Cemeteries and significance tests

H. E. M. Cool and M. J. Baxter

Introduction

This note arises in part from E. Swift's paper, "Late-Roman bead necklaces and bracelets," in *JRA* 16 (2003) 336-49. In it, Swift explored various aspects of the use of bead strings as shown by grave finds; and drew attention to the association between amber beads and children within the empire. This was described in terms of percentages of adult/child graves with such beads, and was illustrated graphically by bar charts. The association was not found in graves beyond the frontier.

These are the sort of associations that can be formally tested by the use of statistical significance tests. We feel that the use of such tests, in addition to the simple description of the data, would often be beneficial. They enable the researcher to establish whether the observed pattern is likely to have come about by chance. If it has not, then the pattern is likely to have some archaeological meaning that can be explored further. They are particularly useful in studies of funerary rites where the information provided by the human bones as to the age and sex of the deceased provides a natural framework for studying pyre and grave good associations. Establishing the patterns in these associations allows us to gain insights into the community's attitudes as to what was appropriate for young and old, as well as for males and females.

Here we present a methodology for using significance tests that was developed during post-excavation analysis of a cemetery of the 3rd c. A.D. at Brougham (N England).[1] It involved looking at not only the number of funerary deposits with a particular feature, but also the number without it. With the aid of the tests, we were able to demonstrate that the whole funerary ritual was very strongly influenced by the age and sex of the person buried. The paper will start by taking two simple examples, one from Swift's paper and one from Brougham, to explain the approach and how significance tests work. There follows a more extended technical consideration of the nature of significance tests and when it is appropriate to use them. The two final case-studies show how the methodology can be used with data derived from settlement sites, and in situations where pattern is to be expected. The examples have been chosen to demonstrate how the methodology works. It is not our purpose to discuss in any detail the implications of the patterns observed. Establishing patterns in data is only the first step of any analysis. Once it can be shown that a pattern is most unlikely to have come about by chance, it is the archaeologist's, not the statistician's, task to provide explanations of why this should be. Explanation is likely to come from setting the pattern within the wider archaeological context of the project from which the data were generated.

Amber beads in the *Barbaricum*: the chi-squared test

In Swift's paper, figures are given[2] for the numbers of graves in the *Barbaricum* which have amber beads amongst their grave goods. She concentrates on looking at the cases where amber beads are present, but to fully understand what is happening it is also important to look at the graves which do not have them. When this is done, the data can be expressed as a 2 x 2 contingency table showing the numbers of adult female and child graves both with and without amber beads (Table 1).

	With amber	*Without amber*	*Total*
Female	30	40	70
Child	23	123	146
Total	53	163	216

Table 1. Incidence of amber beads in the *Barbaricum*

Such tables can be explored using a significance test to see if there is any association between the variables, here the presence/absence of amber and the age of the deceased. The null hypothesis is that there is no association between them. In this case this would mean that adult females were no more or less likely to have amber beads in their graves than children. If the hypothesis is rejected at the 5% level, then there is good

1 H. E. M. Cool, *The Roman cemetery at Brougham, Cumbria: excavations 1966-1967* (Britannia Monograph 21, 2004).

2 Swift p. 343, using data from M. Tempelmann-Maczynska, *Perlen im mitteleuropäischen Barbaricum* (Römisch-Germanische Forschungen Bd. 43, 1985).

evidence to contradict the assumption of no association; at the 1% level, there is strong evidence that the assumption is wrong; at the 0.1% level, there is very strong evidence to contradict the assumption of no association. These percentages correspond to what are called "*p*-values" at 0.05, 0.01 and 0.001. The smaller the *p*-value, the stronger the evidence is to contradict the assumption of no association.

The significance test that is most familiar to the archaeologist is the chi-squared test.[3] When Table 1 is subjected to this test, the *p*-value returned is 0.000. This shows that the deposition of amber beads in the *Barbaricum* is very strongly associated with age, with adult females far more likely to have them than children. Swift had concluded from the percentages that "if anything, amber beads are more likely to be found in adult female than in child graves". The use of the chi-squared test would have enabled her to dispense with the "if anything", and make the much more positive assertion that they were regarded as appropriate for women rather than girls. This would have strengthened her argument that the same artefact type can take on different meanings in different situations.

