7

Analyzing event sequences

7.1 Describing particular sequences: Basic methods

Throughout this chapter, we assume that the reader is interested in event-sequence data. That is, no matter how the data may have been recorded and represented initially, we assume that it is possible to extract event sequences (see section 5.2) from the data, and that the investigator has good reasons for wanting to do so. This means that the data to be analyzed are represented as sequences or chains of coded events (or behavioral states but without time information) and that those events are defined in a way that makes them mutually exclusive and exhaustive. Sometimes the chains will be unbroken, collected all during one uninterrupted observation session. Other times, several sequences may be pooled together for an analysis, either because there were breaks in the observation session or because observation sessions occurred at different times. Data from more than one subject may even be pooled for some analyses (see section 8.4). In all cases, the data to be analyzed consist of chains or sequences of codes.

The codes for event sequences are mutually exclusive and exhaustive, as already noted. In addition, often the logic of the situation does not permit consecutive codes to repeat. For example, when coders are asked to segment the stream of behavior into behavioral states, it follows naturally that two successive states cannot be coded the same say. If they were, they would not be two states, after all, but just one. The restriction that the same code may not follow itself in event sequences occurs relatively often, especially when the events being coded are thought of as states. This restriction affects how some statistics, especially expected frequencies, are computed, as we discuss later.

As an example, let us again assume the coding scheme used in the Bakeman and Brownlee study of parallel play. Behavioral states were classified as one of five kinds: Unoccupied, Solitary, Together, Parallel, and Group. When sequences were analyzed, adjacent codes were not allowed to be identical. Thus there were 20 (or 5×4) different kinds of two-event sequences (not 5^2, which would be the case if adjacent codes could be the

same); 80 (or 5×4^2) different kinds of three-event sequences (not 5^3); 320 (or 5×4^3) different kinds of four-event sequences (not 5^4); etc.

Determining how often particular two-event, three-event, etc., sequences occurred in one's data is what we mean by "basic methods." This involves nothing more than counting. The investigator simply defines particular sequences, or all possible sequences of some specified length, and then tallies how often they appear in the data. For example, Bakeman and Brownlee were particularly interested in transitions from Parallel to Group play. For one child in their study, 127 two-event sequences were observed, 10 of which were from Parallel to Group. Thus they could report that, for that child, f(PG) = 10 and p(PG) = .079 (10 divided by 127). (Note that p(PG) is not a transitional probability. It is the simple or zero-order probability for the two-event sequence, Parallel to Group.) In sum, the most basic thing to do with event-sequence data is to define particular sequences, count them, and then report frequencies and/or probabilities for those sequences.

7.2 Determining significance of particular chains

We might now ask, how should these values be evaluated? One possibility would be to compute expected frequencies, based on some model, for particular observed frequencies, and then compare observed and expected with a chi-square goodness-of-fit test. For example, there are 20 different kinds of two-event sequences possible (U to S, U to T, U to P, U to G, S to U, S to T, etc.). Thus, we might argue, the expected probability for any one kind is .05 (1/20), and so the expected frequency in this case is 6.35 (.05 × 127). What we are doing is assuming a particular model – in this case, a "zero order" or "equiprobable" model, so called because it assumes that the five codes occur with equal probability – and then comparing the expected values the model generates for a particular sequence with those actually observed. We note that the observed value for the Parallel to Group sequence, 10, is greater than the expected value, 6.35. If we were only concerned with the Parallel to Group sequence, we might categorize all sequences as either Parallel to Group (10) or not (117), and compare observed to expected using the familiar Pearson chi-square statistic,

$$X^2 = \sum \frac{(obs - exp)^2}{exp}$$
$$= \frac{(10 - 6.35)^2}{6.35} + \frac{(117 - 120.65)^2}{120.65} = 2.21$$

which, with one degree of freedom, is not significant (we use a Roman X to represent the computed chi-square statistic to distinguish it from a Greek chi, which represents the theoretical distribution).

Alternatively, we might make use of what we already know about how often the five different codes occurred. This "first-order" model assumes that codes occurred as often as they in fact did (and were not equiprobable), but that the way codes were ordered was determined randomly. For the child whose data we are examining, 143 behavioral states were coded; 34 were coded Parallel and 30 Group. (Because 127 two-event sequences were tallied, and 143 states were coded, there must have been 15 breaks in the sequence.) Now, if codes were indeed ordered randomly, then we would expect that the probability for the joint event of Parallel followed by Group would be equal to the simple probability for Parallel multiplied by the simple probability for Group (this is just basic probability theory). Symbolically,

$$\mathrm{p}(PG)_{exp} = \mathrm{p}(P) \times \mathrm{p}(G)$$

The p(P) is .238 (34/143, the frequency for Parallel divided by the total, N).

In this case, however, the p(G) is not the f(G) divided by N. Because a Parallel state cannot follow a Parallel state, the probability of group (following Parallel) is computed by dividing the frequency for Group, not by the total number of states coded, but by the number that could occur after Parallel – that is, the total number of states coded, less the number of Parallel codes. Symbolically.

$$\mathrm{p}(G) = \frac{\mathrm{f}(G)}{N - \mathrm{f}(P)}$$

when adjacent codes must be different and when we are interested in the expected probability for Group following Parallel. Now we can compute the expected probability for the joint event of a Parallel to Group transition. It is:

$$\mathrm{p}(PG)_{exp} = \frac{\mathrm{f}(P)}{N} \times \frac{\mathrm{f}(G)}{N - \mathrm{f}(P)} = \frac{34}{143} \times \frac{30}{143 - 34} = .0654$$

The expected frequency, then, is 8.31 (.0654, the expected probability for this particular two-event sequence, times 127, the number of two-event sequences coded).

The chi-square statistic for this modified expected frequency is

$$X^2 = \frac{(10 - 8.31)^2}{8.31} + \frac{(117 - 118.69)^2}{118.69} = 0.368$$

which likewise is not statistically significant. Neither analysis suggests that Group is any more likely to follow Parallel than an equiprobable or first-order independence model would suggest.

The methods presented in this section are fairly limited. First, as sequences become longer, the number of possible sequences increases exponentially. For example, with just five codes when consecutive codes cannot repeat, there are 20 (5×4) two-event sequences, 80 ($5 \times 4 \times 4$) three-event sequences, 320 ($5 \times 4 \times 4 \times 4$) four-event sequences, etc. Consequently, expected probabilities for any one sequence may become vanishingly small, requiring staggering amounts of data before expected frequencies become large enough to evaluate with any confidence. Second, rarely are investigators interested in just one particular sequence such as the Parallel to Group sequence used here as an example. More general methods, described in subsequent sections and chapters, are required.

7.3 Transitional probabilities revisited

In the last section, we presented data derived from observing one child in the Bakeman and Brownlee study of parallel play. We noted that her event-sequence data contained 127 two-event sequences and that 10 of them represented transitions from Parallel to Group play. Thus we were able to say that p(PG), the probability of a Parallel to Group sequence, was .0787, or 10 divided by 127. We then discussed ways of determining whether this observed probability differed significantly from expected. We also noted that p(PG) was a simple, not a transitional probability. In other words, p(PG) is the probability for this particular sequence; if the probabilities for all 20 possible two-event sequences were summed, they would add up to one.

Occasionally it may be useful to describe probabilities for particular sequences, no matter whether chains are two-event, three-event, or longer. When longer sequences are considered (e.g., 4 or 5 events long instead of just 2), the number of possible sequences increases exponentially. When there are many possible sequences, probabilities for particular sequences can become almost vanishingly small and, as a result, less useful descriptively. Thus usually attention focuses on transitional probabilities involving two events. As discussed in section 6.5, these are usually symbolized t and, unless noted otherwise, refer to lag 1.

For example, consider again the child in the parallel play study. (frequencies and simple and transitional probabilities derived from her data are given in Tables 7.1 through 7.3.) Considering just simple probabilities, we note that the probability of a Parallel to Group sequence, p(PG), was .0787, whereas the probability of a Parallel to Unoccupied sequence,

Table 7.1. *Observed frequencies for two-event sequences*

Given code,	Target code, lag 1					
lag 0	Un.	Sol.	Tog.	Par.	Gr.	Totals
Unoccupied	—	6	5	2	2	15
Solitary	5	—	6	7	5	23
Together	5	6	—	12	10	33
Parallel	2	7	11	—	10	30
Group	2	4	11	9	—	26
Totals	14	23	33	30	27	127

Table 7.2. *Simple probabilities for two-event sequences*

Given code,	Target code, lag 1				
lag 0	Un.	Sol.	Tog.	Par.	Gr.
Unoccupied	—	.0472	.0394	.0157	.0157
Solitary	.0394	—	.0472	.0551	.0394
Together	.0394	.0472	—	.0945	.0787
Parallel	.0157	.0551	.0866	—	.0787
Group	.0157	.0315	.0866	.0709	—

Note: The tabled probabilities do not sum exactly to 1 because of rounding.

p(PU), was .0157. This certainly conveys the information that Group was more common after Parallel than Unoccupied. But somehow it seems both clearer and descriptively more informative to say that the probability of Group, given a previous Parallel, p($G|P$) or t_{PG}, was .333, whereas the probability of Unoccupied, given a previous Parallel, p($U|P$) or t_{PU}, was .067. Immediately we know that 33.3% of the events after Parallel were Group, whereas only 6.7% were Unoccupied.

