ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 269, 2012

Czesław Domański[*]

STATISTICAL TESTS BASED ON EMPTY CELLS

Abstract. In professional literature on statistics, a constantly growing interest is observed in non-parametric methods. Commonly these methods are based on counting, rank or position statistics and on a number or length of series. In this paper, the least popular tests, namely tests based on a number of empty cells, are presented. David-Hellwig test and a two-sample consistency test are considered. Empirical power of the tests is presented in comparison to classic tests: Kolmogorov and Shapio-Wilk test for testing normality of a distribution and t-Student's and Wilcoxon tests for testing consistency of two distributions.

Key words: nonparametric tests, tests empty cells, student t-tests, Wilcoxon test.

I. INTRODUCTORY REMARKS

In the statistical literature a constantly growing interest in the nonparametric methods has been observed. These methods are generally based on numerical statistics, rank statistics, numerical-rank statistics, order statistics, as well as the number and length of runs. The paper will present the test of goodness of fit for two samples based on the number of empty cells. The empirical power of the test compared with the classical tests: the Student t-test and the Wilcoxon test for the hypothesis on consistency of two distributions will be shown.

II. TEST FOR TWO SAMPLES BASED ON THE NUMBER OF EMPTY CELLS

Let two simple samples $X_1, X_2, ..., X_m$ be given and $Y_1, Y_2, ...Y_n$ drawn from a population which has distributions with continuous cumulative distribution function F and G.

The interval where

$$I_i = (X_{(i-1)}, X_{(i)}], \quad (i = 1, ..., m+1) \tag{1}$$
$$X_{(0)} = -\infty, \quad X_{(m+1)} = +\infty$$

will be called the i-th cell.

[*] Professor, Chair of Statistical Methods, University of Lodz.

Let us denote by $r_1,...,r_{m+1}$, the number of elements of the second sample which belong to subsequent cells , respectively.

The number of empty cells is denoted by S:

$$S = card\,\{\, i : r_i = 0\} \tag{2}$$

Wilks (1961) and Csörgö i Guttman (1962) stated that the distribution of probability of the number of empty cells, under the assumption that $F = G$, takes the form:

$$P(S = s) = \frac{\binom{m+1}{s}\binom{n+1}{m-s}}{\binom{m+n}{m}} \tag{3}$$

The expected value and the variance of the random variable S, representing the number of empty cells, are defined by the following formulae:

$$ES = \frac{m(m+1)}{n+m},$$

$$D^2S = \frac{m^2(m^2-1)}{(m+n)(m+n-1)} + \frac{m(m+1)}{m+n} - \frac{m^2(m+1)^2}{m+n} \tag{4}$$

According to the classical empty cell test the hypothesis

$$H_0 : F = G \tag{5}$$

is rejected, when $S \geq s_\alpha$, where s_α denotes the critical value connected with the selected significance level α

$$P(S \geq s_\alpha) \leq \alpha \quad i \quad P(S \geq s_\alpha - 1) > \alpha \tag{6}$$

We will now examine the properties of the test by comparing the test power with the parametric Student t-test and the non-parametric Wilcoxon test.

If F and G are cumulative distribution functions of the distribution with the same variance and the hypothesis (5) holds true, then the statistics

$$t = \frac{\overline{Y} - \overline{X}}{\sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}} \tag{7}$$

where $\overline{X}, \overline{Y}$ oraz S_x^2 i S_y^2 denote mean values and variances from the examined samples, has the Student distribution with $n + m - 2$ degrees of freedom.

In this case

$$G(x) = F(x + \Delta) \tag{8}$$

is therefore equivalent to

$$H_0 : \Delta = 0 \tag{9}$$

and the test based on the statistic (7) is most powerful.

Generally,

$$H_1 : \Delta > 0 \tag{10}$$

is accepted as an alternative hypothesis.

If *F* and *G* differ from cumulative distribution functions of the normal distribution then the equation (8) does not hold true (different variances), and the t-test does not have optimal properties so it makes sense to look for alternative solutions. The test based on the number of empty cells is one of the possible solutions.

In the paper we draw a comparison between the power of the empty cell test with the power of the *t*-test and the Wilcoxon test, which is characterized by a relatively high power in the class of non-parametric tests.

