240. Note: On Estimating the Equilibrium and Transition Probabilities of a Finite-State Markov Chain from the Same Data

Joel E. Cohen

Biometrics, Vol. 24, No. 1. (Mar., 1968), pp. 185-187.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28196803%2924%3A1%3C185%3A2NOETE%3E2.0.CO%3B2-G

Biometrics is currently published by International Biometric Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ibs.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

http://www.jstor.org
Tue Mar 13 18:00:56 2007

McArthur fails to specify whether the sibs having twin children were males or females; the probability she records is that for male sibs. As may be seen in the table the result for the relationship 'X's sister had DZ twins' is $\frac{1}{4}(1 + q)^2$.

ACKNOWLEDGMENT

This work was supported by U. S. Public Health Service Research Grant GM 11803.

DISTRIBUTION, PARMI LES APPARENTES, DE GENOTYPES CONDITIONNANT LA GEMELLITE

RESUME

La probabilité qu'une personne soit d'un génotype donné quand on connait le génotype d'un sujet qui lui est apparenté, est obtenue avec l'aide de matrices de probabilité conditionnelle.

On considère quatre hypothèses avec un gène unique pour expliquer l'hérédité de la gemellité chez l'homme.

REFERENCES

Li, C. C. and Sacks, L. [1954]. The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. *Biometrics 10*, 347–60.

McArthur, N. [1952]. A statistical study of human twinning. *Ann. Eugen. 16*, 338–50.

240 NOTE: On Estimating the Equilibrium and Transition Probabilities of a Finite-State Markov Chain from the Same Data

Joel E. Cohen

Computation Laboratory, Harvard University, Cambridge, Mass. 01238, U. S. A.

SUMMARY

A single set of data that conforms approximately to the condition of stationarity sheds no light on the validity of a certain Markov chain model. Tests for validity should be made before asking whether the stationary distribution has been reached.

The results of a variety of behavioral and biological studies may be represented by a finite string of letters chosen from a discrete, finite alphabet. For example, a study of animal vocalizations may yield a sequence of discretely categorized signals. It is frequently reasonable to see whether the observed string could have been generated by a Markov chain with constant transition probabilities and a finite number n of states equal to the number of distinct letters observed. Each distinct letter corresponds to one state of the chain.

Supposing the string had been generated by an irreducible, aperiodic Markov chain with a matrix $\mathbf{P} = (p_{ij})$ of transition probabilities, it is known that there exists a unique stationary (row) vector $\mathbf{p} = (p_i)$ of probabilities of occurrence which satisfies $\mathbf{p} = \mathbf{pP}$. Some investigators are tempted to suppose that they may test the validity of assuming a Markov chain source by seeing how closely their maximum-likelihood estimates $\mathbf{q}$ of $\mathbf{p}$ and $\mathbf{Q}$ of $\mathbf{P}$ satisfy the condition of stationarity

$$\mathbf{q} = \mathbf{qQ}, \tag{1}$$

where again $\mathbf{q}$ is a row vector. The purpose of this note is to point out that if the same string of observations is used to obtain both maximum-likelihood estimates $\mathbf{q}$ and $\mathbf{Q}$, then (1) *must* be approximately satisfied, with an accuracy which increases as the length of the string of data. Hence the observation that a single set of data conforms nearly to (1) sheds no light on the validity of the Markov chain model.

Let N be the length of the string. Let N_i be the number of times the ith letter occurs in the string and N_{ij} the number of times the ordered pair of (i, j)th letters appears in the string, for $i, j = 1, 2, \cdots, n$. Suppose the i'th letter of the alphabet begins the string and the i''th letter of the alphabet ends it. If $N_{i.} = \sum_{j=1}^{n} N_{ij}$ and $N_{.j} = \sum_{i=1}^{n} N_{ij}$ and if δ_{kl} is the Kronecker delta, then for $i = 1, 2, \cdots, n$,

$$N_i = N_{i.} + \delta_{i,i''} = N_{.i} + \delta_{i,i'} . \tag{2}$$

The maximum-likelihood estimate q_i of the ith probability of occurrence p_i is $q_i = N_i/N$. The maximum-likelihood estimate q_{ij} of the transition probability p_{ij} is $q_{ij} = N_{ij}/N_{i.}$. Hence the jth component of the row vector $\mathbf{qQ}$ is

$$\begin{aligned}\sum_{i=1}^{n} q_i q_{ij} &= \sum_{i=1}^{n} \frac{N_i}{N} \cdot \frac{N_{ij}}{N_{i.}} = \sum_{i=1}^{n} \frac{N_{i.} + \delta_{i,i''}}{N} \cdot \frac{N_{ij}}{N_{i.}} \\ &= \sum_{i=1}^{n} \frac{N_{ij}}{N} + \sum_{i=1}^{n} \frac{\delta_{i,i''} N_{ij}}{N N_{i.}} = \frac{N_{.j}}{N} + \frac{N_{i''j}}{N N_{i''.}} .\end{aligned} \tag{3}$$

But by (2), $N_{.j} = N_j - \delta_{j,i'}$. Substituting into the last member of (3) gives

$$\sum_{i=1}^{n} q_i q_{ij} = \frac{N_j}{N} - \frac{\delta_{j,i'}}{N} + \frac{N_{i'',j}}{N N_{i''.}} = q_j + \frac{1}{N} \left(\frac{N_{i''j}}{N_{i''.}} - \delta_{j,i'} \right) \tag{4}$$

which approaches q_i as N increases. Thus for a given long string, the estimated equilibrium or stationary probabilities $\mathbf{q}$ and estimated transition probabilities $\mathbf{Q}$ must nearly satisfy (1).

The correct tests for deciding whether a string could have been generated by a finite-state Markov chain with constant transition probabilities and of given order are stated explicitly in Billingsley [1961]. These tests should be made before asking whether the stationary distribution has been reached, that is, whether $\mathbf{q}$ and $\mathbf{Q}$, each derived from different data from the same source, satisfy (1).

SUR L'ESTIMATION DES PROBABILITES D'EQUILIBRE ET DE TRANSITION D'UNE CHAINE DE MARKOV A NOMBRE D'ETATS FINI A PARTIR DES MEMES DONNEES

RESUME

Un ensemble unique de données approximativement conformes à la condition de stationnarité n'apporte aucune lumière sur la validité d'un modèle markovien. Les tests de validité devraient être faits avant de se demander si la distribution stationnaire a été atteinte.

REFERENCE

Billingsley, P. [1961]. *Statistical Inference for Markov Processes.* University of Chicago Press, Chicago.

241 NOTE: Equilibrium under Selection at a Multi-allelic Sex-Linked Locus

C. Cannings

Department of Statistics, University of Aberdeen, Aberdeen, Scotland

SUMMARY

A representation suggested by Tallis [1966] is used to find equilibria for a multi-allelic sex-linked locus, the equilibria being a generalization of those for the diallelic sex-linked locus.

Tallis [1966] has demonstrated that by representing the gene frequencies of a population by a vector $\mathbf{p}$, and the genotypic viabilities by a matrix γ, one can find the equilibria of a multiallelic autosomal locus under selection. These are given by $\mathbf{p} = \gamma^{-1}\mathbf{1}\lambda$, where $\lambda = (\mathbf{1}^T\gamma^{-1}\mathbf{1})^{-1}$ and $\mathbf{1}$ is the unit vector. The procedure for finding the value of $\mathbf{p}$ is equivalent to the maximization of the mean viability of the population $\omega = \mathbf{p}^T\gamma\mathbf{p}$.