ANALYZING FOR INDEPENDENCE OF SEQUENTIAL EVENTS WITH MARKOV CHAIN MODELS IN SAS® (vers. 6)

Paul F. Steblein, Syracuse University

INTRODUCTION

Testing for the independence of sequential events is a useful analysis for many disciplines. An event is defined as a discretely identifiable occurrence or incident. Examples of events might be behavioral acts in sociology (Morgan 1976) or succession of plant communities in ecology (Usher 1979). Markov chain models have been suggested as an appropriate approach for identifying patterns in sequential events. The objective of this paper is to describe how to use Markov chain models with SAS (vers. 6) to test for independence of sequential events. An example is provided that employs the technique to examine patterns of behavior in birds; specifically, the detection of sequences in maintenance behavior by nestling broad-winged hawks (*Buteo platypterus*).

METHODS

MARKOV CHAINS

There are situations where one type of event always follows another event (e.g., a heavy object always falls when it is released). When the outcome of event transition can be predicted with certainty, it is a relatively simple matter to construct a web or pattern of event change for that system. However, in many situations the transition of events is not deterministic, but rather discrete-time random processes which can be modeled as Markov chains (Fagen and Young 1978). If the occurrence of two events is truly random, the transition is a zero order Markov chain. Whereas, a first order Markov chain would be defined as an event that is only dependent on the event immediately preceding it. This is the type of relationship which is most interesting (and easily tested). Second order Markov chains are described by cases where a sequence of two events defines the outcome of a third event.

The conditional probability \( p_{ij} \) that an event \( E_j \) will be followed by another event \( E_i \) in a suite of possible outcomes (or repertoire - R) can be calculated for a first order Markov transition as follows:

\[
p_{ij} = P[E_j(n)|E_i(n-1)] \quad \text{for } 1 < i < R, 1 < j < R
\]

where \( n \) is the observation of an event (Fagen and Young 1979). The transition probabilities \( p_{ij} \) can be summarized in a transition matrix where all possible initial events \( E_i \) in a two-event sequence are listed as rows, and all subsequent events \( E_j \) are represented as columns. Thus, a cell value identified by its row \( (i) \) and column \( (j) \) notes the probability of that transition occurring.

\[
\begin{array}{lll}
& 1 & 2 & 3 & \ldots & R \\
1 & p_{11} & p_{12} & p_{13} & \ldots & p_{1j} \\
2 & p_{21} & p_{22} & p_{23} & \ldots & p_{2j} \\
3 & p_{31} & p_{32} & p_{33} & \ldots & p_{3j} \\
& \vdots & \vdots & \vdots & \ddots & \vdots \\
R & p_{R1} & p_{R2} & p_{R3} & \ldots & p_{Rj}
\end{array}
\]

Furthermore, the column and row totals should each equal one, since events are mutually exclusive and all possible sequential events are represented. Depending on the objectives of the study, definition of events should preclude the possibility of homogenous transitions (i.e., where an event is repeated), inasmuch as the "change" is the focus. Under this circumstance, the diagonal elements of the transition matrix would have the value of zero and could be collapsed to simplify the matrix.

Hypotheses concerning the appropriate order Markov Chain model and transition probabilities can be statistically examined using a Chi-square statistic \( (X^2) \) to compare observed and expected transition frequencies. The formula (Fagen and Young 1978) utilized to evaluate the presence of a first order Markov Chain process is as follows:
\[ X^2_{\text{calc}} = \sum_{i=1}^{R} \sum_{j=1}^{R} \frac{(x_{ij} - m_{ij})^2}{m_{ij}} \]

and individual transitions are evaluated by:

\[ Y = \frac{(x_{ij} - m_{ij})}{(m_{ij})^{1/2}} \]

where \( x_{ij} \) is the observed frequency of element \((i,j)\) of the transition table, and \( m_{ij} \) is the expected frequency (row sum \( i \) x column sum \( j \)/grand sum). The individual transition is significant if

\[ |Y| > \sqrt{(X^2 / \text{d.f.})/R^2} \]

IMPLEMENTATION IN SAS (vers. 6)

A program was developed in SAS (vers. 6) to test for first order Markov Chains. The program is organized to first read the data as multiple series of sequential events to tally the frequency of specific transitions. Each line of data is read into an array for temporary storage. Because the event data is coded as a series of whole numbers (1, 2, 3, ...), they can be used as a reference to increment by one the appropriate element of a multidimensional array (i.e., transition table). Another datastep is then used to calculate the significance of individual first-order transitions and an overall test for first order Markov processes.

