2 What Are Data? Archaeological Measurements

This chapter will consider what constitutes archaeological data and how they are measured and presented. Here we will use "measurement" in a very broad sense of the term, including any kind of observation that an archaeologist can make. The description of color, for example, is a measurement and archaeologists often use a standardized scale to measure it. Measurement theory provides a way of describing things and making them comparable to one another. Later parts of the chapter will deal with summarizing and displaying data.

From the initial collection of data through their analysis and publication, there is always a *selection* of data from a theoretically infinite number of possible observations (Coombs, 1964). Although some archaeologists try to argue that you can collect "objectively," as in all sciences, we only see something if we are prepared to see it.

The novel, *Zen and the Art of Motorcycle Maintenance* makes this point quite well:

> According to the doctrine of 'objectivity,' .... We should keep our mind a blank tablet which nature fills for us, and then reason disinterestedly from the facts we observe.
>
> But when we stop and think about it disinterestedly, ... Where *are* those facts? What are we going to observe disinterestedly? ... The right facts, the ones we really need, are not only passive, they are damned *elusive*, and we're not going to just sit back and "observe" them. We're going to have to be in there *looking* for them... (Pirsig, 1974:274-75).

Data are not objects, such as projectile points or potsherds (cf. Sullivan, 1978:189; Thomas, 1976:7). They are observations and **measurements** you make on these objects and on their contexts. In some sense, we do not even merely select facts from an infinite sea of data, but *construct* them by deciding how we will "see" them, and how we will categorize them. Data are theory-laden.

Data in archaeology can include the name or category of an artifact (e.g., "Clovis point"), the density of potsherds on the surface of a site (10 sherds per square meter), the average distance between a site and its nearest contemporary neighbouring site, the invasiveness of retouch on a stone tool, the order of design motives on a pot, or the date of a wooden artifact in years B.C. It can include the spatial provenience of a flake on XYZ coordinates or the stratigraphic context of a sherd (layer 6 in square 4). It can also include practical or methodological information, such as the mesh size of screens used in excavation, the volume of earth excavated, the number of person-hours per hectare of fields surveyed by walking crews (survey intensity), or evidence of such processes as erosion and deposition on a site. All these measures are ones dicatated by the questions archaeologists are trying to answer,

such as, "what is the age of this site?" "who were the people who lived here?" or "what is the social meaning of this design?"

But how do we decide whether or not to measure these things? Some people argue that our perceptive abilities, our interests, our social, economic, and cultural backgrounds, and our unique historical contexts so pervade our choice of questions, and our ways of obtaining answers, that no two archaeologists will ever form the same interpretations of archaeological phenomena. Instead, each archaeologist's interpretation simply makes a target for other archaeologists to reevaluate and perhaps reject. Over time we may come closer to ones that many people, but not all, find broadly acceptable, but the reevaluation process will also encourage the search for new kinds of data that no one considered before. In some cases this will lead to radically new kinds of data and new ways to interpret them.

At the same time, even though we as individuals may not perceive the world in exactly the same way, we as archaeologists operate within a community of archaeologists who can agree about some things at some times. Whether Lewis Binford tells me that a particular pipe stem has a diameter of 3 mm, or Ian Hodder tells me that a particular bone was found in the ditch of a Bronze Age hillfort, I am inclined to accept such information with a rather high level of confidence, or low uncertainty, and with a small margin for error. There are other kinds of reported observations that I am inclined to accept as perfectly reasonable observations, but with somewhat more uncertainty. There are still others I am sure I would have seen differently. I do not have to assume that I would have perceived the evidence in exactly the same way as Dr. Binford or Dr. Hodder to accept them as plausible measurements. This chapter and chapter 4 present some ways we can evaluate these plausibilities.

Some authors suggest an opposition between a "scientific" archaeology that gives primacy to facts and measurements, which are supposed to be neutral or objective (e.g., Shanks, 1992: 26), and other archaeologies in which we suppose facts and measurements to have little or no place. This is a false dichotomy. Even in "hard" sciences, such as physics, data are theory-laden and the kinds of data collected are influenced by the kinds of instruments available. And even in **post-processual** or "interpretive" archaeology, data provide the basis for interpretations, for discerning meaning in "facts" ranging from the order and orientation of incised chevron bands on pottery (Shanks and Tilley, 1987: 160) to the spatial organization of houses and settlements (Hodder, 1990).

Having decided what kinds of "facts" we want, *how* do we measure them? At what level of detail do we stop measuring? We are always sampling and deciding what kinds of data to collect or ignore. We have to make a conscious decision what kinds of data are important for our purposes or those purposes we can reasonably expect our colleagues and successors to have. Whether or not your research strategy specifically outlines the search for particular kinds of data, it is important to stress that all data are "filtered" through the investigator's senses and instruments. We have to "measure" data with our eyes, measuring tapes, calipers, transits, Munsell soil charts, artifact typologies and other instruments.

Measurement consists of *comparison*. You make an observation by comparing an archaeological object or feature with a scale — whether measuring tape, Munsell chart or some classification or typology. Measurement is a process whereby we assign an abstract symbol — a number, colour, name, icon — to represent the object or the value or magnitude of one of its attributes, or qualities. And we can make this assignment by reference to one of several different kinds of scales of measurement.

SCALES OF MEASUREMENT

Measurement scales can be characterized in a number of ways. A common distinction is between *qualitative* scales, with which we can assign observations to categories, and *quantitative* ones, which represent magnitudes. One scale of scales that is quite useful employs four main categories. These are the nominal, ordinal, interval and ratio scales of measurement.

A **nominal scale** consists simply of categories, such as kinds of pottery decoration, that are unordered and of equal weight. The simplest kind of nominal scale is the **dichotomous scale**, with only two categories, such as male and female, or present and absent. Nominal scales are important in archaeology because archaeologists make so much use of classification and typology (see chapter 3), which employ nominal scales. It is possible to count observations in each category of a nominal scale, but it is inappropriate to apply most mathematical operations to such a scale. Consequently the kinds of statistical methods that we can apply to nominal data are limited.

One of the most common types of analysis is to count how many objects we assign to each class in a nominal scale. This is a process called **enumeration**. For example, for a particular archaeological feature, such as a pit, we might count how many sherds are decorated and how many plain, or how many charred seeds are attributable to maize, chenopods, fleshy fruit, and so on. Enumeration is the first step for several other kinds of measures, and particularly for measuring proportions or percentages.

An **ordinal scale** consists of categories that are ordered. Ordinal scales allow you to make the deduction, for example, that if class A is greater than class B and class B is greater than class C, then class A is greater than class C. Archaeologists frequently use ordinal scales to characterize artifacts or sites by their sizes (large, medium and small), for example, but the most common use of ordinal scales in archaeology is in chronology. Archaeological time periods, such as "Archaic," "Middle Woodland," "Neolithic," or "Early Bronze Age," are all categories on ordinal scales. Stratigraphic units also belong to ordinal scales, and archaeological seriation is a method used to create ranked ordinal scales. A rank scale is a special case of an ordinal scale in which each class contains only one member (except in the case of exact ties), and each observation normally has its own class, as opposed to other ordinal scales that have a small number of classes, each containing many observations. An example of the former would be a ranking of all sites in a region from smallest to largest, while an example of the latter would be characterization of all the sites as small, medium or large, or as camps, hamlets, villages and towns.

Some people would consider both nominal and ordinal scales to be, simply, "qualitative" scales of measurement because, even if we assign numbers to represent each class, these numbers are only labels and it is not appropriate to apply most arithmetic functions to them. For example, on a ranked scale, we could say that the site ranked 6 was larger than the one ranked 5, but it would be incorrect to conclude that the site ranked 6 was twice as large as the site ranked 3. Furthermore, it would be wrong to conclude that the difference in size between the sites ranked 6 and 5 was the same as that between sites ranked 3 and 2. Ordinal scales contain no information about the magnitude of such differences in ways that we would normally express in units (e.g., meters, hectares, or degrees). Another characteristic of these scales is that, if they have been constructed properly, no observation could possibly belong to two categories of the same scale simultaneously. This property of being unambiguous is necessary. It would create problems if an artifact were red and black simultaneously, or an artifact belonged to both layer 5 and layer 6. As we will see in chapter 3, which deals with the construction of nominal scales ("classification"), there are unambiguous ways to deal with artifacts that are black over one part of their surface and red over another.

The **interval** and **ratio** scales are ones that many would describe as "quantitative." While interval scales, like ordinal ones, contain an inherent order, there are also consistent intervals between points on the scale. That is to say, it is possible to infer that the distance between 3 and 5 is identical to the distance between 6 and 8 or 7 and 9. This makes the operations of addition and subtraction possible, but is insufficient to allow multiplication or division. Ratio scales contain all the characteristics of the interval scale but, in addition, have zero points that are non-arbitrary and represent an absolute *absence* of some quantity. The Celsius scale of temperature is an interval scale because it is possible to say that the difference between 0° and 10° is the same as that between 30° and 40° but, because

the zero is arbitrary (the freezing point of water) and does not represent an absence of temperature, it makes no sense to say that 40° is twice as hot as 20°. Consequently it is not a ratio scale. The length of a flint blade, by contrast, can be measured on a ratio scale, and we could appropriately say that, within measurement errors, one blade was twice as long as another, so it is appropriate to apply multiplication and division to this kind of measurement.

Measurement Errors and Uncertainty

Nominal and ordinal scales are always discontinuous, or **discrete**, scales because there are theoretically no "grey areas" between categories where you may make observations (as we will see in chapter 3, this can be a problem for typology). Interval and ratio scales, by contrast, frequently are **continuous**. This means that between any two numbers on the scale you can always find other numbers. In fact, you can find an infinity of them. The practical consequence of this property is that you can never have absolute accuracy on a continuous scale. Even if you measure the length of a projectile point to the nearest 0.00001 mm, you have to accept that the "true" length of the projectile point lies somewhere within 0.00001 mm of your measurement, but is not exactly *equal* to your measurement. It is possible, however, to have discrete measures on the interval and ratio scales as well. Integers make a discrete ratio scale that is appropriate for counting objects and other situations where a fractional observation would not make sense. A flint flake might show scars from one, two, or three previous flake removals, for example, but it will never have 2.3 or 1.6 scars because rake is a discrete event (see chapter 8).

Because our senses and instruments are never perfect, measurements always have errors. Furthermore, as we have seen, even a "perfect" instrument could never have absolute accuracy on a continuous scale, because it is always theoretically possible to "magnify" the observation and record it more precisely. Since some degree of error is inevitable, it is important to know the magnitude of error in the measurements. Consequently, in reporting our research, we should always do our best to estimate the size of these errors and report them to our colleagues. If we do not, we make it extremely difficult for others to evaluate the precision and accuracy of our observations or to compare observations. If the measure of the proportion of deer bones at one site is different from that at another, for example, it is impossible for us to tell if the difference is meaningful — are the bone assemblages really different? — unless those reporting on the faunal remains from the sites provide estimates of error in those proportions. Estimating measurement error is not particularly mysterious, and is a basic aspect of scholarly reporting.

A simple way to visualize the problem of measurement errors on continuous scales is to consider the smallest interval you used to distinguish your observations. If you have measured a projectile point with a 12-inch ruler that has no intervals smaller than a quarter-inch, for example, are you justified in reporting that the point is 6-1/16 inches long? Probably only 6 inches ± 1/4 inches is reasonable in this case. Now assume that you had a metric ruler and measured the point to the nearest millimeter. Here you might arrive at 152 mm ± 1 mm. Now if you measured five more projectile points with electronic calipers accurate within 0.01 mm, and wanted to pool your data with the earlier 152 mm one and take an average, would it be reasonable to report a resulting mean of 71.24 mm ± 0.01? Since, in principle, your 152 mm measurement could represent a "true" length anywhere from slightly over 151 mm to slightly less than 153 mm, it would not. It would be best to remeasure the first artifact with the more precise calipers before calculating the mean if you consider precision to 0.01 mm to be important. Alternatively, round the mean off to 71 mm. We will return to the issue of rounding off measurements shortly.

