8

Issues in sequential analysis

8.1 Independence

In classical parametric statistics, we assume that our observations are independent, and this assumption forms part of the basis of our distribution statistics. In the sequential analysis of observational data, on the other hand, we want to *detect dependence* in the observations. To do this we compare observed frequencies with those we would expect if the observations were independent. Thus, dependence in the data is not a "problem." It is what we are trying to study.

The statistical problem of an appropriate test is not difficult to solve. It was solved in a classic paper in 1957 by Anderson and Goodman (see also Goodman, 1983, for an update). Their solution is based on the likelihood-ratio chi-square test.

The likelihood-ratio test applies to the comparison of any two statistical models if one (the "little" model) is a subcase of the other (the "big" model). The null-hypothesis model is usually the little model. In our case, this model is often the assumption that the data are independent (or quasi independent); i.e., that there is no sequential structure. Compared to this is the big, interesting model that posits a dependent sequential structure. As discussed in section 7.6, the difference between the G^2 for the big model (e.g., [01]) and the G^2 for the little model (e.g., [0][1]) is distributed asymptotically as chi-square, with degrees of freedom equal to the difference in the degrees of freedom for the big and little models. "Asymptotic" means that it becomes increasingly true for large N, where N is the number of observations.

When the data have "structural zeros," e.g., if a code cannot follow itself (meaning that the frequency for that sequence is necessarily zero), the number of degrees of freedom must be reduced (by the number of cells that are structural zeros). These cells are not used to compute chi-square (see Goodman, 1983).

We shall now discuss the conditions required to reach asymptote. In particular, we shall discuss assigning probability values to z scores. We

should note that most observational data are *stochastically* dependent. They are called in statistics "m-dependent" processes, which means that the dependencies are short lived. One implication of this is that there is poor predictability from one time point to another, as the lag between time points increases. In time-series analysis, forecasts are notoriously poor if they exceed one step ahead (see Box & Jenkins, 1970, for a graph of the confidence intervals around forecasts). It also means that clumping m observations gives near independence. For most data, m will be quite small (probably less than 4), and its size relative to n will determine the speed at which the asymptote is approached.

We conclude that assigning probability values to pairwise z scores (or tablewise chi-squares) is appropriate when we are asking if the observed frequency for a particular sequence is significantly different from expected (or whether lag 0 and L are related and not independent). We admit, however, that more cautious interpretations are possible, and would quote a paragraph we wrote earlier (Gottman & Bakeman, 1979, p. 190):

> As N increases beyond 25, the binomial distribution approximates a normal distribution and this approximation is rapidly asymptotic if P is close to 1/2 and slowly asymptotic when P is near 0 or 1. When P is near 0 or 1, Siegel (1956) suggested the rule of thumb that $NP(1 - P)$ must be at least 9 to use the normal approximation. Within these constraints the z-statistic above is approximately normally distributed with zero mean and unit variance, and hence we may cautiously conclude that if z exceeds $\pm$ 1.96 the difference between observed and expected probabilities has reached the .05 level of significance (see also Sackett, 1978). However, because dyadic states in successive time intervals (or simply successive dyadic states in the case of event-sequence data) are likely not independent in the purest sense, it seems most conservative to treat the resulting z simply as an index or score and not to assign p-values to it.

As the reader will note, in the chapter just quoted we were concerned with two issues: the assumption of independence and the number of tallies required to justify use of the binomial distribution. On reflection, we find the argument that the categorizations of successive n-event sequences in event-sequence data are not "independent" less compelling than we did previously, and so we are no longer quite so hesitant to assign probability values on this score.

