



Plus ça Change? - 25 Years of Statistics in Archaeology

Clive Robert Orton

Abstract

Taking as its starting point the seminal Mamaia conference of 1970 and the first CAA conference (both attended by the author), this paper examines the changing role of statistics in archaeology, as perceived by archaeologists, statisticians, and the author himself. It concentrates on how and why archaeologists have used statistics, rather than on the particular techniques they have used, although there is of course an interaction between the two. It discusses the questions: do archaeologists understand what statistics is about? do they really want to? do modern computer packages help or hinder? what (if anything) has changed?

1 CAA and the history of statistics in archaeology

The start of the series of CAA conferences did not mark a particularly significant point in the history of the use of statistics in archaeology. Indeed, statistics was not on the agenda at the first conference, which dealt entirely with 'Information Retrieval', only appearing on the agenda for the second conference in 1974. The 'Quantitative Methods' sub-title did not appear until the 1981 London conference (the first to be held outside Birmingham), although statistics had featured strongly in some of the intervening years. So before looking at the period 1973-1997, I shall try to set the activities of CAA in the field of statistics in a wider context, both historically and geographically.

CAA started shortly after the seminal conference *Mathematics in the Archaeological and Historical Sciences* which was held in Mamaia, Romania, in 1970 (Hodson *et al* 1971), and which can be seen as a formal international recognition of the status of quantitative methods in archaeology, or as a rite of passage from a phase of experimentation to one of acceptance and optimism. Participation was mainly from Romania and the UK, but ten other European countries, Canada, India, and the USA were also represented. This conference followed the international conference *Archéologie et Calculateurs*, held by the CNRS in Marseille in 1969 (Gardin 1970), with mainly French participation, but eleven other European countries, Canada, Turkey, the USA, and Venezuela were also represented. The first CAAs were neatly sandwiched between two conferences on the related subject of sampling: *Sampling in Archaeology* in the USA in 1973 (Mueller 1975) and *The Role of Sampling in Contemporary British Archaeology* in 1977 (Cherry *et al* 1978).

The picture given by the Marseille and Mamaia conferences is of widespread but mainly uncoordinated experimentation and development across Europe, the USA and Canada, and a need to bring researchers together to share ideas and experiences. They followed about twenty years of relatively small-scale and isolated research, the start of which I shall mark by some sampling experiments undertaken by Treganza and Cook (1948) and the little-known work of

Myers (1950). The 1950s saw statistics impinging on debates about typology (e.g. Spaulding 1953), and attempts to formalise the previously intuitive technique of seriation (Robinson 1951), both in the USA. The first review paper came in the mid-'50s (Heizer and Cook 1956), followed at the end of the decade by what appears to have been the first quantitative methods conference in archaeology, held in 1959 at Burg Wartenstein, Austria, under the auspices of the Wenner-Gren Foundation for Anthropological Research (Heizer and Cook 1960); many of the contributions would now be recognised as archaeometry.

The papers that really put quantitative methods 'on the map' in the USA and the UK included Binford's study of Mousterian variation (Binford and Binford 1966), Cowgill's overview of techniques (Cowgill 1968), Clarke's study of beaker pottery (Clarke 1962; 1970), and Hodson's work on Iron Age cemeteries and artefacts (Hodson 1969; 1970).

By the early 1970s there was a feeling of 'spring-time' in the air, with green shoots appearing and bringing the promise of a bountiful harvest. Ideas as well as people seemed to be coming together and pointing the way forward. Outright opposition was rare (e.g. Hawkes 1968) and could be dismissed as a Luddite reaction by people who did not really understand what was going on. CAA seemed well placed to be part of this movement. A quotation from Spaulding (1960, 83) sums up this feeling that surrounded the birth of CAA:

"... archaeologists are universally users of quantitative methods in the broadly descriptive and comparative sense and that they frequently resort to more explicit statistical techniques for treating special problems. I... predict that applications of the latter type will become increasing common... The failure of our predecessors to employ elaborate quantitative techniques was not the result of either ignorance or stupidity; they were properly and understandably preoccupied with the gross description and ordering of the grand cultural types, one does not need regression analysis or chi-square to grasp and describe the essential difference between an Acheulian and a Magdalenian assemblage. Similarly, the current trend towards utilization of more elaborate and theoretically

sophisticated quantitative methods is not dilettantism or a naive attempt to appropriate some of the current prestige of the physical sciences by aping their methods. It is rather an effort to discover and communicate subtle differences between closely related assemblages, so that details can be inserted into the broad outline. It is not an exaggeration . . . to predict that the future of archaeology is in large measure bound up with the success or failure of this effort.”

2 Where are we now?

We can contrast this confidence with a range of recent views:

“. . . there seems to be as yet little assimilation amongst archaeologists of the basic notions of experimental design which . . . plays an essential part in all applications of statistics. Perhaps . . . archaeostatistics can be said to be immature, and not yet of the same status as its better-known companion subjects.” (Fieller 1993, 280).

“In certain respects the development of quantitative methodology in archaeology can be . . . seen as imitating developments in other areas such as geography or taxonomy. Such imitation has been decried, with one major concern being the fear that methodologies thus imported embody assumptions inappropriate to archaeological data and problems. This . . . has led to calls for the development of methodology that is ‘congruent’ or ‘concordant’ with archaeological problems (e.g. Carr, 1985; Aldenderfer, 1987a).

