ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Grzegorz Kończak**

ON THE MODIFICATION OF THE EMPTY CELLS TEST

Abstract. In the paper the proposition of the nonparametric test to verify the hypothesis on the distribution of the random variable is presented. The proposed test is the modification of well known empty cells test. In the empty cells test the area of variability of the random variable is divided into some fixed cells. In the proposed modification the cell is moving over the whole area of variability of the random variable.

The analysis of testing the hypothesis of normality is presented. The table with critical values of the test statistic and the comparison of the empty cells test and the proposed modification is presented.

Key words: test, empty cells test, Monte Carlo.

I. INTRODUCTION

Among the other goodness-of-fit tests that are described and discussed in nonparametric statistic books is David's empty cells test (David F.N. 1950, Sheskin D. J., 2004). This test can be used to test the hypothesis of the distribution of random variable. The area of variability of random variable is divided into m cells and the number of elements in each cell is counted. Then the number of empty cells is determined. This number of empty cells is compared to the critical value. The proposition of the modification of the empty cells is presented in the paper. In the proposed modification the cell is moving over the whole area of variability of the random variable. The analysis in the case of verifying the hypothesis of normality is presented. The table with critical values of the test statistic and the comparison of the "empty cells" test and the proposed modification is presented. The Monte Carlo study for comparison properties of the classical form of the empty cells test and the proposed modification were made. The results of this simulation were presented.

*Ph.D., Department of Statistics, Katowice University of Economics, koncz@ae.katowice.pl.

II. THE EMPTY CELSS TEST

Let X be the continuous random variable and let F_0 be the distribution function of this random variable. Let $x_1, x_2, \ldots, x_n$ be an n-element simple sample. We will test the hypothesis that the sample is taken from the F_0 distribution. Let $\mathbf{S}$ denotes the area of variability of the random variable X. In the classical version of the empty cells test the area of variability $\mathbf{S}$ of the random variable X is divided into m cells $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m$ which fulfill conditions:

1. $\mathbf{S} = \bigcup_{i=1}^{m} \mathbf{S}_i$
2. $\mathbf{S}_i \cap \mathbf{S}_j = \varnothing$ for $i \neq j$
3. $P(x \in \mathbf{S}_i) = \frac{1}{m}$ for $i = 1, 2, \ldots, m$.

For each cell $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_m$ we determine the number of elements in the cell. The number of elements in the i-th cell we denote as m_i. Let K_n be the number of empty cells. The statistic K_n can be written as follows

$$K_n = card\{i : m_i = 0\} \tag{1}$$

where m_i is the number of elements in the i-th cell.

The probability function of number of empty cells is known and can be written as follows (Hellwig Z., 1965, Csorgo M. and Guttman I., 1962)

$$p(K_n = k) = \binom{m}{k} \sum_{r=0}^{m-k} (-1)^r \binom{m-k}{r} \cdot \left(\frac{m-k-r}{m} \right)^n \tag{2}$$

where $k = h, h+1, \ldots, m-1$ and $h = \max(0, m-n)$

The cumulative distribution function of the statistics K_n can be written as follows

$$P(k) = \sum_{s=0}^{k} \binom{m}{s} \sum_{r=0}^{m-s} (-1)^r \binom{m-s}{r} \cdot \left(\frac{m-s-r}{m} \right)^n \tag{3}$$

The statistic K_n can be used to test the hypothesis

$$\begin{aligned} H_0 &: F(x) = F_0(x) \\ H_1 &: F(x) \neq F_0(x) \end{aligned} \tag{4}$$

For the assumed significance level α the rejection region can be written as follows

$$Q = \{k : k \geq K_{n,\alpha}\} \tag{5}$$

Where $K_{n,\alpha}$ is taken from the tables (eg. Hellwig Z., 1965, Domański Cz., Pruska K. 2000).

III. THE MODIFICATION OF THE EMPTY CELLS TEST

In the classical form of the empty cells test the cells are fixed. Let us consider the case that the cells are not fixed. In the proposed modification there is one cell which is moving over the whole area **S** of variability of the random variable X. The probability that x_i ($i = 1, 2, \ldots, n$) is in the cell under H_0 is constant. The idea of the proposed modification is presented in the Fig. 1. There are $m = 4$ fixed cells (the classical form of the empty cells test) and the cell S_x (the modification of the empty cells test) which is moving over the set $[a, b]$.

