The Annals of Statistics
1973, Vol. 1, No. 5, 933-939

THE JOINT PROBABILITY GENERATING FUNCTION FOR RUN-LENGTHS IN REGENERATIVE BINARY MARKOV CHAINS, WITH APPLICATIONS[1]

BY I. J. GOOD

Virginia Polytechnic Institute and State University

Gontcharov obtained the joint probability generating function for the numbers of runs of all lengths, both of successes and failures, in a Bernoulli sequence. This is here generalized to a class of regenerative binary Markov processes. For an allied class of Markov processes, the probability generating function is obtained for a "total score" defined in terms of runs of successes only, and asymptotic formulas are derived for the expectation and variance of the score.

1. Bernoulli and other runs. Gontcharov (1943, 1944) obtained numerous properties of the distribution of the numbers of runs of various lengths in a sequence of Bernoulli trials. All his results depended on his showing first that the joint or multivariate PGF (probability generating function) for runs of 1's and 0's, of various lengths, is

$$\frac{(1 + \sum p^r x_r)(1 + \sum q^s y_s)}{1 - \sum p^r x_r \sum q^s y_s} \tag{1}$$

where the summations are for $r, s = 1, 2, \cdots$; the probability of a "success" (a 1) on one trial is $p = 1 - q$, and the coefficient of $\prod x_r^{m_r} y_s^{n_s}$ is the joint probability of m_r runs of successes of length r, and n_s runs of failures of length $s (r, s = 1, 2, \cdots)$ when the Bernoulli sequence is known to be of length $\sum r(m_r + n_r)$. The constant term in (1) is 1 and corresponds to a sequence of length zero. In the terminology of Good (1961), (1) is a "universal" PGF in the sense that it covers sequences of all lengths simultaneously. My purpose is to generalize (1) to some Markov chains, "regenerative" in a sense to be defined, and to draw some deductions. Any application of the results for Bernoulli trials, such as to quality control, is also a potential application for the more general models of the present work.

We consider two classes of chains. The first class is regenerative at the end of *each run* of "successes" and of "failures"; the second class is regenerative after *each failure*. The asymptotic moments of an arbitrary "score", or linear function of the numbers of runs of each length, can be deduced from the universal PGF for both classes. But for the sake of simplicity this deduction is carried out explicitly only for the mean and variance of the score for the second class of

Received February 1972; revised January 1973.

Key words and phrases. Regenerative Markov chains, binary Markov chains, runs in Markov chains.

[1] This work was supported in part by H.E.W. grant No. 1 R01 GM18770-01.

chains, where here the score is defined in terms of the frequencies of runs of successes alone.

2. The PGF for a regenerative chain. We first generalize the joint PGF (1) to a regenerative binary Markov chain defined in the following manner. The conditional probability of a run of r 1's, following the end of a run of zeros, is κ_r, and the conditional probability of a run of s 0's, following the end of a run of 1's is $\lambda_s(r, s = 1, 2, 3, \cdots)$. This defines a process having two regenerative conditions, namely the right-hand ends of runs of 1's and 0's. (Perhaps the process should be called "biregenerative". Note that the process is *not necessarily an ordinary binary Markov chain of finite order.*) We shall consider a segment of such a process, of length L, and we shall refer to this segment as *the sequence.* We shall first find a formula for the probability that the sequence, assumed to be ergodic and in the stationary part of the process, will contain m_r runs of 1's of length r, and n_s runs of 0's of length $s(r, s = 1, 2, 3, \cdots)$. The formula is basically a multinomial expression, but with modifications related to the beginning and end of the sequence. If the sequence begins (or ends) with precisely a 1's or 0's we shall count this as a run of length a although its probability is not equal to κ_a or λ_a.

