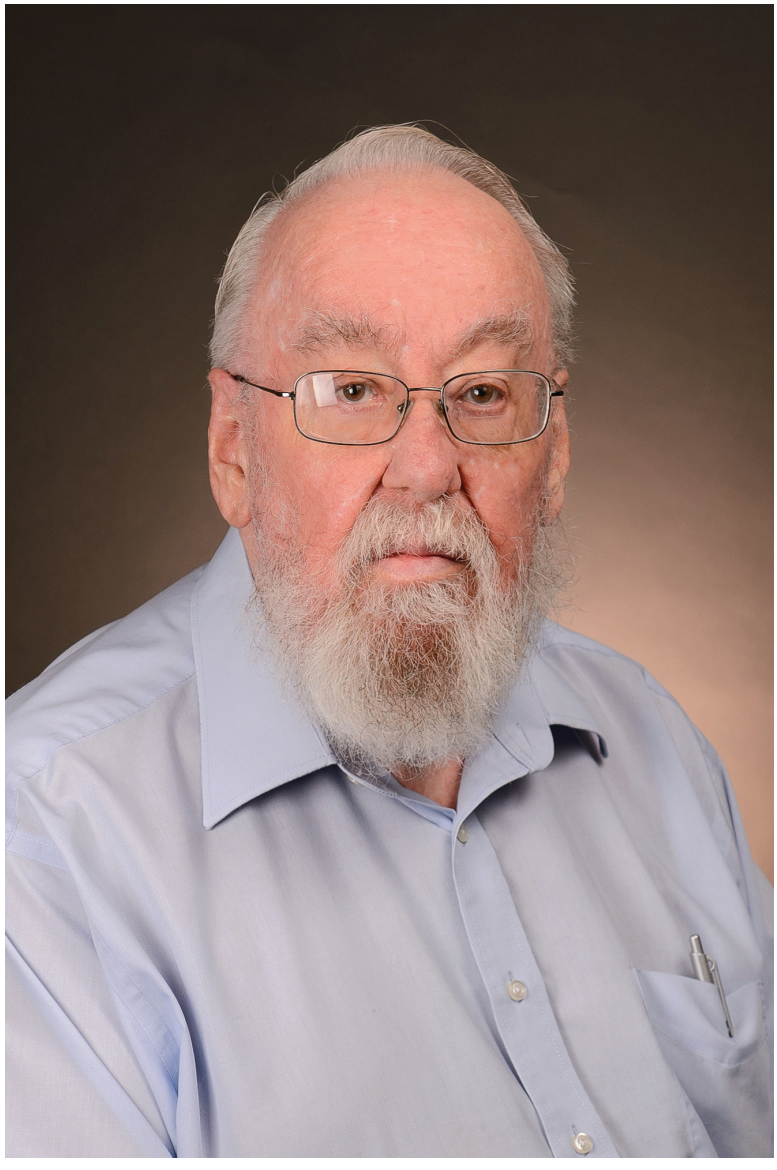

George L. Cowgill

Some Things I Hope You Will Find Useful Even if Statistics Isn't Your Thing

George L. Cowgill

School of Human Evolution and Social Change, Arizona State University, Tempe, Arizona 85287-2402; email: cowgill@asu.edu

Annu. Rev. Anthropol. 2015. 44:1–14

First published online as a Review in Advance on June 5, 2015

The *Annual Review of Anthropology* is online at anthro.annualreviews.org

This article's doi: 10.1146/annurev-anthro-102214-013814

Keywords

statistics, estimation, settlement patterns, Bayes's rule, chronology

Abstract

I emphasize some common misuses of statistics that everyone, whether you do statistics or just read what others write, should be on the lookout for. I next discuss somewhat more complicated issues in archaeological method and theory and then conclude with a qualitative explanation of Bayesian methods and why they are often preferable to the frequentist methods advocated in many introductory statistics texts.

