I want to begin by raising some general questions about the roles of the concept of 'diversity' in archaeological method and theory. The concept has been current for some time in biology, especially in ecology, where there has been time to use it, refine it, and criticize it. Unquestionably it is also useful for archaeologists, but we must not simply import it from the ecologists. We must think rather carefully about just how and why it will be useful for us. Most of the papers in this volume give evidence of such thought, but there are a few that seem unduly bent on applying diversity statistics simply because it is something other archaeologists are doing. There are several cases in which the measures of 'richness' and 'evenness' add little, if anything, to what can be learned from more straightforward approaches to the data presented.

Above all, we must pay more attention to differences in the sorts of entities to which measures of diversity may apply. For the ecologists, it appears that the entities have mainly been the populations of individual organisms existing within some specified tract at a certain time. Basic data are the species (or other taxonomic units) represented, and counts or estimates of the numbers of individuals in each species. In this volume, Rindos often seems to have some such set of entities in mind, but few other chapters are about such entities. Instead, most are about archaeological data sets, and the basic data are the archaeological categories represented and the numbers of examples of each category in various deposits. This difference from ecological studies probably has little importance for the specific mathematical techniques that
might be applied, but it makes a tremendous difference in the meaning of the numbers. The essential point is that diversity in an archaeological data set is an aspect of what Binford (1983a and elsewhere) calls "stries." Diversity in the archaeological record often bears a complex relationship to diversity (or any other aspect) of ancient behavior (Binford's "dynamics"). I do not know how much agreement about the meaning of species diversity there is among ecologists. In any case, that is their concern. We, as archaeologists, do not need simply to develop techniques to measure diversity in archaeological data sets. We must also think much more about the meanings of such diversity.

In the following section, I will leave the problem of meanings aside temporarily and discuss some technical aspects of measuring diversity in the archaeological record. With this preparation, I will then elaborate on some aspects of giving meaning to archaeological diversity. Finally I will take up problems raised by some of the chapters.

Diversity in archaeological data sets

One obvious difference between our data and those of biologists is that our categories are far more problematic. Taxonomic difficulties may seem large to biologists, but I think ours is far greater. If inappropriately defined categories are used, no amount of mathematical finesse can remedy the trouble. There is, of course, no single "right" set of categories for all purposes. The essential thing is that, for a given purpose, we must not lump in ways that obscure important differences and we must not split on the basis of differences that are irrelevant for that purpose. In practice, there are strong pressures to accept categories that have been used previously, often either for obscure reasons or because they work well for time-space systems, and to use these categories for other purposes, such as inferring ancient activities. There are some examples of this kind of use of inappropriate categories in this volume. It is hard to avoid doing this because, at best, to use more relevant categories requires restudy of collections. Often the collections are no longer available, so that one has no choice but to try to make some sense of categories designed for other purposes, or else reject the data as useless for studies of diversity.

The matter of techniques and concepts for arriving at good categories for specific purposes is a large topic (Whallon and Brown 1982), and of great importance, although it receives little analysis in this volume. For that reason, and to keep my own chapter shorter, I too will only underscore its importance and move on to other things.

A central theme that is emphasized in many chapters is that small samples misrepresent assemblage diversity. Richness of categories will be underrepresented, and evenness in the numbers of individuals in the categories is also likely to be lower in the sample than in the parent population. Thus, comparisons between data sets are deeply flawed and likely to be highly misleading if the sizes of the data sets are ignored. However, all is not lost. Although corrections for sample size must remain probabilistic, there are ways of making such corrections, and they often permit one to make comparisons that can be accepted, if not with perfect certainty, then at least with a satisfactorily high level of confidence. Grayson and Kintigh have alternative approaches to this problem, and they and other contributors give a number of examples of the application of these techniques.

Kintigh's method is an appealing application of the "Monte Carlo" approach, wherein a problem that it does not seem feasible to solve by exact methods is solved, to a satisfactory approximation, by probabilistic simulation. I have only a few suggestions. One is that I am teased by the suspicion that there may be an easier approximation, not requiring simulation, that is just as accurate, and that would require substantially less computer power or time. The exact solution can be found, in principle, by expanding the expression \( p_1 + p_2 + p_3 + \ldots + p_n \), where \( m \) is the number of categories and \( n \) is the sample size, and finding the cumulative probability of all the terms for which any given number of the \( p_i \)'s have a zero exponent. If \( m \) and \( n \) are even moderately large, however, the full expansion will involve an enormous number of terms. For example, for \( n = 30 \) and \( m = 10 \), there will be about 20 trillion terms. Many of these terms need not be calculated, however, and I wonder if there is some good approximation that can be computed more easily than running Kintigh's simulation program. However, I offer this only as a suggestion and a challenge to the statistically inclined.

A second point is that I do not think the jagged "stairstep" shapes of Kintigh's graphs are useful or necessary. They are due only in part to the probabilistic aspects of his simulations. The main reason is that, with a small number of types, only a limited number of discrete sample outcomes are possible. For example, if there are 20 types, then in any real data set, the number of types present can only be 20, 19, 18, and so on. Suppose that, in a given case and for a given sample size, Kintigh's technique finds all 20 types present in about 0.4% of the simulations, exactly 19 types present in about 4.0%, 18 types present in about 14.0%, and so on. Suppose we would like to show a symmetrical 80% confidence interval, that is, an interval such that 10% of the simulation outcomes are at or higher than its upper limit, and another 10% are at or below its lower limit. When we try to obtain such an upper limit in the hypothetical case I have described, however, we find that about 4.4% of the outcomes have 19 or more types present, while about 18.4% of all outcomes have 18 or more types present.

We are, thus, faced with a dilemma. Either we adopt an upper limit for the confidence interval that cannot correspond to any possible real outcome (such as 18.6 types present) or we adopt an upper limit that excludes more or less than the uppermost 10% of outcomes (such as 19 types present or 18 types present). Kintigh makes the latter choice. But this means that the lines representing confidence limits in his displays do not actually correspond to the stated confidence interval. Generally, they show intervals that are narrower by amounts that vary considerably and unpredictably as one moves across the graph.

