ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 34, 1984

Czesław Domański[*], *Andrzej Tomaszewicz*[*]

RECURSIVE FORMULAE FOR RUNS DISTRIBUTIONS

1. INTRODUCTION

In the statistical literature combinatorial formulae for probabilities connected with runs distribution [2, 3] have been presented. However, these formulae are not suitable for numerical calculations. Much more efficient appeared to be the recursive formulae, especially in the case when the calculations are made for subsequent values of n.

The presented recursive formulae refer to runs length distribution, number of runs and joint probability distributions and runs length distributions. We shall discuss the case when subsequent observations in a sample are generated by a stationary Markov chain at two states denoted traditionally as A and B and transition matrix

$$\begin{bmatrix} p_{AA} & p_{AB} \\ p_{BA} & p_{BB} \end{bmatrix} = \begin{bmatrix} 1 - q_o & q_o \\ q_1 & 1 - q_1 \end{bmatrix}.$$

Let $P_{n,\Theta}$ be a distribution of this chain for each

$$\Theta \in \Theta = \{(q_o, q_1) : 0 < q_o < 1,\ 0 < q_1 < 1\}$$

[*] Lecturer, Institute of Econometrics and Statistics, University of Łódź.

and let $\Omega_n = \{A,B\}^n$ be a set of all n-element sequences formed of elements A, B. Thus, we shall consider the probability spaces

$$M_{n,\Theta} = (\Omega_n, 2^{\Omega_n}, P_{n,\Theta}), \quad \text{for} \quad \Theta \in \Theta. \tag{1}$$

2. RECURSIVE FORMULA FOR A THREE-DIMENSIONAL RUNS DISTRIBUTION

We assign to each sequence

$$\omega = (x_1, x_2, \ldots, x_n) \in \Omega_n,$$

the following numbers:

$N_A(\omega)$ - number of elements A in sequence ω,

$L_A(\omega)$ - number of runs formed of elements A,

$L(\omega)$ - total number of runs,

$S_A(\omega)$, $S_B(\omega)$ - maximum lengths of runs formed of elements A and B, respectively,

$S_D(\omega) = \min\{S_A, S_B\}$, $S_G(\omega) = \max\{S_A, S_B\}$,

$K_A(\omega)$, $K_B(\omega)$ - number of elements A and B, respectively, placed at the end of sequence ω,

$Z_A(\omega)$, $Z_B(\omega)$ - maximum lengths of runs consisting of elements A and B, respectively, without taking into account the last runs.

These notions are pretty obvious. To avoid, however, the possible ambiguity, we are presenting some examples:

	n	N_A	L_A	L	S_A	S_B	K_A	K_B	Z_A	Z_B
AAAAA	5	5	1	1	5	0	5	0	0	0
AABBB	5	2	1	2	2	3	0	3	2	0
ABAAA	5	4	2	3	3	1	3	0	1	1
AABBBABABB	5	4	3	6	2	3	0	2	3	2
ABBAABBBBB	10	3	2	4	2	5	0	5	2	2
BABABABABA	10	5	5	10	1	1	1	0	1	2

Assume that sequences $\omega \in \Omega_n$ are the realizations of the stationary Markov chain $\{X_1, X_2, \ldots, X_n\}$ with a transition matrix

$$\begin{bmatrix} p_{AA} & p_{BA} \\ p_{AB} & p_{BB} \end{bmatrix},$$

where $0 < p_{AB} < 1$, $0 < p_{BA} < 1$. Therefore, stationary probabilities are given by the formulae

$$p_A = P(X_j = A) = \frac{p_{AB}}{p_{AB} + p_{BA}},$$

(2)

$$p_B = P(X_j = B) = \frac{p_{BA}}{p_{AB} + p_{BA}},$$

for $j = 1, 2, \ldots, n$.

