Brèves communications - Kurze Mitteilungen
Brevi comunicazioni - Brief Reports

Note on the Estimation of the Number of States in a Discrete Markov Chain

In some applications of the theory of finite Markov chains to problems in physics, genetics, and radiobiology[1] a discrete chain with two absorbing states and with transition probabilities which are functions of the number of states in the chain has been considered. In certain cases the number of states in the chain may be known and specified in advance; in others, however, the number of states may be an unknown parameter associated with the chain that we would like to estimate on the basis of experimental observations. In this note a method of obtaining the maximum likelihood estimate of the number of states is given.

We consider a discrete chain with $a+1$ states E_0, E_1, ..., E_a, say and with transition probabilities

$$\left.\begin{aligned} p_{i,i+1} &= Pr\,(E_i \to E_{i+1}) = \gamma_i(a) \\ p_{i,i-1} &= Pr\,(E_i \to E_{i-1}) = 1-\gamma_i(a), \end{aligned}\right\} \tag{1}$$

where $\gamma_i(a)$, $0 < \gamma_i(a) < 1$, $(i = 1, 2, \dots, a-1)$ is a given function of a. Also

$$p_{ii} = Pr\,(E_i \to E_i) = \begin{cases} 1, & i = 0, a \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

This chain is equivalent to a particle executing a random walk on the integers $0, 1, \dots, a$, in which a particle at position i $(i = 1, 2, \dots, a-1)$ has probability $\gamma_i(a)$ of moving one step to the right, and probability $1-\gamma_i(a)$ of moving one step to the left. Once the particle reaches either 0 or a it is absorbed and the random walk is terminated. This random walk representation with $\gamma_i(a) = i/a$ has been used in radiobiology as a model for the spread of radiation damage[2]. When $\gamma_i(a) = 1-i/a$ we have the classical Ehrenfest model of diffusion[3].

Consider the following experiment: At time zero a particle is placed in position i $(i = 1, 2, \dots, a-1)$ and permitted to execute a random walk on the finite set of integers $0, 1, \dots, a$ until it is absorbed at either 0 or a. Let t and u represent the observed times required for the termination of the random walk at 0 and a, respectively. It is obvious that t and u will depend on the initial position i; however we do not introduce this dependence explicitly in the notation. Let us now perform the experiment N times, m of which it ends up at 0, and n of which it ends up at a, $m + n = N$. Call t_j $(j = 1, 2, \dots, m)$ and u_k $(k = 1, 2, \dots, n)$ the observed times required for the termination of the N independent random walks at 0 and a, respectively. If we let σ represent the number of steps or transitions per unit time we can call $t'_j = \sigma t_j$ and $u'_k = \sigma u_k$ the number of transitions required for the particle to pass from some initial state i to 0 and a, respectively. In Markoc chain theory t'_j and u'_k are the first passage times for the absorbing states 0 and a[4].

Now, if the probability of the particle being at position i at time zero, is π_i, $0 \leqq \pi_i \leqq 1$, $(i = 1, 2, \dots, a-1)$ then the probability of the particle being at position k after n transitions is

$$\pi_k^{(n)} = \sum_{i=1}^{a-1} \pi_i \, p_{ik}^{(n)}, \tag{3}$$

where $p_{ik}^{(n)}$ is the probability of passing from position i to position k in n steps. Hence the joint probability density for the m particles being absorbed at zero is given by

$$\prod_{j=1}^{m} \pi_0^{(t'_j)} = \prod_{j=1}^{m} \sum_{i=1}^{a-1} \pi_i \, p_{i0}^{(t'_j)}; \tag{4}$$

and the joint probability density for the n particles being absorbed at a is given by

$$\prod_{k=1}^{n} \pi_a^{(u'_k)} = \prod_{j=1}^{m} \sum_{i=1}^{a-1} \pi_i \, p_{ia}^{(u'_k)}. \tag{5}$$