The Wilcoxon test against the alternative hypothesis (10) is carried out as follows:

a) We order all the values from the sample X_1, X_2,...,X_n and Y_1, Y_2,...,Y_m which constitute the sequence of the form $\{ Z_{(i)}, i = 1,2,...,n+m \}$, where

$$Z_{(1)} \langle Z_{(2)} \langle ... \langle Z_{(n+m)} \tag{11}$$

b) We assign to the elements of the sequence $\{Z_{(i)}\}$ subsequent natural numbers 1, 2,…, $n + m$, the so-called ranks.

c) We determine the value of the statistic which is a sum of ranks in the second sample

$$W = \sum_{i=1}^{n} w_i \tag{12}$$

where $w_1, w_2, \ldots, w_n$ denote ranks of elements coming from the second sample.

d) The hypothesis H_0 is rejected if $W \geq w_\alpha$, where w_α is the critical value ,

$$w_\alpha = \max \quad \{w : P_0(W \geq w) \leq \alpha\} \tag{13}$$

P_0 denotes the probability in case when the H_0 holds true.

Both the number of empty cells S, and the Wilcoxon statistic W are discrete statistics. Therefore, we use randomized tests in order to compare their powers (cf. Domański, C., Pruska K. (2000)).

The randomized empty cell test consists in rejecting H_0 if $s \geq s_\alpha$, with the probability

$$p_\alpha^S = \frac{\alpha - P_0(S \geq s_\alpha)}{P_0(S = s_\alpha - 1)}, \text{ if } S = s_\alpha - 1 \tag{14}$$

and accepting H_0 if $S \langle s_\alpha - 1$.

The size of the randomized test is obviously equal to α :

$$P_0(S \geq s_\alpha) + p_\alpha^s P_0(S = s_\alpha - 1) = \alpha \tag{15}$$

and its power equals:

$$1 - \beta^s = P_1(S \geq s_\alpha) + p_\alpha^s P_1(S = S_\alpha - 1) \tag{16}$$

where P_1 denotes probability in case when H_1 holds true.

The randomized Wilcoxon test is defined analogously and its power is equal to:

$$1 - \beta^W = P_1(W \geq w_\alpha) + p_\alpha P_1(W = w_\alpha - 1) \tag{17}$$

Some quantiles of the distribution of the empty cells number were given by Csörgö and Guttman (1962). In the present paper we make use of the calculated interpolated critical values (interpolated quantiles) of the form:

$$S_{\alpha}^{\text{int}} = s_{\alpha} + p_{\alpha}^{s} \tag{18}$$

which are given in Table 1.

Table 1. Interpolated quantiles of distribution of empty cells number for two samples

Sample size		Significance level	
m	*n*	$\alpha = 0.05$	$\alpha = 0.10$
6	6	5.62667	5.18667
12	12	8.99851	8.68016
18	18	12.59041	11.94178
24	24	15.89123	15.28761
30	30	19.24773	18.56055
8	4	7.85417	7.62500
16	8	13.74160	13.30014
24	12	19.47649	18.89374
32	16	25.01850	24.55238
40	20	30.74623	30.02811

Source: Author's own calculations.

III. THE MONTE CARLO EXPERIMENT

The evaluation of the power of the empty cell test for two samples is based on the results of the Monte Carlo experiment , which included:

- 2 values of $c = 1{,}3$,
- 7 values of $\Delta = 0, 0.5, 1, 2, 2.5, 3$,
- 5 values of $m + n = 12, 24, 36, 48, 60$,
- 2 values of the quotient $m/n = 1, 2$,
- 3 types of distributions of F: normal, exponential, double exponential.

We generated $q = 10\,000$ of samples having $(m + n)$ elements each, with

- m observations from distribution $F(x)$
- n observations from distribution $G(x) = F(cx + \Delta)$.

The obtained results are presented in Figures 1–9. For all the figures the horizontal axis represents values of Δ, in 0.0–3.0 interval. In Figures 1–3 the vertical axis represents the values of the power of the empty cell in the interval 0.0–1.0, while in Figures 4–9 it represents the evaluation of differences between the power of the t-test (Figures 4–6), the power of the Wilcoxon test (Figures 7–9) and the power of the empty cell test in the interval –1.0–1.0. Each of the graphs presents two lines. The continuous line represents power (differences in power) in case of $c = 1$ (equal variances), while the non –continuous line represents cases of different variances ($c = 3$).

