

Never under-estimate the power of a model

Clive Robert Orton

University College London Institute of Archaeology

31-34 Gordon Square

London WC1H 0PY, UK

Phone: +44 (0)20 7679 4749 - Fax: +44 (0)20 7383 2572 - E-mail: c.orton@ucl.ac.uk

Abstract: Statistical models used in archaeology tend to be either implicit and very simple, or complicated, formal, and associated with the use of specialized software. The case is made here for the explicit use of relatively simple models, as a valuable aid to archaeological interpretation. The argument is illustrated by reference to the problem of interpreting the quantities of pottery discarded at a Romano-British pottery production site. The specific question is "how many firing seasons does this pottery represent?" Conventional approaches suggest a range of possible answers that is wider than that given by external evidence (from consumption sites), but the use of successively more refined models can narrow the range to a point where it provided useful interpretative information. The sensitivity of this approach to the initial assumptions is also discussed.

Key words: Statistical models, confidence intervals, archaeological interpretation.

Introduction

The uses of statistical models in archaeology tend to fall into one of two extreme categories. On the one hand, we have the informal, often implicit, role of a model as a basis for the use of a particular statistical test or analytical technique. For example, if we use a t-test to compare two sets of data, we are making implicit assumptions about the Normality of distributions, equality of variances, etc., which together constitute a model of the behaviour of the variables concerned. If we decide (for example) to take logarithms of the data before undertaking an analysis, we are modifying our model to try to bring it more into line with real life. All this takes place so automatically that we scarcely recognize that a model is involved.

At the other extreme, we have the formal models that are needed if we are to employ more specialized statistical techniques. Until recently, these were rare in archaeology, but they are now becoming increasingly common through the growing use of sophisticated analytical packages such as *Oxcal* (Bronk Ramsey 1995) and *Bcal* (Buck *et al* 1999). This increase can be linked to the use of a Bayesian approach to statistical analysis, which requires a more formal model-building stage. Here the archaeologist is forced to be very explicit and detailed about their model of (for example) the stratigraphic relationships between their samples, if they are to benefit from the analytical power that such techniques can bring.

Somewhere between the two lies an area where simple statistical models can be used in the interpretation of relatively simple datasets. The simple 'back of envelope' calculations that archaeologists may use to try to interpret the quantities of material from a site are frequently inconclusive, because the margins of error on their figures are so wide that the final answer is almost meaningless, or could be arrived at by other means. By building a simple but realistic model of the situation, the limits may be reined in to a point at which they may actually become useful. Just such a situation is presented below – the

'Highgate' problem.

Background

The site known as Highgate Wood is a small Romano-British pottery production site, located about 8 km north-west of the Roman city of *Londinium* (Brown and Sheldon 1969). The main phase of production, which is dated to *c.* AD 90 to 140, consists mainly of necked jars, bowls, and beakers in a grey sandy fabric, known as *Highgate Wood C* (Brown and Sheldon 1974). Excavations in the late 1960s and early 1970s revealed six pottery kilns of this phase, together with what felt like an enormous quantity of 'waster' pottery fragments. Post-excavation work showed them to weigh well over a tonne in total.

In recent years discussion has focussed on what these quantities might mean in terms of the organization of pottery production on the site – how much pottery was produced, how it fitted in to the broader picture of *Londinium's* pottery supply, etc. An initial attempt to quantify this problem (Tyers 1997) yielded inconclusive results, but provided the parameters for a more precise analysis.

The 'Highgate' problem

Evidence from *Londinium*, in the form of Highgate Wood pots found in dated contexts, suggests that this phase of production at Highgate Wood lasted for about 50 years (Tyers 1997: 9). On reflection, one tonne or so of wasters is not a lot to show for 50 years' potting, raising questions about the intensity of the use of the site, e.g. was it in use every year? To answer them, we need to estimate the number of firing seasons represented by the waster pottery, and to compare this estimate with the overall life-span of the site.

