15. A Selection of Samplers: Comments on Archaeo-statistics

George L. Cowgill

George Cowgill (Ph.D., Harvard, 1963) has been serving as Associate Professor of Anthropology at Brandeis University. His fieldwork has been principally in Mesoamerica, in the Maya Lowlands, and especially in the Basin of Mexico. He has been engaged in computerized analyses of data from Teotihuacan, Mexico, supported by the National Science Foundation. His major research interests have been in mathematical and computer methods in archaeology, the comparative study of early civilizations, and the relationships between population and other sociocultural variables.

INTRODUCTION

An autobiographical anecdote by A. V. Kidder illustrates one style in archaeological research design. In 1907 he and two companions were introduced to fieldwork by Edgar Lee Hewitt, who took them to a mesa top in the Colorado Plateau country and casually gave them their instructions. “He waved an arm, taking in it seemed, about half the world, ‘I want you boys to make an archaeological survey of this country. I’ll be back in three weeks’” (Woodbury 1973: 2). The present book is one piece of evidence, in case anyone still doubts, that archaeology now is not quite the same as it used to be.

Perhaps less obvious is a difference between most (though not all) of the papers in this volume and much of the “avant-garde” archaeological writing of the 1960s. The polemics, while present here, are subdued, and prophets and True Believers are not much in evidence. There are no claims that a particular study has given us, or that a particular technique is bound to accomplish the results of others, as well as plan their own research. The papers in this book should go a long way toward serving this educational function. However, it would be inadvisable for statistical beginners to try to use the book as either a primer or a source of recipes. It is no substitute for a good advanced book on sampling, such as Cochran (1963) or Kish (1965).

In what follows I will not try to repeat or discuss everything of importance in these papers. I will first discuss some important general issues which I think could be restated more clearly, or have been neglected by other contributors. Finally, I will touch on the relatively few instances where I seriously disagree with writers on points of central importance.

GENERAL ISSUES

First, there is the matter of probability versus nonprobability sampling. I suggest that it is helpful to define two deeper contrasts in strategy. When we know almost nothing about the populations of interest, the most economical strategy may be some form of probing, a sort of preliminary exploration to get some notion of the gross characteristics of the population. Probing can include judgmental or systematic or opportunistic pits or borings in a site, but it can also include such things as exploratory surface reconnaissance of a region. Probing may well be a useful strategy in the early stages of research, but the possibilities of deriving trustworthy generalizations from it to say that probability sampling per se guarantees highly effective research, or that alternative strategies are not sometimes preferable. Individual writers differ considerably in their enthusiasm for probability sampling; but there is a consensus that sometimes at least it can be extremely useful while sometimes at least other strategies are better. Insight and sophistication are needed not merely to apply the techniques, but also to judge which techniques are preferable to realize particular objectives in specific situations.

There is a relatively high ratio of practice to preaching in these papers, and clearly the sampling lessons learned in other disciplines, however instructive for us, cannot entirely substitute for our own experience. Yet these papers indicate that we are well-launched in sampling in archaeology. This is one aspect of applied mathematics in archaeology which, if still very young, is at least beyond infancy; and there are a respectable number of archaeologists now who understand probability sampling fairly well and are doing it, not just talking about it. We will always depend heavily on the advice of professional statisticians. But this must be, and can be, an active collaboration in which archaeological expertise and judgment are brought into creative interaction with mathematical skills.

Nevertheless, the archaeological profession as a whole has only begun to become educated in both the theoretical aspects of probability sampling and its practical role in research design, so that they can evaluate intelligently the results of others, as well as plan their own research. The papers in this book should go a long way toward serving this educational function. However, it would be inadvisable for statistical beginners to try to use the book as either a primer or a source of recipes. It is no substitute for a good general introductory statistics course, preferably followed by a more advanced book on sampling, such as Cochran (1963) or Kish (1965).

In what follows I will not try to repeat or discuss everything of importance in these papers. I will first discuss some important general issues which I think could be restated more clearly, or have been neglected by other contributors. Finally, I will touch on the relatively few instances where I seriously disagree with writers on points of central importance.
are very limited, and it needs to be followed by different strategies as research continues. Incidentally, Kidder's work in 1907 was probably quite useful probing, appropriate for what was then known of Southwestern archaeology.

In contrast to probing are all strategies where the population we wish to study can be fairly well defined, and we already have some information, or at least some good hunches, about its structure and variability. These in turn can be divided into selection and sampling strategies. In all cases, a governing consideration is that we neglect no observations likely to be important. "Importance," of course, must be defined with respect to explicit goals, which do not necessarily have to be scientific. For scientific research, "important" data are those needed for decisive choices between two or more competing hypotheses or models, and those of considerable heuristic value for suggesting new hypotheses or new models.

Sometimes our resources, the nature of the data, and the nature of critical test implications derived from competing hypotheses are all such that we can define obvious criteria of relevance, and use these criteria as a basis for picking a manageable number of intrinsically important observations. In these cases, there are other observations we might make but do not make because we are satisfied that they are relatively unimportant for our purposes (always with the critical proviso that we do not destroy data which may be important for other purposes). This strategy I propose to call purposive selection. In the papers here, strategies at least partly or possibly similar are sometimes called purposive sampling, judgment sampling, or nonprobability sampling. I prefer to emphasize the difference between this strategy (at least in its pure form), and any kind of sampling.

