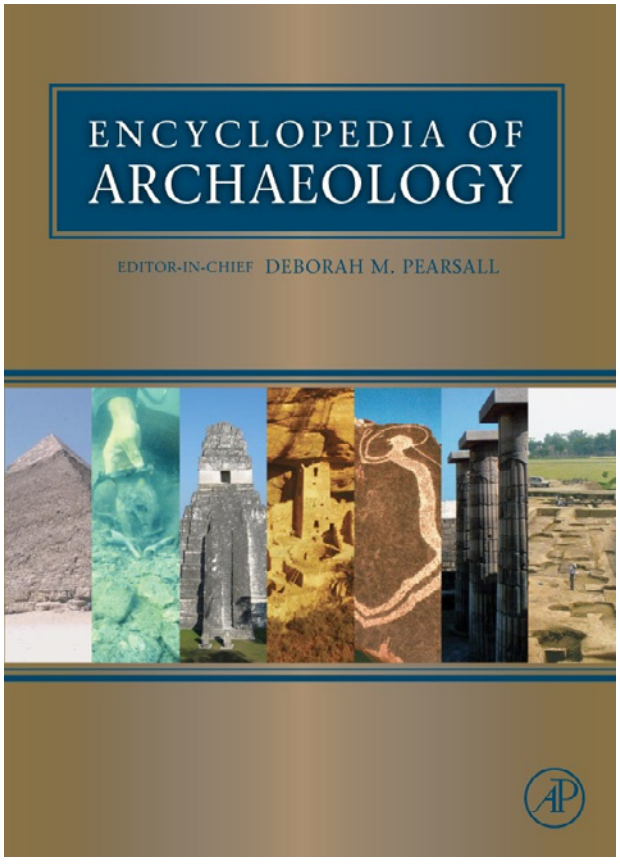

States, Rise of *See:* **Political Complexity, Rise of**.

STATISTICS IN ARCHAEOLOGY

Robert D Drennan, University of Pittsburgh, Pittsburgh, PA, USA

quantitative analysis Analysis of information that comes in the form of numbers, relies heavily on the tools of statistics.

sampling bias Selection of a sample in such a way that some members of the population are less likely to be included than others.

vagaries of sampling The variation that can be expected to occur by pure random chance between different samples selected from the same population.

Archaeological data are irrevocably (although not exclusively) quantitative in nature. The phenomena that archaeologists work with (including artifacts, ecofacts, features, remains of architecture, and many more) are classified, counted, and measured in various ways. The results are numbers, often quite a lot of numbers. Describing things quantitatively, then, along with finding patterns in and comparing numbers, are essential archaeological tasks. Statistical analysis is especially associated with certain schools of thought in archaeology; for example, it was strongly advocated as processual archaeology developed (*see* **Processual Archaeology**). Counting and measuring different kinds of things found in the archaeological record, however, are so fundamental to the process of using material remains to reconstruct past human activities, that statistical analysis cannot be ignored by any school of thought in which it matters to know what people did in the past. The importance of statistical analysis in archaeology is evidenced by the number of books published in recent years with the purpose of introducing and explaining the tools of statistical analysis in a specifically archaeological context. Any of these works can be consulted for further discussion of aspects of statistical analysis touched on here. Traditional statistical tools and terminology developed between about 1920 and 1950 are most often used in archaeology, although the perspectives and vocabulary of the more recent 'exploratory data analysis' school are becoming more widely known.

Description

Categories

The basic tools of descriptive statistics are really quite simple and well known. Classification (especially of artifacts, but also of other things) has long been a quintessential archaeological activity. When things are classified, and how many of them there are in the different categories is determined, the results are one of the two basic kinds of numbers archaeologists work with: counts (as summaries of categorical variables). Counts are often usefully expressed as proportions, or percentages, especially for comparisons – as, for example, when the proportions of different ceramic types recovered from different sites are compared with each other. It is proportions that make it possible to compare two or more sets that contain different numbers of things. Thus, for example, flakes are more strongly represented in an excavation unit where they comprise 25% of the artifacts recovered than in a unit where they comprise only 15%, even if the 25% in the former unit consists of only 10 flakes (out of 40 artifacts) and the 15% in the latter consists of 30 flakes (out of 200 artifacts). This common sense use of percentages, familiar from elementary school, is sometimes referred to as standardizing for different sized collections, or samples, although it is probably better to reserve the term 'standardize' for a more precise and specific statistical use.