Our lack of hesitancy rests in part on a simulation study Bakeman and Dorval (1989) performed. No matter the statistic, for the usual sorts of parametric tests, p values are only accurate when assumptions are met. To those encountering sequential analysis for the first time, the common (but not necessarily required) practice of overlapped sampling (tallying first the e_1e_2 chain, then e_2e_3, e_3e_4, etc.) may seem like a violation of independence. The two-event chain is constrained to begin with the code that ended the

previous chain (i.e., if a two-event chain ends in B, adding a tally to the 2nd column, the next must add a tally to the 2nd row), and this violates sampling independence. A Pearson or likelihood-ratio chi-square could be computed and would be an index of the extent to which observed frequencies in this table tend to deviate from their expected ones. But we would probably have even less confidence than usual that the test statistic is distributed as χ^2 and so, quite properly, would be reluctant to apply a p value.

Nonoverlapped sampling (tallying first the e_1e_2 chain, then e_3e_4, e_5e_6, etc.) does not pose the same threat to sampling independence, although it requires sequences almost twice as long in order to extract the same number of two-event chains produced by overlapped sampling. However, the consequences of overlapped sampling may not be as severe as they at first seem. Bakeman and Dorval (1989) found that when sequences were generated randomly, distributions of a test statistic assumed their theoretically expected form equally for the overlapped and nonoverlapped procedures and concluded that the apparent violation of sampling independence associated with overlapped sampling was not consequential.

8.2 Stationarity

The term "stationarity" means that the sequential structure of the data is the same independent of where in the sequence we begin. This means that, for example, we will get approximately the same antecedent/consequent table for the first half of the data as we get for the second half of the data.

	1st Half		2nd Half	
	HNice	HNasty	HNice	HNasty
WNice	80	10	64	7
WNasty	2	60	12	66

We then compute the pooled estimates over the whole interaction:

	HNice	HNasty
WNice	144	17
WNasty	14	126

To test for stationarity of the data, we compare the actual data to the expected values under the null hypothesis that the data are stationary. Let's assume that we are interested in only the lag-1 antecedent/consequent table for s codes. Let *N(IJ, t)* be the joint frequency for cell *IJ* in segment t, and

P(IJ, t) be the transition probability for that cell. Let *P(IJ)* be the pooled transition probability. Then G^2, computed as

$$G^2 = 2\sum_t N(IJ,t)\log_e\left(\frac{P(IJ,t)}{P(IJ)}\right)$$

is distributed as chi-square. If there are s codes, and T segments, then G^2 has degrees of freedom $(T-1)(s)(s-1)$. A more general formula for the rth-order transition table is given by Gottman & Roy (1990, pp. 62–63), where r is the order of the chain. The sum is across segments of the data. For the example data, the value computed was 6.44, which is compared to $df = 2$; this quantity is not statistically significant. Hence, the example data are stationary.

This test can be used to see if the data have a different sequential structure for different parts of the interaction. This appears to be the case for conflict resolution in married couples; the first third is called the "agenda building segment," the second third is called the "disagreement segment," and the final third is called the "negotiation segment" (Gottman, 1979a). However, a problem with this test is that as the number of observations increases, the power we have to detect violations of absolute stationarity increases, and yet, for all intents and purposes the data may be stationary enough.

In this case we can do a log-linear analysis of the data and evaluate the Q^2 statistic, as recommended by Bakeman and Robinson (1994, p. 102 ff.). For example, let C = consequent (husband nice/husband nasty), A = antecedent (wife nice/wife nasty), and T = segment (first half/second half). If the CAT term is required for a fitting model, this suggests that the association between antecedent and consequent varies as a function of time (i.e., is not stationary). For the present data, the loss in fit when this term is removed is significant at the .10 but not the .05 level ($G^2[1] = 3.11$). The G^2 for the base model (i.e., [C][A][T]) is 227.4, so the 3.11 represents less than 1.4% of the total. Even when the loss in fit is significant at the .05 or a more stringent level, Knoke and Burke (1980) recommend ignoring terms that account for less than 10% of the total. This can be useful when even small effects are statistically significant, as often happens when the number of tallies is large.

