AN INTRODUCTION TO USING STATISTICAL TECHNIQUES FOR
CLASSIFYING STONE TOOLS

Ian Herbertson

INTRODUCTION

The idea for this paper arose from a discussion with students and friends about the difficulties encountered in the application of statistical techniques to the classification of artefacts, and my own conviction that provided simple rules are followed difficulties need not arise. Rather, in the same vein as Ballin (2000), I am offering this as a discussion of the basic ideas of classification and the help available from employing statistical techniques. If “ΣXn/n” leaves you cold, don’t worry it need not. An understanding of elementary descriptive statistics is quite easy and once grasped can lead on to more advanced techniques. As a start the following questions are important:

1. What is the archaeological question being asked?
2. How should the stone tools be classified to answer the question?
3. If elementary techniques are used, have I mastered them?
4. Do I really need to employ more advanced techniques?

If the answer to the last question is yes, then it is worth spending time on the elementary techniques first. Using them provides a feel for the data recorded and how they relate to the artefacts measured. More advanced techniques should not be too readily applied without due consideration, otherwise the statistical treatment can overshadow the archaeological question.

WHY CLASSIFY ARTEFACTS?

Providing order

As Ballin (2000) points out, in any excavation there are usually a large number of finds of similar artefacts. To describe each item individually would be an enormous undertaking so a means by which large amounts of data from the finds can be reduced to manageable proportions is very helpful. Although there might be some loss of detail, if that detail is merely repetition then summarising the data assists rather than obscures the information.

Comparison

The volume of finds presents a problem when comparing with others. Summarising the data on the finds by classifying them is a convenient way of doing this. It also enables comparison with other classified artefacts to see what light each can shed on the other.

1 18 Weir Close, Buckden, St Neots, CAMBRIDGSHIRE. PE19 5TH
Practicalities

Classification data, as Fagan (1991) observes, can be analysed on computers giving the opportunity to process large volumes of data in sophisticated ways, for example, by using statistical software packages. This facilitates comparison with other studies quite readily. Since 1991 computing power, especially of PCs, has increased and become much cheaper; statistical software packages that contain powerful techniques like multivariate analyses (see below) can be run on a laptop on site. Statistical software packages in the form of freeware are available from the Internet. A number of statistical techniques are available in Excel.

Helping to focus an enquiry

Debenath and Dibble (1992) point out that in classification “…the research questions should dictate the choice of measurements taken”. In lithic analysis they point out that choosing variables is not easy because the factors affecting morphological variability are little understood. Deducing a sound system of classification of stone tools may lead to the distinctions early hominids made themselves and that in turn may improve our understanding of their behaviour (see Fagan 1991).

A CONCEPTUAL FRAMEWORK

There are a number of ways of classifying stone tools but first it would be helpful to take a fundamental look at classification because it focuses attention on aims and difficulties. The American philosopher of science, Carl Hempel (1970), begins with the very basic point that classification divides a class (set) of objects into sub-classes. A set of necessary and sufficient conditions are specified against which an object is measured to decide if it falls within the class in question. In practice this will not be a once and for all stipulation of a selection criterion for inclusion with a class; it will be revised or refined as the classification proceeds. But at the end of the work the criteria should be conditions that are usable by others. Other workers have argued that some classification is too complex to be captured in a single typology so specifying the conditions may not be straightforward.

It is important to be clear whether a criterion has been selected beforehand as part of the objective of the research (e.g., looking for retouch at the distal end of a blade) or whether the criterion has emerged objectively from the application of some analytical technique (for example, the ratio of length:width of flakes). In the latter case it is important that any causal relationship is recognised, as inevitably there is between these two attributes.

Hempel points out that clear criteria of application are also necessary before they can be applied to cases. This is a different to the criteria above, which tackle the development of classes per se. It is important for replication of analysis that there is a uniform method of applying the classification scheme.

He also points to the difficulty of deciding cases that seem to fall near the boundaries of different classes. Debenath and Dibble (1992) experienced problems in distinguishing between “…choppers/chopping tools and cores…even some bifaces are very similar to cores…”. They concluded that this shows it is virtually impossible to have too rigorous a system of classification,
although Read and Russell (1996) argue that statistical methods can reveal underlying distinctions not available to the eye.