We approach this by noting that the total number of waster pots must be the product of:
 the number of firing seasons,
 the number of weeks per season,
 the number of firings per week,
 the number of pots per firing, and
 the 'waster rate', i.e. the ratio of pots spoilt to pots fired.
 The number of waster pots can be estimated from the site catalogue, while estimates for the other factors can be obtained from experimental and ethnographic evidence (see below).

Number of waster pots

The sums of the raw values from the catalogue are (Tyers 1997: 2):

weight	1304 kg,
eves	1295 eves,
rim sherds	13,347,

implying a brokenness statistic of just over 10 rim sherds per pot.

The two extreme interpretations of these figures are:

1. each rim sherd represents a single pot; about 90% of each waster pot has either not been recovered or not been recorded. Therefore, about 13,300 pots are represented.
2. each pot represented is substantially complete, but it has not been possible to reconstruct them because of their abraded condition, their dispersal across several contexts and the large quantity of pottery overall. Therefore, about 1300 pots are represented.

Both of these extremes seem unlikely, and the actual number of pots represented on the site must lie somewhere between the two (Tyers 1997: 3).

An alternative approach is through the weight of the vessels, since whatever the problems of measuring the eves of small and abraded rim sherds, all the pottery was weighed and this should be more reliable (Tyers 1997: 5). A regression analysis of the weights and rim diameters of complete or near-complete pots at Highgate (Orton, unpublished) suggests that the average weight of a waster jar is about 650 g, and of a beaker is about 600 g. An overall average of 650 g is estimated, to allow for other forms (such as bowls) which may be slightly heavier. This compares with a figure of about 500 g per eve from sites in *Londinium* (Tyers 1997: 7); it may be that larger pots are more prone to failure in firing than smaller ones. If we accept an average of 650 g per pot, the 1304 kg translates into about 2000 pots. This is again a minimum figure, as it assumes that the pots at Highgate are virtually complete.

An additional complication is that there appears to be under-recording of the Highgate pottery in 'Layer 2' (i.e. from elsewhere than the dumps, kilns and ditches), compared to the non-local pottery (Tyers 1997: 7-8). Tyers (*ibid.*) calculates that adjusting for this would increase the total weight of pottery from the site to about 2000 kg. This would raise the upper and lower limits of the number of pots represented to 3000 and 20,000 respectively. Arguing on the basis that some cross-joins (i.e.

sherds from the same pot in different contexts) can be expected, Tyers (1997: 8) reduces these limits to 2500 and 4500 pots.

Number of weeks per season

Tyers (1997: 9) assumes that the firing season was confined to the period from June to September, giving a maximum of 20 weeks in round figures. He takes a span of four weeks as the minimum.

Number of firings per week

Experimental firings suggest that the firing itself would take the best part of one day and the cooling the best part of another. Taking into account the time taken to load and unload the kiln, Tyers (1997: 9) assumes that a maximum of two firings per week could be achieved. His minimum is one firing per week.

Number of pots per firing

The kilns constructed during the kiln experiments (Anon 1972; 1973) were modelled on those of the Roman potters, and it was estimated that a full load would comprise *c.* 180 to 200 pots. This may be an under-estimate as a temporary superstructure might have allowed a larger load than the permanent dome used in the experiments. Tyers (1997: 9) therefore suggests 250 and 150 as the upper and lower limits on the firing capacity of the kilns.

Waster rate

The first Highgate kiln experiments achieved a waster rate of 78%! The report on the experiment suggests a 20% waster rate, based on the comments of an experienced modern potter and comparison with experiments in firing Roman pottery elsewhere (Anon 1973: 59). The more experienced Roman potters, familiar with their tools and environment, should have achieved a lower figure. Tyers (1997: 8) therefore suggested a lower limit for the waster rate of 5%, and an upper limit of 20%.

The maximum and minimum values for each of these factors are given in Table 1.