BACKGROUND

In 1948 I set out to be a physicist (BS from Stanford University in 1952 and a MS from Iowa State College (now University) in 1954). Doing so involved numerous courses in mathematics, including calculus, differential equations, and matrix algebra. My math skills were adequate for physics but not outstanding, and, after a career-changing experience digging in North Dakota in the summer of 1952, in 1954 I switched to anthropology (AM from the University of Chicago in 1956 and a PhD from Harvard University in 1963). Readers interested in more of my background should see Cowgill (2008) and Cowgill (2013a). I soon saw that most anthropologists and archaeologists were not adept at math—perhaps in many cases being drawn to a field where math could be avoided. Much of my anthropological thinking has not been mathematical and my archaeological research has focused on sociopolitically complex polities, especially in Mesoamerica and particularly at the great ancient city of Teotihuacan in highland central Mexico (Cowgill 2015a). But my math background made it easy for me to deal with elementary statistics and not too hard to deal with statistics at an intermediate level, though my knowledge is far from a professional level. I found myself teaching introductory courses in statistics for archaeologists and anthropologists. I soon discovered that trying to get important and useful statistical ideas across to people who did not have a mathematical bent and had limited prior math schooling was a huge and frustrating task.

Few archaeologists and probably fewer sociocultural anthropologists enjoy mathematics, and most make little use of formal statistics. This limitation is likely to persist as long as there are no penalties for failing to make use of methods for statistical inference, and even outright blunders often go uncriticized. I hope there will increasingly be rewards for statistical competence and widespread penalties for incompetence. However, for the foreseeable future, most will understand little about statistical techniques or the logic behind them. But even those who do not actively work on these topics can hardly avoid evaluating publications that are based on statistical methods. Therefore, I begin with some simple but important advice that I hope will be widely understood and widely heeded. I hope many people will read the first parts of this article, even if they drop out when things get harder. Most of my examples are drawn from archaeology, but everything I write applies to other fields of anthropology and to everything else.

Humans have some amazing mental abilities, such as recognizing faces, which we do without needing to be taught how to do it. Inability to recognize faces is considered pathology. There is no such thing as a Recognizing Faces 101 course, and I doubt very much if there is a book called anything like *Face Recognition for Dummies*.[1] But we have not evolved good statistical skills; e.g., if a coin comes up heads 10 times in 10 tosses, it is obvious that we should be very skeptical of the claim that the coin is honest (i.e., tails just as likely as heads on any one toss). And if another coin comes up heads 6 times in 10 tosses, it is probably obvious that we have no compelling reason to doubt that it is honest. But what if it comes up heads 8 times? I suspect many readers would not know what to think. Could that easily happen with an honest coin, or is it an improbable result? I did not know what to think until I made a very simple pencil and paper computation, using something called the binomial theorem, to find that the probability of exactly 8 heads in 10 tosses of an honest coin is 45/1,024, or about 0.044.[2] It would be better to rephrase the question as the probability of getting 8 or more heads, or 2 or fewer heads if the coin is honest, which turns out

[1] In fact, statistician Herman Chernoff (1973) actually proposed human faces as a device for displaying several attributes of multiple cases in a single picture.

[2] If the probability of heads on any one toss is 1/2, then the probability of exactly 8 heads in 10 independent tosses is $[10!/(8! \times 2!)](1/2)^{10}$, where n! is called n factorial and means $n \times (n - 1) \times (n - 2) \ldots \times 2 \times 1$.

to be about 0.11. So, what should one decide? The sensible answer, as I discuss further below, is that the result is ambiguous—improbable enough to raise doubts but not improbable enough to feel sure that the coin is dishonest. We have not removed ambiguity; we have only become clear on what the odds are, if our assumption of honesty is correct. That, in itself, is useful to know.