Kintigh's graphs would be easier to interpret and more useful if he made the first choice, even though that requires some interpolation between whole numbers, and will lead to values that no real sample can possibly show, such as 18.6 types present. Such graphs should not be hard to interpret or misleading. A real
The concept of diversity in archaeological theory

data set with 18 of 20 types present is just within the 80% confidence interval, while a data set with 19 types present is somewhat above it.

The same argument applies to estimates of the most likely number of types expected in samples of a given size from a specific postulated population. A further advantage of plotting non-integer values is that the curves should move smoothly from one sample size to another. Kinks in the curves will flag points where the simulation runs accidentally led to a poor result.

Instead of Kintigh's simulation, Grayson, Thomas, and others use a much simpler technique for relating richness to sample size. They merely use linear regression to estimate the logarithm of the richness as a function of the logarithm of sample size. The results are often quite good, at least in the sense that the correlations are quite high. It is reasonable to ask, then, whether anything is to be gained by using Kintigh's technique? I suggest there are three possible advantages. First, the log-log method does not work well when there are very few data sets, as Simone and Rothchild point out. Secondly, the log-log regression cannot continue to be linear when the expected richness closely approaches the total number of defined categories. Thirdly, Kintigh's method seems to offer a more sensitive way to identify outliers. Extreme outliers can be spotted by eye on the log-log graphs, but no way is provided to identify cases that are not so extreme, yet significantly more or less rich than expected. One might, of course, derive probability ellipses, but unless the data points are from nearly bivariate-normal populations the meaning of these ellipses will be doubtful. Kintigh's method, however, provides direct estimates of confidence intervals.

A further point on which I suspect that there is conceptual difficulty concerns the nature and size of the population that is being sampled. The situation is clear if we think of the problem as that of estimating the richness and diversity that would be found in an entire archaeological deposit, if only a part of the deposit has been excavated, or if we only have information obtained by surface collecting. To be sure, even in this situation there may be very serious questions about the representativeness of the sample. But, if these can be satisfactorily settled, then I see no conceptual problems in using the Kintigh or Grayson techniques to estimate the richness and/or evenness that would have been found if the rest of the deposit had been excavated by similar techniques.

But suppose that one has *all there is* from a number of deposits, and these total collections differ considerably in size. Does that mean that we should directly compare their richness and evenness, without making any allowances for differences in size? If we only want to compare one piece of the archaeological record with another piece of that record, then there is a sense in which we are comparing populations, and perhaps we need not control for size. If, however, we hope to give cultural meaning to our data, it is essential to control for size. Most contributors to this volume understand this, but the reasons are not discussed very fully.

A nearly universal view is that there is a significant degree of structure, of orderliness, in human behavior. Archaeologists (and others) differ sharply about whether this structure is almost wholly given by the historically contingent manners and customs of specific societies in specific times and places, or whether much of it can be formulated as the consequences of more general laws or principles. I will avoid that interesting debate here because, for present purposes, I emphasize that it doesn't matter where the structure comes from or what explains it; it is the fact of its existence that is important. A consequence is that anything that a finite number of people actually do, in a finite amount of time, can be seen as a *sample* of what more bearers of the same culture, or the same people, given more time, would be likely to do. If a site was occupied only briefly or by a few people, and then forever abandoned (for reasons that might well have been external to their own society and culture), we have *only a sample* of what these same people, or others bearing essentially the same culture, would have added to the archaeological record, if they had continued longer doing the same kinds of things in the same place.

For this reason, our attempts to give meaning to the archaeological record will always go astray if we do not control for size effects in comparing data sets. It makes no difference if we have obtained everything that was ever there. The relevant population is not any finite set of actual objects. The relevant population is the infinite set of objects that would be generated by some set of behavioral propensities.

One point that should be abundantly clear from the discussions of richness in this volume is that it is altogether mistaken to suppose that retreating to presence/absence scoring is generally a good way of coping with poor data. Actually, two situations should be sharply distinguished. One is the case in which counts of materials have not been reported or have been arrived at in ways that make the interpretation of the counts highly problematic, but the total mass of material is large enough to warrant the belief that any category not reported is very probably absent in the sampled population, or at any rate extremely uncommon. In this case, presence/absence tabulations are often reasonable. But if the sample size is so small that many categories that are only moderately uncommon in the population may easily be absent in the sample, then reliance on presence/absence tabulations is no help. If anything, it magnifies the distortions caused by smallish sample sizes and by differences in the sizes of different samples.

The papers in this volume concentrate on the biases small samples cause for estimates of diversity measures such as richness and evenness. Small samples cause other systematic distortions as well. For example, even if all other sources of error are eliminated, counts of numbers of objects in each category in small samples are subject to random sampling errors that are large relative to the counts themselves. This means that correlations will have low reliability, whether the correlations are based directly on counts or on percentages (which are ratios between counts). It might be thought that the effect of a reduction in the reliability of correlations due to small sample size is simply to increase the variance of sample correlations, with values higher than the true population value being just as likely as lower values. In fact, Spearman (1904) showed long ago that unreliability of correlation coefficients systematically biases them toward zero. That is, positive population
correlations are usually represented by lower positive values in samples, while negative population correlations are usually represented by more weakly negative values in samples. Spearman labelled this effect 'attenuation' and gave a formula for estimating its effect, if one also has estimates of the reliability of each variable. If \( r \) is the correlation between variables \( x \) and \( y \) computed from the sample data, \( x_0 \) is the reliability of \( x \), and \( y_0 \) is the reliability of \( y \), then \( \hat{r} = r / \sqrt{x_0 y_0} \) where \( \hat{r} \) is the estimated correlation between \( x \) and \( y \) in the population. Cowgill (1970:165, 1986) and Nance (1987) discuss this further.