In many of the scales that archaeologists use it is not as simple to conceptualize error. As we will see in chapter 3, assessing measurement error in discrete measurements is a little less obvious, dealing as it does with the problem of **misclassification**. But when the measurement is of something less tangible than artifact length or mass, perhaps of social inequality, of motivation, or of the distribution of gender roles, it is perhaps easier to think of *uncertainty*. In these cases, rather than talk about error in terms of

some hypothetical difference between our measurement and a "true" value, we may simply talk about how confident we are in our measurement and interpretations based on it. Although uncertainty in this sense seems rather subjective, it can still be quite useful (see chapter 4).

Accuracy, Precision, and Reliability

It is important to recognize the difference between Accuracy and Precision.

Accuracy concerns the degree of **bias** in measurement — i.e., systematically recording observations that are higher or lower on an ordinal or interval scale than they should be, or systematically classifying objects in the wrong category of a nominal one. If, for example, you measured some stone tools with calipers that were improperly made, or had been filed down, or measured the size of a feature with a tape that had been stretched, you would not get accurate measurements. Consequently, you might consistently underestimate lengths, and the extent of this underestimate would be the bias. Similarly, tending to make the error of classifying Type A pots as Type B, but rarely or never the reverse, would also result in bias.

Precision concerns the range of results you get if you repeat the measurement several times — the "spread" in a set of repeated measurements. If a measurement is precise, you would expect to get a very closely similar measurement if you or someone else re-measured it. A precise measurement, however, will not necessarily be accurate.

Reliability, meanwhile, is the extent to which a measurement gives the same result in different situations, such as when made by different researchers. It is the proportion of the total variability in the measurement that is due to the actual variability in what we are measuring. The remainder is variability due to such things as inter-observer differences.

For example, when the Black Mesa project (Plog, 1986:42-48; Plog et al., 1978:414) had five crews independently measure site size and ceramic density during an archaeological survey in northeastern Arizona, the crews disagreed substantially in their measurements (table 2.1). Presumably these measurements hovered

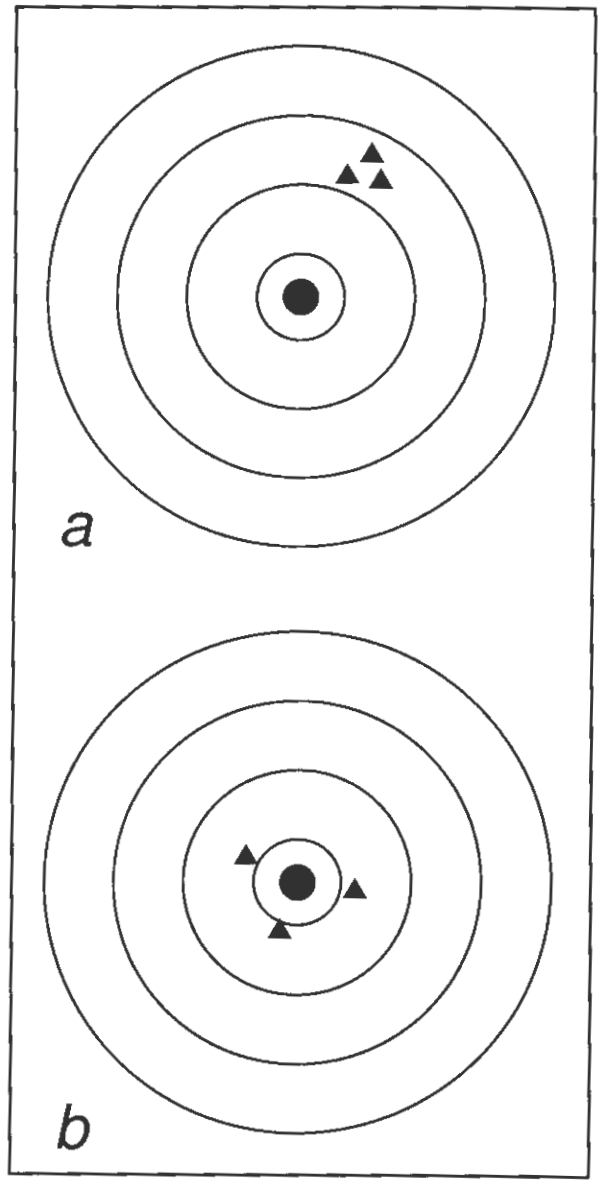

*Figure 2.1. The three triangles in **a** indicate three arrows shot precisely, but not very accurately, while the three in **b** are fairly accurate, but less precise.*

Table 2.1. Variability between crews in estimates of site size and artifact density on Black Mesa, Arizona (after Plog et al., 1978:414)

Crew Number	Mean Site Size (m^2)	Mean Ceramics per m^2	Percentage Site Areas < 100 m^2
1	803	1.2	29.1
2	507	1.7	44.2
3	838	1.6	48.1
4	464	0.9	42.9
Mean	653 ± 195	1.4 ± 0.4	41.1 ± 8.3

around the actual site sizes and densities, and so the average of the measurements would probably be reasonably accurate estimates, but the substantial variation in measurement (here measured by the **standard deviation** after the ±) indicates that they were not very precise or reliable. Below we will consider why this may have happened, but in general it is important to remember that observations on continuous scales are always approximations. Ideally, these are reasonably accurate and precise, as well as appropriate to the questions we are asking (valid).

A useful analogy to help you remember the difference between accuracy and precision is a an archery target (figure 2.1). If you shoot three arrows, and they cluster very tightly, your aim is precise but, unless the arrows hit near the bull's eye, not very accurate. If your shots result in three arrows scattered in the vicinity of the bull's eye, but not very close together, they are reasonably accurate, but not very precise. Of course the ideal is to be both precise and accurate, with all three arrows hitting in or very close to the bull's eye.

When there are several sources of error, they can be **compensating** or **cumulative** errors. If you want to measure the area of a site by overlaying a map with graph paper and counting squares, for example, you can count squares that appear more than half within the site boundaries and exclude ones that appear half out. In this case your errors tend to cancel out, or compensate, and the results are fairly accurate, but not necessarily precise. You can also count only those squares that are completely within the site boundaries. In this case you may get a very precise measurement, but you would underestimate the site size, because the errors would be cumulative, or non-compensating. A common source of this kind of bias occurrs when people doing archaeological survey crudely measure site size by multiplying length times width, which is fine so long as the site is rectangular, but substantially overestimates the size of elliptical and irregularly shaped sites. Again, the errors are non-compensating, leading to substantial bias in the estimate of site size.

If we have been careful and used appropriate measurements, the errors are reasonably small and the observations are accurate enough for our purposes. Sometimes we are forced to use other people's data with unrecorded but possibly large errors, and a useful skill is to be able to assess the reliability of these data.

Significant Digits

This brings us to the issue of significant digits. Many students of archaeology abdicate responsibility for the number of digits to their electronic calculators. Just because your calculator displays a result to 6 decimal places, however, does not mean that it is appropriate to display your results that way.

The problem with such statements as, "the mean blade length is 71.327496 mm" is that it

Table 2.2. Examples of measurements and their significant digits

Measurement	Scientific Notation	Significant Digits
26.01 cm	2.601×10^{1}	4
26010 g	2.601×10^{4}	4
0.0026 kg	2.6×10^{-3}	2
14 artifacts/m^2	1.4×10^{1}	2
14.00 cm	1.400×10^{1}	4
140.0 mm	1.400×10^{2}	4
101 °C	1.01×10^{2}	3
100 dn	1×10^{2}	1

implies a level of precision that is dishonest. In essence, reporting 71.327496 mm constitutes the claim that the "true" mean length has a large probability of being between 71.327495 and 71.327497. In fact the actual standard error on this mean (see below, p. 21) is probably many times larger than ± 0.000001. It also implies that you actually measured artifacts to this level of precision, which seems highly unlikely. If your actual measurements were all made only to the nearest 0.5 mm, then do not report your measurements, or statistical summaries based on them, in such a way as to imply greater precision than that. Simply put, never end up with more significant digits than you started with.

But significant digits are not equivalent to the number of decimal places either. Decimal places are too dependent on the units used to make your measurement. For example, 2.1 km is a measure with two significant digits, meaning essentially that the distance we are trying to represent is somewhere between 2.0 and 2.2 km, but probably close to 2.1. The measure, 5.6 mm also has two significant digits. Both 2.1 km and 5.6 mm are distances measured to one decimal place and two significant digits, but they have substantially different units. If we convert 2.1 km into mm, the resulting measure, 2,100,000 mm, now has no decimal places but it still has two significant digits. The zeroes following the digit, 1, are only there to tell you where the decimal place would go, or the "order of magnitude" of the measurement. Similarly, we could represent 5.6 mm in km. But 0.0000056 km, although it has seven decimal places, still has only two significant digits. Essentially the measure tells you that the actual distance is somewhere between 0.0000055 km and 0.0000057 km. If you find this confusing, or are unsure how to decide how many significant digits a measurement has, convert it into scientific notation. This is the notation that uses powers of 10 to indicate the order of magnitude of the measure. In scientific notation, 2.1 km is 2.1×10^3 m, while 5.6 mm is 5.6×10^{-3} m. Meanwhile, 5.60×10^{-3} m has three significant digits, because the trailing zero is not marking the position of the decimal place, but telling us that the "true" measure is somewhere between 5.59×10^{-3} m and 5.61×10^{-3} m. Scientific notation makes it easy to see how many significant digits there are because everything in front of the multiplication sign is significant. Table 2.1 provides examples that may help to clarify this.

One way to define significant digits is that they consist of *all the certain digits plus the first uncertain one.*

Other Sources of Error

We should not assume that the measuring device, such as a ruler, tape, or caliper, is the only source of error in their measurement and that, consequently, it is appropriate to cite measurements to the smallest increment on that device. This leads to a false sense of accuracy and precision. Suppose, for example, that you are involved in an archaeological survey, like the one that Plog and his colleagues carried out on Black Mesa, and two of your measures are the length and width of each site you encounter. Suppose that you have really well crafted steel tapes and that you can control for the effects of temperature on the steel, so that you can measure even long distances within 1 cm. Would it be appropriate or useful to claim that a site was 93.16 m in length? Here the degree of precision lies, not in the tape, but in our ability to determine consistently where the edge of the site lies. Most archaeological sites have very indistinct edges, such as a gradual diminishment in the density of scattered artifacts, and no two archaeologists are likely to agree on the exact location of a site's boundary or on the orientation of its long axis. Another archaeologist making the same measurement with the same tape might arrive at 106.05 m, or 88.70 m, depending on his or her interpretation of where that boundary lies and what constitutes the long axis, just as the Black Mesa survey teams arrived at different results. Even though the measuring device, the tape, is very precise, there are other factors that would not lead us to expect very precise results.

In cases like this, it is important to ask yourself how much measurement precision is warranted. Will it really alter our intepretation of the site if our measurement of its size is out by 3%, by 5%, or by 10%? Will our ultimate use of the data involve lumping sites into ordinal categories? In most cases, we will find that rounding off such field measurements to only a couple

of significant digits is more honest than implying high degrees of precision and accuracy that are in fact quite meaningless.

You should also be aware that combining measurements that have errors, as when you multiply a length by a width, actually adds a little more error. We will return to this adjustment in the section on descriptive statistics.

Outliers

Sometimes in a set of data we encounter one or a few measurements that are very surprising, so different from the rest of the data that we find it hard to believe that they could be correct. Sometimes it is fairly obvious that they result from human error, such as an error in copying data from a recording sheet into a computer record, or failure to read calipers correctly. When we can plausibly interpret the **outliers** in this way, we may simply omit them from the data set, or remeasure something to check on the surprising readings.

