DE GRUYTER
OPEN

DOI: 10.1515/bile-2017-0010

Biometrical Letters
Vol. 54 (2017), No. 2, 175 - 186

Sample size determination in the Mann–Whitney test

Andrzej Kornacki, Andrzej Bochniak, Agnieszka Kubik-Komar

Department of Applied Mathematics and Computer –Science, University of Life Sciences in Lublin, Akademicka 15, 20-950 Lublin, Poland, e-mail: andrzej.kornacki@up.lublin.pl

SUMMARY

This paper discusses the problem of determining the number of observations necessary to apply the nonparametric Mann–Whitney test. We describe the method given by Noether (1987) for determining a sample size which guarantees that the Mann–Whitney test at a given significance level α has a predetermined power 1–*β*. The presented theory is tested by calculating the empirical power in computer simulations. The paper also raises the issue of the method of rounding the determined sample size to an even number when the sample is divided into two equinumerous subsamples.

Key words: Mann–Whitney test, sample size, empirical power.

1. Introduction and notation

In this paper, we will consider a test statistic whose distribution is asymptotically normal with mean $\mu(S)$ and standard deviation $\sigma(S)$. The mean and standard deviation of the statistic S, when the null hypothesis is true, will be denoted by $\mu_0(S)$ and $\sigma_0(S)$. For simplification, our discussion will concern a right-tailed test. Let Z be a random variable with a standard normal distribution, and let z_α be its right-tailed critical value, that is, a number such that $P(Z > z_\alpha) = \alpha$. Then the power of the test S against the alternative hypothesis H_a has the following form:

$$\begin{aligned} Power &= P\left[S > \mu_0(S) + z_\alpha \sigma_0(S) \middle| H_a\right] \\ &= P\left[\frac{S-\mu(S)}{\sigma(S)} > \frac{\mu_0(S)-\mu(S)+z_\alpha\sigma_0(S)}{\sigma(S)}\right] \\ &= P\left[Z > \frac{\mu_0(S)-\mu(S)}{\rho\sigma_0(S)} + \frac{z_\alpha}{\rho}\right], \text{ where } \rho = \frac{\sigma(S)}{\sigma_0(S)} \end{aligned} \tag{1}$$

As can be seen, the power of this test will be equal to 1–β when the expression on the right side of the equality sign in formula (1) is equal to $-z_\beta$. This can also be formulated as the following condition:

$$Q(S) = \left[\frac{\mu(S) - \mu_0(S)}{\sigma_0(S)}\right]^2 = (z_\alpha + \rho z_\beta)^2 \tag{2}$$

Obviously, the ρ value is generally unknown. But for alternatives that do not deviate too much from the null hypothesis, the assumption that $\sigma(S)$ is close to $\sigma_0(S)$ may often be appropriate. Equivalently, we may assume that ρ=1. Let us refer to $Q(S)$ as the noncentrality parameter of the test S. Then we will obtain an approximation to the sample size sought by comparing the noncentrality parameter $Q(S)$ with $(z_\alpha + z_\beta)^2$ and then solving the resulting equation for the number of observations.

The aim of this paper is to test by simulation the sample size proposed by Noether which guarantees the maintenance of the assumed power of the test, using the example of the Mann–Whitney U test. Attention is also drawn to the effect of the method of mathematical rounding of the derived actual number used to determine the size of two equinumerous samples.

2. The Mann–Whitney U test

We will now illustrate the above theory using the example of a nonparametric goodness-of-fit test for two samples. Given are two independent samples $X_1, X_2, \ldots X_m$ and $Y_1, Y_2, \ldots Y_n$. We want to test the hypothesis that both samples come from the same population, against an alternative hypothesis that they come from different populations. These hypotheses may also be formulated using the following probabilities:

$$\begin{cases} H_0 : P(Y > X) = P(Y < X) = \dfrac{1}{2} \\ H_a : P(Y > X) = p > \dfrac{1}{2} \end{cases} \tag{3}$$

We will use the Mann–Whitney U test in the following form as a test statistic to test the hypothesis H_0:

$$U = \#(Y_j > X_i)\,,\, i = 1,2,\ldots,m;\ j = 1,2,\ldots,n \tag{4}$$

where #(A) denotes the power of set A. It is known that $\mu(U) = mnp$ (Fisz 1967). Moreover, we have

$$\mu_0(U) = \frac{1}{2}mn \text{ and } \sigma_0^2(U) = \frac{mn(N+1)}{12} \tag{5}$$

