Behavior Research Methods, Instruments, & Computers
1984, 16 (3), 335-336

SAMPLE and TEST: Two FORTRAN IV programs for analysis of discrete-state, time-varying data using first-order, Markov-chain techniques

ROBERT B. ARUNDALE
University of Alaska, Fairbanks, Alaska

Discrete-state Markov-chain analysis provides a means of analyzing change over time in any behavior that can be characterized as occupying one of two or more discrete states at any one point in time. Time-referenced sequences of discrete behavior states arise in observational and experimental studies of both individuals and social groups.

A first-order, discrete-state Markov chain is a sequence of behavior states in which the probability of the organism or system's being in a given state at time t_2 depends on the state the entity was in at the immediately preceding point in time, t_1. Time measurement may be either ordinal or interval. The central descriptive device in Markov-chain analysis is an m x m matrix showing the probabilities of transitions from each of the m possible states at point t_1 to each of the m states at t_2. Normally, each individual or group under study will exhibit its own unique sequence of behavior states. The transition probability matrix extracts from this unique sequence the general pattern of transitions among the states. That pattern may be examined in its own right or may be compared with the patterns of other entities behaving under similar or different conditions.

Anderson and Goodman (1957) provided goodness-of-fit and likelihood-ratio statistics both for assessing whether or not a given sequence of behavior states meets the assumptions underlying Markov-chain analysis (see also Hewes, 1980) and for comparing pairs or sets of transition probability matrices. Valid application of first-order Markov-chain techniques requires that the behavior sequence be a first-order chain, as opposed to a zero-order (independence among the states) or a higher order chain (dependence on the two or more preceding states). Markov-chain analysis generalizes to higher order chains, but the number of time points required to obtain stable transition probabilities often becomes excessive. Valid application also requires that the behavior sequence represented in a single transition probability matrix be stationary. That is, the probabilities for each transition must not vary significantly over time within the sequence. Again, Markov analysis generalizes to nonstationary sequences, but the techniques for dealing with them are not well developed. Program SAMPLE allows one to assess whether the order and stationarity assumptions have been met. If the assumptions are valid, program TEST allows one to test both for significant differences between pairs of transition probability matrices and for homogeneity among a group of matrices.

Program SAMPLE

Program description. Program SAMPLE computes both the transition frequency and transition probability matrices for a sequence of behavior states and performs either or both of two types of sampling from the sequence, as specified by the user. Arundale (1977) found that if the behavior of an entity is indexed with high frequency, the probabilities of transitions to the same state may be artificially inflated, resulting in a distorted view of the pattern of changes among the states. Appropriate sampling from the original sequence can correct the artificial inflation. If specified by the user, the program also (1) finds the point of convergence or "steady state" of the process (see Hewes, 1975, for information), (2) constructs a distribution of lengths of time that the sequence remains in each state and fits this distribution to an exponential curve (see Arundale, 1977, pp. 263-267, for information), and (3) calculates both goodness-of-fit and likelihood-ratio statistics for testing stationarity and/or the first-order Markov property.

Input. The program is designed for batch-mode operation. The user provides the total number of time points in the sequence, the number of distinct states in the sequence and their codes (any machine-readable code is acceptable), a value for convergence tests, a descriptive title for identification, and the format of the data, all on three initializing control cards or lines. These are followed by cards or lines containing the sequence itself. On subsequent control cards or lines (which are unlimited in number), the user specifies the desired processing options: full sequence or sampling (with sampling parameters), convergence testing and optional printing of matrices, distribution construction, punching of matrices for subsequent testing, order statistics, and/or stationarity statistics (and parameters).

Output. The output includes all descriptive information, a listing of the entire sequence, and the transition frequency and probability matrices. Subsequent output depends on the processing options selected. If sampling is performed, the sampled sequence is printed, together with its length and associated matrices. Convergence testing may show either all powers of the matrix or only the converged matrix, as specified. Distribution construction includes, for each state, a listing of time lengths and their frequencies, plus mean length, standard deviation, and R^2 for the fit to an exponential distribution. If matrices are punched, they are provided in the format required for input to program TEST, as described below. If order statistics are requested, the output

The author is affiliated with the Department of Speech and Drama, University of Alaska, Fairbanks, AK 99701.

includes degrees of freedom and goodness-of-fit and likelihood-ratio statistics for testing both the null hypothesis of zero order (to be rejected) and the null hypothesis of first order (to be accepted). Finally, a request for stationarity statistics provides degrees of freedom, goodness-of-fit and likelihood-ratio statistics, and the transition probability matrices for each of the segments of the sequence specified by the user. The output for both order and stationarity statistics also includes an analysis of zero- and low-probability values encountered in the computations for identifying possible artificial inflation of the statistics.