Under the above assumptions the probability distribution on set Ω_n can be presented using the formula

$$(3) \qquad P(\omega) = \frac{1}{p_{AB} + p_{BA}} p_{AA}^{n_A - l_A} p_{AB}^{l_A} p_{BA}^{l - l_A} p_{BB}^{n - n_A - l + l_A},$$

where $n_A = N_A(\omega)$, $l = L(\omega)$, $l_A = L_A(\omega)$ were assumed. We have

$$(4) \qquad P(\omega) = P(X_1 = x_1)\, P(X_2 = x_2 | X_1 = x_1) \ldots P(X_n = x_n | X_{n-1} = x_{n-1})$$

and

$$(5) \qquad P(X_1 = x_1) = \begin{cases} p_A, & \text{if } x_1 = A, \\ p_B, & \text{if } x_1 = B. \end{cases}$$

Because l_A is the number of these A's which form new runs, i.e. they follow B (except, maybe, the first element), hence at the right-hand side of (4) there is l_A of factors equal p_{BA} (taking also into account factor (5) in the form (1), where $X_1 = A$). The number of elements A which do not form new runs and therefore following A, is $(n_A - l_A)$, hence, there is the same number of factors p_{AA} at the right-hand side (4). Similarly we can prove that the numbers of factors p_{BB} and p_{BA} are $(n_B - l_B)$ and l_B, respectively (taking also into account factor (5) in the form (2), when $X_1 = B$). Both, when $X_1 = A$ and $X_1 = B$, at the right-hand side there is one factor:

$$\frac{1}{p_{AB} + p_{BA}}.$$

Consider, for a given n, joint three-dimensional distribution

$$(L, S_A, S_B) \tag{6}$$

of the runs number L, maximum length of runs consisting of elements A and maximum length of runs consisting of elements B.

Denote

$$M(n,l,s,t,u) = \operatorname{card}\{\omega \in \Omega_n : l = L(\omega),\ s = Z_A(\omega),\ t = S_B(\omega),\ u = K_A(\omega)\}. \tag{7}$$

The following formulae hold [1]:

$$M(n,l,s,t,u) = \begin{cases} M(n-1,l,s,t,u-1), & \text{for } u > 1, \\ M(n-1,l-1,s,t,0), & \text{for } u = 1, \\ \sum_{v=0}^{t-1} M(n,l,v,s,t) + \sum_{w=1}^{t} M(n,l,t,s,w), & \text{for } u = 0. \end{cases} \tag{8}$$

The first two equalities are obvious. They can be obtained by adding the n-th element A to the (n-1)-element sequence. In the case of u = 0, by changing elements A for B and vice versa, we obtain

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

where summation is extended to these pairs (v,w) for which max (v,w) = t.

Initial conditions for formula (8) have the form

$$M(1,l,s,t,u) = \begin{cases} 1 & \text{when } l = u = 1,\ s = t = 0, \\ 0 & \text{in other cases.} \end{cases} \tag{9}$$

Now, consider the probabilities

$$R_0(n,l,s,t,u) = P(L = l,\ Z_A = s,\ S_B = t,\ K_A = u)$$

and

$$R_1(n,l,s,t,u) = P(L = l,\ Z_B = s,\ S_A = t,\ K_B = u).$$

Of course, when the distribution is symmetrical, i.e. $p_{AB} = p_{BA}$ then probabilities R_0 and R_1 are equal. We shall go on using the more suitable notation

$$q_0 = p_{AB} \quad \text{and} \quad q_1 = p_{BA}.$$

By adding the n-th element to (n-1)-element sequence we obtain for h = 0, 1 and n > 1.

$$R_h(n,l,s,t,u) = R_h(n-1,l,s,t,u-1)(1-q_h), \tag{10}$$

when u > 1 and

$$R_h(n,l,s,t,1) = R_n(n-1,l-1,s,t,0)\, q_{1-h}. \tag{11}$$

For n = 1 we have

$$(12)\qquad R_h(1,1,s,t,u) = \begin{cases} \dfrac{1-q_h}{q_o+q_1} & \text{for } l = u = 1,\ s = t = 0, \\ 0 & \text{in other cases.} \end{cases}$$