Number of firing seasons

The lower limit for the number of seasons is found by dividing the lower limit for the number of pots by the upper limits of all the other factors, i.e.

$$2500 / (20 \times 2 \times 250 \times 0.2) = 1.25 \text{ seasons.}$$

The probable upper limit for the number of seasons is found by dividing the probable upper limit for the number of pots by the lower limits of all the other factors, i.e.

$$4500 / (4 \times 1 \times 150 \times 0.05) = 150 \text{ seasons,}$$

$$\text{and the absolute upper limit would be } 20000 / (4 \times 1 \times 150 \times 0.05) = 667 \text{ seasons.}$$

Since these are, strictly speaking, kiln-seasons, any figure greater than 50 implies that not only did firing take place every year, but more than one kiln was in operation for at least part of the time. We are therefore no nearer the answer to the question of the frequency of use of the kilns.

A statistical approach

Model 1

The initial approach provides us with a simple arithmetical model, namely

$$N = S.W.F.P.R \quad (1)$$

where:

N = number of waster pots,

S = number of firing seasons,

T = number of weeks per season,

F = number of firings per week,

P = number of pots per firing,

R = waster rate.

As we have seen above, applying maximal and minimal values to each variable leads to a very wide range of possibilities for the 'unknown' variable S .

To make further progress, we have to model the variability of the factors concerned, as well as the relationship between them. At its simplest, we could construct a model (model 1) from the assumptions that:

1. each variable has a lognormal distribution,
2. the lower and upper limits given above represent confidence intervals at chosen probability levels (the absolute upper limit for N is not used because it is an absolute, not a confidence, limit),
3. each variable is independent of all the others.

Then equation (1) can be rewritten as

$$\log(S) = \log(N) - \log(T) - \log(F) - \log(P) - \log(R) \quad (2)$$

with

$$\text{var}(\log(S)) = \text{var}(\log(N)) + \text{var}(\log(T)) + \text{var}(\log(F)) + \text{var}(\log(P)) + \text{var}(\log(R)) \quad (3)$$

The expected values of each logged variable, and their variances, can be calculated by assuming that the logged upper and lower limits are the logged means $\pm k$ standard deviations, where k is chosen to give the required probability level. It is possible to choose different probability levels for different variables (perhaps reflecting different levels of confidence in the assigned limits), but for the time being we choose a common level across all variables.

Three levels were chosen: a 'conventional' one of 95% ($k = 1.96$), a 'strict' one of 99% ($k = 2.576$) and a 'relaxed' one of 90% ($k = 1.645$). We also need to specify the confidence limits for S ; here a 'conventional' 95% confidence interval has been chosen. The results are shown in Table 2.

Model 2

Model 1 has been criticized on the grounds that F (number of firings per week) cannot be lognormal, but must have a Bernouilli (either/or) distribution because it can only take the values 1 and 2. This raises an interesting question about the nature of the variability that we are modelling – is it variability in past events (e.g. some weeks there was one firing and some weeks two), or uncertainty in our knowledge (e.g. we know there was likely to be between one and two firing per week, but

cannot be more precise), or both? This criticism also draws attention to problems of modelling the number of firings in a season by multiplying the number of 'active' weeks by the number of firings per week. For example, if work had taken place on an n -day firing cycle (where $n \neq 7$), this would create correlations between successive values of F , which would therefore not be independent.

To overcome this, a more flexible model is suggested (model 2), in which T (weeks per season) and F (firings per week) are replaced by two new variables L (total length of season in days) and M (interval in days between successive firings). The effect is to focus all the chronological variability within a season in a single variable (M). The model thus becomes

$$N = S.(L/M).P.R \quad (4)$$

i.e. $S = N.M/L.P.R$

Using the same assumptions as for model 1, we then obtain $\log(S) = \log(N) + \log(M) - \log(L) - \log(P) - \log(R)$ (5) with

$$\text{var}(\log(S)) = \text{var}(\log(N)) + \text{var}(\log(M)) + \text{var}(\log(L)) + \text{var}(\log(P)) + \text{var}(\log(R)) \quad (6)$$