. . . there is an attendant danger that can be highlighted by analogy with the development of the ‘quantitative revolution’ in geography. There, quantitative methods were associated with a particular theoretical standpoint -- positivism -- that engendered unrealistic expectations about what such methods might achieve. The inevitable disappointment with, and backlash against, this approach encompassed (in some quarters) quantitative methods with the result that methodologically useful babies were unfairly thrown out with the theoretical bathwater.” (Baxter 1994, 7)

“There has been some decline in popularity in recent years because statistical methods are less popular with ‘real world’ archaeologists than twenty years ago. Statistical approaches were very much part of the ‘new archaeology’, now called processual archaeology, and post-processualists seem to feel less at ease with this subject. But there are signs that in the near future the pendulum will swing again in the other direction. The application of ‘hard science’ in archaeology is on its way back. The main reason for this is that much of the funding for scientific archaeological research is by way of ‘hard science’ projects linked to the environment. We are not sure whether statistical applications in archaeology are part of ‘hard science’ but they will certainly benefit from this development.” (Kamermans and Fenema 1996)

“Although simple quantitative and statistical procedures have had a long history of use in archaeology, the decades of the 1960s and 1970s witnessed the “Great Borrowing”, in which archaeologists searched through all sorts of sciences, ranging from physics to ecology, in an effort to find methods to help make sense of archaeological data (Clark and

Stafford 1982). . . there has been much misuse of methods and a fragmentation of effort. There have been frequent expressions of dissatisfaction with the use of quantitative methods in archaeological research; . . . quantitative methods have yet to fulfil the promise we thought they offered.” (Aldenderfer 1987a, 90)

“How archaeologists present the evidence they collect, the hypotheses they make . . ., and the interpretations they construct is about to change. Technological, sociological, and economic forces outside archaeology along with shifts in popular expectations of archaeological presentation are fostering a situation in which the image is soon to become the major component of archaeological argument.” (Ross 1996)

I would add to these my personal observations that most archaeologists are still afraid of statistics, regarding the subject as third, after death and taxes, in a list of inevitables to be avoided for as long as possible. The thought that one day they will have to “do some statistics” on the data they have collected is pushed to the back of their mind until it can be avoided no longer, at which point the answer to their plea of “what can I do?” is probably the apocryphal one to the lost traveller: “if that’s where you want to go, I wouldn’t start from here”.

3 How did we get here?

So how did we get from the optimism of 25 years ago to the fragmented situation of the 1990s, with feelings about quantitative methods that range from at best disappointment to at worst irrelevance? Like true (if old-fashioned) archaeologists, we can try to answer this question by breaking the 25-year period into phases. Aldenderfer (1987b, 12) explored the idea that “topics in a scientific discipline follow a predictable course of four stages: early exploration, discovery, consolidation and accommodation’, suggesting that the late 1970s/early 1980s might mark the transition of quantitative methods from the discovery phase to the consolidation phase. However, he also pointed out that “quantitative methods have penetrated archaeology at different rates in different subject areas” (*ibid.*: 13), and that a more detailed model was therefore needed. We can add that a simple chronological model does not explain the fragmentation of opinion in which we now find ourselves. For both reasons, it seems more useful to look at the past 25 years in terms of the tensions within a series of themes, which lie on a spectrum from the purely statistical to the purely archaeological, for example:

Statistical		
Exploratory statistics	<-->	Confirmatory statistics
Statistical analysis	<-->	Statistical design
Classical statistics	<-->	Bayesian statistics
Numerical presentation	<-->	Graphical presentation
Commercial packages	<-->	Home-grown’ packages
Sources of techniques; archaeology, other disciplines, ‘all-purpose’ sampling metaphor as a basis for interpretation		
	<-->	Taphonomy

Archaeological

New Archaeology' <--> post-processual archaeology.

Although these themes are inter-related, it is worth looking at each to see how it may have contributed to our present situation. The first three themes can cause problems because archaeologists perceive statistics as a monolithic discipline.

It may come as a shock to discover that statisticians disagree with one another, and there are different schools of thought and different emphases within the discipline; this shock can quickly lead to confusion in the face of conflicting advice

.Find type	Context date by latest century								Totals
	to 11th	11th	12th	13th	14th	15th	16th	later	
beads	1	2	1	-	4	2	8	19	37
belt-fittings	5	8	5	12	17	27	4	13	91
buckles	4	14	17	57	55	44	34	33	258
building ironwork	25	20	24	58	39	21	19	24	230
combs	9	12	6	2	2	-	4	10	45
furniture	16	10	6	20	8	4	8	8	80
gaming equipment	8	6	9	10	7	5	4	8	57
hones	17	28	7	42	29	23	5	9	170
horse furniture	5	8	6	26	12	5	4	11	77
horseshoes	23	36	52	104	63	17	19	33	347
jewelry	24	17	13	25	21	10	8	22	140
keys	12	32	23	38	21	15	10	19	170
kitchen equipment	11	6	13	13	9	6	12	10	80
knives	38	56	42	66	48	22	34	44	350
lighting equipment	5	20	35	50	19	7	2	6	144
locks	7	9	6	12	10	5	5	3	57
mortars	-	-	-	7	14	14	4	8	47
points	-	-	6	3	3	13	43	77	145
querns	7	18	7	10	6	-	1	1	50
reliquary fittings	8	21	25	24	15	17	6	6	122
sewing equipment	2	-	4	2	8	1	3	21	41
sewing pins	1	1	1	16	15	39	64	242	379
shoes	60	27	6	13	-	-	-	-	107
tacks	19	16	20	21	7	13	5	13	114
tenter-hooks	2	3	9	43	18	5	-	2	82
textile tools	16	34	30	36	19	5	1	9	150
toggles	2	2	2	5	-	1	-	1	13
tools	26	52	11	20	16	10	7	12	154
weapons	5	2	5	25	6	5	5	11	64
weights	1	5	4	3	5	3	6	6	33
Totals	359	465	405	763	496	339	326	681	3834

Figure 1: Numbers of finds from medieval Winchester by type and context date (from Biddle, Barclay and Orton 1990)