8.3 Describing general orderliness

The material already presented in the last chapter assumes that investigators want to know how particular behavioral codes are ordered. For example, they may want to confirm that a particular, theoretically important sequence (like Touch, Nurse, Groom) really occurs more often than expected, given

base rates for each of these codes. Or, in a more exploratory vein, they may want to identify whichever sequences, if any, occur at greater than chance rates. However, there is another quite different kind of question investigators can ask, one that concerns not how particular codes in the stream of behavior are ordered, but how orderly the stream of behavior is overall.

In this section, we shall not describe analyses of overall order in any detailed way. Instead, we shall suggest some references for the interested reader, and shall try to give a general sense of what such analyses reveal and how they proceed. Primarily, we want readers to be aware that it is possible to ask questions quite different from those discussed earlier in this chapter.

One traditional approach to the analysis of general order is provided by what is usually called "information theory." A brief explication of this approach, along with appropriate references and examples, is given by Gottman and Bakeman (1979). Although the classical reference is Shannon and Weaver (1949), more useful for psychologists and animal behaviorists are Attneave (1959) and Miller and Frick (1949). A well-known example of information theory applied to the study of social communication among rhesus monkeys is provided by S. Altmann (1965). A closely related approach is called Markovian analysis (e.g., Chatfield, 1973). More recently, problems of gauging general orderliness are increasingly viewed within a log-linear or contingency-table framework (Bakeman & Quera, 1995b; Bakeman & Robinson, 1994; Bishop, Fienberg, & Holand, 1975; Castellan, 1979; Upton, 1978).

No matter the technical details of these particular approaches, their goals are the same: to determine the level of sequential constraint. For example, Miller and Frick (1949), reanalyzing Hamilton's (1916) data concerning trial-and-error behavior in rats and 7-year-old girls, found that rats were affected just by their previous choice whereas girls were affected by their previous two choices. In other words, if we want to predict a rat's current choice, our predictions can be improved by taking the previous choice into account but are not further improved by knowing the choice before the previous one. With girls, however, we do improve predictions concerning their current choice if we know not just the previous choice but the one before that, too.

If data like these had been analyzed with a log-linear (or Markovian) approach, the analysis might have proceeded as follows: First we would define a zero-order or null model, one that assumed that all codes occurred with equal probability and were not in any way sequentially constrained. Most likely, the data generated by this model would fail to fit the observed data. Next we would define a model that assumed the observed probabilities for the codes but no sequential constraints. Again, we would test whether the data generated by this model fit the observed. If this model failed to

fit, we would next define a model that assumed that codes are constrained just by the immediately previous code (this is called a first-order Markov process). In terms of the example given above, this model should generate data that fit those observed for rats but not for girls. Presumably, a model that assumes that codes are constrained by the previous two codes should generate data that pass the "fitness test" for the girl's data.

In any case, the logic of this approach should be clear. A series of models are defined. Each imposes an additional constraint, for example, that the data generated by the model need to take into account the previous code, the previous two codes, etc. The process stops when a particular model generates data similar to what was actually observed, as determined by a goodness-of-fit test. The result is knowledge about the level of sequential constraint, or connectedness, or orderliness of the data, considered as a whole. (For a worked example, analyzing mother–infant interaction, see Cohn & Tronick, 1987.)

8.4 Individual versus pooled data

In the previous chapter, we discussed two different uses of sequential statistics such as z scores (i.e., adjusted residuals and Yule's Q's). First, assuming that successive codings of events are independent of previous codings, and assuming that enough data points are available, we have suggested that z scores can be tested for significance (see section 7.4). Second, when data are collected from several different "units" (e.g., different participants, dyads, or families), we have suggested that scores such as Yule's Q can be used in subsequent analyses of individual or group differences (see section 7.7). Because the familiar parametric techniques (e.g., analyses of variance) are both powerful and widely understood, such a course has much to recommend it.