(Fisz 1967) where N=m+n. Putting $m = cN$, we obtain

$$Q(U) = \frac{12c(1-c)N^2(p-\frac{1}{2})^2}{N+1}\,,$$

and hence by approximation:

$$N = \frac{(z_\alpha + z_\beta)^2}{12c(1-c)(p-\frac{1}{2})^2} \tag{6}$$

For samples with the same size $m = n$, we have $c = 1/2$, and the formula (6) takes the following form:

$$N^* = \frac{(z_\alpha + z_\beta)^2}{3(p-\frac{1}{2})^2} \tag{7}$$

3. Materials and methods

The first step in estimating the required sample size N* was to calculate the probability p = P(Y>X).

For this purpose, the properties of a two-dimensional normal distribution were used. If the random variable (X,Y) has a two-dimensional normal distribution $N(\mu, \Sigma)$, the density function is expressed by the following formula:

$$f(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y}\cdot\exp\left[-\frac{1}{2(1-\rho^2)}\cdot q(x,y)\right], \tag{8}$$

where

$$q(x,y)=\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2} \tag{9}$$

and ρ denotes the correlation between X and Y,

$$\mu=\begin{pmatrix}\mu_X\\ \mu_y\end{pmatrix},\ \Sigma=\begin{pmatrix}\sigma_X^2 & \rho\sigma_X\sigma_Y\\ \rho\sigma_X\sigma_Y & \sigma_Y^2\end{pmatrix} \tag{10}$$

where $-\infty < x < \infty$, $-\infty < y < \infty$, $\sigma_X > 0$, $\sigma_Y > 0$, $-1 \le \rho \le 1$, and the constants μ_X and μ_Y are arbitrary. In this study, independent samples were generated, and hence $\rho = 0$.

Finally

$$p = \mathrm{P(Y>X)} = \int_{-\infty}^{\infty}\int_{x}^{\infty} f(x,y)dydx. \tag{11}$$

After the value of p was calculated from the formula (11) and subsequently the value of N* was determined from the formula (7), for the determined distribution parameters two $N^*/2$-element samples were generated, and the Mann–Whitney test was used to test the hypothesis that they both came from the same population (3). Simulations were repeated 50,000 times, using three significance levels α (0.01, 0.05 and 0.1) and two levels of the power of the test $1-\beta$ (0.9 and 0.8).

Based on these 50,000 sample pairs, the empirical power of the Mann–Whitney test was determined through simulation by counting the number of cases when the hypothesis H_0 was rejected. To increase the calculation precision of this value, each determination of the empirical power was repeated 10 times, and the mean value based on those 10 repetitions was presented as the final result.

Samples in the simulations were randomly selected from a population with a normal distribution. The first sample, from the distribution N(0.1), was compared with the second sample from a population having a distribution with a standard deviation equal to 1 and with the mean determined by the effect size ES, ranging from 0.1 to 1.0. The value of ES measures the deviation of the alternative hypothesis from the null hypothesis. The effect size was defined by Cohen as "the degree to which the null hypothesis is false" (Cohen, 1992). In our case, this value is based on means and calculated according to the following formula (Cohen, 1988):

$$ES = \frac{\mu_2 - \mu_1}{\sigma} \tag{12}$$

Although the effect size may exceed a value of 1, Cohen referred to small, medium and large effect sizes, and in the case of formula (12) the values corresponding to these terms are 0.2, 0.5 and 0.8 respectively. Therefore, in our simulation test the value of ES does not exceed the value 1.

The simulations were performed using MathWorks MatLab 2014a software and our own code, as well as built-in procedures to generate random numbers and compute the nonparametric Mann–Whitney test.

4. Results

The graphs for the determined (averaged) empirical power are shown in Figure 1. The values on the Y axis were restricted to an interval close to the expected value of the power. The simulations show that the power of the test maintains its value for small effect values, but when the effect size increases, the test's power is slightly overestimated. The curves for different values of alpha and beta change similarly. Hence, only an increase in the parameter ES is responsible for an increase in the test's power relative to its set value.

The graphs for the determined size *N** are shown in Figure 2. Large variation between the curves determined by different choices of α can be observed only for small values of ES. It decreases with increasing effect size.

