STATISTICAL ANALYSIS OF CERAMIC ASSEMBLAGES

1. INTRODUCTION

This paper describes the results of the first year of a two-year project 'Statistical analysis of ceramic assemblages' at University College London Institute of Archaeology. It is funded by a grant from the Science-based Archaeology Committee of the Science and Engineering Research Council (SERC-SBAC), with C. R. Orton as principal investigator and Dr. P. A. Tyers as research assistant. In the first year we have concentrated on the theory of the subject and on collecting suitable data; in the second year we intend to analyse as many datasets as possible, to gain experience in the application of the theory to a wide range of material.

1.1 *Quantification of ceramics*

By the quantification of a ceramic assemblage we mean the giving to it of a value which expresses the 'amount' of material in it. Such a value we here call a *measure* (of quantity). We suppose that, for a given investigation, the assemblage can be divided into subsets, here called *types*, by an exclusive and exhaustive classification, i.e. every piece of ceramic material belongs to one, and only one, type. This is not a restrictive requirement as it is permissible to have an 'unclassified' or 'unknown' type. Different aims may require different definitions of type, e.g. shape types (forms), fabric types, decorative types, or some combination of two or more of these (see Section 4.3). Quantification consists of assigning a measure to each type in the assemblage, i.e. saying 'how much' of each type there is.

Why should archaeologists want to do this? There seem to be three main reasons — for chronological, spatial or functional/social reasons.

CHRONOLOGICAL REASONS

Seriation is a long-established archaeological technique (see MARQUARDT 1978 for a history) for sorting groups of archaeological material (e.g. grave-groups, ceramic assemblages) into a linear order, which is assumed (or sometimes shown) to correspond to chronology. Sometimes the presence or absence of types in groups is used (e.g. DUNCAN *et al.* 1988), but for ceramic assemblages the proportions of different types in different assemblages form the usual starting point. For example, Millett (1979a) used the proportions of different forms in various assemblages to seriate the pottery from a small Roman town. The use of pottery in urban seriation has been discussed by Carver (1985).

SPATIAL REASONS

Variations in the proportions of a particular type of ceramic at different sites around its production centre can give valuable information about possible means of marketing and transport. In the classic study of this type, Fulford and Hodder (1974) showed differential variation in the fall-off rate (the rate at which the proportion of a type decreases with increasing distance from the production centre) of Oxfordshire ware according to whether the direction was along a river or overland. This suggested that the main means of transport was by water. The competing influence of the New Forest centre was also highlighted by a study of the residuals from a regression analysis.

FUNCTIONAL/SOCIAL REASONS

Functional differences in the composition of assemblages have been used to infer differences in activities or social status, both within and between sites. For example, Millett (1979b) looked for evidence for different activities in different pit groups at the late Roman site of Portchester, and Redman (1979) used ceramic assemblages as evidence for different social areas in the medieval town of Qsar es-Seghir, Morocco.

OTHER REASONS

Quantification can also give an insight into site formation processes (see Section 1.4).

In all cases the need is for a measure of quantity of ceramics, which can be broken down by different groupings, e.g. by chronological, geographical or functional types. The main use for such figures is to compare them with other assemblages; figures relating to one assemblage in isolation are rarely of use.

1.2 *Measures of quantity*

Several ways of measuring quantities of pottery have been used and suggested over the years. Two broad approaches can be discerned, one asking the question 'how many vessels?' and the other the question 'how much pottery?'. Within the latter are two sub-approaches, depending on whether all vessels should in some way be considered equivalent, irrespective of size, or whether (for example) larger or more valuable vessels should be given more weight. Problems have also arisen because some measures can be calculated directly, while others are in the nature of 'unknown' parameters, and have to be estimated. Discussion of the 'best' estimator for a particular parameter can obscure the issue of whether it is the 'best' parameter for the purpose in hand.

HOW MANY VESSELS?

The first approach sees the aim of quantification as saying how many vessels

are present on a site or in an assemblage, i.e. from how many vessels do the sherds that we have come? This is here called the *vessels represented* approach. Some writers (e.g. MOORHOUSE 1986, 86) seem to believe that this question can be answered directly, by bringing together all the sherds that belong to the same vessel. Others (e.g. VINCE 1977, 63) seem to be less sanguine, and to see this parameter as something to be estimated. The answer probably depends on the nature of the site or deposit, the technology of the pottery (e.g. hand-made or wheel-thrown), and the skill and time available to the archaeologist. It seems likely that the cases in which a direct count of vessels can be made are in the minority, and are therefore unlikely to be useful in intra- or inter-site comparisons.

We therefore have to look at ways of estimating the number of vessels represented in an assemblage. One estimate in the *minimum number of vessels* (VINCE, *ibid*), which is analogous to the MNI statistic of animal bone studies (GRAYSON 1984, 27-48). Vessels are reconstructed as far as possible; sherds which do not physically join the reconstructions but which could feasibly belong to the same vessel are assigned to them. Groups of sherds belonging to the same vessel are referred to as *sherd families*. This gives a figure equal to, or less than, the actual number of vessels represented. An alternative approach, *maximum number of individuals*, assigns unmatched sherds to separate vessels, giving rise to an estimate greater than or equal to the number of vessels represented. Because the first approach tends to under-estimate the true number, and the second to over-estimate it, hybrid estimates, e.g. the average of the two, have been suggested. For reasons of speed or practicability, it has at times been suggested that estimates should be based on rims, bases or other distinguishing features, rather than on whole vessels. We here use the term *estimated vessels represented* (*evrep*) to mean any estimate of the number of vessels represented in an assemblage.

AMOUNTS OF POTTERY

If we decide to count all vessels as equivalent in some way, then we need a measure under which a whole vessel will counts as '1' and fragments will be represented as fractions, according to the amount of each vessel present. For example, a vessel of which half is present would have a measure of 0.5. Such a measure is called a *vessel-equivalent* (v.e.). Ideally, an assemblage would be sorted into its vessels represented, and the v.e. of each measured. Even when it is possible to sort to vessels represented, it is not clear how the v.e. should be measured. Schiffer (1987, 283) suggests using the ratio of the weight of the reconstructed fragment to that of a comparable complete vessel. This would be a good measure provided the pottery is so standardised that one can tell what

the weight of a complete vessel would be, and if a high proportion of the sherds in an assemblage could be assigned to a known vessel type.

In general, however, it is not possible to measure the v.e. directly, and it must be estimated in some way, giving rise to the *estimated vessel equivalent* or *eve* (see ORTON 1975 for the term; the idea seems to be due to BLOICE 1971 or EGLOFF 1973). Eves are most commonly calculated from the percentage of a complete rim represented by each rim sherd or family of rim sherds. In the simplest form, it is not necessary to identify rim-families. The eve statistic is likely to be more reproducible than evreps, as it does not rely on the identification of individual pots. Other distinctive features can be used if appropriate, e.g. bases, handles, spouts, or some combination of two or more of them.

Alternatively, one may wish to give more value to larger vessels. The simplest way to do this is to record the weight of sherds of each type present. Thus heavy vessels count more than light ones, in both the complete and the fragmentary state. Similar measures that have been suggested include surface area (GLOVER 1972, 92-6; HULTHÉN 1974) and displacement volume (HINTON 1977, 231).

Confusion has sometimes arisen because the eve gives a measure which is always less than or equal to the evrep; in fact, it is less than or equal to the minimum number of vessels. For this reason it has at times been called 'minimum number of vessels' and used as a 'floor' estimate of vessels represented. It is important to see it as an estimate of the measure of the amount of pottery, not as a rather bad estimate of the number of vessels represented.

Finally, we have what is probably the most widespread measure, the sherd count. It is highly variable, depending on the degree of breakage of pottery in an assemblage. For example, in an assemblage of complete vessels all would have the same value (one), while in an assemblage of fragments, the most breakable vessels would have the highest count and hence the greatest measure.

1.3 *Comparisons of different measures*

In this section we look at attempts that have been made to compare the measures described in section 1.2. We shall concentrate on four measures:

(i) sherd count
(ii) weight (and related measures such as surface area)
(iii) estimated vessel-equivalent (eve)
(iv) vessels represented (evrep).

The choice of a suitable measure depends on the use to which it will be put. The criteria to be used in assessing and comparing measures are:

(i) archaeological — what is the archaeological purpose?

(ii) formal — does the measure satisfy formal criteria, e.g. concerning bias, variance?
(iii) practical — is a measure quick and easy to use, giving consistent results between different workers, and likely to be free from gross errors?

Following reasoning of Section 1.1 we take the main purpose of quantification as the comparison of the proportions of different types (however defined) in different assemblages.