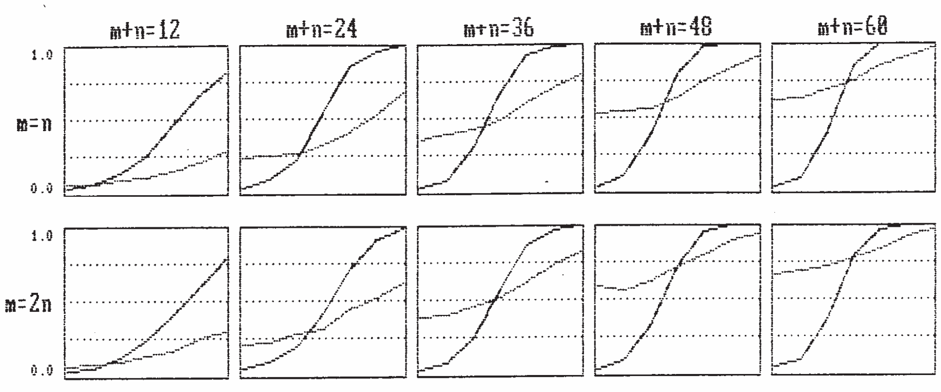

Fig. 1. Power of the empty cell test – normal distribution

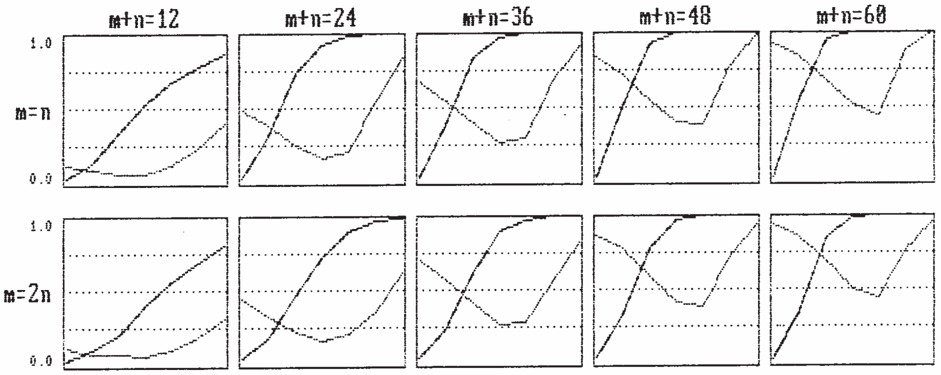

Fig. 2. Power of the empty cell test – exponential distribution

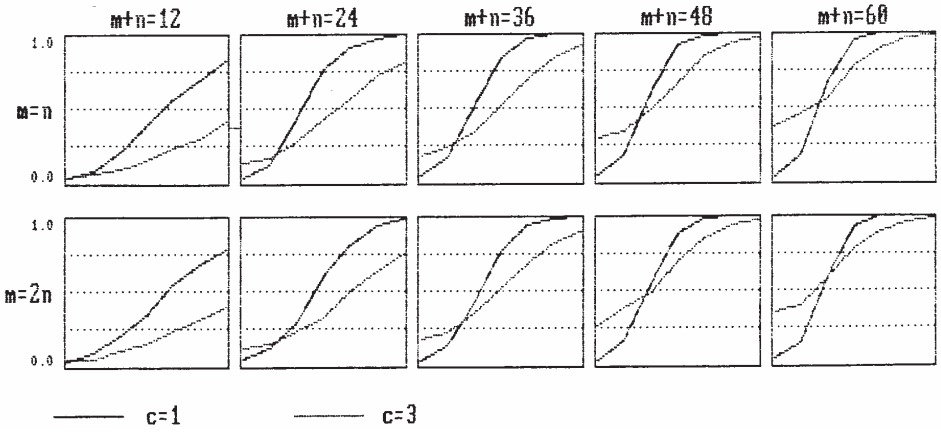

Fig. 3. Power of the empty cell test – double exponential distribution

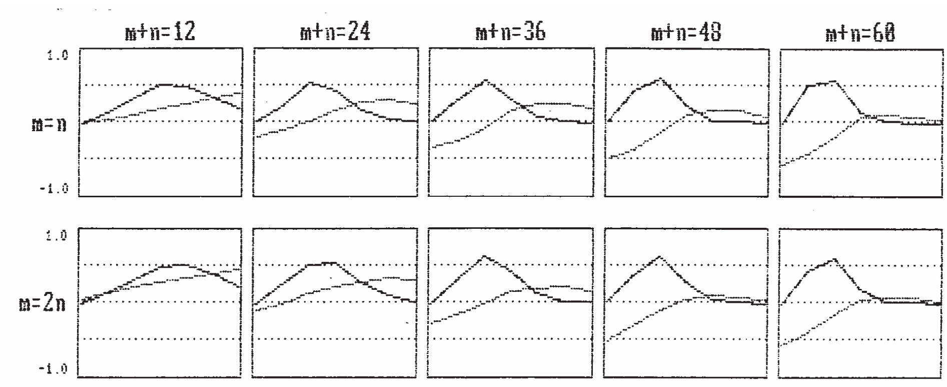

Fig. 4. Differences in power of the t-Student test and empty cell test – normal distribution

