10

Analyzing cross-classified events

10.1 Describing cross-classified events

Earlier, in chapter 3, we distinguished between continuous and intermittent recording strategies, noting that because this is a book about sequential analysis, we would emphasize continuous approaches. Four of these strategies (coding events, recording onset and offset times, timing pattern changes, coding intervals) are continuous both in the sense that observers are continuously alert, ready to record whenever required, and in the sense that the data recorded reflect a continuous record of some aspect of the passing stream of behavior. Such data are typically represented as event, state, timed-event, or interval sequences (see chapter 5). We also mentioned a fifth strategy (cross-classifying events), which is continuous only in the sense that observers are continuously alert, ready to record information about certain events whenever they occur. This strategy results in sequential data only if the information coded represents logically sequential aspects of the event.

This is an appealing strategy for a number of reasons. When an investigator knows beforehand that a particular kind of event is of interest, this approach provides focused information about that event with little extraneous detail. Moreover, the data can be represented and analyzed in simple, well-known, and quite straightforward ways.

As an example, let us again use the Bakeman and Brownlee (1982) study of object struggles, referred to earlier. The event of interest, of course, was an object conflict or, to put it in more neutral terms, a "possession episode." As they defined it, a possession episode required that one child (the taker) touch and attempt to take away an object currently being held by another child (the holder). Observations took place in both a toddler (ages ranged from 12 to 24 months) and a preschool (ages ranged from 40 to 48 months) classroom. In each, a target or focal child strategy was used, that is, observers would focus first on one child for a period of time (in this case, 5 minutes), then on another, switching their attention from child to child according to a predetermined random order. Each time the current

Table 10.1. *Resistance cross-classified by prior possession and dominance for toddlers and preschoolers*

Age group	Dominance	Prior possession	Resistance Yes	Resistance No
Toddlers	Yes	Yes	19	7
		No	42	30
	No	Yes	16	4
		No	61	13
Preschoolers	Yes	Yes	6	5
		No	18	5
	No	Yes	9	6
		No	27	4

Note: Counts for the toddlers are based on 20.7 hours of observations; for the preschoolers, on 16.6 hours; see Bakeman and Brownlee (1982).

target child was involved in a possession episode, observers recorded (a) the name of the taker and the name of the holder; (b) whether the taker had had "prior possession," that is, had played with the contested object within the previous minute; (c) whether the holder resisted the attempted take; and (d) whether the taker was successful in gaining the object.

In addition, a linear dominance hierarchy was determined for the children in each class and so, for each episode, it was possible to determine whether the taker was dominant to the holder. Thus each episode was cross-classified as follows: taker dominant (yes/no), taker had prior possession (yes/no), taker resisted (yes/no), and taker successful (yes/no). These were viewed as sequential, in the order given, although, properly speaking, dominance is a "background variable" and not a sequential aspect of the possession episode.

Bakeman and Brownlee chose to regard both resistance and success as outcomes, as "response variables" to be accounted for by the "explanatory variables" of dominance and prior possession. As a result, they presented their data as a series of 2 × 2 × 2 tables (dominance × prior possession × success or dominance × prior possession × resistance) instead of including resistance and success as dimensions in the same table. Because our purpose here is to discuss, in a general way, how to describe (and analyze) cross-classified event data, we shall use an an example just data on dominance, prior possession, and resistance, ignoring success.

Describing cross-classified event data is almost embarrassingly simple. According to time-honored tradition, such data are presented as contingency tables (see Table 10.1). Data can be presented as raw counts or tallies as here, or various percentages can be computed. For example, 72% of the possession episodes met with resistance among toddlers; the corresponding percentage for preschoolers was 75%, almost the same. We could then go on and report percentages of possession episodes encountering resistance when the taker had had prior possession, was dominant, etc., but a clearer way to present such data is as conditional probabilities.

The simple probability that a take attempt would encounter resistance was .72 among toddlers. The conditional probability for resistance, given that the toddler taker had had prior possession, $p(R/P)$, was actually somewhat higher, .76, but the conditional probability for resistance, given that the taker was dominant, $p(R/D)$, dropped to .62. This suggests that, at least among toddlers, the likelihood that the taker will encounter resistance is affected by dominance but not by prior possession.

