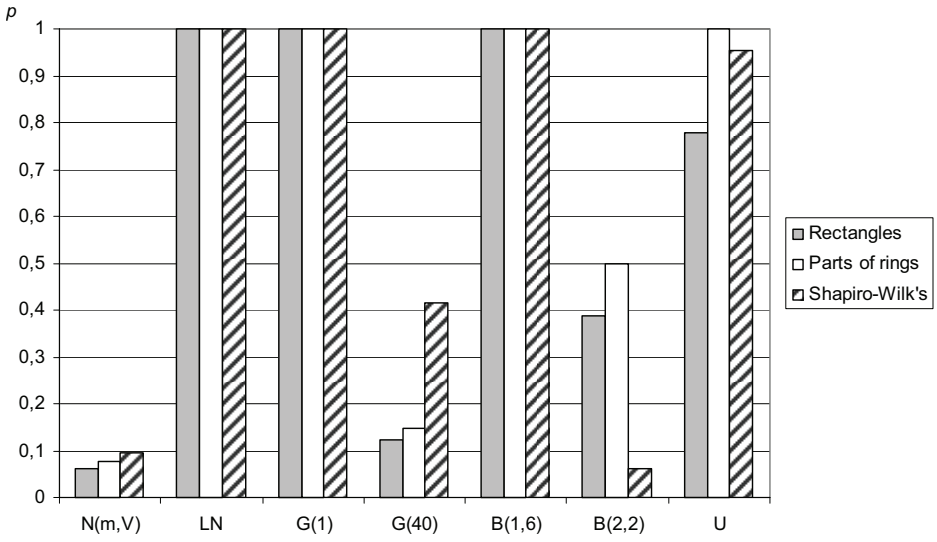

Figure 5. The estimated probabilities of rejection of H_0 for the rectangle cells, the part of rings cells and for the Shapiro-Wilk's test.

Source: table 7

We can notice that for analyzed samples the empirical power of the Shapiro-Wilk test is higher than the power of the chi-square goodness-of-fit tests for gamma distribution and for uniform distribution. In the case of beta distribution the power is higher in the chi-square goodness-of-fit test case (tables 7).

IV. CONCLUDING REMARKS

The generalization of the chi-square goodness of fit test into multivariate case was considered. Two sets of cells in the multivariate case were analyzed. The sets of rectangles cells and the set of the parts of the rings cells were considered. These two cases were compared using simulation study. The power of the chi-square goodness-of-fit test in these two cases were analyzed. The simulation study have shown that the chi-square goodness of fit test is most powerful when the parts of rings cells are used.

REFERENCES

Domański Cz. (1998), Własności testu wielowymiarowej normalności Shapiro-Wilka i jego zastosowanie. *Cracow University of Economics Rector's Lectures*, No. 37.

Royston P. (1982), Algorithm AS 181: *The W Test for Normality*. *Applied Statistics*, 31, 176–180.

Rubinstein R. Y., Kroese D. P. (2007) *Simulation and the Monte Carlo Method*, Wiley – Interscience. New Jersey.

Tallis G. M. (1963), *Elliptical and radial truncation in normal population*. *Annals of Mathematical Statistics* vol. 34, pp. 940–944.

Tallis G. M. (1965), *Plane truncation in normal population*. *Journal of the Royal Statistical Society* Series B, vol. 27 pp. 301–307.

Grzegorz Kończak, Janusz L. Wywiał

O MOCY TESTU CHI-KWADRAT DLA HIPOTEZY O NORMALNOŚCI ROZKŁADU WIELOWYMIAROWEGO

Ogólnie znany test zgodności chi-kwadrat jest wykorzystany do weryfikacji hipotezy o normalności rozkładu prawdopodobieństwa zmiennej losowej wielowymiarowej. Najczęściej cele testu konstruuje się w kształcie prostokątów. W artykule rozważono elipsoidy, których wspólny środek ma współrzędne wyznaczone przez oceny z próby wartości średnich zmiennych losowych. Analizę mocy testu przeprowadzono z wykorzystaniem symulacji komputerowej. Porównywano moc testu dla różnych liczebności próby oraz dla różnych od normalnego alternatywnych rozkładów prawdopodobieństwa. Przeprowadzono również porównanie z wielowymiarowym testem Shapiro-Wilka.