We next need to consider appropriate limits for L and M . Starting from Tyers' maximum of 20 firing weeks (140 days) per season as the mean calendar length of a season (assuming no time is 'lost'), a range of from 90 to 160 days might be reasonable as a confidence interval for L . For M , Tyers gives an absolute minimum value of 3 days, so a lower bound of 4 days for the confidence interval might be reasonable. The maximum value is more difficult to assess; assuming a maximum of 7 consecutive days' 'down-time' would give an upper bound of 11 days. However, this leads to a mean of about 18 firings per season, compared to only about 13 under model 1. To maintain the same number of firings per season under model 2 would require an upper limit of 22 days (i.e. a maximum 'down-time' between firings of 18 days, which seems rather high). Using bounds of 4 and 11 days for the confidence interval of M leads to table 3; bounds of 4 and 22 days give results very similar to table 2, as might be expected.

Archaeological conclusion

Under model 1, the number of firing seasons is likely to lie between about 4 and 40, depending on the reliance that is placed on the limits assigned to the other variables. This is a reduction on the span of possible outcomes implied by the original analysis (about 25% of the original span of 150 years). Under model 2, with its rather more intensive firing regime, the upper limit of S is reduced to about 20/30 seasons and the span of possible outcomes to about 30% of the original. In either case, it is unlikely that there was a firing every year of the site's existence. This is an important conclusion, because it contradicts the common assumption that 'fixed' kilns are not associated with 'itinerant' potters.

Statistical discussion

Apart from improving our estimate of the 'dependent' variable (in this case, S), this approach also yields useful diagnostic in-

formation about the 'independent' variables. The variances of the logged variables show their 'importance', i.e. the relative contributions that they make to the variability of the dependent variable. The orders of importance (greatest to least) under the two models are:

model 1	model 2
T	R
R	M
F	L)
N	N)
P	P

These show that, if we want to improve our estimate, we should concentrate on reducing the variability of the 'seasonal' factors (*T* and *F* or *M* and *L*), and the waster rate *R*, rather than the 'potting' factors (*N* and *P*). For example, even if we knew the value of *P* exactly, it would make very little reduction in the variability of our estimate of *S*.

We also need to consider the possible effects of mis-specifying the model. This has two aspects: the shapes of the distributions of the variables and the assumption that the variables are uncorrelated.

Distribution of the variables

The calculations of the means and variances from the ranges are based on the assumption that all have lognormal distributions, which was made purely for computational convenience. We therefore need to take a closer look at the variables and assess the likely form of the distribution of each. It seems reasonable to assume that *N* and *P* have normal distributions, but given the sizes of the standard errors (= standard deviation/mean) of these variables, the difference between normal and lognormal will not be great. *R* seems more likely to be genuinely lognormal, reflecting a potter's wish to minimize the waster rate, and the occasional higher values that are likely to occur. The distribution of *F* has already been discussed above. *T* is more difficult; one might suppose that there was a 'natural' firing season, reduced by a random number of inclement weeks, which can be expressed by $T = 22 - T_L$, where T_L has a lognormal distribution. This approach is not followed up here, because of the decision to replace *F* and *T* by *L* and *M* in moving from model 1 to model 2. Since *L* is the overall length of the firing season, a normal distribution seems most appropriate, but for the same reason as for *N* and *P*, a lognormal approximation does not seem unreasonable. *M* seems much more suited to a lognormal distribution, as skewness is likely to be a notable feature of its distribution.

Thus, although much more work could be done on finding the most appropriate distribution for each variable, the use of the lognormal seems reasonable as a first step.

