THE FREQUENCY COUNT OF A MARKOV CHAIN AND THE TRANSITION TO CONTINUOUS TIME

By I. J. Good

Admiralty Research Laboratory, Teddington, England

1. Introduction. Consider a chain, N letters long, generated by a discrete-time Markov process that has a finite number, t, of available states. Each state will be called a "letter", and the set of t states the "alphabet". I shall discuss the joint probability distribution of the frequencies of the t letters of the alphabet, in other words the probability distribution of the "frequency count" of the chain, by making use of what may be called a "pseudo probability generating function". The discussion makes use of the interesting method of multiple contour integration, previously used by Whittle for another problem concerning Markov chains. I shall then apply a transition to continuous time. For the case $t = 2$, the result for continuous time is already known, but our result is more general; and it is of interest to relate the theories of discrete and continuous time.

The main results are given by formulae (3), (8), (9), and (10). Formula (8), for example, gives the covariance between the frequencies of any pair of letters when the chain is ergodic and is in its stable state; formula (9) gives a neat expression for the variance of the number of 0's when $t = 2$, and shows clearly how it differs from the familiar result for binomial sampling; and formula (10) provides, in principle, the joint density function for the durations of the t states when time is continuous, and where the chain is not necessarily in a stable state.

I believe this paper is of interest largely for its methods. I have not found it convenient to present it in the conventional theorem-proof form.

2. Frequency Counts of a Markov Chain. Let the matrix of transition probabilities be $Q = (q_{\mu,\nu})$ $(\mu, \nu = 0, 1, \cdots, t-1)$. Let p_r be the probability that the first letter of the chain is r $(r = 0, 1, \cdots, t-1)$. These need not be stable-state probabilities. Let $p(\mathbf{n})$ be the probability that the letter frequency count will be $\mathbf{n} = (n_0, n_1, \cdots, n_{t-1})$, where $n_0 + n_1 + \cdots + n_{t-1} = N$. The probability generating function (P.G.F.) of the frequency count is (cf., [1])

$$\sum p(\mathbf{n})\mathbf{x}^{\mathbf{n}} = \sum p(n_0, n_1, \cdots, n_{t-1}) x_0^{n_0} x_1^{n_1} \cdots x_{t-1}^{n_{t-1}}$$

summed over all $\mathbf{n}$ for which $n_0 + \cdots n_{t-1} = N$. If however the summation is over all $\mathbf{n}$ for which $n_0 + \cdots + n_{t-1}$ is positive, then the result may be called the "universal" P.G.F., and it serves for all positive values of N simultaneously.

Let $\mathbf{e}$ be the column vector consisting of t 1's, and let $\mathbf{X}$ be the diagonal matrix diag $(x_0, \cdots, x_{t-1})$. Then it is easy to check that the P.G.F. is (cf., [1]

$$(1) \qquad (p_0x_0, p_1x_1, \cdots, p_{t-1}x_{t-1})(\mathbf{QX})^{N-1}\mathbf{e},$$

Received September 29, 1959; revised October 29, 1960.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

so that the universal P.G.F. is

$$(p_0x_0, p_1x_1, \cdots, p_{t-1}x_{t-1})(\mathbf{I} - \mathbf{QX})^{-1}\mathbf{e}$$
$$= (p_0x_0, p_1x_1, \cdots, p_{t-1}x_{t-1})\mathrm{adj}(\mathbf{I} - \mathbf{QX})\mathbf{e}/\det(\mathbf{I} - \mathbf{QX}). \tag{2}$$

I shall outline below[1] a proof that the coefficient of $\mathbf{x}^{\mathbf{n}}$ in (2) is equal to that of $\mathbf{z}^{\mathbf{n}}$ in

$$(z_0 + z_1 + \cdots z_{t-1}) \prod_{r=0}^{t-1} (f_r(\mathbf{z}))^{n_r-1} \sum_{r=0}^{t-1} p_r C_r, \tag{3}$$

where

$$f_r(\mathbf{z}) = b_r + \sum_{s=0}^{t-1} q_{s,r} z_s,$$

in which b_r is an arbitrary constant which can be allowed to be zero if $n_r > 0$ $(r = 0, 1, \cdots, t-1)$, and where the C_r are the cofactors of the diagonal elements of the matrix