In (4) and (5) $p_{i0}^{(t'_j)}$ and $p_{ia}^{(u'_k)}$ are the probabilities of entering the states 0 and a for the first time in t'_j and u'_k steps, respectively. These probabilities are functions of the unknown parameter a. From (4) and (5) we have, as the joint probability density, for the N independent observations

$$F_N(t'_1, \dots, t'_m;\ u'_1, \dots, u'_n;\ a) = \left(\prod_{j=1}^{m} \pi_0^{(t'_j)}\right) \left(\prod_{k=1}^{n} \pi_a^{(u'_k)}\right). \tag{6}$$

Taking the logarithm of (6) we obtain the likelihood function

$$G_N = \log F_N = \sum_{j=1}^{m} \log \pi_0^{(t'_j)} + \sum_{k=1}^{n} \log \pi_a^{(u'_k)}. \tag{7}$$

Hence, the maximum likelihood estimate of a is given by the root of the equation

$$\frac{\partial G_N}{\partial a} = 0. \tag{8}$$

---

[1] A. T. BHARUCHA-REID, Biometrics *9*, 275 (1953). – A. T. BHARUCHA-REID and H. G. LANDAU, Bull. math. Biophysics *13*, 153 (1951). – W. FELLER, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1950). – G. MALÉCOT, C. r. Acad. Sci. Paris *219*, 379 (1944).

[2] A. T. BHARUCHA-REID, Biometrics *9*, 275 (1953). – A. T. BHARUCHA-REID and H. G. LANDAU, Bull. math. Biophysics *13*, 153 (1951).

[3] W. FELLER, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1950).

[4] W. FELLER, *An Introduction to Probability Theory and Its Applications* (Wiley, New York, 1950). – M. S. BARTLETT, Proc. math. Cambridge, philos. Soc. *49*, 263 (1953). – T. E. HARRIS, Trans. Amer. Soc. *73*, 471 (1952).

Let $P = \| p_{ij} \|$, i, j 0, 1, ..., a, be the matrix of transition probabilities. Since the chain has two absorbing states P can be decomposed by eliminating the first and last rows and columns of P. Denote by Q the decomposed matrix. Since P is decomposable we have

$$\left.\begin{aligned} p_{i0}^{(k)} &= p_{i1}^{(k-1)}\, p_{i0} = p_{i1}^{(k-1)}\, [1-\gamma_i(a)] \\ p_{ia}^{(u)} &= p_{i,\,a-1}^{(u-1)}\, p_{a-1,\,a} = p_{i,\,a-1}^{(u-1)}\, [\gamma'_i(a)] \,. \end{aligned}\right\} \quad (9)$$

Now $p_{i1}^{(k-1)}$ is the element in the i-th row and first column of $Q^{(k-1)}$, and $p_{i,\,a-1}^{(u-1)}$ is the element in the i-th row and (a–1)-st column of $Q^{(u-1)}$. Hence, for given $\gamma'_i(a)$ the above probabilities can be found by ordinary matrix multiplication of Q, or by the method of generating functions[3]. Having obtained these probabilities equation (8) can be solved to obtain the maximum likelihood estimate of the number of states in the Markov chain.

A. T. Bharucha-Reid

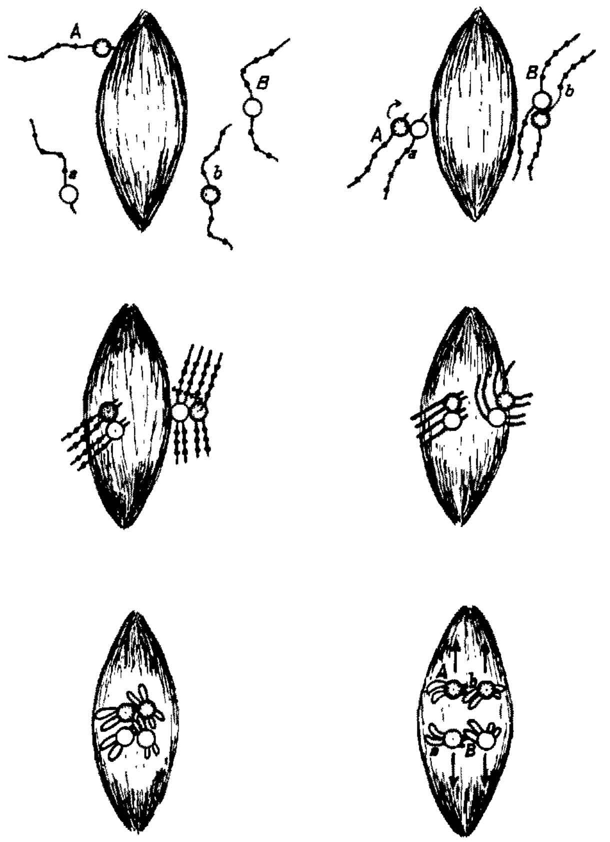
Department of Statistics, University of California, Berkeley 4, December 14, 1955.