For correlations based on counts or on percentages derived from counts, an upper limit to the reliability can be computed from the numbers of objects in the relevant categories. The implications of the attenuation effect are far more serious than has been recognized for either bivariate or multivariate statistics based on small samples. In this connection, it is unfortunate that one contributor to this volume (Rice) suggests that 'adequate' sample sizes may be somewhere in the range of 30 to 100. Recently (Cowgill 1986) suggested that we are often in trouble with samples less than about 300. This vague suggestion deserves elaboration.

What really matters is not the total sizes of the samples, but the numbers of objects in the least abundant categories about which one would like to be able to make good inferences. To take an extreme example, suppose two samples consist of 1000 sherds each. One sample has 996 sherds of plainware, three sherds of decorated type A, and one sherd of decorated type B. The second sample also has 996 plainware sherds, one of type A, and three of type B. The samples suggest, of course, that the first comes from a population with three times as much type A as type B, while the second comes from a population with the reverse proportions. However, in assessing the confidence we should place in these suggestions, the 1992 plainware sherds don't count; they simply provide extremely strong evidence that both decorated types are quite uncommon in both populations. With respect to inferences about type A relative to type B, for practical purposes we have two samples of size four each, and these eight sherds warrant only very tentative suggestions about the true ratios of type A to type B in the two populations. A \( 2 \times 2 \) contingency table and a Fisher's exact test of the hypothesis that both samples are from populations with identical proportions of types A and B yields a two-tailed significance level of 0.49. Thus, the observed differences could very easily be accidental differences in samples from populations with identical proportions of types A and B.

I have presented an extreme example to illustrate the point. I fear that with less extreme data sets, the problem is very often not recognized. One approach to the question 'how much is enough?' from the viewpoint of correlation studies, is to get some estimate of the average proportion of a category in the entire set of populations being sampled (where each collection represents a different population). One might reasonably do this, as Kintigh does, by pooling data from all the samples in one's study. A criterion for a minimal sample size, then, could be that it is large enough so that it offers the possibility of reliably detecting evidence that the sample is from a population with a substantially below-average proportion of the given category.

To illustrate, suppose that pooling the data from all collections yields a grand total of 10,000 specimens, of which 100 belong to category \( x \). A reasonable null hypothesis is that category \( x \) constitutes 1% of each of the different populations represented by each sample. Notice that this does not lead us into circular reasoning. We expect that, in fact, category \( x \) may constitute considerably more than 1% of some populations, and less than 1% of other populations. However, it is by comparing the number of specimens actually found in a particular sample with the number expected if the sample were from a population with 1% of category \( x \) that we can identify the samples that really give evidence that the actual population proportion is not close to 1%.

Suppose, then, we have a sample with a total of 1,000 objects. If it is from a population with 1% of category \( x \), the sample should include about ten examples of category \( x \). Assuming that the sampling procedure reasonably approximates simple random sampling, we can compute (using the normal approximation to the binomial expansion) that the probability that the sample will have four or fewer examples of category \( x \) is 0.04. If, for example, the sample actually has only three examples of category \( x \), then it provides strong evidence that the population proportion of category \( x \) is not 1%, but is instead considerably less.

Now suppose that another sample has a total size of 100. We expect that it will include one item of category \( x \). But, using the binomial expansion, we can compute that 36.6% of such samples from populations with 1% of category \( x \) will have no examples at all of category \( x \). Hence, it is not possible, with only 100 objects, to obtain strong evidence that the true proportion of category \( x \) is less than 1%. Samples no larger than 100 are necessarily ambiguous with respect to this question. A further calculation shows that if we want to reduce to 5% the risk of incorrectly concluding that a category is absent from a population when in fact the category constitutes 1% of the population, we will need an approximately simple random sample of 298 or more objects. To reduce the risk to 1%, we will need at least 458 objects.

In some situations, it would be quite possible to get larger archaeological collections, once the need for larger numbers is understood. In many other cases, of course, this is just not possible. In these cases, one must apply corrections, such as those discussed in this volume for richness and evenness and the attenuation correction discussed above for correlations.

It seems worth calling attention to a relationship between the Shannon \( H \) statistic and the 'likelihood ratio' version of chi-square. This version, often called \( G^2 \), is defined as \( 2 \sum f_i \times \ln (f_i/j_i) \), where the \( f_i \) are observed frequencies, the \( j_i \) are expected frequencies, and 'ln' stands for natural logarithms (to base \( e \)) (Blalock 1979:281). After some algebraic struggle, I have established that, if we have two samples, where \( n_i \) is the size of one sample, \( n_j \) is the size of the other sample, \( N \) is the combined size of the two samples, \( H_1 \) pertains to sample 1 and \( H_2 \) to sample 2, and \( H \) is computed from the pooled samples, then, if the \( H_i \) are computed using natural logarithms, \( G^2 = 2(NH - n_i H_i - n_j H_j) \). Since \( G^2/N \) is very close to chi-square/\( N \), which equals the strength of association measure phi-square, phi-square is about \( H - n_i H_i \).
The concept of diversity in archaeological theory

One consequence of this is that if \( H_1 = H_2 \), they must both equal \( H \), and \( G^2 \) must be zero. If phi-square is nearly zero, then \( H_1 \) must be close to \( H_2 \). In addition, I conjecture that if \( n_i \) is not very different from \( n_j \), but \( H_i \) is quite different from \( H_j \), phi-square is also necessarily large. This is another challenge that I set before the mathematically inclined. Further development of these relations may add useful insight into the \( H \) measure of diversity.

Among other things, since the approximate sampling distribution of \( G^2 \) is known, it may be possible to get a good approximation to the sampling distribution of \( H \), which could be used in place of Kintigh's simulation technique. Also, when there are three or more collections, some useful generalization of my analysis may be possible. At least, one can compare any one collection with the pooled numbers for all the other collections.