To use the program, a user must first calibrate the model for the number of possible events \( R \), the maximum number of events in an observational bout, and the degrees of freedom. Expansion of the program to analyze for second order Markov Chains is discussed later in the paper.

AN EXAMPLE OF MARKOV CHAIN APPLICATION IN ANIMAL BEHAVIOR

Behavioral observations were recorded for nestling and adult broad-winged hawks at nests in the central Adirondack Mountains of New York State during the summer of 1987. Five hours of observations were recorded on video tape from nearby tree blinds for detailed analysis of behavioral patterns and to document the different types of behavior. Nestling behavior was coded from the tape as a sequence of mutually exclusive behavioral acts. For example, an observational bout might include the following sequence:

- preen breast
- tail wag
- head shake
- preen breast;

and would be coded for data entry as follows:

1 -> 2 -> 3 -> 1

The original data was recorded with a repertoire consisting of more than 30 distinguishable behaviors, but was aggregated to comply as well as possible with the restrictions of the analysis. Nine individual behaviors were used for the analysis, which included: preening; scanning; positioning wing; reposition body; head scratch; stretching; tail wagging; shaking; and miscellaneous behaviors (calling, feeding, sleeping, and aggression).

RESULTS AND DISCUSSION

ANALYSIS OF ANIMAL BEHAVIOR EXAMPLE

Scanning was clearly the most common behavior nestling broad-winged hawks, followed by preening, wing positioning, repositioning the body, and tail wagging. Preening at this stage in the nestlings is common because the development of juvenile plumage.

First order Markov Chains tested significantly \( (X^2 = 1707.210, \ p < 0.000) \) in this example. However, caution should be used in interpreting the results because of the assumptions in this technique (see below). Fourteen sequences were positive (obs > exp) and had highly significant transitions \( Y > 3.0 \). These sequences were used to develop a flowchart to show common patterns in behavioral sequences for nestling broad-winged hawks (Fig. 1). The importance of scanning in the sequence is well illustrated here, since most behaviors are connected to it. The most common sequence observed could be described by a pattern of preening, positioning the wings, looking around, and continuing the preening. Tail wagging frequently followed a facial scratch or moving around the nest. There were also shorter behavior chains (e.g., scanning to stretching, or scan-stretch-scan). This approach towards analyzing behavior is being used in another study to compare nestling and adult behavior to examine developmental patterns.
CONSIDERATION OF SAMPLE SIZE AND ASSUMPTIONS

Researchers applying Markov Chains to analyses of sequential events should be highly attentive of the sample size and assumptions. Fagen and Young (1976) recommend a sample size equivalent to $10^R\cdot R$, where $R$ is the suite of all possible events. For example, if $R = 5$ the minimum sample size should be 250, or if $R = 10$ then 1000 observations should be recorded. Obviously the sample size required escalates quickly as the complexity of possible events increases. Fagen and Young (1976) further state that slightly less observations ($7.3R^2$) may be just adequate, and that $5R^2$ observations are borderline for statistical analyses.

There are several other restrictions for the valid application of Chi-square analysis to Markov Chains (Fagen and Young 1976). First, none of the expected values should be less than one. Furthermore, cells with expected values of less than five should not exceed 20% of the total number of cells. Cells with low values may be pooled if appropriate. In the example presented here, 43% of the cells contained expected values less than five even though the sample size ($N = 1673$) was more than twice the amount required for Markov Chains ($10^9\cdot 9^2 = 810$). This points out the difficulty of achieving an appropriate sample size when there are a large number of rare sequences.

Markov models also assume that the transition probabilities are stationary (i.e., independent of the observation period or position in the sequence). This may be an unreasonable expectation in some cases. However, separate transition matrices (and subsequent analyses) may be created for each group.

HIGHER ORDERED MARKOV CHAINS

The paper has been oriented towards analysis of first order Markov chains because that is a frequent and easily tested transition dependence. However, second and third order dependencies can be tested for within the limits of Chi-square goodness of fit analyses (Chatfield and Lemon 1970). Extending the transition matrix for second order tests requires the addition of another dimension for the second preceding event (n-2). The SAS (vers. 6) program presented here can be easily modified for accommodation of third dimension in the matrix, but sample size and restrictions for expected values in Chi-square analysis will quickly limit its usefulness.