**CLASSIFICATION METHODS**

Fagan (1991) describes typology as classification based on the construction of types. He calls it a search for structure amongst the artefacts or amongst the attributes of those artefacts, and through which further questions can be answered by statistical analysis. This kind of analysis does lead to types that have statistically significant similarities and dissimilarities. The results still have to be interpreted; the ratio of the length:width of an Oldowan stone chopper may be exactly the same as that of an *Homo erectus* femur, but would such a contrast be meaningful? The precision in the mathematical treatment of measurements does not in itself improve understanding of what an artefact is or its use.

Classification of data about artefacts often leads to a conception of an average artefact for a given type. This might coincide with an actual artefact but it might lead to an 'average' type that is used as a definition of that artefact (Fagan 1991). This is helpful because it can be used for comparison with others for both common and different features. On the other hand averages require cautious interpretation because they can mask distinguishing details.

Fagan (1991) talks of four types of classification, being descriptive, chronological, functional and stylistic, but usually used in conjunction with each other. They could all include technological attributes such as raw material and hammer type.

Two more contrasting approaches to classification are the agglomerative and divisive approaches. In the former individual artefacts are compared to others and discernible types are established. A divisive strategy starts the other way around by beginning with the population and sub-dividing it into groups.

Dissection and modal approaches are a further pair of methods. In dissection the set of artefacts are divided up by matching them to an arbitrarily divided set of measurements of one variable attribute, such as scrapers grouped by length into, 0-10cm, 11-20, 20-30 etc. With the modal approach artefacts are sorted into groups where the measured variable is matched up to a modal value (a statistical measure of commonness); the scrapers might fall into four groups by length, say, 0-12 cm, 40-55 cm, 60-80cm and >90cm. Focusing on just one attribute will certainly produce groupings but whether they are meaningful depends on the objective. They do lend themselves to graphical representation that are useful for interpretation.

Bordes (1988 plate 1 ff) provides an extensive classification of stone tools analysed by shape, marshalling the tools into categories such as, *Nucléus Levallois* or *Biface légéniforme*. Of course, this presents the reader with an extensive set of morphological benchmarks. But this presentation has certain limitations for analysis (apparent similarity of shape can mask metrical differences that can be revealed by precise measurement) that can be remedied by measurements. Whilst two tool types may look similar perhaps further interpretation is available based upon measurements and comparison of these measurements. Statistical treatment of these measurements may reveal associations and, moreover, their significance rather than chance association can be tested (see below).
STATISTICAL METHODS AND TECHNIQUES

Basic terms and scales
There are certain fundamental terms that need to be understood, before exploring further, the first being the bedrock of most statistics – the normal distribution. This simply shows that for any set of objects with the measure X, most of the object’s readings will naturally cluster around the mean (the sum of measurements divided by the number of items measured) and that those whose measurements are increasingly further from the mean occur less frequently (a bell-shaped curve). The mean measures how central or common the measurement is. How good a measurement it is, is expressed by the standard deviation which is the mean difference of each measurement from the mean itself. The variability of the measurements may be expressed by the (statistical) variance, which is the square of the standard deviation.

There are also four different kinds of scale: nominal; ordinal; interval; and ratio (Andrefsky 1998). These levels of measurement (Marsh 1988) have different properties; some simply allow for categories to be distinguished whereas others allow for more precision in measurement. It is important to be clear which of these levels is being used before the statistical method is chosen.

- A nominal scale simply represents a pair of mutually exclusive attributes such as the wear of a stone tool; it is either present or absent.
- An ordinal scale puts things in order without telling us anything about the measurement in the ordering, eg. size - artefacts are small, medium or large.
- Interval scales are like ordinal scales but they tell us something about the relationship between the artefacts ordered. These scales have an arbitrary base (or zero) points. These can be useful to researcher but not directly related to underlying artifacts – a bit like Celsius we just chose “0” as freezing point.
- A ratio scale is like an interval scale but the base or zero point is a real. Computations can be performed on the measurements, eg. lengths of stone tools, conceptually, start at 0 cm and continue up to N cm; a 10 cm blade is half the length of a 20 cm blade.

Statistical techniques relevant to archaeologists tend to fall into two camps: those that demonstrate difference and those that demonstrate association. This broad distinction should always be borne in mind as most classification covered here turns on this distinction. Measures of association can be further divided into those that use continuous or ordinal data (testing whether two variables are related); and those that use categorical data (are two characteristics related?).