Finally, I offer some suggestions about further concepts related to diversity. I think these will be especially useful for discussions of ceramic production and style, but they will probably also be useful for considering diversity in other phenomena. Figure 14.1 sketches these ideas. First, of course, is richness, the number of categories present. Second is evenness, which expresses the extent to which the categories are represented by similar quantities of objects. A third concept is range, by which I mean the amount of difference between the most different categories. Range can easily be measured separately for each interval variable used to describe a data set. For nominal variables, an indicator of range is the number of distinct categories represented for each variable. There is an obvious and quite simple sense in which an assemblage that exhibits wider ranges in descriptive variables is more diverse than an assemblage that exhibits narrower ranges. For example, are all pots in a collection more or less the same size, or do they range from tiny to huge? Preoccupation with richness and evenness can blunt one's sensitivity to such a simple question. A fourth concept is standardization. This has been used too loosely, to mean several different things, including relatively low richness. I suggest we distinguish between richness and standardization, and use the latter term to refer to low variation within categories. Fifthly, since some categories in a data set may show high standardization while other categories show low standardization, it seems worthwhile defining uniformity of standardization as the extent to which some categories are more standardized than others. At Teotihuacan, for example, Paula Krotzer points out (personal communication) that some wares appear to be far more standardized than others. This adds to other evidence that suggests there were several different ceramic production subsystems simultaneously in operation.

Equipped with these terms, one can suggest several possible reasons for relatively low richness in the ceramics produced by a workshop. Among these are that the consumers of the pots used for a limited number of activities; that the workshop did not produce all the kinds of pottery locally used; and that some kinds of pots served multiple purposes, so that a few kinds of pots were used for a large number of different activities. In all these cases, the common theme is that the producers had a limited number of distinct kinds of pots in mind. I will not rehearse lengthy arguments on this matter, but will simply point out that one can regard ancient ideas as a subject for research without falling into a 'normative' model of prehistory. Also, although we cannot be sure that we have correctly captured essential aspects of ancient ideas (namely, how many different kinds of things people thought there were), we cannot be scientifically sure of anything else either, and there are scientifically valid procedures for making and improving inferences about the anciently recognized categories.

Low standardization, in contrast, as I urge we define it, means that there is considerable variation between different examples of a given category. There are at least three plausible reasons for low standardization. One is that there is simply little value placed on standardization by the culture. A second is relatively low skill, and/or conditions not conducive to uniformity of products or raw materials. A third is relatively high skill and control over techniques, which can be taken advantage of to vary monotonically. Greater skill and greater control over materials and techniques doubtless appear when producers spend a higher proportion of their time making ceramics (or lithics, or whatever), but low standardization, per se, may reflect low skill, high skill, or have little to do with skill. Even distinguishing richness and range from standardization is not enough. Studies of artifact production must take explicit account of more than this, such as sensitive indicators of skill. Unfortunately, many attempts to deal with craft production and specialization, such as those of Rice (1981, this volume) have been seriously flawed by vagueness about some of the distinctions I have discussed.

**Diversity in ancient phenomena**

The papers in this volume bear on diversity in at least four aspects of ancient life: diet, raw materials, style, and activities. These are such different phenomena that, when we turn from simply measuring diversity in the archaeological record to giving meaning to that diversity, quite different concepts seem to be needed. My discussion here will focus on diversity in activities, and I will make only a few observations about diversity in diet or style. Concerning diet, I wonder about the relevance of biological taxa. It seems to me that more relevant categories would be those reflecting differences in microenvironments exploited, in techniques used, and in nutrition. Among other things, does low diversity in food remains reflect low diversity in what was available, or concentration on a limited part of what was available? As to style, I suggest that attention to range, standardization, and uniformity of standardization may usefully supplement measures of richness and evenness. But I am especially struck by the absence, in this volume, of any reference to the concepts current in art history and related fields. Surely they have at least as much to offer as ecology or information theory.

Concerning ancient activities, it is natural to assume that there is a reasonably simple relationship between diversity in the archaeological record and diversity in the activities that generated that record. This would probably be so, if for each activity that we wished to distinguish, we could identify at least one 'diagnostic' category that (a) was practically never used for any other activity, and (b) as a consequence of this activity, entered the archaeological record in such numbers that absence of the category in a de-
posit would assure us that the activity in question had contributed practically nothing to the deposit in question.

Often, we do not seem to have any categories of objects that meet these criteria. However, to judge from everything in the literature and my own struggles to derive formal techniques for dealing with less than ideal situations, I conclude that we should do our utmost to identify good "diagnostics." Note that by a diagnostic category I do not mean simply a formal or stylistic category: I mean any possible combination of form, style, material, use-wear, and spatial pattern of discard that can be shown to be
strongly associated with a specific activity. Studies in ethnoarchaeology, experimental archaeology, and taphonomy must be pursued to the fullest in order to expand the number and quality of recognized 'diagnostics.'

Consider the situation when no single category is diagnostic of any single activity. That is, given some set of ancient activities that we would like to be able to identify in the archaeological record, and given some set of recognized categories of objects, there is no one category whose presence or absence in a deposit provides strong evidence that a specific activity did or did not contribute to that deposit. In general, two or more categories may have been used for each activity, and each category may have been used to some extent in two or more activities. Note, incidentally, that low richness in the archaeological record can be due to low richness of ancient activities, or it can reflect multiple uses for many categories, together with a possibly high richness of ancient activities.