Let us begin at the beginning. The probability that the first element of the sequence, which we regard as a randomly selected trial, being a success is equal to $\bar{\kappa}/(\bar{\kappa} + \bar{\lambda})$, where $\bar{\kappa} = \sum r\kappa_r$ is the average length of a success run in an infinite chain, and similarly for $\bar{\lambda}$. Therefore the probability that the first element is *within* a success run of length r (in the *infinite* chain) is

$$\pi_r = \frac{r\kappa_r}{\sum r\kappa_r} \cdot \frac{\bar{\kappa}}{\bar{\kappa} + \bar{\lambda}} = \frac{r\kappa_r}{\bar{\kappa} + \bar{\lambda}}.$$

Therefore the probability that the sequence starts with exactly r consecutive 1's (which we are going to count as a run of length r) is

$$\sum_{\mu=r}^{\infty} \frac{\pi_\mu}{\mu} = \frac{K_r}{\bar{\kappa} + \bar{\lambda}}, \tag{2}$$

and similarly for s 0's, it is

$$\frac{\Lambda_s}{\bar{\kappa} + \bar{\lambda}}, \tag{3}$$

where

$$K_r = \sum_{\mu=r}^{\infty} \kappa_r, \qquad \Lambda_s = \sum_{\nu=s}^{\infty} \lambda_\nu. \tag{4}$$

The end of the sequence is easier to deal with. If a run of 0's ends on the $(L - r)$th element of the sequence, where $r < L$, then the probability that the sequence ends with a run of r 1's is

$$K_r, \tag{5}$$

with an obvious modification corresponding to the interchange of 0's and 1's.

By using (2) and (5) we see that the probability that the sequence begins with a run of a 1's, ends with b 1's, and contains m_r runs of 1's of length r, and n_s runs of 0's of length $s(r, s = 1, 2, \cdots; \sum r(m_r + n_r) = L)$, is

$$\frac{1}{\bar{\kappa} + \bar{\lambda}} \mathscr{C}(\mathbf{x}^{\mathbf{m}}\mathbf{y}^{\mathbf{n}})(\sum \kappa_r x_r)^{M-2} K_a x_a K_b x_b (\sum \lambda_r y_r)^{M-1} \tag{6}$$

where $a \geqq 1$, $b \geqq 1$, $a + b < L$, $\mathbf{x}^{\mathbf{m}} = x_1^{m_1} x_2^{m_2} \cdots$, $\mathbf{y}^{\mathbf{n}} = y_1^{n_1} y_2^{n_2} \cdots$, $\mathscr{C}(\mathbf{x}^{\mathbf{m}}\mathbf{y}^{\mathbf{n}})$ means "the coefficient of $\mathbf{x}^{\mathbf{m}}\mathbf{y}^{\mathbf{n}}$ in", $M = \sum m_r$, and the summations run from $r = 1$ to ∞. If a and b are unspecified, the probability is therefore

$$\frac{1}{\bar{\kappa} + \bar{\lambda}} \mathscr{C}(\mathbf{x}^{\mathbf{m}}\mathbf{y}^{\mathbf{n}})(\sum \kappa_r x_r)^{M-2} (\sum K_a x_a)^2 (\sum \lambda_r y_r)^{M-1} \tag{7}$$

where we can allow the middle summations to run from $a = 1$ to ∞ if $M > 1$. We shall deduce that if the initial and final elements of the sequence are unspecified the PGF is

$$1 + \frac{1}{\bar{\kappa} + \bar{\lambda}} \left\{ K^* + \Lambda^* + \frac{K^2 Y + \Lambda^2 X + 2K\Lambda}{1 - XY} \right\} \tag{8}$$

where

$$\begin{aligned} X &= \sum \kappa_r x_r , & Y &= \sum \lambda_r y_r \\ K &= \sum K_r x_r , & \Lambda &= \sum \Lambda_r y_r \\ K^* &= \sum (K_r + K_{r+1} + \cdots) x_r , & \Lambda^* &= \sum (\Lambda_r + \Lambda_{r+1} + \cdots) y_r . \end{aligned}$$

In (8), the first term 1 corresponds to $L = 1$, the terms K^* and Λ^* within the braces correspond to the cases $m_L = 1$ and $n_L = 1$, and the other terms are obtained by summing expressions like (7) from $M = 2$ to ∞. The expression for K^* is easily proved to be appropriate by writing it in the form

$$\sum (\kappa_r + 2\kappa_{r+1} + 3\kappa_{r+2} + \cdots) x_r .$$

Gontcharov's PGF (1) can be deduced from (8) by writing $\kappa_r = p^{r-1}q$, $\lambda_r = q^{r-1}p$, $K_r = p^{r-1}$, $\Lambda_r = q^{r-1}$, and $\bar{\kappa} + \bar{\lambda} = 1/(pq)$.