There is little question that purposive selection is preferable to sampling whenever selection is feasible, sufficient for one's research objectives, and not wasteful. One of the reasons why improved techniques of data detection (aerial photography, magnetometers, etc.) are so important is that they offer new possibilities for feasible and effective data selection. Also, detection and selection offer our best safeguards against what Daniels (1972: 205) calls gross errors. These are instances in which our sample statistics fail to reveal or reflect the existence of rare but extremely unusual and intrinsically important elements of the population. Formally, they can be thought of as drastic failures to recognize population heterogeneity. Missing the Pyramid of the Sun in a survey of the Teotihuacan Valley would be a gross error; happily, an error not easily made. Missing the tomb of Tutankhamen was a far more easily made gross error, repeated by grave robbers for many centuries and corrected partly by luck in the 1920s. The nature of archaeological data is such that the risks of missing key data items of major importance are often hard to evaluate. They are greatest for archaeology-as-treasure-hunting, but often not negligible for archaeology-as-science. This is not an argument against sampling, but it is a reminder that purposive selection and new developments in data detection techniques are also important.

In contrast to selection, some form of sampling is preferable in situations (1) where there are more potential observations than our resources permit us to make, or we have reason to think that we do not need to make all possible observations in order to obtain convincing tests of competing hypotheses or acceptable estimates of important parameters, and (2) where there are no indications that tell us which of the possible observations are unimportant or unnecessary. In these cases, we can call the set of all possible relevant observations the population, and the set of observations actually picked the sample. For better or worse, our criteria for picking observations to include in the sample must be something other than obvious irrelevance of what we omit; in sampling we omit observations because we think they are redundant, rather than because they are intrinsically less relevant than the observations we include. Our interest is less in the intrinsic properties of the sample than in what it permits us to infer about the entire population.

In the case of selection, in a scientific context, the aim is to select observations sufficient for decisive tests of nomothetic hypotheses—lawlike generalizations of some sort. In the case of sampling, an additional stage intervenes: we attempt to pick observations in a way which enables us to make idiographic, essentially descriptive and phenomenological, inferences about the sampled population which are sufficiently precise, unbiased, and accurate that they can in turn be used for decisive tests of nomothetic hypotheses.

Very often, of course, purposive selection strategies are not feasible or, if we are confident we have not greatly underestimated population heterogeneity, may seem inefficient, and some form of sampling is required or preferable. The catch, then, is to pick the sample so that we have a sound basis for idiographic inferences about the population. Note that, in these terms, purposive sampling is practically impossible. In particular, the notion of picking a "representative" sample is ambiguous and apt to be misleading. If one takes it to mean picking items that exhibit extremes in the ranges of variation, or picking at least some examples of all types, or all "diagnostic" types, then this sort of picking is certainly possible and probably useful for limited purposes. The result, however, is most unlikely to be "representative" in any other sense. If, on the other hand, "representativeness" is taken to mean that we are positive that the sample statistics are virtually identical to the corresponding values for the population, there is no way this can be guaranteed without already knowing the population values. Attempts to pick a sample "representative" in this sense by judgment, opportunistic picking, or any other nonprobabilistic strategy founder on the difficulties of estimating bias in the procedures, or prove at best to be some form of covert (and often ill-conceived) selection, rather than sampling at all. In order to know where we stand with respect to idiographic inferences about the population, a reasonable approximation to some form of probability sampling is required.

As numerous chapters in this volume attest, simple random sampling is only one form of probability sampling. Often specific research goals, prior knowledge or strong hunches about the population, or considerations of data-collecting efficiency will argue for some form of cluster sampling, stratified sampling, systematic sampling with random elements, or other
variants of probability sampling. Also, research designs that include probability sampling are completely compatible with purposive selection, so long as one keeps good track of when one is selecting and when one is sampling.

For the most part, these papers are reasonably hardheaded about the limitations of sampling, and there are only occasional hints of the notion that in some obscure way sampling per se provides a guarantee that our conclusions will be correct. But among archaeologists in general, there is probably still a tendency either to expect too much or to be disenchanted with sampling because it does not fulfill impossible expectations.

One should always remember that there is no way, short of including most or even all the observations in the population, that we can absolutely guarantee the truth of most idiographic inferences of any interest about the population, quite apart from the difficulties of confirming nomothetic generalizations. The advantage of probability sampling is not that it very often enables us to be sure about the population, but that it helps us, as I said, to know where we stand in relation to various inferences.

One of the most prevalent abuses of statistical reasoning is the tendency to crank out some test of a hypothesis about a population, look up in a table to see if "p" is greater than or less than .05, and then make one of two interpretations. If p is greater than .05, the hypothesis is not rejected, and one feels free to act as if it were certain, for all practical purposes, that all alternative hypotheses about the population characteristic in question were false. If p is less than or equal to .05, the hypothesis is rejected, and one feels entitled to act as if it were certainly false. Aside from the patent absurdity of treating odds of one in twenty as somehow the magic threshold number that separates confident acceptance from confident rejection (a point noted here especially by Redman), there is another less often understood difficulty. Statistical inference is never complete if we only ask whether sample results are consistent with one specific idiographic hypothesis about the population. We always have to think also about what other hypotheses the results may also be consistent. This concept tends to be underemphasized in elementary statistics texts because there are often an unlimited number of possible alternative hypotheses, and there is no single number that expresses the probability of the observed sample statistic under all possible alternatives. Instead, a power function, a mathematical function of all these alternatives, is required. This is not the place for a lengthy discussion of this problem (cf. Rozeboom 1960 and Morrison and Henkel 1970), but the upshot is that, especially for small samples, a result that does not enable us to reject one hypothesis at a 5% significance level may also be quite consistent, and even more consistent, with very different hypotheses about the population. In general, it seems wiser to focus less on tests of specific hypotheses (especially the traditional "null" hypothesis that what we're looking for isn't there), and more on best estimates and confidence intervals. For very small samples (Asch gives some examples in chap. 10 in this volume) the confidence intervals may prove to be wide enough to warn us that the sample results are consistent with an extremely wide range of hypotheses. On the other hand, very large samples may give results that would be highly improbable if the null hypothesis were exactly true, yet provide estimates and confidence intervals which imply that the null hypothesis almost surely differs from the true value by only an inconsequential amount.

