Unknown sampling bias is not a license to ignore statistical theory

George L. Cowgill

Most of us want to know how we can gain secure knowledge. What are valid reasons to think our ideas are right, and what should we demand of others in order to be convinced? Perhaps one of the deepest divisions in contemporary styles of thought is between persons who want very much to divide ideas into those that are true and those that are false, and who furthermore think this is generally possible (Type I people); and persons who are generally sceptical of such absolute claims and are comfortable with a large grey zone of ideas that have varying degrees of credibility (Type II people). This does not mean that Type II people (I am one) are happy to leave all ambiguities unresolved forever. It does mean that we see research as only occasionally punctuated by breakthroughs, when something previously considered doubtful, or even something previously unthought of, suddenly seems sure beyond all reasonable doubt. Much of the time, however, it is a matter of gathering evidence and reflecting on evidence in ways that only shift the odds somewhat in favour of some ideas and against other ideas.

If you accept this view about how ideas and research should change one another, it means that we cannot rely entirely on finding new evidence so overwhelming that its connections with ideas are obvious and unarguable; evidence that 'can only mean' this or that. Instead, we have to be rather sophisticated about connecting ideas and evidence. Among other things, we must recognise that there is no such thing as 'pure' evidence - all evidence is experience, conditioned by our personal pre-existing ideas. This is a valid insight emphasised by many post- or anti-processualists. I will not elaborate on it here, except to emphasise that evidence is conditioned by pre-existing ideas but is not created by them. Hence, acknowledgment of individual subjectivity does not lead to or warrant an unlimited 'anything goes' subjectivity. What comes from outside us counts, and it counts a lot.

I will, instead, concentrate on some uses of statistical theory in connecting experience and ideas. In the palmer days of the New Archaeology, statistical inference was very fashionable. More recently it has been less well regarded, even among archaeologists who are enthusiastic about other quantitative methods. A good example is the rather negative discussion of statistical inference in chapter five of Stephen Shennan's Quantifying Archaeology (1988), a weak section in an other-
wise mostly excellent book. The gist of his argument is that, since archaeological data often do not conform to assumptions of traditional kinds of statistical hypothesis testing, the results of such tests are apt to be misleading. Therefore, Shennan argues, it will generally be better to calculate various kinds of descriptive statistics, then assess their meaning by judgment and intuition, avoiding formal significance tests.

I agree to the extent that I think it is best when the descriptive results are so strongly patterned that any formal test of significance is superfluous. If, for example, a set of points in a two-dimensional space is as sharply clustered as those in Fig. 1, a formal test of the hypothesis that they are randomly distributed is unnecessary and rather foolish. However, our descriptive data are often not that nicely patterned. Even so, is it the case that we cannot do better than to use our personal intuition?

Fig. 1: A formal test of the hypothesis that the points are distributed randomly is superfluous.

It is all too easy to do worse than 'commonsensical' intuition if statistical inference is misunderstood and misapplied. Thus far, I agree with Shennan. However, our intuitions are often not very good either, and can be quite treacherous and misleading. On one hand is the danger of thinking our data more ambiguous than they are, and being too unsure of whether they support any one interpretation more than others. On the other hand is the danger, perhaps less often recognised but probably more prevalent, of intuitively placing greater confidence in the data than is warranted. This is especially the case for Type I persons. Simply, if the data were obtained by a technique so biased or otherwise problematic that no formal statistical inference is justified, how can we claim to have intuitively allowed for these problems in making 'commonsense' interpretations? By relying on intuitive judgments, we are likely to sweep problems of data quality under the rug, rather than facing up to them. This is no solution at all to problems of inference.

Doing inference right

Among the things that give statistical inference a bad name are the oversimplified and somewhat wrong versions of the logic that are common in introductory statistics texts. Quite a while ago I tried to list most of the problems (Cowgill 1977) but that article seems to have had almost no impact on practice, and I suspect that it suffered by being too involved and by not saying enough about better alternatives. In this section I will concentrate on one common fallacy: the belief that an arbitrary significance level, such as 5%, is a sure guide for securely sorting hypotheses into two categories which, whatever circumlocations we have learned to apply, such as 'disconfirmed', 'provisionally accepted' and so on, amount for practical purposes to true and false. This can make Type I people happy, but we ought to know better. Why should we confidently hold exactly opposite beliefs about the truth of a hypothesis, depending on whether our data have a 4.9% probability of occurring if the hypothesis is true, or whether they have a 5.1% probability?

Indeed, a sensible recognition of this absurdity is one of the reasons for a negative attitude about statistical inference in general. The remedy, however, is not to toss out all formal inference and 'wing it' intuitively, but to recognise that this and other problematic kinds of reasoning do not have to be part of formal inference. There are better alternatives.