The earliest attempts to compare measures concentrated on archaeological and practical criteria. Solheim (1960) compared sherd count and weight, concluding that both together gave more information than either separately (see Section 1.4). Hinton (1977) compared total sherd count, rim sherd count, weight and volume, concluding that total sherd count was probably the most accurate, but weight was the fastest, while rim sherd count seemed unreliable and the measurement of volume was messy. However, the lack of formal criteria in these assessments is a serious drawback, as the outcomes could only be compared with the writer's expectations.

Formal criteria, in the form of correlation coefficients, were used by Glover (1972, 93-6) and Millett (1979c). Both concluded that apparently high correlation coefficients between different measures indicated that they were more-or-less equivalent, and that a decision could be based on practical criteria. Weight or surface area (measured by a quick but approximate technique, rather than by Hulthén's (1974) laborious adjustment of weight) seemed to be the favoured approach. Once again, the value of using more than one measure was noted.

A formal assessment, based largely on the criterion of bias, was made by Orton (1975). Use of sampling theory showed that measure (iii) (eves) was unbiased, and that measures (i) and (ii) (sherd count and weight) could give unbiased comparison in favourable circumstances. Weight seemed preferable to sherd count because the circumstances were less restrictive. Measure (iv) (evreps) gave biased estimates of both the proportions in an assemblage and of comparisons between assemblages, and was not recommended. Further work (ORTON 1982a), based on simulation, suggested that evreps gave rise to smaller standard deviations than the other measures; there seemed to be little to choose between sherd count, weight and eves in this respect.

1.4 *Statistics based on two or more measures*

It has long been known that quantifying the same assemblage by two different measures can give information that is not apparent from either taken in isolation (e.g. SOLHEIM 1960). For example, dividing weight by sherd count gives the average weight of a sherd of a particular type, which can give indications

of the post-depositional history of an assemblage. Bradley and Fulford (1980) gave an example of the contribution of such statistics to the interpretation of a site. The general use of derived statistics has been discussed by Schiffer (1987, 282-5) in the broader context of the study of site formation processes.

If we take the four measures compared in Section 1.3, there are twelve possible pairings. Two of them have been found to be particularly useful:

(i) sherd count ÷ vessel-equivalent, called *brokenness* (ORTON 1985),
(ii) vessel-equivalent ÷ vessels represented, called *completeness* (ORTON 1985) or the *completeness index* (CI) (SCHIFFER 1987, 282).

BROKENNESS

Brokenness is a function of both ceramic type and post-depositional history. Given the same history, large or fragile vessel types tend to have a greater value of brokenness. i.e. more sherds per eve, than small or robust vessel types. For example, in a study of groups of early Roman pottery (BEDWIN, ORTON 1984) the mean value for coarse wares was over three times that of fine wares (110 and 32 sherds/eve respectively), because the fine wares were present as much smaller vessels than the coarse wares. At the same time, the longer and more complicated the post-depositional history of an assemblage, the more broken the pottery is likely to be, and the greater the brokenness. In the same study, the brokenness in 'reliable' contexts (excluding very small assemblages) varied by a factor of more than two, from 59 to 140 sherds/eve. It may be difficult to disentangle the two effects when comparing assemblages.

COMPLETENESS

Completeness differs from brokenness in that it is in general a function only of post-depositional history, and not of ceramic type (but see Section 2.5). In other words, all types in an assemblage should have the same value of completeness, within limits due to sampling theory (and excluding factors such as differential scavenging of particular sherds). Completeness starts from a value of one (whole pots) and decreases at every subsequent 'event'. At the site mentioned above, lower fills of a ditch gave values of 0.09 to 0.12 eves/evrep, while upper fills gave values of only 0.04 to 0.06, suggesting recutting of the ditch (the composition of the pottery in the lower and upper fills was otherwise indistinguishable).

The value of derived statistics is thus one aspect that must be kept in mind in studying the properties and behaviour of various measures, and in devising procedures for their calculation or estimation.

1.5 *Recording pottery*

Despite various recommendations (e.g. YOUNG 1980), there is no generally-accepted method of cataloguing pottery, and many are in use. We need to develop a theory that will be applicable to as many different systems as possible, while pointing out the advantages of some and the disadvantages of others (see section 4.2).

Catalogues can be made as simple hand-written lists, card indices, entries on pre-printed cataloguing sheets, or records in a computer database. Whatever the system, we need to forge a link between it and statistical theory by establishing what entity, in the system, corresponds to the statistical concept of the 'observation'.

What we need for this purpose is the smallest possible catalogued quantity for which the measure is known. For example, if the measure is weight, and sherds are weighed individually, then this unit is the sherd; if they are weighed in batches, then the unit is the batch. Similarly, if the measure is in rim-eves, and rim sherds are measured individually, then the unit is the rim sherd; if they are measured in rim-families, then the rim-family is the unit.

Since we are interested in estimating proportions of different types in different assemblages, the catalogued unit should not mix pottery of different types or from different assemblages. In other words, the broadest possible cataloguing unit consists of all the pottery of one type in one assemblage. The smallest possible unit is the individual sherd. A catalogue can be seen as broken down into a number of entries, or *records*, each of which contains (at least) the context (assemblage), type and measure of a group of pottery. The totality of pottery of one type in one assemblage may thus be represented by a single record, or several records, depending on the system used.

1.6 *The problem*

A major problem is that none of the measures described in Section 1.2 is, as it stands, suitable for statistical analysis.

(i) sherd counts appear superficially to be suitable for analysis as discrete data. This appearance is however deceptive, as the 'observations' (i.e. individual sherds) are not independent. We can demonstrate this by a simple *reductio ad absurdum*: if we broke all sherds in half we would (if the use of statistical theory is permitted) halve all our variances and correspondingly increase the precision of our estimates. Clearly this is nonsense and shows that the theory cannot be applied.

(ii) sherd weight and eves are very similar in this respect: neither can generate estimates of variances internally. Such estimates must therefore be based

on repeated observations, i.e. we require several assemblages which we can assume with confidence to be samples from the same population. But this begs the very question we are trying to answer, and is therefore inappropriate.

(iii) the 'vessels represented' statistic, like sherd count, could be treated as discrete data, but suffers from such serious biases (ORTON 1975) as to rule it out.

2. THEORY

2.1 *Aims and notation*

There are two main theoretical aims of the project:

(i) to be able to set confidence limits on the proportions of a ceramic assemblage that belong to different types,

(ii) to be able to compare, both numerically and graphically, the compositions of two or more assemblages in terms of the proportion of each type present in each assemblage, and to assess the statistical significance of the differences between them.

The practical aim of the project is to apply the theory to assemblages from a wide range of sites, of different types and periods, to assist in their interpretation, and hence the interpretation of the sites themselves. We expect that the work will also lead to recommendations about the recording of ceramic assemblages.

In the theory, the term 'type' is used in a perfectly general sense, to mean a categorical variable that takes a value on all the pottery from an assemblage (but it is accepted that one of the categories may of necessity be 'unknown'). In practice, we use 'type' to mean either fabric, form, or the combination fabric-by-form. However, the theory does not depend on the particular meaning assigned to the term 'type', and users are free to assign their own meaning to it. For example, decoration may be an important aspect of the definition of type in some circumstances.

The data consist of the values of a chosen measure (see Section 1.2) on each *record* (see Section 1.5) of one or more assemblage. The implication of the theory for the choice of measure are set out in Section 4.1. Unless otherwise stated, the theory that follows is perfectly general. Sections which depend on characteristics of a measure, and which therefore will not be appropriate for all measures, will be indicated.

The number of assemblages making up a dataset is denoted by A, and the number of types by T. The numbers of records of the jth type in an assemblage

are denoted by m_j (j = 1,. . ., T), and the total number of records by m. The measure of the ith record of the assemblage is denoted by w_i, (i = 1,. . ., m). The total measure of a type is denoted by W_j (j = 1,. . ., T), and the overall total by W (note that upper case is used for type and assemblage totals, and lower case for individual values). The proportions of types in an assemblage are denoted by $\mathbf{p} = (p_1,. . ., p_T)$.

The symbol ~j refers to all types *except* the jth, and Σ_j means summation over the jth type. A 'bar' symbol is used to denote a mean (e.g. $\bar{x}$). The unadjusted sum of squares $\Sigma_j w_i^2$ is denoted by S_j^2.

The approach used is to treat each assemblage as a sample from a different population of vessels. Our task then becomes

(i) to make point and interval estimates of the proportions of different types in each population, and

(ii) to test the significance of the differences between the estimates of proportions obtained from the different samples. This can also be seen as testing whether the assemblages could reasonably have come from the same population.

(iii) to display the relationships between the different assemblages and the different types graphically.