Many statistical techniques test an observation against an expected result; a hypothesis is formulated, which rejects the proposition under consideration (null hypothesis) and tests against a mutually exclusive alternative hypothesis.
There are statistical tests that allow artefacts’ measurements to be used to tell whether the artefacts come from the same or different population. They are useful when differences are not obvious from visual inspection, and even when visual distinctions can be drawn, statistics can help quantify the strength of the difference (see Mitchell 1997). This approach establishes the distribution of the finds, and assumes that the data exhibit a normal distribution. Having calculated the distributions and variances, a test statistic can be calculated which shows how significant the difference is (ie not just by chance). The value of the test statistic is looked up in the relevant statistical table and the answer in terms of significance read off.

Some statistical tests that are likely to be encountered are ones that point out the significance of differences between a sample mean and an actual mean (student’s t-test) or differences in variances (chi²). An analyst will set a significance test to avoid rejecting a null hypothesis when it is true. The underlying idea behind such tests is to identify unlikely results as having a very low probability.

Sometimes it is useful to understand the relationship between attributes or variables, achieved through a correlation coefficient. This is a number whose range is -1 to +1, where values approaching +1 are positively correlated, those approaching -1 are negatively correlated, and where there is no correlation (ie no relationship) the coefficient approaches 0.

As well as using statistical techniques to answer questions about whether there is an association between artefacts they can also be used to tell the strength of association. Classification can be carried out from visual inspection but it would take a very keen eye to order the artefacts to reflect precisely how similar or dissimilar they are.

A technique known as cluster analysis can be used and the strength of the association measured by a chi² test and its significance revealed by looking up the result in the appropriate statistical table. This is a simple method that could be used to compare actual frequencies of artefacts, perhaps between assemblages.

A good introduction to more advanced techniques is Stephen Shennan’s Quantifying Archaeology (1997). He describes a number of more sophisticated techniques for analysing data where it is not possible to make assumptions about whether or not the variables are independent or dependent. These are helpful in classifying data where even after classification the volume of results prohibits seeing the wood for the trees.

The first of these techniques, cluster analysis (Everitt 1993), looks for significant groupings in the items and to account for their patterns of association. Items are grouped according to similarity and these clusters are themselves clustered to produce an overall analysis of the artefacts. The objective is to discover patterns in the data. The techniques are applied to measurements of variables such as length, breadth and thickness. One potential problem is that considering all the data can obscure significant information. For example, two sets of bifaces may be compared using 20 variables but any difference may be detected by concentrating on, say, just two.
A family of techniques known as multivariate analyses can help in this situation. Bryan Manly's *Multivariate Statistical Primer* (1994) provides a good introduction. Principal Components Analysis (PCA) produces indices or artificial variables that are uncorrelated to the original variables. These indices are called the principal components and they can be arranged in descending order, $Z_1$, $Z_2$, ..., $Z_n$. The aim is to reduce the number of variables and to capture the variability in a few. In a sense the first principal component might account for much of the variability of the original variables. Although a stone tool might be measured across many variables, (eg. Roe, below) the variability between those stone tools might rest on length. Although the other measures might vary, the variance is so slight compared to variation of length that focusing on length is enough to provide an understanding of the stone tools. A similar but more statistically rigorous technique is Factor Analysis.

A technique useful to help distinguish artifacts from assemblages is discriminant function analysis, where, for example, two different types of industry might be distinguishable, allowing the artefacts to be placed in two groups. This could be cross-checked with other evidence to conclude that they are indeed from the two different industries. This statistical technique lends itself to being tested for significance. Instead of looking for evidence for difference, an archaeologist may be looking for similarity across two sets of variables. An appropriate technique here might be canonical correlation analysis. The aim of the analysis is to focus on correlations between the different measurements of the stone tools to see if there is any association between them.

Sometimes techniques have been used in conjunction. Orton (1980, 54) describes a “dendrogram” approach to cluster analysis, which assesses similarity or dissimilarity and plots the result on a chart. He gives as an example a dendrogram of spearheads where similarity was assessed using seven measurements. The resulting dendrogram is rather unattractive to read. A principal components analysis showed that the first two variables (total length and length of head) account for 83% of the variability so there was significance attached to a few variables, although these are difficult to see in the dendrogram.