In other cases, however, it is not so obvious that the outlier has simply resulted from human error. As we will see below and in chapters 4 and 6, even in well-behaved data we can expect extreme values to occur from time to time, in which case the surprising values are not really outliers at all. In other cases, the outliers may be the result of **contamination**: they are measurements or observations pertaining to something that does not really belong in the group (or population) we are analysing (Barnett and Lewis, 1994). In archaeology this is quite common. For example, we might be trying to date a particular layer in a Middle Woodland site in Ohio, perhaps by radiocarbon dates on pieces of charcoal (see chapter 15). Perhaps most of the dates that result are fairly close to 1600 B.P., but one date, surprisingly, is 2700 ± 70 B.P. Statistically it is highly improbable that a piece of charcoal from wood that was cut around 1600 B.P. would produce such a date, and the more likely scenario is that the outlier belongs to charcoal that was **residual** in the sediment (see chapter 12). That is, the charcoal had already been on the site for a long time, perhaps the remnant of some ancient forest fire, when Middle Woodland people began to occupy the site and cut some trees for fuel or building material. Of course contamination is usually only noticeable when it shows surprising values, as in this example. If a residual piece of charcoal yielded a radiocarbon date fairly close to the other dates, it would usually be accepted quite happily, because we would have no way of knowing that it did not belong in the data set. Fortunately, in that case, it at least does not seriously affect our interpretations.

Direct and Indirect Measurement

Not only do we need to consider the scale, precision and accuracy of archaeological measures, we need to consider what they really mean. Are we measuring what we are interested in measuring?

Direct measurement involves straightforward measurements such as length, where we can directly compare an object of interest with a standard scale. For example, we can measure the length of a projectile point by reference to a ruler or calipers. A quantity can be measured directly only if it can be measured without measuring some other quantity (Kyburg 1984: 90-112).

Indirect measurement involves measuring one phenomenon as a way of deriving a measure of some other concept. It is an extremely common type of measurement in some fields, such as social anthropology, but crops up frequently in archaeology as well. Even measurements of speed and temperature are indirect, being based on measurements of distance and time in the former and the length of a column of mercury in the latter case (Kyburg, 1984:100, 113-42).

The simplest and most straightforward indirect measures are ratios. For example, we generally do not measure sherd or lithic density on a site directly, instead you usually count the sherds (**enumeration**), and also measure the area where we collected the sherds (usually by measuring directly the lengths and widths of a rectangle or the radius of a circle). We then divide the number sherds by area and create a new measure, "sherds per square meter."

Often ratios make no reference to any units, because they cancel out during division. So the ratio of the frequencies of lithic blades to flakes (e.g., 1:1.4) or the proportion of Deverel-Rimbury pottery in a ceramic assemblage (e.g., 0.29 or 29%) is a unitless indirect measure ("percent" is not a unit, but simply shows that you have multiplied a proportion by 100).

Systematic inaccuracies (biases) in direct measures are compounded by arithmetic operations such as division and multiplication, so that the accuracy and precision in your original measurements are doubly important when you convert them into ratios.

In addition to ratios, there are conceptually indirect measures, which are sometimes called proxy or surrogate measures. The following measurements, all to be found in archaeological literature, are impossible for prehistorians to measure directly, and so they instead measure a proxy that they think might be associated with the phenomenon of interest:

- Number of people who inhabited an Anasazi pueblo
- Wealth of a Bronze Age household
- Social status of the occupant of a grave or tomb
- Degree of interaction between neighbouring settlements
- Volume of traffic in an obsidian exchange system
- Amount of deer meat in a prehistoric diet
- Variation in the magnitude of prehistoric rainfalls
- Conservatism in pottery decoration

To measure prehistoric population sizes, for example, archaeologists have proposed many different indirect measures based on such direct and indirect measures as site area, roofed floor area, number of hearths, number of burials per unit of time, length of longhouse, and even average volume of cooking pots multiplied by number of hearths (Hassan, 1984). Each of these measures is supposed to have some predictable relationship with the site's population size. For example, you would have to account for average life expectancy, whether all or only some people were buried, and the duration of a site's occupation to base population on the number of burials in a cemetery. Alternatively, you might multiply house area by a mean floor area per person in an ethnographic sample. We would expect all these measures to have relatively large, but quite different, sources of error.

Even if we could measure site size very precisely and accurately, for example, estimates of population based on site size would have errors introduced by the following:

- There may be a large degree of variation in the ethnographic examples used as a basis of the estimate
- There may be errors in the selection of ethnographic examples. You may have inappropriate analogues for prehistoric settlement or a biased sample of them
- Possibly not all of the site was occupied simultaneously, leading you to overestimate the ancient population size
- Possibly not all of site was used for domestic settlement. There may have been special industrial areas, gardens, etc., so that you overestimate population size
- Possibly the site we have measured is simply not very typical. It is an **outlier** that does not obey whatever equation we have for calculating population size
- The correlation between site size and the number of people occupying sites is not very strong, or other factors, such as the length of time people anticipate they will occupy the site ("anticipated mobility") may be more strongly correlated (Kent, 1991:39).

In some archaeological jargon, measuring things we cannot observe directly is met with "operational definitions" that simply use easily measurable or observable criteria. For example, some archaeologists who have adopted "decision theory" or "optimal foraging theory" to help them understand the behavior of prehistoric hunter-gatherers, being unable to measure the total "cost" or "benefit" of acquired game, instead estimate the amount of energy used to capture it or contained in its meat. Finding it

equally difficult to measure "taste," they instead measure the amount of fat in the meat. The operational definitions are proxies, different from, but related to, the thing that interests us. Energy is not all there is to the cost and benefits of game, and fat content is only one aspect of taste, but we expect these to be reasonable substitutes.

The extent to which an indirect or proxy measurement is a good approximation of the measurement of interest is its **validity**. We have "face validity" if we simply have widespread agreement among researchers that a measure is valid, not a very strong indication of validity. "Content validity" occurs if the measure appears to contain all the important concepts and behaviors we would expect to find in the phenomenon of interest. This is rather difficult to assess, so content validity is also a somewhat weak measure of validity. "Criterion validity," by contrast, involves comparison of the measure with a standard or testing its ability to predict. For example, we can test a measure of population based on roofed floor area by applying it to ethnographic cases where we already know the number of inhabitants.

Even in cases where the indirect measure may be appropriate or "valid," you should not forget to consider errors intrinsic to it. For example, many people like to use Narroll's (1962) constant of 10 m^2 of roofed floor area per person to estimate the number of people that occupied a house or a settlement. But the relationship between roofed floor area and number of people, even in Narroll's original data, is not very "tight" — the points are scattered widely around the **regression** line — so that the 10 m^2 constant allows us to estimate population size only with a large margin of error (see figure 2.12). Archaeologists who have adopted Narroll's formula or similar ones do not always cite errors for the population estimates that result, although this should be routine.

Uncertainty in estimating the wealth or status of a prehistoric family, or the proportion of its subsistence produced by women's labor, would be even greater. As indirect measures of wealth or status, for example, we might look at the number or quality of grave goods, the size and cost of houses, or osteological evidence for nutrition, none of which, by itself, is a completely valid measure of wealth or status. It is not difficult to find in modern cultures many cases of wealthy people having simple burials, the occupants of mansions falling on hard times, or people whose nutrition and general health in no way reflect their wealth or status. Furthermore, some archaeologists argue that material culture, and particularly burial practice, can often be used as a social strategy to contradict or deny social reality. For example, a very egalitarian ethic in burial practices could be meant to obscure or deny in death extreme wealth differences that existed in life. In spite of such difficulties, however, we make the most we can of indirect measures, combining different ones where possible, because they often help us to understand some of the most interesting aspects of past cultures.

Lest I seem to put indirect measurement in a bad light, it is worth mentioning that there are many examples where indirect measurement is even to be preferred to direct measurement. Quite often we can measure more easily or more precisely with indirect measurement. For example, even though we can measure area directly, perhaps by tiling the area with unit rectangles or triangles and counting them, for many geometric shapes, and especially rectangles, we find it more convenient instead to measure length and width. Even length is measured indirectly when we want to be very precise: we measure temperature to calibrate a steel tape or we measure the time it takes for a laser beam to reflect off a target. Theoretically, we could measure temperature directly, but we can do so much more precisely indirectly, by measuring the length of a mercury column (Kyburg, 1984:141).

Possibly you are wondering why I stress these basic measurement concepts. First, you should never begin research without clearly thinking about how you will measure the variables that are important to your research goals. Later, if you do statistical tests (see chapter 6) and cannot reject the hypothesis that patterns in the data are due only to measurement error, you have a serious problem. Second, too many people do not think about what different researchers' measures really mean when they pool data

for regional synthesis or comparative analysis. They may be "comparing apples with oranges." For example, perhaps project A measured site population size by multiplying the number of pueblo rooms they found by a constant, project B measured site population by multiplying the number of hearths by a constant, and project C measured it by multiplying site area and a constant. If a researcher uses data from all three projects to infer population change over time or differences in population density between regions, it is possible — in fact quite likely — that any differences this researcher sees are simply due to variations in the measurement methods, and not in actual population sizes. Third, too few archaeological publications describe measurement methods in detail or provide estimates of measurement errors. This makes it virtually impossible for us to evaluate the reliability or validity of the results or to tell whether differences in the measurements are statistically or intellectually significant.

A final aspect of data collection that archaeologists need to consider is, how many and what kinds of measurements (data) are relevant to their research problems. This is a matter of research design, and we will return to it in chapter 4.

Descriptive Statistics

Inevitably, archaeologists want to summarize their data, if only because the number of individual artifacts or observations is far too large to show them all in a publication, or even too large for us to recognize any patterns in the data. There are several major ways we can produce such summaries, some of which will be the subject of subsequent chapters. Broadly speaking, we can summarize data verbally, numerically, or graphically.

Verbal summaries are common in archaeological reports and once constituted the bulk of archaeological reporting. In a verbal summary, an archaeologist simply describes "typical" data, such as the most common kinds of pottery in a site, as well as noteworthy anomalies that give the reader some impression of variation in the data. Sometimes verbal summaries can be richer, and usually they are more interesting, than numerical summaries, and they have the advantage that they can convey some of the researcher's thought processes and goals (Hodder, 1989), but they are not amenable to accurate comparison of data sets and it is virtually impossible to apply statistical tests to them. They have an important place in archaeology, but should not be the exclusive means of archaeological reporting.

Numeric summaries of interval or ratio-scale data are what archaeologists usually mean by **descriptive statistics**. These are measures intended to sum up, in only one or two numbers, the "typical" or "central" characteristics of the data, or the amount of "spread" in the data. They include such common numerical summaries as the **mean** or "average," as well as other measures of central tendency, and measures of spread in the data, such as the **range**. We often summarize nominal-scale data with measures of relative abundance such as percentages, but we are more limited in how we can show central tendency or "spread" in nominal data. Although descriptive statistics may seem somewhat tedious, it is necessary to review them here.

Measures of central tendency, indicate the position along a scale about which the data tend to be centred. Measures of dispersion, meanwhile, refer to the amount of variation, or spread, around this location.

The most common measure of central tendency in interval and ratio scales is the average, or **arithmetic mean**. We use this statistic in all kinds of everyday applications, such as when we refer to mean annual rainfall in mm, average income in dollars, or average fuel consumption in miles per gallon. The mean is easy to calculate, as we only have to sum the values of all our observations and then divide by the total number of observations. For a **population** (the totality from which we may have taken a sample), the statistical expression for this simple process is,

$$\mu = \frac{\Sigma X_i}{N}$$

where μ is the population mean (average), ΣX_i is the sum (Σ) of all the data (X_i) values in the population and N is the number of observations or measurements in the population. For sam-

ples that we can expect to have characteristics similar to the population (see chapter 4), a good estimate of μ is the sample mean, $\bar{x}$:

$$\bar{x} = \frac{\Sigma x_i}{n}$$

where Σx_i is the sum of all the data values in the sample, and n is the number of observations in the sample.