To test how much the sample size *N** can be reduced while maintaining the assumed power of the test, an additional simulation was carried out. The value of *N** was reduced by 1 until two consecutive values were below the assumed value 1–β. The graph shows the mean number, determined based on 10 repetitions, by

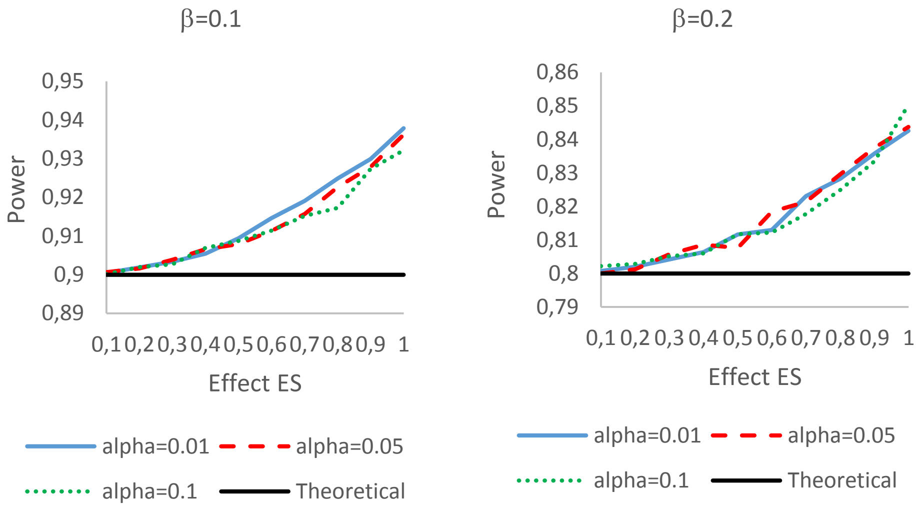

Figure 1. Empirical power for the determined N^* for different levels of α, β depending on effect size

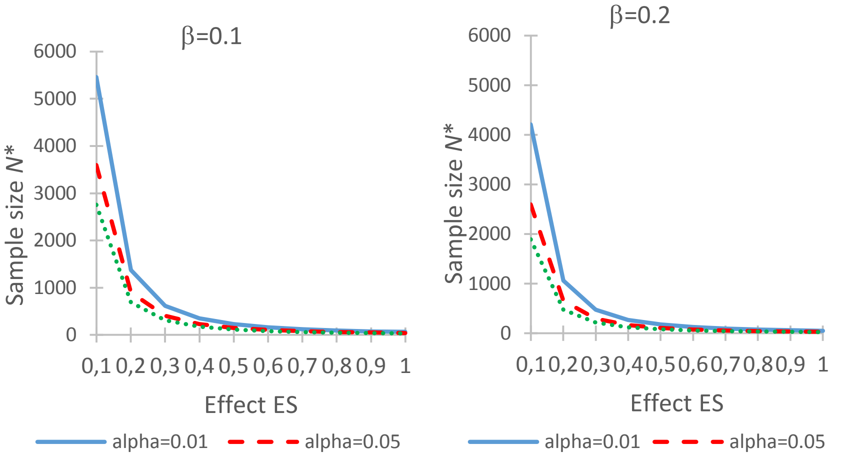

Figure 2. Size N^* for different levels of α, β depending on effect size

which the number N^* can be reduced (Figure 3). The mean value by which N^* was reduced is denoted as N_r, and it was determined as the mean difference between N^* and the sample size that maintains the assumed level of the test's power depending on effect size.

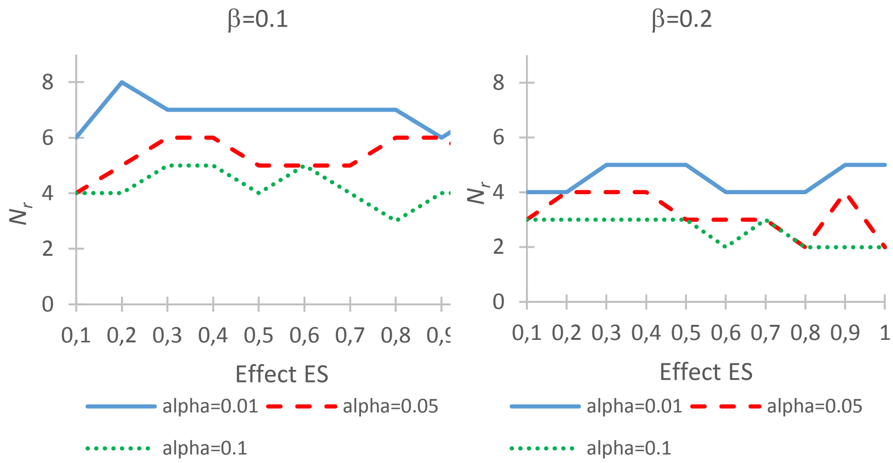

Figure 3. Difference N_r between N^* and the sample size maintaining the assumed level of the test's power for different values of α, β depending on effect size

It can be seen on the graphs that N^* is overestimated by about 3–7 elements in the sample, regardless of effect size and thus regardless of the sample size N^*, and these differences increased with decreasing values of α and β.