3.1 Explore or confirm?

The first theme covers the thorny issue of hypothesis tests, which are much mis-understood by archaeologists, and which generate much discussion among statisticians as to

when and whether they should be used. What a hypothesis test does is to construct a very simple hypothesis (the 'null' hypothesis), usually representing 'no change' (e.g. when a new treatment is applied to a material), 'no difference' (e.g. between two C14 dates), 'no association/correlation' (e.g. between two variables) or 'randomness' (e.g. of a

spatial pattern). The 'test' itself then ascertains whether it is reasonable to believe this null hypothesis in the light of data which are subsequently collected. Two inherent problems are:

1. it is a conservative procedure, i.e. it favours the null hypothesis since the data have to be strong enough to force us to change our mind about the null hypothesis,
2. it is heavily dependent on sample size: if we take a big enough sample, we are likely to be able to 'reject' a null hypothesis as implausible, even though the observed difference or pattern in the data is so small as to be of no practical importance.

When used in archaeology, I see hypothesis testing as a necessary precaution that precedes analysis and interpretation. For example, if there is no statistical evidence that the dates of two 'events' are not the same, it would be unwise to build an interpretation on the observed difference between the two C14 determinations that represent them. We hope that the null hypothesis will be rejected, so that we can proceed to interpret our data with our backs covered.

Unfortunately, this statistical procedure has sometimes been linked with a rigid 'hypothetico-deductive' approach to archaeology (Shennan 1988: 4) which may require the testing of complicated hypotheses that are quite beyond the procedure's scope. This has led to two equally erroneous and contrasting perceptions:

1. it may be fine in theory, but it doesn't work in practice,
2. it's only for people who believe in the 'hypothetico-deductive' approach.

A balanced view is lost.

The opposite approach, of exploratory data analysis, also has its dangers. Its objective is to make apparent patterns that may be present in the data, but are not immediately obvious (this is especially true of multivariate data). This is valuable, e.g. when using a correspondence analysis plot to expose the patterns hidden in a contingency table. As an example, contrast Figs. 1 and 2: the chronological pattern of the artefact types is present in the table (Fig. 1), but is very hard to see. The plot (Fig. 2) contains less information than the table (only 64%), but several chronological patterns are immediately apparent, for example the contrast between querns, which are 'early' at the bottom of the plot, and mortars, which fulfil the same function and are 'late' (at the top of the plot).

The danger is that the human eye is good at seeing patterns, *even when they are not there*, leading to the risk of over-interpretation. A quick check of the statistical significance of the overall pattern could remove this risk. What one may not do is to detect a possible pattern, e.g. in the value of a correlation coefficient, and then test the null hypothesis that it takes that value, *on the same data*. To take this route, one needs to explore half the dataset for a possible pattern, and then test it on the other half.

Archaeological criticism of the 'hypothetico-deductive' approach, combined with statistical warnings about the

dangers discussed above, can create feelings of confusion which can deter archaeologists from using statistics at all.

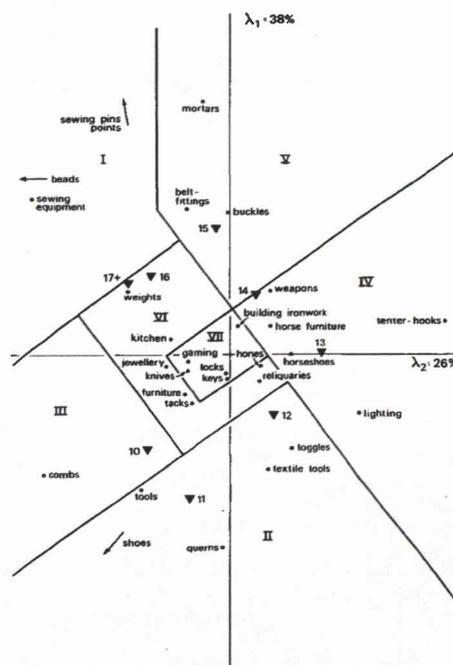


Figure 2: Correspondence analysis of numbers of finds from medieval Winchester by find type and context date. (from Biddle, Barclay and Orton 1990)

3.2 Analysis by design

Most archaeologists see statistics simply as an analytical tool: you collect your data, you analyse them, and you interpret the results. The idea that statistics has anything to do with the collection of data, or of the preceding stage of research design, is alien to them. Given the wish to postpone involvement with statistics for as long as possible, this is hardly surprising, but it has unfortunate consequences. It is very galling for an archaeologist to be told by a statistician that they have collected "the wrong data" or "not enough data" or, most puzzling of all, "too many data". "What's it to do with them?" they ask, "all I wanted was an analysis."

As an illustration, suppose an archaeologist wants to distinguish between different pottery fabrics on the basis of the statistical distribution of the sizes of the inclusions in the clay that are visible in a cross-section, and suppose that the differences lie in the relative proportions of the larger inclusions. There is a choice of method for describing the distributions: (s)he can either (a) count the numbers of inclusions in each of a series of size ranges, or (b) count the numbers of inclusions of each size range that are 'hit' by a sample of points in the cross-section. In either case, the data can be (correctly) compared by use of the Kolmogorov-Smirnov test. But if method (a) is chosen, it may well suggest that fabrics that are visually quite distinct are statistically "not significantly different". A few minutes statistical thought would suggest that method (b), with its greater emphasis on the larger inclusions (because, being larger, they have a higher probability of being 'hit' by the sampling points), is more likely to result in such differences being assessed as "significantly different". This

actually happened, resulting in much extra work for the archaeologist (who here remains nameless to avoid embarrassment). The point is that, by not bringing in statistical considerations until the analysis stage, the archaeologist chose the 'wrong data' (for this particular problem) and created much unnecessary work. Choice of the correct analytical technique for the data in hand was not enough.

