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 131, 1993

*Czesław Domański**, *Andrzej S. Tomaszewicz**

CORRELATION BETWEEN TESTS BASED ON LENGTH OF RUNS

1. INTRODUCTION

Statistical theory based on number-of-runs distributions belongs undoubtedly to the best known applications of the run theory. Nevertheless, many valuable tests applying length-of-run distributions can be constructed, especially tests based on

- the maximal length of runs on one side of median,
- the smaller one from the maximum lengths of runs above and below the median,
- the larger one among the maximum lengths of runs above and below the median.

These tests can be applied in verification of hypotheses about independence of the sequence of observations, in determination of the trend in uni- or multivariate time series, in verification of the hypothesis that a regression model with one or more independent variables is linear etc.

Numerous applications of tests based on the length of runs are limited by the lack of appropriately precise tables to be employed in the construction of critical regions of the tests which would contain, simultaneously, information about the first type error probabilities. For the same reason it is difficult to construct randomized tests.

* Professors at the Institute of Econometrics and Statistics, University of Łódź.

The aim of this work is the analysis of relationships between the three mentioned variants of length-of-run tests. Our results were achieved due to the construction of tables for length-of-run distribution, using recursive formulae.

2. RECURSIVE FORMULAE

Consider a sequences of n independent realizations of a variable with binary symmetrical distribution

$$P(A) = P(B) = \frac{1}{2}.$$

Let us define the statistic

$$M(n, s, t, u) \tag{1}$$

which is the number of n-element samples composed of elements A, B such that

- the maximal length of runs of element B is t;
- the number of elements A in the last run is u;
- the maximal length of runs of elements A, excluding the last run is s.

Let us assume that function M is defined for all quadruples of non-negative integer arguments. Of course, M is equal to 0 for all arguments not fulfilling the condition

$$s + t + u \leq n. \tag{2}$$

For $u > 0$ (for samples ending with element A) we have the identity

$$M(n, s, t, u) = M(n - 1, s, t, u - 1). \tag{3}$$

If $u = 0$ then, changing elements A into B and vice versa, we obtain

$$M(n, s, t, u) = \sum_{v,w} M(n, v, s, w) \tag{4}$$

where the sum is taken for all these pairs (v, w) for which $w \geq 1$, as the last element is A, and

$$\max \{v, w\} = t, \tag{5}$$

since the maximum length of runs of elements A is t. Thus, the formula (4) can be written in the following form

$$M(n, s, t, 0) = \sum_{v=0}^{t-1} M(n, v, s, t) + \sum_{w=1}^{t} M(n, t, s, w) \tag{6}$$

(the last term of the first sum is $t - 1$, since writing t we would count the element $M(n, t, s, t)$) twice.

Let

$$R(n, s, t, u) = \frac{1}{2^n} M(n, s, t, u) \tag{7}$$

denote the probability of observing the triple (s, t, u). Then the recursive formula

$$R(n, s, t, u) = \frac{1}{2} R(n - 1, s, t, u - 1) \quad \text{for} \quad u \geq 1,$$

$$R(n, s, t, 0) = \sum_{v=0}^{t-1} R(n, v, s, t) + \sum_{w=1}^{t} R(n, t, s, w) \tag{8}$$

holds with the initial conditions

$$R(1, 0, 1, 0) = R(1, 0, 0, 1) = \frac{1}{2},$$

$$R(1, s, t, u) = 0$$

for the remaining triples (s, t, u).

Let us define, for the fixed n, the following random variables

S_A - maximal length of run composed of elements A,

S_B - maximal length of run composed of elements B,

$S_L = \min \{S_A, S_B\}$,

$S_U = \max \{S_A, S_B\}$.

Using these symbols, the easiest way to write the joint probability of distribution function for variables S_A and S_B is

$$P(S_A = s, S_B = t) = R(n, s, t, 0) + R(n, t, s, 0) \tag{9}$$

(the first component represents the probability under the condition that the last element is B, and the second one - that the sample ends with A).

3. BIVARIATE LENGTH-OF-RUN DISTRIBUTIONS

The object of our analysis are four bivariate distributions

$$(S_A, S_B), \quad (S_A, S_L), \quad (S_A, S_U), \quad (S_L, S_U) \tag{10}$$

Due to symmetry, the distributions of (S_B, S_L) and (S_B, S_U) are the same distributions as (S_A, S_L) and (S_A, S_U), respectively

Let F_{AB} be the cumulative distribution function of (S_A, S_B) defined as follows

$$F_{AB}(s, t) = P(S_A \leq s, S_B \leq t) \tag{11}$$

Analogously, we define cumulative distribution functions F_{AL}, F_{AU}, F_{LU} for remaining variables (S_A, S_L), (S_A, S_U), (S_L, S_U). Note that F_{AB} is symmetrical:

$$F_{AB}(s, t) = F_{AB}(t, s) \tag{12}$$

but the others do not possess this property.

