6

Analyzing sequential data: First steps

6.1 Describing versus modeling

If all the steps described in previous chapters – developing coding schemes, recording behavioral sequences reliably, representing the observational data – are in order, then the first fruits of the research should be simple description. Introductory textbooks never tire of telling their readers that the basic tasks of psychology are, one, description, and two, explanation. Similarly, almost all introductory textbooks in statistics distinguish between descriptive statistics, on the one hand, and inferential statistics, on the other. This distinction is important and organizes not just introductory statistics texts but this and the next four chapters as well.

Much of the material presented in this and the following four chapters, however, assumes that readers want first to describe their data, and so description is emphasized. Problems of inference and modeling – determining if data fit a particular model, estimating model parameters – are touched on only slightly here. These are important statistical topics and become especially so when one wants to move beyond mere description to a deeper understanding of one's data. That is why so many books and courses, indeed huge specialized literatures, are devoted to such topics. We assume that readers will use scores derived from observing behavioral sequences as input for anything from simple chi-square or analyses of variance, to log-linear modeling, to the modeling approach embodied in programs like LISREL.

Our task, fortunately, is not to describe all the modeling possibilities available. Instead, we have set ourselves the more manageable task of discussing how to derive useful descriptive scores from sequential data. Still, we describe some simple instances of model testing and try to point out when sequential data analysis presents particular problems for statistical inference. Throughout, we attempt to maintain the distinction between description and modeling. Thus we would never talk of "doing" a Markov analysis. Instead we would describe how to compute transitional probabilities, on the one hand, and how to determine if those transitional probabilities fit a particular Markov model, on the other (see section 8.3).

In the remainder of this chapter, we note some of the simpler descriptive statistics that can be derived from sequential observational data.

6.2 Rates and frequencies

Some statistics are so obvious that it almost seems an insult to the reader's intelligence to mention them at all. Still, they are so basic, and so useful, that omitting them would seem negligent. Certainly, logical completeness demands their inclusion. One such statistic is the rate or frequency with which a particular event occurred.

What is recorded is the *event*. This could be either a momentary or a duration event. Just the occurrence of the event might be recorded, or its onset and offset times might be recorded as well, or the occurrence of events might be embedded in another recording strategy such as coding intervals or cross-classifying events.

What is tallied is the *frequency* – how often a particular event occurred. The data could be represented as simple event sequences, as state sequences, as timed-event sequences, or as cross-classified events. In all these cases, it is possible to count how often particular events occurred. Only interval sequences pose potential problems. Two successive intervals that contain the same code may or may not indicate a single bout; thus frequencies of intervals that contain a particular code may overestimate the number of instances, and this needs to be taken into account when interpreting frequency data for interval sequences.

In most cases, raw frequencies should be transformed to *rates*. frequencies, of course, depend on how long observation continues, whereas rates have the merit of being comparable across cases (individuals, days, etc.). The total observation time needs to be recorded, of course, even for event sequences, which otherwise do not record time information; otherwise rates cannot be computed. For example, Adamson and Bakeman (1985) observed infants at different ages (6, 9, 12, 15, and 18 months) and with different partners (with mothers, with peers, and alone). A coder recorded whenever infants engaged in an "affective display." For a variety of reasons, the observation sessions (with each partner, at each age) varied somewhat in length. Thus Adamson and Bakeman divided the number of affective displays (for each observation session) by the observation time (in their case, some fraction of an hour), yielding a rate per hour for affective displays. This statistic, just by itself, has descriptive value (there were 59 affective displays per hour, on the average), but further, it can be (and was) used as a score for traditional analysis of variance. Thus Adamson and Bakeman were able to report that the rate of affective displays was greater when infants were with mothers instead of peers, and that this rate increased with age.

6.3 Probabilities and percentages

A second obvious and very useful descriptive statistic is the simple probability or percentage. It can be either event-based or time-based. When event-based, the simple probability or percentage tells us what proportion of events were coded in a particular way, relative to the total number of events coded. Initial procedures are the same as for rates. Events are recorded (using any of the recording strategies described in chapter 3) and the frequencies tallied. The only assumption required is that codes be mutually exclusive and exhaustive. Then the number of events coded in a particular way is divided by the total number of events coded. This gives the simple probability for that particular kind of event. Or, the quotient can be multiplied by 100, which gives the percentage for that particular kind of event. For example, in the Bakeman and Brownlee study of parallel play, 12% of the events were coded Unoccupied, 18% Solitary, 26% Together, 24% Parallel, and 21% Group. (Recall that they used an interval-coding strategy; thus a single event was defined as any contiguous intervals coded the same way.)