Further, each statistical technique carries its assumptions, and if a dataset does not fit in with those assumptions, the technique should not be used. Conversely, if a certain technique is to be used, the data must be chosen to fit its assumptions. Also, statistical input at the design stage can offer useful advice on the quantity of data needed to answer a certain question to a certain level of confidence: too few data and the results are inconclusive, too many and resources are wasted.

Resistance to statistical input at any stage other than "analysis" thus hamstring the archaeologist's endeavour, causing inefficiency in the use of both time and money, and leads to disappointment with their results and disenchantment with statistics as a whole (see Fieller, quoted above).

3.3 Bayes rules ok?

There are some things that conventional ('Classical') statistical theory does not do very well. To appreciate the sort of problem that can arise, try working out the effect on two C14 dates of being told that one is stratigraphically earlier than the other. Suddenly, along came a new type of statistician, the Bayesians, who say that this sort of problem is grist to their mill, and start talking about "prior knowledge" and the like. They have a view of statistics that is in many ways closer to the archaeologists': that probability is subjective, and that statistics is about "the orderly influencing of opinions by data". Just when the archaeologist is getting excited, they let slip that the mathematics is hideously complicated, the software is not user-friendly and anyway you need at least a Sun workstation to run it. Feelings of let-down, almost of betrayal, can result. Statistics was bad enough, but two rival versions of the discipline are just too much.

Fortunately, areas where this approach is likely to yield pronounced benefits are beginning to be identified (e.g. interpretation of scientific dating techniques). Software is now available to perform the more routine tasks (e.g. OxCal, see Ramsey 1995), and ways of providing help for the more difficult problems are being developed (Buck and Christen, this volume).

3.4 How graphic are graphics?

"One picture is worth a thousand words", they say, but how many numbers is it worth? One of the developments in statistics in the past 25 years has been the growth of visual approaches to data, initially focused on the 'exploratory data analysis' (eda) school (Tukey 1977), and more recently under the name of 'data visualisation'. This is all to the good; we have already seen (section 3.1) how much easier it is to spot a pattern in (for example) a scatter plot

than a table of figures. But we have also seen that the plot cannot, by itself, tell us whether it actually means anything: for that, we need some serious number-crunching. We need to remember what numbers are good at and what graphics are good at. At its simplest, numbers give information and graphics convey ideas or impressions. If I want to catch a train, I look up a timetable, not a timegraph, but if I want to show the relationship between the rows and columns of a table I use a scatter plot, not the table itself (compare Figs. 1 and 2 again). The reader of the graphic has to take on trust that there are enough data behind it to give it credibility.

Archaeologists can miss out in two ways here: they may have discovered analytically a very interesting pattern, but be unable to display it visually in a way that 'grabs' their reader, or they may plot a very pretty picture that means absolutely nothing because it is based on a handful of sherds (or coins, or flints, . . .). Once again, there may be a feeling that statistics has let them down.

3.5 Off the shelf or out of your head?

A very strong external trend over our period has been the increasing availability of commercial computer packages for data analysis and visualisation. At the start of our period, commercial software for data analysis was almost non-existent, and researchers had to write their own. Indeed, this was often a major part of a research project, and archaeologists required a proficiency in Basic, Pascal, C, etc. Today, there is a wide choice of software available for statistical analysis, from simple spreadsheet bolt-ons (like the Analysis Toolpak in Excel), through serious analytical packages like SPSS and BMDP, to very serious ones like Statistica, S-plus, GLIM and Genstat. Specialised archaeological software is now only needed for techniques that are unique to archaeology, such as seriation (e.g. Winbasp, iagraves, etc.), analysis of dating evidence (e.g. OxCal) and 'fragmentary data' (e.g. Pie-slice).

What have we gained, and what have we lost? We have gained freedom from an endless re-inventing of the wheel, and the need to acquire low-level programming skills. We have not acquired a sense of discrimination: even in Excel, dozens of techniques are available at the click of a button. The problem is -- which to choose? Here archaeologists are still on their own, and easy availability does not make their choice any easier. We have freed them from the need to know how to do the calculations (though I would claim there are still benefits in knowing), but not from knowing which calculations. Some have tried to avoid the decision by hitting the 'all options' button (e.g. in cluster analysis) and have drowned in their print-out. Ideas that this aspect could be handed over to an expert system have quietly disappeared.

Histogram

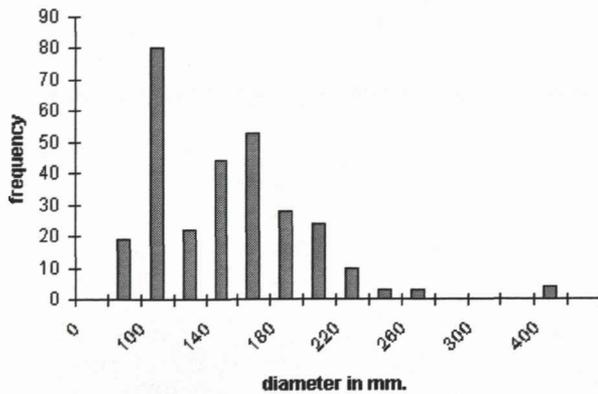


Figure 3: Example of so-called histogram produced by the use of Excel.

Also, I think we have lost quality. As graphical output has become widely available, it has often become worse in a communicative sense. For example, Excel cannot produce true histograms; it produces bar charts with a label saying “histogram” (Fig. 3). This can be seen as simple innumeracy, or as an example of our Alice in Wonderland world, where people believe that by attaching labels to things they make them become what the label says. Another bug-bear is the use of 3-dimensional graphics for 2-dimensional datasets; they obscure rather than illuminate, and can even mislead (Fig. 4) (see also Lockyear 1994). The comments of Tufte (1983) should be compulsory reading for all archaeologists.