We just considered transitions from the same behavioral state (Parallel) to different successor states (Group and Unoccupied), but the descriptive value of transitional probabilities is portrayed even more dramatically when transitions from different behavior states to the same successor state are compared. For example, the simple probabilities for the Unoccupied to Solitary, p(US), and for the Together to Solitary, p(TS), transitions are both .0472. Yet the probability of Solitary, given a previous Unoccupied, p($S|U$) or t_{US}, is .400, whereas the probability of Solitary, given a previous Together, p($S|T$) or t_{TS}, is .182. The transitional probabilities "correct"

Table 7.3. *Transitional probabilities for two-event sequences*

Given code,	Target code, lag 1				
lag 0	Un.	Sol.	Tog.	Par.	Gr.
Unoccupied	—	.400	.333	.133	.133
Solitary	.217	—	.261	.304	.217
Together	.152	.182	—	.364	.303
Parallel	.067	.233	.367	—	.333
Group	.077	.154	.423	.346	—

Note: Rows may not add to 1 because of rounding.

for differences in base rates for the "given" behavioral states and, therefore, clearly reveal that in this case, Solitary was relatively common after Unoccupied, and considerably less so after Together, even though the Unoccupied to Solitary and the Together to Solitary transitions appeared the same number of times in the data.

Moreover, as already noted in section 6.5, transitional probabilities form the basis for state transition diagrams, which, at least on the descriptive level, are a particularly clear and graphic way to summarize sequential information. The only problem is that, even with as few as five states, the number of possible arrows in the diagram can produce far more confusion than clarity. The solution is to limit the number of transitions depicted in some way. In this case, for example, we could decide to depict only transitional probabilities that are .3 or greater, which is what we have done in Figure 7.1. This reduces the number of arrows in the diagram from a possible 20, if all transitions were depicted, to a more manageable 9.

The nine transitions shown in Figure 7.1 are not necessarily the most frequent transitions; this information is provided by the simple probabilities for two-event sequences (see Table 7.2). Nor are the transitions necessarily significantly different from expected; to determine this we would need to compute and evaluate a z score for each transition (see next section). What the state transition diagram does show are the most likely transitions, taking the base rate for previous states into account. In other words, it shows the most likely ways of "moving" from one state to another. For this one child, Figure 7.1 suggests frequent movement from Unoccupied to both Solitary and Together, from Solitary to Parallel, and reciprocal movement among Together, Parallel, and Group.

One final point: Transitional probabilities can be used to describe relationships between two nonadjacent events as well. Not only can we compute, for example, the probability of Group in the lag 1 positions given an immediately previous Parallel: $p(G_{+1}|P_0)$, but we can also compute the

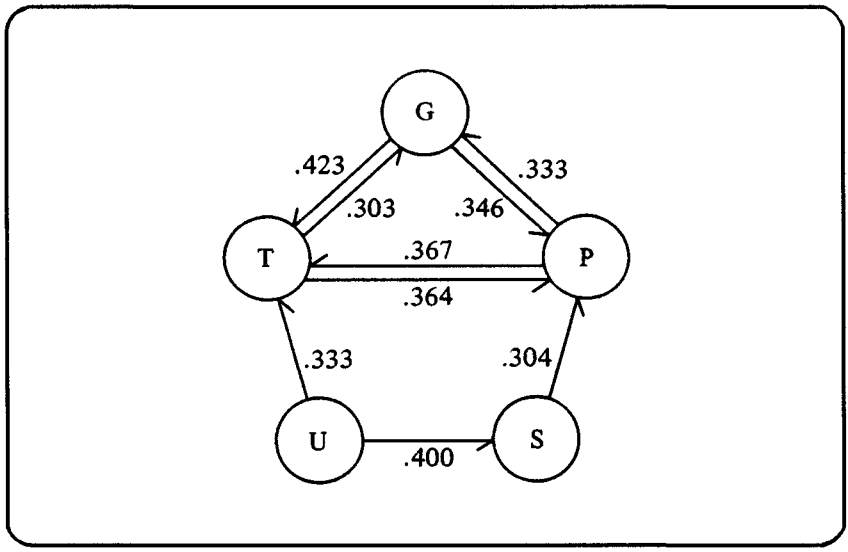

Figure 7.1. A state transition diagram. Only transitional probabilities greater than 0.3 are shown. U = unoccupied, S = solitary, T = together, P = parallel, and G = group.

probability of Group in the lag 2 position: $p(G_{+2}|P_0)$, the lag 3 position, etc., with as many events intervening between the "given" and the "target" code as desired. The ability of transitional probabilities to describe such lagged relationships, in fact, forms the basis for the lag sequential method described in section 7.5.

Transitional probabilities, although useful descriptively, also have limitations. In particular, when several "subjects" take part in a study, and when transitional probabilities are computed separately for each subject, those transitional probabilities should not, under most circumstances, be used as scores for testing for individual or group differences.

The reason is as follows: A transitional probability is valuable descriptively, to be sure, and certainly reflects the "experience" of a particular subject. For example, the experience of the child whose data are given in Table 7.3 was that one-third of the time after Parallel, Group followed. This may be important, but it may or may not be "significant." It depends on how probable Group was for the child. If Group was especially probable for one child, then a score of .33 for the transitional probability might not be very high; it could even be less than expected. On the other hand, if Group was unlikely for another child, then a score of .33 might be quite high, considerably above expected.

The major problem, then, with using transitional probabilities when analyzing for individual or group differences, is that similar numeric values may have quite different meanings for different subjects, or for the same subject at different times, rendering the results difficult to interpret at best. A secondary problem is that the values of transitional probabilities are "contaminated" with the values for the simple probabilities, which means that what appear to be analyses of transitional probabilities may in fact reflect little more than simple probability effects.

An example may help to clarify this. Imagine that we have two kinds of children, farm kids and city kids, and that the mean value for the simple probability of Group computed for the city kids is significantly higher than the mean value computed for farm kids. If we then turn around and test the transitional probability, $p(G_{+1}|P_0)$, we would expect that mean values for city and farm kids would differ significantly as well. After all, the expected value for this transitional probability is directly related to the simple probability for Group. The more often the Group code appears in the data, the more often we would expect it to follow the Parallel code as well.

Symbolically, what we are saying is that the expected value for $p(T|G)$ – where T stands for "target" and G for "given" – is directly related to $p(T)$, the probability of the target code. It is not necessary, of course, that observed agree with their expected values. Still, when analyses of both $p(T|G)$ and $p(T)$ reveal significant group differences, it seems unjustified to us to claim that the differences with respect to $p(T|G)$ are explained by anything more complex than the differences detected for $p(T)$. In sum, for both reasons given above, analyses that use transitional probabilities as scores rarely provide much insight into sequential aspects of the data.

Almost always, when individual or group differences are at issue, the appropriate score to use is not the transitional probability, but some index of the strength of the effect, like Yule's Q (as discussed in section 7.7). Analyzing such scores, we would be able to determine, for example, whether the Parallel to Group sequence was significantly more characteristic of city kids, on the average, than of farm kids, or whether the extent to which Group tended to follow Parallel was a stable characteristic of children measured at two different ages.

In the previous edition of this book, we suggested that z scores (see next section) be used as an individual index for such analyses, but the magnitude of z scores is affected by the number of tallies. For an effect of a specific size, the z score becomes larger as the number of tallies increases. This is an example of the well-known fact that power increases with sample size. But it also makes the z score an inappropriate choice for analyzing individual differences. Unless the number of tallies is the same for all

subjects (dyads, etc.) analyzed, z scores conflate how strongly a particular transitional probability deviates from its expected value with the number of tallies available for the analysis.

7.4 Computing z scores and testing significance

Once lagged events have been tallied in a table like Table 7.1, where rows represent lag 0 and columns lag 1, and after we have examined transitional probabilities like those in Table 7.3, next we usually want to identify which transitional probabilities deviate significantly from their expected values. Commonly, a z score of some sort has been used for this purpose. Assuming that a computed score is distributed normally, then z scores larger than 1.96 absolute are often regarded as statistically significant at the .05 level. But there are a variety of ways to compute z scores, and simply calling one a z score hardly guarantees that it will be normally distributed.

Assume that the behavioral event of interest is called the "target" event, T, and that we want to relate that to another event, called the "given" event, G. In other words, we are interested in $p(T_1|G_0)$ or t_{GT}, the probability of the target even occurring after the given event.

A z score compares observed to expected, so the first task is to compute the expected value for x_{GT}, the observed value for the transition from given to target behavior. When consecutive codes may repeat (and so the upper-left to lower-right diagonal tallies would not all be zero as in Table 7.1), expected values are computed using the familiar formula,

$$m_{GT} = \frac{x_{+T}}{x_{++}} x_{G+} = \frac{x_{G+} x_{+T}}{x_{++}} \tag{7.1}$$

where m_{GT} is an estimate of the expected frequency (m because often expected values are means), x_{G+} is the sum of the observed frequencies in the Gth or given row, x_{+T} is the sum of the observed frequencies in the Tth or target column, and x_{++} is the total number of tallies in the table (also symbolized as N_2 or the number of two-event chains tallied, as compared to N_1, the number of single events coded). This formula yields expected values assuming independence, that is, no association between the rows and columns of the table.