Zusammenfassung

Diese Arbeit behandelt das Problem der statistischen Schätzung der Anzahl der Zustände einer diskreten Markoff-Kette, wenn die zugehörigen Übergangswahrscheinlichkeiten bekannte Funktionen der Anzahl der Zustände sind. Die Methode benutzt das Maximum-Likelihood-Prinzip.

Nonrandom Assortment of Centromeres with Implications Regarding Random Assortment of the Chromosomes*

The genes analyzed by Mendel[1] showed no linkage, and he proposed that the hereditary characters are assorted independently. The discovery of linked genes provided exceptions and "Mendel's second law" in its current form is stated by Sturtevant and Beadle[2] as follows: "Independent assortment is, in fact, the rule that applies to genes carried by separate chromosome pairs." Since crossing-over randomizes the assortment of genes distal to the centromere, the principle of random assortment applies specifically to the centromeres of separate chromosome pairs; if the centromeres are assorted independently (in the absence of crossing-over) the genes on the separate chromosome pairs will be independently assorted. Even in the most sophisticated texts[3] the inference (random assortment of centromeres) is represented as a fact by showing the paternal chromosomes as solid diagrams which are assorted at random to the gametes. Since the centromere itself has no "character" preferential segregation can be detected only by the behavior of the genes nearest the centromeres. In *Drosophila*, however, chiasmata in the gene-free region of heterochromatin (between the centromeres and the euchromatin of the first, second and third chromosomes) would randomize the effects of any directed assortment of centromeres which might occur.



A diagram showing the effects of nonrandom assortment of the centromeres of two different chromosome-pairs at meiosis in an $AB \times ab$ hybrid. Synapsis of the chromosomes begins at the centromeres and as the synapsed chromosomes travel to the spindle the like centromeres (stippled *versus* open) are attracted to each other. The tetrads line up on the spindle so that like centromeres pass to the same pole. In the absence of crossing-over only recombinant genomes are formed. The diagram describes the events which produce reverse linkage; the same mechanism would produce direct linkage if the two stippled centromeres had originated from the same parent.

Tetrad analysis of *Saccharomyces* has revealed that the centromeres of chromosomes are not assorted independently of each other. R-distributions[4] have been interpreted to result from the synapsis of the centromeres of chromosomes at meiosis I originating from different parents of the hybrid thus directing recombinant chromosomes to the same poles and reducing the frequency of original combinations except by crossing-over between either of the genes and their respective centromeres (Figure).

These facts led to an analysis of the conventional method of sampling single chromatids from mixed matings of diploid parents otherwise genetically identi-

---

* This work has been supported by research grants from the National Cancer Institute of the National Institutes of Health, U.S. Public Health Service C-1179 and C-2140.

[1] G. Mendel, Verhandlungen Naturforschender Verein in Brunn, Abhandlungen IV (1865).

[2] A. H. Sturtevant and G. W. Beadle, *An Introduction to Genetics* (W. B. Saunders Co., Philadelphia, 1939).

[3] E. W. Sinnott, L. C. Dunn, and Th. Dobzhansky, *Principles of Genetics* (McGraw-Hill Book Company, Inc., New York 1950).

[4] E. E. Shult and C. C. Lindegren, Genetica (in press).