$$(\delta_r^s f_r(\mathbf{z}) - q_{r,s}z_s). \tag{4}$$

(The double-suffix summation convention is not used in the present paper.) We may regard (3) as a *pseudo* P.G.F. It is not an ordinary P.G.F. since it depends on $\mathbf{n}$. In principle it may be used in order to obtain the asymptotic behaviour of $p(\mathbf{n})$, by invoking, for example, the saddle-point theorem, Theorem 6.3 of Good [6]. It could also be used in order to obtain the exact expectation of a function of $\mathbf{n}$, $\varphi(\mathbf{n})$, when $\sum_{\mathbf{n}} \varphi(\mathbf{n})\mathbf{w}^{\mathbf{n}}$ can be neatly expressed as a function of $\mathbf{w}$. For example, the moments could be obtained by this method. But the expectation and variance will be obtained below by a more standard method.

The last factor of (3) is a polynomial of degree $t - 1$, and therefore, when extracting the coefficient of $\mathbf{z}^{\mathbf{n}}$, its effect is at worst to complicate the algebra. The values of C_0 when $t = 2$ and $t = 3$ are

$$q_{01}z_0 \tag{5}$$

and

$$q_{01}q_{02}z_0^2 + q_{02}q_{21}z_0z_2 + q_{01}q_{12}z_0z_1. \tag{6}$$

In both these cases, and perhaps for all values of t, the coefficients in C_r are all non-negative. (A proof of this conjecture may well involve a direct proof of (3), without the help of (2).) When t is such that the conjecture is true, and in particular when $t = 2$ or 3, we have $p(\mathbf{n}) \sim A \cdot q(\mathbf{n})$, where A is mathematically independent of $\mathbf{n}$, and $q(\mathbf{n})$ is the coefficient of $\mathbf{z}^{\mathbf{n}}$ in

$$\prod_{r=0}^{t-1} (f_r(\mathbf{z}))^{n_r-1}. \tag{7}$$

[1] The proof is postponed to Section 4 in order that the continuity of the present discussion should not be interrupted.

Thus when the conjecture is true the algebraic complications mentioned above are so to speak asymptotically immaterial. It is to be understood that all components of $\mathbf{n}$ are to tend to infinity in the definition of asymptotic equivalence.

Unfortunately the work required in order to apply the saddlepoint method to this problem seems to be heavy. The above discussion is however of some mathematical interest, and may be of use if very accurate estimates of $p(\mathbf{n})$ are required. It will also be used below in order to discuss continuous-time processes. Meanwhile a less accurate approximation can be obtained by the following method, provided that the process is ergodic.

Let us break off for a moment in order to clarify some terminology.

If, in a discrete-time stochastic process, the transition probabilities at any stage depend only on the previous k states, and do not otherwise depend on time, then the process is called a kth-order Markov process. It is an ordinary Markov process when $k = 1$.

A succession of m states or letters occurring in a sequence in a chain is called an m-plet. A kth-order Markov process may be thought of as a first-order process by regarding its k-plets as states, the successor state of a k-plet being made up of its last k-1 letters together with the next letter of the chain. (Compare, for example, Good [5], de Bruijn [2], Bartlett [1].) If, for this revised interpretation of a state, the process is ergodic, then it is called an ergodic kth-order Markov process. If $l > k$, then a kth-order Markov process is also an lth-order Markov process. If it is an *ergodic* kth-order process, then it is easily seen to be an ergodic lth-order process.

If an m-plet occurs ν times in a chain, then ν is called the *frequency* of the m-plet, and $\nu/(N - m + 1)$ the *relative frequency*. The entire set of all m-plets in a given chain, with m fixed, is called the *frequency count* of the m-plets. The joint distribution of the relative frequencies of the m-plets of a kth-order ergodic Markov chain will, with probability 1, tend to a limit as the length of the chain tends to infinity.