If there is not a one-to-one relationship between activities and categories, it is easy to show that evenness of activities need not be the same as evenness of categories in deposits generated by these activities. That is, suppose that activity $A$ uses categories $X$ and $Y$, and, on the average, generates twice as many fragments of $X$ as of $Y$. Suppose activity $B$ uses the same two categories but on the average generates four times as many fragments of $Y$ as of $X$. Thus, the contribution of activity $A$ to a deposit will, on average, consist of $1/5X$ and $4/5Y$. If half the material in a given deposit was generated by activity $A$ and half by activity $B$, the expected assemblage proportions will be $(1/2)(2/3)X + (1/2)(1/5)X$, or $(13/30)X$, and $(1/2)(1/3)Y + (1/2)(4/5)Y$, or $(17/30)Y$. The assemblage proportions are not equal, and hence do not show the highest possible evenness. On the other hand, if $9/14$ of the material in the deposit was generated by activity $A$ and only $5/14$ by activity $B$, then the expected assemblage proportions are $(9/14)(2/3)X + (5/14)(1/5)X$, or $(13/14)(1/3)Y + (5/14)(4/5)Y$, or $(17/30)Y$. Thus, assuming that both activities add absolute amounts of material to the archaeological record at about the same rate, a less even mix of the two activities can generate a less even mix of the two activities can generate the highest possible evenness. On the other hand, if $9/14$ of the material in the deposit was generated by activity $A$ and only $5/14$ by activity $B$, then the expected assemblage proportions are $(9/14)(2/3)X + (5/14)(1/5)X$, or $(13/14)(1/3)Y + (5/14)(4/5)Y$, or $(17/30)Y$. Thus, assuming that both activities add absolute amounts of material to the archaeological record at about the same rate, a less even mix of the two activities can generate more even proportions in the archaeological record.

The above numerical example is, of course, merely illustrative, and I am not suggesting that any ancient people made such calculations, or were necessarily aware of the point I have made.

To deal more fully with the connections between ancient activities and the archaeological record, let us begin with a model of activities and deduce consequences for the archaeological record, rather than beginning with archaeological data and trying to imagine what might have caused them. Consider the following general model. Suppose that there were $L$ different activities, $A_1$, $A_2$, ..., $A_L$. Suppose we have defined $M$ types, and $i$ is the index over types. Suppose also we have $N$ archaeological deposits, and $j$ is the index over deposits. From a formal viewpoint, it does not matter whether the deposits are whole sites, whole layers, or spatial locali within layers. Assume that each activity has a set of proportions to contribute different proportions of each category to the archaeological record. Let the average proportion of activity $k$'s total contribution to the archaeological record that consists of category $j$ be symbolized by $p_{jk}$. The value of $p_{jk}$ will be the resultant of several factors, including the extent to which category $j$ was used in performing activity $k$, the wear and tear on this category typically produced by this activity, people's preferences for discarding or continuing to use damaged examples of this category, different ways of dealing with discarded or lost examples of this category, and so on. All these factors will vary from situation to situation. Moreover, at best, the $p_{jk}$'s represent statistical tendencies; expected proportions rather than actual proportions, so actual proportions in actual cases will show variance about the average $p_{jk}$s.

In order to take this model seriously, we have to assume that, for each pair of activities, $A_1$ and $A_2$, there is at least one category, $j$, for which the difference between $p_{1j}$ and $p_{2j}$ is large relative to the variances of $p_{1j}$ and $p_{2j}$. In view of what is now known of the many factors that can affect the quantities and proportions of items that enter the archaeological record, this may seem like a large assumption. However, it seems to me that to deny the reasonableness of this assumption is tantamount to rejecting the possibility of ever identifying ancient activities by any possible archaeological technique.

Incidentally, notice that if, in this model, we split what is, for practical purposes, one category into two categories, $j$ and $r$, we will find that $p_{1j}$ is practically the same as $p_{1r}$ for all $k$. Of course, it would still be a good idea to avoid excessive splitting, if for no other reason than that splitting will reduce category counts and thus increase effects of sampling errors.

Also notice that, if some category $j$ is 'diagnostic' of activity $k$, we can express this formally by saying that $p_{jk}$ equals zero, with negligible variance, for all $p_{jk}$ for which $i$ is not equal to $k$.

Since the $p_{jk}$ are proportions, they add to unity for any activity. That is, $\sum_{j} p_{jk} = 1$, for each value of $k$. Consider, now, that each deposit may be the sum of contributions from each activity, where $A_k$ is the proportion of the material in deposit $i$ contributed by activity $k$. Note that $\sum_{j} A_k = 1$. Now, the expected proportion of category $j$ in deposit $i$ is $e_{ij} = \Sigma_{k} A_k p_{jk}$; that is, the sum of the products of the proportion of each activity times the average proportion of that activity's contribution to the archaeological record that consists of category $j$. In fact, because of variance, the observed proportion of category $j$ in deposit $i$, $o_{ij}$, will not generally equal $e_{ij}$.

In a real situation, we observe the $o_{ij}$, and our problem is to infer the $A_k$ and the $p_{jk}$. There are NL activities, ML categories, and NM observed proportions of categories in deposits. However, since the proportions of activities, categories, and observed proportions each must add to unity, the actual number of unknowns will be $N(L - 1) + L(M - 1)$, and the number of independent known quantities will be $N(M - 1)$. If $N(M - 1)$ is greater than $N(L - 1) + L(M - 1)$; that is, if $N$ is greater than $L(M - 1)/(L - 1)$, there will be more known than unknowns, and at first it seems that a unique solution is possible. For example, we could, by analogy with multiple regression models, solve for the values that minimize the mean squared deviations between observed and expected frequencies. That is, we could minimize $\Sigma_{i=1}^{N} (o_{ij} - e_{ij})^2$ for each $j$, which is the same
as \( \sum \hat{q}_k (q_k - \hat{q}_k) \). Note, however, that we have to estimate the \( p_k \) as well as the \( A_k \). An appealing approach would be to assume that only one activity accounted for all the observed deposits, find the best solution, then repeat the process for two postulated activities, three activities, and so on, until the postulation of a further activity failed to yield any statistically significant further reduction in mean squared error.