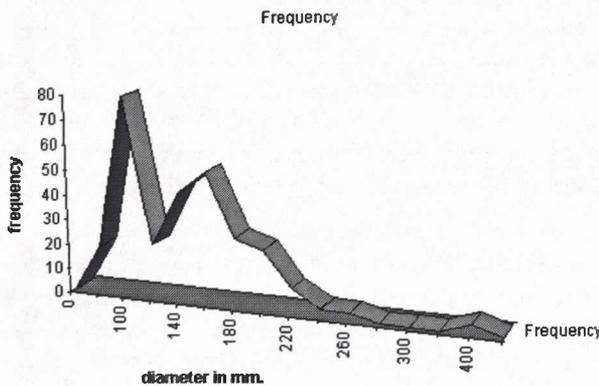


Figure 4: Example of three-dimensional line chart produced by the use of Excel.

Thus the undoubted technological and commercial advances have not made life easier for the user or for the reader, although they have made it faster to obtain results. The clarity that could come from well-displayed data is still too often missing.

3.6 The Borrowers

Many statistical techniques have arisen in a particular field of application, and have later been transferred, more-or-less satisfactorily, to other disciplines. One of the beauties of mathematics (of which statistics is just part) is the ability of abstract systems to simultaneously represent widely differing physical realities. For example, the same mathematical model can be used to represent the condition

of objects in a museum store, and the fate of heart-transplant patients (Orton 1996). This means that many of the ‘problems’ of analysis that archaeologists encounter may have already been met, and overcome, by someone working in a quite different discipline, and that a literature search may therefore save months or even years of work. Archaeologists have been well aware of this, and are in fact notorious ‘borrowers’ from other disciplines, as the quotation in section 2 from Aldenderfer (1987a, 90) suggests.

The start of our period saw Aldenderfer’s “Great Borrowing” in full swing, with geography and ecology as favoured source disciplines, but the net stretching far wider, into areas as unlikely as astronomy. Some spectacular failures, for example the prediction by trend-surface analysis that North Oxfordshire pottery was probably made in South Wales or the North Sea led to a reappraisal (Hodder and Orton 1976: 166), and to a recognition that there had to be a measure of ‘concordance’ in the underlying models.

Inevitably the pendulum swung, in this case to the NIH (not invented here) syndrome, which argued that techniques should only be used in the discipline for which they were originally designed. This led to some bizarre ‘home-grown’ applications being used in situations where perfectly acceptable ‘all-purpose’ statistical techniques were available.

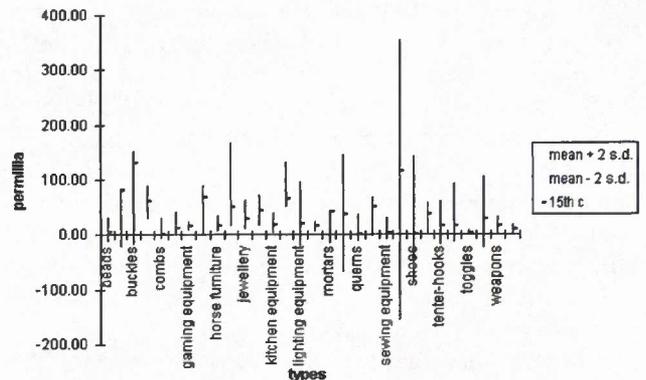


Figure 5: Example of a ‘permillia’ plot. The column of figure 1 representing the 15th century is plotted against the other columns across the range of artefact types. The bars show the ranges of ± 2 s.d. about the overall mean values. To display all of figure 1 would require a separate plot for each column.

A good example comes from the study of coin hoards, which usually involve simple datasets consisting of the counts of different types of coins in different hoards. The data can therefore be studied by techniques appropriate for contingency tables, such as the chi-squared test, log-linear analysis and correspondence analysis (e.g. Lockyear 1996). Dissatisfaction with existing techniques led Reece (1981) to devise his own technique, which he called “permillia”. It first expresses the composition of each hoard as a series of “permillia” (i.e. per thousands, instead of the more usual percentages) of coins of each type. The permillia of the types are plotted, and compared to the mean value of the permillia of each type across all hoards, together with a range, usually the mean \pm two standard deviations. An

example of this approach, applied to the data of Figs. 1 and 2, is shown as Fig. 5. Values that lie outside their range are taken as evidence of divergence from the underlying pattern. This approach has three statistical flaws:

1. the permillia values from which each standard deviation is calculated do not have the same statistical distribution. Even under the implicit null hypothesis that all the hoards have the same pattern (and hence all hoards have the same mean permillia for any chosen coin type), the values all have different standard deviations because they are based on hoards of different sizes. The calculation of the standard deviation in this way (i.e. on percentages or permillia), and its use in constructing a hypothesis, is invalid.
2. probability arguments based on “ ± 2 s.ds” are implicitly based on normal distributions. The extent to which permillia tend not to be normal is demonstrated by the high proportion of ranges that include negative values,
3. the use of multiple significance tests (which is implicit in this approach) alters the significance levels of such tests. For example, in a small dataset of ten types in ten hoards, there would be 100 such tests. Even in purely random (artificial) data, the chance of at least one difference appearing to be statistically significant at the 5% level is very high (99.4%), and even at the 1% level it is better than ‘evens’ (63.4%).

Reece withdrew this technique under statistical pressure (Reece 1988, 22-3), but it has recently re-appeared (van Arsdell 1996) and the argument must be repeated (Orton 1997a). Re-inventing square wheels is a peculiarly futile pursuit.