Proportions are commonly represented graphically in many different ways, some much more effective than others. Bar graphs are perhaps the simplest and clearest of such graphics. The ease with which commonly available computer programs can be used to draw very complicated forms of bar graphs, often with showy three-dimensional effects, sometimes undermines the utility of this, and other graphic techniques. Figure 1 illustrates good and bad approaches to representing a relatively simple set of percentages graphically for comparative purposes.

Measurements

The other, fundamentally different, kind of number archaeologists often work with (in addition to counts) is measurements, such as length, width, height, thickness, area, weight, etc. Measurements are made in

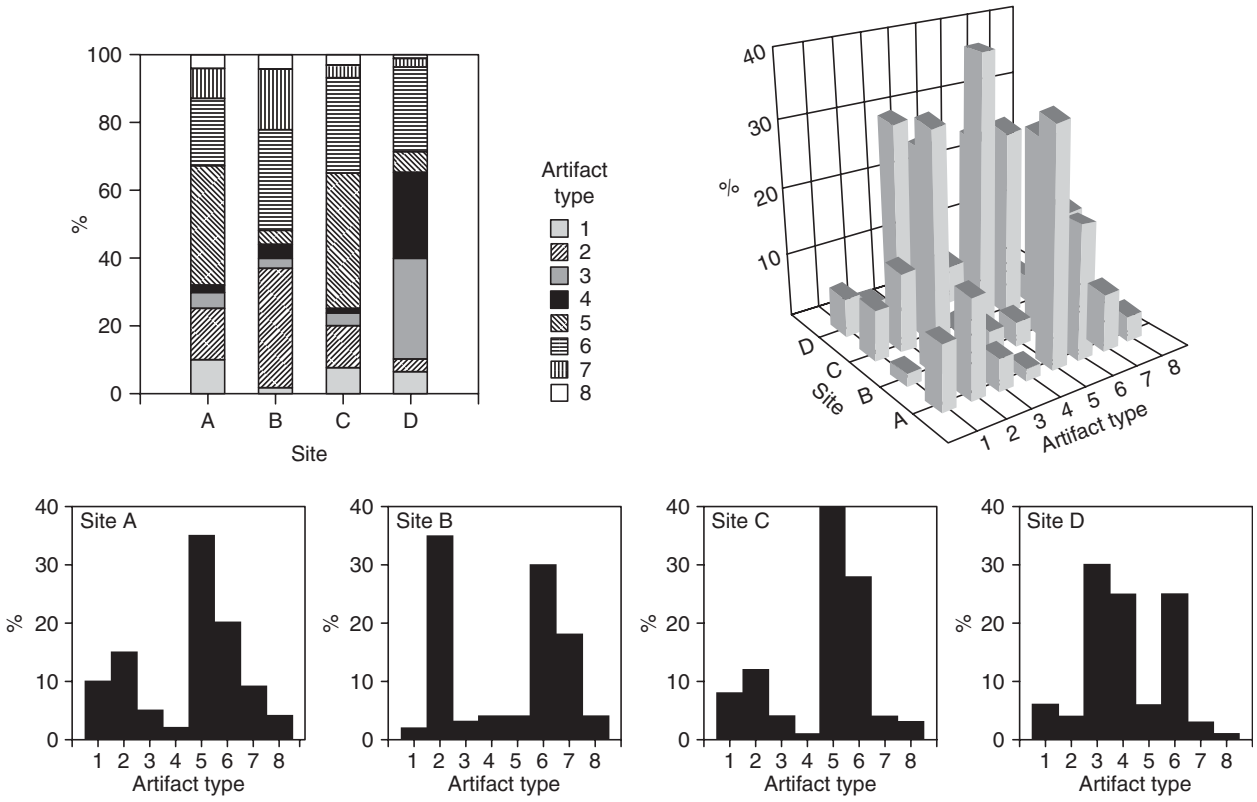

Figure 1 Good and bad ways to illustrate the proportions of artifact types for comparison of the assemblages from different sites. Stacked bars at the upper left are confusing and do not make it easy to recognize patterns. A bar chart in three dimensions at the upper right is almost entirely undecipherable. It can be difficult to resist the temptation to use such glamorous graphics, but simple, flat bar charts like those at the bottom are much more effective presentations of information. In these charts it is easy to recognize that sites A and C have similar assemblages, dominated by artifact types 5 and 6, while the artifact proportions at sites B and D differ sharply, both from sites A and C and from each other.

defined units along scales that are, in principle at least, infinitely subdivisible. Sets of measurements, often called 'batches', have several properties it is important to be aware of when engaging in fundamental description. In the first place, measurements of a single kind of thing tend usually to bunch up around a clear central point along the measurement scale. In what is called a 'normal' shape, there is a single central bunch of numbers with a symmetrical shape or distribution tapering off on either side of the central point when displayed as a stem-and-leaf plot or as a histogram, the two most common ways to represent the shape of a batch of measurements graphically (Figures 2 and 3).