However, when consecutive codes cannot repeat, resulting in what are called structural zeros on the diagonal (*structural* because logical definition and not data collection resulted in their being zero), expected frequencies cannot be computed with a simple formula but require an iterative procedure, best performed with a computer. Two of the most widely used are iterative proportional fitting (IPF, also called the Deming–Stephan algorithm)

Table 7.4. *Expected frequencies for two-event sequences*

Given code,	Target code, lag 1				
lag 0	Un.	Sol.	Tog.	Par.	Gr.
Unoccupied	—	2.82	4.69	4.04	3.45
Solitary	2.65	—	7.83	6.75	5.76
Together	4.40	7.83	—	11.21	9.56
Parallel	3.80	6.75	11.21	—	8.24
Group	3.14	5.59	9.27	8.00	—

and the Newton–Raphson algorithm; for descriptions see Bishop, Fienberg, & Holland (1975) and Fienberg (1980). Both methods yields identical values. Expected frequencies for the data shown in Table 7.1 are given in Table 7.4.

When consecutive codes may repeat, z scores are computed as follows:

$$z_{GT} = \frac{x_{GT} - m_{GT}}{\sqrt{m_{GT}(1 - p_{G+})(1 - p_{+T})}} \tag{7.2}$$

where p_{G+} is $x_{G+} \div x_{++}$ and p_{+T} is $x_{+T} \div x_{++}$. In the log-linear literature, this is called an adjusted residual (Haberman, 1978, p.111). When consecutive codes cannot repeat, matters are more complex. The formula (actually several formulas, many of which are used by the Newton–Raphson algorithm) is given in Haberman (1979, p. 454; but see footnote 1 in Bakeman & Quera, 1995b); a number of relatively complex matrix operations are involved. The most practical way for you to compute these values, given structural zeros, is to use a computer program like GSEQ or a general-purpose log-linear program. Adjusted residuals, computed with the SPSS for Windows General Log-Linear routine, are given in Table 7.5.

If you are new to sequential analysis, you may want to skip the following paragraphs, which are included largely for historical purposes and for readers who wish to reconcile the preceding paragraphs with the first edition of this book and with earlier literature. Early on, Sackett (1979) suggested that z be computed as follows:

$$z_{GT} = \frac{x_{GT} - m_{GT}}{\sqrt{m_{GT}(1 - p_T)}} \tag{7.3}$$

where

$$m_{GT} = \mathrm{p}(T) \times \mathrm{f}(G) = \frac{x_T}{N_1} x_G = \frac{x_G x_T}{N_1} \tag{7.4}$$

which is almost but not quite the same as Equation 7.1 because it is based

Table 7.5. *Adjusted residuals for two-event sequences*

Given code,	Target code, lag 1				
lag 0	Un.	Sol.	Tog.	Par.	Gr.
Unoccupied	—	2.25	0.19	−1.29	−0.96
Solitary	1.71	—	−0.94	0.13	−0.42
Together	0.37	−0.94	—	0.37	0.22
Parallel	−1.17	0.13	−0.10	—	0.88
Group	−0.79	−0.88	0.85	0.51	—

Note: Row and column totals may not add exactly to those shown in Table 7.1 because of rounding.

on simple frequencies and probabilities for given and target behaviors, not on values from two-dimensional tables as is Equation 7.1. Equation 7.3 is based on the normal approximation for the binomial test,

$$z = \frac{x - NP}{\sqrt{NPQ}} \tag{7.5}$$

where N is f(G) or x_G, P is p(T) or $x_T \div N_1$ (also symbolized as p_T), and Q is $1 - p_T$. Almost immediately, however, Allison and Liker (1982) objected, noting that Equation 7.5 would only be appropriate if p_T were derived theoretically instead of from the data at hand. They wrote that Equation 7.3 should be

$$z_{GT} = \frac{x_{GT} - m_{GT}}{\sqrt{m_{GT}(1 - p_G)(1 - p_T)}} \tag{7.6}$$

instead, which is almost but not quite the same as Equation 7.2 because, like Equation 7.3, it is based on single occurrences and not two-event chains.

We prefer Equation 7.2 to 7.6 because it seems more grounded in a well-developed statistical literature, that dealing with log-linear models (e.g., see Bishop, Fienberg, & Holland, 1975; Fienberg, 1980; Wickens, 1989), and because it is based on two-event chains, which seems more faithful to the situation at hand. True, if a sequence of N_1 consecutive events is tallied using overlapped sampling (i.e., tallying first the e_1e_2 chain, then e_2e_3, e_3e_4, and so forth, where e stands for an event), so that $N_1 - 1$ chains are tallied, then x_G and x_{G+}, for example, will differ by at most 1. But overlapped sampling, while common, is not always used; moreover, often breaks occur in sequences and then the number of two-event chains tallied is $N_1 - S$, where S is the number of separate segments coded. In such cases, x_G and x_{G+} could differ by quite a bit.

Moreover, the log-linear tradition, from which Equation 7.2 is derived, offers a statistically based solution when consecutive codes cannot repeat (see Bakeman & Quera, 1995b). Sackett (1979), recognizing that Equation 7.1 would not compute expected frequencies correctly when structural zeros occupied the diagonal, suggested

$$m_{GT} = \frac{x_T}{N_1 - x_G} x_G = \frac{x_G x_T}{N_1 - x_G}. \tag{7.7}$$

He reasoned that when consecutive events cannot repeat, the expected probability for the target code at lag 1 (assuming independence) is the frequency for that code divided by the number of events that may occur at lag 1, which is N_1 minus the frequency for the given code. Thus expected frequencies on the diagonal are set to zero and off-diagonal ones are the frequency for the given code times this probability, as indicated by Equation 7.7. However, expected frequencies, when summed across rows and down columns, should equal the observed row and column totals, which expected frequencies computed per Equation 7.7 do not (Bakeman & Quera, 1995b), whereas expected frequencies computed with an iterative procedure do (see Table 7.4). Thus, as mentioned earlier, when consecutive codes cannot repeat, we would compute adjusted residuals using log-linear methods (and an appropriate computer program), not Equations 7.6 and 7.7.

7.5 Classic lag sequential methods

So far in this chapter, much of our discussion and most of our examples have been confined to two-event sequences. We have mentioned how longer sequences can be described and tested for significance (mainly in sections 7.1 and 7.2), but we have also noted that such tests may require prohibitive amounts of data. The approaches already discussed are "absolute" in the sense that they define particular sequences and then tally how often each occurred. As we attempt to investigate longer and longer sequences, the expected frequencies for particular sequences become vanishingly small, the number of possible sequences increases at a staggering rate, and it becomes almost impossible to make sense out of the wealth of information produced about so many different sequences. Clearly, a less absolute, more probabilistic and more flexible approach to the investigation of sequences comprising more than two events would be useful. One such approach is usually called the "lag sequential method." It was first developed by Sackett (1974, 1979, 1980) and later described by others (Bakeman & Dabbs, 1976; Bakeman, 1978; Gottman & Bakeman, 1979).

The reader of this book will already be familiar with the basic elements of the lag sequential method. As an example, assume that our code catalog defines several events, five of which are:

1. Infant Active
2. Mother Touch
3. Mother Nurse
4. Mother Groom
5. Infant Explore

(These codes are suggested by Sackett's work with macaque monkeys. The example here is based on one given in Sackett, 1974.) Assume further that successive events have been coded so that, as throughout this chapter, we are analyzing event-sequence data. Finally, assume that we are particularly interested in what happens after times when the infant is active, that is, we want to know whether there is anything systematic about the sequencing of events beginning with Infant Active episodes.

To begin with, the investigator selects one code to serve as the "criterion" or "given" event. In this case, that code would be Infant Active. Next, another code is selected as the "target." For example, we might select Mother Touch as our first target code. Then, a series of transitional probabilities are computed: for the target immediately after the criterion (lag 1), after one intervening event (lag 2), after two intervening events (lag 3), etc. Symbolically, we would write these lagged transitional probabilities as $p(T_1|G_0)$, $p(T_2|G_0)$, $p(T_3|G_0)$, etc. (remember, if we just write $p(T|G)$, target at lag 1 and given at lag 0 are assumed). The result is a series of transitional probabilities, each of which can then be tested for significance. For example, given Infant Active at lag "position" 0, if we had computed transitional probabilities for Mother Touch at lags 1 through 6, but only the lag 1 transitional probability significantly exceeded its expected value, we would conclude that Mother Touch was likely to occur just after Infant Active, but was not especially likely in the other lag "positions" investigated.

If we stopped now, we would have examined transitional probabilities for one particular target code, at different lags after a particular criterion code. This is not likely to tell us much about multievent sequences. The next step is to compute other series of transitional probabilities (and determine their significance), using the same criterion code but selecting different target codes. For example, given a criterion of Infant Active at lag 0, we could compute lag 1 through 6 transitional probabilities for Mother Nurse, Mother Groom, and Infant Explore. Imagine that the transitional probabilities for Mother Nurse at lag 2, for Mother Groom at lag 3, and for Infant Explore at lags 4 and 5 were significant. Such a pattern of results could suggest the four-event sequence: Infant Active, Mother Touch, Mother Nurse, Mother

Table 7.6. *Results required to confirm the Active, Touch, Nurse, Groom sequence*

		Lag				
Criterion	Target	1	2	3	4	5
Infant Active	Mother Touch	p*	p	p	p	p
	Mother Nurse	p	p*	p	p	p
	Mother Groom	p	p	p*	p	p
Mother Touch	Mother Nurse	p*	p	p	p	p
	Mother Groom	p	p*	p	p	p
Mother Nurse	Mother Groom	p*	p	p	p	p

Note: Asterisks indicate transitional probabilities whose values significantly exceed expected. Numerical values for transitional probabilities have not been given for this hypothetical example.