In fact Bartlett [1] proved that, for an ergodic kth-order process, the joint distribution of the $(k + 1)$-plet relative frequencies in a chain of length N is asymptotically normal when N tends to infinity. But a linear combination of normal variates is again normal, hence the l-plets also have a joint normal distribution if $l < k + 1$. (The same is true if $l > k + 1$ since a Markov chain of order k is also one of any higher order.) In particular, the letter frequencies $(l = 1)$, $n_0, n_1, \cdots, n_{t-1}$, have asymptotically a joint normal distribution when the process is of order 1, as I shall assume again from now on. (This fact was also proved by Kolmogorov [8].) In order to approximate to the probability $p(\mathbf{n})$ it is therefore adequate first to note that the expectation of n_r is Nq_r (where $q_0, q_1, \cdots, q_{t-1}$ are the stable-state probabilities of letters: this result is exact and not merely asymptotic if the chain is in its stable state, which I shall assume, during the remainder of this section, for the sake of simplicity); and second to compute the covariance matrix, $(\mathrm{cov}(n_r, n_s))$.

Let $x_{i,r} = 1$ if the ith letter of the chain is an r, and let $x_{i,r} = 0$ otherwise.

Then

$$E(n_r n_s) = \sum_{i,j}^{1,2,\cdots,N} E(x_{i,r}x_{j,s})$$
$$= \delta_r^s q_r N + q_r \pi_{r,s} + q_s \pi_{s,r},$$

where $\pi_{r,s}$ is the (r, s) element of the matrix

$$(N-1)\mathbf{Q} + (N-2)\mathbf{Q}^2 + \cdots + \mathbf{Q}^{N-1}.$$

Since the process is ergodic, the only eigenvalue of $\mathbf{Q}$ of modulus 1 is 1 itself, and it has multiplicity 1. It corresponds to the left and right eigenvectors $\mathbf{q}' = (q_0, q_1, \cdots, q_{t-1})$ and $\mathbf{e}$, i.e., $\mathbf{q}'\mathbf{Q} = \mathbf{q}'$, $\mathbf{Q}\mathbf{e} = \mathbf{e}$. Note also that $\mathbf{q}'\mathbf{e} = 1$, and that

$$\mathbf{Q} = \mathbf{e}\mathbf{q}' + \mathbf{R},$$

where $\mathbf{R}$ is singular and has all its eigenvalues of modulus less than 1. It follows that $\mathbf{q}'\mathbf{R} = \mathbf{q}'\mathbf{Q} - \mathbf{q}'\mathbf{e}\mathbf{q}' = \mathbf{0}'$, and $\mathbf{R}\mathbf{e} = \mathbf{0}$, and hence that

$$\mathbf{Q}^n = \mathbf{e}\mathbf{q}' + \mathbf{R}^n \quad (n = 1, 2, 3, \cdots),$$

so that

$$(N-1)\mathbf{Q} + (N-2)\mathbf{Q}^2 + \cdots + \mathbf{Q}^{N-1} = \binom{N}{2}\mathbf{e}\mathbf{q}' + N\cdot\mathbf{R}(\mathbf{I}-\mathbf{R})^{-1} + O(1).$$

Therefore the required covariance is

$$\operatorname{cov}(n_r, n_s) = \delta_r^s q_r N + 2\binom{N}{2} q_r q_s + N q_r \{\mathbf{R}(\mathbf{I}-\mathbf{R})^{-1}\}_{r,s}$$
$$+ N q_s \{\mathbf{R}(\mathbf{I}-\mathbf{R})^{-1}\}_{s,r} - N^2 q_r q_s + O(1) \tag{8}$$
$$= N\{\delta_r^s q_r - q_r q_s + q_r(\mathbf{R}(\mathbf{I}-\mathbf{R})^{-1})_{r,s}$$
$$+ q_s(\mathbf{R}(\mathbf{I}-\mathbf{R})^{-1})_{s,r}\} + O(1).$$

When $t = 2$, the covariance matrix is determined completely by its top left-hand element, which reduces to

$$\operatorname{var}(n_0) = N q_0 q_1 (2\beta^{-1} - 1), \tag{9}$$

where β is the "association factor" between 0's and 1's, $\beta = q_{10}/q_0 = p_{10}/q_0 q_1 = p_{01}/q_0 q_1$, where p_{rs} is the stable probability of the 2-plet (r, s). For a random sequence the association factor is 1 and (9) reduces to the usual formula for a binomial variance. If the association factor is less than 1 there will be a tendency for 0's and 1's to occur in runs, and the variance of the number of 0's (and 1's) will be greater than for a random sequence. The association factor cannot exceed 2.