The mean turns out to be something like a "centre of gravity" for the distribution. Consequently, the values to either side of it "cancel out" or sum to zero when expressed as differences from the mean. For example, if our data consist of the nine numbers 25, 32, 45, 22, 28, 38, 5, 12 and 18, their mean is 25. The deviations from the mean turn out to be:

25 - 25 = 0
32 - 25 = 7
45 - 25 = 20
22 - 25 = -3
28 - 25 = 3
38 - 25 = 13
5 - 25 = -20
12 - 25 = -13
18 - 25 = -7, $\Sigma = 0$

When our data are not measured at the interval or ratio scales, or when the data are distributed very unevenly, we need to use a different measure of central tendency. The simplest alternative is the **mode**, which is simply the value of the highest peak in a frequency distribution. For data measured at the nominal or ordinal scale, the mode is the most common category, or the category with the greatest number of observations. The **median**, on the other hand, is the value that divides the total number of observations so that half of the observations are greater, and half are less than the median. On a frequency distribution, then, the median is the point on the x-axis that has equal numbers of observations to its left and right. The median is only applicable to ordinal, interval and ratio scales because you cannot have higher or lower values on a nominal scale. When you have an even number of observations, the median is the average of the central two values; when you have an odd number of observations, it is equal to the middle value.

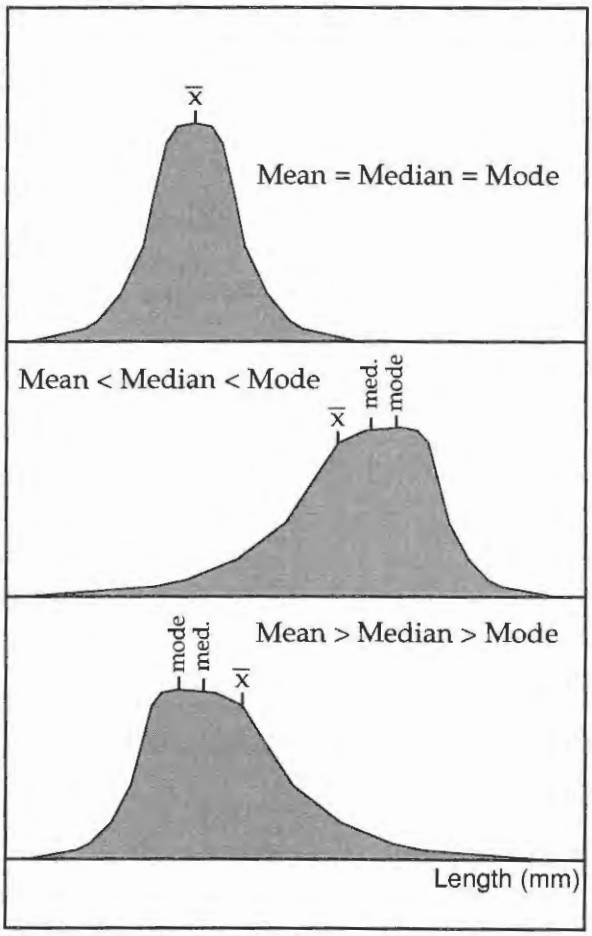

Figure 2.2. The effect of symmetry and skew on measures of central tendency.

Even when you have interval or ratio measurements, the mean is not always the best measure of central tendency, because it is much affected by extreme values and by the symmetry of the distribution of observations. In a perfectly symmetrical distribution peaked in the middle (figure 2.2), the mean, median and mode are all equal. This happens in the **normal distribution** or "bell curve," to which we will return in chapter 6. If, instead, the distribution has some extreme values toward the right, that is, it is skewed to the right or positively skewed, the mean will be greater, sometimes much greater, than the median, which will itself be greater than the mode. If it is negatively skewed, with extreme values to the left, the mean will be less than the median, which in turn is less than the mode. In these cases, since the values in the skewed "tail" of the distribution pull the mean away from the main part of the distribution, the mean no longer

Central Tendency for Grouped Data

Sometimes when we want to measure central tendency, we have to use someone else's data that have been published only in graphic or tabular summaries. We can still calculate a mean (or mean centre, in the case of choropleth maps) quite easily. We take advantage of the fact that we know the frequency of observations that fall within each interval, that the sum of all these observations is equal to n, and that within each interval the average of all the observations should be close to the midpoint of that interval. For the simple histogram in figure 2.3, for example, where x_i = the value of the midpoint of a bar, and f_i = the frequency (height) of each bar:

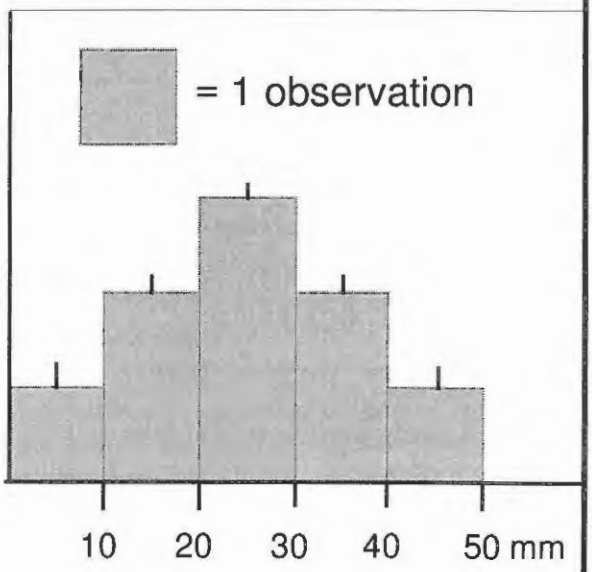

Figure 2.3. A histogram showing lengths to be used (ticks on top of bars) in calculating a mean from grouped data.

$$\bar{x} = \frac{\Sigma x_i f_i}{\Sigma f_i} = \frac{5+(2*15)+(3*25)+(2*35)+45}{1+2+3+2+1} = \frac{225}{9} = 25$$

gives us a very good idea of where most of the data tend to be clumped. Social scientists will typically use median income, rather than mean income, as a measure of central tendency in income distributions because only a few cases of enormous incomes can pull the mean far to the right. In extremely irregular distributions, the mode may be more useful than either mean or median, while some distributions may be bimodal or even multi-modal, requiring us to record more than one mode.

So far we have looked at measures of central tendency in one dimension, but it is also possible to measure central tendency in two or more dimensions simultaneously. In the case of two dimensions, the mean, for example, will be at the coordinates $(\bar{x}, \bar{y})$, where $\bar{x}$ is the mean along the x-axis and $\bar{y}$ the mean along the y-axis. When we find these coordinates on a map, the measure of spatial central tendency that results is called the mean centre. It is even more analagous to a centre of gravity, as it is the location on a map where all the distances away from it sum to zero when divided into positive and negative halves, or the point where a map would balance when all the data points on it have equal weight. The areal median is difficult to define, and can only be approximated with a tedious process of trial and error. The areal mode, meanwhile, is quite easy to determine. Once a map has been divided up into units similar to the bar widths on a histogram (but in two dimensions instead of one), as happens whenever we put a grid on a map, we can make a "stepped statistical surface," the height or colour of each quadrat corresponding, perhaps, to the number of sites or artifacts found in that quadrat (see figure 2.15). The areal mode is the grid quadrat with the highest frequency.

Measuring Dispersion

If we are going to represent a large body of data with a single statistical summary, such as the mean, then we are going to have a rather poor idea of how the data are distributed unless we also include some measure of how they are spread out. This is where measures of dispersion come in.

The simplest measure of dispersion is the **range**, which is simply the difference between the highest and lowest values in the distribution. Because it is based on only two values, and these values are extremes, the range does not give a very good impression of how "clumped" the data are, and so is not very useful except in cases where extreme values are very important. For example, the range in annual rainfall is very important in marginal environments because a large range may make agriculture very risky and unreliable.

A better measure of where the data are clumped is the **interquartile range**. This is the range that includes the middle 50% of the data values. It is found by taking the median and upper and lower quartiles (25th percentile and 75th percentile), that divide the data into four equal numbers of observations (lower 25%, next 25%, next 25% and upper 25%).

One measure of spread in data that may seem obvious would be an average deviation from the mean. But because all the deviations from the mean sum to zero, this doesn't work. We could take the absolute values of all the deviations from the mean (simply removing the minus-signs) to solve this problem, but the result is a measure that turns out to have poor statistical usefulness.

A better way to get rid of the minus-signs is to square the deviations from the mean and sum them. This results in a useful measure of dispersion called **variance**. Algebraically, we usually represent variance as σ^2 (for populations) or s^2 (for samples), and indicate that it is the sum of the squared differences from the mean divided by number of observations or sample size:

$$\sigma^2 = \frac{\Sigma(\mu - X_i)^2}{N} \text{ and } s^2 = \frac{\Sigma(\bar{x} - x_i)^2}{n-1}$$

Note that for samples we divide by n-1 instead of n.

However, the units of variance are not the same as those of the mean. For example, if the mean is measured in meters, the variance would be in square meters. Fortunately, it is easy to compensate for this simply by using the square root of variance. We call the measure of dispersion that results the **standard deviation**, represented as either σ or s (for samples):

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(\mu - X_i)^2}{N}} \text{ and}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\Sigma(\bar{x} - x_i)^2}{n-1}}$$

You may find it useful to think of it as something like an average deviation, with the minus signs removed by squaring, but the standard deviation retains the useful statistical properties of the variance (see chapter 6).

For interval data that we would summarize with a mean, the standard deviation is an extremely useful measure of dispersion. In many applications, however, we will use instead a similar measure that is essentially a standard deviation that takes sample size into account. This measure is known as the **standard error**, and, as we will see later, is extremely important when we want to make predictions on the basis of samples.

The standard error of the mean is simply the ratio of the standard deviation to the square root of the sample size (n). Consequently we can represent it algebraically as

$$SE = \frac{\sigma}{\sqrt{n}}$$

but in samples, since we can only estimate σ, we must estimate SE as,

$$SE = \frac{s}{\sqrt{n}}$$

Finally, it is important to mention that when, as often happens in archaeology, we are interested in densities, such as the number of lithics per square meter, or the number of radiocarbon decays per five minutes, the standard deviation is rather different. We will return to this point in discussion of the Poisson distribution in chapter 6 and radiocarbon dates in chapter 15. For now I will simply mention that in these cases the standard deviation is simply the square root of the mean. To remind you of this difference, I will refer to the mean of counts per unit area or unit time as λ, instead of μ, and to its standard deviation as $\sqrt{\lambda}$.

Accumulating Errors

Now that we have seen some basic measures of dispersion, we can return briefly to the problem of carrying out arithmetic operations on measurements that have errors. We should expect the error on the result of adding, subtracting or multiplying two measurements to be a little larger than the errors on the original measurements, because there is a chance that the two errors will be cumulative. An estimate of the total error on the sum of two measurements is the square root of the sum the squares of the individual errors. For example, for the operation, 10.0 + 5.0 mm, if both measurements had estimated errors of ±0.5 mm, we would estimate the error of the sum as $\sqrt{(0.5^2 + 0.5^2)} = \sqrt{(0.25 + 0.25)} = 0.7$. Consequently, we should report the sum as 15.0 ± 0.7 mm. For multiplication and ratios, we sum the squares of the relative or proportional errors before taking the square root. For an artifact density, for example, we might count 100 ± 10 artifacts in an area of 20 ± 1 m^2. That makes relative errors of 10% and 5%, respectively. The density, then, would be 100/20 or 5.0 artifacts/m^2, with a relative error of $\sqrt{(0.1)^2 + (0.05)^2} = \sqrt{0.01 + 0.0025} \cong 0.112$. Consequently we could report the density as 5.0 ± 0.6 artifacts/m^2. Estimating cumulative errors for more than two measurements simultaneously is somewhat more complicated.

GRAPHIC SUMMARIES OF DATA

> Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations (Tufte, 1983:13, italics in original).