When simple probabilities or percentages are time based, the interpretation is somewhat different. These widely used statistics convey "time-budget" information, that is, they indicate how the cases observed (animals, infants, children, dyads, etc.) spent their time. The recording strategy must preserve time information; thus simple event coding would not work, but recording onset and offset times of events, or even coding intervals (remembering the approximate nature of the statistics estimated) would be fine. Similarly, data would need to be represented as state, timed-event, or interval sequences, not simple event-sequences. (Cross-classified event data would also work if duration had been recorded, but frequently when this recording and representation approach is used, the proportion of time devoted to the event being cross-classified may not be recorded.) One final note: When proportion of time coded in a particular way is of interest, codes need not be mutually exclusive and exhaustive.

The Bakeman and Brownlee study of parallel play, just used as an example of event-based probabilities, also provides an example of time-based probabilities, or percentages as well. Using an interval-coding strategy, they reported that 9% of the 15-second intervals were coded Unoccupied, 25% Solitary, 21% Together, 28% Parallel, and 17% Group. Thus, for example, although 24% of the events were coded Parallel play, Parallel play occupied 28% of the time. (These are estimates, of course. Recording onset times for behavioral state changes would have resulted in more accurate time-budget information than the interval-recording strategy actually used; see section 3.7.) A second example could be provided by the Adamson and Bakeman study of affective displays. They recorded not just the occurrence of affective displays, but their onset and offset times as well. Thus we were able to compute that affective displays occurred, on the average, 4.4% of the time during observation sessions. Put another way, the probability that the infant would display affect in any given moment was .044.

Event-based (e.g., proportion of events coded Solitary) and time-based (e.g., proportion of time coded Solitary) probabilities or percentages provide different and independent information; there is no necessary correlation between the two. Which then should be reported? The answer is, it depends. Whether one or both are reported, investigators should always defend their choice, justifying the statistics reported in terms of the research questions posed.

6.4 Mean event durations

Whenever time information is recorded, mean event durations can be reported as well as (or instead of) proportions of total time devoted to particular kinds of events. In fact, mean event durations provide no new information not already implied by the combination of rates and time-based percentages. After all, mean event (or bout, or episode) durations are computed by dividing the amount of time coded for a particular kind of event by the number of times that event was coded. But in some cases, mean event durations may be more useful descriptively than time-based probabilities or percentages.

Because of the clear redundancy among these three descriptive statistics (rates or frequencies, time-based probabilities or percentages, mean event durations), we think investigators should report, or at least analyze, only two of them. The question then is, which two? The answer will depend on whatever an investigator thinks most useful descriptively, given the sort of behavior coded. However, we suspect that when behavioral states are being coded (see section 3.2), time-based percentages and mean durations are more useful than rates, but that when the events being coded occur just now and then and do not exhaustively segment the stream of behavior as behavioral states do, then rates and mean event durations may prove more useful than time-based percentages.

For example, although in the previous section we computed the (time-based) probability of an affective display from the Adamson and Bakeman data, they in fact reported just rates and mean durations for affective displays. As with rates, mean durations were used as a score for subsequent analysis of variance. Thus Adamson and Bakeman reported, not just that the average length of an affective display was 2.5 seconds, but that the length became shorter as the infants became older. This was an effect Adamson and Bakeman had predicted on the basis of a hypothesized shift in the function of affective displays during these infant ages.

6.5 Transitional probabilities: An introduction

The statistics discussed in the three preceding sections can be extremely useful for describing aspects of sequential observational data (and all can be computed with the GSEQ program), but they do not themselves convey anything uniquely sequential. Perhaps the simplest descriptive statistic that does capture a sequential aspect of such data is the transitional probability; however before transitional probabilities can be described, some definitions are in order.