We note that for standard statistical theory to be applicable, the 'observations' (in our notation, records) must be independent of each other. This aspect is discussed in more detail in Section 4.2.

2.2 *Estimates of proportions in a single assemblage*

The proportion p_j is estimated by

$\hat{p}_j = W_j/W$, for j = 1,. . ., T.

We define two new variables x(j) and y(j) for *all* records by

$x_i(j) = w_i$

$y_i(j) = w_i$ if the ith record relates to the jth type,

$= 0$ otherwise.

Then $\hat{p}_j = W_j/W = \Sigma y_i(j)/\Sigma x_i(j)$, a *ratio estimate*.

Cochran (1963, 30-1) gives a formula for the variance of a ratio estimate, leading to $\text{var}(\hat{p}_j) = (m/(m - 1)W^4)\{W_{\sim j}^2 S_j^2 + W_j^2 S_{\sim j}^2\}$ — **(1)**

and $\text{cov}(\hat{p}_j, \hat{p}_k) = -(m/(m - 1)W^4)\{WW_j S_k^2 + WW_k S_j^2 - W_j W_k S^2\}$ — **(2)**

SUMMARY

This section enables us, *for the first time*, to estimate the variances and covariances of the proportions of different types in a single assemblage. We can therefore attach confidence intervals to the estimate of the proportion of a type or any combination of types.

2.3 *Comparing proportions in two or more assemblages*

Before we can tackle this problem, we need to develop some preliminary theory.

Given any type j, we can compare var $(\hat{p}_j)$ with the variance of an estimate based on a binomial model, i.e. on an assemblage of complete vessels. In the latter case, the formula is $var'(\hat{p}_j) = \hat{p}_j \hat{q}_j/n$, for a population of size n, where $\hat{q}_j = 1 - \hat{p}_j$.

So the variances would be the same if $var(\hat{p}_j) = \hat{p}_j \hat{q}_j/n$.

We can turn this round and define $n_j = \hat{p}_j \hat{q}_j/var(\hat{p}_j)$,

so that n_j is the number of whole vessels that would give the same value of var $(\hat{p}_j)$ as our sample of m measurable records.

The full formula is $n_j = (W_j/W)(W_{\sim j}/W)((m - 1)/m)\{W_{\sim j}^2 S_j^2 + W_j^2 S_{\sim j}^2\}$

$= ((m - 1)/m)W_j W_{\sim j} W^2/\{W_{\sim j}^2 S_j^2 + W_j^2 S_{\sim j}^2\}$ — **(3)**

It is important to note that n_j is just a number which, when applied to the binomial formula for variance, give the same result as the formula (1) above. It is *not*, for example, an estimate of the number of vessels in the original population.

Suppose that all types have the same mean and variance of w.

Recall that $var(\hat{p}_j) = (m/(m - 1)W^4)(W_{\sim j}^2 S_j^2 + W_j S_{\sim j}^2)$ — (1)

and that $n_j = \hat{p}_j \hat{q}_j / var(\hat{p}_j)$,

and $\hat{p}_j = W_j/W$.

We pool our estimates of the mean and sum of squares of w,

obtaining W/m and S^2 respectively,

and replace W_j by $W(m_j/m)$, S_j^2 by $S^2(m_j/m)$.

So $n_j = ((m-1)/m)(m_j/m)(m_{\sim j}/m)W^4\{W^2(m_{\sim j}/m)^2 S^2 (m_j/m) +$

$W^2 (m_j/m)^2 S^2 (m_{\sim j}/m)\}$

$= ((m - 1)/m) W^2/\{m_{\sim j}/m) S^2 + (m_j / m)S^2\} = ((m - 1)/m)W^2/S^2$ = **(4)**

So that if all the types have the same mean and variance of w, we can by pooling estimates obtain a common value $n = n_j$ for all j.

SEVERAL ASSEMBLAGES

We at last have the background theory we need to look at the comparison of several assemblages, say A of them.

We have vectors of measures $\{W_{r1}, \ldots, W_{rT}\}$,

of numbers of observations $\{m_{r1}, \ldots, m_{rT}\}$,

of sums of squares $\{S^2_{r1}, \ldots, S^2_{rT}\}$,

and estimates of proportions $\{\hat{p}_{r1}, \ldots, \hat{p}_{rT}\}$,

and variance-covariance matrices $\| \text{cov } (\hat{p}_{rj}, \hat{p}_{rk}) \|$,

all for $1 \leq r \leq A$.

We want to compare the vectors of estimated proportions, e.g. to test a

hypothesis H_o: all assemblages are 'the same', i.e. can be thought of as sample from the same parent population.

We assume that each assemblage is homogeneous (see Section 2.5) and so has a single n-value, which we call n_r, for $1 \leq r \leq A$.
We replace each W_{rj} by $W'_{rj} = n_r(W_{rj}/W_r) = (n_r/W_r)W_{rj}$
for $j = 1,\ldots, T$ and $r = 1,\ldots, A$.
so that $W'_r = n_r$ for all assemblages r.
The estimates of proportions are unchanged:
$\hat{p}'_{rj} = W'_{rj}/W'_r = W_{rj}/W_r = \hat{p}_{rj}$,
and so are their variances and covariances.
Recalling (4), that $n_r = ((m - 1)/m)(W_r^2/S_r^2)$,
and writing $m_{rj}/m_r \approx \hat{p}_{rj}$, etc., since the assemblages are homogeneous,
we have $\text{var}(\hat{p}_{rj}) \approx \hat{p}_{rj}\,\hat{q}_{rj}/n_r$
and $\text{cov}(\hat{p}_{rj}, \hat{p}_{rk}) \approx -\hat{p}_{rj}\,\hat{p}_{rk}/n_r$.
But these are exactly the same as the variance and covariances we would obtain from a multinomial distribution with parameter **p** and sample size n.

CONCLUSION

This is a very important result. It means that, as a large-sample approximation, we can treat the transformed data as a series of samples from multinomial distributions. We can therefore treat the data collectively as a contingency table and use any of the theory appropriate to contingency tables (e.g. log-linear models, see Sections 2.4 and 3.2; correspondence analysis, see Section 3.1). For the first time, this approach enables to make proper statistical comparisons between the proportions of different types in different assemblages.

We refer to the transformed values W'_{rj} as *pseudo-counts*. They are not integers, but can be treated for statistical purposes as if they were. We shall develop the notation further when we look at specific measures is Section 4.1.

2.4 *Log-linear and quasi-log-linear models*

Suppose we have a table of pseudo-counts (see Section 2.3) **n**, which may be two-way (e.g. context-by-fabric, context-by-form) or three-way (context-by-fabric-by-form).

NOTATION

To follow standard notation (e.g. FIENBERG, 1977) we replace n by x; the subscripts i, j, and k refer to context, fabric and form respectively. We can construct a set of nested models of increasing complexity and archaeological reality.
Model 1: « complete independence ».
Model 2a: « fabric-by-form interaction only », i.e. fabric and form are « com-

pletely independent » of context.
Model 2b: « context-by-form interaction only », i.e. context and form are « completely independent » of fabric.
Model 2c: « context-by-fabric interaction only », i.e. context and fabric are « completely independent » of form.
Model 3a: « fabric-by-form and form-by-context interactions », i.e. fabric is « conditionally independent » of context, given the form.
Model 3b: « fabric-by-form and fabric-by-context interactions », i.e. form is « conditionally independent » of context, given the fabric.
Model 3c: « form-by-context and fabric-by-context interactions », i.e. fabric is « conditionally independent » of form, given the context.
Model 4: « all pairwise interactions », or « partial association » of form and fabric, fabric and context, and form and context.
Model 5: « all interactions », the saturated model.

There are six routes from Model 1 (complete independence) to Model 5 (saturated) — via Models 2a and 3a, 2a and 3b, 2b and 3a, 2b and 3c, 2c and 3b, and 2c and 3c (see Fig. 1) — corresponding to different archaeological needs.

Within each model, we can estimate the parameters and carry out a goodness-of-fit test. The four levels of unsaturated models to form a « nested hierarchy », and their goodness-of-fit can be measured by the likelihood ratio statistics (BISHOP *et al.* 1975, 125) $G^2(1)$, $G^2(2)$, $G^2(3)$, $G^2(4)$. Both the G^2 and the differences $G^2(1) - G^2(2)$, etc., are distributed as chi-squared statistics. We can partition the overall chi-squared statistic by writing
$G^2(1) = (G^2(1) - G^2(2)) + (G^2(2) - G^2(3)) + (G^2(3) - G^2(4)) + G^2(4)$.