For any value of t, the part of the equation (8) that depends on $\mathbf{R}$ may be regarded as the part of the covariance that is attributable to "Markovity". For

weak Markovity, i.e., with all elements of **R** small, this contribution will be small, and also easy to approximate numerically.

3. Continuous-Time Processes. In order to avoid difficulties of rigour I shall take the point of view of some one, whom I shall call a "physicist", who wishes to apply the results to a particular physical problem. He will consider it adequate to regard time as continuous if he considers that the following assumption is also adequate for his problem. There is a smallest time unit, $d\tau > 0$, so small that its size cannot be determined, but only upper bounds on its size. (There will also be a smallest space unit.) All time intervals are then integer multiples of $d\tau$.

Thus the physicist will demand that the solution of the (discrete-time) problem with an assumed small enough value of $d\tau$ must be experimentally indistinguishable from the solution with any smaller value of $d\tau$. He will then be satisfied *a fortiori* if all physically measurable aspects (measured in standard units, not as multiples of $d\tau$) of the solution tend to limits as $d\tau \to 0$. If this condition is met we may say, by definition, that we have obtained the solution of the continuous-time problem in a form adequate for the physicist. I do not know whether this definition is more or less realistic than the usual one.

Consider then a t-state continuous-time Markov process, with constant infinitesimal transition probabilities. Let the total times in the t states, $0, 1, \cdots, t-1$ be $\tau_0, \tau_1, \cdots, \tau_{t-1}$. These must be integer multiples of the time element, $d\tau$, and we may write $\tau_r = n_r\, d\tau$ $(r = 0, 1, \cdots, t-1)$. From the previous results concerning discrete time, we may deduce the joint distribution of $(\tau_0, \cdots, \tau_{t-1})$. The total time, $\tau_0 + \tau_1 + \cdots + \tau_{t-1} = \tau$, is regarded as given. The joint distribution is not normal: it would be fallacious to argue that "it must be because that of $n_0, n_1, \cdots, n_{t-1}$ is normal." This argument fails because if we divide time up into N small intervals each of length $d\tau$, where $Nd\tau = \tau$, and then let $d\tau \to 0$ and $N \to \infty$, the transition probabilities do not remain constant.

We may write

$$q_{r,s} = \alpha_{r,s}\, d\tau \ (s \neq r), \qquad q_{r,r} = 1 - \sum_s \alpha_{r,s}\, d\tau,$$

where, by convention, $\alpha_{r,r} = 0$. The probability density of $(\tau_0, \tau_1, \cdots, \tau_{t-1})$, when $\tau_0 > 0, \tau_1 > 0, \cdots, \tau_{t-1} > 0$, can be obtained from the pseudo P.G.F., formula (3), with all the b_r's equal to zero. We write $\tau_r/d\tau$ for n_r, and take the limit of the probability after dividing by $(d\tau)^{t-1}$, since we are in the $(t-1)$-dimensional simplex $\tau_0 + \tau_1 + \cdots + \tau_{t-1} = \tau$. We find that the density is the limit, if this limit exists, of the constant term (the term mathematically independent of the z's) in

$$\frac{\sum z_r}{\prod z_r} \sum p_r D_r \prod \left\{1 + d\tau \sum_s \left(\frac{\alpha_{sr} z_s}{z_r} - \alpha_{rs}\right)\right\}^{\tau_r/d\tau},$$

where $D_0, D_1, \cdots, D_{t-1}$ are the cofactors of the diagonal elements of the matrix

$$(\delta_r^s \sum_u \alpha_{ur} z_u - \alpha_{rs} z_s).$$

(For the present, summations and products of unspecified range are from $r = 0$ to $r = t - 1$.) Thus the probability density, when $\prod \tau_r \neq 0$, is equal to the constant term in

$$\exp\left(-\sum_{r,s} \alpha_{r,s}\tau_r\right) \frac{\sum z_r}{\prod z_r} \sum p_r D_r \exp\left(\sum_{r,s} \alpha_{s,r}\tau_r z_s/z_r\right). \tag{10}$$