Of course, this is not the kind of problem our early ancestors would have often encountered. But even in more realistic situations we are not innately skilled at determining odds. We evolved to be instinctively amazingly skilled at many other mental things, but maybe there was survival value for our ancestors in not knowing the odds they often faced. That is a question I pose but will not try to answer.

SOME DATA AND OPINIONS ABOUT THE USE OF STATISTICS BY ARCHAEOLOGISTS

The use of statistics and other quantitative methods in archaeology has gone through several stages. Before the 1950s, its use was rudimentary and rarely consisted of more than tabulating some counts and (generally) making sure the numbers added up correctly. Beginning in the 1950s, archaeologist George Brainerd (1951), together with statistician W.S. Robinson (1951), proposed a formal method for seriation of artifact assemblages, in place of the trial-and-error methods formerly used. Albert Spaulding (1953) became a strong voice for statistics and published examples, mainly using chi-square statistics. In the 1960s, devotees of the "new archaeology" saw statistics as a major way forward. But far too often their attempted applications were seriously flawed—what I like to call the "you can't help but admire someone who can spell Tuesday, even if he can't spell it right" syndrome, after something in *Winnie-the-Pooh* (Milne 1926). Disenchantment ensued and even, especially among some postprocessualists, downright aversion. We are now in a phase where enthusiasm is rising for computer-intensive systems and modeling approaches. It may be too soon to say how productive this trend will be. But, in any case, it will always be the domain of specialists. For the most part, except a little toward the end of this article, I am not addressing the hotshots at the cutting edge, although I would not be surprised if I have a few unexpected things even for them. But I am writing mainly for those who have had maybe one statistics course that left them uncomfortable. I hope to encourage these readers by concentrating on simple things that every archaeologist should know—opportunities they should be aware of and pitfalls they should avoid.

There were 334 abstracts of sessions at the 2014 annual meeting of the Society for American Archaeology. Of these session abstracts, five (1.5%) include some word related to statistics. There were roughly 3,025 abstracts of individual papers and posters. Of these, 73 (2.4%) contain some mention of statistics. It is unlikely that many papers or posters that made more than minor use of statistics failed to mention this in their abstracts, so it seems that more than 97% of the presenters made little or no use of statistics in the work they presented. Of the 73 abstracts that do mention statistics, 33 are very vague. Another 9 refer vaguely to multivariate statistics. Seven refer to significance testing. Eight refer to Bayesian methods. Another 21 mention other specific techniques, including t-tests, analysis of variance, regression or correlation, chi-square, autocorrelation, principal components, cluster analysis, correspondence analysis, biodistance, and a few others. Only two mention estimation, and three criticize misuses of statistics or at least mention a need for improvement. There seems to have been no presentation that dealt with the teaching of statistics.

In early 2014 I sent a few questions about the teaching and use of statistics to a small very nonrandom sample of archaeologists I knew. There were no surprises. There is a general perception, which I share, that emphasis on statistics has declined since the 1970s, when it was often associated with processual theory.

Some US universities encourage archaeology graduate students to get some statistical training, but others do not, and only a few require it. Very rarely is there any pressure, or even opportunity, to take more than a one-semester course. In general, statistical learning is given little priority, which creates a self-perpetuating situation. I summarize here comments from respondents to my little survey. At the University of California, Berkeley, statistics is encouraged but not required. At the University of California, San Diego, students must take either statistics or geographic information systems (GIS), a requirement generally fulfilled in another department by advanced undergraduates. The archaeologists would love to develop a course in quantitative analysis within the department; however, other subdisciplines are not interested, and there is no hope of such a course. At Boston University (Archaeology Department), undergraduate majors in archaeology are required to take an introductory statistics course, but at the graduate level it is encouraged only for some students. At Arizona State University there is a somewhat flexible requirement. At the University of California, Los Angeles, graduate students are encouraged but not required to take a course in statistics. At the University of Pittsburgh, two statistics courses are required for archaeology graduate students. At the University of Michigan, a course on analytical methods is required, as is one course outside the department of anthropology. At Statistical Research, most in the Cultural Resource Management (CRM) field come with some specialized skill and know how and when to apply specific quantitative techniques, but only a few understand the underlying statistical theory. They have addressed the lack of quantitative expertise by hiring specialists in this field. At Brigham Young University, statistics is considered essential. One course is required of undergraduates and graduates. At Tulane University, two courses are required: a basic course for all anthropology undergraduates and graduates and an advanced course for archaeologists. At Yale University, all graduate students specializing in archaeology are required to take some sort of statistics course. At Purdue University, a four-field quantitative course at the MA level has been required but is no longer required of cultural anthropology students, which means the course will die for lack of demand. Some students are encouraged to take additional quantitative courses. Younger faculty are less knowledgeable about quantitative methods and do not expect their students to develop these skills.