If one must think in the significance test framework, one should abandon the simplistic notion that we are committed to, or even entitled to, a twovalue yes/no decision strategy. This commitment is necessary if one is deciding, for example, whether to extend an unproductive exploratory trench another meter. But in terms of idiographic hypothesis testing with a view toward nomothetic theory confirmation, something at least as complicated as pretty-surely-yes/maybe/pretty-surely-no is preferable. This is not an evasion or a cop-out. Rather it is appropriate use of statistical reasoning to express what the sample results tell us about hypotheses, rather than to provide a phony aura of certainty for ambiguous data. Ambiguity is where we begin. If we were sure of the answers, there would be no point to research. We try for a research design that we think will settle the questions we started out with, and hope will also lead to new questions. But if some of the answers to our original questions remain ambiguous our response should not be to try to deny this, but to design further research (if it's still possible) that will provide whatever more is needed to shift "maybe" to "pretty-surely-yes" or "pretty-surely-no."

In this vein of belaboring misconceptions, it is probably impossible to attack too often the persistent delusion that there is some special merit in a 10% sample, or in any other preconceived sampling fraction for that matter. A point repeatedly made in these papers is that unless the sampling fraction is more than 20% of the total population, the proportion of the population included in the sample is of negligible importance. What is virtually all-important is the absolute size of the sample; that is, the actual number of independent cases (or, in cluster sampling, the effective number) included in the sample. For a population of 40, a 50% sample may be barely adequate, while for a population of 200 million, a 0.01% sample (20,000 cases) may be more than ample. Sometimes I visualize an afterlife in which people who think 10% sampling is a good rule of thumb are set to work drawing such samples from infinite populations. In fact, as several papers here illustrate, selection of a suitable sampling fraction is an important problem in good research design. It depends on prior knowledge or good guesses about the variability in the population, the precision of estimates required for good tests of important hypotheses, and the research resources obtainable.

Five concepts which are used frequently in these papers are accuracy, precision, bias, representativeness, and skewing. They are generally defined and used correctly (although Read, chap. 4, confuses precision with accuracy), but the definitions may be too brief and possibly confusing. Therefore, I will try to explain them more fully. This seems particularly necessary because many archaeologists appear to confuse bias and unrepresentativeness. Bias is also sometimes confused with the existence of structured relationships within the population, or with skewing.

My explanation of the meanings of accuracy, precision, and bias is intended to be a commentary in less technical terms on the discussion by Cochran (1963: 12-16). I will follow standard practice in calling sample
values statistics and the corresponding population values parameters.

Consider the problem of estimating the true population mean, \( \mu \), from a sample mean, \( \bar{X} \). If we were to draw a large number of independent samples, following the same sampling procedure in each case, each sample would in general have a different mean – \( \bar{X}_1, \bar{X}_2, \bar{X}_3, \) etc. – and these sample means will generally not be exactly equal to \( \mu \), the population mean. We can also think of the mean of all the means of the independent samples, and call this mean of means \( m \). It may or may not be true that as the number of independent samples becomes infinite, the probability that \( m \) differs by any finite amount from \( \mu \) will approach zero. If the sampling technique is biased, or if \( \bar{X} \) were a "biased estimator" of \( \mu \), the expected value of \( m \) for a very large number of samples may be substantially different from \( \mu \).

Precision is a measure of how widely individual sample means, \( \bar{X}_i \), are spread out around \( m \), the mean of these means. In other words, precision is a measure of the probability that the one sample mean we actually obtain differs by more than some given amount from \( m \), the mean of all sample means we would obtain, if we were to repeat, independently, the same sampling procedure a large number of times (independence here implies that inclusion of an observation in one sample has no effect on its chances of being included in another sample). High precision means that our sample results are highly repeatable, whether or not they are correct. For many forms of probability sampling, the amount of spread to be expected in sample means can be deduced mathematically from known population characteristics, without actually having to draw any samples at all, and it turns out for simple random sampling to be directly proportional to the amount of spread of individual values in the population, and inversely proportional to the square root of the size of the individual samples. Precision can be increased either by increasing the sample size (but 4 times the sample size is needed to double the precision if population variance is known) or by using prior knowledge to define subpopulations which are relatively homogeneous with respect to the variable in question. This is one reason for stratified sampling designs, as is pointed out by many papers here.

Accuracy, in contrast to precision, is a measure of the extent to which the individual sample means are spread out around \( \mu \), the true population mean, rather than \( m \). If \( m \) is equal to \( \mu \), then accuracy and precision are equal, but if \( m \) is not equal to \( \mu \), then the accuracy will be less than the precision.