One better alternative is to report the actual computed significance level: that is, the probability of obtaining data at least as different as the observed data are, from what would be expected if the hypothesis in question, and all the needed auxiliary assumptions, were true. Assuming we agree that all the key auxiliary assumptions are sufficiently well met (I will say more about this in the next section), everyone will also agree that a probability of 0.0001 is very strong evidence against the hypothesis and that a probability of 0.30 hardly counts against it at all. What about a probability of 0.04, or 0.06, or 0.11? Well, ... these are unavoidably more ambiguous cases. This fact may cause great pain to the Type IIs, but I leave it up to them to save themselves. My recommendation to them is to try to become a Type II. We who are already Type IIs can live happily with these ambiguous results.
Maybe it is this way; then again, maybe it is that way. We do not have to leave it at that, however, because we have a statement of what the odds are, and the odds are rather unequal. Nevertheless, one person can reasonably be more sceptical of the hypothesis than is another person.

This does not mean that we can never get any further. On the contrary, it is just these ambiguous cases that are one of the strongest spurs to further research (other spurs are awareness of deep ignorance, important practical consequences of getting it right, and recognition of paradoxical situations with regard to data and theory). If we get a probability that we feel is too low to ignore but too high to provide overwhelming evidence against a hypothesis, we have every good reason to seek more data that are likely to provide stronger evidence one way or the other.

However, even properly understood, the traditional hypothesis-testing / significance-level approach has limitations. Typically it takes the form of testing for absence of a difference between two or more sampled populations, or the absence of a relationship among two or more variables in a sampled population. But, if we feel that there is good sample evidence against these absences, merely feeling justified in saying 'Yes, there is a difference between these populations,' or 'Yes, there is a relationship between these variables in this population' does not get us very far. It is mind-boggling to think of cheerfully doing statistics on radiocarbon data only to test the hypothesis that the specimen died before 1950 (Cowgill 1977). We would not dream of doing radiocarbon studies without estimating the ages of the specimens.

Why is estimation not used more often in statistical inference? For example, assuming a situation in which it makes sense to compute a correlation coefficient, $r$, for a sample, it is often silly to test the hypothesis that the corresponding population correlation, $\rho$, is zero. It makes far more sense to compute an estimate of $\rho$, and a confidence interval for the estimate. To be sure, the computations are more difficult and the assumptions about the population and the sampling techniques may be more stringent, but it is not too difficult, with a desktop computer, to get a serviceable rough and ready estimate of $\rho$ and its confidence interval. In this and many other situations, estimation is far more useful than hypothesis-testing.

Estimates also tend to be easier than significance tests to interpret, or somewhat harder to misinterpret, both for initiates and for the statistically untrained. One reason is that estimates and confidence intervals provide a direct picture of what it is reasonable to think certain properties of the sampled population are, while hypothesis tests, at best, assess the probability that the properties of the population aren't something - typically the uninteresting situation that two populations aren't different, or that certain variables aren't related to one another.

A second reason is that in hypothesis testing the concept of 'power' (very roughly, how improbable the observed data would be if some other hypothesis were true of the population) is murky and often simply overlooked. With estimation, the width and location of the confidence interval graphically and succinctly show, in a single picture, the range of all hypotheses about the relevant population property that look reasonable on the basis of the sample observations.

I should add that, because the relevant probabilities often approximately follow a Normal distribution, 'one-sigma' intervals usually encompass a range of estimates that has about a 2/3 probability of including the correct population value. Presenting just the one-sigma range invites the highly over-optimistic interpretation that there is negligible probability that the true value is outside the range shown. I suggest a variant of box-and-whisker plots, with a box to show the one-sigma range and whiskers extending on either side to span the full two-sigma range - there is usually only about a 4% probability that the true value is outside the two-sigma range.

Coping with problematic assumptions

Another common objection to formal statistics, which applies to estimation as well as to hypothesis testing, is that in the real world of archaeological practice critical auxiliary assumptions are often violated, and often violated to unknown extents. Most, if not all, these assumptions come under two headings: assumptions about the population sampled and assumptions about the sampling process.

It assuredly behoves us to be aware of the assumptions about populations presupposed by any procedure of statistical inference we use. All I will say about this here is that there are often alternative procedures that may require weaker and more plausible assumptions, and that it is
possible to get too concerned about this matter. There may be some procedures for which even small violations of certain assumptions can have large effects, but in many cases moderately large deviations from assumptions about the population have small effects. For each statistical procedure we use, we need to know which assumptions about populations really matter, and how much they matter. It is not like ritual purity; it is not an all-or-none matter. Some formal procedures are not too sensitive to moderate violations of some assumptions, and will still yield usefully accurate results under these conditions.

The principal critical assumption about sampling is that the sampling process is unbiased. It is worth reminding ourselves that it is not samples that are biased - a sample may, by bad luck, be unrepresentative of the population from which it was drawn - but it is procedures that can be biased. A biased procedure is one that, if the results of many repetitions are averaged, tends to give an average value of the sample statistic in question that is different from the corresponding population value. There is no question that many archaeological data-gathering procedures are apt to be biased, and sometimes all the surviving data that we could even potentially gather may be biased by various vicissitudes of deposition and survival (Cowgill 1970; Schiffer 1976).