Glass cups at Brougham: Fisher's exact test

The second example is concerned with the pattern of deposition of glass cups as grave goods at Brougham. They were found only in the graves of the adults and, in those graves where the sex of the occupant could be ascertained, they were all associated with males (Table 2). Could this have come about by chance, or were glass drinking cups really the preserve of adult males at Brougham?

Sex	*With*	*Without*	*Total*
Male	6	8	14
Female	0	17	17
Total	6	25	31

Table 2. The occurrence of glass drinking cups in the graves of men and women at Brougham

The validity of the chi-squared test is questionable here. As discussed in more detail in the next section, the chi-squared test involves a comparison between the observed data and the values to be expected if the null hypothesis of no association is true. If the expected values are too small, as can happen when row or column totals are small, the test may be inappropriate. Statistical software will often return a warning message if an expected value is less than 5.

A chi-squared test applied to Table 2 rejects the null hypothesis of no association at the 1% level, but with such a warning. In such cases another type of test, Fisher's exact test,[4] can be used. Details are given in the next section. Applying the test to the data in Table 2 gives a *p*-value of 0.004, significant at better than the 1% level and confirming that the association is unlikely to have arisen by chance. At Brougham, glass drinking vessels were seen as something belonging to the male world.

Fisher's exact test is very useful where sample sizes are small, as can often be the case when comparing the incidence of a particular grave good with individuals of a particular age. Even quite large cemeteries may yield only small numbers of individuals securely sexed and of a particular age. At the Late Roman cemetery at Lankhills (Winchester), for example, 409 individuals have been identified in the re-appraisal of the skeletal collection.[5] If, however, one wished to compare the incidence of an item in the graves of members of the community over 50 years old, the total numbers of individuals would only be 30 (11 male, 19 female).

Chi-squared, Fisher's exact test, and 'significance'

Introductory textbooks on statistics for archaeologists[6] invariably contain an exposition of the use of the chi-squared statistic for testing the null hypothesis of no association in cross-classified data, of which the 2 x 2 tables considered here are an important special case. Fisher's exact test is not discussed in any of these

3 S. Shennan, *Quantifying archaeology* (2nd edn., Edinburgh 1997) 104-26.

4 B. S. Everitt, *The analysis of contingency tables* (2nd edn., London 1992) 14-19.

5 R. Gowland, "Playing dead: implications of mortuary evidence for the social construction of childhood in Roman Britain," in G. Davies *et al.* (edd.), *TRAC 2000* (Oxford 2001) Table 13.1.

6 R. D. Drennan, *Statistics for archaeologists* (New York 1996) 187-92; M. Fletcher and G. R. Lock, *Digging numbers: elementary statistics for archaeologists* (Oxford 1991) 115-20; Shennan (supra n.3) 109-15.

texts, and a 'continuity-corrected' version of chi-squared, which often gives similar results to Fisher's test, is also either not discussed or mentioned only in passing.

The basic calculations underlying the X^2 are very simple, and can if necessary be computed with calculator and paper, though modern statistical and spreadsheet packages such as Excel will calculate it for one. Here we explain the basic concepts. For those readers who prefer to see how the calculations proceed, a worked example is given in the Appendix.

A 2 x 2 table consists of 4 cells with associated row, column and overall totals. Let the observed value in a cell be denoted by O. If there is no association between the variables used in constructing the table, then the expected value (E) is given as

$$E = (\text{row total} \times \text{column total}) / \text{overall total}$$

Under the hypothesis of no association, the chi-squared statistic as usually described in archaeological texts, $X^2 = \Sigma (O - E)^2 / E$, should be 'small' and follow, *approximately*, a chi-squared distribution with one degree of freedom. Any particular value of X^2 can be referred to the distribution and the probability of exceeding it obtained. This is the *p*-value. Large values of X^2 correspond to small *p*-values. If the *p*-value is too small, the hypothesis of no association can be rejected. Conventionally, significance levels such as 5% and 1% are often used. If, for example, the *p*-value is 0.03, the hypothesis of no association would be rejected at the 5% level but not at the 1% level (see above).

A drawback of the chi-squared statistic is that the approximation to the chi-squared distribution may be poor if any of the expected values (E) are too small. The rule of thumb that E must be bigger than 5 is often suggested. In Table 2, the E values for males with cups, and for females with cups, were 2.71 and 3.29; thus the chi-squared test is unsafe.