**A CASE STUDY**

Roe (1994) carried out a metrical analysis of handaxes and cleavers found at Olduvai Gorge. He analysed these morphologically with the assistance of statistical tabulations to demonstrate both differences and similarities between various artefacts. He used largely the same method as his study of British Lower and Middle Palaeolithic handaxes, but with some adjustments to take account of the greater distinction between handaxes and cleavers in Africa than in Britain. He defined a cleaver as possessing a characteristic transverse or oblique cutting edge at the tip end, where the length of the oblique edge is greater than half its breadth. The measurements he used were defined as follows:

<table>
<thead>
<tr>
<th>Cleaver edge length (CEL)</th>
<th>The length of the straight line between the two extremities of the cleaver edge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cleaver edge angle</td>
<td>Angle ($\leq 90^\circ$) between CEL and long axis of cleaver.</td>
</tr>
<tr>
<td>L</td>
<td>Length</td>
</tr>
<tr>
<td>B</td>
<td>Breadth</td>
</tr>
<tr>
<td>T</td>
<td>Thickness</td>
</tr>
<tr>
<td>T1</td>
<td>Thickness at a point one fifth of the length distant from the tip end.</td>
</tr>
<tr>
<td>B1</td>
<td>Breadth at a point one fifth of the length distant from the tip end.</td>
</tr>
<tr>
<td>B2</td>
<td>Breadth at a point one fifth of the length distant from the butt end.</td>
</tr>
<tr>
<td>L2</td>
<td>Distance from the butt end to the point along the long axis at which maximum breadth occurs.</td>
</tr>
</tbody>
</table>
These measurements were not all used in isolation but were sometimes combined as ratios (e.g. Th/B or CEL/B). For each measurement or ratio a table was produced which showed sample details, means, standard deviations, frequency distributions etc.

Roe also used measures of significance (t-values page 154. In a series of tables (1994, 158-201) he illustrated which pairs of sites showed consistent or frequent differences and which pairs did not differ ‘significantly’ from each other. From these he tabulated ‘difference scores’ (maximum 100). For example, the developed Oldowan bifaces at Bk Bed II and Acheulian tools excavated by John Waechter, gave a high score of 86. In contrast, the artefacts from sites Bk and PDK trenches I-III scored 0.

Overall, he suggested that that three successive groups of Acheulian industry could be identified in the Olduvai stratigraphy:

1. Bed II contained heavier, thicker handaxes and rarely cleavers compared to Bed IV;
2. Bed IV contained handaxes with regularity of shape and cleavers more elegantly made;
3. post-Bed IV artefacts demonstrate a higher level of competent manufacture (1994, 204).

He further argued that the analysis showed a great variation in the Acheulian artefacts in Bed IV.

Such analyses may come up with surprises and require careful interpretation. Roe points out that it is important to check the frequency tables for any large values that might be distorting the mean, although the standard deviation should reflect this. It is important that the results can be used to check the data and the statistics. The main body of the report comes from these numerous tables which have to be studied carefully and attention given to the sample sizes and levels of probability.

As Roe rightly points out, statistics are not enough on their own. It is easy to succumb to the risk of taking as fact that which is only a possibility. Roe also warns that his conclusions are “...based entirely on a study of biface morphology, and not on full analyses of complete lithic industries”. Statistical techniques can support archaeological analysis; they are not as noted in this paper’s introduction the whole or final story. The extent to which the variety can be seen through visual inspection would provide a measure of the utility of the statistical analysis.

**CONCLUSION**

Most would agree that classification is an essential aspect of lithic research, whether it be of individual artefacts or groups of artefacts. It provides order, a means of comparison, facilitates the use of statistics through computerisation and can often help to focus research. It is important, though to understand some of the underlying principles of classification. Stipulation of clear criteria for inclusion in a class is essential, together with clear criteria of application. It is important to be clear whether an order is being imposed or unearthed. Equally, the archaeological significance of attributes needs to be fully appreciated.

Classification can be assisted through the appropriate use of statistical analysis. Simple techniques are often sufficient to answer quite complex questions. The more complex the technique the more necessary it is to understand the archaeological significance of the attributes. As ever,
statistics provide a means of understanding a given set of data, but which will only facilitate interpretation, if the archaeological meaning of that data is fully understood.

BIBLIOGRAPHY