It is useful for basic description, and fundamental for further statistical analysis, to have a single index of the center (or 'measure of central tendency') for a batch of measurements. The mean, or average, which is as familiar as percentages, is most often used. If the batch of measurements has 'outliers', as in Figure 3, the mean may not provide a very sensible index to represent the center. In such cases, the median (essentially the middle number in the batch) may be a more useful index. Similarly, if the batch has a very asymmetrical shape (is 'skewed'), as in Figure 3 the mean may not be a sensible index of its center. The median may, again, be more useful, or, for statistical analyses that depend on the special properties of the mean, the batch can be transformed to make its shape more symmetrical, for example, by using the squares or square roots or logarithms of all the numbers instead of the original measurements. Occasionally, a batch of measurements may show two or more clear bunches of numbers, as in Figure 4. Once again, simply calculating the mean of such a batch produces nonsensical results. The batch must be separated into two or more sub-batches, because the presence of multiple clear bunches of numbers is an unmistakable indication that more than a single kind of thing is represented. Exploration of the shape of a batch of measurements is an essential first step in almost any statistical analysis of it, so as to be sure that a sensible and relevant index of its center can be identified.

Inferences from Samples

Beyond simply exploring and describing the patterns observed in batches of measurements or counts of

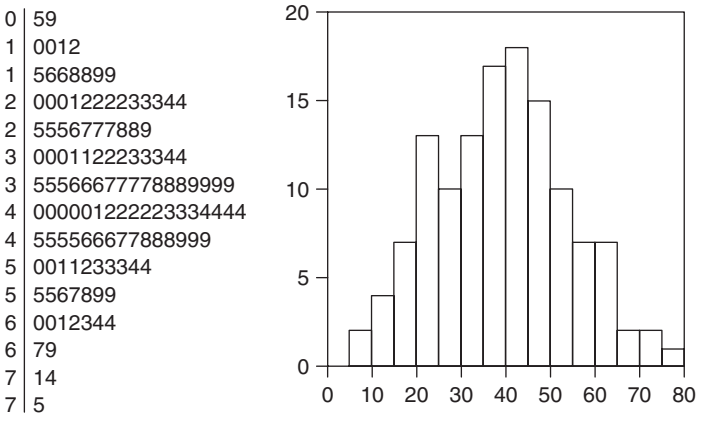

Figure 2 The distribution of a batch of 128 measurements ranging from 5.7 to 75.7, displayed graphically in a stem-and-leaf plot (left) and a histogram (right). The mean of 39.1 provides a sensible index of the center of this batch because it falls at the middle of the main bunch of numbers. The shape is not perfectly symmetrical, and there is a small 'valley' in the 25–30 interval. Such small valleys in a distribution of site sizes have sometimes been taken to indicate two different kinds of sites and thus settlement hierarchy, but such an interpretation is not warranted. Batches of measurements rarely conform perfectly to theoretical distributions, and this example is as close to a normal shape as can be expected in a sample of this size.

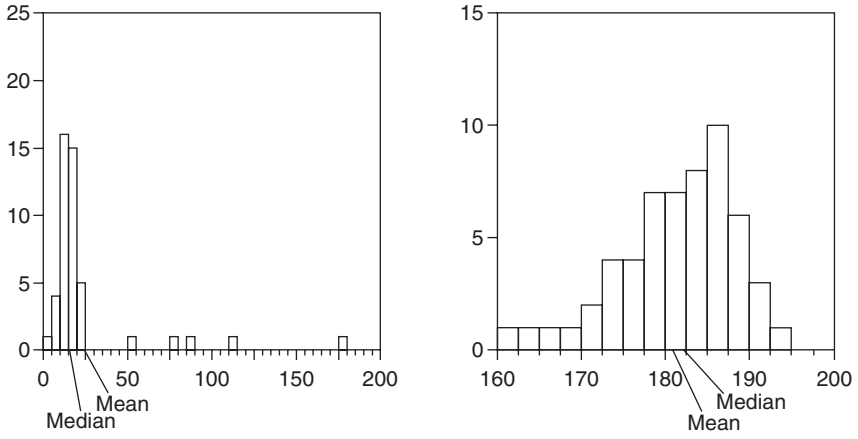

Figure 3 When a batch has outliers (the high numbers scattered far from the other measurements in the histogram at left), the mean (24.1) may be strongly affected and no longer provide a good indication of the center of the main bunch of measurements. In such a case the median (15.4) is a better index. Similarly, when a batch is skewed (as in the histogram at right), the asymmetrical shape produces a mean (181.4) that does not indicate the center of the main bunch of numbers, and the median (182.4) is not much better. A transformation is required (see text).