A simple (or unconditional) probability is just the probability with which a particular "target" event occurred, relative to a total set of events (or intervals, if time based). For example, if there were 20 days with thunderstorms last year, we could say that the probability of a thunderstorm occurring on a particular day was .055 (or 20 divided by 365).

A conditional probability, on the other hand, is the probability with which a particular "target" event occurred, relative to another "given" event. Thus if it rained 76 days last year, and if on 20 of those 76 days there were thunderstorms, then we would say that the probability of a thunderstorm occurring, given that it was a rainy day, was .263 (or 20 divided by 76). If T stands for thunderstorms and R for a rainy day, then the simple probability for thunderstorms is usually written $p(T)$, whereas the conditional probability for thunderstorms, given a rainy day, is usually written $p(T|R)$; in words, this is "the probability of T, given R."

A transitional probability is simply one kind of conditional probability. It is distinguished from other conditional probabilities in that the target and given events occur at different times. Often the word "lag" is used to indicate this displacement in time. For example, if data are represented as event sequences, then we might want to describe the probability, given event A, of the target event B occurring immediately after (lag 1), occurring after an intervening event (lag 2), etc. These event-based transitional probabilities can be written $p(B_{+1}|A_0)$, $p(B_{+2}|A_0)$, etc.

Similarly, if data are represented as state, timed-event, or interval sequences, then we might want to describe the probability, given event A, of the target event B occurring in the next time interval (often written $t + 1$, where t stands for time), in the time interval after the next ($t + 2$), etc. Such time-based transitional probabilities are often written $p(B_{t+1}|A_t)$, $p(B_{t+2}|A_t)$, etc.

For example, imagine the following data (assume that each letter stands for an event and that the coding scheme contains three mutually exclusive and exhaustive codes):

$$B\ C\ A\ A\ A\ B\ B\ C\ B\ C\ A\ C$$

In this case, it turns out that each code occurred four times. That is, $f(A) = f(B) = f(C) = 4$ (where "f" stands for frequency), and $p(A) = p(B) = p(C) = 4/12 = .33$, because 12 events were coded in all. The transitional frequency matrix for these data is given in Figure 6.1. Note that the labeling of rows and columns is somewhat arbitrary. We could just as well have labeled rows lag-1 and columns lag 0. Either way, rows refer to events occurring earlier, columns to events occurring later in time. This is the usual convention (probably because we are a left-to-right reading society) and one we recommend following.

Transitional frequency matrices are easy to construct. Each cell indicates the number of times a particular transition occurred. For example, B was followed by C three times in the sequence given in the preceding paragraph; thus the cell formed by the Bth row and the Cth column in the lag-one frequency matrix contains a "3." Symbolically, $f(C_{+1}|B_0) = 3$, or $f_{BC} = 3$; this if often written x_{BC}, letting x represent a score in general or, less mnemonically but more conventionally, x_{23}. Note that if N successive events or intervals are coded, then there will be $N - 1$ lag-one transitions, $N - 2$ lag-two transitions, etc. (The reader may want to verify the other frequencies for the transitional frequency matrices given in Figure 6.1.)

To tally transitions, we use a "moving time-window." For the lag 1 transition matrix, we slide the two-event moving time-window along as follows:

$$(B\ C)\ A\ A\ A\ B\ B\ C\ B\ C\ A\ C$$

Then,

$$B\ (C\ A) A\ A\ B\ B\ C\ B\ C\ A\ C$$

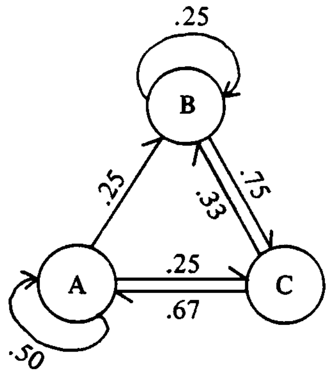
Then,

$$B\ C\ (A\ A)\ A\ B\ B\ C\ B\ C\ A\ C$$

etc. The first position adds a tally to the x_{BC} cell of the first table in Figure 6.1. The second position of the window adds a tally to the x_{CA} cell, and so forth. Note that the consequent code takes its turn next as an antecedent. This raises questions of independence of tallies, and may