Inquiries on the program can be addressed by writing to: Paul F. Steblein, Research Consulting Services, Academic Computing Services, Syracuse University, Syracuse, NY 13244-1260; or on BITNET (PFSTEBLE@SUVM).
**LITERATURE CITED**


SAS is a registered trademark of SAS Institute Inc., Cary, NC, USA.

**APPENDIX**

`/* ANALYSIS OF 1ST ORDER MARKOV CHAINS (5 EVENTS) WITH SAS (VERS. 6)*/`

DEVELOPED BY:
PAUL F. STEBLEIN
Research Consulting Services
ACADEMIC COMPUTING SERVICES
SYRACUSE UNIVERSITY
SYRACUSE, NY 13244

DATA COUNT records the number of event transitions within observation periods. Observations are read in and assigned to an array and increments the appropriate event seq. by one for each seq. observed. The loop should be set for max. number observations in a record minus one (20-1=19).

```
DATA COUNT;
ARRAY ACT [5,5] ACT11 - ACT15
ACT21 - ACT25
ACT31 - ACT35
ACT41 - ACT45
ACT51 - ACT55;
R=5;
DO I = 1 TO R;
    DO J = 1 TO R;
        ACT[I,J]=0;
    END;
END;
```

INPUT (X1 - X19) (2.);
ARRAY BOUT[19] X1 - X19;
DO I = 1 TO 31;
   J = I+1; A = BOUT[I];
   B = BOUT[J];
   IF B = . THEN RETURN;
   ACT[A,B] = ACT[A,B] + 1;
END;
CARDS;
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
RUN;

PROC MEANS calculates total frequency of sequential events, i.e., generates values for the transition matrix.

```
PROC MEANS SUM;
VAR ACT11 - ACT15
ACT21 - ACT25
ACT31 - ACT35
ACT41 - ACT45
ACT51 - ACT55;
OUTPUT OUT=TEMPFREQ
SUM = ACT11 - ACT15
ACT21 - ACT25
ACT31 - ACT35
ACT41 - ACT45
ACT51 - ACT55;
TITLE 'OBSERVED FREQUENCIES OF BEHAVIORAL SEQUENCES';
RUN;
```

DATA CELLCHI calculates the probability statistic for individual event transitions and overall presence of first order Markov Chain. The data step should be calibrated similar to above for the particular application.

```
DATA CELLCHI;
SET TEMPFREQ;
R=5;
ARRAY ACT [5,5] ACT11 - ACT15
ACT21 - ACT25
ACT31 - ACT35
ACT41 - ACT45
ACT51 - ACT55;
ARRAY ROW [5] ROWTOT1 - ROWTOT5;
ARRAY COL [5] COLTOT1 - COLTOT5;
ARRAY CELLCHI [5,5] CELCHI11 - CELCHI15
CELCHI21 - CELCHI25
CELCHI31 - CELCHI35
CELCHI41 - CELCHI45
CELCHI51 - CELCHI55;
DO I = 1 TO 5;
    END;
DO I = 1 TO 5;
    END;
```
DO J = 1 TO 5;
            ACT[4,J] + ACT[5,J];
END;

DF = (R - 1) * (R - 1);
DO I = 1 TO 5;
   DO J = 1 TO 5;
      MIJ = (ROW[I] * COL[J]) / GRANDSUM;
      CELLCHI[I,J] = ((ACT[I,J] - MIJ) / SQRT(MIJ));
      Y = CELLCHI[I,J];
      CHIVALUE = (CELLCHI[I,J] * R * R)**2;
      PROB = 1 - PROBCHI(CHIVALUE, DF);
      PARTCHI = ((SEQ[I,J] - MIJ)**2) / MIJ;
      TOTCHI = TOTCHI + PARTCHI;
      ACTONE = I; ACTTWO = J; FREQ = SEQ[I,J];
   END;
END;
RUN;
PROC PRINT;
VAR ACTONE ACTTWO FREQ MIJ Y PROB TOTCHI;
TITLE 'Individual (Y) and Overall (TOTCHI) Chi-square Values';
TITLE 'For Analyses of First Order Markov Chains';
RUN;

1023