categories, and perhaps comparing them with other batches, looms the issue of using a batch (or 'sample') to characterize a larger set (or 'population') of things not available for study. Sometimes archaeologists consciously select samples from larger populations, as when they choose some squares in a site grid for excavation or some artifacts for raw material sourcing from those recovered in an excavation. Sometimes the process is less obvious, as when an archaeologist describes, say, Formative period flaked stone tools. In this case, the population is very large and vaguely defined – all the flaked stone tools made during the Formative period in some region – and it is characterized on the basis of those that happen to have been recovered and are thus available for observation. Even if 100% of the flaked stone tools that have been recovered are studied, it is still a question of samples and populations since the description is ordinarily taken to apply to all Formative projectile points. Many archaeologists do not recognize this, and for this and other reasons, the archaeological literature is rife with misunderstanding and misuse of statistical tools.

Sampling Bias

Whenever statements are made about a population on the basis of a sample, there is at least some risk of

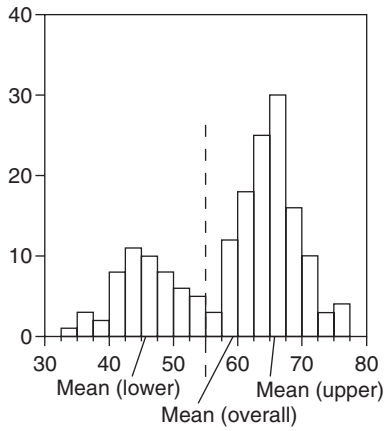

Figure 4 A batch of measurements whose distribution shows, in the form of two clear peaks or bunches of values, that fundamentally two different kinds of things are included. It makes no sense to calculate the overall mean (59.2) of such a batch because it indicates, not the center of anything, but rather the valley between the two peaks. Such a 'bimodal' batch must be split into two separate parts for analysis. When this is done, using a measurement of 55 as a dividing point (dashed line), the shapes of the two resulting batches are single peaked and symmetrical enough that their means (45.7 and 65.3, respectively) provide a good index of each of the two centers. While it might be tempting to separate out a third batch below about 38 and a fourth above about 73, such minor deviations from a normal shape are to be expected as a consequence of random variation and are not meaningful.

error. The tools of inferential statistics cannot eliminate this risk, but they provide powerful ways of assessing it and working with it. Practically every analysis in archaeology (whether statistical or not) involves characterizing a larger set of things than are actually observed, so the perspectives of inferential statistics have implications for archaeological analyses that reach far beyond the quantitative contexts they were designed for. The risk of error in characterizing a population on the basis of a sample arises from two fundamentally different sources. One is that the process of selecting the sample from the population may systematically produce a sample with characteristics different from those of the population at large. This is known as 'sampling bias'. It happens, for example, when lithic debitage is recovered by passing excavated deposits through screens with 6 mm mesh. Lithic debitage is known to include very small waste flakes, many of which will pass through mesh of this size, so the sample recovered from the screen will be systematically larger than the complete population. The mean weight of such a sample of waste flakes would be higher than that of the population as a whole, and any statement made on the basis of this sample about the weight of waste flakes in general would be inflated as a direct consequence of this sampling bias.

Precisely the same is true even in entirely nonquantitative analyses. An archaeologist might subjectively characterize the lithic technology of the Archaic period in some region as reflecting a broad application of a high degree of technical skill. If this characterization of the large, vaguely defined population consisting of all lithic artifacts produced in the region during the Archaic period is based on a sample recovered by artifact collectors who value well-made bifacial tools and never bother to keep anything as mundane as a utilized flake, then the sample is clearly biased toward well-made tools, and the breadth of application of high technical skill will be overvalued as a direct consequence of sampling bias.