Thus archaeologists must tread a path between two dangers -- the inappropriate use of borrowed techniques on one side and the unnecessary recreation or parodying of existing techniques on the other. Unless they are skilled in modelling, it may not be obvious to them why a technique is inappropriate, and if they do not familiarise themselves with a wide literature, they may be surprised to find “it’s been done before”. Small wonder if they feel “damned if they do, and damned if they don’t”, and confused about the whole business.

3.7 What’s in a sample?

Sampling, as a means of learning about the characteristics of a population by studying some of its members, is a well-established statistical approach to many practical problems, from opinion polls to industrial quality control. Its *raison d’être* is that it is simply not necessary, and sometimes impossible, to examine all of a population to find out what we need to know about it. There is no point in devoting more resources to a problem than are needed to resolve it -- the result is just a waste of time and/or money.

Archaeologists have always been samplers *par excellence*, in a practical sense. Problems of resources require sampling to be employed on excavations, in scientific dating, and in many forms of non-intrusive fieldwork.

Traditionally, it has been done on an intuitive basis, using ‘professional judgement’, although there were some surprisingly early attempts to examine the effects of sampling on data collection (e.g. Treganza and Cook 1948). Around the start of our period, the ‘New Archaeology’, with its emphasis on research design, led to this process becoming more explicit and statistical, as evidenced by the two conferences mentioned in section 1 (Mueller 1975; Cherry *et al* 1978). The over-rigid use of formal schemes caused some absurdities (Hole 1980), but it became generally accepted that this was an area where statistics had much to offer. The question of professional judgement remained, and under the influence of PPG 16 (DoE 1990) many archaeologists have sought the Holy Grail of a ‘minimum percentage’ for field evaluations (English Heritage 1997, 57) and have been upset when statisticians tell them that it does not exist. A more fruitful approach would be to agree a level of probability that a site contains no significant archaeological remains, that would entitle one to declare it ‘clean’, and design fieldwork projects to achieve this level. Bayesian statistics can help here too, by allowing archaeologists to incorporate their ‘prior knowledge’ (the closest to professional judgement it can get) into the calculations, thus reducing sample size (Orton 1997b). The price is a distinct step up in complexity, which would require specialised software and training.

But there is another, more controversial, way in which archaeology is sometimes thought of as sampling. Even if we excavate a complete site (supposing we could define what that meant), we would not discover all the artefacts and ecofacts that once existed on that site. One of the more perceptive questions that visitors ask is “where did all the missing bits [pottery, bone, etc.] go?” The answer is that we don’t really know, but we haven’t got them. This creates a very different sort of sample, since it is neither designed nor purposively selected: it just happened. Do the same rules of inference apply? Almost certainly not. But it is easy to unthinkingly behave as if they did. For example, if we compare the proportions of a certain type of pottery at several sites in a region, we implicitly assume that those proportions reflect in some way the proportions in some ‘parent’ populations from the sites. Why else would we compare them? Ideally, we want our estimates of the ‘parent’ proportions to be free of the effects of the post-depositional histories of the assemblages. This may mean attempting to model those effects (e.g. by recursive sampling, see Orton 1982) or by looking for ‘invariants’ -- parameters which are likely to survive the post-depositional histories (Orton 1993). Either way, the outcome is likely to be a set of statistical criteria on (for example) how pottery should be quantified, which may be welcomed by the archaeologist, or may be seen as an unwarranted interference. The tendency of statistics to get in the way of archaeologists’ wish to “do their own thing” may cause resentment.

3.8 Newer than ‘New’

It was a great misfortune for archaeology that the use of statistical techniques became associated in the popular mind with the school of thought known as the ‘New

Archaeology' (aka processual archaeology). The link was claimed by the New Archaeologists as part of their bid for academic respectability, although statistical techniques were in use in archaeology well before the rise of New Archaeology, and some of their most cogent advocates were very scathing about New Archaeology: "we find its claims greatly exaggerated and therefore dangerous. . . . a bizarre mixture of naiveté and dogmatism" (Doran and Hodson 1975, 5). When processual archaeology came under criticism from the post-processual school, statistics tended to be found guilty by association. This is regrettable, and is perhaps an example of what has been called "The curse of originality". For a new approach to make itself appear distinctive, it may be necessary to reject certain aspects of an older paradigm with which it has really no quarrel (just as politicians tend to dismiss *all* their opponents' policies, even the ones with which they secretly agree). There may also be an element of genuine misunderstanding, based on the over-stated claims for statistics made by the New Archaeologists.

So what does statistics still have to offer archaeology in a post-processual era? Above all, it offers method. Even post-processual archaeologists use statistical techniques (e.g. Shanks and Tilley 1992: 137-240), but it seems to me that they have a problem with method. Archaeological analysis is primarily about *data* -- we may argue about which data, and how to collect or measure them, or from which theoretical perspective we view them, but ultimately they are data like any other. The view that archaeologists create data is an attack of *hubris*: the data are there, and our role is to select and record. If you create your data and I create mine, then we are in the position of the witnesses to a traffic accident that consisted of a collision between two stationary vehicles.

Conversely, what statistics does not do is to impinge directly on archaeological theory, although theory can impinge on statistics by creating a need for new forms of analysis, such as social status analysis (socistat), devised by Hodson for his analysis of the Hallstatt cemetery (Duncan *et al.* 1988). The role of statistical method is to mediate between data and theory; the creation of that theory is entirely the archaeologist's responsibility. But the acid test for any theory is whether it stands up to the data (i.e. whether the data support it or not), and this is where statistical method can have its say, not only in the analysis but also in the selection and collection of data.