Bias, in this example, is the difference between \( m \) and \( \mu \). More generally, it is the difference between the expected long-run mean value of the sample statistic, if the same sampling procedure were to be repeated a great many times, and the true value of the population parameter which the sample statistic is used to estimate.

There are, so far as I know, two ways in which bias can occur. The first is that, even under probability sampling, the theoretical long-run expected value of the sample statistic is not the same as the corresponding population parameter. For example, for simple random sampling the long-run expected value of the sample variance differs in a known way from the true population variance, and this is why the sample variance has to be multiplied by \( n/(n-1) \) in order to obtain an unbiased estimate of the true population variance. However, for simple random sampling, \( \bar{X} \) is an unbiased estimator of \( \mu \).

The other way in which bias can occur is that our actual procedure for picking observations from the population differs seriously from whatever probability sampling scheme is presupposed by our computations. For example, if we wished to estimate mean size of sites in a region, we might assume simple random sampling, in which the probability of picking any site is independent of its size, and use mean site size in our sample as the estimator of mean site size in the population. But if in fact our procedures are not very carefully designed to give small sites nearly as good a chance of inclusion as larger and more obvious sites, the sample mean is much more likely to be larger than the true population mean, rather than smaller. Even though many repetitions of the same sampling procedure might show quite high precision, the mean of all these sample means would be substantially higher than the true population mean—that is, there would be substantial bias and the accuracy would be low. In fact, if the possibility of bias is not recognized, and the theoretical precision is taken to be a good measure of the actual accuracy, the results will be downright misleading.

Note that it is meaningless to speak of bias with respect to the population. "Bias" should not be confused with structure or the existence of nonrandom relationships between variables in the population. Note also that (as Asch, chap. 10, and Read, chap. 4, observe), it is not samples that are biased, it is procedures that may be biased. If a particular sample happens to be such that \( \bar{X} \) is considerably different from \( \mu \), it would be best to describe it as unrepresentative. I suggest that closeness of specific sample statistics to specific population parameters is the best way of sharpening the ill-defined concept of "representativeness." No sampling procedure guarantees a representative sample. The accuracy of a particular procedure is a measure of the risks of obtaining a sample which is unrepresentative by more than some specified amount.

Finally skewing, strictly speaking, has no direct connection at all with bias. It refers to the shape of a frequency distribution, either in the population or in a sample. Its importance is that some statistical procedures presuppose that the population distribution is approximately a symmetrical bell-shaped "normal" curve, while many other procedures do not require this assumption about the population. When the assumption is clearly unjustified (as it often is for archaeological data) or in serious doubt, one should be sure to avoid reasoning which depends heavily on this assumption. In general, the assumption is far more critical for small samples than for large samples.

The most refractory source of bias in archaeology is that the probability or even the possibility of making many observations once potentially available has been drastically altered by events and processes over which archaeologists have no control. This is because some kinds of data are inherently more durable than others, or more durable in some contexts than in others, and because post-occupation events may bury, redeposit, or destroy data in quite nonrandom ways. These are difficulties I have discussed elsewhere (Cowgill 1970), and they are elaborated on here especially by Chenhall.
• tion readers against overreacting to them. Clearly we are just deceiving
ourselves if we try to talk them away or pretend they are not real. However
this does not mean that archaeology is hopeless or that uncontrollable bias
is so bad that we can comfortably forget all about probability sampling and
careful research design, and continue to rely happily on judgmental or
opportunistic samples.

What can often be done is to make reasonable estimates of either the
maximum likely absolute magnitude of the bias, its direction, or both. If
the probable magnitude of the bias is not great relative to the accuracy we
require, then its effects will not be too serious (cf. Cochran 1963: 15). If
the direction of the bias can be deduced, even though we are uncertain of
its magnitude, then we know that tests of significance which ignore bias
will be conservative tests of hypotheses which assert that the population
differs from the sample in the direction opposite to the bias. The troubles
in this case are that (a) we cannot tell if we are being far too conservative,
and (b) we may be stuck with an interesting hypothesis which predicts
differences in the same direction as the bias. At any rate, the things to do
are to try to recognize sources of bias and to gain far more experimental
evidence about these sources and their effects. We can then design detection
and research strategies to overcome the effects where possible. Where
this is not possible we still need not despair. Often we can make sensible
allowances for bias if the sources of bias are understood.

Another topic alluded to in several papers and which calls for emphasis
is the matter of "grain size" of sampling units. It is important that sampling
units be substantially larger than significant spatial dimensions of the phe-
nomena we are interested in studying.

If we are primarily concerned with aggregates of items (such as sites or
artifacts), the grain size need not be much larger than the individual items.
But whenever we are concerned with patterned or systemic relations be-
tween diverse items, as we increasingly often are, we require grain sizes
(as well as sampling fractions) sufficient to make entire patterns pretty clear.
A grain size and sampling fraction quite adequate and efficient for settling
the question of whether sites are substantially more abundant in one eco-
logical zone than in another may be very inadequate for revealing settlement
systems.

It is relevant here to discuss some instructive results reported by Rohn
(n.d.). Density plots of surface sherd counts and weights were compared
with the results of total excavation of a Basketmaker III pithouse village
which covered about 0.224 hectares (0.55 acres) and was only about 20 cm.
in depth. In spite of the shallowness of the deposit, surface sherd concen-
trations show only a moderate correlation with architectural features. Fur-
thermore, simulations of randomly located 2 × 2 meter squares using 10%,
25%, and 50% sampling fractions indicated that even a 50% sample of test
pits would not yield a particularly good picture of site features. One lesson
I draw from this is that, as numerous other experiences have demonstrated,
test pits are not the way to find out about structures larger than the test
pits. In some cases pits may be the best way to locate features, but they
should then be expanded to reveal more or less the entire feature. A second
lesson is that even on relatively undisturbed shallow sites surface artifact
concentrations cannot be expected to pinpoint underground structures
more closely than within 10 or 20 meters. For a settlement whose maximum
dimension is about 50 meters this is a serious limitation. The validity of
surface data, relative to the overall scale, seems much greater on sites cover-
ing more than a few hectares.