Groom (we shall return to Infant Explore in a moment), even though the lagged transitional probabilities examined only two codes at a time.

As stated before, the lag sequential is a probabilistic, not an absolute approach. To confirm the putative Active, Touch, Nurse, Groom sequence, we should do the following. First, compute lagged transitional probabilities with Mother Touch as the criterion and Mother Nurse and Mother Groom as targets, then with Mother Nurse as the criterion and Mother Groom as the target. If the transitional probabilities for Mother Nurse at lag 1 and Mother Groom at lag 2, with Mother Touch as the lag 0 criterion, and for Mother Groom at lag 1 with Mother Nurse as the lag 0 criterion, are all significant, then we would certainly be justified in claiming that the Active, Touch, Nurse, Groom sequence was especially characteristic of the monkeys observed (see Table 7.6).

Recall, however, that Infant Explore was significant at lags 4 and 5 after Infant Active. Does this mean that we are dealing with a six-event instead of a four-event sequence? The answer is, not necessarily. For example, if the transitional probabilities for Infant Explore at lags 3 and 4 given Mother Touch as the criterion, at lags 2 and 3 given Mother Nurse, and at lags 1 and 2 given Mother Groom were not significant, then there would be no reasons to claim that Infant Explore followed the Active, Touch, Nurse, Groom sequence already identified. Instead, if the results were as suggested here, we would conclude that after a time when the infant was active, next we would likely see either the Touch, Nurse, Groom sequence or else three more or less random events followed by Infant Explore in the fourth or fifth position.

If we let X stand for a "random" event, then in effect we have detected the following sequence: Active, X, X, X, Explore, Explore. More accurately, because in this example adjacent codes must be different, we have detected the following two sequences: Active, X, X, X, Explore, X and Active, X, X, X, X, Explore. Given other data, we might have detected a sequence like Active, X, Nurse, Groom, which we would interpret as follows: Whatever happens after times when the infant is active is not systematic – it could be almost any code, randomly chosen. After a random event in lag 1, however, the Nurse, Groom sequence is likely (in lag positions 2 and 3). Such a sequence would not be easily detected with "absolute" methods. One advantage, then, of the lag sequential approach is the ease with which sequences containing random elements can still be detected. The main advantage of this approach, however, remains its ability to detect sequences involving more than two events without requiring as much data as absolute methods would. When interpreting lag sequential results, a number of cautions apply. First, it is important to keep in mind whether adjacent codes can be the same or not, because this affects how expected frequencies or probabilities are computed. In the previous section, we noted that when consecutive codes cannot repeat, expected frequencies for lag 1 are best computed using an iterative procedure, although Equation 7.7 from the lag sequential literature provides an approximation. A similar approximation is suggested by Sackett (1979) for lags greater than 1. It is

$$m_{GT} = \frac{x_T - x_{GT}}{N_1 - x_G} x_G \tag{7.8}$$

where m_{GT} represents the expected frequency for the target behavior at lag L when preceded by the given behavior at lag 0, and x_{GT} the observed frequency for the target at lag $L-1$ preceded by the given behavior at lag 0.

Sackett (1979) reasoned that when adjacent codes cannot repeat, the expected probability for a particular target code at lag L (assuming a particular given code at lag 0) is the frequency for that target code diminished by the number of times it appears in the lag $L-1$ position (because then it could not appear in the L position, after itself) divided by the number of events that may occur at lag L (which is the sum of the lag L minus the lag $L-1$ frequencies summed across all K target codes). Simply put, this sum is the number of all events less the number of events assigned the given code. As with Equation 7.7, Equation 7.8 assumes overlapped sampling; and again like Equation 7.7, marginals for expected frequencies based on Equation 7.8 do not match the observed marginals.

Nonetheless, when consecutive codes cannot repeat, traditional lag sequential analysis (Sackett, 1979; the first edition of this book) has

estimated expected frequencies at lag 1 with Equation 7.7 and at longer lags with Equation 7.8, and then has determined statistical significance based on the z computed per Allison and Likers's (1982) Equation 7.6. As already noted, at lag 1 we recommend the log-linear methods described in the previous section, and in the next section we develop log-linear methods that apply at longer lags. But note, when consecutive codes may repeat, and when overlapped sampling is used, at lag 1 traditional lag sequential (Equations 7.4 and 7.6) and log-liner (Equations 7.1 and 7.2) analyses produce almost identical results. The same is essentially true at longer lags, although then log-linear analyses offers certain additional advantages, as described in the next section.

Two additional cautions should be mentioned. As is always true, no matter the statistic, before significance is assigned to any z score, the investigator should determine that there are sufficient data to justify this (see section 8.5). Finally, as always, the investigator should keep in mind the type I error problem (see section 8.6).

This is probably not a serious problem if sequences beginning with just one, or at most a few, criterion codes are investigated in the context of a confirmatory study. However, if many codes are defined, and if all serve exhaustively as criteria and targets, with many lags, then interpretation of such exploratory results should be guided by the almost certain knowledge that some chance findings are contained therein.

We end this section with a second example of the lag sequential method. For a study of marital communication Gottman, Markman, and Notarius (1977) coded the sequential utterances of a number of nondistressed and distressed couples observed discussing marital problems. Among other questions, these investigators wanted to know what happened after the husband complained about a marital problem. In other words, given Husband Complaint as the criterion code, they computed lagged transitional probabilities for a number of target codes, including Wife Agreement, Wife Complaint, Husband Agreement, and Husband Complaint. (The same code may serve as both criterion and target; however, when adjacent codes must be different, both its observed and expected frequencies at lag 1 will be 0. This is an example of a "structural" zero.) Although 24 codes were defined, the four just listed always included the highest z scores.

An interesting difference was noted between nondistressed and distressed couples. For nondistressed couples, significant z scores occurred only when Wife Agreement (at lags 1, 3, and 5) and Husband Complaint (at lags 2 and 4) served as targets. The process of cycling between Husband Complaint and Wife Agreement, Gottman et al. called "validation." For distressed couples, on the other hand, significant z scores occurred only when Wife Complaint (at lags 1 and 3) and Husband Complaint (at lags 2,

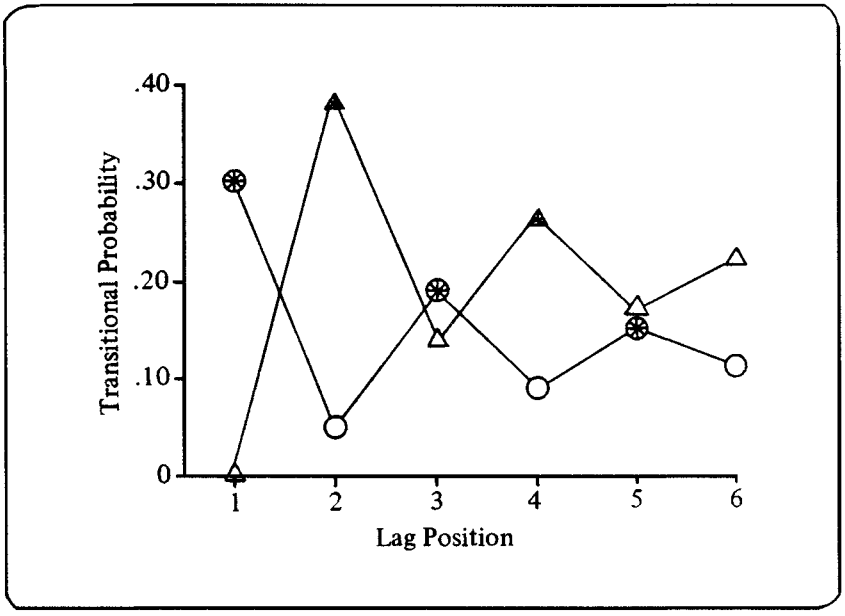

Figure 7.2. A lagged probability profile for nondistressed couples. Triangles represent transitional probabilities for Husband Complaint at the lag specified, given Husband Complaint at lag 0. Circles represent transitional probabilities for Wife Agreement at the lag specified, given Husband Complaint at lag 0. Asterisks (*) indicate that the corresponding z score is significant.

4, and 6) served as targets, a process that Gottman et al. termed "cross-complaining." Lagged probability profiles for these results are presented in Figures 7.2 and 7.3.