Not all studies include readily discernible "units" however. For example, just one individual or couple might be observed, or the animals observed might not be easily identifiable as individuals. In such cases, the issue of pooling data across units does not arise; there are not multiple units. In other cases, for example, when several different children are observed, so few data might be collected for each child that pooling data across all children could seem desirable, if for no other reason than to increase the reliability of the summary statistics reported. At the same time, assuming enough data, it might then be possible to test z scores for significance on the basis of the pooled data. Properly speaking, however, any conclusions from such analyses should be generalized just to other behavior of the group observed, not to other individuals in the population sampled.

Thus, even though investigators who pool data over several subjects usually do so for practical reasons, it has some implications for how results are interpreted.

How seriously this last limitation is taken seems to vary somewhat by field. In general, psychologists studying humans seem reluctant to pool data over subjects, often worrying that some individuals will contribute more than others, thereby distorting the data. Animal behaviorists, on the other hand, seem to worry considerably less about pooling data, perhaps because they regard their subjects more as exemplars for their species and focus less on individuality. Thus students of animal behavior often seem comfortable generalizing results from pooled data to other members of the species studied.

As we see it, there are three options: First, when observations do no derive from different subjects (using "subject" in the general sense of "case" or "unit"), the investigator is limited to describing frequencies and probabilities for selected sequences. Assuming enough data, these can be tested for significance. Second, even when observations do derive from different subjects, but when there are few data per subject, the investigator may opt to pool data across subjects. As in the first case, sequences derived from pooled data can be tested for significance, but investigators should keep in mind the limits on interpretation recognized by their field.

Third, and again when observations derive from different subjects, investigators may prefer to treat statistics (e.g., Yule's Q's) associated with different sequences just as scores to be analyzed using standard techniques like t test or the analysis of variance (see Wickens, 1993). In such cases, statistics for the sequences under consideration would be computed separately for each subject. However, analyses of these statistics tell us only whether they are systematically affected by some research factor. They do not tell us whether the statistics analyzed are themselves significant. In order to determine that, we could test individual z scores for significance, assuming enough data, and report for how many participants z scores were significant, or else compute a single z score from pooled data, assuming that pooling over units seems justified.

For example, Bakeman and Adamson (1984), for their study of infants' attention to people and objects, observed infants playing both with their mothers and with same-age peers. Coders segmented the stream of behavior into a number of mutually exclusive and exhaustive behavioral states: Two of those states were "Supported Joint" attention (infant and the other person were both paying attention to the same object) and "Object" attention (the infant alone was paying attention to some object). The Supported Joint state was not especially common when infants played with peers. For

that reason, observations were pooled across infants, but separately for the "with mother" and "with peer" observations.

The z scores computed from the pooled data for the Supported Joint to Object transition were large and significant, both when infants were observed with mother and when with peers. This indicates that, considering these observations as a whole, the Supported Joint to Object sequence occurred significantly more often than expected, no matter the partner. In addition, an analysis of variance of individual scores indicated a significant partner effect, favoring the mother. Thus, not only was this sequence significantly more likely than expected with both mothers and peers, the extent to which it exceeded the expected was significantly higher with mothers, compared to peers.

In general, we suspect that most of our colleagues (and journal editors) are uneasy when data are pooled over human participants. Thus it may be worthwhile to consider how data such as those just described might be analyzed, not only avoiding pooling, but actually emphasizing individuality. The usual parametric tests analyze group means and so lose individuality. Better, it might be argued, to report how many subjects actually reflected a particular pattern, and then determine whether that pattern was observed in more subjects than one might expect by chance.

For such analyses, the simple sign test suffices. For example, we might report the number of subjects for whom the Yule's Q associated with the Supported Joint to Object transition was positive when observed with mothers, and again when observed with peers. If 28 infants were observed, by chance alone we would expect to see the pattern in 14 (50%) of them, but if the number of positive Yule's Q's was 19 or greater ($p < .05$, one-tailed sign test), we would conclude that the Supported Joint to Object transition was evidenced by significantly more infants than expected. And if the Yule's Q when infants were observed with mothers was greater than the Yule's Q when infants were observed with peers for 19 or more infants, we would conclude that the association was stronger when with mothers, compared to peers. The advantage of such a sign-test approach is that we learn, not just what the average pattern was, but exactly how many participants evidenced that pattern.