Under these circumstances an alternative is needed, and Fisher's exact test is one such. It operates in a rather different manner. Table 2, for example, showed one possible distribution of numbers within the body of the table, given the row and column totals. Fisher's test can be thought of as constructing all possible tables that could have arisen from these row and column totals.[7] Table 3 shows an example of another possible table that could be constructed using the same row and column totals as Table 2. It is less extreme than Table 2 in the sense that there are more possible ways of obtaining it than for Table 2. The *p*-value reported for Fisher's test is essentially based on the proportion of possible tables that are as, or more, extreme than that actually observed.

Sex	*With*	*Without*	*Total*
Male	5	9	14
Female	1	16	17
Total	6	25	31

Table 3. A hypothetical table with the same row and column totals as Table 2

Since it is not usually mentioned in statistical texts for archaeologists,[8] the chi-squared test with a continuity correction is also worth noting here. It is obtained by adjusting each value of O by 0.5, the adjustment being made so that it becomes closer to the associated E value. This should make very little difference to the *p*-values when the Es are reasonably large, but it improves the approximation to the chi-squared distribution when the Es are small. The test can also be viewed as a close approximation to Fisher's exact test except when the expected values are very small.[9]

We have attempted here to explain the ideas behind these tests in a fairly non-technical fashion. Apart from illustrating some potentially useful applications of what many will regard as a standard test (chi-squared), we have also drawn attention to a test (Fisher's), valid when chi-squared is not, that we suspect many archaeologists will be less familiar with. Some further aspects, including that of computation, are discussed in the Appendix.

It would, however, be dishonest to leave the reader with the impression that the appropriate use of these tests is a settled issue in the statistical community. For such a simple data structure, discussion of the appro-

7 For a worked example see D. G. Altman, *Practical statistics for medical research* (London 1991) 255.

8 Fletcher and Lock (supra n.6) 118 is an exception.

9 F. Yates, "Tests of significance for 2 x 2 contingency tables (with discussion)," *J. Royal Statistical Soc.* A147 (1984) 427.

priate way to analyse 2 x 2 tables has what has been described as "an extraordinary history".[10] (See Yates [supra n.9] and the discussion of his paper for a flavour of the, sometimes arcane, debate.) One view is that Fisher's test is only strictly valid if the row and column totals are fixed by the 'experimental design', and this is clearly not the case for archaeological data of the kind used here. A more forgiving view is that it is valid to proceed *as if* the row and column totals are fixed, and this is the position we adopt. It might also be noted that the use of the chi-squared statistic (and many other statistical tests) strictly requires assumptions along the lines that the data represent a random sample from some well-defined population, and it is questionable whether much archaeological data ever satisfies such assumptions. Again, one proceeds *as if* the assumptions are reasonable.

A merit of using a formal statistical approach to analysis is that it can guard against reading too much into a data-set, especially if the data-set is small. That is, if *non-significant* results are obtained, then one can be fairly sure there is nothing there to get excited about.

What if statistically significant results are obtained? There are several considerations here. One is that, for very large samples, statistically significant results may have little archaeological significance. This is because, in the context of chi-squared, any association, however small, will be found to be significant with a large enough sample.[11] In Table 1, a significant association between age and the presence of amber beads was detected. One way of looking at this is that the proportion of adult females with beads is 0.43, and the proportion of children with beads is 0.18. This difference, of 0.25, is statistically significant; is it archaeologically interesting? The answer here is probably yes; but would a statistically significant difference of 0.10 or 0.01 attract the same interest? The answer is almost certainly no in the latter case.

A second consideration is that a statistically significant, and possibly interesting, result may arise because of bias in the 'sampling process'. A golden rule in statistical analysis is that if one gets a 'very good' (i.e., highly significant) result that one was not expecting, there is a fair chance that something has gone wrong. There is no simple panacea here, but if careful scrutiny of the data collection procedure reveals no obvious problems, results can be reported with a clear conscience in the spirit that they suggest hypotheses that later study might usefully address.