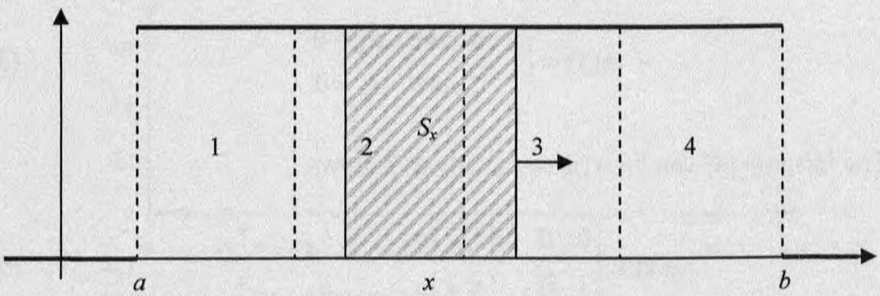

Fig. 1 The idea of the proposed modification (uniform random variable case)

Let us consider a set S^* of cells S_x which satisfy following two conditions:

1. $x \in \left[q_{\frac{p}{2}}; q_{1-\frac{p}{2}} \right]$

2. $S_x = \left[q_{\beta-\frac{p}{2}}; q_{\beta+\frac{p}{2}} \right]$

Where q_a is the quantile of order a of the random variable X, β denotes the order of the quantile of x $\left(\frac{p}{2} < \beta < 1 - \frac{p}{2} \right)$ and $0 < p < 1$.

We can notice that S_x is a cell in which x is a mid-point.

Let $x_1, x_2, \ldots, x_n$ be an i.i.d. sample. The hypothesis (4) will be tested. For each $x \in \left[q_{\frac{p}{2}}; q_{1-\frac{p}{2}} \right]$ under H_0 we have $P(x_i \in S_x) = p = const$ $(i = 1, 2, \ldots, n)$.

Therefore the probability that the cell S_x is empty can be written as follows:

$$P(card\{S_x\} = 0) = P\big((x_1 \notin Sx) \wedge (x_2 \notin S) \wedge \ldots \wedge (x_n \notin S)\big) \tag{6}$$

Under the assumption that $x_1, x_2, \ldots, x_n$ are independent it can be written as follows

$$\begin{aligned} P(card\{S_x\} = 0) &= P(x_1 \notin S_x) \cdot P(x_2 \notin S_x) \cdot \ldots \cdot P(x_n \notin S_x) = \\ &= (1-p) \cdot (1-p) \cdot \ldots \cdot (1-p) = (1-p)^n \end{aligned} \tag{7}$$

Let us consider the function $h: \left[q_{\frac{p}{2}}; q_{1-\frac{p}{2}} \right] \to \{0,1\}$ given as follows

$$h(x) = \begin{cases} 0 & \text{if } card \;\; S_x > 0 \\ 1 & \text{if } card \;\; S_x = 0 \end{cases} \tag{8}$$

The formula (8) can be written equally as follows

$$h(x) = \begin{cases} 0 & \text{if} \quad \underset{i}{\exists} x_i \in [q_{\beta_i - p/2}; q_{\beta_i + p/2}] \\ 1 & \text{if} \quad \underset{i}{\forall} x_i \notin [q_{\beta_i - p/2}; q_{\beta_i + p/2}] \end{cases} \tag{9}$$

where β_i $\left(\frac{p}{2} \le \beta_i \le 1 - \frac{p}{2} \right)$ is given as follows

$$\beta_i = \begin{cases} \dfrac{p}{2} & x_i < F_0^{-1}\left(\dfrac{p}{2}\right) \\ F_0(x_i) & x_i \in \left[F_0^{-1}\left(\dfrac{p}{2}\right); F_0^{-1}\left(1 - \dfrac{p}{2}\right) \right] \\ 1 - \dfrac{p}{2} & x_i > F_0^{-1}\left(1 - \dfrac{p}{2}\right) \end{cases} \quad \text{for } i = 1, 2, \ldots, n.$$

That's mean that the value $h(x)$ is equal to 1 if and only if the cell S_x is empty. The statistic K_n from the classical form of the empty cells test (1) can be rewritten as follows $K_n = \sum_{i=1}^{m} h(x_{(i)})$ where m is the number of cells and $x_{(i)}$ is the mid-point of the i-th cell. Therefore the proposed modification can be treated as a generalization of the classical form of the empty cells test.

The function $h(x)$ is equal to 1 if and only if the corresponding to x cell is empty, that's mean

$h(x) = 1 \Leftrightarrow card\{S_x\} = 0$. It can be written as follows

$P(h(x)=1) = P(card\{S_x\}=0) = (1-p)^n$ for each $x \in \left[q_{\frac{p}{2}}; q_{1-\frac{p}{2}} \right]$.