The approach presented earlier in section 8.2 can also be applied to the issue of pooling over subjects. Again, the question we ask is one of homogeneity, that is, whether the sequential structure is the same across subjects, or groups of subjects (instead of across time as for stationarity). The formula to test this possibility is similar to the formula for stationarity (see Gottman & Roy, 1990, pp. 67 ff.). In the following formula the sum is across $s = 1, 2, \ldots S$ subjects, and $P(IJ)$ represents the pooled joint

probability across the s subjects:

$$G^2 = 2\sum_s N(IJ, s) \log_e \left(\frac{P(IJ, s)}{P(IJ)} \right)$$

The degrees of freedom are $(s-1)$ (NCODES)^r (NCODES-1), where r is the order of the chain (in our case, with lag-1, $r = 1$), where NCODES $=$ the number of codes. Note that this approach is quite general. For example, we can test whether a particular married couple's sequential data is best classified with one group of couples or with another group. Although tedious, this could be a strategy for grouping subjects.

8.5 How many data points are enough?

The issue of the number of tallies – of determining how many data points are enough – remains an important matter. The question the investigator needs to ask is this: How many events need to be coded in order to justify assigning significance to a computed z score associated with a particular cell or a chi-square statistic associated with a table?

At issue is not just the total number of events, but how they are distributed among the possible codes. It is worth reflecting for a moment why – and when – we need be concerned with sufficient tallies for the various codes in the first place. The bedrock reason is stability. If a summary statistic like a chi-square, a z score, or Yule's Q is based on few tallies, then quite rightly we place little confidence in the value computed, worrying that another time another observation might result in quite a different value.

For example, if one of the row sums of a 2×2 table is a very small number (such as 1 or 2), then shifting just one observation from one column to the other can result in a big change in the summary statistic. As a specific example, imagine that 50 observations were classified A|not-A and B|not-B, as follows:

	B	~B	
A	2	0	2
~A	24	24	48
	26	24	50

The Pearson chi-square for this table is 1.92 (so the z score is its square root, 1.39) and its Yule's Q is $+1$. After all, every A (all two of them) was followed by a B. However, if only one of the A's were followed by a B,

	B	~B	
A	1	1	2
~A	24	24	48
	25	25	50

then all statistics (Pearson chi-square, z score, and Yule's Q) would be zero. This example demonstrates summary statistics' instability when only a few instances of a critical code are observed.

To protect ourselves against this source of instability, we compute summary statistics only when all marginals sums are 5 or greater, and regard the value of the summary statistics as *missing* (too few instances to regard the computed values as accurate) in other cases. This is only an arbitrary rule, of course, and hardly confers absolute protection. Investigators should always be alert to the scanty data problem, and interpret results cautiously when summary statistics (e.g., z scores, Yule's Q's) are based on few instances of critical codes.

If stability is the bedrock reason to be concerned about the adequacy of the data collected, correct inference is a secondary but perhaps more often mentioned concern. This matters only when investigators wish to assign a p value to a summary statistic (e.g., a X^2 or a z score) based on assumptions that the statistic follows a known distribution (e.g., the chi-square or normal). Guidelines for amount of adequate data for inference have long been addressed in the chi-square and log-liner literature, so it makes sense to adapt those guidelines – many of which are stated in terms of expected frequencies – for lag-sequential analysis.

Several considerations play a role, so absolute guidelines are as difficult to define as they are desired. Summarizing current advice, Wickens (1989, p. 30) noted that (a) expected frequencies for two-dimensional tables with 1 degree of freedom should exceed 2 or 3 but that with more degrees of freedom some expected frequencies may be near 1 and with large tables up to 20% may be less than 1, (b) the total sample should be at least four or five times the number of cells (more if marginal categories are not equally likely), and (c) similar rules apply when testing whether a model fits a three- or larger-dimensional table.