If unique solutions could be found for the above set of equations, the method would constitute a truly enormous breakthrough in the linked problems of inferring ancient activities, ancient activity sets (or 'tool kits'), and, to a certain extent, the spatial organization of ancient activities. The principal difficulty about spatial organization, as I see it, is that we cannot assume that all the debris generated by the activity was simply dropped in place. We must at least assume that often some fraction was tossed out of the way (Binford 1983a: 153) or collected and dumped elsewhere. In many cases, it may be adequate to distinguish between the average composition of 'dropped' fractions left in place and the different average composition of 'dumped' fractions accumulated elsewhere. If the compositions of these fractions themselves remain reasonably stable, we can perfectly well think of the dropped fraction of the materials generated by a given activity as one activity set, while the dumped fraction of the materials generated by the same activity can be regarded, formally, as a second activity set. As long as our archaeological techniques permit us to judge whether the contents of a given deposit (or excavation unit within a deposit) were generated by dropping or dumping, the fact that a fraction of the material resulting from an activity is dropped and the remainder is dumped should cause no insurmountable problems. The greatest worry is that, it seems, the fractions dumped or dropped may be highly dependent on factors such as whether the activity was carried out indoors or outdoors, plans to conduct other activities in the vicinity, and the intended duration of occupation of a site. If this is the case, the formal implication is that the variances about average values may be excessively large, thus frustrating the attempted analysis.

Assuming the variances are not too large, however, notice that there is no need to begin by identifying deposits or segments of deposits whose contents were generated by a single activity. We could simply bypass the vexing problems discussed by Whallon (1984), by Carr and others (1985), and in this volume by Simek. We could take provenience units and their contents as given, solve the above equations for the \( A_k \) and the \( p_k \), and, for each activity, plot spatially the proportion it contributed to the contents of each provenience unit. To be sure, the smaller the provenience units are, the more detailed and meaningful will be the pictures given by such spatial plots. But problems of overlapping activity drop areas, and problems of dumps that pool a fraction of the materials from different activities, would no longer be nearly so difficult.

Unfortunately, after this tantalizing glimpse of what the approach seems almost to offer, I have to report that investigation of imaginary 'perfect' data in which observed proportions of categories in deposits are identical to expected proportions shows that, if two or more activities are postulated, there is usually no unique solution for the \( A_k \) and the \( p_k \). It does seem that one can often show that the \( A_k \) and the \( p_k \) have to lie within certain limits. Table 14.1 shows the expected proportions of four artifact categories for eight deposits generated by two activities. It is assumed that the fraction of a deposit generated by activity 1 consists of 20% of category 1, 30% of category 2, 40% of category 3, and 10% of category 4, while the fraction of a deposit generated by activity 2 consists of 30% of category 1, 20% of category 2, 30% of category 3, and 45% of category 4. It is further assumed that the fraction contributed by activity 1 is none of deposit 1, 10% of deposit 2, 20% of deposit 3, 30% of deposit 4, 40% of deposit 4, 50% of deposit 6, 60% of deposit 7, and 100% of deposit 8. The remainder of each deposit is contributed by activity 2. Finally, it is assumed that the observed values are identical to the computed expected values.

The problem, then, is to take the observed values and see if the values of the \( A_k \) and the \( p_k \) actually used can be recovered. In fact, one can show that there are some proportional relationships that necessarily hold among the various \( p_k \), but these are not strong enough uniquely to determine their values. For example a set of 'wrong' values of \( p_k \) and the \( A_k \) is shown in parentheses in Table 14.1. These wrong values generate the observed values perfectly. They are among the most extreme wrong values that will do this. They include the largest workable values for \( p_{11} \) and \( p_{12} \) and the smallest workable values for \( p_{21} \) and \( p_{22} \). One could find the range of all workable values by finding the largest values of \( p_{11} \) and \( p_{12} \) that will generate the correct observed values.

My feeling about these results is disappointment that unique solutions are not generally possible, tempered with gratification that the method seems to do well enough to be of some use. The range of possible values of \( p_k \) and the \( A_k \) is, after all, fairly narrow. Notice that the assumed true values imply that, except for category 4, the two activities generate rather similar proportions of the various categories. An example in which the two activities contrasted more sharply in their contributions to deposits would probably yield better results.

Consideration of Table 14.1 suggests two important further points. First, whenever the proportion of any one of the four categories in a deposit is specified, the proportions of the other three categories can be predicted. This would not be the case if three or more activities were contributing to the contents of the deposits. Thus, the structure of the 'observed' data tells us that two, and only two, activities contributed to these deposits. Secondly, the 'wrong' values allow for the possibility of a deposit in which category 1 is wholly absent (if the deposit represents only activity 2) and for the possibility of a deposit in which category 1 constitutes 24% of the contents (if the deposit represents only activity 1). The true values, however, imply that the observed proportional contribution of category 1 will not be less than 5% in any deposit nor more than 20%. Similarly for the other categories, the true values predict limits on observable proportions that are different from the limits predicted by various wrong guesses. If one has data on a large number of deposits, it can be assumed that it is unlikely that the possible ranges...
The concept of diversity in archaeological theory

Table 14.1. Eight hypothetical deposits, derived from different proportional contributions of four artifact categories by two activity sets. Numbers in parentheses are one set of incorrect values that are consistent with the observed data values.