To find the 'best' model for the data, we start at the 'bottom' end (i.e. model 4) and test $G^2(4)$.
If it fits the data (i.e. $G^2(4) <$ critical value), we move up to level 3 (i.e. either model 3a, 3b or 3c, as appropriate) and test *both* $G^2(3)$ and $G^2(3) - G^2(4)$, with appropriate degress of freedom.
If *both* fits are acceptable, we move to level 2;
If *either* fails, we accept model 4.
At model 2, we test *both* $G^2(2)$ and $G^2(2) - G^2(3)$.
If *both* fits are acceptable, we move up to model 1;
if *either* fails, we accept model 3.
At model 1, we test *both* $G^2(1)$ and $(G^2(1) - G^2(2))$.
If *both* fits are acceptable, we accept model 1;
if *either* fails, we accept model 2.

This approach enables us to find the simplest model, out of a chosen hierarchy of models, that fits the data reasonably.

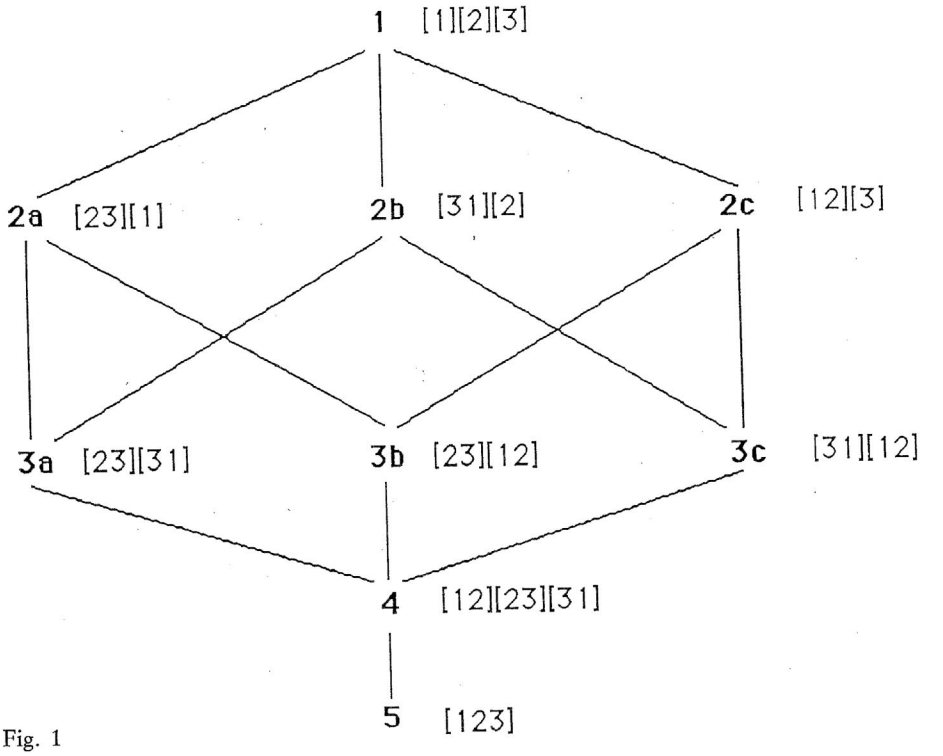

Fig. 1

model	X^2	G^2	df	prob
1	742.9	393.3	204	0.0%
2a	337.2	267.1	140	0.0%
3a	202.0	182.8	54	0.0%
3b	110.8	89.5	110	46.2%
4	21.8	17.5	26	69.8%

goodness-of-fit statistics

model	G^2	df	prob	
4	17.5	26	89.3%	
3a	182.8	54	0.0%	FAILED

accept model 4 (all pairwise interactions)

goodness-of-fit statistics

model	G^2	df	prob	
4	17.5	26	89.3%	
3b	89.5	110	92.5%	
3b*4	71.9	84	82.3%	
2a	267.1	140	0.0%	FAILED

accept model 3b (fabric-by-form and context-by-fabric interactions)

Table 1

model	X^2	G^2	df	prob
1	555.8	261.5	108	0.0%
2b	223.6	188.3	62	0.0%
3a	107.7	106.8	42	0.0%
3c	87.4	81.3	16	66.7%
4	13.6	10.5	16	49.3%

goodness-of-fit statistics

model	G^2	df	prob	
4	10.5	16	84.1%	
3a	106.8	42	0.0%	FAILED

accept model 4 (all pairwise interactions)

goodness-of-fit statistics

model	G^2	df	prob	
4	10.5	16	84.1%	
3c	81.3	16	0.0%	FAILED

accept model 4 (all pairwise interactions)

Table 2

model	X^2	G^2	df	prob
1	466.6	232.9	38	0.0%
2c	90.1	65.7	19	0.0%
3b	7.9	9.4	10	63.7%
3c	50.9	44.7	7	0.0%
4	6.4	7.8	7	49.3%

goodness-of-fit statistics

model	G^2	df	prob	
4	7.8	7	35.1%	
3c	44.7	7	0.0%	FAILED

accept model 4 (all pairwise interactions)

goodness-of-fit statistics

model	G^2	df	prob	
4	7.8	7	35.1%	
3b	9.4	10	49.6%	
3b*4	1.6	3	66.3%	
2c	65.7	19	0.0%	FAILED

accept model 3b (fabric-by-form and context-by-fabric interactions)

Table 3

The simplest statistically is complete independence of fabric, form and context (model 1), i.e. the proportions of the different forms are the same in all fabrics, and the proportions of both are the same in all contexts. This is archaeologically incredible.

The next simplest (model 2a) is that different forms occur in different proportions in different fabrics, but that the proportions of fabrics and forms are the same in all contexts. This may occur if all the context are 'similar', or the assemblages are so small that differences between them are not significant.

In model 2b the proportions of forms may vary from context to context, but the proportions of fabrics may not. This corresponds to a 'functional' interpretation of the differences between the contexts. This interpretation depends on a suitable level of definition of forms (see Section 4.3). By contrast, in model 2c the proportions of fabrics may vary from context to context, but not the proportions of forms. This corresponds to either a 'chronological' or a 'spatial' interpretation of the differences between the contexts, depending on their spatial relationships. This interpretation depends on a suitable definition of fabric, and on the assumption that different sources have limited life-spans within the period being considered. This is true for (e.g.) Roman and medieval pottery in South-east England, but may not hold elsewhere.

Model 3a introduces the possibility that the proportions of fabrics and forms may vary from context to context; but the proportions of fabrics only as a side-effect of variations in forms and the preference of certain forms for certain fabrics.

Model 3b allows the proportions of fabrics to vary, and the proportions of forms to vary as a side-effect.

Model 3c allows the proportions of fabrics to vary between some contexts, and the proportions of forms to vary between others.

Model 4 allows proportions of fabrics and of forms to vary independently of each other from context to context, but also to interact on each other. This would allow for functional variability (forms) as well as chronological variability (fabrics and forms) and geographical variability (fabrics).

Model 5 is the most complicated, and one hopes it would not be needed as it would be difficult to interpret. It is here as a 'backstop' should all other models fail to fit the data.

In practice there are many fabric-by-form combinations that cannot exist, and many fabric-by-context and form-by-context ones that do not exist. In statistical terms, the design of the data is *incomplete*. The theory for handling incomplete tables is known as the theory of *quasi-log-linear models* (BISHOP *et al.* 1975, 177-228). There is not the space to go into the details here: the main points to note are that the parameters have to be estimated by a more compli-

cated, iterative, process, and that the calculation of the number of degrees of freedom must be adjusted to take account of the numbers of 'empty' cells in the tables.

2.5 *Practical problems (i)*

Work on assemblages of Roman pottery from London and Silchester immediately revealed a serious problem: different types j gave different estimates $\hat{n}_j$, making it apparently impossible to transform a whole assemblage. This arises when not all types have the same statistical distribution of the values w_i, i.e. they are not statistically *homogeneous*. One reason could be a lack of archaeological homogeneity in the assemblage.

We say that an assemblage is *archaeologically homogeneous* if all the types in it have the same post-depositional history, i.e. have been subjected to the same series of 'events' (ORTON 1982a, 3). While archaeological inhomogeneity implies statistical inhomogeneity, the reverse is not always the case.

A simple numerical example should clarify this point:
Suppose an assemblage consist of two types, one of which breaks into 100 fragments and the other of which does not break at all. The history consists of a single breakage event followed by a 50% sampling. For the first type, some fraction of almost every vessel will be retrieved, leading to an average completeness of about 50%. For the second, about 50% of the vessels will be retrieved, with an average completeness of 100%. The assemblage is therefore archaeologically homogeneous but statistically inhomogeneous.