Among preschoolers, the situation appears reversed. For them, the simple probability that a take attempt would meet with resistance was .75. The conditional probability for resistance, given that the preschooler taker had had prior possession, was only .58, whereas the conditional probability for resistance, given that the taker was dominant was .71, close to the simple or unconditional probability for resistance. Thus preschool takers were less likely to encounter resistance when attempting to "take back" a contested object – as though prior possession conferred some current right. However, this presumed "right" was evident only among the preschoolers, not the toddlers. (The reader may want to verify that the conditional probability values were correctly computed from the data given in Table 10.1.)

This is only descriptive data, however. The question now is, are the differences between simple and conditional probabilities just described larger than one would expect, based on a no-effects or chance model?

10.2 Log-linear models: A simple example

Cross-classified categorical data like those in Table 10.1 can be analyzed relatively easily with what is usually called a log-linear modeling approach (e.g., see Bakeman & Robinson, 1994; Wickens, 1989; for applications see Bakeman, Adamson, & Strisik, 1989, 1995; Bakeman, Forthman, & Perkins, 1992). In principle, this approach is fairly simple and flexible. In addition, results can be expressed in familiar analysis-of-variance-like terminology. Finally, few assumptions need to be made about the data. For all these reasons, this approach, or some variety of it, can prove useful.

Although there are a variety of different kinds of models (e.g., log-linear, logistic, etc.), the general approach is the same. The investigator defines a set of hierarchical modes ("hierarchical" in the sense that simpler models are proper subsets of more complex ones). The simplest model – the null or equiprobale model – contains no terms at all and generates the same expected value for each cell in the contingency table. The most complex model – the saturated model – contains sufficient terms to generate expected values for each cell that are identical to the values actually observed. The idea is to find the least complex model that nonetheless generates expected values not too discrepant from the observed ones, as determined by a goodness-of-fit test. Sometimes the investigator begins with the null or some other simple model. If the data generated by it fail to fit the actual data, then more complex models are tried. If all else fails, the saturated model will always fit the data because it generates values identical to the observed ones. Alternatively, and more typically, the investigator begins with the saturated model and deletes terms until the simplest fitting model is found.

The simplest example of this logic is provided by the familiar chi-square test of independence, although introductory textbooks rarely present it in these terms. The model typically tested first – the no-interaction model – consists of two terms, one for the row variable and one for the column variable. In other words, the row and the column totals for the data generated by the model are forced to agree with the row and column totals actually observed. In introductory statistics, students almost always learn how to compute these expected frequencies, although they are rarely taught to think of them as generated by a particular model.

For example, if we look just at the 2×2 table detailing prior possession and resistance for dominant toddlers (shown in the upper right-hand corner of Table 10.1), most readers would have no trouble computing the expected frequencies. The total tally for row 1 is 26 ($19 + 7$), for row 2, 72; similarly, the total for column 1 is 61 ($19 + 42$), for column 2, 37; the grand total is 98. Thus the expected frequencies are 16.2 (61 times 26 divided by 98) and 9.8 for row 1 and 44.8 and 27.2 for row 2. In this case, the generated data fit the observed quite well, and the chi-square would be small. In the context of a chi-square analysis just of this 2×2 table, the investigator would conclude that, among dominant toddlers, prior possession and resistance are not related.

However, if chi-square were big, meaning that data did not fit this particular model, the investigator would need to invoke a more complex model. In the case of a 2×2 table, the only model more complex than the no-interaction model is the saturated model, which includes (in addition to the row and column effects) an interaction term. To recapitulate, the expected frequencies computed for the common chi-square test of independence are in fact generated by a no-interaction model. If the chi-square is significantly large, then this model fails to fit the data. An interaction term is required, which means that the row and column variables are not independent of each other, but are in fact associated. Often this lack of independence is exactly what the investigator hopes to demonstrate.

Following usual notational conventions (e.g., Fienberg, 1980), the no-interaction model mentioned in the previous paragraph would be indicated as the [R] [C] model, meaning that cell values are constrained by the model to reflect just the row and column totals. The saturated model would be represented as [R] [C] [RC], meaning that in addition to the row and column constraints, the cell totals must also reflect the row × column [RC] cross-classification totals. In the case of a two-dimensional table, this means that the model is saturated and in fact generates data identical with the observed. Because meeting the [RC] constraints means that necessarily the [R] and [C] constraints are met, the saturated model is usually indicated simply by [RC]. In general, to simplify the notation, "lower"-level terms implied by higher-order terms are usually not included when particular models are described. They are simply assumed.