Exploring identity at Claydon Pike in the Gloucestershire countryside

Thus far, our examples have been taken from data provided by cemeteries. As noted in the introduction, contingency tables are an obvious way of exploring the kind of data recovered from graves; they can, however, also be useful when exploring data from settlement sites. An example will show how this might be done. The material comes from the excavations of a large rural site at Claydon Pike (Glos.).[12] The occupation of the site ran from early in the 1st to the 4th c. A.D. The Roman conquest in this part of England can be placed around A.D. 43/44, so the site was well-established prior to that. Being part of the empire seemed to have had very little impact on the settlement's inhabitants, as building styles and the material culture assemblage show very little change throughout the 1st c. (Phase 2 contexts). In the early to mid-2nd c., however, there was a major re-organisation of the landscape with new boundaries and new styles of buildings. There were also changes in the material culture recovered from the contexts belonging to this re-organised phase (Phase 3).

One intriguing aspect of the changes was that iron hobnails started to be found in Phase-3 contexts, whereas they had been absent in Phase-2 contexts. As this pattern could not be related to any bias in site formation processes, it suggested that the change in the landscape was matched by changes in lifestyle. It has been argued that the adoption of Roman shoe styles, made of the vegetable-tanned leather that was one of the great technological innovations of the Roman period, was not simply the adoption of a new style to fulfil old functions. There was a great range of Roman shoe types to choose from, and C. van Driel-Murray has argued that starting to use them implies the adoption of "a new way of using clothing in social communication".[13] If the inhabitants were adopting nailed shoes only in Phase 3, then there were some very profound changes occurring in this settlement between the phases, but could the pattern have come about by chance?

The method adopted to explore this was to compare the incidence of hobnails with another common dress accessory which was known to have been in use from the start of the occupation. The item chosen was the

10 D. R. Cox and E. J. Snell, *Analysis of binary data* (2nd edn., London 1989) 102.

11 Shennan (supra n.3) 113-15.

12 D. Miles *et al.*, *Iron Age and Roman settlement in the Upper Thames Valley: excavations at Claydon Pike and other sites within the Cotswold Water Park* (Thames Valley Landscapes Monograph, forthcoming).

13 C. van Driel-Murray, "Vindolanda and the dating of Roman footwear," *Britannia* 32 (2001) 185-86.

brooch, as in this part of the country the wearing of brooches was very common from the late Iron Age onwards. The null hypothesis was that there was no association between the phase and the presence of these artefacts. The data are presented in Table 4:

	Phase 2	*Phase 3*	*Total*
Brooch	7	18	25
Hobnails	0	28	28
Total	7	46	53

Table 4. A comparison of the numbers of brooches and hobnails in Phase-2 and Phase-3 contexts at Claydon Pike

Given the expected small values in some cells, the Fisher's exact test is the appropriate one to use. Applied to the table, it gives a p-value of 0.003, strongly suggesting that the null hypothesis should be rejected. The wearing of nailed shoes, and all that they imply, does seem to be a feature of Phase 3.

The incidence of diffuse idiopathic skeletal hyperostosis at Poundbury: a more complex case

In the analyses reported so far, it was the chi-squared or Fisher's exact test that alerted us to the existence of patterns in the data that invited archaeological interpretation. Since such patterns were not expected in advance of data collection and analysis, a significance test based on the null hypothesis of no association was a natural starting point. It is, however, often the case that some form of pattern is expected, so that the null hypothesis of no association is of no interest (it is to be anticipated that it will be rejected). Our final example explores what might be done in such circumstances.

Skeletons can preserve information about the lifestyle of the deceased as certain conditions lead to changes in the bones. An interesting area is the changes brought about by over-eating and over-drinking, as patterns observed here can reflect the community's attitude to food. Did both sexes have equal access to high-living, or can differences be observed between them? Diffuse idiopathic skeletal hyperostosis (DISH) is a condition whose specific causes are uncertain, but which is generally accepted to be associated with obesity and diabetes. It generally affects older individuals. The disease results in much new bone formation and the gradual and complete fusing of the spine.[14] Table 4 shows the incidence of the conditions amongst the sexed skeletons aged over 45 in the 4th-c. cemetery at Poundbury (Dorset).[15]

	Male	*Female*	*Total*
With DISH	11	0	11
Without DISH	84	129	213
Total	95	129	224

Table 4: The incidence of DISH in sexed individuals over the age of 45 at Poundbury

It would be possible to test the null hypothesis of no association here, using any of the tests previously described, and the hypothesis would be rejected. Modern research suggests, however, that the incidence of DISH is more prevalent in males than females. Amongst modern N American populations, for example, the ratio is 65 : 35;[16] thus an association is to be expected, assuming no radical difference between now and the Roman period.