SIGNIFICANCE TESTING VERSUS ESTIMATION

Probably the most common misuse of statistics is using some test of significance to decide whether some feature or relationship in a sample is really true of the larger population from which the sample was drawn or whether the feature or relationship could easily occur in the sample as an accidental result of sampling vagaries.[3] Typically this computation is carried out by setting up the null hypothesis that the feature or relationship does not exist in the population, and then computing the probability that, even if so, the feature or relationship would nevertheless occur in the sample. Typically the result is considered significant if this probability is less than 5%, in which case the sample is taken to be true of the population, whereas any probability greater than 5% is considered to be easily explainable as due to sampling vagaries and thus not convincingly true of the population. There used to be a practical reason to rely on levels such as 5% and 1% and 0.1%. They were often the only ones tabulated in printed tables. But now, electronic packages will often compute exact significance levels.

[3] "Sampling vagary" is a felicitous term I adopt from Drennan (2009), in preference to "sampling error," which is somewhat misleading.

As I have said, however, statistical analysis is not a way to arrive at certainty; it is a powerful aid in discerning what your data suggest and how strongly your data suggest it. This task is often done more effectively with an estimation approach rather than by hypothesis testing. Above all, do not mindlessly rely on some significance level (p-value) for assurance that a hypothesis is true beyond reasonable doubt or false beyond reasonable doubt. You must take into account many additional things, including the size of the sample, how representative the sample is of the population about which you want to make inferences, and all the tacit assumptions built into the test you are using, as well as relevant prior knowledge. Do not use $p = 5\%$ as a talisman to replace thinking but instead realize how arbitrary it is.

One difficulty with hypothesis testing is that strength of the effect being tested is a somewhat complicated concept and is often ignored. Does the sample suggest that the population differs only slightly from the null hypothesis value, or does the sample suggest that the population differs a great deal from the null hypothesis? Also, you should take into account the size of your sample—that is, how many independent cases are in your sample? With a small sample, you will not get a small p-value (i.e., a significant result) unless the sample statistic differs a great deal from the null hypothesis value. With a large sample, you will get a small p-value even if the sample statistic differs so little from the null hypothesis value that the difference is unimportant.

These difficulties can be largely avoided by computing a confidence interval, which is usually a bell-shaped curve with a single peak and low "tails" on either side. The height of the curve at any point gives you an idea of how likely that point might be the true value in the population. A confidence interval consists of the range of possible population values that includes some specified proportion of the total probability, such as 95%. Getting a 95% confidence interval that just fails to include the null hypothesis value is equivalent to getting a 5% p-value, but the confidence interval tells you much more than that. It suggests that the true population value is close to the peak of the sample's confidence interval. In addition, a wide confidence interval indicates that your sample is quite ambiguous, whereas a narrow confidence interval means you can feel pretty confident that the population value is pretty close to the sample value. With a large sample, a weak relationship may be highly significant but not very important. With a small sample, a result with low statistical significance may suggest a strong relationship in the population and should be followed up with a larger sample, if that is possible.