A final important general point, only touched on here by Judge, Ebert,
and Hitchcock (chap. 6) is the relevance of sophistication about sampling
for salvage, contract, or conservation archaeology. Increasingly archae-
ologists are being called on to design, to justify budgets for, and to establish
guidelines for work in regions where archaeological data are seriously
threatened by contemporary human activities. There is a risk that projects
which look adequate, judged by conventional archaeological wisdom, will
in fact prove inadequate and leave important questions forever moot be-
cause the relevant data have been destroyed. Clear definition of objectives
and statistical expertise at the planning stage are needed to minimize this
risk. At the same time, we have to make the best use of scarce resources,
and avoid overemphasis on certain regions, or certain kinds of data, at the
expense of other kinds of data or other regions. Our research designs and
our budgets should be defensible against the criticism that we are afflicted
with a mindless compulsion to find everything, dig everything, and save
everything. It is true that whenever we make any judgments that further
data would be redundant we have to face the very real possibility that,
sometime in the future, they will assume an importance not now recognized.
But it should be of very great value to be able to offer a reasoned set of
predictions about payoffs for various investments of resources. That is,
we should be able to say that, in order to have good prospects of reason-
ably firm answers to certain basic questions about the prehistory of a given
region, approximately X investment of dollars, time, and manpower will
be necessary and also probably sufficient; while there is little prospect
of getting conclusive answers to such-and-such further questions, unless Y
further investment is made. Individuals and institutions would then have a
much more informed basis for judging or justifying the adequacy of research
plans and the merits of budgets.

The outcome of such calculations will probably be that, at least whenever
we are concerned with currently important questions about the systemic
aspects of the data (let alone when we worry about unimagined future ques-
tions), the scale of research required will be considerably larger and the
sampling more intensive than has customarily been thought adequate. Cer-
tainly the notion that a 10% sample is generally both necessary and sufficient
will crumble. A strong argument for more intensive work and a higher level
of financial support will be the demonstration that such effort is needed in
order to solve problems whose importance is already recognized; it is not
simply called for because of wasteful research designs or in order to keep
archaeologists employed.
**SPECIFIC COMMENTS**

Chenhall (chap. 1) is pessimistic about sampling, and concludes that in his test case something like a 50% sample was required in order to get representative data. Unfortunately, there are two serious errors in his argument. First, he has converted actual frequencies of sites into percentages, and then computed chi square statistics based on these percentages. This is invalid.

For example, from a known population of 76 sites classified into 14 types, he selected a simple random sample of 30 out of 95 tracts into which the region was divided — a sampling fraction of about 32%. The sample proved to contain 17 sites. Chenhall tabulates the number of sites of each of his 14 types in the population, the number of sites of each type in the sample, converts to percentages, and emerges with a chi square of 98.64. He notes that, with 13 degrees of freedom, such a large chi square has a probability of less than .001, and he concludes that in this case sampling about 32% of the area has yielded a sample of sites which is highly unrepresentative of the population. A second trial, using 48 of the 95 tracts, a sampling fraction of slightly over 50%, was treated in the same way and yielded a chi square of 21.3, which has a probability less than .10 but greater than .05. He concludes that even this 50% sample of tracts is only marginally successful in getting a representative sample of sites.

If chi square for Chenhall's 30/95 sample is recomputed using frequencies rather than percentages, it is only 22.0, rather than 98.64, and the tabulated significance level is between .10 and .05. Actually chi square is a quite inaccurate approximation here, since of 28 expected frequencies, 23 are less than 5, and 14 of these are less than 1. I think the effect is to increase considerably the probability of a large chi square. At any rate, the observed sample seems within the range of probable sampling variation predicted by theory for this situation. An approximate recomputation of chi square for his 50% sample indicates that the value is between 8 and 10. This implies a significance level between .80 and .70 and this sample is very well within the theoretical range of sampling variation.

Parenthetically, if Chenhall's chi square of 98.64 were valid, it would imply that his procedure had not merely failed to pick a representative sample, but that he had succeeded in picking a sample so unusual that, if simple random sampling of sites had indeed been approximated by his simple random sample of tracts, a comparably unrepresentative sample could be expected only once in many thousands of trials. The implication would not be that the sampling fraction was inadequate, but that there was probably some drastic failure of method.

The second serious difficulty with Chenhall's study is that the entire "population" that he sampled consisted of only a 4 square mile region, in which 76 sites were located.

Judge, Ebert, and Hitchcock (chap. 6), who also compared samples with a known population, got considerably better results than Chenhall. In part this is because they avoided his computational errors, but it is also because the absolute sizes of their samples are more adequate. Their population covered 5 times the area of Chenhall's and included 1,130 sites. They used a 20% sampling fraction and located 214 to 227 sites in their various trials. This means that each of their samples covered as much territory as Chenhall's population, and included about 3 times as many sites. David Thomas' survey area (chap. 5) was nearly 34 times as large as Chenhall's, and the area in his sample (10% of the total) was about 3.4 times the area of Chenhall's population. Matson and Lipe (chap. 7) plan to sample a region about 50 times the size of Chenhall's at an ultimate rate of only about 1.75%, but this still calls for covering about 84% of the area, and several times as many sites, as in Chenhall's total population.