At this point, we hope the reader has a clear idea of the logic of this approach, at least with respect to two-dimensional contingency tables. The situation with respect to tables with three or more dimensions is somewhat more complex. For one thing, computing expected frequencies for various models is not the simple matter it is with two-dimensional tables; usually, computer programs are used. For another, the notation can become a little confusing. Our plan here is to demonstrate the approach with the three-dimensional dominance × prior possession × resistance tables for toddlers and preschoolers, as given in Table 10.1. This hardly exhausts the topic, but we hope it will given interested readers a start, and that they will then move on to more detailed and specialized literature (e.g., Bakeman & Robinson, 1994).

For the 2 × 2 × 2 tables for the toddlers and preschoolers, it seems clear that the null or equiprobable model (all cells have the same value) would not fit these data. In fact, we would probably not test the null model first, but, as with the chi-square test of independence, would begin with a more complex one. Earlier we said that resistance (R) was regarded as a response or outcome variable, and that dominance (D) and prior possession (P) were regarded as explanatory variables. That is, substantively we want to find out if dominance and/or prior possession affected resistance. Thus we would begin testing with the [R] [DP] model. This model simply constrains the generated data to reflect the resistance rates and the dominance cross-classified by prior possession rates actually observed. In particular, it does not contain any terms that suggest that either dominance, or prior possession, or their interaction, is related to the amount of resistance encountered.

In analysis-of-variance terms, the [*R*][*DP*] model is the "no effects" model. In a sense, the [*DP*] term just states the design, whereas the fact that the response variable, [*R*], is not combined with any of the explanatory variables indicates that none affect it. If the [*R*][*DP*] model failed to fit the data, but the [*RD*][*DP*] model did, we would conclude that there was a main effect for dominance – that unless dominance is taken into account, we fail to make very good predictions for how often resistance will occur. Similarly, if the [*RP*][*DP*] model fit the data, we would conclude that there was a main effect for prior possession. If the [*RD*][*RP*][*DP*] model fit, main effects for both dominance and prior possession would be indicated. Finally, if only the [*RDP*] model fit the data (the saturated model), we would conclude that, in order to account for resistance, the interaction between dominance and prior possession must be taken into account.

In the present case, the no-effects model failed to fit the observed data. For both toddlers and preschoolers, the chi-square comparing generated to observed was large and significant (values were 11.3 and 7.2, $df = 3$, for toddlers and preschoolers, respectively; these are likelihood-ratio-chi-squares, computed by Bakeman & Robinson's [1994] ILOG program). However, for toddlers the [*RD*][*DP*] model, and for preschoolers the [*RP*][*DP*] model generated data quite similar to the observed (chi-squares were 1.9 and 0.8, $df = 2$, for toddlers and preschoolers, respectively; these chi-squares are both insignificant, although in both cases the difference between them and the no-effects model is significant). This is analogous to a main effect for dominance among toddlers and a main effect for prior possession among preschoolers. In other words, the dominance of the taker affected whether his or her take attempt would be resisted among toddlers, but whether the taker had had prior possession of the contested object or not affected whether he or she would meet with resistance among preschoolers. Thus the effects noted descriptively in the previous section are indeed statistically significant. Bakeman and Brownlee (1982) interpreted this as evidence for shared possession rules, rules that emerge somewhere between 2 and 3 years of age.

10.3 Summary

Sometimes investigators who collect observational data and who are interested in sequential elements of the behavior observed seem compelled both to obtain a continuous record of selected aspects of the passing stream of behavior and to analyze exhaustively that continuous record. An alternative is to record just selected aspects of certain kinds of events. The kind of event is defined beforehand (for example, possession struggles) as well as the aspects of interest. Each aspect corresponds to a codable dimension. In the case of the possession struggles described in the previous sections, the dimensions were dominance, prior possession, and resistance. The codes for each dimension were the same (either yes or no), but in other cases the mutually exclusive and exhaustive codes for the various dimensions could well be different. In all cases, cross-classified event data result.

A major advantage of such data is, first, cross-classified categorical data are hardly exotic or new, and second, ways of analyzing such data are relatively well understood (e.g., Bakeman & Robinson, 1994; Kennedy, 1983, 1992; Upton, 1978; Wickens, 1989). In this chapter, a simple example was given, showing how log-linear results can be expressed in analysis-of-variance-like terms. The technique, however, is not confined just to contingency tables, but can also be applied to event-sequence and time-sequence data; this was mentioned earlier in section 7.6. The interested reader is advised to consult Bakeman and Quera (1995b) and Castellan (1979).