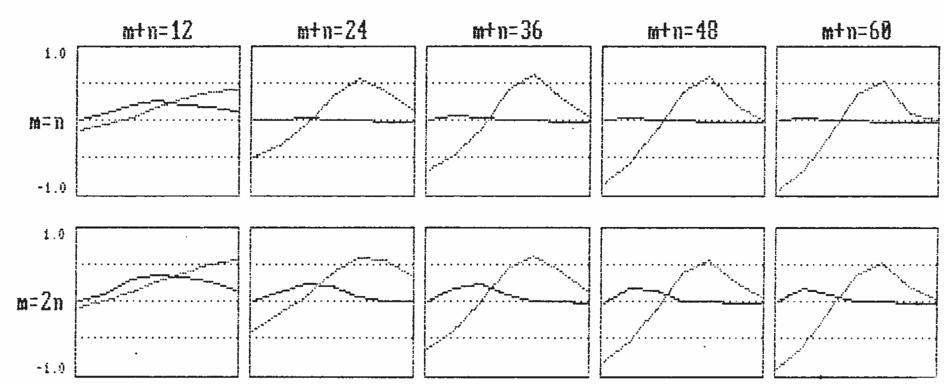

Fig. 5. Differences in power of the t-Student test and empty cell test – exponential distribution

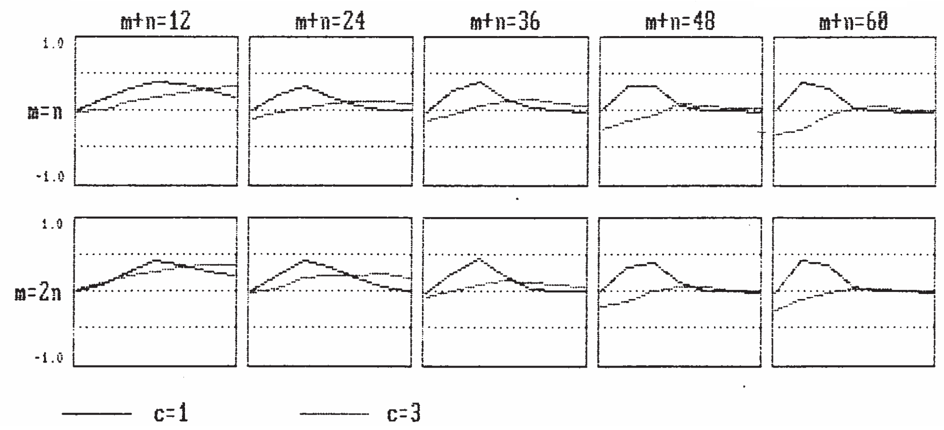

Fig 6. Differences in power of the t-Student test and empty cell test – double exponential distribution

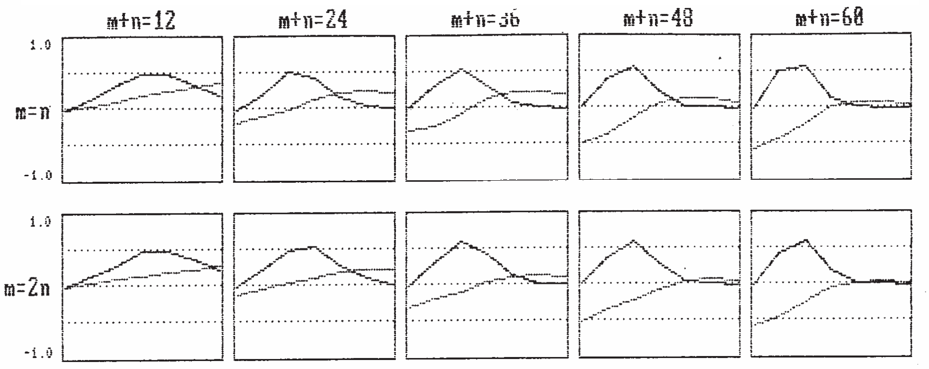

Fig. 7. Differences in power of the Wilcoxon test and empty cell test – normal distribution