Rigorously random procedures for sample selection are designed to avoid bias, and neither of the sampling procedures in the examples above is random. There are no statistical tools for removing bias from a sample once it has been selected, whether by screening deposits through large mesh or by collectors. Indeed, the tools of inferential statistics are often said to 'assume' that samples are unbiased, and thus to be inapplicable to the analysis of biased samples. This is not a useful way to approach statistical analysis in archaeology, because archaeologists are often forced to work with samples they know to be biased. The prescription offered in other disciplines (collect another sample with rigorously random procedures that avoid bias) is often impossible in archaeology. Fortunately, there are at least two common ways of working with biased samples. Archaeologists may want to say things about populations that would not be affected by the bias present in the available sample. It might, for example, be perfectly possible to make unbiased conclusions about the proportions of raw materials represented in lithic debitage on the basis of the screened sample discussed above. It would only be necessary to assume that different raw materials would not occur among the small waste flakes that fell through the screen in proportions very different from those among the larger flakes that would not escape the sample in this way. Biased samples can also often be accurately compared to each other if the same biased operated to the same degree in the selection of all samples to be compared. Thus, two samples of lithics recovered from 6 mm screen may show rather different flake weights. Such a difference cannot be attributed to sampling biases if the sampling biases were the same in both cases, and it is valid to say that the flakes in one sample are heavier than those in the other.

Precisely, the same principles apply to subjective and qualitative comparisons. To compare the application of technical flint-knapping skill to the production of lithic tools in the Archaic with that in the

Formative, it may well be possible to work successfully with biased samples. (This is likely to be necessary in any event.) As long as the same sampling biases operated in the recovery of both Archaic and Formative artifacts, then they can be compared. If, however, the Formative tools come from systematic excavations and the Archaic ones are those accumulated by artifact collectors, then the sampling biases are different and likely to affect very strongly precisely the characteristics of interest. Such a difference in sampling biases affects any comparison based on the abundance of 'nice' well-made tools, whether the comparison is quantitative or not. To repeat, statistical tools cannot eliminate bias once a sample has been selected (and the utter elimination of all kinds of sampling bias from the process of archeological recovery is an unrealistic goal in any event). Statistics, however, does provide two very useful things: first, a reminder of the important effects that sampling bias can have, and, second, some useful concepts for thinking about sampling bias and how serious a worry it is in the specific context of particular observations of potential interest.

Vagaries of Sampling

Sampling bias, then, is one of the two principal sources of error in making conclusions about a population on the basis of a sample. The other is that samples selected entirely without bias still differ from each other and from the population they were selected from because of pure random chance. This is often referred to as the 'vagaries of sampling' and is easily approached by imagining tossing a coin. When a coin is tossed honestly four times, it is a completely unbiased sample of four from the infinitely large population composed of all the times the coin could be tossed. Assuming no prior knowledge at all about the principles of coin tossing, this sample of four could be used to infer the proportions of heads and tails in the large population of all possible coin tosses. The inference made would not always be the same, because while common sense tells us that the proportion of heads in that large population is 50%, it also tells us that in any given sample of four, it might well not turn out exactly that way. Sometimes, in a sample of four coin tosses, the proportion of heads would be 50%, sometimes 25%, or 75%, and sometimes even 100% or 0%. An analyst with a sample of two heads and two tails would conclude that the proportion of heads among coin tosses in general was 50%. An analyst with a sample of one head and three tails, however, would have to conclude that the proportion of heads among coin tosses in general was 25%, and one with a sample of four tails would have to conclude that the proportion of heads among coin tosses in general was 0%. It is easy to see that an analyst making conclusions about a population on the basis of a sample from it will sometimes be right and sometimes wrong. This is true even if sampling bias can be completely ruled out (as it can in this hypothetical example); the erroneous conclusions are the result of pure blind luck – the completely random vagaries of sampling (*see* **Sampling Methods, Theory and Praxis**).

It is the vagaries of sampling that the tools of inferential statistics deal with. They are derived ultimately from a consideration of the range of possible outcomes in selecting a sample of a given size from a given population. In the coin tossing example, the 16 possible (and equally probable) outcomes for a sample of four are easily enumerated: HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT. The 'correct' outcome of a 'representative' sample (one that represents its parent population accurately) is easily seen to occur six times; four times heads are 1/4 or 25%; four times 3/4 or 75%; once 0/4 or 0%, and once 4/4 or 100%. Several of the important principles underlying inferential statistics are easily seen in the samples of four coin tosses. First, an unbiased sample gives an accurate answer more often than it gives any other single answer. Second, it is possible for a completely unbiased sample to represent its parent population inaccurately. Thus, contrary to common belief, random (unbiased) sample selection is no guarantee that a sample is 'representative'. Third, a sample very strongly different from its parent population occurs less often than one fairly similar to the parent population. Thus, when reasoning from samples to populations, serious errors are less common than moderate errors.