To the extent that biased data collection is a problem, it is hard to see how reliance on intuitive judgments can help much. Avoidance of formal estimation or other inference techniques is apt to make it easier to overlook the problem. Facing the problem, there are two things we can do. One is to think about the likely direction of bias; the other is to think about its likely magnitude. For example, silting or heavy vegetation will militate against the discovery of sites in a survey, and will have a stronger effect on the discovery of sites that are small and lack prominent architecture. We might, thus, expect that our estimates of the proportion of small sites without prominent remains, relative to sites that are large and/or have mounds, will be lower in heavily vegetated areas, as a consequence of surface conditions. Similarly, in surface collecting, it stands to reason that small objects whose colour is similar to the local ground or to non-artefactual objects such as local pebbles, are less likely to be spotted than are larger objects that contrast more strongly with 'background' materials. In such cases, just knowing the direction of the expectable bias is useful.

It is even more useful to have some idea of the magnitude of the bias, especially the magnitude relative to the width of the confidence interval computed on the assumption that the sampling technique is unbiased. If the magnitude of the likely bias is not large relative to the width of the interval computed on the assumption of unbiased sampling, then the effect is not serious. For example, Cochran (1963:15) shows that if bias is less than about a fifth of the standard error of an estimate, the effect of bias will be only moderate. Probably we can rarely assume such a small bias in archaeological data.

Suppose the computed standard error of an estimate is SE, but we believe there may be a bias, of unknown sign, and magnitude possibly as great as B. A crude 'worst case' correction would be to replace SE with SE+B as the standard error of our estimate. I suspect (though I have not worked out the math in detail) that if we assume that the probable magnitude of the bias is approximated by a normal curve with a standard deviation of B, a more reasonable allowance for bias would be to replace SE with $\sqrt{(SE^2+B^2)}$. Again, it would usually be better to think of a two-sigma interval, and replace 2SE with $2\sqrt{(SE^2+B^2)}$.

The narrower the unbiased interval, the more one has to worry about bias. In other words, if SE is quite small, it may give a very misleading impression of the accuracy of the estimate. Nevertheless, if $\sqrt{(SE^2+B^2)}$ is not too large, relative to whatever question we are asking, the possibility of bias need not undermine the ability of the data to provide strong evidence.

It takes no great leap of imagination to see where this argument is heading. It is certainly good to reduce bias as much as it can be without excessively increasing the cost of research. But some bias is unavoidable. In such cases, we should not just shrug it off and fall back on intuition. Instead, we should find out a great deal more about likely magnitudes of bias, directions of bias, or preferably both, in the procedures we are using.

In 1985 I set out to do this for one project. During the 1960s and early 1970s the Teotihuacan Mapping Project, directed by René Millon, collected about 900,000 sherds (as well as several hundred thousand other objects) from the surfaces of over 5,000 tracts within that huge ancient Mesoamerican city. Although surface survey of about 30 square km was intensive, only diagnostic and feature sherds, a small proportion of all visible sherds, were collected. It was reasonable to think that there might have been a bias toward collecting a higher proportion of sherds of certain very significant categories, and also reasonable to

Cowgill
wonder if different field parties had shown different degrees of bias. Fortunately, truly intensive collections of all sherds larger than about 2.4 cm. had been made from a few tracts. I and Brandeis University graduate students Joseph Lokaj and James Chiarelli and UCLA graduate student Martin Biskowski studied these collections in various ways, including comparing proportions of certain categories in the original collections and in the more intensive collections. A few of the results were summarised in a paper given at the 1986 annual meeting of the Society for American Archaeology.

Although the results were reassuring and suggested that the biases were actually smaller than I had feared, and were not large enough to materially affect interpretations based on the less intensive collections, all of us were drawn into other things and our tabulations have still not been fully analysed and published. In this regard we have not set a good example, but the very fact that we put several person-months into it at all is far more than most archaeologists seem to have done. And the fact that we have not felt great pressure to publish is, itself, because we knew that the archaeological community would not be particularly sceptical of our sherd count statistics even if we had never carried out such tests. In fact, archaeologists should have been sceptical of our untested tabulations, as well as of innumerable other archaeological data sets where such tests do not seem to have even been contemplated. It is wonderful how many problems of data quality can be overlooked by most readers if the problems are just not mentioned.

There have been a few other archaeological efforts to use experimental or other methods to actually study bias in various procedures, rather than merely speculating about it, but they have been pitifully few. I began by arguing for the value of formal methods of statistical inference, especially estimation, and I end with a plea for far better empirical data on magnitudes and directions of biases in various archaeological procedures. Without controlled knowledge of bias, it may hardly matter whether we use formal inference or informal intuition.