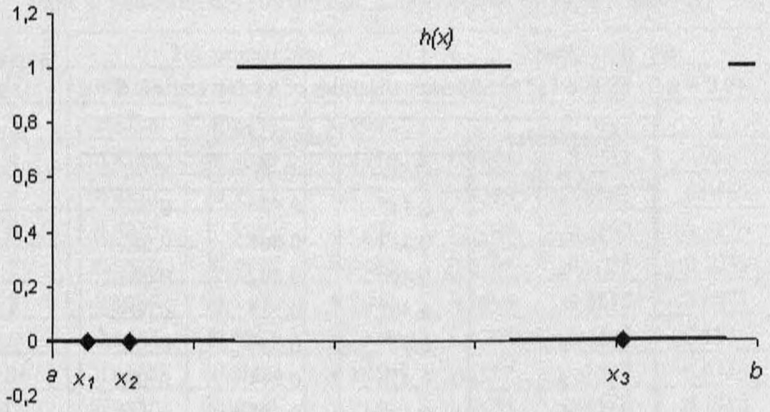

Fig 2. Function $h(x)$ for 3 element sample.

The idea of the function $h(x)$ is presented in the Fig. 2. The $n = 3$ element sample is taken. Function $h(x)$ is equal to 1 if and only if $x_i \notin S_x$ for $i = 1, 2, 3$.

To test the hypothesis (4) it can be used following statistic

$$T = \frac{1}{q_{1-p/2} - q_{p/2}} \int_{p/2}^{1-p/2} h(x)dx \qquad (10)$$

It can be notice that $0 \le T < 1$. The value of the statistic T represents the relative length of the empty cells area and is equal to the area under $h(x)$. We reject the hypothesis if $T \ge T_\alpha$.

IV. THE CASE OF NORMAL DISTRIBUTION

Let us assume that $X \sim N(\mu,\sigma)$ and x_1, x_2, ..., x_n is the n element i.i.d. sample and let $p = \frac{1}{n}$. To obtain the critical values for test the hypothesis that random variable X is normally distributed the Monte Carlo simulation were made. For sample sizes of n = 3, 4, ..., 15 there were found quantilies of the statistic T. They were found for the significance levels α 0.10, 0.05 and 0.01. There are following steps in computer simulations:

1. The values x_1, x_2, ..., x_n (n = 3, 4, ..., 15) were generated from normal distribution with mean μ = 100 and standard deviation σ = 5.
2. For each sample the value of the T statistic was calculated.
3. The steps 1-2 were repeated 10 000 times.
4. The empirical quantilies 0.90, 0.95 and 0.99 were accepted as estimates of quantiles of the statistic T.

Table 1. The estimates quantiles of the test statistic T

Sample size	Quantil $q(1-\alpha)$		
n	0.90	0.95	0.99
3	0.542	0.627	0.767
4	0.512	0.565	0.682
5	0.490	0.542	0.634
6	0.481	0.523	0.603
7	0.477	0.519	0.587
8	0.473	0.511	0.581
9	0.470	0.508	0.574
10	0.468	0.501	0.572
11	0.465	0.499	0.564
12	0.463	0.497	0.561
13	0.463	0.498	0.556
14	0.458	0.494	0.551
15	0.457	0.490	0.550

Source: Monte Carlo study

The results of Monte Carlo study are presented in table 1. For sample size from 3 to 15 there are presented estimates quantiles of the statistic T (10).

V. MONTE CARLO STUDY – COMPARISON OF THE EMPTY CELLS AND THE PROPOSED MODIFICATION

To compare the classical form of the empty cells test and the proposed modification the series of computer simulations was made. The samples of size n was taken from normal distribution N(105, 5). There were test the hypothesis $H_0 : F(x) = F_0(x)$ against $H_1 : F(x) = F_1(x)$, where $F_0(x)$ is the cumulative distribution function of the random variable $X \sim$ N(100,5) and $F_1(x)$ is the cumulative distribution function of the random variable $X \sim$ N(105,5).

For every n 10 000 samples were generated and for each sample the value of the statistic T was calculated. The critical values of the statistic T was taken from table 1. The estimates of probabilities of rejection the hypothesis H_0 are presented in table 2.