Data: B C A A A B B C B C A C

Lag 0 \ Lag 1	A	B	C	
A	2	1	1	4
B	0	1	3	4
C	2	1	0	3
				11



Data: A B A B C B A C B C A C

Lag 0 \ Lag 1	A	B	C	
A	0	2	2	4
B	2	0	2	4
C	1	2	0	3
				11

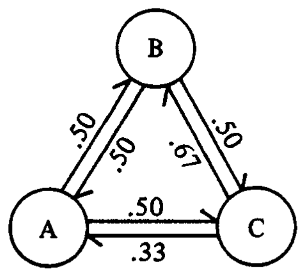


Figure 6.1. Examples of transitional frequency matrices and state transition diagrams.

matter for some subsequent inferential statistical procedures; this issue is discussed further in section 8.1, although Bakeman and Dorval (1989) believe it is not practically consequential.

Similarly, transitional probability matrices are easy to compute. For example, for the data sequence given earlier, the lag 1 transitional probability matrix would be as given in Table 6.1. As the reader can see, a

Table 6.1. *Transitional probability matrix for the first data sequence given in Figure 6.1*

	Lag 1		
Lag 0	A	B	C
A	.50	.25	.25
B	.00	.25	.75
C	.67	.33	.00

transitional probability is just the frequency for a particular cell divided by the frequency for that row. (A consequence of this definition is that the transitional probabilities in each row sum to 1.) Symbolically,

$$t_{ij} = \frac{x_{ij}}{x_{i+}}$$

For example, the probability of code C occurring, given that code B just occurred, is $p(C_{+1}|B_0) = t_{BC} = x_{BC} \div x_{B+} = 3/4 = .75$. This means that 75% of the time, code C followed code B.

Transitional probabilities are often presented graphically, as state transition diagrams (for examples, see Bakeman & Brown, 1977; Stern, 1974). Such diagrams have the merit of rendering quite visible just how events (or time intervals) were sequenced in time. Circles represent the codes, and arrows represent the transitional probabilities among them. Examples are given in Figure 6.1, and the reader may want to verify that they were drawn and labeled correctly.

Figure 6.1 contains a second data sequence, along with its associated transitional frequency matrix and state transition diagram. For both, the simple probabilities are the same, that is, for both sequences each different code occurred four times. The point of presenting these two examples is to show that, even when simple probabilities indicate no differences, events may nonetheless be sequenced quite differently. And when they are, transitional probabilities and their associated state transition diagrams can reveal those differences in a clear and informative way.

One final point: The discussion here treats transitional probabilities as simple descriptive statistics, and state transition diagrams as simple descriptive devices. Others (e.g., Kemeny, Snell, & Thompson, 1974) discuss transitional probabilities and state transition diagrams from a formal, mathematical point of view, as parameters of models. Interesting questions for them are, given certain models and model parameters, what sorts of outcomes would be generated? This is a formal, not an empirical exercise, one in which data are generated, not collected. For the scientist, on the other hand, the usual question is, first, can I accurately describe my data, and second, does a particular model I would like to support generate data that fit fairly closely with the data I actually got? The material in this chapter, as noted in its first section, is concerned mainly with the first enterprise – accurate description.

6.6 Summary

This "first steps" chapter discusses four very simple, very basic, but very useful, statistics for describing sequential observational data: rates (or frequencies), simple probabilities (or percentages), mean event durations, and transitional probabilities.

Rates indicate how often a particular event of interest occurred. They are probably most useful for relatively momentary and relatively infrequent events like the affective displays Adamson and Bakeman (1985) described. Simple probabilities indicate what proportion of all events were of a particular kind (event based) or what proportion of time was devoted to a particular kind of event (time based). Time-based simple probabilities are especially useful when the events being coded are conceptualized as behavioral states (see section 3.2). Indeed, describing how much time individuals (dyads, etc.) spent in various activities (or behavioral states) is a frequent goal of observational research.

With the knowledge of how often a particular kind of event occurred, and what percentage of time was devoted to it, it is possible to get a sense of how long episodes of the event lasted. Mean event durations can be computed directly, of course. But because these three statistics provide redundant information, the investigator should choose whichever two are most informative, given the behavior under investigation.

Finally, transitional probabilities capture sequential aspects of observational data in a simple and straightforward way. They can be presented graphically in state transition diagrams. Such diagrams have the merit of rendering visible just how events are sequenced in time.