A final fundamental principle of inferential statistics is firmly embedded in common sense as well: larger samples are more reliable than smaller samples. The specific implications of this principle for the coin tossing example can be seen by expanding the thought experiment to samples of eight. That is, instead of imagining tossing the coin four times, we imagine tossing it eight times. For eight coin tosses, there are 256 equally probable different outcomes. The most seriously 'erroneous' sample results for estimating the proportion of heads in coin tossing are less common than with a sample of four. Samples with a proportion of heads of 25% or less or 75% or more, for example, represent only 74/256 or 29% of the outcomes, compared to 10/16 or 63% for samples of four coin tosses. Larger samples, then, are more reliable, not because beyond some point they are guaranteed to give an absolutely accurate result, but rather because, as sample size increases, results will be fairly

accurate a steadily growing proportion of the time. The effect of the vagaries of sampling, then, is reduced when sample size is increased. This effect depends entirely on the size of the sample; contrary to widespread opinion; it has nothing to do with the size of the population or the proportion of the population included in the sample.

Estimates, Confidence, and Significance

These principles are often summed up and expressed as statistical confidence. For example, an archeologist recovers a sample of sherds from a site, and 35% of the sherds recovered are from serving bowls. If the sample is unbiased (i.e., if the recovery process would not systematically privilege collecting either bowl sherds or non-bowl sherds), then the best guess that can be made (the one most likely to be near the 'real' population value) is that 35% of the sherds at the site are from bowls. Clearly, however, it is entirely possible that the vagaries of sampling could produce a sample of 35% bowls, even though the proportion in the parent population was different from that figure. Just how much risk there is of error in the estimate of the proportion of bowls in the ceramic assemblage as a whole, can be expressed as an error range for a given confidence level: say, 35% ± 6% at the 95% confidence level. This means that one can be 95% confident that the proportion of bowl sherds in the ceramic assemblage as a whole is between 29% and 41%. Statistical confidence is not just a yes-or-no concept; it is a continuous scale. Choosing to speak at the 95% confidence level means being right 95% of the time (and, inevitably, wrong 5% of the time). Being 95% confident that a ceramic assemblage consists of 35% ± 5% bowls means there is a 5% chance that the ceramic assemblage actually consists of more than 41% bowls or less than 29% bowls. One could speak at a higher confidence level, say 99%, on the basis of the same sample, but only at the cost of less precision, reflected in a larger error range: with the same sample, one can be 99% confident that the proportion of bowl sherds in the assemblage is 35% ± 8%. Conversely, one can speak more precisely on the basis of the same sample, but only at the cost of statistical confidence: one could say that the proportion of bowl sherds in the assemblage is 35% ± 4%, but only with 80% confidence (i.e., a 20% chance of being wrong). Statistical confidence, then, takes the form of a statement (or 'estimate') about a population with an error range for a specified confidence level.

Estimates like these can be compared graphically, as in Figure 5. Here the proportion of bowl sherds at a site, as estimated from the sample discussed above, is compared with the estimate for another site. The error ranges for different confidence levels are represented graphically, showing that either site's estimate falls outside even the 99% confidence level error range for the other. One can thus be more than 99% confident (based on the two samples) that the proportions of bowl sherds in the complete assemblages from the two sites do indeed differ. The 99% confidence level means that there is less than a 1% chance that the statement is actually wrong (i.e., that bowl sherd proportions are actually the same in the two complete site assemblages). Precisely this same issue can be discussed in terms of statistical significance, which is simply the mirror image of statistical confidence. The difference between these two ceramic assemblages could be said to be highly significant ($p < 0.01$), meaning that there is less than a 1% chance that two unbiased samples with such different bowl sherd proportions would be selected from populations that did not differ at all with regard to bowl sherd proportions. In sum, one would have very good reason to explore what the difference in bowl sherd proportions might mean, because it is extremely likely that the difference observed between the two samples does indeed reflect a difference between the two parent populations that are ultimately the objects of interest. That is to say, it is extremely unlikely that the difference between the two samples is attributable to just the vagaries of sampling. The two samples are of adequate size to make it possible to talk with high confidence about differences between the two sites in regard to bowl sherd proportions. If the samples had been smaller, the error ranges would have turned out to be larger for any given confidence level, and differences