Taking into account the large variation in the value of N^* for different effects, this difference is shown in relative values (Figure 4).

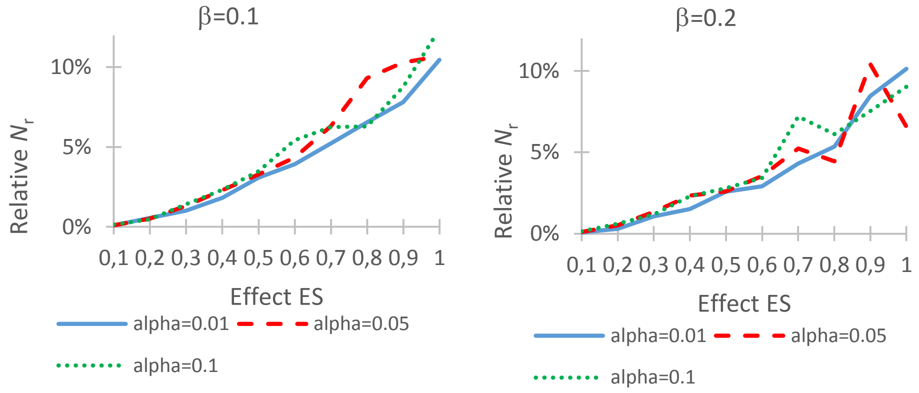

Figure 4. Relative difference N_r between N^* and the sample size maintaining the assumed level of the test's power for different values of α, β depending on effect size

It can be noted that for small values of ES this difference is insignificant, but for large values of the effect size, due to the small number of observations required to detect it, this difference amounts to 10–12% depending on the set power of the test.

After the value of N_r reducing the sample size N^* was determined, it was tested whether the newly determined size $N^{**} = N^* - N_r$ maintains the assumed level of the test's power. The value of N_r was different depending on the assumed levels of significance α and β (Table 1).

Table 1. Mean values of N_r depending on the values of the parameters α and β

	β	
α	0.1	0.2
0.01	6.9	4.5
0.05	5.3	3.2
0.1	4.2	2.6

An additional problem was the method of rounding the value of N^{**}. The value of N^* determined according to the formula (7) is an actual number and the total size of both samples. In our work, we wanted to give the most effective size N^{**} being an even natural number. To this end, we tested eight methods of rounding the N^{**} to an integer value, as defined in Table 2.

Table 2. Rounding options in sample size determination

Denotation	Rounding method
P1	round(N^*/2) – floor(N_r/2);
P2	ceil(N^*/2) – floor(N_r/2);
P3	round(N^*/2) – round(N_r/2);
P4	ceil(N^*/2) – round (N_r/2);
P5	ceil((N^*– N_r)/2);
P6	round((N^* – N_r)/2);
P7	round((N^* – N_r + 1)/2);
P8	round((N^* – N_r + 2)/2);

The empirical power of the test for the size *N*** derived by rounding methods P1–P8 is shown on the graphs below (Figures 5 and 6).