Modern research also suggests that between about 2.4% and 5.4% of adults of the age of 40 have DISH; the figure increases with age to 11.2% for those at age 70 and 28% for those at 80.[17] For illustration, let us take 5%. Assuming equal numbers of males and females in the population leads to the expectation that 6.5% of males and 3.5% of females of age 40 will have DISH.

---

14 C. Roberts and K. Manchester, *The archaeology of disease* (2nd edn., Stroud 1995) 120-21.

15 T. Molleson in D. E. Farwell and T. L. Molleson, *Excavations at Poundbury 1966-80*, vol. II: *the cemeteries* (Dorset Nat. Hist. & Arch. Soc. Monog. 11, 1993) Tables 46 and 60.

16 Figures from the Arthritis Society of Canada, http://www.arthritis.ca/types%20of%20arthritis/dish

17 J. Rotes-Querol, "Clinical manifestations of diffuse idiopathic skeletal hyperostosis (DISH)," *Brit. J. Rheumatology* 35 (1996) 1193-94.

What we have here is a *model* for the data. Under this model, the expected values can be calculated as 6.175 and 88.825 for males with and without DISH, with corresponding values of 4.515 and 124.485 for females. Fisher's exact test is not now available, but a modified version of chi-squared (with or without the continuity correction) is (see Appendix for details).

Using the continuity correction gives $X^2 = 7.16$, with a *p*-value of 0.028 (without the continuity correction, the *p*-value is 0.013). The null hypothesis, which assumes a population incidence of DISH of 5% with the male to female ratio of incidence of 65 : 35, can be rejected at the 3% level. We are primarily interested in whether or not the cemetery evidence can be interpreted to mean that the ratio was different for the population represented by the Poundbury burials, but rejection may be occurring because the 5% assumption is wrong. It is possible to vary this, and also the assumed male to female ratio of incidence. Fig. 1 shows a plot of the *p*-value as the assumed population incidence varies, for assumed ratios of 65 : 35, 70 : 30, and 75 : 25.