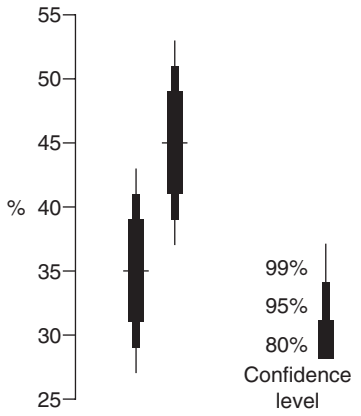

Figure 5 A 'bullet graph' comparing proportions of bowl sherds in samples from two sites. The proportions differ by 10%, and since each proportion falls outside even the 99% confidence level error range for the other, one can have greater than 99% confidence that bowl sherd proportions do differ in two site assemblage populations the samples came from. Expressed in significance terms rather than confidence terms, the difference between bowl sherd proportions at the two sites is very significant ($p < 0.01$).

between the sites would have been identifiable only with less statistical confidence. The differences, in such a case, would be said to be less significant.

It is, unfortunately, a common practice to report results with a significance probability (p) of greater than 0.05, as simply 'not significant' and then ignore them. A p value greater than 0.05 simply indicates more than a 5% chance that the results observed are only a consequence of random processes operating in samples too small to detect the quantitative effect of interest. This is the same as saying one has something less than 95% confidence in the results. Clearly, results with very low significance levels (indicated by very high p values) do not merit much attention. On the other hand, p values modestly greater than 0.05, are decidedly worth knowing. A significance probability of 0.10 (twice the level often taken as the threshold of 'not significant') is equivalent to statistical confidence of 90%. While a higher confidence level would be desirable, results one can be only 90% confident of are definitely worth being aware of. Colloquial speech recognizes this by cautioning that some things must be 'taken with a grain of salt'. Results we are only 90% confident of should, indeed, be taken with a grain of salt, but they should not be ignored; results we are only 80% confident of must be taken with an even larger grain of salt, but they, too, may well be worth attention; and so on. By the same token, it is important to remember that confidence in excess of 95% or even 99% is still not equivalent to certainty. Providing the actual confidence level (70%, 80%, 90%, 95%, 99%, etc.) of results is common practice. Providing the actual significance level ($p = 0.30$, 0.20, 0.10, 0.05, 0.01, etc.) is less common, but makes considerably more sense than just branding results as 'significant' or 'not significant'.

Relationships between Variables

Error ranges for specific confidence levels are attached to estimates made for populations on the basis of samples. These estimates can either be of proportions (corresponding to counts of categories of things, as in the examples above) or of means (if the observation of interest is a measurement). A completely different, but fully equivalent, way of talking about this task is in terms of relationships between variables. Comparing proportions of bowl sherds between two sites is equivalent to studying the relationship between two variables: 'site' and 'vessel form'. For each sherd, there are two pieces of information contained in two categorical variables. The categories for the variable site are A and B; the categories for the variable vessel form are bowl and non-bowl. It is usually easier to think about this information as proportions of bowl sherds at different sites, but the problem can also be formulated as one of the relationship between the variables site and vessel form. If there is a very strong relationship between the two variables, then one site will have a substantially higher proportion of bowl sherds than the other. The bigger the difference in proportions, the stronger the relationship between the two variables is said to be. When the samples involved are large enough to permit talking about the difference between the sites with high statistical confidence, then the relationship between the two variables is highly significant. 'Strength' and 'confidence' (or 'significance') are two related but quite different concepts.

To repeat, in this example, strength has to do with how big the difference between the samples is; confidence (or significance) has to do with whether, given the strength of the difference, the samples are large enough to make it possible to speak of a difference with much confidence that it exists, not just between the samples, but also between the populations the samples come from. If tossing a coin twice turns up two heads, the result (100% heads) is very strongly different from our theoretical expectation of 50% heads. The result is not, however, very significant since it is quite likely that such a sample could be produced by nothing more than the vagaries of sampling. On the other hand, if we toss a coin 2000 times and turn up 75% heads, the result is not as strongly different from our theoretical expectation of 50% heads, but it is considerably more significant because it is very unlikely that just pure random luck would produce such a high percentage of heads in such a large sample.