Graphics allow us to display data visually. If used effectively, graphs can help us to communicate complex information easily, in ways that the viewer can interpret accurately. If used inappropriately, however, graphs can also be confusing or downright misleading. There are many kinds of graphs that have archaeological applications and it is important to select the right graph for the kind of data you have and the point you are trying to make. Many of the computer graphing packages typically designed for business applications make it very easy to generate inappropriate and sometimes very misleading graphs. Remember that sales and marketing people have very different goals than archaeologists, and communicating accurately is not necessarily one of them!

Among the criteria that you should use to help you select the correct kind of graph are the following:

- What scale or scales of measurement have you used in your data?
- Are the data continuous or discrete?
- How many dimensions (scales) must you show on a single graph? Most graphs are only appropriate for showing one or two dimensions.
- Who will use the graph? Is it for publication of your final results, or to help you plan your own research strategy for analysis of the data?
- How will viewers want to use the graph? Will they want to be able to extract detailed information from it, perhaps to use in a statistical test, or just compare it with another graph? Will they need to know what proportion of the data lies within a particular range of values, or just get an impression of how the data are distributed?

Tables

Sometimes all you really need to display some data is a simple table. When you are only showing a few values for a small number of attributes, a table can be a very effective way to make a point. The rows and columns in the table should be labelled in such a way that viewers can interpret it easily, and you can always put the values to which you would like to draw particular attention in bold type, or perhaps in color.

One thing you should *not* do, unless you want to lose your viewers, is to present a huge table with a sea of numbers that no one can interpret without considerable time and effort. Large data tables have their place, principally in storing large quantities of information (see chapter 3), but they are not very useful for disseminating complex information clearly and quickly. In these cases it is advisable to substitute a graph

of one of the types described below. Alternatively you could make several small tables with excerpts or summaries of the data that emphasize the points you want to make.

Box-and-Dot Plots and Stem-and-Leaf Plots

For comparing batches of data quickly and without losing detail in the data, a box-and-dot plot or stem-and-leaf plot can be very useful. Both are often used as tools in "exploratory analysis"(Tukey, 1977) because they can help you visualize your data in ways that might help you plan a research design.

Both kinds of plot are used for data along a single dimension, measured on an interval scale.

The stem-and-leaf plot is appropriate if you have a small body of interval data and you want a quick way to tally it manually, with pencil and paper, that does not lose any of the original measurements. The plot begins with a scale ranging from just below the lowest measurement to just above the highest one, consisting of a list of values at ten- or five-unit increments, and omitting the last significant digit (figure 2.4). You then build up the plot by recording the last significant digit of each measurement next to its appropriate interval. For example, in figure 2.4, there are two artifacts with recorded lengths of 127 mm, and those are each indicated by a "7" next to the upper "12" on the plot. There are also two artifacts measuring 128 mm, and those are represented by the two "8"s next to the upper "12." An artifact measuring 124 mm and another two measuring 122 mm are represented by the "4" and two "2"s next to the lower "12." Meanwhile, a large artifact with a recorded length of 207 mm is represented by a "7" next to the "20" at the top of the graph. The result is something like a tally or a histogram, but it preserves all of the original measurements.

A box-and-dot plot summarizes the data more completely, so that you lose the individual measured values but retain a fairly good impression of how the data are distributed (figure 2.5). Now we see only the position of the highest 25% of values, the lowest 25% of values, the upper quartile, median and lower quartile. The box encloses the middle 50% of values (i.e., it is bounded by the upper and lower quartiles), and the line segment subdividing the box marks the median. This kind of plot is useful both in exploratory analysis and for some kinds of archaeological presentations, such as when you want to compare the size distributions of houses from several sites (figure 2.5) and the number of houses from each is small or may not be normally distributed.

Bar Graphs

Archaeologists sometimes confuse bar graphs, intended for discrete (usually nominal) data, with histograms, which are intended for continuous data. Bar graphs can be effective graphic replacements for small tables of data because they visually present the numbers or proportions of observations, or both, in a number of categories. Most computer spreadsheets and statistics packages, as well as dedicated graphing software, will generate a bewildering array of bar graphs, often incorrectly labelled as histograms. You should avoid selecting bar graphs that are too complicated or with too many embellishments that will distract your viewers.

A bar graph shows you how observations are distributed across a number of categories or discrete intervals, so it is a kind of frequency distribution. The bars are separated from each other to signal that the observations are discrete, not continuous. The height of each bar is proportional to both the number and proportion of observations for the category or value that the bar is intended to represent. At least one scale, usually on the left side, allows users to measure the heights of bars to infer the number of observations represented. It is possible to use two scales, however. Where the viewer will find it useful, you may want to have a scale for number of observations on the left and proportion or percentage of observations on the right. Since this may make your graph somewhat "busy," use an extra scale only if it has a purpose.

The y-axis is usually a linear scale, but could be logarithmic, square-root, or something else. Make sure that the label on the scale and title make this clear if it is not a linear scale. Below we will discuss cases where "transforming" the data with a non-linear scale may be useful.

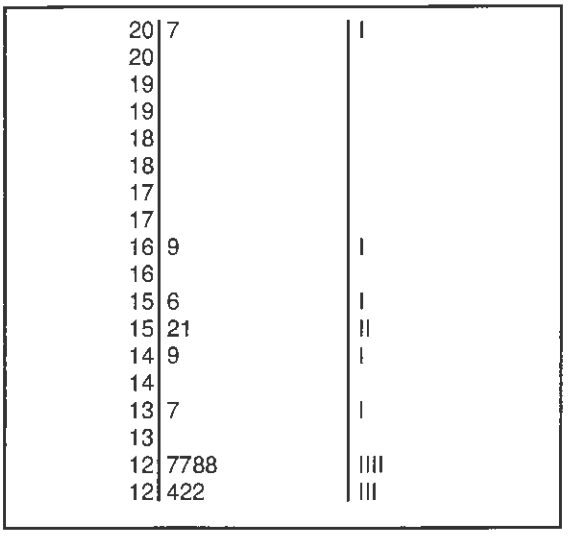

Figure 2.4. Comparison of a stem-and-leaf plot (left) with a simple tally (right). The former includes all the raw data that were measured, while the latter, like a histogram, only tells you how many observations fell within each interval.

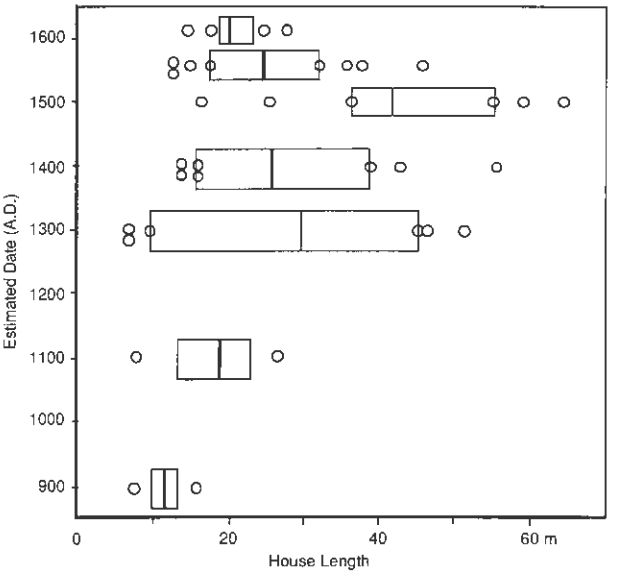

Figure 2.5. Box-and-dot plots of house lengths at seven Iroquoian sites of varying ages (after Warrick, 1996: 18). Note how the chronological relationships of the box-and-dot plots are portrayed.

Sometimes it is tempting to put too many kinds of observations on a single bar graph. This usually only confuses the viewer and makes it difficult to compare data in meaningful ways. Each bar graph should really only show a single dimension. If you need to show several dimensions, it is usually better to draw several bar graphs to the same scale than to put bars with different colours or hatching on a single bar graph. Think carefully about what kinds of comparisons viewers will probably want to make and set up your graph or graphs to facilitate those comparisons (figure 2.6). Most viewers will want to compare bar heights, so put bars that are likely to be compared on the same horizontal axis, rather than offsetting them with a "3-D" look that is popular in business graphics packages. If you have several groups of data to compare in the same graphic, divide them into meaningful sub-graphs and arrange them in ways that facilitate comparison, rather than shuffling all the data into a single bar graph.

One of the more common archaeological variations on the bar graph is the "battleship plot" used in seriation (figure 2.7). This is really a series of bar graphs rotated 90° and reflected about the x-axis so that each bar is actually shown twice. Arguably this reflection is visually appealing, but it is also redundant, and makes the graph take up nearly twice as much space on a page as is necessary. It could easily be omitted (figure 2.8). The rotation is a useful device, however, as it allows us to communicate visually that some categories are older (lower or deeper) than others.

Histograms

Histograms are an appropriate means for displaying the frequency distributions of continuous data in a single dimension on an interval or ratio scale. Although they look much like bar graphs, one of the important characteristics of histograms is that their bars are adjacent to one another to signal to the viewer that the data are continuous, rather than having separate bars (figure 2.9-2.10). Another extremely important characteristic is that the frequencies of observations are indicated by *area* on the graph, and not height of the bars. This is easy to overlook, which can lead to substantial misinterpretation of data.

In a histogram, the horizontal x-axis is a continous interval or ratio scale, and if you were to construct one manually, rather than with a com-

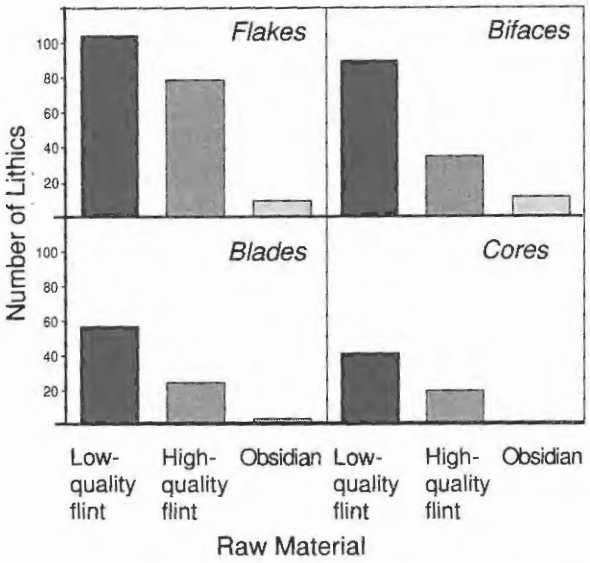

Figure 2.6. Four separate bar graphs are used here to compare the distributions of items measured on nominal scales.

puter, you would have to group your data much as you would to create a bar graph. Because the data are continuous, however, you have to make an arbitrary decision of what interval to use to group your data. The procedure is as follows.

- Select an interval. Generally you will want an interval that is narrow enough to give you an idea of variation in the data, but large enough to create peaks and valleys. You may have to try several different intervals before you get one that presents your data effectively, but you might start with an interval that will give you 10 or 12 bars from the lowest to highest value among your observations.
- Uniquely define the groups. For example, it does not help to say that your intervals will have values of 0-5, 5-10, 10-15, and so on, because some of your recorded observations may lie exactly on the boundaries (at the level of precision you used), creating ambiguity. Instead define your intervals in a way that will separate your data even at the greatest number of significant digits you recorded. For example, you might have intervals of 0-4.9, 5.0-9.9, 10.0-14.9, and so on.
- Tally up the number of observations in each interval by simply counting them.
- Draw the histogram so that the area of each bar is proportional to the number of observations in each interval. Of course the easiest way to do this is to measure the heights and keep the widths of the intervals constant but, as we will see below, it is not always appropriate for the widths to be constant. Make sure that the bars are contiguous, to signal that the data are continuous. Technically the height of the bar is measured in units of observations per interval.
- Draw a rectangle somewhere in a blank area of the graph that has an area proportional to one unit, or ten units, or the like, to indicate the scale of the areas on the graph. Most archaeologists would instead show the scale as increments along the vertical y-axis, but this assumes that the heights of the bars are proportional to the number of observations that they represent, which is not always the case. If you must indicate magnitude on the y-axis, show it in "observations per interval" units to account for the effect of bar width. For example, you might have "number of artifacts per 5 mm interval." You can indicate the scale in numbers of observations, proportions, or both. Tell the viewer how many observations were used to make the graph with an indication like the "n=22" on the graph in figure 2.9.
- As with all graphs, make sure that the scales are labelled, including their units, where appropriate, and give the graph a title. Do not clutter the x-axis with too many labelled increments. Three or four labels at regular intervals are usually sufficient for the reader to appreciate the scale of measurement.