The point is that, irrespective of sampling fractions, Chenhall's samples are extremely small in absolute size, whether measured by total land surface covered or by number of sites discovered. In fact, he is right, although for the wrong reasons, in arguing that, for the population he has defined, something much more than a 30% sample is badly needed. This is true, but it is because the numbers of cases in his samples are so small that they are consistent with widely differing hypotheses about the population. The problem with his samples is not that they are biased, but that they permit only extremely imprecise estimates of his population parameters. For that matter, his population itself seems much too small to be meaningful except as a pilot study. It is difficult to believe that the inhabitants of any of his sites confined their significant activities and environmental interactions to this 4 square mile tract.

Chenhall also raises difficulties about testing associations between cultural activities and environmental variables. Obviously it is easy to think of enough conceivably important environmental variables to generate, from their logically possible combinations, far more micro-environmental strata than could possibly be used in any feasible research design. But there is no need to prestratify the population by every conceivably important combination of variables. Stratification is, after all, just an intelligent use of prior information or hypotheses to increase sampling efficiency without too greatly increasing design complexity. The effects and possible interactions of many environmental variables can be investigated by a number of multivariate techniques. Chenhall himself does this to some extent.

I do not follow Collins' argument that we always must verify estimates derived from a first sample by taking a second sample. This may be an unnecessary overextension of the quite sound warning that one cannot calculate a large number of association coefficients between many different pairs of variables, selectively weed out the majority which are nonsignificant, and then take seriously the cookbook significance levels of the remaining high coefficients. Scaulding (1973) discusses this well.

Mueller (chap. 3) underlines the point that many of the feasible procedures for archaeological sampling are forms of cluster sampling, which undoubtedly is the case. He also stresses the distinction between spatial and nonspatial calculations. The former are those in which some quantity of excavated volume or surveyed area figures explicitly in the computation of ratios, and the latter are frequencies or ratios in which neither volumes nor areas appear. Mueller argues that nonspatial statistics do not conform to cluster sampling. His point is that if one, for example, excavated 30 rooms in a 100-room pueblo, one should not simply add up all the sherds of type X in all the rooms, find a grand total of sherds of type X, compare this to the grand total of all sherds in all rooms, and treat these numbers as if they were derived from any kind of sampling scheme in which individual sherds were
the sampling units. However, this hypothetical case could be treated as
cluster sampling with 30 clusters, and cluster sampling concepts and for­
mulas would then be appropriate for this situation.

Read's chapter 4 contains a discussion of bias, errors, precision, accuracy,
and considerations entering into choices of sampling schemes and sampling
fractions. In general his discussion is very good, although at one point he
momentarily confuses precision and accuracy, and he describes small sam­
ples as tending to be biased, where I think it would be more correct to say
that they tend to be unrepresentative. He goes more deeply than most
contributors to this book into the mathematics of cluster sampling, and
this should be quite useful for readers who make the effort to follow his dis­
cussion here. Especially valuable is his demonstration that cluster sampling
may be either more or less efficient than simple random sampling, depending
on differences in how data items are distributed in different populations.
This leads him to emphasize a point which is also made in many other chaps­
ters: the more prior knowledge one has of the population, the better one can
design further research. If prior knowledge is not very full, a multistage
strategy is advisable, and information gained at each stage can be used in
planning subsequent stages.

David Thomas' chapter 5 is another useful case study, coupled with a
technically sophisticated discussion. It seems open to at least one spurious
criticism: Thomas asks whether his sample data support the existence of
certain population differences predicted by his theoretical model, computes
certain statistics, and finds that the sample differences are not statistically
significant. He then reformulates the question slightly, computes other
statistics, and obtains highly significant differences. A skeptic might wonder
if this is anything but hocus-pocus, and even be reminded of the unkind
aphorism that there are lies, damn lies, and statistics. Put more carefully,
the question goes like this: "If we compute 3 different statistics intended
to ask substantially the same question of a specific sample, and 2 of them
give results that are not highly significant while the 3rd is highly significant,
what can we conclude?"

We should remember that computing a multiplicity of statistical tests that
ask closely related questions of the same sample is not analogous to com­
puting associations between a great many different variables. The statistical
tests, even though different, will give highly correlated results. We cannot
expect that if we ask the same question of random data in 20 slightly different
ways, at least one result will be significant at the 5% level just by chance.
In fact, Thomas asks only 2 or 3 variants of his basic questions, and gets a
number of results which are significant not only at the 5% level, but at the
1% level or better. In the strict sense of testing idiographic hypotheses about
phenomenological differences in his subpopulations, I find his results
convincing.

Again, it might be well to shift away from one-sided emphasis on tests
of significance and think more in terms of estimation of population para­
meters. I suspect that Thomas' results that were consistent with the hypoth­
esis of no differences between subpopulations might also be consistent with
the alternative hypothesis of substantial differences. That is, the computa­
tions that do not yield high significances may not be in conflict with those
that do give high significance; they may simply be inconclusive. In ques­
tions like these reliance on the "cookbook" approach can be very misleading,
as Thomas realizes, and the need for full insight into the techniques and the
(often partly implicit) reasoning behind them is acute.