3. Scoring. If scores of s_r and s_r' are associated with runs of length r of 1's and 0's, and the total score is $S = \sum (s_r m_r + s_r' n_r)$, then the PGF of S is obtainable from (8) by replacing x_r by $x^{s_r} z$, y_r by $x^{s_r'} z$ and then extracting the coefficient of z^L to allow for the constraint $\sum r(m_r + n_r) = L$.

It would be possible to obtain the moments of the total score S from the generating function, but the calculations would be very heavy in the general case. Instead I shall exemplify the methods by means of a slightly different problem which is almost a special case of the above one.

4. Runs of successes. In the present section we shall be concerned with runs of ones only, but we shall include runs of zero length. We shall abbreviate the expression "run of success" simply to "run". Our model will be of the "second class" (see Section 1).

Let ρ_n and ρ_n' be the probabilities that respectively *true* and *apparent* runs of length n occur at a specified place in the (infinitely long) chain, where a true run is a run of *precisely* n successes (preceded and followed by a failure) and an apparent run is one that may or may not be preceded and followed by failures. A true run is regarded as "occurring" at its first success, and a true run of length n contains $n - m + 1$ apparent runs of length $m (m \leqq n)$. The distinction between true and apparent runs (and half-true ones: see Section 5) and formulas (10) to (13) were drawn to my attention in 1940 by A. M. Turing, who invented the regenerative model of the second class for the analysis of certain binary processes. We adopt the natural convention that

$$\rho_0' = 1 . \tag{9}$$

The following identities are easy to prove once they are pointed out:

$$\rho_n = \Delta^2\rho_n' = \rho_n' - 2\rho_{n+1}' + \rho_{n+2}' = \nabla^2\rho_{n+2}' \tag{10}$$

$$\rho_n + \rho_{n+1} + \rho_{n+2} + \cdots = \rho_n' - \rho_{n+1}' \tag{11}$$

$$\rho_0 + \rho_1 + \rho_2 + \cdots = 1 - \rho_1' \tag{12}$$

$$\rho_n + 2\rho_{n+1} + 3\rho_{n+2} + \cdots = \rho_n' \tag{13}$$

$$\rho_n + 3\rho_{n+1} + 6\rho_{n+2} + 10\rho_{n+3} + \cdots = \rho_n' + \rho_{n+1}' + \cdots \tag{14}$$

$$\textstyle\sum_{n=0}^{\infty} (n+1)^2\rho_n = 1 + 2(\rho_1' + \rho_2' + \rho_3' + \cdots) . \tag{15}$$

These formulas depend on the assumption of stationarity. They reduce to simple algebraic identities for the case of a random sequence (Bernoulli trials), for which $\rho_n' = p^n$, $\rho_n = (1 - p)^2p^n$, p being the probability of a success.

We shall assume further that the Markov chain is *regenerative after each failure*, and we shall think of the whole chain as starting with a failure. This model is almost a special case of the model of the first class considered in Section 2, and may be regarded as a compromise between that model and a Bernoulli process. But we are now further assuming that the sequence starts with a failure so as to avoid "initial end effects". This will not affect the asymptotic results. The probability of any specified sequence, of length N, can be readily expressed in terms of the ρ_n's.

We now define the *total score* as

$$S = s_0m_0 + s_1m_1 + s_2m_2 + \cdots , \tag{16}$$

where s_n is a score arising from each true run of length n. The advantage of considering total scores with the generality of S is that it includes various special cases of some interest. For example, the following formulas can be readily obtained from the subsequent formulas (20) and (21).