7.6 Log-linear approaches to lag-sequential analysis

Since lag-sequential analysis was first developed, log-linear analyses have become more widely understood and used by social scientists (Bakeman & Robinson, 1994; Wickens, 1989). They offer a number of advantages over traditional lag-sequential methods. As already noted, structural zeros are handled routinely and do not require the ad hoc formulas described in the previous section. But primarily, use of log-linear methods allows integration of sequential analysis into an established and well-supported sta-

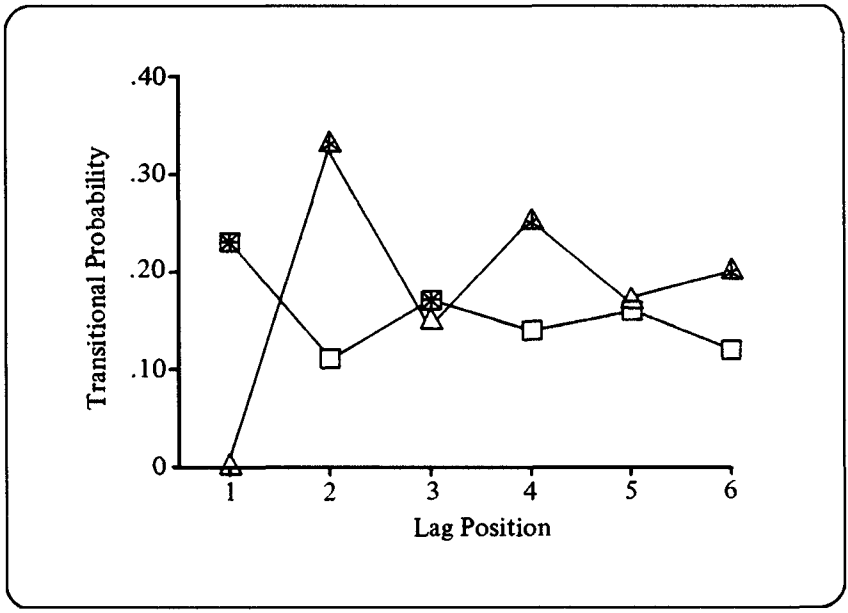

Figure 7.3. A lagged probability profile for distressed couples. As in Figure 7.2, triangles represent transitional probabilities for Husband Complaint at the lag specified, given Husband Complaint at lag 0. Squares, however, represent transitional probabilities for Wife Complaint, given Husband Complaint at lag 0. Again asterisks (*) indicate that the corresponding z score is significant.

tistical tradition (Bakeman & Quera, 1995b). As Castellan (1979) pointed out some time ago, almost always sequential questions can be phrased in terms of multidimensional contingency tables, which log-linear analysis was developed to analyze.

In this section, we describe three advantages of a log-linear view to sequential questions. First, log-linear analysis promotes a whole-table view, whereas often traditional lag-sequential analysis focused, almost piecemeal, on individual transitions in a table, which invites type I error. Moreover, log-linear analysis provides ways of disentangling the web of connected results in a table, as we demonstrate shortly. Finally, log-linear analysis, using well-established techniques, provides an integrated method for determining whether there are effects at various lags, no matter whether consecutive codes may or cannot repeat and no matter whether or not overlapped sampling was employed.

Omnibus tests

Used in an exploratory way, traditional lag-sequential analysis invites type I error (although among statistical techniques it is hardly unique in this respect). When 10 codes are defined, for example, the lag 0 by lag 1 table contains 100 cells when consecutive codes may repeat and 90 when not (K^2 and $K[K-1]$ generally, where K is the number of codes defined). Assuming the usual .05 value for alpha, if there were no association between lag 0 and lag 1 behavior, approximately 5 of 100 transitions would be identified, on average and incorrectly, as statistically significant. One solution is to take an omnibus or whole-table view. Absent specific predictions that one or just a few transitions will be significant, individual cell statistics should be examined for significance only when a tablewise statistic, such as the Pearson or likelihood-ratio chi-square (symbolized as X^2 and G^2, respectively), is large, just as post hoc tests are pursued in analysis of variance only when the omnibus F ratio is significant (Bakeman, 1992; Bakeman & Quera, 1995b).

Applying this test, we would not have examined the data presented in Table 7.1 further. For these data,

$$X^2(11, N = 127) = \sum_{G}^{K} \sum_{T}^{K} \frac{(\mathrm{x}_{GT} - \mathrm{m}_{GT})^2}{\mathrm{m}_{GT}} = 11.0 \tag{7.9}$$

and

$$G^2(11, N = 127) = 2 \sum_{G}^{K} \sum_{T}^{K} \mathrm{x}_{GT} \log \frac{\mathrm{x}_{GT}}{\mathrm{m}_{GT}} = 10.3 \tag{7.10}$$

where *log* represents the natural logarithm (i.e., the logarithm to the base e); X^2 and G^2 both estimate chi-square, although usually G^2 is used in log-linear analyses (for a discussion of the differences between them, see Wickens, 1989). These estimates fall short of the .05 critical value of 19.7. Moreover, only 1 of 20 adjusted residuals exceeded 1.96 absolute (Unoccupied to Solitary; see Table 7.5), which suggests it was simply a chance finding, unlikely to replicate.

In log-linear terms, we ask whether expected frequencies generated by the model of independence (i.e., Equation 7.1), which is symbolized [0][1] and indicates the independence of the lag 0 and lag 1 dimension, are similar to the observed frequencies. If they are, then the chi-square statistic will not exceed its .05 critical value, as here (i.e., observed frequencies fit those expected tolerably well). However, if the computed chi-square statistic is large, exceeding its .05 critical value, then we reject the model of independence and conclude that the dimensions of the table are in fact related and not independent.

Table 7.7. *Observed frequencies and adjusted residuals for 100 two-event sequences*

Given code,	Observed frequencies: Target code, lag 1				Adjusted residuals: Target code, lag 1		
lag 0	A	B	C	Totals	A	B	C
A	23	5	15	43	2.02	−1.82	−.056
B	11	1	7	19	1.56	−1.78	−0.12
C	8	14	16	38	−3.32	3.30	0.66
Totals	42	20	38	100			

Note: This example was also used in Bakeman and Quera (1995b).

Winnowing results

As a second example, consider the data given in Table 7.7 for which K is 3; the codes are labeled A, B, and C; and consecutive codes may repeat. For these data, $X^2(4, N = 100)$ is 15.7 and $G^2(4, N = 100)$ is 16.4. Both exceed the .05 critical value of 9.49, which suggests that lag 0 and lag 1 are associated. Moreover, three of nine adjusted residuals exceed 1.96 absolute (again, see Table 7.7). But now we confront a different dilemma. Adjusted residuals in a table form an interrelated web. If some are large, others necessarily must be small, and so, rather than attempting to interpret each one (thereby courting type I error), now we need to determine which one or ones should be emphasized.

The initial set of all statistically significant transitions in a table can be winnowed using methods for incomplete tables (i.e., tables with structural zeros). Assume the C–B chain, whose adjusted residual is 3.30, is of primary theoretical interest. In order to test its importance, we declare the C–B cell structurally zero, use an iterative procedure to compute expected frequencies (e.g., using Bakeman & Robinson's, 1994, ILOG program), and note that now the [0][1] model fits the remaining data ($G^2[3, N = 86] = 5.79$; .05 critical value for 3 $df = 7.81$; $df = 3$ because one is lost to the structurally zero cell). We conclude that interpretation should emphasize the C–B chain as the other two effects (decreased occurrences for C–A, increased occurrences for A–A) disappear when the C–B chain is removed.

Had the model of independence not fit the reduced data, we would have declared another structural zero and tested the data now reduced by two cells. Proceeding stepwise (but letting theoretical considerations not raw empiricism determine the next chain to delete, else one risks capitalizing on chance as with backward elimination in multiple regression and compro-

mising type I error control), we would identify those chains that prevent the [0][1] model from fitting. (A logically similar suggestion, not in a log-linear context, is made by Rechten & Fernald, 1978; see also Wickens, 1989, pp. 251–253.) The ability to winnow results in this way is one advantage of the log-linear view over traditional lag-sequential analysis.

Sequential log-linear models

Perhaps the major advantages of a log-linear approach to sequential problems are the statistical foundation and generality provided. Assuming an interest in lags no greater than L, we begin by assembling $(L + 1)$-event chains. For example, if we were interested in lags no greater than 2, we would collect three-event chains in which the first event was associated with lag 0, the second with lag 1, and the third with lag 2.

The three-event chains might be derived from just one or a few longer sequences using overlapped sampling, selecting first $e_1e_2e_3$, then $e_2e_3e_4$, then $e_3e_4e_5$, etc., where e_i represents an event in the longer sequence. In such cases, if S segments were coded comprising N events in all, the number of three-event chains derived would be $N - SL$ (assuming none of the segments consisted of fewer than $L + 1$ events). For example, in the simplest instance, if one segment consisting of N events were coded, $N - 2$ three-event chains would be tallied. Alternatively, the three-event chains might be derived from a few longer sequences using nonoverlapped sampling (selecting first $e_1e_2e_3$, then $e_4e_5e_6$, etc.), in which case the number of three-event chains derived would be $N \div 3$ (assuming all segments are multiples of three). Or the three-event sequences might be sampled directly from a population of three-event sequences.

Overlapped sampling is often used, and usually assumed, in traditional lag-sequential analysis. But nonoverlapped sampling is both quite common and useful. For example, imagine that we are only interested in tallying speaker 1 (e.g., husband) to speaker 2 (e.g., wife) lag 1 transitions. Then, assuming two-event chains were derived from a single segment of N coded events, the number of speaker 1 to 2 transitions would be $N \div 2$ if the speakers always alternated turns (tallying first e_1e_2, then e_3e_4, etc.), or some smaller number if a speaker's turn may contain more than one thought unit (which could result in tallying e_2e_3, e_5e_6, e_7e_8, $e_{11}e_{12}$, etc.; see Bakeman & Casey, 1995).

No matter the sampling strategy, the $(L + 1)$-event chains are tallied in a K^{L+1} table; thus each chain adds a tally to one, and only one, cell. To demonstrate the log-linear approach, let us begin with an example representing the simplest of circumstances, assuming codes of A, B, and C

	1: A	B	C
0:A	21	25	49
B	23	26	21
C	50	19	15

Figure 7.4. Observed frequencies for 249 two-event chains derived from a sequence of 250 events.