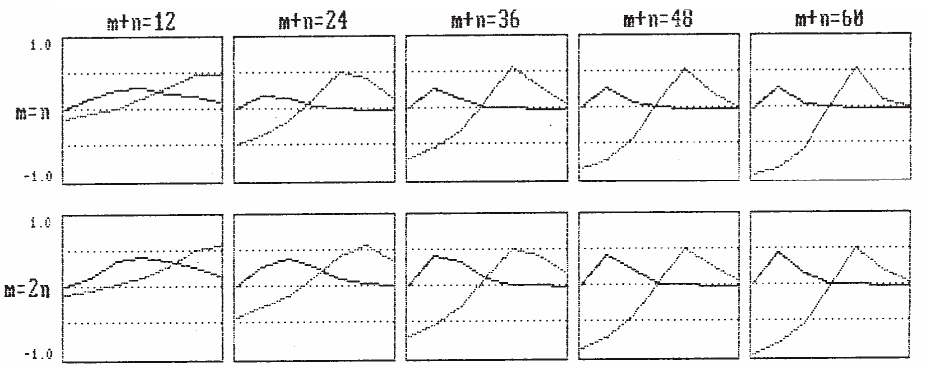

Fig. 8. Differences in power of the Wilcoxon test and empty cell test- exponential distribution

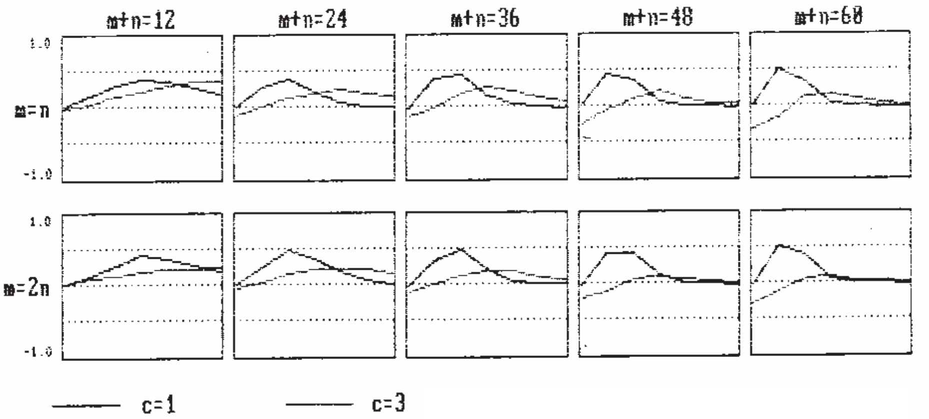

Fig 9. Differences in power of the Wilcoxon test and empty cell test – double exponential distribution

IV. CONCLUSIONS

For equal variances ($c = 1$) the Student t-test is more powerful than the empty cell test. This is the case when the difference in mean values $\Delta \geq 2.5$ and $m + n \geq 24$ then the power of the Student t-test exceeds 0.9 for all the examined distributions. This fact seems to support the thesis that the Student t-test is an optimal test for two samples with equally powerful variances.

In case of populations with different variances ($c = 3$) the power of the empty cell test exceeds the power of the Student t-test and the Wilcoxon test. The empty cell test proves to be useful in case when population variances differ significantly while differences in mean values are not substantial.

Obviously, the power of all the examined tests increases when the sample size $m + n$ increases.

The obtained results encompass a few special cases, however, they point to the usefulness of the empty cell test. In the light of the above considerations it seems that the properties of the empty cell tests as well as other nonparametric tests are worth further detailed research.

LITERATURE

Csörgö M., Guttman J. (1962), *On the empty cell test*, „Technometrics" 4, 235–247.

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

Willes S. S. (1961), *A combinatorial test for the problems of two samples from continuous distribution*. Proc. 4-th Berkeley Sympos. Math. Stat. and Probability, vol. 1, Berkley-Los Angeles, 707–717.

Czesław Domański

TESTY STATYSTYCZNE OPARTE NA PUSTYCH CELACH

W literaturze statystycznej obserwujemy stały wzrost zainteresowania metodami nieparametrycznymi. Metody te bazują na ogół na statystykach liczących rangowych, licząco-rangowych, pozycyjnych oraz na liczbie i długości serii. W pracy tej przedstawimy mniej znane testy oparte na liczbie pustych cel. Rozważany będzie test zgodności dla dwóch prób oparty na pustych celach. Przedstawiona zostanie empiryczna moc tego testu w porównaniu z testami: t-studenta i Wilcoxona dla hipotezy o zgodności dwóch rozkładów.