Let r_0 and r_1 be the numbers of failures and successes in the segment. Then, for large N,

$$E(r_1) = NA + (A + B)(1 - \tfrac{1}{2}A) + o(1) \tag{17}$$

$$\text{Var } r_0 = \text{Var } r_1 \sim N(B - A^2 - 2AB + A^3 + A^2B) , \tag{18}$$

where

$$A = \sum n\rho_n , \qquad B = \sum n^2\rho_n .$$

The variance of the number of runs of length ν is asymptotically

$$N\{\rho_\nu' + 2(\rho'_{\nu+1} + \rho'_{\nu+2} + \cdots) - (2\nu + 1)\rho_\nu'^2 - 4\rho_\nu'(\rho'_{\nu+1} + \rho'_{\nu+2} + \cdots) + 2(\rho_1' + \rho_2' + \cdots)\rho_\nu'^2\} . \tag{19}$$

More generally, under assumptions (38),

$$E(S) = N\sum_{n=0}^{\infty} t_n + \sum (n + 1)t_n - \sum t_n \sum_{n=1}^{\infty} \rho_n' + o(1) \tag{20}$$

$$\operatorname{Var}(S) = N\{\sum s_n t_n - \sum (2n + 1)t_n \sum t_n + 2(\rho_1' + \rho_2' + \cdots)(\sum t_n)^2\} + O(1) \tag{21}$$

where

$$t_n = \rho_n s_n . \tag{22}$$

5. A generating function for the score. We can regard "the sequence" as composed entirely of subsequences ("half-true runs") all of the form 0 1 1 1 $\cdots$ 1, juxtaposed, where each subsequence ends with a run of 1's but this run can be of zero length. By virtue of the regenerative assumption, this way of regarding the sequence reduces it to a random sequence, not of course to be confused with the Bernoulli sequence that we mentioned before as a special case. (Our "alphabet" is countably infinite.)

Let κ_n be the probability of the selection of a subsequence with n 1's on it. Then

$$\kappa_n = \rho_n/(1 - \rho_1') \tag{23}$$

$$\kappa_0 + \kappa_1 + \kappa_2 + \cdots = 1 \tag{24}$$

$$\kappa_1 + 2\kappa_2 + 3\kappa_3 + \cdots = \rho_1'/(1 - \rho_1') \tag{25}$$

$$m_0 + 2m_1 + 3m_2 + \cdots = N , \tag{26}$$

provided we count the right-most subsequence as a "true" run (but we shall allow for its correct probability). Let

$$m_0 + m_1 + m_2 + \cdots = M , \tag{27}$$

but we regard N as given, rather than M.

The probability of our sequence, if it ends with just c 1's, is

$$\kappa_0^{m_0}\kappa_1^{m_1}\cdots\kappa_{c-1}^{m_{c-1}}\kappa_c^{m_c-1}\kappa_{c+1}^{m_{c+1}}\cdots(\kappa_c + \kappa_{c+1} + \cdots) . \tag{28}$$

The PGF (probability generating function) of $(m_0, m_1, m_2, \cdots)$ is therefore

$$\sum_{m_0,m_1,\cdots}^{m_0+2m_1+\cdots=N} \sum_{c=0}^{\infty} \frac{(M-1)!}{m_0!\, m_1! \cdots m_{c-1}!\,(m_c - 1)!\, m_{c+1}! \cdots} (\kappa_0 x_0)^{m_0} \cdots (\kappa_{c-1} x_{c-1})^{m_{c-1}}\{(\kappa_c x_c)^{m_c-1}(\kappa_{c+1}x_{c+1})^{m_{c+1}} \cdots (\kappa_c + \kappa_{c+1} + \cdots)x_c\} . \tag{29}$$