(thus $K = 3$) that may repeat and an initial interest in lag 1. Then occurrences of each of the nine possible two-event chains (AA, AB, etc.) would be tallied in one of the cells of a 3^2 table. For example, we (Bakeman & Quera, 1995b) generated a sequence of 250 coded events and tallied the 249 overlapped two-event chains; the results are shown in Figure 7.4. For this two-dimensional table, the [0][1] model (implying independence of rows and columns) fails to fit the data (G^2[4, N=249]=35.2) and so we would conclude that events at lag 0 and lag 1 are associated and not independent. This much seems easy and, apart from the preliminary omnibus test, not much different from traditional lag-sequential methods.

Next, assume that our interest expands from lag 1 to lag 2. Still assuming $K = 3$ and consecutive codes that may repeat, then each of the 27 possible 3-event chains (AAA, AAB, etc.) would be tallied in one of the cells of a 3^3 table. Tallies for the 248 overlapped three-event chains derived from the same sequence used earlier are shown in Figure 7.5.

Cells for the 3^3 table shown in Figure 7.5 are symbolized x_{ijk}, where i, j, and k represent the lag 0, lag 1, and lag 2 dimensions, respectively. Traditional lag-sequential analysis would test for lag 2 effects in the collapsed 02 table, that is, the table whose elements are x_{i+k}, where

$$x_{i+k} = \sum_{j}^{K} x_{ijk}$$

This table is shown in Figure 7.6. For this two-dimensional table, the [0][2] model (implying independence of rows and columns) fails to fit the data (G^2[4, $N = 248$] = 10.70) and so, traditionally, we would conclude that events at lag 0 and lag 2 are associated and not independent. But this fails to take into account events at lag 1.

A hierarchic log-linear analysis of the 3^3 table shown in Figure 7.5 provides more information, and in this case leads to a different conclusion, than a traditional lag-sequential analysis of the collapsed table shown in Figure 7.6. The complete or *saturated* model for a three-dimensional table

		2: A	B	C
0:A	1:A	7	6	8
	B	10	9	6
	C	31	8	10
B	1:A	3	6	14
	B	6	11	9
	C	12	4	4
C	1:A	11	13	26
	B	7	6	6
	C	7	7	1

Figure 7.5. Observed frequencies for 248 three-event chains derived from the same sequence of 250 events used for Figure 7.4.

	2: A	B	C
0:A	48	23	24
B	21	21	27
C	25	26	33

Figure 7.6. Observed frequencies for the collapsed lag 0×lag 2 table derived from the observed frequencies for three-event chains shown in Figure 7.5.

is represented as [012] and includes seven terms: *012, 01, 12, 02, 0, 1,* and *2*. The saturated model is not symbolized as [012][01][12][02][0][1][2] because the three two-way and three one-way terms are implied by (we could say, nested hierarchically within) the three-way term, and so it is neither necessary nor conventional to write them explicitly.

Typically, a hierarchic log-linear analysis proceeds by deleting terms, seeking the simplest model that nonetheless fits the data tolerably well (Bakeman & Robinson, 1994). Results for the data shown in Figure 7.4 are given in Table 7.8. The best-fitting model is [01][12]; the term that represents lag 0–lag 2 association (i.e., the *02* term) is not required. Thus

Table 7.8. *Hierarchic log-linear analysis of the data shown in Figure 7.4*

Model	G^2	G^2 df	Term Deleted	ΔG^2	ΔG^2 df
[012]	0.0	0			
[01][12][02]	7.67	8	012	7.67	8
[01][12]	11.52	12	02	3.85	4
[01][2]	45.88*	16	12	34.35*	4
[0][1][2]	81.34*	20	01	35.46*	4

$^{**}p < .01$

the log-linear analysis reveals that, when events at lag 1 are taken into account, events at lag 0 and lag 2 are not associated, as suggested by the analysis of the 02 table, but are in fact independent. Such conditional independence – that is, the independence of lag 0 and lag 2 conditional on lag 1 – is symbolized $0 \perp\!\!\!\perp 2|1$ by Wickens (1989; see also Bakeman & Quera, 1995b), and the ability to detect such circumstances represents an advantage of log-linear over traditional lag-sequential methods. Readers who wish to pursue the matter of conditional independence further should read Wickens (1989, especially chapter 3).

As just described, when consecutive codes may repeat log-linear but not traditional lag-sequential methods detect conditional independence. Additional advantages accrue when consecutive codes cannot repeat because log-linear methods handle structural zeros routinely and do not require ad hoc and problematic formulas such as Equations 7.7 and 7.8. As an example, we (Bakeman & Quera, 1995b) generated a sequence of 122 coded events and tallied the 120 overlapped three-event chains. Tallies for the 12 permitted sequences are given in Figure 7.7. Cells containing structural zeros are also indicated; when consecutive codes cannot repeat, the 012 table will always contain the pattern of structural zeros shown.

A summary of the log-liner analysis for the data given in Figure 7.7 is shown in Table 7.9. When K is 3, and only when K is 3, the [01][12][02] model is completely determined; its degrees of freedom are 0 and expected frequencies duplicate the observed ones (as in the [012] model, when consecutive codes may repeat). Unlike in the previous analysis, for these data the model of conditional independence – [01][12] – fails to fit the data ($G^2[3, N = 120] = 10.83, p < .05$). Thus we accept the [01] [12] [02] model and conclude that events at lag 0 and lag 2 are associated (and both are associated with lag 1).

		2: A	B	C
0:A	1:A	–	–	–
	B	15	–	6
	C	10	9	–
B	1:A	–	12	8
	B	–	–	–
	C	9	12	–
C	1:A	–	8	11
	B	5	–	15
	C	–	–	–

Figure 7.7. Observed frequencies for the 12 possible three-event chains derived from a sequence of 122 events for which consecutive codes cannot repeat. Structural zeros are indicated with a dash.

Table 7.9. *Hierarchic log-linear analysis of the data shown in Figure 7.7*

Model	G^2	G^2 df	Term Deleted	ΔG^2	ΔG^2 df
[01][12][02]	0.0	0			
[01][12]	10.83*	3	02	10.83*	3
[01][12]–CBC	1.64	2	cell x_{CBC}	9.19*	1

*$p < .05$
**$p < .01$

Moreover, we can winnow these results, exactly as described for the earlier example that permitted consecutive codes to repeat. For theoretic reasons, assume that the *CBC* chain is of particular interest. An examination of the residuals for the [01] [12] model (i.e., the differences between observed frequencies and expected frequencies generated by the [01] [12] model) showed that 4 of the 12 chains were associated with quite large absolute residuals (the *ABA, ABC, CBA,* and *CBC* chains), which made us think that the observed frequencies for these chains, in particular,

might be responsible for the failure of the [01] [12] model to fit the observed data. Because the *CBC* chain is of primary interest, we replaced the x_{CBC} cell (which contained a tally of 15) with a structural zero. As shown in Table 7.9, the model of conditional independence now fit the data ($G^2[2, N = 105] = 1.64, NS$), and so we conclude that the *CBC* chain can account for the lag 2 effect detected by the omnibus analysis described in the previous paragraph. This can be tested directly with a hierarchic test, as indicated in Table 7.9. The difference between two hierarchically related G^2s is distributed approximately as chi-square with degrees of freedom equal to the difference between the degrees of freedom for the two G^2s; in this case, $\Delta G^2(1) = 9.19, p < .01$. (Replacing a different chain with a structural zero might also result in a fitting model, which is why it is so important that selection of the chain to consider first be guided by theoretic concerns.)

Minimizing data demands

Quantitative data analysis always requires sufficient data, and log-linear and traditional lag-sequential approaches are no exception. Still, the data required for the multidimensional tables of log-linear analysis can be quite intimidating. Several rules of thumb for log-linear analysis are available, usually stated in terms of expected frequencies or even degrees of freedom for hierarchic tests like the one for cell x_{CBC} shown in Table 7.9, but one suggested requirement (a necessary minimum, but not necessarily sufficient) is that the total sample be at least 4 or 5 times the number of cells not structurally zero (Wickens, 1989, p. 30). This number is K^{L+1} when consecutive codes may repeat and $K(K-1)^L$ when they cannot (e.g., when $K = 3$ and $L = 2$, the number of cells is 27 and 12 when codes may and cannot repeat, respectively). Thus the number of cells, and so the total sample desired, increases exponentially with increases in K and L (although the increase is somewhat less pronounced when consecutive codes cannot repeat).

Especially for larger values of L, unless the number of events observed is almost astronomically large, the average number of events per cell may be distressingly small. Further, expected frequencies for far too many of the cells may be near zero, which is problematic for log-linear analysis. To minimize data demands, Bakeman and Quera (1995b) have suggested a sequential search strategy for explicating lagged effects. Although the details are somewhat different, the general strategy is the same when consecutive codes may and cannot repeat, which once again demonstrates the generality of the log-linear approach.

Consider first the strategy when consecutive codes may repeat. Lag 1 effects, which require only the two-dimensional 01 table, are unproblematic,

although of course we would test for lag 1 effects. Thus the search begins by looking for complex lag 2 effects in the 012 table. At each lag ($L > 1$), the complex effects we seek first implicate lags 0 and L with lag $L - 1$. If present, collapsing over the $L - 1$ dimension (which would reduce the number of cell and so data demands) is unwarranted. For example, if L is 2, then complex effects are present if the simplest model that fits the 012 table includes any of the following:

1. [012] because then lag 0 and lag 2 are associated and interact with lag 1 (three-way associations), or
2. [01][12][02] because then lag 0 and lag 2 are associated with each other and lag 1 but do not interact with lag 1 (homogeneous associations), or
3. [01][12] because then lag 0 and lag 2 are independent conditional on lag 1.