<table>
<thead>
<tr>
<th>Activity set 1:</th>
<th>Activity set 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_{31} = 0.20 (0.24) )</td>
<td>( p_{31} = 0.40 (0.43) )</td>
</tr>
<tr>
<td>( p_{21} = 0.30 (0.33) )</td>
<td>( p_{42} = 0.10 (0.00) )</td>
</tr>
<tr>
<td>( p_{22} = 0.05 (0.00) )</td>
<td>( p_{32} = 0.30 (0.27) )</td>
</tr>
<tr>
<td>( p_{42} = 0.20 (0.17) )</td>
<td></td>
</tr>
</tbody>
</table>

Proportions of each activity in each deposit:

<table>
<thead>
<tr>
<th>Activity set I:</th>
<th>Activity set 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_{11} = 0.00 (0.21) )</td>
<td>( A_{11} = 0.05 (0.00) )</td>
</tr>
<tr>
<td>( A_{12} = 1.00 (0.79) )</td>
<td>( A_{12} = 0.20 (0.17) )</td>
</tr>
<tr>
<td>( A_{13} = 0.20 (0.33) )</td>
<td>( A_{13} = 0.30 (0.27) )</td>
</tr>
<tr>
<td>( A_{14} = 0.30 (0.39) )</td>
<td>( A_{14} = 0.45 (0.57) )</td>
</tr>
<tr>
<td>( A_{21} = 0.40 (0.45) )</td>
<td>( A_{21} = 0.05 (0.00) )</td>
</tr>
<tr>
<td>( A_{22} = 0.50 (0.52) )</td>
<td>( A_{22} = 0.30 (0.27) )</td>
</tr>
<tr>
<td>( A_{23} = 0.60 (0.58) )</td>
<td>( A_{23} = 0.40 (0.42) )</td>
</tr>
<tr>
<td>( A_{24} = 1.00 (0.82) )</td>
<td>( A_{24} = 0.60 (0.48) )</td>
</tr>
</tbody>
</table>

Observed data (identical to expected values in this example):

<table>
<thead>
<tr>
<th>Activity set I:</th>
<th>Activity set 2:</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_{11} = 0.05 )</td>
<td>( A_{11} = 0.21 )</td>
</tr>
<tr>
<td>( A_{12} = 1.00 )</td>
<td>( A_{12} = 0.79 )</td>
</tr>
<tr>
<td>( A_{13} = 0.21 )</td>
<td>( A_{13} = 0.73 )</td>
</tr>
<tr>
<td>( A_{14} = 0.30 )</td>
<td>( A_{14} = 0.67 )</td>
</tr>
<tr>
<td>( A_{21} = 0.40 )</td>
<td>( A_{21} = 0.45 )</td>
</tr>
<tr>
<td>( A_{22} = 0.50 )</td>
<td>( A_{22} = 0.55 )</td>
</tr>
<tr>
<td>( A_{23} = 0.60 )</td>
<td>( A_{23} = 0.73 )</td>
</tr>
<tr>
<td>( A_{24} = 1.00 )</td>
<td>( A_{24} = 0.67 )</td>
</tr>
<tr>
<td>( A_{31} = 0.10 )</td>
<td>( A_{31} = 0.27 )</td>
</tr>
<tr>
<td>( A_{32} = 0.30 )</td>
<td>( A_{32} = 0.30 )</td>
</tr>
<tr>
<td>( A_{33} = 0.50 )</td>
<td>( A_{33} = 0.50 )</td>
</tr>
<tr>
<td>( A_{34} = 1.00 )</td>
<td>( A_{34} = 0.50 )</td>
</tr>
</tbody>
</table>

In observed proportions are much greater than the observed ranges. This can be used to narrow further the range of probably correct guesses about the true values of the \( p_{a} \) s and \( A_{a} \) s.

It remains to be seen how this approach will fare when it is applied to real data, where (a) there will be more or less large differences between observed and expected proportions, and (b) there may well be three or more activities contributing to the contents of deposits. At any rate, I think it is finally possible to say that we have a model that is concordant (in the sense of Carr 1985) with the situation. I believe it offers a valid and appropriate way for thinking of the connections between the archaeological record and ancient human activities. I am, however, far from satisfied with my current mathematical analysis of the model. I hope very much that others will be challenged to carry this work further.

In this connection, it is useful to look at the results of applying some alternative approaches to the data in Table 14.1. For example, R-mode correlations between proportions of the four categories in the eight deposits show (Table 14.2) that all pairs of the first three categories have correlations of plus 1 with one another, while category 4 has correlations of minus 1 with all the other categories. All correlations are plus or minus 1, of course, because in this example all observed values are identical to expected values. A straightforward principal components analysis would yield one component, which would account for 100% of the variance. The variables representing proportions of categories 1, 2, and 3 would have loadings of plus 1 on the one component, and the variable for category 4 would have a loading of minus 1. Up to a point, such an analysis seems appropriate and useful. The data in Table 14.1 have been constructed in such a way that all the variation in observed category frequencies can be attributed to just one thing: the proportion of the contents of each deposit that was generated by activity 1. Since there are only two activities in this example, the remainder of the contents of each deposit is, necessarily, due to activity 2. The fact that all the observed variation can be attributed to just one factor is just what a principal components analysis would tell us. Indeed, it seems highly likely that principal components analysis would be an appropriate and effective way to determine how many distinct activity sets are implied by other data sets. I hope to test this soon with invented examples based on three or more distinct activities.

In other respects, however, customary interpretations of a principal components analysis of the data in Table 14.1 are more
problematic. Characterizing the deposits by their scores on component 1 may or may not prove to be appropriate. I am especially concerned, however, that one might conclude that categories 1, 2, and 3 constituted one activity set (or "tool kit"), while category 4 constituted a second activity set. In fact, of course, this is very different from the activity sets used to generate the data of Table 14.1. The principal components analysis is quite correct in the implications of loadings; the more activity 1 contributes to a deposit, the higher the proportions of categories 1, 2, and 3 there will be in the deposit, and the lower the proportion of category 4. The difficulty is not that the principal components analysis will give a wrong answer, but that we may be mistaken about what the question was. My earlier analysis showed one way to get an approximation to the actual constitution of the activity sets responsible for the data. It is not clear at present whether principal components analysis could also aid in this task. It is certain, however, that it is wrong simply to interpret a set of categories whose integrations would yield one large cluster consisting of deposits n y other deposit. This is correct, but it docs not seem to offer a very effective lead for further analysis of the observed data.