If $u = 0$, then by replacing elements A by B and vice versa, we obtain

$$(13)\qquad R_h(n,l,s,t,u) = \sum_{v=0}^{t-l} R_{1-h}(n,l,v,s,t) + \sum_{w=1}^{t} R_{1-h}(n,l,t,s,w).$$

Formula (13) can be transformed in such a way that instead of R_h there are four-argument functions

$$(14)\qquad Q_h(n,l,s,t) = R_h(n,l,s,t,0)$$

for $h = 0,1$. From (10) and (11) it follows that

$$(15)\qquad R_h(n,l,s,t,u) = R_h(n-u,l-1,s,t,0)q_{1-h}(1-q_h)^{u-1}$$

for $h = 0, 1$ and $u < n$. If, however, $u = n$, then from (12) we have

$$(16)\qquad R_h(n,l,s,t,u) = \begin{cases} \dfrac{1}{q_o+q_1}(1-q_h)^n & \text{for } l = 1,\ s = t = 0, \\ 0 & \text{in other cases.} \end{cases}$$

Thus if we take

$$(17)\qquad Q_h(0,l,s,t) = \begin{cases} \dfrac{1-q_h}{(q_o+q_1)q_{1-h}} & \text{for } l = s = t = 0, \\ 0 & \text{in other cases,} \end{cases}$$

$h = 0, 1$, then, instead of (15) and (16) we can write

$$(18)\qquad R_h(n,l,s,t,u) = Q_h(n-u,l-1,s,t)q_{1-h}(1-q_h)^{u-1}.$$

Therefore, on the basis of (14)

$$(19)\quad Q_n(n,l,s,t) = \sum_{v=0}^{t-1} Q_{1-h}(n-t,l-1,v,s)q_h(1-q_{1-h})^{t-1} + \sum_{w=1}^{t} Q_{1-h}(n-w,l-1,s,w)q_h(1-q_{1-h})^{w-1}.$$

Formula (19) under initial conditions (17) is the basis for the efficient algorithm of determining functions Q_0 and Q_1.

From the obvious equality

$$P(L = l, S_A = s, S_B = t) = P(L = l, S_A = s, S_B = t, X_n = A) + P(L = l, S_A = s, S_B = t, X_n = B)$$

we obtain finally

$$(20)\quad P(L = l, S_A = s, S_B = t) = Q_0(n,l,s,t) + Q_1(n,l,t,s).$$

Hence we proved:

T h e o r e m 1. Joint distribution of random variables (L, S_A, S_B) determined on probabilistic space $M_{n,\Theta}$ is given by formulae (17), (19) and (20).

3. RECURSIVE FORMULAE FOR TWO AND ONE-DIMENSIONAL RUNS DISTRIBUTIONS

The obtained recursive formula (eqs. (17), (19) and (20), allows us theoretically to determine the function of joint probability distribution (L, S_A, S_B), and thus numerical analysis of dependences between statistics L, S_A, S_B, S_D and S_G.

Now we shall give recursive formulae resulting from Theorem 1, for probabilities of two-dimensional distribution (S_A, S_B) (Theorem 2) and one-dimensional distributions S_A, S_B, S_G (Theorems 3 and 4). Distribution S_D can be obtained from the dependence

$$P(S_D \leqslant s) = P(S_A \leqslant s) + P(S_B \leqslant s) - P(S_G \leqslant s).$$

Proofs for these theorems, as of little interest, are omitted. In all cases it is sufficient to sum up both sides of each relation (17), (19) and (20). It is also possible to prove them directly, similarly (but in a less complicated way) as proof to Theorem 1.