4. Conclusion

What can we say now, in the late 1990s, about this long but uneasy relationship between archaeology and statistics? It has gone through various phases; some represent the usual oscillations that accompany the adoption of a new methodology, while some were caused by the vagaries of archaeological fashion, in which schools of thought have over- or under-played the role of statistics to suit their own agendas. It has also seen the introduction of many new techniques, most notably correspondence analysis in the 1980s (Bølviken *et al.* 1982) and kernel density estimation in the 1990s (Baxter and Beardah 1996); this aspect will be dealt with in more detail by Wilcock (this volume).

But other things, I suggest, have not changed. Many archaeologists still approach statistics with the air of a soldier about to enter a minefield. They sense rewards on the far side, but also sense traps for the unwary en route. Even when they are trying very hard, something may blow up in their face (see van Arsdell 1996; Orton 1997a). I think this is to some extent a self-fulfilling prophecy. Archaeologists are, as a rule, independent people who like to "do their own thing" and do not like the imposition of external constraints. To them, statistics should be "on tap, but not on top" (an old Civil Service saying, referring to the role of specialists in government, but equally applicable here). This means that while they are happy to borrow tools from the statistical bag, they are not keen to adopt the statistical ethos of data, for example as portrayed in the Statistical Cycle (Fig. 6). To do so would be to submit themselves to an outside authority, and the attitude of "it's my data and I'll do what I like with it" is very strong. It follows that although they may know how to perform techniques, they do not fully understand the implications of their data, and thus lay themselves open to the very sorts of errors that they fear they will make.

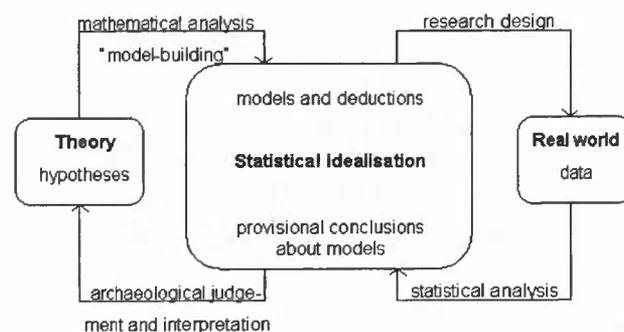


Figure 6: The Statistical Cycle.

What can be done? At the risk of sounding like a 'data fascist', I think that if archaeologists want the benefits of statistical techniques, they must submit to the disciplines that the use of such techniques inevitably bring. Once acquired, such disciplines can actually be liberating rather than constricting, since they free the archaeologist from a wide range of inappropriate decisions and actions, and from the need to try to think things out from first principles. Archaeological data are not uniquely difficult (whatever archaeological pride may say), and many of their features crop up in other subjects (where, of course, they may be handled just as badly). Recognition that statistics is not just a bag of techniques, but encompasses issues like the appropriateness of models, the quality and survival of data, and efficient design, can open up a wide range of relevant experience to archaeologists.

This will not come about overnight; some archaeologists may be too 'old' to change, others may cling to their philosophical uniqueness and purity, while others may clothe their fear in posturing. Perhaps only the real education of the next generation of archaeologists will put this rocky relationship on a proper footing.

Acknowledgements

I am grateful to Sue Laflin for suggesting this topic, to John Wilcock for agreeing to share it, and to Kris

Lockyear for commenting on a draft of this paper.