On the level of nomothetic theory testing, Thomas' paper is not so satis­
fying. He presents test implications derived from Steward's theory of Basin
ecological adaptations, and presents good evidence from his sample that
most of these implications are true of his population. But he does not present
test implications derived from any theory other than Steward's, and there­
fore we are quite unable to judge whether his idiographic inferences about
the population might not be just as consistent, perhaps even more consistent,
with the implications of some contrary theory. As it stands, he has done only
part of the job. Of course, many other reputedly scientific archaeological
"tests" of theory have suffered from this same flaw.

Judge, Ebert, and Hitchcock's chapter 6 comparison of results from 4
different sampling designs for the 20 square-mile (51.8-square km.) Chaco
Canyon area is also very instructive. However, we should not generalize
their results unquestioningly. We need further replications of these same
designs on the same population, replications of other designs, and repli­
cations on other populations. Read's paper makes this point also. A shift
to stratified sampling with unequal fractions, which Judge et al. suggest,
would probably produce an improvement in precision far greater than the
differences between any of the 4 designs they actually tested. Their Canyon
Bottom stratum, which alone included over half the sites in the population,
gave estimates of total site density within 20% of the true value for all 4 of
their sampling strategies, and in 3 cases within 10%. But the South Mesa,
with less than 10% of the area and 1% of the sites, in 3 of 4 cases gave esti­
mates that were off by 60% to 200%. For the other 3 strata nearly half the
estimates of site density were off by 50% to 200%. Less intensive sampling of
the Canyon Bottom (assuming it is indeed homogeneous enough to justify
being treated as a single stratum—note here Thomas' experience with
poststratification) and more intensive sampling of the other strata would
yield much better overall precision for the same total sampling effort. Nev­
evertheless it is clear that overall deviations of estimates of site density cannot
be brought below 10% to 20% for any strategy that does not require a larger
total sample. Precision of estimates of site density by phase or other less
gross data features will be still lower. As in other papers, the implication
is that if we want very high precision, we must draw rather larger samples
than we have tended to think.

Judge et al. suggest that interval transects, across the grain of major
ecological zones, are a good way to begin a regional survey. This sounds
good for most situations. One obvious exception is where tributaries join
a main stream at roughly equal intervals. Interval transects might either miss
the side valleys and junctions, or else overemphasize them. Common sense
should indicate whether this is likely to be a problem in specific instances.

Three technical points in Judge et al.'s chapter need correction. The dis­
tribution of sample means does not approach normality when many samples
are taken unless either the population distribution itself is normal or the sizes
of individual samples are fairly large. The Law of Large Numbers does not
require a normally distributed population. Simple random sampling also does not require that population parameters have a normal distribution.

Matson and Lipe’s work (chap. 7) looks first rate and is certain to provide valuable experience. My only reservation is that I wonder whether their plans to cover about 1.75 square km. in each of several 25 square km. drainages will prove adequate. I am mildly uneasy about this, and await further results.

Redman (chap. 8) offers generally excellent advice, although he leans a little heavily on 10% samples. He gives an example of results from a 10% sample of 1850 obsidian blades, where the sample mean and standard deviation were nearly identical to those from the population. This is so, but note that the standard error is about 3 times as great. It is .49 mm. for the population (thinking of the 1850 blades as themselves a sample), but 1.48 mm. for his sample. In this case the gain in precision is probably not worth the effort needed to measure an additional 1665 blades. Nevertheless, precision requires more explicit attention than it received in this example.

Brown describes efforts to make the best of what remain very refractory problems. For deeply buried strata, both probability sampling and any well-informed purposive selection are practically impossible. I think his strategies alleviate the situation, but the best thing to be hoped for is some drastic breakthrough in sensing devices which will greatly increase the feasibility of “seeing” deeply buried phenomena. Otherwise, there is still no satisfactory alternative to the heavy investments required to clear large overlying volumes without irresponsible destruction of important data in these upper layers.

Asch’s (chap. 10) general discussion is lucid and cogent, and I agree with most of what he says. However, he is pessimistic about the feasibility of getting probability samples large enough for acceptably precise estimates. Some readers will be disconcerted by the confidence limits he presents, since in several cases they imply that there is a real possibility that the true frequency of sherdsof segments in the Macoupin site is a negative number. This is a consequence of having (apparently) used the t-distribution, which presupposes normality of the population distribution. It is true that confidence limits based on some distribution-free statistic (such as recommended by Matson and Lipe) would be more meaningful here, but I am certain that Asch’s point would stand: many of his subsamples would still yield extremely imprecise estimates. Positive skewing in the population means that the lower limit of a confidence interval computed on the assumption of population normality will be too low, but the upper limit will also be too low, and the net effect is generally an underestimate of the total range of the confidence interval (Cochran 1963: 40).

Nevertheless, I think much of Asch’s pessimism is unjustified. It seems overoptimistic to have hoped to get useful estimates of internal differences in sherd frequencies in a site covering over 5 acres (about 2.16 hectares) by putting in 155 test pits, each 2.5 x 5 feet (12.5 square feet, or 1.16 square meters). This adds up to a total area of 1937.5 square feet (180 square meters), the equivalent of a single large square 44 feet on a side. The sampling fraction is .83%. Asch does not give the number of sherds recovered. His overall mean of 50.68 for the 155 pits implies a total of 7,855. Means

for the 24 individual strata imply totals per stratum ranging from 56 to 1,490, and a grand total of 13,580. It is not asking too much to suggest that we must and also can, excavate considerably larger areas and deal with several times as many sherds in order to understand 5 acre sites.