When examining the relationship between two variables, it is important to consider both the strength and the significance of the relationship. When the two variables are both categories of things which we count, a common approach to evaluating strength and significance is the chi-square test. It yields a significance probability (p), as discussed above, and various measures of strength, including Cramér's V and phi. Cramér's V provides a number on a scale from 0 to 1, where 0 means a relationship of no strength between the variables (i.e., no relationship) and 1 means the strongest possible relationship. In the example of **Figure 5**, where two sites have 35% and 45% bowl sherds, respectively, the value of Cramér's V is 0.1, indicting a strong enough relationship to be meaningful, even though it seems not very different from 0. A Cramér's V of 1 in this example would be produced only if one site had 100% bowl sherds and the other had none. Phi is the same as Cramér's V when there are only two categories for each variable, as in this example: two sites (A and B) and two vessel

forms (bowls and non-bowls). When more than two categories are involved for one variable or the other, Phi becomes an open-ended scale and is much more difficult to interpret in absolute terms.

Different statistical tools are needed if the two variables under consideration include one categorical variable and one measurement, as, for example, comparing utilized flake lengths between two sites. In this case, the two pieces of information for each artifact are which site it came from (two categories) and how long it is (a measurement along a continuous scale). The strength and significance of any difference can be characterized and represented graphically as in Figure 5, except that the vertical scale, instead of proportion of bowl sherds, represents mean flake length. Alternatively, the strength and significance of the relationship between the two variables 'site' and 'flake length' can be evaluated with a *t* test. The *t* test yields a significance probability (p) and a measure of strength (t), on an open-ended scale. When the categorical variable consists of more than two categories, an analysis of variance (ANOVA) can be used to produce the usual significance probability (p) and a measure of strength (F), on an open-ended scale.

If the two variables are both measurements, no categories are involved, and the problem cannot be formulated as a comparison between categories. The most powerful statistical approach is regression analysis, which again produces a significance probability (p) and, as a measure of strength, the correlation coefficient (r). The scale of r ranges from 0 to 1 (in both positive and negative directions), with 0 indicating a relationship of no strength and 1 (or -1) indicating the strongest possible relationship. If the two variables were ceramic vessel wall thickness and rim diameter, an r value of 1 would indicate a perfect positive correlation – the larger the vessel, the thicker the wall. An r value of -1 would indicate a perfect negative correlation – the smaller the vessel the thicker the wall. The square of r is often taken as an indication of the proportion of variation in one variable that is 'explained' by the relationship with the other variable. If the correlation (r) between rim diameter and wall thickness were 0.85, this would be a strong positive correlation, in which 72% ($r^2 = 0.72$) of the variation in wall thickness is accounted for by rim diameter.

See also: **Archaeology Laboratory, Overview**; **Processual Archaeology**; **Sampling Methods, Theory and Praxis**.

Further Reading

Aldenderfer MS (ed.) (1987) *Quantitative Research In Archaeology: Progress and Prospects*. Newbury Park, CA: Sage Publications.

Baxter MJ (1994) *Exploratory Multivariate Analysis in Archaeology*. Edinburgh: Edinburgh University Press.

Cowgill GL (1977) The trouble with significance tests and what we can do about it. *American Antiquity* 42: 350–368.

Drennan RD (1996) *Statistics for Archaeologists: A Commonsense Approach*. New York: Plenum Press.

Fletcher M and Locke GR (1991) *Digging Numbers: Elementary Statistics for Archaeologists*. Oxford: Oxford University Committee for Archaeology.

Shennan S (1997) *Quantifying Archaeology*, 2nd edn. Edinburgh: Edinburgh University Press.

Thomas DH (1986) *Refiguring Anthropology: First Principles of Probability and Statistics*. Prospect Heights, IL: Waveland Press.

Wilkinson L (2000) Cognitive science and graphic design. In: *Systat 10 Graphics*, pp. 1–18. Chicago: SPSS, Inc.

Stone Tools *See:* **Lithics**: Manufacture.

Stratified Societies *See:* **Political Complexity, Rise of**; **Social Inequality, Development of**.

Presentation on theme: "COMPUTERS AND STATISTICS IN ARCHAEOLOGY Week 4. Geographic Information Systems (GIS) - 2 Â© Richard Haddlesey www.medievalarchitecture.net."â€” Presentation transcript: 1.Â Geographical Information Systems in Archaeology: Cambridge Manuals in Archaeology. Cambridge University Press. Cambridge Conolly J, Lake M 2006. Geographical Information Systems in Archaeology: Cambridge Manuals in Archaeology. Cambridge University Press. Statistics in archaeology. Robert D Drennan, University of Pittsburgh, Pittsburgh, PA, USA. Ã£ 2008 Elsevier Inc. All rights reserved.Â The importance of statistical analysis in archaeology is evidenced by the number of books published in recent years with the purpose of introducing and explaining the tools of statistical analysis in a specifically archaeological context. Any of these works can be consulted for further discussion of aspects of statistical analysis touched on here.