Table 2. The estimates probabilities of rejection H_0 hypothesis under H_1

Sample size	The proposition			Empty cells test		
	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	0.3498	0.2435	0.0992	0	0	0
4	0.4166	0.3089	0.1216	0.1552	0.1552	0
5	0.4812	0.3537	0.1595	0.4388	0.0563	0.0563
6	0.5213	0.3953	0.1851	0.2193	0.2193	0.0176
7	0.5536	0.4118	0.2030	0.4701	0.1035	0.1035
8	0.5768	0.4402	0.2048	0.2815	0.2815	0.0471
9	0.5953	0.4502	0.2135	0.4905	0.1544	0.1544
10	0.6151	0.4808	0.1958	0.3117	0.3117	0.0713
11	0.6270	0.4790	0.2140	0.4924	0.1750	0.1750
12	0.6460	0.4938	0.2138	0.3406	0.3406	0.0982
13	0.6590	0.4923	0.2247	0.5093	0.5093	0.2062
14	0.6864	0.5121	0.2315	0.6637	0.3560	0.1182
15	0.7054	0.5458	0.2302	0.5198	0.5198	0.2387

Source: The results of the Monte Carlo study.

As we can see it is impossible to reject the null hypothesis in classical form of empty cells test for n = 3 element sample (for α = 0.1. 0.05 and 0.01). The proposed modification can be used for small sample.

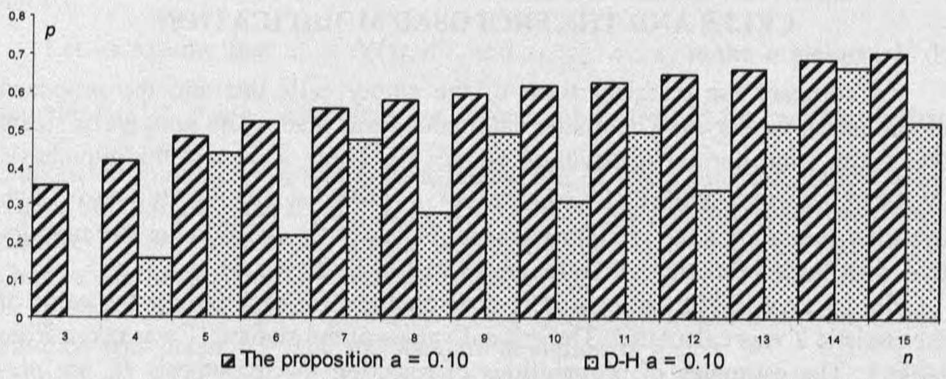

Fig. 3. The estimates probabilities of rejection H_0 hypothesis under H_1 ($\alpha = 0.10$)

The results for the significance level $\alpha = 0.10$ of the Monte Carlo study are presented in the Fig. 3. It can be noticed that use of the modification of the empty cells test leads more often to rejection the H_0 hypothesis (under H_1).

VI. CONCLUDING REMARKS

The proposed modification of the empty cells test can be used to test the hypothesis in statistical control quality procedures. It can be especially used in process monitoring using Shewhart's control chart to test the hypothesis of normality distribution in small sample cases.

The Monte Carlo study have been made. In the first part of the simulation the critical values of the proposed statistic have been derived. In the second part the comparison of the classical empty cells test and the proposed modification has been done. If the null hypothesis is false then the proposed modification more often leads to the rejection of the null hypothesis. The proposed modification of the empty cells is natural enhancement of the classical form of this test and is easy to use.

REFERENCES

Csorgo M., Guttman I. (1962) *On the Empty Cell Test*, Technometrics, vol. 4, No. 2, p. 235–247.

David F.N. (1950) *Order Statistics*. J. Wiley & Sons Inc.. New York.

Domański Cz. Pruska K. (2000) *Nieklasyczne metody statystyczne*. PWE Warszawa.

Hellwig Z. (1965) *Test zgodności dla małej próby*. Przegląd Statystyczny. 12. p. 99–112.

Sheskin D. J. (2004) *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall /CRC, Boca Raton.

Grzegorz Kończak

O PEWNEJ MODYFIKACJI TESTU PUSTYCH CEL

W artykule przedstawiono propozycję nieparametrycznego testu do weryfikacji hipotezy o postaci rozkładu badanej zmiennej. Proponowany test jest modyfikacją znanego testu pustych cel. W teście pustych cel obszar zmienności jest dzielony na ustalone cele i sprawdza się w ilu celach nie ma żadnego elementu z próby. W proponowanej modyfikacji położenie celi jest zmienne. Wyznaczana jest funkcja podająca czy dla danego położenia celi jest ona pusta, a następnie na podstawie przebiegu tej funkcji podejmowana jest decyzja odnośnie weryfikowanej hipotezy. Przedstawiono rozważania dla szczególnego przypadku gdy testowana jest hipoteza o normalności rozkładu. Wyznaczone zostały wartości krytyczne dla proponowanego testu oraz porównania tej metody z testem pustych cel. Proponowana modyfikacja została porównana z klasycznym testem pustych cel.