Limitations. Up to 99 behavior sequences, each containing up to 3,000 time points, can be processed in a single run, with each sequence subject to all processing described above. No sequence may contain more than 15 distinct behavior states. Convergence testing will stop at the 50th power if convergence is not reached. Distribution construction encompasses no more than 125 different time lengths per state, although more may be present in the data. Users may specify from two to eight segments of the sequence for calculating stationarity statistics.

Program TEST

Program description. As specified by the user, program TEST will (1) calculate goodness-of-fit and likelihood-ratio statistics for comparing any one transition probability matrix (the expected values) with any other (the observed values), (2) calculate both types of statistics for testing the homogeneity among a set of matrices, and (3) call a user-supplied FORTRAN subprogram to perform special operations on a set of matrices before proceeding to the tests specified in (1). The program is supplied with code for averaging a set of matrices and treating the results as the observed values.

Input. The program is designed for batch-mode operation. The user provides the number of data sets contained in the group to be processed, together with the number of distinct states and their codes (any machine-readable code is acceptable), all on two initializing control cards or lines. These are followed by the cards or lines containing the group of data sets (one data set consists of a transition frequency matrix and its associated probability matrix). Matrices punched by program SAMPLE follow the required input format. Each data set is identified by two optional numerical descriptors, by the number of states in the matrix, and by an optional descriptive title. On subsequent control cards (which are unlimited in number), the user specifies the desired processing: If pairs of matrices are to be compared, the user specifies only the sequence numbers of the data sets to be treated as expected and observed; if homogeneity statistics are desired, the user specifies the sequence numbers of the matrices to be included.

Output. The output includes the number of data sets entered and, for each set, a listing of all descriptive information and of the transition frequency and probability matrices. If the user specifies comparison of two matrices, the output identifies the expected and observed data sets, provides degrees of freedom and goodness-of-fit and two likelihood-ratio statistics, and includes both an analysis of the goodness-of-fit statistic by rows and the difference in transition frequencies between the expected and observed. Specifying homogeneity statistics provides a listing of the identifying information for the data sets included, the degrees of freedom, the goodness-of-fit and likelihood-ratio statistics, and the master transition probability matrix used in the computations. Both outputs also include an analysis of zero- and low-probability values encountered, as in program SAMPLE.

Limitations. Up to 99 separate groups, each consisting of 2 to 20 data sets, can be processed in a single run, with each group subject to all processing described above. No matrix may exceed 15 x 15 states, and all matrices in one group must have the same number of states. Homogeneity statistics must be computed on at least 2 and no more than 20 data sets.

Program Specifications

SAMPLE and TEST are written in standard FORTRAN IV and are shipped in their IBM 360 versions. Modifications to one or two statements have permitted installation on CDC, UNIVAC, and Honeywell systems. Both programs are formatted so that output may be viewed conveniently on a CRT device, as well as in standard listings.

Availability

SAMPLE and TEST, together with test data for each, are available as listings, on cards, or on nine-track tape. Documentation is available separately. All may be obtained at cost of reproduction and shipping from Robert B. Arundale, Department of Speech and Drama, University of Alaska, Fairbanks, Alaska 99701.

REFERENCES

Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about Markov chains. *Annals of Mathematical Statistics*, **28**, 89-110.

Arundale, R. B. (1977). Sampling across time for communication research: A simulation. In P. M. Hirsch, P. V. Miller, & F. G. Kline (Eds.), *Strategies for communication research*, Beverly Hills, CA: Sage.

Hewes, D. E. (1975). Finite stochastic modeling of communication processes: An introduction and some basic readings. *Human Communication Research*, **1**, 271-283.

Hewes, D. E. (1980). Stochastic modeling of communication processes. In P. R. Monge & J. N. Cappella (Eds.), *Multivariate techniques for communication research*. New York: Academic Press.

(Manuscript accepted for publication June 8, 1984.)