R. D. Drennan's excellent introductory statistics text for archaeologists ably explains why estimation is preferable to significance testing and explains how to do estimation (Drennan 2009). Steven Shennan's (1997) introductory text, also excellent in most respects, mentions estimation but gives more emphasis to hypothesis testing and significance levels. He says, "By convention, the two most commonly used significance levels are. . .0.05 and. . .0.01" (Shennan 1997, p. 53), but he does not dwell on how arbitrary those levels are or why they should not be blindly used as a basis for "yes/no" decisions.

Estimation is routinely used in reporting radiocarbon dates in 1-sigma (68% confidence interval) and/or 2-sigma ranges (~95% confidence interval), so it is not something unfamiliar to archaeologists. But estimation can and should be used much more widely in addressing other topics.

OTHER STATISTICAL THINGS TO REMEMBER, WHATEVER ELSE YOU FORGET

I expand on a little article that I published as a sidebar in the *Archaeological Record* (Cowgill 2005). I suspect that few readers paid much attention to it. I hope it will have more impact here. This list of topics does not cover everything you need to know. It just lists some issues I am especially aware of.

Darrell Huff's little book, *How to Lie With Statistics* (1954), has a light-hearted title but is full of good advice, both for spotting misleading presentations in others' publications and for doing your own good job. I also recommend Drennan (2009), which is excellent in most respects except that it never mentions Bayesian methods, and Shennan (1997), also excellent except for saying little about estimation.

First, just look at your data using simple tables and pictures. Edward Tufte (1983) published a classic book on the visual presentation of data, and he and others have published several good books about this since then. Among other things, he coined the phrase "chart junk." Do not make your pictures fancier than they need to be. Do not clutter them with extra bells and whistles just because you have graphics packages that encourage you to use them. In particular, if you need only two dimensions to show something, do not add a needless third dimension. It may look sexier, but it often makes it difficult to tell what the actual values are. Also, please avoid pie charts. Drennan (2009, pp. 73–74) shows how much easier it is to see the patterns using bar charts. For tables, if your computer obligingly turns out numbers such as 0.236795 and 0.375622 but you know that only the first two digits are meaningful, please round them to 0.24 and 0.38. Not only does this save space and allow you more room for discussion before exceeding limits set by your publisher, but it also makes the tables much easier to read and it is much easier to spot the really sizable differences that matter most. Box plots, described in many texts, are also an excellent way of visually displaying results if you do not make the plots too complicated. By looking at pictures of two variables with one plotted on the horizontal axis and the other on the vertical axis, you can see whether the plotted points deviate so much from a straight line that computing a simple correlation coefficient would be very misleading.

Often just making good tables and good pictures tells you everything that is important. If they do not, they will tell you what is sensible or not sensible to do next. Do not rush to apply advanced techniques while overlooking the messages of simple methods.

It is not the sampling fraction that matters; it is the size of the sample. A well-chosen sample with 300 cases that is 1% of a large population can tell you a lot, but a sample with 10 cases that is 20% of a small population will tell you little. Actually, for reasons I explain below, it is usually best to assume that the population is infinitely large, which is in fact the usual assumption. It is hardly ever advisable to make any small population modification.

Archaeologists can rarely choose the sizes of their samples. But we can, and should, think ahead of time about how large a sample would be needed to get a reasonably clear answer to specific questions. Often this will amount to thinking about how narrow a confidence interval you would like. This desirable sample size will often turn out to be much larger than you can get because of limits on what is available or funding limitations. Your desirable sample size will often turn out to be much more than 100 cases, and likely 300 or more. This is especially so if you want to address subsets of the total collection, such as chert versus obsidian artifacts. I hope that the recognition of the need for sizable collections can create a culture of higher requirements for funding. By all means, do not believe that a sample consisting of 100 cases will often be sufficient.

Proportions, percents, and ratios represent something relative to something else. They are all fractions, with a numerator and a denominator. Always report the denominator. Sometimes the reader can guess what the denominator is, but you should never require the reader to be a detective and figure it out. A numerator standing alone without a denominator is an anomaly, like the sound of one hand clapping.