Moreover, let us accept, the symbols F_A, F_L, F_U for the marginal distribution functions

$$\begin{aligned} F_A(s) &= P(S_A \leq s) = P(S_B \leq s), \\ F_L(s) &= P(S_L \leq s), \\ F_U(s) &= P(S_U \leq s). \end{aligned} \tag{13}$$

As

$$\begin{aligned} F_A(s) &= F_{AB}(s, n) = F_{AB}(n, s), \\ F_U(s) &= P(\max\{S_A, S_B\} \leq s) = P(S_A \leq s, S_B \leq s), \end{aligned}$$

hence

$$F_U(s) = F_{AB}(s, s)$$

thus,

$$F_L(s) = 2F_{AB}(s, n) - F_{AB}(s, s) = 2F_A(s) - F_U(s).$$

The distribution function F_{AU} can be expressed as follows

$$F_{AU}(s, t) = P(S_A \leq s, S_U \leq t) = P(S_A \leq s, S_A \leq t, S_B \leq t),$$

hence

$$F_{AU}(s, t) = \begin{cases} F_{AB}(s, t) & \text{for } s \leq t \\ F_{AB}(t, t) = F_U(t) & \text{for } s \geq t. \end{cases} \tag{14}$$

Similarly we can find F_{AL} and F_{AU}:

$$F_{AL}(s, t) = \begin{cases} F_A(s) & \text{for } s \leq t \\ F_{AB}(s, t) + F_A(t) - F_U(t) & \text{for } s \geq t. \end{cases} \tag{15}$$

$$F_{LU}(s, t) = \begin{cases} 2F_A(s, t) - F_U(s) & \text{for } s \leq t \\ F_U(s) & \text{for } s \geq t. \end{cases} \tag{16}$$

4. RANDOMIZED TESTS

Consider three marginal distributions, i.e. S_A, S_L and S_U statistics. Let S describe one of them and let F be its cumulative distribution function:

$$F(s) = P(S \leq s).$$

The left-hand and right-hand critical values - the integer quantiles - are defined as follows:

$$s_\alpha^L = \max \{s: P(S \leq s) \leq \alpha\} \tag{16}$$

$$s_\alpha^R = \min \{s: P(S \geq s) \leq \alpha\} \tag{17}$$

The randomized left-hand test is based on the following procedure:

if $S \leq s_\alpha^L$ then mull hypothesis should be rejected,

if $S \geq s_\alpha^L + 1$ then null hypothesis should be accepted,

if $S = s_\alpha^L + 1$ then null hypothesis should be rejected with the probability

$$r_\alpha^L = \frac{\alpha - P(S \leq s_\alpha^L)}{P(S = s_\alpha^L + 1)} . \tag{18}$$

An analogous rule is applied in case of the right-hand randomized test:

if $S \geq s_\alpha^R$ then null hypothesis should be rejected,

if $S \leq s_\alpha^R - 1$ then null hypothesis should be accepted,

if $S = s_\alpha^R - 1$ then null hypothesis should be rejected with the probability

$$r_\alpha^R = \frac{\alpha - P(S \geq s_\alpha^R)}{P(S = s_\alpha^R - 1)} . \tag{19}$$

The size of the above defined randomized test is equal to the chosen significance level α.

5. CORRELATION BETWEEN TESTS BASED ON LENGTH OF RUN

We shall now deal with the problem of correlation between the three tests based on S_A, S_L and S_U statistics.

We reduce our analysis to randomized tests. Comparison of tests with different first type error may always arise doubts.