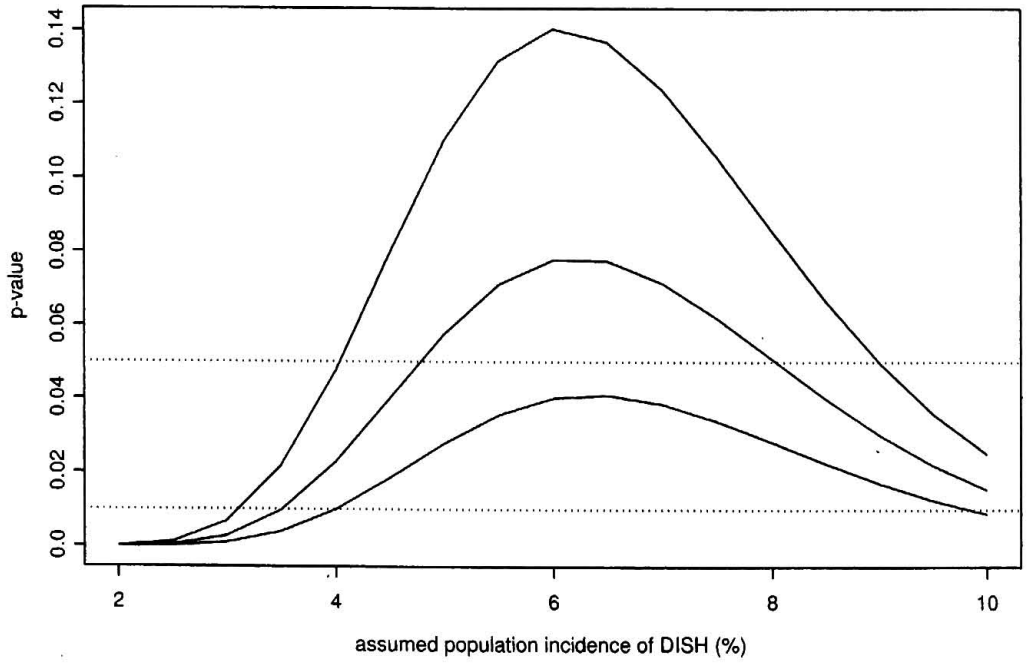

Fig. 1. For different ratios of incidence of DISH for males and females, the figure shows how the *p*-value obtained from a chi-squared test with continuity correction varies, as the assumed population incidence of DISH varies. The lowest curve corresponds to an assumed ratio of 65 : 35, the middle curve to a ratio of 70 : 30, and the upper curve to a ratio of 75 : 25. See the text for explanation and interpretation.

The horizontal dotted lines correspond to the 5% and 1% levels of significance. The graph shows that the assumption of a male to female ratio of 65 : 35 would be rejected at the 5% level for any assumed population incidence. The assumed ratio of 70 : 30 would not be rejected at the 5% level for an assumed population incidence between *c.*4.75 and 8%. The assumed ratio 75 : 25 would not be rejected at the 5% level in the range of *c.*4 to 9%. Taken as a whole, the evidence suggests that, in the population represented in the Poundbury cemetery, DISH was more prevalent among males than is the case in our modern reference population.

As already noted, though the specific causes of DISH are unknown, there does seem to be an association with conditions that can result from over-eating. The evidence would seem to suggest that in the community that buried its dead at Poundbury, a disproportionate number of males had developed the condition. This opens up fascinating avenues to explore about gender relations. If the pattern is correct, was this because males either demanded or were freely given the richest food? Were there foodstuffs or patterns of eating that were considered 'manly'? Or could this pattern be reflecting special circumstances of the Poundbury community? There are good grounds for suspecting that it was Christian,[18] and the relationship between females and fasting within a Christian milieu is well-established, not just for the mediaeval period but also in the

18 Woodward in Farwell and Molleson (supra n.15) 236-37.

early Church.[19] Elite women in the Poundbury community might have chosen frugal diets of their own accord. Comparison with other Late Roman populations from cemeteries where there is no evidence that they belong to Christian communities would be helpful here, but that lies beyond the confines of this paper.

Conclusions

We hope to have shown that significance tests can be a useful weapon in the armoury of the archaeologist, and that different ones are available to address the varying sizes of data-sets. Certainly in an area as data-rich as Roman archaeology, they can be a useful tool. We have often observed that most archaeologists are much more comfortable with simple graphical displays and with figures presented as percentages. These have their place, but we would argue that a better understanding of the data can often be acquired by the application of relatively simple statistical techniques.

As with all patterns, once it has been established that the one observed seems to have meaning, it has to be evaluated taking into consideration other factors. In the case of the hobnails at Claydon Pike, for example, it is possible that the inhabitants of Phase 2 had adopted Roman footwear but of a type that did not employ hobnails. To assess this, the other changes in material culture between Phases 2 and 3 were considered, and it was judged to be unlikely. The introduction of nailed shoes seemed to be matched by the adoption of other changes in the way the inhabitants presented themselves to the world (such as females starting to dress their hair in styles that needed hair pins). Real changes seem to be taking place at that time which, when viewed against a regional background, appear widespread. The use of the contingency tables does not enable us to explain these and other changes, but it does allow the patterns to be examined in a more rigorous way.

Acknowledgements

We are most grateful to Alex Smith of Oxford Archaeology for allowing us to present the Claydon Pike data here in advance of full publication of the site.

Appendix: Computational considerations

Here we give a worked example based on Table 1 to show how the X^2 statistic is calculated. Table 1 was as follows:

	With amber	*Without amber*	*Total*
Female	30	40	70
Child	23	123	146
Total	53	163	216

A 2 x 2 table can be thought of thus:

	Column 1	*Column 2*	
Row 1	O_1	O_2	Row 1 Total
Row 2	O_3	O_4	Row 2 Total
Total	Column 1 Total	Column 2 Total	Overall Total

The values in each cell $O_{1\text{-}4}$ are the observed values. The first step is to calculate the expected values (E) for each cell. This is achieved by applying the following equation:

E = (row total x column total)/overall total.