Because histograms express frequency by area, you will probably get into trouble if you try to use a non-linear scale for any of your measurements. If you want to use transformations (see below), you should express them on a histogram with transformed units, so that the proportionality of area will be preserved.

As I have already hinted, there are cases when it does not make sense for the intervals

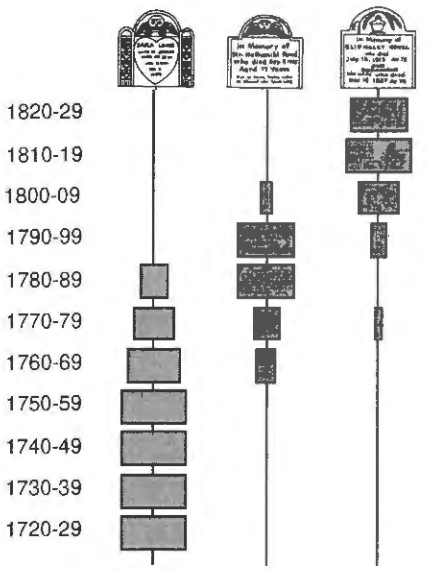

Figure 2.7. Vertical bar graphs with symmetrical bars used in a seriation of headstones from New England (after Deetz, 1967).

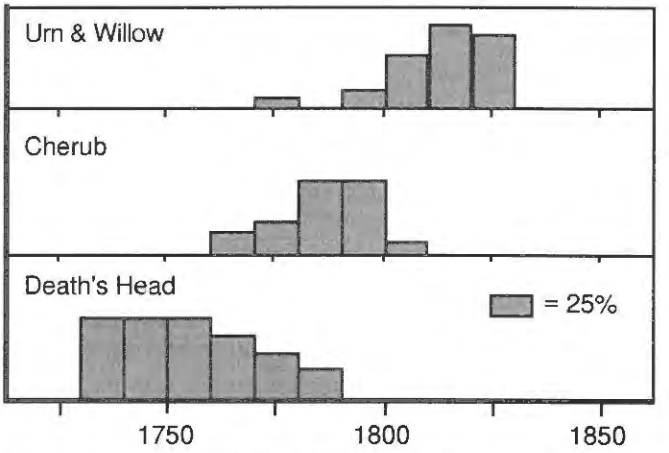

Figure 2.8. The same data rotated 90° and shown, without the symmetry, as a histogram to reflect the fact that the dates along the x-axis, in this case, are actually on a continuous scale.

along the x-axis to be equal. This has also led to some grievous errors in histograms published in archaeological reports. Two kinds of situations in which this often occurs involve chronological scales and grain-size distributions.

Commonly, archaeologists want to show change is some variable, such as settlement intensity, over time but we are often forced to assign observations to intervals of time that are not equal in length. Let us say, for example, that we want to show how the number of sites (our indirect measure of human population perhaps) varies over a period of many centuries, but we can date the sites much more precisely in recent times than we can in more ancient ones, and this is reflected in our grouping of sites into "periods" or "phases" of varying numbers of years (an interval scale). If we display the data in a histogram with bars of equal width, as in figure 2.9a, we will give the incorrect impression that human population was actually greater in the earlier (longer) periods than in the more recent (shorter) ones. If we correct for the interval widths, but indicate number of sites by bar height instead of area, we will still give misleading impression of changes in the intensity of settlement (figure 2.9b). In figure 2.9c, the intervals are shown correctly and the area of each bar, not its height, is used to indicate the number of sites in each period.

Similarly, if we want to make a histogram showing the distribution of soil particles, or the like, sorted by size, typically we would pass the soil through a series of screens of gradually decreasing aperture. Even if we have an impressive collection of screens, the increment in screen aperture will not be constant. In fact, it is likely that we would consider a difference between 1.0 mm and 1.4 mm to be much more significant than the difference between 10.0 mm and 10.4 mm, and so would not want to make all the increments equal even were this possible. For this reason soil scientists regularly use special non-linear particle-size scales instead of scales in mm (see chapter 11). But if we showed the abundance of soil caught on each screen with bars of equal width in mm on a histogram, it would give the impression that large particles (which typically are captured by screens with larger increments in aperture) are much more common than small ones, when in fact the reverse may be the case.

Most of the commercial computer packages do not give you the option of having unequal intervals. If we need to draw a histogram with unequal intervals manually, the procedure is the same as before but with a small complication. We need to transform the data into a ratio

measurement, such as "sites per year" for the first example or "soil mass per mm" for the second. This will give us the heights we need to preserve the proportionality of area when the bars' widths are made proportional to the number of years of each period or the size increment between screens. You then need to "stretch" each bar to the appropriate width, using a graphics program. If, however, all of our increments are equal except for one or two, we can simplify the process. If, for example, one interval is twice the width of all the others, we simply halve its height so that the area of a bar with twice the width will still have the correct area.

Before leaving histograms, it is necessary to mention how the arbitrary selection of interval width and boundary of intervals affects overall histogram shape. Even a small change in one or both of these can sometimes make a bimodal (two-peak) distribution seem unimodal, or change the location of the modes dramatically (cf. Whallon, 1987:144-47). Given this, you should be wary of making too much of histogram shape, especially if sample size is small or the intervals are very small or large relative to dispersion in the data. For example, it might be ill advised to use three modes in a histogram of data in a small pilot sample to define the ordinal categories you will use for a much larger project unless those modes are fairly reproducible in several changes of interval width. At the same time, this dependence of histogram shape on interval placement and width does not prevent the histogram from being a very useful graphic device. It is an excellent way to display general trends in continuous data, as long as you view the graphs with a critical eye, and also has close similarity to the probability distributions we will see in chapter 6. In cases where it is critical to avoid this flaw in histograms, one should instead consider using an ogive or cumulative frequency distribution (below).

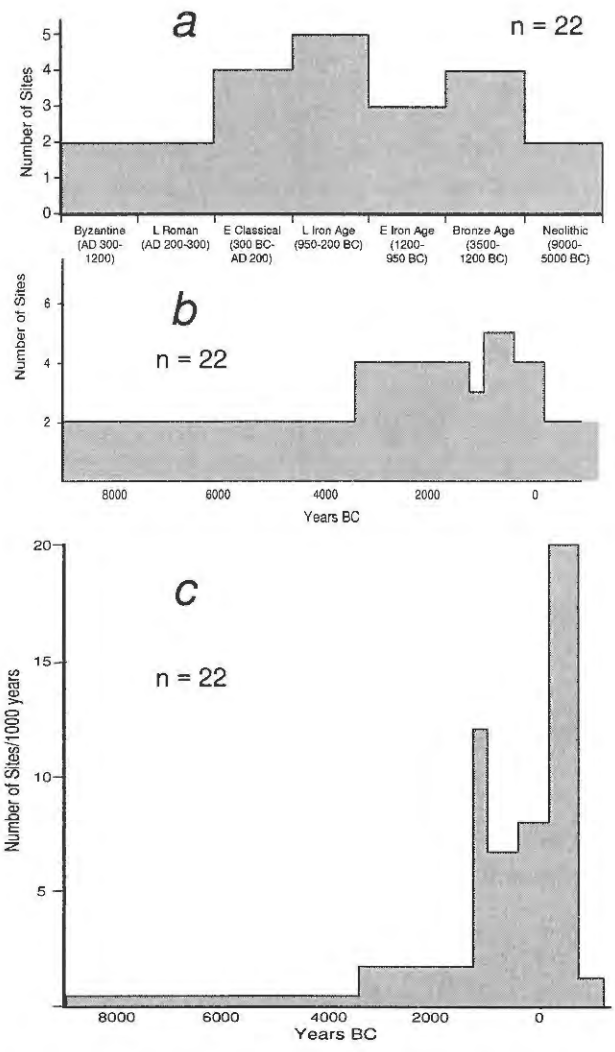

*Figure 2.9. The histogram in **a** is misleading because the time intervals are of equal width, the time scale is in reverse order, and site number is given by height, rather than area. In **b** the interval width is corrected but the graph still misrepresents changes in settlement intensity by using heights. The graph in **c** gives a more realistic indication of relative settlement intensity through proportionality of area.*

Line Graphs

Line graphs are correctly used in cases where we want to show changes over a continuous scale. In some ways they are similar to histograms, but they are not frequency distributions. That is, rather than showing how many or how much of some item occurs in each interval along the scale, it shows how a particular statistic (such as population size, proportion of decorated pottery, or mean temperature or house size) varies up and down as we vary the value of the x-axis, and, in archaeology, usually the x-axis represents time. In that case, the line graph can be called a **"time-series."** Another difference between line graphs and histograms, which stems from the fact that they are not frequency distributions, is that it is the height of any point

along the line, as measured by the y-axis, that determines its value, and not the area under the line. Usually you can get away with showing two or even three different measurements on the same line graph without making it too confusing, as long as each line is clearly labelled or coded with a key.

Often, it is impossible for us to make continuous measurements along this x-axis, and instead we have a sample of measurements made at different times and have to interpolate between them. A typical, non-archaeological example might be measurements of outdoor temperature at noon over a period of several months, but made not quite daily. We would simply mark points at the coordinates appropriate for each combination of x- and y-values (date and temperature), then join the points with line segments to show how noon temperature varies over the period in question.

Appropriate uses of line graphs are not as common in archaeology as those of other graph types. We could use a line graph to indicate, for example, changes in population over time (figure 2.10). The fact that we are usually unable to measure population (or one of its proxies) at a single point in time, however, makes it difficult or impossible for us to know where each point on the graph should lie, and we should show **error bars** on the points. These are line segments that extend out from the point to show the size of the estimated error. The graph in figure 2.10 has bars to show the estimated error in population size, but not in date.

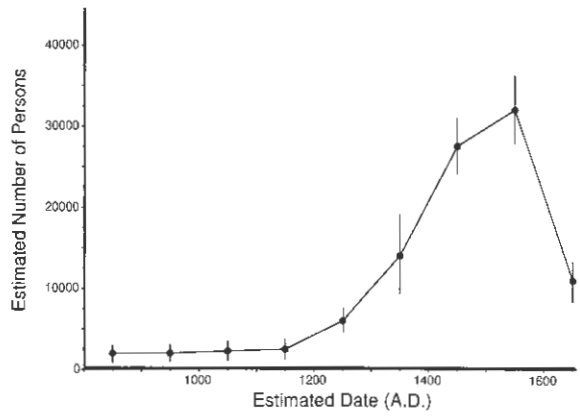

Figure 2.10. A Line graph showing changes in estimated human population over time (after Warrick, 1996:17).

Cumulative Frequency Graphs (Ogives)

The cumulative frequency graph, or **ogive**, is very useful if you want to know what percentage or proportion of your data lie above or below a certain value, or what proportion lies between two values. It is also very useful when it is important to know whether a particular distribution is relatively even or uneven, and for comparing distributions.

The graph has at least an ordinal, and usually a continuous interval scale along the x-axis and a measure of relative abundance (proportion or percentage) along the y-axis. It works much like a histogram except that the values accumulate, rather than going up and down. This characteristic is described as "monotonic" (one-way) "increasing." The graph is relatively easy to draw. You go through most of the procedure you would follow for a histogram, except you make a y-axis ranging from 0 to 1.0 or from 0% to 100% and you add each new value, going from left to right across the x-axis, to the previous ones so that the line gradually rises from near 0 at the bottom left corner of the graph, to 100% at upper right (figure 2.11).