This situation arises whenever one (or more) type(s) is 'chunky', i.e. breaks into fewer measurable fragments than the general run of types, and is here called the *quantum effect*. It seems likely that whenever we encounter values of $\bar{w}$ that are significantly above average for their assemblage, we have 'chunky' types, but whenever the values of $\bar{w}$ are below average for their assemblage they indicate *residual* types (i.e. types which have a longer post-depositional history, and have been subjected to more events). In some cases there may be a 'high' group of types and a 'low' group, and it may not be clear whether the former is chunky or the latter residual. The answer is to compare across several assemblages; a 'chunky' type is likely to be chunky wherever is occurs, but residual types are likely to be residual is some assemblages but not others.

To be able to use the theory, we must detect the chunky types (if any) and remove their effect from the transformation. We proceed in three stages:

(i) detect the presence of chunky types by use of an F-test (KENDALL, STUART 1973, 522) on the w-values of all the types. Theory suggests, and experience confirms, that the test should be carried out on the logarithms of the values.

(ii) if the result of the F-test is significant, detect which types are the chunky ones. This is the problem known as *multiple comparisons* — the sorting out from an inhomogeneous set of $\{\bar{w}_j\}$ those particular values of j which cause the set to be inhomogeneous. Much has been written on this topic: a good general account is given by Miller (1981) and the latest overview is by Hochberg and Tamhane (1987).

After considerable experimentation, we found that the best approach for our problem was to carry out a t-test between the $\bar{w}$ for each type and the overall mean — a version of the LSD test (FISHER 1935). This test can be used to detect both chunky forms and chunky fabrics, as well as to compare contexts to help in the elucidation of site formation processes.

(iii) Here we must depart from out usual practice of dealing with general measures, and look at one in particular. We shall see in Section 4.1 that the most satisfactory measure for our purposes is the *vessel equivalent* (v.e.), which measures each sherd family as a proportion of the parent vessel. This is rarely practical, and it is usually necessary to estimate the v.e. from some easily recognisable part of the vessel, e.g. the rim, base or handle. This gives rise to an *estimated vessel equivalent* (eve). A measure based on part of a vessel is likely to be more chunky than one based on the whole vessel, because the part will always break into fewer fragments than the whole.

In this section, therefore, we look at the situation where the v.e.s of the types are not themselves chunky, but cannot be measured, but the eves, which can be measured, are chunky. It may be useful to think of eves as based on some specific part of the vessel, e.g. the rim, although the following argument will be more general.

We assume that for the 'ordinary' types, the eve provides an adequate estimate of the true vessel-equivalent, though even here some slight overestimation of the mean value seems likely. We need to find an estimate of the distribution of the v.e.s of the chunky types, and especially of the mean, which is better than that provided by the eves. If the assemblage is archaeologically homogeneous, then all types have the same distribution of the v.e.; we can therefore use the 'ordinary' distribution as an estimate of the distribution of the v.e.s of the 'chunky' types.

It might be argued that the difference in distribution between chunky and ordinary types removes the grounds for our assumption of archaeological homogeneity. This is not so: departures from archaeological homogeneity are signalled by unusually low values of $\bar{w}$, not by the high ones exhibited by the chunky types (see above).

What this means in practice is that we retain the total W for each chunky type, but replace the individual values by a set of values which have the same

total, but have the frequency distribution of the ordinary types. We remember (4), $n_j = ((m - 1)/m)W^2/S^2$ for all j.
Suppose the first m_j records are chunky and the other $m_{\sim j}$ are ordinary.
Treating the chunky types as unknown, we need to estimate their contributions to m, W and S^2.
Note that we need to estimate their contribution to m because of *ghost records*, i.e. vessels which are represented but not by measurable fragments (e.g. rims).
(i) m: estimate $\hat{m}_j = m_{\sim j}(W_j/W_{\sim j})$ and $\hat{m} = m_{\sim j} + \hat{m}_j = m_{\sim j}(W/W_{\sim j})$
(ii) W: *either* estimate $\hat{W} = W_{\sim j}(m/m_{\sim j})$
or take it as 'given' because we know W.
(equivalent since $\hat{W} = W_{\sim j}(\hat{m}/m_{\sim j}) = W_{\sim j}m_{\sim j}(W/W_{\sim j})/m_{\sim j} = W$).
(iii) S^2: $\hat{S}^2 = S_{\sim j}{}^2(\hat{m}/m_{\sim j}) = S_{\sim j}{}^2(W/W_{\sim j})$.
So (4) becomes $\hat{n} = ((\hat{m} - 1)/\hat{m})WW_{\sim j}S_{\sim j}{}^2$
$= (1 - W_{\sim j}/m_{\sim j}W)WW_{\sim j}S_{\sim j}{}^2$ — **(8)**.
This equation enables us to accomodate the quantum effect within our definition of a statistically homogeneous assemblage, and still use the transformation to pseudo-counts and contingency table theory.

2.6 *Practical problems (ii)*

Attempts to treat assemblage data from London and Silchester as three-way contingency tables had led us to adopt the theory of quasi-log-linear models because of the large numbers of zeros in the tables. Even so, the presence of many 'small' cells gave rise to two problems:
(i) frequently, the large number of such cells would contribute greatly to the number of degrees of freedom, but only a little to the overall X^2 or G^2 statistic, thus masking any potential significance of other parts of the table,
(ii) occasionally, the presence of a very small expected value together with a small observed value would give an abnormally high contribution to the X^2 or G^2 statistic.

The answer seemed to be to merge or delete rows and/or columns of the tables to remove 'small' cells. Conventional theory sets a general minimum 'expected' value of 5 per cell with an absolute minimum of 1 for the chi-squared test to be appropriate (COCHRAN 1954; CRADDOCK, FLOOD 1970 suggest a limiting average value of around 1 in the case when the expected values are roughly equal).

It was therefore decided to adopt the criterion that all cells should have an expected value of at least 1.0, and that merging of contexts, fabrics and forms should be carried out with this aim in view. Studies carried out on Silchester phase 1 showed that this level seems to give very satisfactory results.

The technique srd (simultaneous reduction of dimension) was devised to merge or delete rows and columns of a two-way table in a rational way. Full details are to be published (ORTON, TYERS 1990); the general principles are set out below:

(i) only rows or columns containing 'small' cells are 'flagged', i.e considered for merging (though they may merge with a row or a column which does not contain 'small' cells),
(ii) only merges which make sense archaeologically are permitted,
(iii) rows or columns are selected for merging on the basis of the reduction in X^2 brought about by merging them: those causing the smallest reduction are merged first,
(iv) when all permissible merges have been made, any rows or columns which still have 'small' cells are deleted from the analysis.

This is basically a two-way approach: to extend it to the three-way table we introduce the idea of a 'doubly-reduced' table, which is constructed as follows:

(i) suppose we have reduced the fabric-by-form marginal table, i.e. we are working with model <2><3> (see below: exactly analogous procedures hold for the other models).
(ii) we construct a new three-way table in which the rows are the unflagged fabric-by-form combinations and the columns are contexts.
(iii) we then reduce this table by srd, *except* that we allow only columns (i.e. context) merges. To allow row-merges would destroy the two-way nature of the marginal table.
(iv) when this procedure stops, all columns, consisting entirely of flagged cells are deleted, and all rows consisting entirely of flagged cells are themselves flagged. This flagging is carried back to the marginal table (in this case <2><3>), and any rows or columns in this table which now consist entirely of flagged cells are deleted.

The quasi-log-linear analysis is then carried out on the doubly-reduced tables. The models for the reduced table are listed below:

1A: <2><3>: independence of fabric and form in a reduced table,
1B: <3><1>: independence of form and context in a reduced table,
1C: <1><2>: independence of context and fabric in a reduced table.
IIA: <23><1>: independence of fabric-by-form and context it a doubly-reduced table,
IIB: <31><2>: independence of form-by-context and fabric in a doubly-reduced table,
IIC: <12><3>: independence of context-by-fabric and form in a doubly-reduced table.