Some people argue that frequency always means count, but in practice this is just not so. Some people use it to mean percentage. Often it is not clear whether frequency means a count or a percentage. I urge you to avoid the term. Instead, say explicitly whether you mean a count or a percentage.

If you are worried about data quality, reducing data to "present" and "absent" just makes the problem worse, unless you are sure that absence in the sample unambiguously implies absence in the relevant population. But a category that is scarce but present in the population will be totally absent in many samples from that population. The chance that it is absent in any one sample strongly depends on the size of that sample. This makes "presence versus absence" a very unstable statistic. If you want to be intentionally vague and conservative, it would be much better to use terms such as "way below average," "about average," and "way above average."

Do not rely too much on dendrograms, so-called tree diagrams. They are appealingly simple and they can be a good way to get started, but beware of them as the last word. They make too little use of all the data, and they can be arbitrary and misleading. For discerning groups and subgroups of cases, discriminant analysis and other multivariate methods are much more powerful and less likely to be misleading. If you must stop with dendrograms, try more than one similarity coefficient and try more than one clustering method. See if you get basically similar results with different methods.

A "NO FALSE MODESTY" BIOGRAPHICAL INTERLUDE

Over the years I have published a few articles on statistical or other math topics, as well as on many other topics. The technical level of these writings varies, but I hope many readers will find some of them useful. Among those most relevant is a paper I published in 1964 that I thought was a nifty little application of the binomial theorem, showing how just a few pencil-and-paper computations settled a controversy among cultural anthropologists, naively expecting they would catch on and start making good use of the theorem for other purposes (Cowgill 1964). Nothing of the sort happened. Those on one side of the debate just used it as ammunition, while the other side ignored it.

The following is a list of additional statistical publications I think are still worth reading: Cowgill (1970), which sketches some taphonomic issues similar to those Michael Schiffer (1976) subsequently dealt with much more elaborately; Cowgill (1972), which discusses major concepts, techniques, and issues in chronological seriation, pointing out, among many other things, that it is a fallacy to think that all units must have about the same duration (some faulty notation was inserted by an editor without my knowledge); Cowgill (1975a), which reviews sampling in archaeological survey; Cowgill (1977), which describes problems with traditional significance testing, now largely superseded by Bayesian methods; Cowgill (1989), which provides a superior method for discerning ancient tool kits and activity areas that avoids the inherent problems of k-means and other clustering methods pointed out by Baxter (2003, p. 7); Cowgill (1990a), which is a lengthy reflection on artifact typology; Cowgill (1990b), which discusses some inconvenient truths about survey that I suspect are often ignored; Cowgill (2006), which provides an example of wringing convincing conclusions from a large body of messy data; and Cowgill et al. (1984), written with Jeffrey Altschul and Rebecca Sload, which describes a variety of quantitative methods applied to ancient Teotihuacan in central Mexico.[4]

WHAT IS THE RELEVANT POPULATION?

Sometimes the target population is considered to be only the cases of some phenomenon that actually exist or existed, such as all the houses that ever existed in some settlement. In this situation,

[4]More publications are available at **https://sites.google.com/a/asu.edu/george_cowgill/home**.

one might think that estimates based on a sample can be made more precise by a small population correction. But, in most cases involving theory, this is not the population of interest. The really interesting population is all the cases that might have existed if a particular theory were correct, or if some proposed set of cultural practices were being followed, and this population is indefinitely large. Shennan (1997, p. 64) aptly says that in many contexts "we might conceive of the evidence as one specific empirical outcome of a system of behavior based on social rules." Unless otherwise stated, all the common statistical procedures tacitly assume an infinite population. Never use a finite population correction unless you have a compelling argument for using it in a special case, and do not assume you can avoid estimation just because you think you have nearly all the cases that ever existed.