For example, when $t = 2$, the probability density of τ_0 (or of τ_1), when $\tau_0\tau_1 > 0$, is equal to the constant term in

$$\exp(-\alpha_{01}\tau_0 - \alpha_{10}\tau_1)(z_0^{-1} + z_1^{-1})(p_0\alpha_{01}z_0 + p_1\alpha_{10}z_1) \times \exp(\alpha_{01}\tau_1 z_0/z_1 + \alpha_{10}\tau_0 z_1/z_0),$$

i.e., the density is

$$e^{-\alpha_{01}\tau_0 - \alpha_{10}\tau_1}\{(p_0\alpha_{01} + p_1\alpha_{10})I_0(2(\alpha_{01}\alpha_{10}\tau_0\tau_1)^{\frac{1}{2}}) + (\alpha_{01}\alpha_{10})^{\frac{1}{2}}(p_0(\tau_0/\tau_1)^{\frac{1}{2}} + p_1(\tau_1/\tau_0)^{\frac{1}{2}})I_1(2(\alpha_{01}\alpha_{10}\tau_0\tau_1)^{\frac{1}{2}})\}, \tag{11}$$

where I_0 and I_1 are the Bessel functions of imaginary argument of orders 0 and 1. If the Markov process starts in its stable state we have $p_0 = \alpha_{10}/(\alpha_{01} + \alpha_{10})$, $p_1 = \alpha_{01}/(\alpha_{01} + \alpha_{10})$. If the initial state is known to be 0 then the density is obtained by putting $p_0 = 1$, $p_1 = 0$ in (11). The cumulative distribution can be deduced with the aid of Erdélyi *et al.* [4], p. 201 (16), together with a formula obtained from it by partial integration. Formula (11) is given by Dobrušin [3], with an acknowledgement to F. I. Karpelevitch and V. A. Uspensky. See also Takács [10], who gives the cumulative distribution. My method, and the methods used in these references are all distinct, and in the references the usual definition of a continuous-time process is used.

Formula (10) may be used in order to obtain the expectation of a function $\varphi(\boldsymbol{\tau})$, and the moments of the distribution of $\boldsymbol{\tau}$ could thus be obtained. If we denote the multidimensional Laplace transform of φ by φ^*, where

$$\varphi^*(\mathbf{x}) = \int_0^\infty \cdots \int_0^\infty \varphi(\boldsymbol{\tau}) exp\,(-\mathbf{x}'\cdot\boldsymbol{\tau})\, d\boldsymbol{\tau},$$

then the expected value of $\varphi(\boldsymbol{\tau})$ is equal to the constant term in

$$\frac{\sum z_r}{\prod z_r} \sum p_r D_r \varphi^*\left(\sum_s \alpha_{0s} - \sum_s \alpha_{s0} z_s/z_0, \cdots, \sum_s \alpha_{t-1,s} - \sum_s \alpha_{s,t-1} z_s/z_{t-1}\right). \tag{12}$$

These methods can be at least formally extended to the case of a Markov process having a continuous infinity of states, with discrete or continuous time, by making use of probability generating functionals or characteristic functionals.

4. Proof of formula (3). In Section 2, I postponed the proof of (3). This proof can be based on a generalization to several variables of Lagrange's expansion of

an implicit function as a power series. (See Good [7], which is related to earlier work by Whittle [11].) Leaving aside here the finer points of rigor, the coefficient of $\mathbf{z}^{\mathbf{n}}$ in a function $h(\mathbf{z})$, analytic in a neighbourhood of the origin, $\mathbf{z} = \mathbf{0}$, is equal to

$$\left(\frac{1}{2\pi i}\right)^t \oint \cdots \oint \frac{h(\mathbf{z})\, d\mathbf{z}}{z_0^{n_0+1} \cdots z_{t-1}^{n_{t-1}+1}} = \frac{1}{(2\pi i)^t} \oint \cdots \oint \frac{h(\mathbf{z}(\mathbf{x}))}{\prod (z_r(\mathbf{x}))^{n_r+1}} \cdot \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \cdot d\mathbf{x},$$