T h e o r e m 2. Joint distribution of random variables S_A, S_B determined on $M_{n,\Theta}$ can be presented using the recursive formula

$$P(S_A = s, S_B = t) = Q_0^{AB}(n,s,t) + Q_1^{AB}(n,s,t), \tag{21}$$

where for $h = 0, 1$

$$Q_h^{AB}(n,s,t) = \sum_{v=0}^{t-1} Q_{1-h}^{AB}(n-t,v,s) q_h (1 - q_{1-h})^{t-1} + \tag{22}$$

$$+ \sum_{w=1}^{t} Q_{1-h}^{AB}(n-w,s,t) q_h (1 - q_{1-h})^{w-1},$$

under initial conditions

$$Q_h(0,s,t) = \begin{cases} \dfrac{1}{q_0 + q_1} & \text{for } s = t = 0, \\ 0 & \text{in other cases.} \end{cases} \tag{23}$$

T h e o r e m 3. The distribution of variable S_A determined on $M_{n,\Theta}$ is expressed by the recursive formula:

$$P(S_A = s) = Q_0^A(n,s) + Q_1^A(n,s), \tag{24}$$

where for $h = 0, 1$.

$$Q_1^A(n,s) = \sum_{v=0}^{s-1} Q_0^A(n-s,v) q_1 (1 - q_0)^{s-1} + \tag{25}$$

$$+ \sum_{w=1}^{s} Q_o^A(n-w,s)q_1(1 - q_o)^{w-1},$$

under initial conditions

$$(26) \qquad Q_o^A(0,0) = Q_1^A(0,0) = \frac{1}{q_o + q_1}.$$

Replacing A by B and vice versa, 0 by 1 and vice versa, we shall obtain a formula for the distribution of S_B.

T h e o r e m 4. The distribution of variable S_G determined on $M_{n,\theta}$ is expressed by the recursive formula:

$$(27) \qquad P(S_G = s) = Q_o^G(n,s) + Q_1^G(n,s),$$

$$(28) \qquad Q_h^G(n,s) = \sum_{v=0}^{s-1} Q_{1-h}^G(n-s,v)q_h(1 - q_{1-h})^{s-1} +$$

$$+ \sum_{w=1}^{s} Q_{1-h}^G(n-w,s)q_h(1 - q_{1-h})^{w-1},$$

under initial conditions

$$(29) \qquad Q_o^G(0,0) = Q_1^G(0,0) = \frac{1}{q_o + q_1}.$$

T h e o r e m 5. The distribution of variable L determined on $M_{n,\theta}$ is expressed by the recursive formula:

$$(30) \qquad P(L = 1) = Q_o^L(n,1) + Q_1^L(n,1),$$

where for $h = 0, 1,$

$$(31) \qquad Q_h^L(n,1) = Q_h^L(n-1,1)(1 - q_{1-h}) + Q_{1-h}^L(n - 1, s - 1)q_h$$

under initial conditions

$$(32) \qquad Q_0^L(0,0) = Q_0^L(0,0) = \frac{1}{q_0 + q_1}.$$

The distributions of one-dimensional random variables given in Theorems 3, 4 and 5, can be a basis for the analysis of test powers based on these distributions. Theorem 2 can be a basis for analysing the dependence between the tests being considered.

REFERENCES

[1] Domański Cz., Tomaszewicz A. (1978): *Rozkłady długości serii i ich własności*, Łódź, mimeo.

[2] Moode A. (1940): *The Distribution Theory of Runs*, Ann. of Math. Statist., 11, p. 367-392.

[3] Olmsted P. (1958): *Runs Determined in a Sample by an Arbitrary Cut*, Bell System Techn. Journ., 37, p. 55-58.

Czesław Domański, Andrzej Tomaszewicz

WZORY REKURENCYJNE DLA ROZKŁADÓW SERII

Rozważmy przestrzeń prób generowanych przez stacjonarny łańcuch Markowa o dwóch stanach A, B. Na tej przestrzeni można określić trójwymiarową zmienną losową (L, S_A, S_B), gdzie L oznacza liczbę serii, S_A, S_B - maksymalną długość serii złożonych z elementów odpowiednio A, B. W pracy podane są wzory rekurencyjne dla funkcji rozkładu prawdopodobieństwa zmiennej (L, S_A, S_B), a także rozkładów (S_A, S_B), S_A, S_B, $\max(S_A, S_B)$ i L.

Prezentowane wzory są łatwe do zaprogramowania i przez to mogą być z powodzeniem wykorzystane do obliczeń numerycznych związanych z badaniem niektórych własności (między innymi mocy i odporności) testów serii.