It should also be noted that in this case the precision can increase by much more than the square root of the increase in sample size. This is because in the examples of really wide confidence intervals that Asch gives, much of the trouble comes from the fact that the population variance, as well as the population mean, has to be estimated from the sample statistics. Six of his 24 strata consist of 4 or fewer pits, while another 6 strata include only 5 or 6 pits. Much of the imprecision in the estimates of subpopulation means comes from the imprecision in estimates of subpopulation variance, which makes the t-distribution much broader than the corresponding normal distribution. For sample sizes less than about 6 or 7, the precision increases with increased sample size at a rate much faster than the square root. I believe Asch could quadruple the sampling rate in strata where either the small number of original pits or unexpected heterogeneity has led to very wide confidence intervals, and yet get very substantial improvements in precision within strata by digging less than another 155 pits—although in fact a still larger sample would probably be desirable. Furthermore, I think this could be done without invalidating the assumptions of simple random sampling within strata, and without complicating the analysis or interpretation of results.

One further disagreement with Asch concerns his suggestion that we can consider discarding or stripping off material in the plow zone. This layer, however disturbed, may contain unique information about the latest phase of occupation, and we surely risk gross errors by ignoring it altogether. Finally Asch seems to include certain idiographic inferences about population parameters under the heading of “descriptive” statistics, while I have understood this term to refer strictly to the summarization of information explicitly present in the sample observations.

In sum, Asch’s is a good paper. I agree with most of his discussion, and his experiences at the Macoupin site are instructive. I agree that we should not blindly expect probability sampling to take the place of judicious selection in all situations. But I think he expected far too much from a probability sample whose absolute size (quite apart from the sampling fraction) was small relative to the interpretive burden it was asked to bear.

Morris (chap. 11) presents a complex strategy for a very large and complex site. It seems to be an excellent pragmatic blend, in several stages, of purposive selection and probability sampling. It is the only paper in this volume which grapples with excavations of the scale and complexity required for the study of truly urban settlements. It seems likely that possible questions about the strict applicability of various computational models will be more than offset by the clarity of the patterns which are emerging. Certainly this is a major advance in the archaeology of complex societies, and will provide experience which cannot be wholly duplicated by regional surveys or excavations in smaller sites.

Benfer’s chapter 13 is mainly about multivariate analysis, rather than sampling. He recorded some 60 variables for each of 26 obsidian blade
fragments in a “nonprobability opportunistic collection” from a larger collection of over 50,000 blades. Aside from the unresolvable questions about the implications for the population of any results based on an opportunistic sample, the tiny sample size would create serious difficulties even if it had been picked by probability sampling. A feature has to be present on at least 2 or 3 thousand blades in the collection of 50,000 in order not to have a good chance of being absent on all blades in a random sample of size 26. So even moderately unusual features may be totally absent in the sample. And since these features are the direct objects of investigation we cannot use them as a basis for stratifying the population and improving the chances of their inclusion. Also, Benfer concentrates on disproving null hypotheses of the kind which state that there are no patterned relationships or well-defined subgroups in his population. He finds few relationships in his sample which are so strong that they could not easily be due to chance. But, as I took some pains to point out earlier, with a sample this small it is also extremely difficult to disprove alternative hypotheses about the population. Benfer’s failure to find strong evidence for clusters or subgroups in the sample does not prove that they are not there and important in the population. It is an inconclusive result, suggesting that there are no immensely strong associations among features common in the population. But there may be moderately strong associations among common features and possibly even very strong associations among moderately uncommon features. One wonders, incidentally, whether any associations strong enough to have attained high statistical significance in a sample of this size would not also have been “obvious” to any archaeologist handling a larger number of the blades and using old-fashioned intuitive and vaguely defined techniques of pattern recognition. Multivariate techniques are very useful when the raw data are too numerous or relationships are of kinds not easily perceived by inspection, but they do not somehow squeeze rich understanding out of tiny samples.

In sum, reflecting on all these papers, they confirm my prior conviction that probability sampling strategies are very useful, but that they should never be applied naively or with a blind faith that they will guarantee “scientific” certainty to our conclusions. Often the best strategies are multistage designs with intelligent mixtures of purposive selection and probability sampling. Further, a number of the case studies suggest to me that we have tended to hope for somewhat too much from quite small probability samples. The size and complexity of the sample required depends very strongly on the kinds of questions we ask. For rather simple questions fairly small samples will serve, but for many of the questions we are asking today, especially concerning internal structure and systemic aspects of regions or sites, the samples need to be rather large. The implication is neither that we should forget about probability sampling and rely wholly on intuition, nor that meaningful research is impossible. Rather, it is that statistical expertise can make our research more efficient, but it cannot enable us to work miracles with tiny budgets. These papers help to document the upper limits of what can be achieved by small projects, but they also demonstrate some of the minimal levels of support we need for effective work, however well-designed.
References


Bailey, Vernon, 1913. Life Zones and Crop Zones of New Mexico. United States Department of Agriculture, Bureau of Biological Surveys, North American Fauna, No. 35.


REFERENCES


REFERENCES


REFERENCES


REFERENCES


REFERENCES

(Spaulding. Continued)
—. 1971. Prehistoric Subsistence-settlement Patterns of the Reese River Valley, Central Nevada.
(Thomas, Continued)
Ph.D. dissertation, University of California, Davis.