The reduced tables do not necessarily share the same groupings of the three

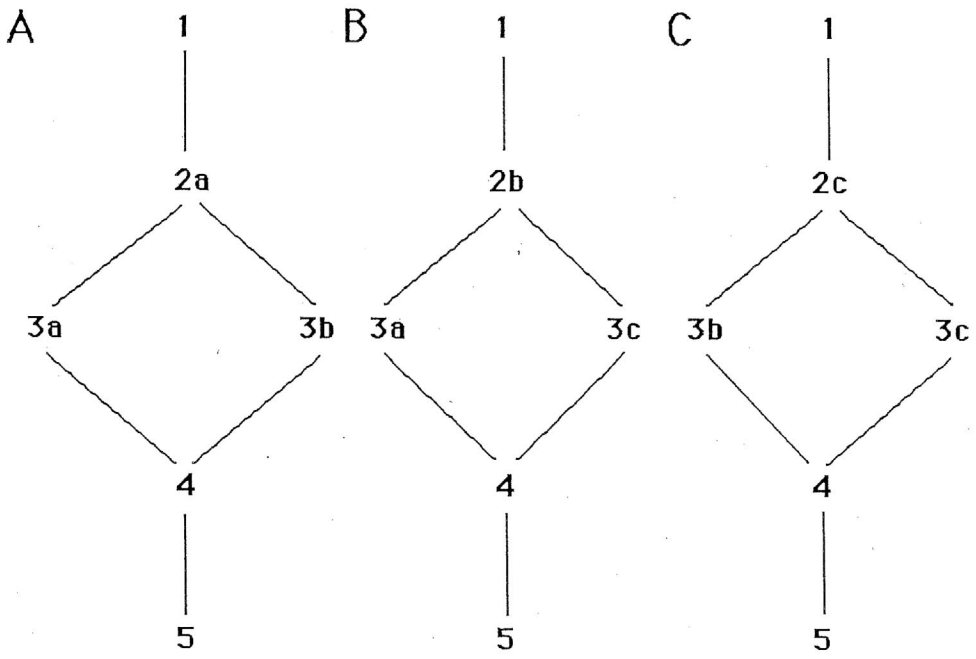

Fig. 2

variables. These models correspond to models 2a-c in the log-linear analysis.

Having ensured compatibility in the data structures, we can now integrate srd into the overall (quasi-)log-linear approach. We break the overall hierarchy of models (see Section 2.4) into three sub-hierarchies (see Fig. 2):

A: models 1, 2a, 3a and 3b, 4, 5;
B: models 1, 2b, 3a and 3c, 4, 5;
C: models 1, 2c, 3b and 3c, 4, 5.

We then run quasi-log-linear analysis on the doubly-reduced table of model IIA, using sub-hierarchy A, on the doubly-reduced table of model IIB, using sub-hierarchy B, and on the doubly-reduced table of model IIC, using sub-hierarchy C.

This leads to three 'accepted' models (see Section 2.4), which we must interpret. But before we do so we must reject any 'collapsed' tables (i.e. doubly-reduced tables with only one value of one or more variables). If all three tables collapse, we reject the entire dataset as inadequate. It would be too simple to look for the 'best' overall model, since the different hierarchies may be telling us different things. For example, model 2b might be the accepted model for one

grouping of contexts, while model 2c might be accepted for a completely different grouping. This would indicate form-by-context and fabric-by-context interactions cross-cutting each other, suggesting functional differences between some contexts but chronological differences between others. To choose one model as 'best' would lose one of these interpretations.

3. Results

3.1 *Correspondence analysis*

We here present the results of a correspondence analysis carried out on data from the Lime street site, after transformation to pseudo-counts but without any special treatment, e.g. no types were omitted from the calculation of the pseudo-totals. The program used is part of the iastats package (Duncan *et al.* 1988), based on one published by Greenacre (1984). Two analyses were made: forms by phase and fabrics by phase.

FORMS BY PHASE

The 1st principal axis (50% of total inertia) is dominated by FINE BOWL (90%) and phase 6 (97%). The 2nd principal axis shows a contrast between FLASK (6%), BEAKER (32%) and perhaps BOWL (11%) against AMPHora (13%) and FINE CUP (13%), matched by a contrast between phase 3 (50%) and phase 4 (46%). These hint a possible functional differences which need further investigation. One would not expect functional differences to appear clearly, if at all, at the level of phase-assemblages.

FABRICS BY PHASE

The 1st principal axis (35% of total inertia) is dominated by fabrics SHEL (56%) and SAND (20%), and by a contrast between phases 1 and 2 (77% and 9% respectively) and phase 4 (11%). On the 2nd principal axis, fabrics BB2 (54%) and KOLN (4%) stand out, as does the contrast between phases 5 (70%) and 3 (23%).

The picture is clearer when both are seen together (Fig. 3). Here we can see the characteristic parabola shape of a chronological sequence (Madsen 1988, 24). Only phase 6 is out of order; its data point is of low quality (lies 'off the plot') and there may be a functional difference. An 'early' fabric (SHEL) is at the beginning of the sequence and three 'late' ones (for this site) — BB1, BB2 and KOLN — are at the end. MORT occupies a central position; it is a 'rag-bag' type comprising a variety of rare and unidentified fabrics.

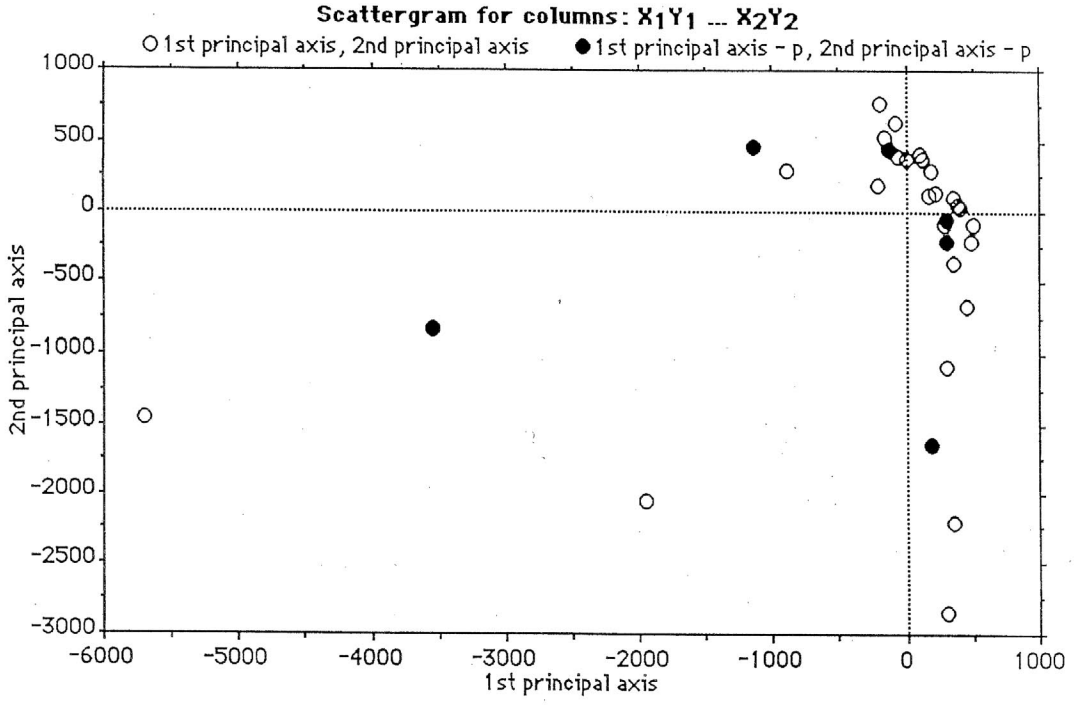

Fig. 3 — Hollow symbols represent types, solid symbols represent phases.

3.2 *Silchester phase 2*

Quasi-log-linear analysis with srd as a preliminary data-reduction technique technique was carried out on this dataset. At this stage we have allowed any groups of fabrics, forms or context to be merged as necessary: later versions of the program will allow for the input of specialist archaeological knowledge at this stage. There were initially 79 contexts, 42 fabrics and 11 forms.

Using the <23><1> model, the double srd reduces the dataset to 15 context-groups, 3 fabric-groups and 5 form-groups.

The quasi-log-linear analysis shows good fits for models 4 and 3b, but very bad fits for models 1, 2a and 3a. This is conclusively in favour of model 3b, i.e. of fabric-by-form and context-by-fabric interactions, but no direct context-by-form interaction. Such an interpretation is supported intuitively by a visual inspection of the final marginal tables.

Using the <31><2> model, the double srd reduces the dataset to 3 form-groups, 14 context-groups and 3 fabric groups.

The quasi-log-linear analysis shows a good fit for model 4, but very bad fits for models 1, 2a, 3a and 3c. This is conclusively in favour of model 4.

Using the <12><3> model, the double srd reduces the dataset to 6 fabric-groups, 5 form-groups and 4 context-groups.

The quasi-log-linear analysis shows good fits for models 3b and 4, but very bad fits for models 1, 2a and 3c. This is conclusively in favour of model 3b.

CONCLUSION

Model 3b is clearly the preferred one; it is accepted conclusively in *both* the hierarchies in which it occurs, and in the one in which it does not occur, *neither* model at level 3 is accepted. We conclude that « form is conditionally independent of context, given the fabric ».