If complex effects are found, we would explicate them, as demonstrated in the previous section. However, if simpler models fit (e.g., [01][2] or any others not in the list just given), which means no complex effects were found, then collapsing over the $L - 1$ dimension is justified (Wickens, 1989, pp 79–81, pp. 142–143), resulting in the $0L$ table of traditional lag-sequential analysis (e.g., the 02 table when $L = 2$).

Assuming no complex effects are found in the 012 table, after collapsing we would first test whether $0 \perp\!\!\!\perp 2$ (unconditional independence) in the 02 table and, if not, examine residuals in order to explicate the lag 0–lag 2 effect just identified (exactly as we would have done for the 01 table). Next we would create a new three-dimensional table by adding the lag 3 dimension, tally sequences in this 023 table, and then look for lag 3 effects in the 023 table exactly as described for the 012 table. This procedure is repeated for successive lags. In general terms, beginning with lag L, we test whether the three-dimensional $0(L - 1)L$ table can be collapsed over the $L-1$ dimension. If so, we collapse to the $0L$ table, add the $L+1$ dimension thereby creating a new three-dimensional table, increment L, and repeat the procedure, continuing until we find a table that does not permit collapsing over the $L - 1$ dimension. Once such a table is found, we explicate the lag L effects in this three-dimensional table. If data are sufficient, we might next analyze the four-dimensional $0(L-1)L(L+1)$ table, and so forth, but further collapsing is unwarranted because of the lag L effects just found. Nonetheless, this strategy may let us examine lags longer than 2 without requiring tables larger then K^3 when consecutive codes may repeat.

The sequential search strategy described in the previous paragraph applies when consecutive codes cannot repeat with one modification. When consecutive codes may repeat, and no complex lag L effects are found (i.e., each table examined sequentially permits collapsing) then the test series

becomes $0 \perp\!\!\!\perp 2$, then $0 \perp\!\!\!\perp 3$ and so forth (i.e., $0 \perp\!\!\!\perp L$ is tested in the $0L$ table), as just described. When consecutive codes cannot repeat, the unconditional test makes no sense because it fails to reflect the constraints imposed when consecutive codes cannot repeat. Then, when no complex lag L effects are found, the analogous series becomes $0 \perp\!\!\!\perp 2|1$, $0 \perp\!\!\!\perp 3|2$, and so forth [i.e., $0 \perp\!\!\!\perp L|L-1$ is tested in the $0(L-1)L$ table]. Models associated with these tests include the $(L-1)L$ term. The corresponding marginal table has structural zeros on the diagonal, which reflect the cannot-repeat constraint. This strategy may let us examine lags longer than 2 without requiring tables larger than $K^2(K-1)$ when consecutive codes cannot repeat ($K[K-1]^2$ when $L = 2$). These matters are discussed further in Bakeman and Quera (1995b).

7.7 Computing Yule's Q or phi and testing for individual differences

Often more than a single individual, dyad, family, or whatever, is observed; these *units* are embedded in a design (e.g., a two-group design might include clinic and nonclinic couples), and investigators want to ask questions about the importance of their research factors (e.g., is a particular sequential pattern more characteristic of clinic than nonclinic couples). The previous edition of this book suggested than z scores might serve as scores for subsequent analyses (e.g., analyses of variance, multiple regression, etc.), but that was not sound advice. The z score is affected by the number of tallies (if the number of tallies doubled but the association remained the same, the z score would increase), and so is not comparable across experimental units (subjects, dyads, families, etc.) unless the total number of tallies remains the same for each. Some measure that is unaffected by the number of tallies, such as a strength of association or effect size measure, should be used instead (Wampold, 1992).

Strength of association or effect size measures are especially well developed for 2×2 tables (to give just two examples from an extensive literature, see Conger & Ward, 1984, and Reynolds, 1984; much of the material in this and subsequent paragraphs is summarized from Bakeman, McArthur, & Quera, 1996). This is fortunate, because when interest centers on one cell in a larger two-dimensional table, the larger table can be collapsed into a 2×2, and statistics developed for 2×2, tables can be used (as Morley, 1987, noted with respect to phi). Assume, for example, that we want to know whether event B is particularly likely after event A. In this case, we would label rows A and $\sim A$ and columns B and $\sim B$ (where rows represent lag 0, columns lag 1, and $\sim$ represents *not*). Then the collapsed 2×2

table can be represented as

	B	~B
A	*a*	*b*
~A	*c*	*d*

where individual cells are labeled a, b, c, and d as shown and represent cell frequencies.

One of the most common statistics for 2×2 tables (perhaps more so in epidemiology and sociology than psychology) is the odds ratio. As its name implies, it is estimated by the ratio of a to b divided by the ratio of c to d,

$$est.\ odds\ ratio = \frac{a/b}{c/d} \tag{7.11}$$

(where a, b, c, and d refer to observed frequencies for the cells of a 2×2 table as noted earlier; notation varies, but for definitions in terms of population parameters, see Bishop Fienberg, & Holland, 1975; and Wickens, 1993). Multiplying numerator and divisor by d/c, this can also be expressed as

$$est.\ odds\ ratio = \frac{ad}{bc}. \tag{7.12}$$

Equation 7.12 is more common, although Equation 7.11 reflects the name and renders the concept more faithfully. Consider the following example:

	B	~B	
A	10	10	20
~A	20	60	80
	30	70	100

The odds for B after A are 1:1, where as the odds for B after any other (non-A) event are 1:3; thus the odds ratio is 3. In other words, the odds for B occurring after A are three times the odds for B occurring after anything else. When the odds ratio is greater than 1 (and it can always be made ≥ 1 by swapping rows), it has the merit, lacking in many indices, of a simple and concrete interpretation.

The odds ratio varies from 0 to infinity and equals 1 when the odds are the same for both rows (indicating no effect of the row classification). The natural logarithm (ln) of the odds ratio, which is estimated as

$$est.\ log\ odds\ ratio = \ln\left(\frac{ad}{bc}\right) \tag{7.13}$$

extends from minus to plus infinity, equals 0 when there is no effect, and is more useful for inference (Wickens, 1993). However Equation 7.13 estimates are biased. An estimate with less bias, which is also well defined when one of the cells is zero (recall that the log of zero is undefined), is obtained by adding 1/2 to each count,

$$y = \ln \frac{(a + 1/2)(d + 1/2)}{(c + 1/2)(b + 1/2)} \tag{7.14}$$

(Gart & Zweifel, 1967; cited in Wickens, 1993, Equation 8). As Wickens (1993) notes when recommending that the log odds ratio computed per Equation 7.14 be analyzed with a parametric t test, this procedure not only provides protection for a variety of hypotheses against the effects of intersubject variability when categorical observations are collected from each member of a group (or groups), it is also easy to describe, calculate, and present.

Yule's Q

Yule's Q is a related index. It is a transformation of the odds ratio designed to vary, not from zero to infinity with 1 indicating no effect, but from -1 to $+1$ with zero indicating no effect, just like the familiar Pearson product – moment correlation. For that reason many investigators find it more descriptively useful than the odds ratio. First, c/d is subtracted from the numerator so that Yule's Q is zero when a/b equals c/d. Then, a/b is added to the denominator so that Yule's Q is $+1$ when b and/or c is zero and -1 when a and/or d is zero, as follows:

$$\text{Yule's Q} = \frac{\frac{a}{b} - \frac{c}{d}}{\frac{c}{d} + \frac{a}{b}} = \frac{\frac{ad - bc}{bd}}{\frac{bc + ad}{bd}} = \frac{ab - bc}{ad + bc} \tag{7.15}$$

Yule's Q can be expressed as a monotonically increasing function of both the odds and log odds ratio; thus these three indices are equivalent in the sense of rank ordering subjects the same way (Bakeman, McArthur, & Quera, 1996).

Phi

Another extremely common index for 2×2 tables is the phi coefficient. This is simply the familiar Pearson product–moment correlation coefficient computed using binary coded data (Cohen & Cohen, 1983; Hays, 1963).

One definition for phi is

$$\phi = \frac{z}{\sqrt{N}} \tag{7.16}$$

where z is computed for the 2×2 table and hence equals $\sqrt{\chi^2}$. Thus phi can be viewed as a z score corrected for sample size. Like Yule's Q, it varies from -1 to $+1$ with zero indicating no association. In terms of the four cells, phi is defined as

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{7.17}$$

Multiplying and rearranging terms this becomes

$$\phi = \frac{ad - bc}{\sqrt{(ac + bd + ad + bc)(ab + cd + ad + bc)}} \tag{7.18}$$

If we now rewrite the expression of Yule's Q, first squaring the denominator of Equation 7.15 and then taking its square root

$$\text{Yule's Q} = \frac{ad - bc}{\sqrt{(ad + bc)(ad + bc)}} \tag{7.19}$$

the value of Yule's Q is not changed but similarities and differences between phi and Yule's Q (Equations 7.18 and 7.19) are clarified.