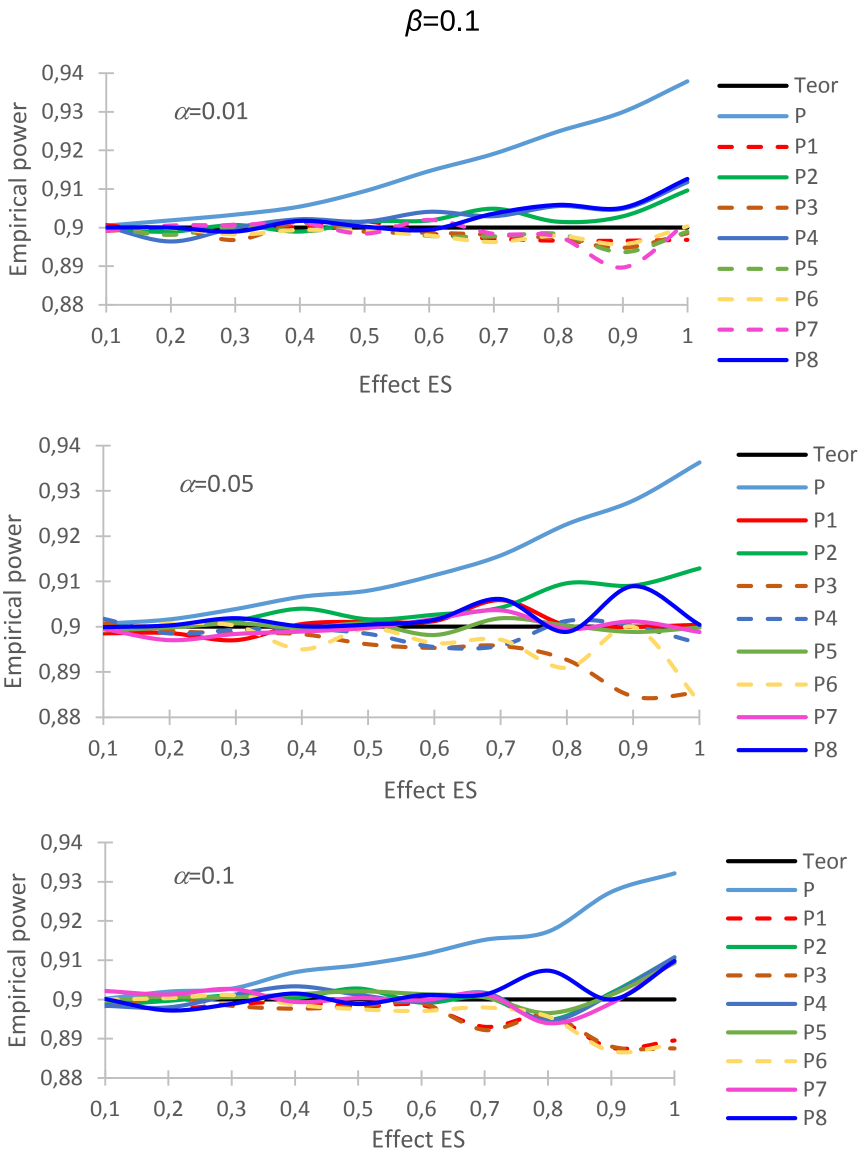

Figure 5. Empirical power of the test for the size *N*** and the value β=0.1

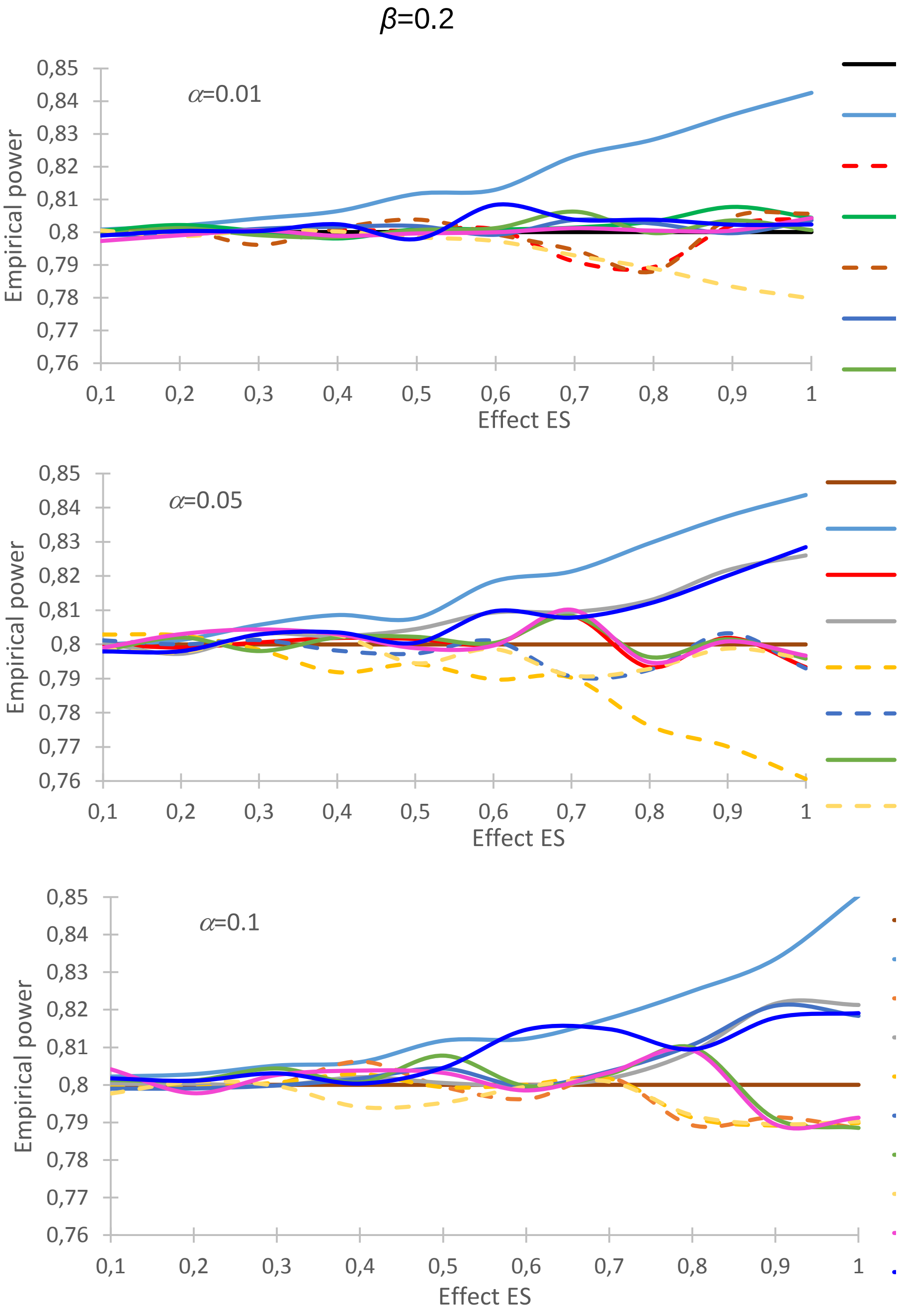

Figure 6. Empirical power of the test for the size *N*** and the value β=0.2

The P value shown in the graphs is the value of the power for the calculated value of N^* (without reduction in observations and rounding), and *Teor* represents theoretical values of the power test. Rounding methods that certainly do not meet the assumptions are marked with dashed lines.

The above graphs show that it is impossible to indicate unambiguously a best method for rounding the size regardless of the assumed level of significance and the test's power. Given that we observe the largest changes in the power of N^{**} for higher values of ES, in selecting the best method our intention was that in these areas the plot should be as close as possible to the assumed value of the test's power and, at the same time, it should not be below this value. It seems that P2, P4 and P6 can be indicated as the most effective rounding methods for β=0.1 and α=0.01, while for α=0.05 the best choice is P2, and for α=0.1 it is P8. For β=0.2 and α=0.01 and α=0.05, the most optimal choices would be the methods P2 and P8, while for α=0.1 they would be P2 and P4.

As can be seen, method P2 appears almost everywhere among the indicated methods for determining N^{**}, apart from the case in which the values of both parameters were 0.1, where P8 was indicated as the most efficient method for obtaining the desired number of observations.

5. Conclusions