In archaeological terms, the main differences between contexts are the fabrics present in them; any differences in forms just reflect the fabrics. This would seem to imply that the main differences between the contexts are chronological rather than functional.

The results are encouraging in that they are coherent — all the hierarchies « tell the same story ». However, this need not in general be the case. A dataset could, for example, have strong interactions between some contexts and fabrics, and between other contexts and forms. They should show up more strongly in different hierarchies, leading to the acceptance of (e.g.) model 3b in one and model 3c in another.

4. Practical implications

4.1 *Comparison of the behaviour of different measures*

In this section we look at the various measures described in Section 1.2 in the light of the theory developed in Section 2.3.

We recall equation (3) of Section 2.3, giving an expression for n_j, the equivalent sample size for type j of an assemblage. We saw that if a common value n existed, then we could transform the data to pseudo-counts and make use of the theory associated with data in the form of counts, e.g. contingency table theory and correspondence analysis. We saw that we could do this if all types had the same mean and variance of the measure w, after discounting chunky types.

We now look at the four main measures discussed in Section 1.2:

(i) vessels represented, (ii) vessel equivalents, (iii) weight, (iv) sherd count.

(i) vessels represented: in this case, $w_i = 1$ for all records, so it follows that the mean and variance are the same for all types j.

(ii) vessel equivalents: in this case, the condition holds if the completeness statistic has the same mean and variance for all types. This provides the rationale for the search for homogeneity and for means of achieving it.
(iii) weight: the condition holds provided that each type has the same mean weight and variance of weight. If weights are to be used, they must therefore be scaled to a common mean weight for all types. This in practice leads to a vessel equivalent based on weight.
(iv) sherd count: the condition holds if the statistic 'sherds present per vessel represented' has the same mean and variance for all types in an assemblage. This is unlikely to be the case in practice.

To sum up, only measures for which a complete vessel has the same value, regardless of its type, are suitable. The two such are vessels represented and vessel equivalents. However, earlier work (ORTON 1975; see Section 1.3) has shown vessels represented to have a serious problem of bias. The only suitable measure for statistical comparison of assemblages is vessels-equivalents. This justifies the use of eves, but leaves open the question of the best way of estimating vessel equivalents.

We here introduce the term *pottery information equivalents (pies)* for the transformed values of eves when applied to ceramic assemblages. The reason is that an assemblage totalling n pies has the same error structure, and therefore gives the same level of information about the proportions of its constituent types, as an assemblage of n complete vessels.

It should be noted that there is no simple relationship between pies and eves or any other statistic. The same number of eves can, for different types, give rise to different pies, depending on the distribution of the completeness of the type. Within an assemblage, the pie of a type depends not only on the type itself, but also on all other types. The pie is therefore a thoroughly contextual variable.

4.2 *The level of recording*

For the theory of Section 2.2. to hold, the observations (in our notation, records) must be independent of one another. If two records, relating to the same assemblage, also relate to the same vessel, then those records are correlated and the general formulae we have derived in Section 2 do not hold. We can envisage this by considering an assemblage that consists of two types of complete vessels. The variance of the population of each type can be calculated through the standard binomial formula, or equation (1) of Section 2.2, which gives the same result in this case. If we then break each vessel in half and record the halves separately, applying equation (1) gives a value for the variance which

is 1/4 that of the original value. In other words, we have halved the standard deviations of our estimates by simply breaking the vessels. Since we cannot actually increase the amount of information we have by breaking vessels, there is a nonsense here. We conclude from this *reductio ad absurdum* that the formula no longer holds — it must be modified to take account of the correlations between the records of the two halves of each broken vessel.

This implies that smallest unit that is permissible to record (for statistical purposes) is the *sherd family* (SMITH 1983), i.e. the set of all sherds from the same vessel in the same context or assemblage. In practice, we can relax this requirement; since sherds which have no measure do not contribute to the formula, they need not be included in the family, which becomes the *measurable sherd family*. For example, if the measure is rim-eves, then only rim sherds contribute to the formula, and for the purposes of calculating proportions we need only sort the rim sherds into families. There may be other reasons for the more difficult and time-consuming task of sorting body sherds into families, but nevertheless there are potential savings. If, however, the measure is weight, then we cannot avoid the task of sorting all sherds into families, since all sherds have a weight.

It might be thought that, by saying that eves should be recorded by measurable sherds families, we are re-introducing the vessels-represented approach by the back door, since the number of records can be taken to be the number of vessels represented. This is not so, for three reasons:

(i) since eves are based on measurable (e.g. rim) sherds, the number of records is less than (or, in rare cases, equal to) the number of vessels represented. The difference is the number of vessels represented only by sherds of measure zero (e.g. body sherds).

(ii) the treatment of the two statistics is different if contexts are merged. The number of record for the merged context is the sum of the numbers of the original contexts, while the numbers of vessels represented is reduced to allow for cross-joining sherds.

(iii) the reason for rejecting the vessels-represented statistic, namely its serious and unpredictable bias, is unaffected if we happen to be able to deduce it from other statistics.

Two sorts of records may cause problems:

(i) *the conflated record*: a record which contains measures of more than one vessel, e.g. the total measure of one type in one context. The data can be used but they will over-state the true variance. The seriousness depends on the degree of conflation.

ii) *the over-detailed record*: a record containing less than a complete measurable sherd family. Variances cannot be calculated unless the data have been

flagged so that sherd families can be put together by merging records. If sherd families cannot be put together, the formula given here cannot be used. The data must be aggregated (e.g. to type-within-context level) and treated as an example of case (i).

This argument might be seen as a case for using rim-eves as the measure, rather than (rim + base) eves or weight, especially if there is doubt about which base or body sherds belong to which rim. However, it is only one voice in a complex decision, and may be over-ruled by other factors.

A general point is that care and extra attention at the recording stage can give benefits in terms of reduced standard deviations of estimates of proportions. The extent to which the extra precision is worth the extra work is a matter of judgement.

Some theoretical work, confirmed by simulation, shows that split families, resulting in over-detailed records, are not likely to be a serious problem unless we are dealing with assemblages in which *both* completeness and brokenness are high. At least 25%, and possibly at least 40%, of an assemblage would have to be recorded as one or more records per family to lead to 'serious' (greater than 10%) errors in the estimates of standard deviations of proportions.

4.3 *Definition of fabrics and forms*

Throughout this work, there has been a tension between the need to aggregate data to make datasets acceptable for statistical purposes, and the need to maintain a fine enough level of detail for useful archaeological interpretation. In general, it is likely that some grouping of both fabrics and forms, as defined by conventional archaeological methods, will be needed before statistical analysis can be undertaken.

While it would be possible to leave this grouping in the hands of srd, under archaeological guidance as to which merges are allowed, it may be better to grasp the nettle and attempt to form preliminary groupings before starting the statistical analysis. This must be done carefully, with the aims properly defined.

The study of chronology has aspects other than the quantitative comparison of ceramic assemblages, which may frequently over-ride it. The two most important are (i) archaeological stratigraphy and (ii) 'absolute' dates, provided by (e.g.) coins or particularly diagnostic types of pottery such as samian ware, often in the form of TPQs. A single piece of such qualitative data, e.g. the stratigraphic relationship between two contexts, or the presence of a single dated sherd in a context, may give more precise information than any amount of quantitative data. On the other hand, if such data do not exist, we must fall back on the quantitative data for (e.g.) seriation. Further, a quantitative analysis of well-stratified assemblages may well provide a framework for the dating

of less well stratified or isolated assemblages (see for example ORTON 1982b).

The combining of different sorts of data in a chronological study is a topic in itself, and is beyond the scope of this project. Nevertheless, we offer some general guidelines, while aware that they may need to be over-ruled in some practical situations:

(i) forms should where possible be grouped according to style or decoration, as these are the aspects most likely to reflect chronological change.

(ii) it may make sense to group fabrics according to source and, if possible, phases within sources.

If, however, we are looking for spatial (inter-site) differences, we should concentrate on groups of fabrics based on source. Groupings of forms may not be possible unless forms distinctive of sources can be identified.

A search for functional or social differences demands a third approach. A grouping of fabrics according to technological aspects might be more appropriate, e.g. fine and coarse wares, or perhaps a finer division based on the degree of tempering. Forms should be grouped into functional types, e.g. cooking pots, drinking vessels.

It is clear from this discussion that no one typology, of fabrics or of forms, will serve for all purpose. The recorder is thus faced with a dilemma — which to use for the basic recording of the pottery? The ultimate uses of the data will not be known at the time of recording, and it seems undesirable to straitjacket the data by immediately-perceived needs. The answer is to record in as fine a level of detail as is possible within the resources available, and to indicate ways in which types may be grouped for different purposes. The same data can then be analysed in different ways according to the groupings employed.