If $N \geqq 1$, the PGF is the coefficient of y^N in the expression obtained from (29) by replacing x_n by $x_n y^{n+1}$ $(n = 0, 1, 2, \cdots)$, and where the condition

$$m_0 + 2m_1 + \cdots = N$$

is no longer required. Hence the PGF equals

$$\mathscr{C}(y^N)G(x_0, x_1, x_2, \cdots; y)\,,$$

where $\mathscr{C}(y^N)$ means "the coefficient of y^N in" and

$$G(x_0, x_1, \cdots; y) = \frac{\sum_{n=0}^{\infty} K_n x_n y^{n+1}}{1 - \sum_{n=0}^{\infty} \kappa_n x_n y^{n+1}} \tag{30}$$

where

$$K_n = \kappa_n + \kappa_{n+1} + \cdots = \frac{\rho_n' - \rho_{n+1}'}{1 - \rho_1'}\,. \tag{31}$$

From (24) we have

$$(y + y^2 + y^3 + \cdots)(1 - \sum \kappa_n y^{n+1}) = \sum K_n y^{n+1}\,, \tag{32}$$

so that

$$G(1, 1, 1, \cdots; y) = y + y^2 + y^3 + \cdots\,, \tag{33}$$

which is a check of (30).

If we replace x_n by x^{s_n} we deduce that the PGF of S is $\mathscr{C}(y^N)F(x, y)$ where

$$F(x, y) = \frac{\sum_{n=0}^{\infty} K_n x^{s_n} y^{n+1}}{1 - \sum_{n=0}^{\infty} \kappa_n x^{s_n} y^{n+1}}\,. \tag{34}$$

6. The moments of S: outline of derivations. The moment generating function of S is $\mathscr{C}(y^N)F(e^u, y)$ where u is the dummy variable. The derivation of the moments involves some points of mathematical rigor, but we give only an outline because the kinds of mathematics required are already in the literature (see, for example, Smith (1957)) and because the full details are decidedly intricate. We first mention the easily proved lemma that if

$$f(y) = \sum_{n=0}^{\infty} (\alpha n + \beta + \varepsilon_n) y^n\,,$$

where $\varepsilon_n \to 0$ as $n \to \infty$, then $\alpha = \lim_{y \to 1} (1 - y)^2 f(y)$. Now we can show, by means of (32), that

$$\frac{\partial}{\partial u} F(e^u, y)\bigg|_{u=0} = \frac{\sum K_n s_n y^{n+1} + (y + y^2 + \cdots) \sum \kappa_n s_n y^{n+1}}{1 - \sum \kappa_n y^{n+1}}\,. \tag{35}$$

Hence, assuming $E(S)$ is of the form

$$E(S) = \alpha N + \beta + o(1)\,, \tag{36}$$

we can deduce $\alpha = \sum t_n$ from the lemma. By a more lengthy but similar argument we can show that

$$\beta = \sum (n + 1) t_n - \sum t_n \sum_{n=1}^{\infty} \rho_n'\,. \tag{37}$$

By means of an even more lengthy argument (which makes use of Abel's theorem for power series), we can derive formula (21) for Var (S) under the assumptions

$$\kappa_n = O(\lambda^n) \quad \text{for some} \quad \lambda < 1 \tag{38}$$
$$|s_n| < n^c \quad \text{for some} \quad c\,.$$

REFERENCES

[1] GONTCHAROV, W. (1943). On runs of events in a series of independent trials of the Bernoullian type. *C. R. (Doklady) Acad. Sci. URSS* **38** 283–285. *MR* **5** 124.

[2] GONTCHAROV, W. (1944). On the field of combinatory analysis. *Bull. Acad. Sci. URSS Ser. Math.* **8**, no. 4 1–48. *MR* **6** 88.

[3] GOOD, I. J. (1961). The frequency count of Markov chain and the transition to continuous time. *Ann. Math. Statist.* **32** 41–48. *MR* 23A no. 4241.

[4] SMITH, W. L. (1957). Contribution to the discussion of a paper by J. G. Skellam and L. R. Shenton. *J. Roy. Statist. Soc. Ser. B* **19** 114–115. *MR* 19 990.

DEPARTMENT OF STATISTICS
VIRGINIA POLYTECHNIC INSTITUTE
& STATE UNIVERSITY
BLACKSBURG, VIRGINIA 24061