As noted earlier, when lag 1 effects are studied, the number of cells is K^2 when consecutive codes may repeat and $K(K-1)$ when they cannot. Thus, at a minimum, the total number of tallies should exceed K^2 or $K(K-1)$, as appropriate, times 4 or 5. Additionally, marginals and expected frequencies should be checked to see whether they also meet the guidelines. When lag L effects are studied, the number of cells is K^{L+1} when consecutive codes may repeat and $K(K-1)^L$ when they cannot. As L increases, the product

of these values multiplied by 4 or 5 can become discouragingly large. However, if attention is limited to three-dimensional $0(L-1)L$ tables, as suggested in section 7.6, then the number of cells is no greater than K^3 when consecutive codes may repeat and $K^2(K-1)$ when they cannot. But remember, these values multiplied by 4 or 5 represent a minimum number of tallies. Marginals and expected frequencies still need to be examined. Further, these products provide a guideline for the minimum number of events that should be coded.

Without question, considerable data may be required for a sequential or log-linear analysis. Following the strategy that limits attention to three-dimensional tables (section 7.6), K is the determining factor. For example, the numbers of cells when K is 3, 5, 7, and 9 and consecutive codes may repeat are

	L:1	2
K:3	9	27
5	25	125
7	49	343
9	81	729

(values for higher lags are the same as when $L = 2$; the general formula is K^2 when $L = 1$ and K^3 when $L = 2$ or higher) and when consecutive codes cannot repeat, numbers of cells (excluding structural zeros) are

	L:1	2	3
K:3	6	12	18
5	20	80	100
7	42	252	294
9	72	576	648

(values for higher lags are the same as when $L = 3$; the general formula is $K(K-1)$ when $L = 1$, $K(K-1)^2$ when $L = 2$, and $K^2(K-1)$ when $L = 3$ or higher). Taking into account only the times-the-number-of-cells rule, in order to compute the N needed these values should be multiplied by 4, 5, or whatever factor seems justified (Bakeman & Quera, 1995b). Moreover, as Bakeman and Robinson (1994) note, such guidelines should not be regarded as minimal goals to be satisfied, but as troubled frontiers from which as much distance as possible is desired.

If the numbers still seem onerous, reconsider expected frequencies and note that Haberman (1977; cited in Wickens, 1989, p. 30) suggested that

requirements for tests based on the difference between two chi-squares – such as the tests of $0 \perp\!\!\!\perp L|L-1$ described in section 7.6 – may depend more on frequency per degree of freedom than on the minimum expected cell size. When doubts arise, it may be best to consult local experts. But for many analyses, especially when the number of codes analyzed (K) is large, expect that the number of events coded should number in the thousands, not hundreds.

The guidelines described in the preceding paragraphs assumed that the computed test statistics would be distributed exactly as some theoretical distribution (e.g., the chi-square or normal), thus permitting p values to be based on the appropriate theoretic distribution. Such tests are often called *asymptotic* because as the amount of data on which the test statistic is based increases, its distribution approximates the theoretical one ever more closely. Asymptotic tests may be either parametric like those based on z or nonparametric like those based on chi-square. Permutation tests (Edgington, 1987; Good, 1994), although less well known, provide an alternative. Such tests construct sampling distributions from the data observed. No reference is made to another, theoretic distribution, so no minimum-data assumption to justify the reference is required.

Especially when data are few, investigators should consider permutation tests as an alternative to the usual asymptotic ones. They yield an exact, instead of an asymptotic, p value, and render minimum-data requirements unnecessary. If data are few, statistical significance will still be unlikely, but that is at it should be. Interested readers are urged to consult Bakeman, Robinson, and Quera (1996) for more details concerning permutation tests in a sequential context. But no matter whether asymptotic or permutation tests are used, you still should expect that the number of events coded will often need to number in the thousands.