To calculate the expected value for the cell occupied by O_1, the Row 1 total is multiplied by the Column 1 total and divided by the overall total. In Table 1 the O_1 cell value is 30, the Row 1 total is 70, the Column 1 total is 53. The calculation is thus E = (70 x 53) / 216 which is 3710 ÷ 216, which equals 17.18. For the expected value for the cell occupied by O_2, the Row 1 total would be multiplied by the Column 2 total and divided by the overall total (70 x 163 / 216). The calculation for the cell O_3 would be (53 x 46) / 216 and for O_4 (163 x 146)/ 216.

19 C. M. Counihan, *The anthropology of food and body* (London 1999) 97-98; S. Elm, *'Virgins of God': the making of asceticism in late antiquity* (Oxford 1994) 115.

The expected values for Table 1 would be as follows:

	With amber	*Without amber*	*Total*
Female	17.18	52.82	70
Child	35.82	110.18	146
Total	53	163	216

Once the expected values have been calculated, the X^2 statistic can be computed. The equation $X^2 = \Sigma (O - E)^2/E$ would thus be:

$X^2 = ((30 - 17.18)^2 / 17.18) + ((40 - 52.82)^2 / 52.82) + ((23 - 35.82)^2 / 35.82) + ((123 - 110.18)^2 / 110.18).$

$X^2 = 9.575 + 3.113 + 4.591 + 1.493.$

$X^2 = 18.771.$

Tables giving the X^2 distribution with one degree of freedom can then be consulted; they will, for example, tell one whether the *p*-value for 18.771 is less than 0.05 or less than 0.01. It should be stressed that modern statistical software will do the calculations and provide the exact *p*-value: all the archaeologist has to do is to know how to interpret the *p*-value.

In the tables above, if the row and column totals are fixed, then once any entry in the body of the table is known, the rest can be determined by simple subtraction (e.g., once the expected value for females with amber is calculated as 17.18, that for females without amber is calculated as (70 – 17.18) = 58.82). The one entry for which it is necessary to use the formula given above for E is the 'one degree of freedom' for the test. More generally, for larger tables, the degrees of freedom are $(r - 1)(c - 1)$, where r and c are the number of rows and columns in the table.

Computations in this paper have been carried out using the open-source package R. An introduction to statistics in R is provided by Dalgaard.[20] This includes discussion of the chi-squared and Fisher's test, as well as information on how to obtain and install R.[21]

The **chi-squared test** is a two-sided test. One way of thinking about this is that the null hypothesis of no association is equivalent to the assumption that (using Table 1 as an example) the proportions of females and infants in the population that are buried with beads are the same. If this assumption is wrong and if, in advance of data analysis or inspection, it could be wrong in either direction (i.e., one does not know whether females or infants will have the greater proportion), then a two-sided test should be carried out. This will often be the appropriate course of action.

Fisher's test can be carried out as a one-sided or two-sided test; the default in R, the two-sided test, is that used here.[22] The default chi-squared analysis in R is to apply a continuity correction; in other packages, the uncorrected version may be the default, and the continuity corrected version may not be available. The general message is to read any documentation carefully in advance of applying a test so that one knows what one is getting.

In our first example, the standard form of the chi-squared test for 2 x 2 tables was used, with one degree of freedom. A simple way of thinking about this is that the row and column totals are treated as fixed. Once a single E value has been estimated, other E values are determined by the fact that row and column totals have to be respected. Hence the single degree of freedom.

In our DISH example, a model was proposed for the data. This model postulated specific probabilities of males and females having DISH. The E value for males having DISH is calculated by multiplying the total number of males by the appropriate proportion; and females are similarly treated. The E values for those not having DISH are determined by the fact that column totals (the number of males and females) are respected, but the same is no longer true of row totals. Hence, chi-squared is tested with two degrees of freedom, rather than one.

hilary.cool@btinternet.com Barbican Research Associates, 16 Lady Bay Rd, Nottingham NG2 5BJ
michael.baxter@ntu.ac.uk School of Biomedical & Natural Sciences, Nottingham Trent University

20 P. Dalgaard, *Introductory statistics with R* (New York 2002).

21 http://cran.r-project.org/

22 The *p*-value for a two-sided test can be defined in ways other than that described in the paper. See Yates (supra n.9) or Altman (supra n.7) 255-56 for discussion.