Table 14.3 gives the Brainerd-Robinson similarity coefficients between the eight deposits of Table 14.1. An attempt to cluster deposits 1 through 7 that could not be decomposed into sub-clusters, and a second cluster consisting of the somewhat distinctive deposits 8. These results are valid, but they reflect no more than that deposits I through 7 represent evenly varying proportional contributions from activity 1, while deposit 8 stands somewhat apart because it represents a much higher contribution from activity 1 than does any other deposit. This is correct, but it does not seem to offer a very effective lead for further analysis of the observed data.

Comments on some papers
All of the chapters in this volume are interesting and useful. I will not attempt to summarize them individually, or to adjudicate between all the differences that crop up. There are, however, a few problems that call for discussion.
As will be apparent from the preceding section, I cannot agree with Simek's insistence that the attempt to infer activity sets from archaeological data is not warranted unless it can be shown that there is considerably less diversity within deposits (or segments of deposits) than between them. I am also baffled by his discussion of global versus local patterning and his distinction between "homogeneous" patterns and "heterogeneous" structures. I agree with Whallon (1984) and Carr (1985) that archaeological records are built up by local, rather than global (i.e., site-wide) processes. But what I understand by this is that different activities will be concentrated in different places and that the dropped and dumped materials resulting from these activities will be concentrated in different places, and that the concentrations themselves will be quite variable in size and shape and density and will often overlap. All this is consistent with the model I discussed in the preceding section of this chapter. Simek, however, seems to argue that localized processes are inconsistent with high site-wide correlations, and that it is good news when correlations between categories, across all the deposits (or segments of deposits) of a site are low. This is simply not so.
At the same time, Simek has a good point in arguing that it is useful to identify segments of a site that are distinctive or anomalous in their contents. However, before employing measures of richness and evenness to identify anomalies, it would be better to begin with simpler techniques such as contingency table analysis. Such an investigation of Simek's nine spatial clusters for Couche V of Le Flageolet I leads to an overall chi-square of 72.3, which, for 40 degrees of freedom, is significant at better than the 1% level. Thus, we are justified in thinking that not all the differences in composition among the nine clusters are accidental. However, the most striking difference is that cluster 5 has 12 bladelets, rather than the 3.5 that would be expected if the distribution of bladelets across clusters were random. It is also notable that cluster 3 has no bladelets at all, while 5.1 would be expected. Cluster 7 has 9 end-scrapers, whereas only 3.7 would be expected. Cluster 8 has 9 retouched pieces, but only 4.4 would be expected. There are a few other cases where there are somewhat more or fewer examples of a category in a cluster than would be expected, but the ones I have listed are all the striking anomalies. It seems to me that efforts to give meaning to Simek's data would do better to focus on these observations rather than to concentrate on richness and evenness indices. Of course, the approach I sketched earlier might be tried, but the low counts for many categories in many clusters suggest that large sampling errors will cause difficulties.
I wonder, indeed, how much meaning one can hope to extract from the six categories used by Simek. Since his chapter is intended merely to illustrate an approach, this is not a major criticism. Nevertheless, I wonder if it would not be highly useful to distinguish different kinds of burins, perhaps to use finer subdivisions of some of the other categories, and to use evidence of use-wear, if it is available.
Some of the other chapters also call for comments. I am puzzled by the passage in which Rice says that diversities should be compared only between assemblages of approximately equal sizes. A central theme of this book is the possibility of controlling for wide ranges of sample size in making diversity comparisons.
I am also puzzled by Rothschild's figures for her New York City samples. Like children in Lake Wobegon, all three seem to be above average in both richness and diversity.
The concept of diversity in archaeological theory

Thomas, like Jones, Beck, and Grayson, shows different regression slopes for different data subsets. Unlike them, he does not provide information about the significance of these slope differences. The method is excellent in principle, but I seriously doubt whether these particular differences are significant at, say, the 5% level.

I wonder if the lower richness of Basketmaker raw materials found by Leonard, Smiley, and Cameron may be simply because most of the Basketmaker materials came from a relatively short distance, within which one would expect less diversity of sources.

Some of Kintigh’s examples seem to use collections known to combine material from several periods. This is unobjectionable as long as the examples merely illustrate the technique, but when one is more interested in the meaning of the analyses it is important to use collections believed to pertain to single periods.

Jones, Beck, and Grayson ask whether evenness should be computed on the basis of all categories used in an analysis, or only those actually present in a given collection. It seems to me that the former choice is preferable. This will, of course, make the maximum possible evenness depend on the sample size whenever the size of the sample is smaller than the number of categories used. However, there is no escaping the fact that such very tiny samples must be problematic. The techniques discussed in this volume for controlling for variable sample sizes cannot work miracles.

To sum up, I think the greatest contributions of the studies in this book are, first, that they show that comparisons between samples of small and large size will be very misleading if corrections for sample size are not made, and, secondly, they show us some good ways to make those corrections. Proper allowance for different sample sizes is a necessary condition for giving meaning to diversity in the archaeological record, but it is not a sufficient condition. The most interesting chapters in this volume deal with various possible meanings of various kinds of diversity. It is along these lines that future work should proceed.

Notes

1. The horizontal and vertical axes in Figure 14.1 are intended to suggest a multivariate descriptive space. Each dot stands for an object and each cluster of dots stands for a category of objects.
2. My ideas on ceramic production have benefitted from discussions of this topic with Mary R. Hopkins.
3. In the terms used by Binford (1983a:147), what I refer to in this chapter as ‘activities’ should probably be called ‘tasks.’
4. There is, in fact, a substantial literature on this topic under the name of ‘linear unmixing.’


1983a. *In Pursuit of the Past: Decoding the Archaeological Record*. New
References

York. Thames and Hudson.


Davis, W.M. 1981. Reply to P. Rice, Evolution of specialized pottery


References


References


References