Let us take into account two from among the analyzed statistics and call them S_1 and S_2. Let F be its joint cumulative distribution function. Let us assume, moreover, that we are verifying a hypothesis using randomized test based on these statistics at significance level α. To be more specific, let us assume that we apply left-hand sided tests which correspond to critical values $s_1 = s^L_{1\alpha}$ and $s_2 = s^L_{2\alpha}$ and randomization probabilities r_1 and r_2. Let R_1 and R_2 denote the events consisting in rejecting the null hypothesis by means of these tests. The cumulative binary distribution function may be characterized by the coefficient

$$\varkappa = P(R_1, R_2) - P(R_1)P(R_2). \tag{20}$$

In our opinion, this measure (being covariance) is more intuitive than other frequently used measures of relationship between binary distributions (see e.g. S i e g e l 1956). Nevertheless, to compare correlations between tests at different significance levels it is better to apply the common coefficient

$$\rho = \frac{\varkappa}{\alpha(1 - \alpha)}. \tag{21}$$

Simple calculations lead to the formula

$$\varkappa = (1 - r_1)\,((1 - r_2)F(s_1, s_2) + r_2F(s_1, s_2 + 1)) + \\ + r_1((1 - r_2)F(s_1 + 1, s_2) + r_2F(s_1 + 1, s_2 + 1)) - \alpha^2. \tag{22}$$

Analogously, we can derive the coefficient for right-hand tests:

$$\varkappa = r_1(r_2F(s_1 - 2, s_2 - 2) + (1 - r_2)F(s_1 - 2, s_2 - 1)) + \\ + (1 - r_1)\,(r_2F(s_1 - 1, s_2 - 2) + (1 - r_2)F(s_1 - 1, \\ s_2 - 1) - (1 - \alpha)^2. \tag{23}$$

Values ρ calculated for $n = 6, 7, \ldots, 120$, $\alpha = 0.01$, 0.05, 0.10 for all four analyzed bivariate distributions and for left- and right-hand test are presented on Figures 1-4.

The obtained results allow to formulate the following conclusions concerning correlations between the analyzed tests under null hypothesis when the sample does not exceed 120 observations.

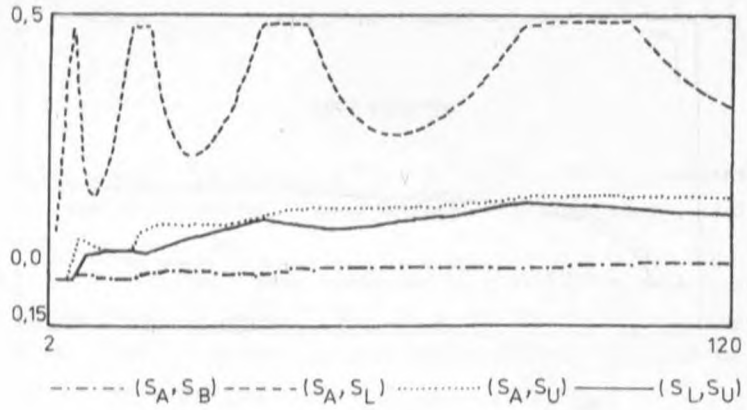

Fig. 1. Correlation coefficient rho as a function of sample size n, left-hand sided tests, $\alpha = 0.05$

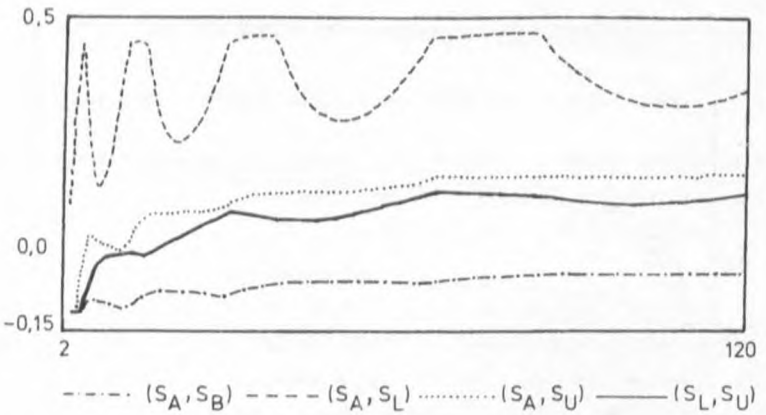

Fig. 2. Correlation coefficient rho as a function of sample size n, left-hand sided tests, $\alpha = 0,10$

1. Tests based on S_A and S_B statistics are weakly correlated. The left-hand tests are characterized by a negative correlation, while for the right-hand tests (at least for large n) the correlation is positive.

2. Correlation between tests based on S_L and S_U statistics is similar to that between left-hand S_A and S_L tests and S_A and S_U right-hand ones. These correlation are positive and quite strong (at least for larger sample sizes).

