‘To Err is Human’, but to Really Foul Things up You Need a Computer

Clive Orton

University College London Institute of Archaeology,
31–34 Gordon Square, London WC1H 0PY, U.K.
c.orton@ucl.ac.uk

Abstract. The aim of this paper is to focus on one specific point of general interest within the field of quantitative archaeology: the question of errors. Three classes of error are identified and discussed: errors in data, errors in models, and errors in strategy, with comments on how they might be detected or prevented. The overall problem is seen as one of education, and the main question is how best practice can be spread amongst archaeologists who use quantitative methods but do not attend conferences at which such topics are discussed.

1. Statistics and Quantitative Methods

It is simply not possible to summarise the ‘state of the art’ of a topic which is not so much a single coherent subject as a broad approach to archaeological questions, encompassing many different techniques and linked only by their common qualitative idiom.

Having said that, it is nevertheless worth noting that many of the contributions to this session focus on one aspect or another of the subject of classification.This has been an important part of traditional archaeological activity since the nineteenth century, and was in the forefront of the quantified approaches of the 1960s. Since then it has continued to grow in sophistication, but the results have not always lived up to expectations (Aldenderfer 1987: 20). As has happened in other fields (e.g. spatial analysis, see Orton 2004) archaeological practice and the theoretical development of the parent subject seem to have diverged, and a period of convergence, or of inter-disciplinary approaches, may be needed, see e.g. Journal of Classification.

2. What’s not on the Programme

Rather than spread my attention thinly over a wide field, I thought it might be more useful to focus on one apparently small topic that does not feature explicitly on the conference programme. This is the subject of error in archaeology, or, to put it more positively, data quality. Despite well known slogans such as GIGO (garbage in, garbage out), errors do not seem to be taken very seriously in archaeology. It may therefore be valuable to look at where and how they occur, what their effects are, and what might be done about them.

Errors can be found at three levels (at least):

- errors in data,
- errors in models,
- errors in strategy.

2.1 Errors in Data

The first point to accept is that we all make errors; they are a natural part of any human activity. The biggest error of all is to believe that we never make any. Whenever we count, measure or record anything, there is the possibility of error. The next step is to consider how errors arise, which leads to a recognition of different sorts of errors:

Random errors: small fluctuations about a ‘true’ value. They may be due to rounding.

Systematic errors: errors which tend to ‘all go the same way’, e.g. all too high or all too low. They may be due to faulty human perception, or poor calibration of equipment. They lead to the phenomenon of bias.

Gross errors: really large and ‘out of line’ errors. They are often due to mistakes in recording, e.g. a zero omitted or added, or digits entered in the wrong order.

Examples of all three sorts of errors can be found in an exercise that I often give to ‘Statistics for Archaeologists’ classes, to introduce students to these ideas. Each member of a class is asked to measure the lengths and breadths of a sample of 35 hand-axes from the Humbla Collection at University College London Institute of Archaeology. They are typically from 70 to 160 mm long and from 35 to 55 mm broad. When all students have independently measured all the axes, the outcomes are distributed. It is immediately obvious that there is no exact agreement about any of the measurements, which typically have a range of about 6% of the mean length or breadth. The bulk of this variation represents random errors, due to differences in exactly where length and breadth are measured on an axe, problems of parallax, and perhaps even differences between measuring instruments. Standardisation of equipment and instructions would do much to reduce error from this source. Gross errors, which may comprise less than 1% of the data, are easily recognised; they may be due to transposition of digits in a measurement, or transposition of measurements (a length recorded as a breadth, and vice versa). Systematic errors are the least obvious, and may only become apparent when one starts to calculate means for different students, and some appear to have consistently higher readings than others. The most clear-cut cause of bias is the small blank area at the end of a plastic ruler: if it is ignored, all the measurements will be about 5 mm too small. Sometimes the differences may be more psychological in nature, perhaps due to a tendency to round up or round down.

Their Effects. The most obvious and expected effect of errors is to blur patterns in data. For example, random errors added to two sets of measurements (e.g. lengths) will increase their variances, possibly to the point at which any differences between their means become 'statistically insignificant). Similarly, a relationship (e.g. linear) between two measurements on a set of objects may be obscured by errors in one or both of them, to the point where a formal F-test shows no significant relationship.

Less obviously, errors (particularly systematic and gross errors) can create spurious patterns. Systematic differences between two workers may lead to apparent differences between sets of measurements, which may have archaeological interpretations if they have measured different groups of material. For example, if two people measure the rim diameters of two assemblages of pottery, and one 'rounds down' when uncertain while the other 'rounds up', this may create an apparent difference between the means of the two assemblages. Gross errors can, in some circumstances, create spectacularly 'good' but completely spurious patterns, for example in the matching of a set of tree-ring widths to a master curve.

The use of automatic recording equipment may reduce the level of 'human' error (e.g. transposition of digits), but may introduce its own form of systematic error through 'drift', and need frequent recalibration to make sure that this does not occur.