One of the important applications of this kind of graph is in a non-parametric statistical test, the Kolmogorov-Smirnov test (Shennan, 1988:53-61) that is useful for many archaeological problems that involve comparing two distributions of observations on the ordinal, interval or ratio scale. In fact, the reason that the cumulative frequency graph became popular with lithic analysts was that you could put two or more cumulative step-lines (one for each assemblage) on a single graph and use the maximum vertical difference between the two lines as a measure of how different the assemblages were.

A more common use of such a graph is to allow viewers easily to see what proportion of houses were greater than 30 m^2, for example, or what percentage of graves had less than three grave goods. In the former case, we might find the place along the x-axis corresponding to 30

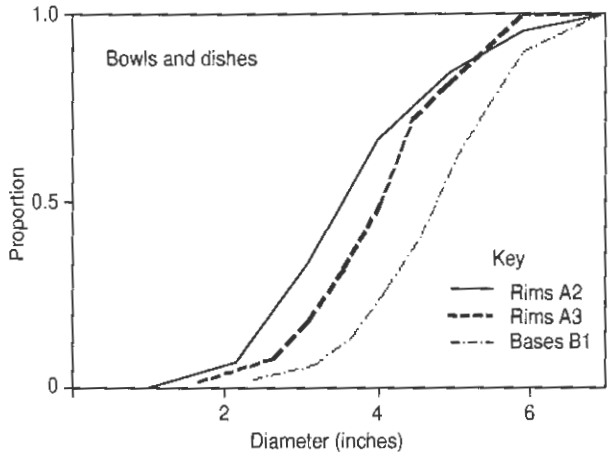

Figure 2.11. An ogive, or cumulative frequency graph, for the diameters of the rims and bases of pot sherds (after Orton, 1980:171).

m^2, draw a line vertically until it meets the step-line, and then horizontally from that point to the y-axis, where we might read off 80%. This would tell us that 100-80 or 20% of houses were bigger than 30 m^2. In the latter case, a line drawn upwards from 3 on the x-axis and then across to the y-axis might give 65%, telling us that 65% of the graves had less than three grave goods.

Clive Orton (1980) shows a good example of how a cumulative frequency graph, with rim and base diameter on the x-axis, can help us to compare differences in the proportions of size classes in pottery assemblages (figure 2.11). Orton makes the somewhat bold attempt to figure out what bases go with what rims by arguing that, if x% of rims are smaller than 6 inches and x% of bases are smaller than 3 inches, then probably the 6-inch rims belong to the same vessels as the 3-inch bases. The shapes of the cumulative frequency plots would be very different if this theory for associating bases with rims was far off, while close similarity between the two step-lines would lend it some support.

In non-archaeological applications, social scientists often use this type of graph to show income distributions. A fairly straight line would show that income varies fairly evenly, with relatively equal numbers of poor, middle-income and rich people or families, a concave distribution would indicate that there were very many poor families, and a convex distribution would indicate that there was an abundance of middle-income or rich families. Social scientists often then use something called the Gini coefficient to measure the degree of convexity or concavity in the distribution. Some archaeologists have used the Gini coefficient in a similar way to measure the degree of inequality in grave goods in a cemetery (e.g., Morris, 1987: 141-43).

Cumulative frequency graphs have often been used to publish distributions in lithic assemblages since François Bordes popularized his approach in the 1960s. This is actually an inappropriate use of them (Thomas, 1976: 52) because the x-axis on the graph is only an arbitrarily ordered list of artifact types. If you changed the ordering of the type-categories, the shape of the cumulative frequency graph could change dramatically.

Scatterplots

Scatterplots are the most appropriate means for showing data along two, or occasionally three, dimensions on an interval or ratio scale. They are particularly useful in exploratory analysis, when we are attempting to group data or find relationships between dimensions.

Usually a scatter plot has two axes, both with continuous interval or ratio scales. The plot then consists of a number of points, or dots, at the intersections of all the (x, y) combinations. For example, following an analysis of the trace elements in a number of obsidian artifacts, we might plot the abundance of zirconium in parts per million (ppm) in each along the x-axis and the abundance of barium, also in ppm, along the y-axis. Each point on the graph represents the particular combination of zirconium and barium for each artifact, and we can search the graph for signs that points tend to cluster into groups (figure 3.7). We can show errors of the measurements by small bars (error bars) extending out from the dots. We can also add a third dimension, usually on a nominal scale, by using different symbols for the points to indicate the category to which each belongs, such as the site it came from or the source of the obsidian, if known. If we are very lucky, points with the same symbol might fall into the same cluster on the graph,

while points with different symbols fall in different groups. This is a useful exploratory way to see what pairs of characteristics (in this case pairs of elemental abundances) might help us to define meaningful groups (see "grouping" in chapter 3).

Scatterplots are especially useful when we want to look for meaningful relationships between two dimensions. For example, if we plot flake width along the x-axis and flake length along the y-axis and the result is a group of points stretched roughly along a diagonal line, stretching upward from left to right, we would conclude that all of the flakes are fairly constant in shape (or at least width/length ratio), and simply vary in size. This kind of relationship is linear association, or **correlation**, and the extent to which two variables are correlated can be described with a descriptive statistic called the correlation coefficient, represented by the letter *r*.

One famous example of the use of a scatterplot in archaeology involves the attempt to find a relationship between the roofed area of dwellings and the number of people who inhabited them (figure 2.12, Narroll, 1962). The plot shows a fairly linear relationship between these two measures, and allowed Narroll to extract an equation for the line that fits the data observations with the minimum squared error. This equation predicts that there will be one inhabitant roughly for each 10 m^2 of roofed area. This fitting of a "best fit" line to the data in a scatter of points to extract an equation for predicting y when x is known is called **regression**. Note, incidentally, that the points are scattered rather widely around the $y = 10x$ line, indicating that the correlation is not especially good, and we would expect a fairly large degree of error in this particular indirect measure of human population size (LeBlanc, 1971; cf. Read, 1987:162).

Pie Charts (Circle Graphs)

Pie charts are a popular way to show proportions of things measured on a nominal scale. They are now easy, perhaps too easy, to generate with computers. They are constructed by radiating a number of line segments from the center of a circle, with the angles between radii proportional to the proportion of each category. They are particularly common to display faunal data in the archaeological literature (figure 2.13).

The trouble with pie charts is that, although the relative areas of the pie slices are proportional to the proportion of each category represented, just as the area of bars on a bar graph, the human eye finds it much easier to estimate and compare the areas of rectangles than circles or wedges (Tufte, 1983:55). Consequently, pie charts are usually not as easy to interpret accurately as bar graphs and are easily replaced by them in most cases. Often, users of pie charts try to compensate for this deficiency by putting labels on each wedge, showing the proportions numerically. This, however, makes the graph redundant, and begs the question of why the author did not simply use a table. In addition, the frequency of this type of graph in the popular media probably has more to do with the editors' low opinion of readers' sophistication than with the graphs' effectiveness (note especially the insulting "slice of your tax dollar" type of graph). In most cases, I would suggest that you use bar graphs instead of pie charts.

One exception is when you want to show how the abundance and distribution of items varies at a number of sites on a map. Assume, for example, that you have four pottery types that occur at ten sites in a region. You can draw a circle at the location of each site on the map such that the area of the circle is proportional to the total number of sherds in your sample from that site. Then you can divide each circle into four wedges, coloured or hatched to indicate which type each wedge represents and proportional to that type's abundance in the site assemblage. When used sparingly, this can sometimes be an effective way to present complex data and allow the viewer quickly to compare the distributions of artifact types across space.

Windflowers and Rose Plots

Windflowers are named for their use to indicate prevailing wind directions. They are essentially radial bar graphs, each bar radiating out from the center and with its height proportional to the amount of time the wind blew from a particular direction. Of course the direction in which the

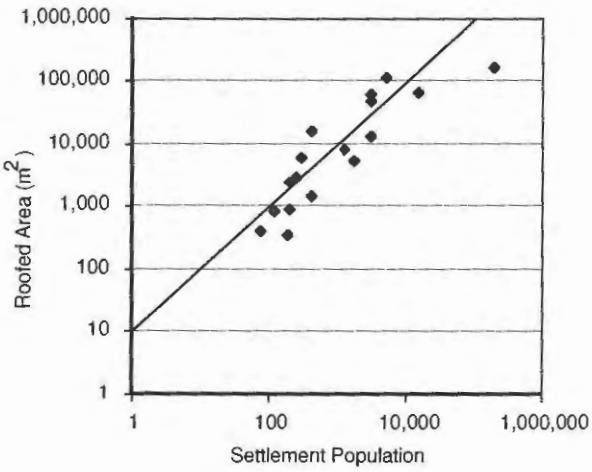

Figure 2.12. Scatterplot to display a relationship between roofed area and numer of people in an ethnographic sample, with the regression line for y=10x (after Narroll, 1962). Note that both axes have logarithmic scales in this example.

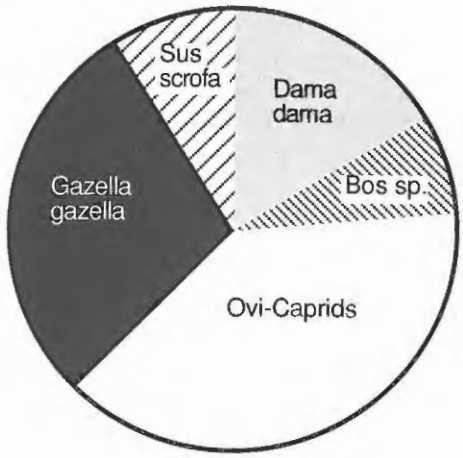

Figure 2.13. Example of a pie chart used to illustrate proportions of faunal bones.

bar radiates is analogous to the wind direction itself, so the windflower is appealing in that it conveys the idea of wind direction better than a standard bar graph would, even though it contains exactly the same information as a bar graph.

Some archaeologists have used windflowers and similar graphs called rose plots to show, for example, the distribution of orientations of the long axis in houses, the entryways in temples or houses, and the head position in graves. When excavators have recorded the individual orientations of lithics on surfaces, they can also be used to show the orientations or dip angles of the distal ends of flakes and blades (figure 2.14). Their attraction over a bar graph is that they convey the actual directionality of the patterns quite well.

Spatial Histograms and Isopleth Maps

Many kinds of archaeological data have an explicitly spatial component, and so it is often useful to combine the concept of a probability distribution with that of a map. Instead of showing variation in artifact size, for example, by using the usual histogram, we may want to show their distribution across a house floor we have excavated according to some kind of grid.

These maps, which we might call **spatial histograms** or stepped statistical surfaces, show variations in the density of artifacts or other materials across the grid. We can think of the density of artifacts across the surface as continuous, even though the actual artifacts themselves are discontinuous, and, as with a histogram, on each axis of the grid we decide on an interval (often we must select the interval in the field), and count the number of artifacts in each square. We could then represent this number by a bar that extends upward from the square, as in a three-dimensional bar graph, or simply shade or color each square according to a key to indicate the abundance of artifacts in each (figure 2.15). Note that, as with regular histograms, if the spatial units are not of equal size, we need to ensure that the data are shown as densities -- numbers of artifacts per unit of area or volume — to avoid giving an inflated impression of abundance for the larger spatial units.

An alternative that is better at capturing the feeling of continuity in the data, where that is appropriate, is to use an **isopleth map**. This looks much like a contour map used to illustrate topographic variation except that the contour lines represent, not equal heights above sea

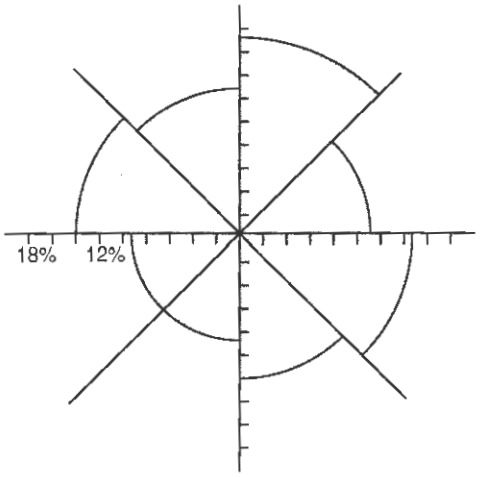

Figure 2.14. A type of rose plot used to indicate the angles of inclination (dip) of lithics in an Upper Paleolithic site (after Petraglia, 1993:102). The radial increments are 2%.