Correlations between variables

The formulae for $\text{var}(\log(S))$ – equations (3) and (6) – carry an implicit assumption that the 'independent' variables are not correlated with each other. The formula if this assumption is not made is of the form

$$\text{var}(A - B - C) = \text{var}(A) + \text{var}(B) + \text{var}(C) - \text{cov}(A, B) - \text{cov}(A, C) + \text{cov}(B, C) \quad (\text{Stuart and Ord 1994: 351}).$$

This means that, under model 1, positive correlations between *N* and any of the other variables will *decrease* $\text{var}(S)$, while positive correlations between the other variables will *increase* $\text{var}(S)$. Since *N* is an estimate made by archaeologists from a fixed body of data, it is not likely to be correlated with other variables. There are no obvious correlations between *T*, *F*, *P* and *R*, except that certain potting strategies (e.g. a need to produce a given number of pots in a season) might suggest a negative correlation between *T* and *F*. This would have the effect of reducing $\text{var}(\log(S))$, but as it is only one correlation amongst many possible ones, the effect is not likely to be great.

A similar argument would hold for model 2, when we should consider possible correlations between *L*, *M*, *P* and *R*. The possible negative correlation between *T* and *F* has disappeared, and there seems to be no reason for replacing it by one between *L* and *M*.

There is thus no reason to suppose that the model is so seriously mis-specified as to cast doubt on the likely values of *S*, and the outcome seems to be robust.

General discussion

The use of a simple stochastic model has here made the difference between an outcome that tells us nothing that we did not already know from other sources of evidence, and an outcome which, although still quite vague, makes a significant contribution to the interpretation of the site. Such a use could be equally valuable in many other interpretative situations. The model can be very easily implemented as a spreadsheet, which allows rapid 'what if' calculations based on different values of the variables. It is a potentially valuable to the archaeologist's toolkit.

Acknowledgements

I am grateful to Paul Tyers for asking me to look at this problem, and for his forbearance for the errors made *en route* to this response to it. Mike Baxter and Ioulia Papageorgiou commented on earlier versions of this paper.

References

- Anon. 1972. The Horniman Museum Kiln Experiment at Highgate Wood – Part 1. In *London Archaeol* 2, no. 1, 12-17.
- Anon. 1973. The Horniman Museum Kiln Experiment at Highgate Wood – Part 2. In *London Archaeol* 2, no. 3, 53-59.
- Bronk Ramsey, C. 1995. Radiocarbon Calibration and Analysis of Stratigraphy: the Oxcal Program. In *Radiocarbon* 37, 425-430.
- Brown, A. E. and Sheldon, H. L. 1969. Early Roman Pottery

Factory in North London. In *London Archaeol* 1, no. 2, 38-44.

Brown, A. E. and Sheldon, H. L. 1974. Highgate Wood: the Pottery and its Production. In *London Archaeol* 2, no. 9, 222-231.

Buck, C. E., Christen, C. E. and James, G. N. 1999. Bcal: an on-line Bayesian radiocarbon calibration tool. In *Internet Archaeol* 7 (<http://intarch.ac.uk/journal/issue7/buck>)

Stuart, A. and Ord, K. J. 1994. *Kendall's Advanced Theory of Statistics Volume 1 Distribution Theory* (6th ed.) London: Edward Arnold.

Tyers, P. 1997. *How much pottery is there at Highgate?* unpublished discussion document.

Tables

	lower limit	probable upper limit	absolute upper limit
pots represented	2500	4500	20,000
weeks per season	4	20	
firings per week	1	2	
pots per firing	150	250	
waster rate (%)	5	20	

Table 1: maximum and minimum values for each of the factors

	Probability level			
	90%	95%	99%	Original limits
Lower limit	3.3	4.2	5.6	1.25
Probable upper limit	56.1	44.7	33.7	150

Table 2: 95% confidence limits on the value of S (number of firing seasons) under various assumptions about the confidence levels implied by the limits assigned to other variables, under model 1.

	Probability level			
	90%	95%	99%	Original limits
Lower limit	3.0	3.6	4.5	1.25
Probable upper limit	31.0	25.6	20.3	73

Table 3: 95% confidence limits on the value of S (number of firing seasons) under various assumptions about the confidence levels implied by the limits assigned to other variables, under model 2.