if the vector function $\mathbf{z}(\mathbf{x})$ is also analytic in the neighbourhood of the origin $\mathbf{x} = \mathbf{0}$. Let the relationship between $\mathbf{z}$ and $\mathbf{x}$ be $x_r = z_r/f_r(\mathbf{z})$, where the f's are defined just below formula (3), and must not vanish at the origin (so that the b_r's must not vanish). In this case it is a simple matter to compute the inverse Jacobian, and, on writing $h = k \cdot f^n$, where k is another function of $\mathbf{z}$, we find that the coefficient of $\mathbf{x}^{\mathbf{n}}$ in $k(\mathbf{z}(\mathbf{x}))/\det(\mathbf{I} - \mathbf{QX})$ is equal to that of $\mathbf{z}^{\mathbf{n}}$ in $k \cdot f^n$. We now select k so that

$$k(\mathbf{z}(\mathbf{x})) = (p_0 x_0, \cdots, p_{t-1} x_{t-1}) \cdot \operatorname{adj}(\mathbf{I} - \mathbf{QX}) \cdot \mathbf{e}.$$

The multiplier of p_r in this expression is easily seen to be the determinant obtained from the matrix $(\delta_r^s - q_{rs} x_s)$ by replacing each element in its rth column by x_r. Now express the x's in terms of the z's, and (3) follows on noting that

$$\begin{vmatrix} z_0, & -q_{01} z_1 & , & -q_{02} z_2, & \cdots \\ z_0, & q_{01} z_0 + q_{21} z_2 + \cdots, & & -q_{12} z_2, & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ z_0, & -q_{t-1,1} z_1 & , & -q_{t-1,2} z_2, & \cdots \end{vmatrix}$$

$$= \begin{vmatrix} z_0 + z_1 + \cdots + z_{t-1}, & 0 & , & 0 & , \cdots \\ z_0 & & , & q_{01} z_0 + q_{21} z_2 + \cdots, & -q_{12} z_2, \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \end{vmatrix},$$

as we may see by adding to the top row of the first determinant the multiples z_1/z_0, z_2/z_0, $\cdots$ of the remaining rows.

A similar, but shorter, proof can be supplied for MacMahon's "Master Theorem," (MacMahon [9], pp. 93–123). I hope to publish it elsewhere.

REFERENCES

[1] M. S. Bartlett, "The frequency goodness-of-fit test for probability chains," *Proc. Camb. Philos. Soc.*, Vol. 47 (1951), pp. 86–95.

[2] N. G. de Bruijn, "A combinatorial problem," *Nederl. Akad. Wetensch., Proc.*, Vol. 49 (1946), pp. 758–764 and *Indagationes Math.*, Vol. 8 (1946), pp. 461–467.

[3] R. L. Dobrušin, "Limit theorems for a Markov chain of two states," *Izvestiya Akad. Nauk. SSSR. Ser. Mat.*, Vol. 17 (1953), pp. 291–330 (in Russian).

[4] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Tables of Integral Transforms*, Vol. I. Based in Part *on Notes Left by Harry Bateman*, McGraw-Hill, New York, 1954.

[5] I. J. Good, "Normal recurring decimals," *J. London Math. Soc.*, Vol. 21 (1946), 167–169.

[6] I. J. Good, "Saddle-point methods for the multinomial distribution," *Ann. Math. Stat.*, Vol. 28 (1957), pp. 861–881.

[7] I. J. GOOD, "Generalizations to several variables of Lagrange's expansion, with applications to stochastic processes," *Proc. Camb. Philos. Soc.*, Vol. 56 (1960), pp. 367–380.
[8] A. N. KOLMOGOROV, "A local limit theorem for classical Markov chains," *Izvestiya Akad. Nauk. SSSR. Ser. Mat.*, Vol. 13 (1949), pp. 281–300 (in Russian).
[9] P. A. MACMAHON, *Combinatory Analysis*, Vol. I, Cambridge, University Press, 1915.
[10] L. TAKÁCS, "On certain sojourn time problems in the theory of stochastic processes," *Acta Math. Acad. Sci. Hung.*, Vol. 8 (1957), pp. 169–191.
[11] P. WHITTLE, "Some distribution and moment formulae for the Markov chain," *J. Roy. Statist. Soc.*, Ser. B, Vol. 17 (1955), pp. 235–242.