4.4 *Definition of assemblages*

Assemblages can be defined at many levels — the individual context, or groupings of contexts such as features, phases, sites or even a whole town or region. Again, there is a tension between the need to aggregate to create groups of the size needed for statistical analysis, and the need to preserve the fine detail of individual contexts. As in the cases of fabrics and forms, it may be preferable to carry out a preliminary grouping rather than leaving it all to srd.

As before, there is scope for choosing different groupings to meet different needs. If chronology is the main concern, grouping contexts into stratigraphic phases will make sense. For inter-site spatial analysis, aggregation to site-groups is an obvious choice, but has a pitfall if sites are not exactly contemporaneous. Different proportions of different fabrics on the sites may then represent chronological as well as spatial differences. Grouping by phase within site may

then be a safer option. To look for functional differences, groupings should be based on the supposed 'function' of contexts, though there is a danger of circular argument here, and a finer level of detail may be safer. Social differences may be marked by differences between assemblages at the level of individual buildings or features (e.g. pits or associated groups of pits).

On any site, there is likely to be more than one such need. Pottery should therefore be recorded according to the finest level of stratigraphic detail (usually the context), with indications of which groupings of contexts would be appropriate for particular purposes. It may be desirable to sub-divide extensive layers spatially (e.g. by grid squares), but this should not be seen as an endorsement of « digging by spits », which can wreck an attempt at ceramic analysis (see e.g. GREEN 1980, 39, where division of a thick layer of dumping into three horizontal layers (nos. 412, 430 and 433, see Fig. 11) prevented any statistical analysis).

There are statistical implications in the merging of contexts to create larger assemblages. If a vessel occurs in more than one of the merging contexts, this leads to the situation of non-independent observations described in section 4.2. In principle, one should re-sort the material and re-catalogue any such vessels, but this is at best very time-consuming and at worst impossible, and very rarely done in practice. However, theoretical considerations and the outcome of simulations (see Section 4.2) suggest that in most practical situations the merging of contexts is unlikely to cause serious increases in standard deviations.

CLIVE R. ORTON - PAUL A. TYERS
Institute of Archaeology
University of London

Acknowledgements

This work has been supported by SERC grant GR/E 95873. We are grateful to the Museum of London Department of Urban Archaeology for permission to use and publish data from Lime Street, and to Jane Timby and Michael Fulford for allowing us to use data from the recent Silchester excavations. We thank Nick Fieller for valuable theoretical advice.

BIBLIOGRAPHY

BEDWIN O., ORTON C. R. 1984, *The excavation of the eastern terminal of the Devil's Ditch (Chichester Dykes), Boxgrove, West Sussex*, « Sussex Archaeological Collections », 122, 63-74.

BISHOP Y. M. M., FIENBERG S. E., HOLLAND P. W. 1975, *Discrete Multivariate Analysis*, Cambridge, Massachusetts, MIT Press.

BLOICE B. J. 1971, *Note*, in G. J. DAWSON, *Montague Close Part 2*, « London Archaeologist », 1, 250-251.

BRADLEY R., FULFORD M. G. 1980, *Sherd in the analysis of occupation debris*, « Bulletin of the Institute of Archaeology », 17, 85-94.

CARVER M. O. H. 1985, *Theory and practice in urban pottery seriation*, « Journal of Archaeological Science », 12, 353-366.

COCHRAN W. G. 1954, *Some methods for strengthening the common chi-squared tests*, « Biometrics », 10, 417-451.

COCHRAN W. G. 1963, *Sampling Techniques*, (2nd edition), New York, Wiley.

CRADDOCK J. M., FLOOD C. R. 1970, *The distribution of the chi-squared statistic in small contingency tables*, « Applied Statistics », 19, 173-181.

DUNCAN R. J., HODSON F. R., ORTON C. R., TYERS P. A., VEKARIA A. 1988, *Data analysis for archaeologists: the Institute of Archaeology programs*, London, Institute of Archaeology.

EGLOFF B. J. 1973, *A method for counting ceramic rim sherds*, « American Antiquity », 38(3), 351-353.

FIENBERG S. E. 1977, *The Analysis of Cross-classified Categorical Data*, Cambridge, Massachusetts, MIT Press.

FULFORD M. G., HODDER I. R. 1974, *A regression analysis of some late Romano-British fine pottery: a case study*, « Oxoniensia », 39, 26-33.

GLOVER I. C. 1972, *Excavation in Timor*, PhD thesis, Austrialian National University.

GRAYSON D. K. 1984, *Quantitative Zooarchaeology*, London, Academic Press.

GREEN C. 1980, *Roman Pottery*, in D. M. JONES, *Excavations at Billingsgate Buildings 'Triangle' Lower Thames Street, 1974*, London and Middlesex Archaeological Society, Special Paper n. 4.

GREENACRE M. J. 1984, *Theory and Applications of Correspondence Analysis*, London, Academic Press.

HINTON D. A. 1977, *'Rudely made earthen vessels' of the twelfth to fifteenth centuries AD*, in D. P. S. PEACOCK (ed.), *Ceramics and Early Commerce*, 221-238, London, Academic Press.

HOCHBERG Y., TAMHANE A. C. 1987, *Multiple Comparison Procedures*, New York, Wiley.

HULTHÉN B. 1974, *On choice of element for determination of quantity of pottery*, « Norwegian Archaeological Review », 7(1), 1-5.

KENDALL M. G., STUART A. 1973, *The Advanced Theory of Statistics*, Vol. 2 (3rd edition), London, Griffin.

MADSEN T. (ed.) 1988, *Multivariate Archaeology, Numerical Approaches to Scandinavian Archaeology*, Jutland Archaeological Society Publications, 21, Århus, Århus University Press.

MARQUARDT W. 1978, *Advances in archaeological seriation*, in M. SCHIFFER (ed.), *Advances in Archaeological Method and Theory*, 1, 257-314, New York, Academic Press.

MILLER R. G. 1981, *Simultaneous Statistical Inference* (2nd edition), New York, Springer-Verlag.

MILLETT M. 1979a, *The dating of Farnham (Alice Holt) pottery*, « Britannia », 10, 121-77.

MILLETT M. 1979b, *An approach to the functional interpretation of pottery*, Institute of Archaeology, Occasional Papers 4, 35-48.

MILLETT M. 1979c, *How much pottery?*, Institute of Archaeology, Occasional Papers 4, 77-80.

MOORHOUSE S. 1986, *Non-dating uses of medieval pottery*, « Medieval Ceramics », 10, 85-124.

ORTON C. R. 1975, *Quantitative pottery studies: some progress, problems and prospects*, « Science and Archaeology », 16, 30-35.

ORTON C. R. 1982a, *Computer simulation experiments to assess the performance of measures of quantity of pottery*, « World Archaeology », 14, n. 1, 1-20.

ORTON C. R. 1982b, *Pottery evidence for the dating of the revetments*, in G. MILNE, C. MILNE, *Medieval Waterfront Development at Trig Lane, London*, London and Middlessex Archaeological Society, Special Paper n. 5, 92-99.

ORTON C. R. 1985, *Two useful parameters for pottery research*, in E. WEBB (ed.), *Computer Applications in Archaeology 1985*, London, Institute of Archaeology, 114-120.

ORTON C. R., TYERS P. A. 1990, *A technique for reducing the size of sparse contingency tables*, submitted to *Computer and Quantitative Methods in Archaeology 1990*.

REDMAN C. L. 1979, *Description and inference with the late medieval pottery from Qsar es-Seghir, Morocco*, « Medieval Ceramics », 3, 63-79.

SCHIFFER M. B. 1987, *Formation Processes of the Archaeological Record*, Albuquerque, University of New Mexico Press.

SMITH M. F. 1983, *The study of ceramic function from artifact size and shape*, Unpublished doctoral thesis, University of Oregon.

SOLHEIM W. G. 1960, *The use of sherd weights and counts in the handling of archaeological data*, « Current Anthropology », 1, 325-329.

VINCE A. G. 1977, *Some aspects of pottery quantification*, « Medieval Ceramics », 1, 63-74.

YOUN C. J. (ed.) 1980, *Guidelines for the processing and publication of Roman Pottery from excavations*, Directorate of Ancient Monuments and Historical Buildings, Occasional Paper n. 4.

ABSTRACT

As well as solving two long-standing theoretical problems, this work shows great potential for the interpretation of ceramic assemblages, and has implications for the way in which pottery is catalogued. Different sorts of interpretation (functional, chronological, distributional) are possible at different levels of grouping (context, phase and site assemblages).