Bibliography

- Aldenderfer, M S, 1987a On the structure of archaeological data in Aldenderfer M. S. (ed.). *Quantitative Research in Archaeology*, 89-113. Newbury Park Sage Publications.
- Aldenderfer, M S, 1987b Assessing the impact of quantitative thinking on archaeological research: historical and evolutionary insights in Aldenderfer M. S. (ed.). *Quantitative Research in Archaeology* 9-29. Newbury Park Sage Publications.
- Baxter, M J, 1994, *Exploratory Multivariate Analysis in Archaeology*. Edinburgh Edinburgh University Press.
- Baxter, M J, and Beardah, C C, 1996 MATLAB routines for Kernel Density Estimation and the Graphical Representation of Archaeological Data in Kamermans H. and Fenema K. (eds.). *Interfacing the Past Computer Applications and Quantitative Methods in Archaeology CAA95* (Analecta Praehistorica Leidensia 28), 179-184.
- Biddle, M. Barclay K, and Orton, C, 1990 The chronological and spatial distribution of the objects in Biddle M. *Object and Economy in Medieval Winchester* (Winchester Studies 7ii), 42-73. Oxford Clarendon Press.
- Binford, L R, and Binford, S R, 1966 A preliminary analysis of functional variability in the Mousterian of Levallois facies *American Anthropology* 68, 238-95.
- Bølviken, E E, Helskog K, Helskog K, Holm-Olsen I M, Solheim, L, and Bertelsen, R, 1982 Correspondence analysis: an alternative to principal components *World Archaeology* 14, 41-60.
- Carr, C, (ed.) 1985 *For Concordance in Archaeological Analysis*. Kansas City: Westport Publishers.
- Cherry, J F, Gamble, C, and Shennan, S, 1978 *Sampling in Contemporary British Archaeology*. (British Archaeological Reports British Series 50).
- Clark, G, and Stafford, R, 1982 Quantification in American Archaeology *World Archaeology* 14, 98-119.
- Clarke, D L, 1962 Matrix analysis and archaeology with particular reference to British Beaker pottery *Proceedings of the Prehistoric Society* 28, 371-82.
- Clarke, D L, 1970 *Beaker Pottery of Great Britain and Ireland*. Cambridge Cambridge University Press.
- Cowgill, G L, 1968 Archaeological applications of factor cluster and proximity analysis *American Antiquity* 33, 367-75.
- DoE 1990 *Planning Policy Guidance: Archaeology and Planning*. London Department of the Environment.
- Doran, J E, and Hodson F R, 1975 *Mathematics and Computers in Archaeology*. Edinburgh Edinburgh University Press.
- Duncan, R, Hodson, F R, Orton, C R, Tyers P A, and Vekaria A, 1988 *Data analysis for archaeologists: the Institute of Archaeology programs*. London Institute of Archaeology.
- English Heritage 1997 *English Heritage Archaeology Division Research Agenda* [draft April 1997]. London English Heritage.
- Fieller, N R J, 1993 Archaeostatistics: old statistics in ancient contexts *Journal of the Royal Statistical Society series D* 42, 279-95.
- Gardin, J-C, (ed.) 1970 *Archeologie et Calculateurs*. Paris CNRS.
- Hawkes, J, 1968 The proper study of mankind *Antiquity* 42, 255.
- Heizer, R F, and Cook S F, 1956 Some aspects of the quantitative approach in archaeology *Southwestern Journal of Anthropology* 12, 229-248.
- Heizer, R F, and Cook S F, (eds.) 1960 *The Application of Quantitative Methods in Archaeology* (Viking Fund Publications in Anthropology 28).
- Hodder, I, and Orton C, 1976 *Spatial Analysis in Archaeology*. Cambridge Cambridge University Press.
- Hodson, F R, 1969 Searching for structure within multivariate archaeological data *World Archaeology* 1, 90-105.
- Hodson, F R, 1970 Cluster analysis and archaeology: some new developments and applications *World Archaeology* 1, 299-320.
- Hodson, F R, Kendall D G, and Tautu P, (eds.) 1971 *Mathematics in the Archaeological and Historical Sciences*. Edinburgh Edinburgh University Press.
- Hole, B L, 1980 Sampling in archaeology: a critique *Annual Review of Anthropology* 9, 217-34.
- Kamermans, H, and Fenema K, 1996 Preface in Kamermans H and Fenema K (eds.). *Interfacing the Past Computer Applications and Quantitative Methods in Archaeology CAA95* (Analecta Praehistorica Leidensia 28) xi-xiii.
- Lockyear, K, 1994 Ranters corner -- PUPs *Archaeological Computing Newsletter* 41, 1-2.
- Lockyear, K, 1996 Dmax based cluster analysis and the supply of coinage to Iron Age Dacia in Kamermans H, and Fenema K, (eds.). *Interfacing the Past Computer Applications and Quantitative Methods in Archaeology CAA95* (Analecta Praehistorica Leidensia 28), 165-178.
- Mueller, J W, (ed.) 1975 *Sampling in Archaeology*. Tucson University of Arizona Press.
- Myers, O H, 1950 *Some Applications of Statistics to Archaeology*. Service des antiquities de l Egypte Cairo Government Press.
- Orton, C R, 1982 Computer simulation experiments to assess the performance of measures of quantity of pottery *World Archaeology* 14, 1-20.
- Orton, C R, 1993 How many pots make five? - an historical review of pottery quantification *Archaeometry* 35, 169-184.
- Orton, C R, 1996 Markov models for museums in Kamermans H. and Fenema K (eds.). *Interfacing the Past Computer Applications and Quantitative Methods in Archaeology CAA95* (Analecta Praehistorica Leidensia 28), 149-153.
- Orton, C R, 1997a Testing significance or testing credibility? *Oxford Journal of Archaeology* 16, 219-225.

- Orton, C R, 1997b A bayesian approach to a problem of archaeological site evaluation in Lockyear K. Sly T J T and Mihailescu-Birliba V (eds.). *Computer Applications and Quantitative Methods in Archaeology*. CAA96 (Editura Demiurg Iasi).
- Ramsey, C B, 1995 *OxCal. A radiocarbon calibration and analysis program*. Oxford Research Laboratory for Archaeology University of Oxford.
- Reece, R, 1981 The "Normal" Hoard in Carcassonne C. and Hackens T. (eds.). *Numismatique et Statistique* (Strasbourg PACT 5), 299-308.
- Reece, R, 1988 *My Roman Britain*. Cotswold Studies 3. Cirencester Cotswold Press.
- Robinson W S 1951 A method for chronologically ordering archaeological deposits *American Antiquity* 16 293-301.
- Ross S 1996 Archaeology's new language: multimedia in Higgins A Main P and Lang J (eds.). *Imaging the Past* (British Museum Occasional Paper 114) 286-7.
- Shanks M and C Tilley 1992 *Re-constructing Archaeology* (2nd edn). London Routledge.
- Shennan S 1988 *Quantifying Archaeology*. Edinburgh Edinburgh University Press.
- Spaulding A C 1953 Statistical techniques for the discovery of artifact types *American Antiquity* 18, 305-13
- Spaulding A C 1960 Statistical description and comparison of archaeological assemblages in Heizer R F and Cook S F (eds.) *The Application of Quantitative Methods in Archaeology* (Viking Fund Publications in Anthropology 28) 60-83.
- Treganza A E and Cook S F 1948 The Quantitative Investigation of Aboriginal Sites: Complete Excavation with Physical and Archaeological Analysis of a Single Mound *American Antiquity* 13 207-297.
- Tufte E R 1983 *The Visual Display of Quantitative Information*. Cheshire Conn. Graphics Press.
- Tukey J W 1977 *Exploratory Data Analysis*. Reading MA Addison-Wesley.
- Van Arsdell R 1996 A statistical analysis of Icenian coin hoards *Oxford Journal of Archaeology* 15 235-242.

Contact details

Clive Robert Orton
 UCL Institute of Archaeology
 31-34 Gordon Square
 London WC1H 0PY
 UK
 email: c.orton@ucl.ac.uk