These issues highlight the need for procedures and/or software to detect different sorts of errors. Credibility tests (limits on the acceptable values of a variable), and verification (the comparison of two measurements of the same quantity) may help to pick out gross errors. Randomisation of the tasks allotted to different workers may minimise the effects of any systematic differences between them.

2.2 Errors in Models

This is a rather more subtle class of error. Many statistical techniques in common use are based on a particular model of the data, i.e. a belief that the data behave in a certain way. The most widespread model is that we are dealing with data from a Normal distribution; this has some theoretical justification, since the Central Limit Theorem states that the behaviour of the mean of a sample tends to behave in this way, even if the underlying 'parent' distribution is not Normal (Fletcher and Lock 1991: 67). Nevertheless, if such (often unrecognised) assumptions do not hold, then we are in relatively uncharted territory (see, for example, Baxter 2003: 224–6; Shennan 1988: 101–9). A good example of an assumption-ridden technique is simple least-squares linear regression, which assumes that:

- All the errors are in the dependent variable and none are in the independent variable (and in archaeology, can we tell which is which?).
- The errors are independent, identically distributed normal random variables with zero mean and constant variance (Baxter 2003: 51).

If we are fortunate, our chosen technique may be robust, and we may be able to use it; on other occasions, the technique may be invalidated by the failure of its assumptions (such as the use of the F-test for equality of variances in different samples (Fletcher and Lock 1991: 82)).

How often do archaeologists know the models or assumptions that lie behind the techniques they use?

Implied Needs. It is not enough for archaeologists to know how to apply routine statistical techniques; indeed, it could be argued that modern statistical software makes this task too easy. They also need to know about the assumptions that underlie the techniques, and to be able to assess whether their data are likely to satisfy these assumptions. In other words, they need basic model-building skills, and some knowledge of the robustness (or lack of it) of the more common techniques. Since it is likely that many archaeological datasets may not follow standard distributions or models, a working knowledge of the field of non-parametric statistics (Fletcher and Lock 1991: 74) would also be useful. Such techniques, although sometimes less powerful than the corresponding parametric techniques, can remove the need to make unjustifiable assumptions about datasets.

2.3 Errors in Strategy

The previous class of errors might be thought of as 'tactical' errors; the archaeologist has a reasonable dataset, well-thought-out questions, and some idea of how to go about answering them, but falls at the hurdle of choosing an appropriate statistical technique. Occasionally, we come across situations even these basic foundations have serious flaws, and these we might call 'strategic errors'. Many of us know of such 'horror stories' that we could recount. Examples that come to mind include the use of percentages in situations when counts would be the appropriate form of data, and the inclusion (and analysis) of numerical labels in real data. As above, the ready availability and ease of use of modern statistical software make such howlers very easy to commit. Another possible source of such errors is the 'home-grown' statistical technique, created when well-meaning but misinformed archaeologists believe that archaeology is so different from other disciplines that they need to invent their own techniques (see, for example, Orton 1997). There sometimes seems to be a view that possession of a dataset entitles the originator to do whatever they like with it. The challenge is to instill a necessary level of discipline in quantitative archaeology, without stifling innovation and originality.

3. The Wider World

One of the problems of giving a paper such as this, at a conference like CAA, is that one has a distinct impression of 'preaching to the converted'. What we hear at a conference like this one may be at the 'cutting edge' of our subject, but we have to acknowledge that that is not where most of our colleagues are. Their place has been aptly described by one of my colleagues at UCL as the 'bleeding edge' of the subject (D. Chapman, pers. comm.), where most of the work is done but little of the credit is gained. In out wish to develop new techniques, or to exploit the latest advances in statistics or other application areas, we must not overlook this fact. There is a saying 'the speed of a convoy is that of the slowest ship'; we need to put as much effort into speeding up the slowest as into improving the fastest. But how can we reach them? Having special statistical sessions in more general conferences (e.g. the Institute of Field Archaeologists Conference in Britain) does not seem to work, as we all end up talking to each other again. Perhaps we could discuss a strategy for outreach at a future CAA?

References

Aldenderfer, M. S., 1987. Assessing the Impact of Quantitative Thinking on Archaeological Research: Historical and Evolutionary Insights. In Aldenderfer, M. S. (ed.), *Quantitative Research in Archaeology Progress and Prospects*. Newbury Park, Sage Publications, 9–29.

Baxter, M., 2003. *Statistics in Archaeology*. London, Arnold.

Fletcher, M. and Lock, G. R., 1991. *Digging Numbers*. Oxford, Oxford Committee for Archaeology.

Orton, C. R., 1997. 'Testing significance or testing credulity?' *Oxford Journ Archaeol* 16, 219–225.

Orton, C. R., 2004. Point Pattern Analysis revisited. *Archeologia e Calcolatori* 15, 00–00.

Shennan, S., 1988. *Quantifying Archaeology*. Edinburgh, Edinburgh University Press.