The obtained simulation results demonstrate that the method given by Noether is an effective method for determining sample size in the Mann–Whitney test. The determined empirical power of the test for the obtained sample size is closest to the theoretical power for small effect sizes, whereas for large effects the empirical power is far greater. This suggests that it is possible to reduce slightly the determined sample size, which can be important in the case of expensive experiments.

The determined value of N^* is an actual value. Due to the fact that it is the total sample size of both samples, with the assumption that they should be equinumerous, we expect that it will be an even natural number. Therefore, it is necessary to round this value. The attempt undertaken in this study to reduce the size and obtain an even value for the determined N^* suggests that in almost all cases the best rounding method was to round up the value ($N^*/2$) and to subtract half of the N_r value, rounded down. The rounding method expressed by the formula ROUND(($N^* - N_r + 2$)/2) proved to be more effective only in the case of 90% theoretical power and a significance level of 0.1.

REFERENCES

Cohen, J. (1988) Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992) A power primer. Psychological Bulletin, Vol 112(1), 155–159.

Fisz M. (1967) Rachunek Prawdopodobieństwa i Statystyka Matematyczna. PWN, Warsaw.

Hamedani, G. G.; Tata, M. N. (1975) On the determination of the bivariate normal distribution from distributions of linear combinations of the variables. The American Mathematical Monthly, 82 (9): 913–915.

Noether G. E. (1987) Sample Size Determination for Some Common Nonparametric Tests. Journal of the American Statistical Association 82: 645–647.