Does it matter which index is used, Yule's Q or phi? The multiplier and multiplicand in the denominator for Yule's Q (Equation 7.19) consist only of the sum of ad and bc, whereas multiplier and multiplicand in the phi denominator (Equation 7.18) add more terms. Consequently, values for phi are always less than values for Yule's Q (unless b and c, or a and d, are both zero, in which case both Yule's Q and phi would be $+1$ and -1, respectively). Yule's Q and phi differ in another way as well. Yule's Q is $+1$ when either b or c is zero and -1 when either a or d is zero (this is called weak perfect association, Reynolds, 1984), whereas phi is $+1$ only when both b and c are zero and -1 only when both a and d are zero (this is called strict perfect association). Thus phi achieves its maximum value (absolute) only when row and column marginals are equal (Reynolds, 1984). Some investigators may regard this as advantageous, some as disadvantageous, but in most cases it probably matters little which of these two indices is used (or whether the odds ratio or log odds ratio is used instead). In fact, after running a number of computer simulations, Bakeman, McArthur, and Quera (1996) concluded that, when testing for group differences, it does not matter much whether Yule's Q or phi is used since both rank-order cases essentially the same. Transformed kappa, a statistic proposed by Wampold

(1989, 1992), however, did not perform as well. For details see Bakeman, McArthur, and Quera (1996).

Type I error considerations

In an ideal confirmatory world, investigators would pluck the one transition from a larger table needed to answer their most important research question; compute a single Yule's Q or phi based on a collapsed A, ~ A|B, ~ B table, such as the one shown earlier; and proceed to test for group differences (or other questions as their design permits). But much of the world is rankly exploratory. Indeed, it is tempting to compute some index for each of the K^2 cells of a table ($K[K-1]$ cells when consecutive codes cannot repeat), one set for each subject, and then subject all K^2 scores to standard parametric tests (t test, analyses of variance, etc.). This courts type I error in a fairly major way. At the very least, no more indices should be derived than the degrees of freedom associated with the table, which is $(K-1)(K-1)$ when consecutive codes may repeat and $(K-1)(K-1)-K$ when not (assuming a table with structural zeros on the diagonal). This is somewhat analogous to decomposing an omnibus analysis of variance into single-degree-of-freedom planned comparisons or contrasts.

One systematic way to derive indices from a larger table requires that one code be regarded as something of a baseline, or base for comparison, such as *unengaged* or *no activity*. For example, imagine that codes are labeled *A*, *B*, and *C*, and that code *C* represents some sort of baseline. Then following Reynolds's (1984) suggestion for decomposing the odds ratio in tables larger than 2 × 2, and labeling the cells in the 3^2 table as follows:

	A	B	C
A	*a*	*b*	*c*
B	*d*	*e*	*f*
C	*g*	*h*	*i*

four 2 × 2 tables would be formed for each subject, as follows:

	A	C
A	*a*	*c*
C	*g*	*i*

	B	C
A	*b*	*c*
C	*h*	*i*

	B	*d*	*f*
	C	*g*	*i*

	B	*e*	*f*
	C	*h*	*i*

and a Yule's Q or phi computed for each. These statistics could then be subjected to whatever subsequent analyses the investigator deems appropriate.

In this section we have suggested that sequential associations between two particular events (e.g., an A to B transition) be assessed with an index like Yule's Q or phi. These statistics gauge the magnitude of the effect and, unlike the z score, are unaffected by the number of tallies. Thus they are reasonable candidates for subsequent analyses such as the familiar parametric tests routinely used by social scientists to assess individual differences and effects of various research factors (e.g., t tests, analyses of variance, and multiple regression). But the events under consideration may be many in number, leading to many tests and thereby courting type I error.

It goes without saying (which may be why it is so necessary to restate) that guiding ideas provide the best protection against type I error. Given K codes and an interest in lag 1 effects, a totally unguided and completely exploratory investigator might examine occurrences of all possible K^2 two-event chains (or $K[K-1]$ two-event chains when consecutive codes cannot repeat). In this section, we have suggested that a more justifiable approach would limit the number of transitions examined to the $(K-1)^2$ degrees of freedom associated with the table (or $[K-1]^2 - K$ degrees of freedom when consecutive codes cannot repeat) and have demonstrated one way that this number of 2×2 subtables could be extracted from a larger table. Presumably a Yule's Q or some other statistic would be computed for each subtable. Positive values would indicate that the pair of events associated with the upper-left-hand cell is associated more than expected, given the occurrences observed for the baseline events associated with the second row and second column of the 2×2 table. The summary statistic for the 2×2 tables, however many are formed, could then be subjected to further analysis.

Investigators are quite free – in fact, encouraged – to investigate a smaller number of associations (i.e., form a smaller number of 2×2 tables). For example, a larger table might be collapsed into a smaller one, combining some codes that seem functionally similar, or only those associations required to address the investigator's hypotheses might be subjected to analysis in the first place. Other transitions might be examined later, and those analyses labeled exploratory instead of confirmatory. For further discussion of this "less is more" and "least is last" strategy for controlling type I error, see Cohen and Cohen (1983, pp. 169–172).

7.8 Summary

Investigators often represent their data as sequences of coded events. Sometimes, data are recorded as event sequences in the first place; other times,

in order to answer particular questions, event sequences are extracted from data initially recorded and represented in some other way. The purpose of this chapter has been to describe some ways of analyzing such event sequences, although much of what has been presented here can apply to the analysis of time sequences as well.

Sometimes consecutive events cannot be assigned the same code in event sequences. For example, when coders are asked to segment the stream of behavior into mutually exclusive and exhaustive behavioral states, often adjacent states cannot be coded the same way, by definition. It they were the same, they would be just one state. However, we can imagine other ways of defining event boundaries that would allow adjacent codes to be the same. (The codes used to categorize events would still be mutually exclusive and exhaustive, but that is a different matter.) For example, if utterances were being coded, two successive utterances might both be coded the same. Whether adjacent codes can be the same or not is an important matter because it affects the way expected frequencies, expected probabilities, and hence adjusted residuals (i.e., z scores) are computed.

One approach to sequence detection we have called "absolute." Investigators define particular sequences of some specified length, categorize and tally all sequences of that length, and report the frequencies and probabilities for particular sequences. A z score can be used to gauge the extent to which an observed frequency (or probability) for a particular sequence exceeds its expected value. However, if the z score is to be tested for significance, its computation should be based on sufficient data to justify the normal approximation to the binomial distribution.

In theory, absolute methods apply to sequences of any length. In practice, certain limitations may prevail. In particular, the number of possible sequences increases dramatically as longer and longer sequences are considered. Unless truly mammoth amounts of data are available, expected frequencies for a particular sequence may be too small to justify assigning significance. Moreover, the number of occurrences for a particular sequence may be so few that the investigator has little confidence in the accuracy of the observed frequency, even descriptively. Another exacerbating circumstance is the number of codes defined. In general, when there are more codes, the expected frequencies for particular sequences are likely to be smaller, and hence more data will be required.

Even when z-score computations are based on sufficient data, the type I error problem remains. This is usually not a problem for confirmatory studies, assuming, of course, that just a few theoretically relevant tests of significance are made. But when the number of tests is large, as it typically is for exploratory studies, then some thought should be given to ways to control the type I error rate. As discussed in this chapter, the number

of initially significant transitions can be winnowed using structural zeros and log-linear methods, and the number of transitions examined in the first place can be limited to those of clear theoretic interest. Nonetheless, interpretation of results may need to take into account that some of the apparently significant findings are due simply to chance.

One way to describe two-event sequences is to report their simple probabilities. Another way is to report (lag 1) transitional probabilities instead. Of the two, transitional probabilities (which "control" for differences in the base rate of the first or "given" code) often seem more informative descriptively. The values of both, however, are affected by the values for the base rates of the two codes involved. This does not affect their descriptive value, but in cases in which transitional probabilities have been computed separately for different subjects, it does make such scores poor candidates for subsequent analyses of individual or group differences. First, depending on base rates, similar numerical values for the same transitional probability may have quite different meanings for different subjects. And second, what appear to be analyses of transitional probabilities may in fact reflect little more than simple probability effects. The z scores are not hampered by these problems, but have additional problems of their own. Their values are affected by the number of tallies, and so larger values may reflect, not a larger effect, but simply more data. Whenever individual or group differences or effects of research factors generally are of interest, magnitude of effect statistics, and not z scores, should be used. Examples include the odds ratio, the log odds ratio, Yule's Q, and phi.

A second approach to sequence detection, the lag-sequential method, we have characterized as "probabilistic" instead of "absolute." Like that of absolute methods, its purpose is to detect commonly occurring sequences, but because it examines codes pairwise only, it can detect sequences longer than two events without invoking the same restrictions and limitations involved with absolute methods. Moreover, sequences containing random elements can be detected as well. The method is based on an examination of z scores associated with transitional probabilities computed for various lags; thus any limitations and cautions (including the type I error problem) that apply to two-event transitional probabilities also apply to the lag-sequential approach as well.

Perhaps the most adequate approach to sequence detection is log-linear. Log-linear analysis promotes a whole-table view, whereas often traditional lag-sequential analysis focused, almost piecemeal, on individual transitions in a table. This is not necessarily problematic when, in the context of a confirmatory study, only a few transitions are of interest, but a narrow focus on repeated tests tied to all cells in the context of an exploratory study invites type I error. Additionally, log-linear analysis provides ways

of disentangling the web of connected results in a table, and makes routine the analysis of sequences in which, for logical reasons, consecutive codes cannot repeat. Finally, log-linear analysis, using well-established statistical techniques, provides an integrated method of broad generality for determining whether there are effects at various lags, no matter whether consecutive codes may or cannot repeat and no matter whether or not overlapped sampling was employed. Whenever possible, it seems the analytic approach of choice for the analysis of coded sequences.