8.6 The type I error problem

Even when enough data are collected to justify significance testing for the various scores computed, the problem of type I error – of claiming that sequences are "significant" when in fact they are not – remains. The reason type I error is such a problem with sequential analyses such as those described earlier is that typically investigators have many codes. These many codes generate many more possible sequences, especially when anything longer than two-event sequences is considered. The number of sequences tested for significance can rapidly become astronomical, in which case the probability of type I error (the alpha level) approaches certainty. When tests are independent, and the alpha level for each is .05, the "investigationwise"

or "studywise" alpha level, when k tests are performed, is $1-.95^k$ (Cohen & Cohen, 1983). Thus if 20 independent tests are performed, the probability of type I error is really .64, not .05.

What each study needs is a coherent plan for controlling type I error. The best way, of course, is to limit drastically the number of tests made. And even then, it makes sense to apply some technique that will assure a desired studywise alpha level. For example, if k tests are performed and a studywise alpha level of .05 is desired, then, using Bonferroni's correction, the alpha level applied to each test should not be alpha, but rather alpha divided by k (see Miller, 1966). Thus if 20 tests are performed, the alpha level for each one should be .0025 (.05/20).

When studies are confirmatory, type I error usually should not be a major problem. Presumably in such cases the investigator is interested in (and will test for significance) just a few theoretically relevant sequences. Exploratory studies are more problematic. Consider the parallel play study discussed earlier. Only five codes were used, which is not a great number at all. Yet these generate 20 possible two-event and 80 possible three-event sequences. This makes us think that unless very few codes are used (three or four, say) and unless there are compelling reasons to do so, most exploratory investigations should limit themselves to examining just two-event sequences, no longer – even if the amount of data is no problem.

Even when attention is confined to two-event sequences, the number of codes should likewise be limited. For two-event sequences, the number of possible sequences, and hence the number of tests, increase roughly as the square of the number of codes. For this reason, we think that coding schemes with more than 10 codes border on the unwieldy, at least when the aims of a study are essentially exploratory.

Two ways to control type I error were described in section 7.6 when discussing log-linear approaches to lag-sequential analysis. First, exploratory studies should not fish for effects at lag L in the absence of significant lag L omnibus tests. And second, the set of seemingly significant sequences at lag L should be winnowed into a smaller subset that can be viewed as responsible for the model of independence's failure to fit. Still, as emphasized in section 7.7 when discussing Yule's Q, guiding ideas provide the best protection against type 1 error. Investigators should always be alert for ways to limit the number of statistical tests in the first place.

8.7 Summary

Several issues important if not necessarily unique to sequential analysis have been discussed in this chapter. Investigators should always worry

whether summary indices (means, Yule's Q's, etc.) are based on sufficient data. If not, confidence in computed values and their descriptive value is seriously compromised. Further, when inferential statistics are used, data sufficient to support their assumptions are required. Guidelines based on log-linear analyses were presented here, but the possibility of permutation tests, which require drastically fewer assumptions, was also mentioned. Again, investigators should always limit the number of statistical tests in any study, else they court type I error. Of help here is the discipline provided by guiding ideas and theories, clearly stated. In contrast, issues of pooling may arise more in sequential than other sorts of analyses because of data demands. Pooling data over units such as individuals, dyads, families, etc., is rarely recommended, no matter how necessary it seems. When data per unit are few, a jackknife technique (computing several values for a summary statistic, each with data for a different unit removed, then examining the distribution for coherence) is probably better than pooling. Finally, common to almost all statistical tests is the demand for independence (or exchangeability; see Good, 1994). When two-event chains are sampled in an overlapping manner from longer sequences, this requirement might seem violated, but simulation studies indicate that the apparent violation in this particular case does not seem consequential.