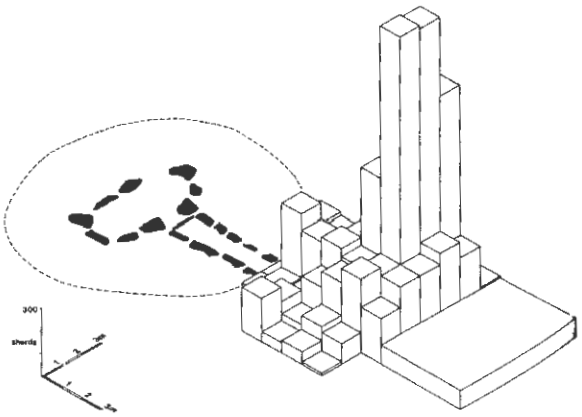

Figure 2.15. Spatial histogram or "stepped statistical surface" of potsherd densities at the entrance to grave 9 at Fjälkinge (Shanks and Tilley 1987:166).

level, but equal densities of artifacts, or isopleths. This type of map is appropriate in cases where an archaeological site-formation process has resulted in reasonably continuous changes in artifact density over space, so that we can reasonably interpolate artifact densities even in places where we did not actually do any measurements. What this means is that our measurements of artifact density across the surface must, of necessity, be discontinuous — normally we would count the number of artifacts in some unit of area at several places across the surface either systematically or opportunistically — but we then interpolate between our discontinuous measurements to estimate the densities in the space between them. This will only work in cases where we would expect fairly even fall-off in artifact density away, for example, from an activity area. In many archaeological cases the distributions are in reality very discontinuous. As with histograms, a small shift in the placement or interval of observations can have substantial effect on the form of the resulting "surface." In these cases, creating an isopleth map requires artificial, mathematical "smoothing" methods if we are to avoid misleading results.

General Principles in Creating Graphs

This summary does not exhaust the graphing methods you could use to display archaeological data, but does present principles for the use of the graph types most commonly encountered in the archaeological literature.

In addition to using the right kind of graph for your particular application, you should think about some general principles. It is easy to make even the correct kind of graph confusing, misleading, or downright dishonest.

An excellent guide for improving the effectiveness of graphs is Edward Tufte (1983; 1990). He emphasizes that graphs are to communicate information, and that they should do so clearly, precisely and efficiently. They should make large bodies of data accessible and coherent in a small space, balance fine detail with overall pattern, and encourage people to draw comparisons or identify patterns. They should have a clear purpose and should be integrated with the statistical and textual presentations with which they are associated. They certainly should not distort the data or mislead the viewer, and should encourage viewers to think about the data's substance rather than about the graph itself (Tufte, 1983: 13).

Among the aspects of graphs that affect their effectiveness, in his view, are the "data-ink" ratio and graphical integrity (Tufte, 1983:53-77, 91-106).

When he criticizes graphs for having a low data-ink ratio, what he means is that the graph is far too cluttered with lines, labels, or anything else that really is not necessary for communicating information. In practice, this may involve labelling too many increments on the x- and y-axes, adding bilateral symmetry to a graph (as in seriation's "battleship curves"), making bar graphs too decorative by making them appear three-dimensional, adding many and sometimes jarring combinations of color or hatching, or adding grid lines that are not really necessary. When graphs are decorative in the extreme, Tufte (1983:107-121) refers to them as "chartjunk." Unfortunately, commercial graphing software, whose primary market is the business communit,y makes it is all too easy for chartjunk to dominate. All this does is distract viewers from the information that the graph is intended to convey. Keep it simple.

On the other hand, it is also a mistake to remove so much labelling that the viewer has to search through a lengthy caption to interpret the graph. Include enough increment labels on your axes that what the axis is measuring, its scale, and its units are clear. Give your graph a title, and show other pertinent information, including, where relevant, a key, a scaling icon, or one or two descriptive statistics (such as n [sample size] or median).

One of the greatest threats to graphical integrity results from scaling the data in misleading ways. There are several ways in which presenters of data can mislead us. One is to hide the baseline of a bar graph or histogram so that, not only is zero not at the bottom of the graph's y-axis, but it is not very clear where zero is. This technique can be used both to inflate the impression of differences between bars (i.e., only showing the top part of the bars and truncating the bottom), and in an attempt to hide important differences and giving the graph a greater impression of stability (Tufte, 1983:54). In an honest bar graph, the height of the bars is proportional to the magnitude of the measurements they indicate. In a line graph with interval rather than ratio data, however, it may not be necessary to show the zero value, as it is not a "real" zero anyway. For example, in a line graph displaying changes in temperature over time, 0°F and 0°C are only conventional temperatures and need not be shown if they are outside the range of observations. Similarly, a time scale across an x-axis need not include A.D. 1 (there is no such year as A.D. 0) or 0 B.P. unless it is included in the time range of interest.

Another threat to graphical integrity is to distort the graph by removing the proportionality of area. As we have seen already, in our discussion of histograms, it is the area of a portion of the histogram (not the height) that should be proportional to the number of observations in a particular interval. As it happens, most people find it easier to compare lengths and rectangular areas than to compare curves, circles, and angles, and systematically tend to underestimate growth in a circle's area (Tufte, 1983:55). Some of the computer graphic packages offer the option of making graphs look three-dimensional. Unless you are actually displaying three-dimensional data, however, this distorts the proportionality of area on the graph, and sometimes also offsets bars so that you cannot even compare heights properly. It is inadvisable to use such an option unless, as in spatial data, the third dimension actually contributes to the viewer's understanding.

Tufte suggests two basic principles to preserve graphical integrity. First, the representation of numerical quantities should be directly proportional to those quantities. Second, the axes and units on the graphs should be clearly labelled, with labels on the graph itself, rather than only in a caption or key, and important "events" in the data, such as modes and sudden changes of slope, should also be labelled, to call the viewer's attention to what is important (Tufte, 1983:55-56).

Many of the options that modern computer graphics software offers, including perspective views with foreshortening, distort the area of bars considerably, or automatically shift the zero (origin) of the graph out of view. You should ensure that you overrule such distortions.

Transformations and Problems of Scale

Sometimes it is difficult to make a single graph display a set of data effectively because some of the observations are many orders of magnitude larger than others. If we follow Tufte's advice not to distort the scale of the y-axis on a bar- or line-graph, in these cases, either the large observations are lost off the top of the graph, or the small observations become nearly invisible.

A common way to resolve this problem is to put a break in the y-axis, indicated by a jagged or wavy line. Although this would violate Tufte's principal of proportionality, the jagged line signals to viewers that part of the scale has been removed.

Another way to handle the problem is to transform the data. A data **transformation** typically involves using the square root or the logarithm of your measurements, either of which has the effect of "pulling in" the high values and "stretching up" the low values. Other kinds of transformations may also be desirable. Essentially, transformation is like expressing your data in different units. The logarithmic scale, however, has no zero (it increments in powers of ten), and consequently cannot be used in cases where you need to display zero values (see figure 2.12). In order to make it clear to the viewer that you have transformed the data, and that the heights displayed on the graph are consequently not proportional to the linear values of the measurements, you should label the axis and the graph accordingly. Sometimes patterns that are not clear in the raw data become much clearer or easier to interpret after data transformation, as in cases where an exponential relationship appears linear after transformation.

In addition, there may be more substantive reasons for transforming data than making the data fit the graph. Some relationships between two or more variables can form extremely clear patterns only when they are transformed. In an attempt to discover pattern in many linear measurements made on a number of artifacts, for example, we may find that the untransformed data suggest groupings by size categories, while a simple transformation results in data that suggest groupings by shape. Data transformation is a standard procedure in some kinds of analysis, such as principal components analysis (chapter 3).

References Cited

Barnett, V., and Lewis, T., 1994, *Outliers in Statistical Data.* John Wiley & Sons, New York.

Coombs, C. H., 1964, *A Theory of Data.* John Wiley, New York.

Deetz, J., 1967, *Invitation to Archaeology.* Doubleday & Co, New York.

Hodder, I., 1989, Writing archaeology: site reports in context. *Antiquity* 63:268-74.

— 1990, *The Domestication of Europe. Structure and Contingency in Neolithic Societies.* Blackwell, Oxford.

Kent, S., 1991, The relationship between mobility strategies and site structure. In *The Interpretation of Archaeological Spatial Patterning*, edited by E. M. Kroll and T. D. Price, pp. 33-59. Plenum Press, New York.

Kyburg, H. E., 1984, *Theory and Measurement.* Cambridge University Press, Cambridge.

LeBlanc, S., 1971, An addition to Narroll's suggested floor area and settlement population relationship. *American Antiquity* 36:210-11.

Morris, I., 1987, *Burial and Ancient Society. The Rise of the Greek City-state.* Cambridge University Press, Cambridge.

Nance, J. D., 1987, Reliability, validity, and quantitative methods in archaeology. In *Quantitative Research in Archaeology. Progress and Prospects*, edited by M. S. Aldenderfer, pp. 244-93. Sage Publications, Newbury Park, CA.

Narroll, R., 1962, Floor area and settlement population. *American Antiquity* 27:587-89.

Orton, C., 1980, *Mathematics in Archaeology.* Cambridge University Press, Cambridge.

Petraglia, M. D., 1993. The genesis and alteration of archaeological patterns at the Abri Duflaure: An Upper Paleolithic rockshelter and slope site in southwestern France. In *Formation Processes in Archaeological Context*, edited by P. Goldberg, D. T. Nash, and M. D. Petraglia, pp. 97-112. Prehistory Press, Madison, WI.

Pirsig, R. M., 1974, *Zen and the Art of Motorcycle Maintenance: An Enquiry into Values.* Morrow, New York.

Plog, S., ed., 1986, *Spatial Organization and Exchange: Archaeological Survey on Northern Black Mesa.* Southern Illinois University Press, Carbondale.

Plog, S., Plog, F., and Wait, W., 1978, Decision making in modern surveys. *Advances in Archaeological Method and Theory* 1:143-82.

Read, D., 1987, Archaeological theory and statistical methods: Discordance, resolution and new directions. In *Quantitative Research in Archaeology, Progress and Prospects,* edited by M. S. Aldenderfer, pp. 151-184. Newbury Park, CA: Sage Publications.

Shanks, M., 1992, *Experiencing the Past. On the Character of Archaeology*. Routledge, London.

Shanks, M., and Tilley, C. 1987, *Re-Constructing Archaeology. Theory and Practice*. Cambridge University Press, Cambridge.

Stapert, D., 1994, Intrasite spatial analysis and the Maglemosian site of Barmose I. *Paleohistoria* 33/34:31-51.

Sullivan, A. P., 1978, Inference and evidence in archaeology. *Adances in Archaeological Method and Theory* 1:183-222.

Thomas, D. H., 1976, *Figuring Anthropology*. Holt, New York.

Tufte, E. R., 1983, *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT.

— 1990, *Envisioning Information*. Graphics Press, Cheshire, CT.

Tukey, J. W., 1977, *Exploratory Data Analysis*. Addison-Wesley.

Warrick, G., 1996, Evolution of the Iroquoian longhouse. In *People Who Lived in Big Houses: Archaeological Perspectives on Large Domestic Structures,* edited by G. Coupland and E. Banning, pp. 11-26. Prehistory Press, Madison, WI.

Whallon, R., 1987, Simple statistics. In *Quantitative Research in Archaeology, Progress and Prospects,* edited by M. S. Aldenderfer, pp. 135-50 . Sage Publications, Newbury Park, CA.