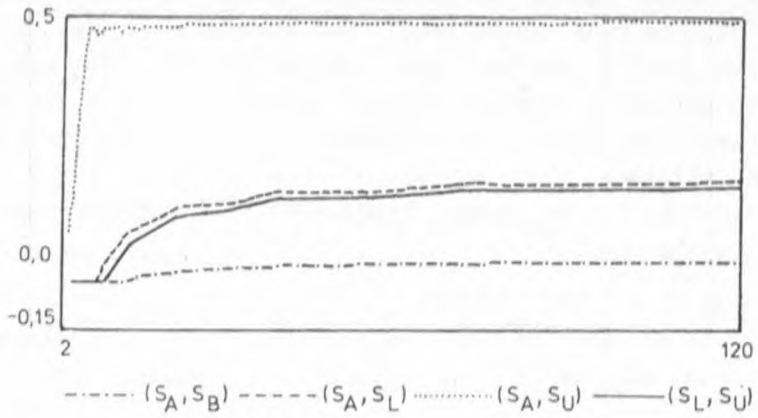

Fig. 3. Correlation coefficient rho as a function of sample size n, right-hand sided tests, $\alpha = 0,05$

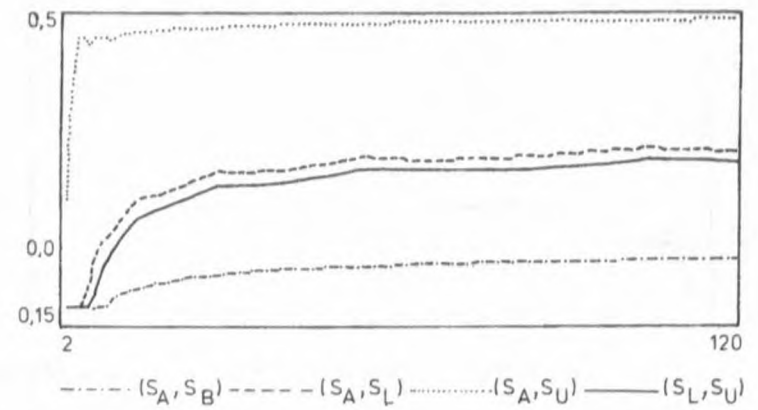

Fig. 4. Correlation coefficient rho as a function of sample size n, right-hand sided tests, $\alpha = 0,10$

3. We can observe a very strong positive correlation between left-hand sided S_A and S_L test similar to that between S_A and S_U tests.

Since the results of tests based on maximum run length on one side of the median (S_A) are strongly correlated with the test based on the smaller from maximum run lengths above and below the median (S_L), and larger from the lengths (S_U), it is sufficient in practice to apply one of them, provided we suppose that the actual

distribution differs very much from the distribution under the null hypothesis.

REFERENCES

D o m a ń s k i C., T o m a s z e w i c z A. S. (1984), *Recursive Formulae for Runs Distribution*, "Acta Universitatis Lodziensis", Folia oeconomica, No. 34, p. 19-28.

O m s t e d P. S. (1958), *Runs Determined in a Sample by an Arbitrary Cut*, "Bell System Technical Journal", No. 37, p. 55-58.

O w e n D. B. (1962), *Handbook of Statistical Tables*, Adolison-Wesley Publishing Co. Inc., Reading.

S i e g e l S. (1956), *Nonparametric Statistic for the Behavioral Sciences*, McGraw-Hill Book Company, Inc., New York.

Czesław Domański, *Andrzej S. Tomaszewicz*

ZWIĄZKI POMIĘDZY TESTAMI OPARTYMI NA DŁUGOŚCI SERII

Teoria serii daje się wykorzystać przy badaniu różnych testów statystycznych, służących na przykład do weryfikacji hipotez o liniowej postaci funkcji regresji, o określonej postaci funkcji trendu lub o losowości próby.

Na uwagę zasługują testy oparte na:

- maksymalnej długości serii po jednej stronie mediany,
- mniejszej z maksymalnych długości serii poniżej i powyżej mediany,
- większej z maksymalnych długości serii poniżej i powyżej mediany.

W artykule analizowane są wzajemne związki pomiędzy wymienionymi testami. Osiągnięte rezultaty prowadzą do wniosku, że testy oparte na maksymalnej długości serii po jednej stronie mediany są ściśle skorelowane z testami opartymi na mniejszej z maksymalnych długości poniżej i powyżej mediany oraz z testami opartymi na większej z tych długości. Wobec tego, w praktyce wystarczy zastosować jeden z nich, przy założeniu, że rzeczywisty rozkład różni się istotnie od rozkładu hipotetycznego.