TESTING RELIABILITY AND VALIDITY

In many research fields there is something called meta-analysis, which is carried out by making a comparative analysis of the statistical results of multiple studies of a topic, assessing their quality, investigating discrepancies, and obtaining conclusions that are more reliable and more accurate than those of any one study. As far as I know, there is no such thing in archaeology or anthropology, although some publications, such as Inomata et al. (2014), amount to an informal meta-analysis. To be sure, not all sites can be redug or resurveyed. Some sites can be, and even when that is not possible, curated collections can be restudied; the frequent need for restudying them is a major reason why it is essential to make collections and is irresponsible to imagine that recording observations without making collections is adequate.

In fact, archaeologists have been astonishingly ready to assume that their findings are highly reliable (repeated studies would get almost the same results) and very valid (their findings are a nearly unbiased sample of the population of interest). Leaving these assumptions unchallenged is nothing short of a conspiracy of silence, and it is not going too far to call it a dirty little secret. Over the years I have compiled a bibliography of more than 4,000 publications on a wide variety of archaeological topics. Among other things, I have been on the lookout for publications reporting studies of reliability and/or validity, wherein sites were recollected or resurveyed or different observers independently classified the objects in a collection. I have located just 20, of which the most recent and most telling is that by Heilen & Altschul (2013). The bad news is that the level of reliability and/or validity is often shockingly low. If you believe your results can be trusted without any checking, you are fooling yourself and doing the profession a disservice.

I have carried out studies of reliability and validity of data acquired by the Teotihuacan Mapping Project, beginning in the 1960s (Cowgill 1968), but many later studies of Teotihuacan data remain unpublished (e.g., Lokaj et al. 1986), although they suggest that in this case discrepancies are not great enough to have any significant effect on interpretations.

MORE ADVANCED TECHNIQUES

Michael Baxter (2003, 2015) is a statistician who has worked closely with archaeologists for several decades. He has a refreshing and (literally) down-to-earth and common-sensical awareness of real archaeological problems and concerns. His books present methods that are useful for archaeologists who have mastered an introductory course in statistics, including many multivariate techniques, among other things. His discussions can be read with profit by beginners as well as by those more advanced.

SETTLEMENT PATTERNS

The pathbreaking work of Gordon Willey in the Virú Valley of Perú (Willey 1953) opened up a very productive line of archaeological research. This settlement pattern research has depended, above all, on good maps, sometimes with statistical techniques largely borrowed from geographers (e.g., Hodder & Orton 1976). I do not feel qualified to discuss much of the recent quantitative work along these lines. But just looking at good maps can be very useful. They are two-dimensional, or, when altitude is taken into account, three-dimensional (e.g., Carballo & Pluckhahn 2007). Peterson & Drennan (2005), Drennan & Peterson (2004, 2008), and Ossa (2013), among other examples, make good quantitative use of two dimensions.

Some archaeologists have relied too much on the number of peaks in one-dimensional histograms of site sizes within a region as a clear indication of whether the regional polity was a so-called "chiefdom" or a "state." I doubt whether there is, in fact, any clear threshold that distinguishes chiefdoms from states. This has been questioned not only by me, but also by Brumfiel (1995) and Yates (1997), among others. However, that is not a statistical issue. The statistical shortcomings are the failure to use more than one dimension and, especially, the fact that detection of peaks in the histogram has usually been based on statistically naïve subjective judgments that the reader is asked to simply accept. A histogram with three peaks is taken to be good evidence of a chiefdom, whereas a histogram with four peaks suggests a state (or, sometimes, two peaks for a chiefdom and three for a state). For example, Flannery (1998, pp. 16–17) acknowledges that it is only a useful rule of thumb but nevertheless accepts it for southwestern Iran. But shifting only two sites to an adjacent bin (vertical bar) in the histogram in Flannery's illustration is enough to make an alleged gap disappear. His figure shows excellent separation of three peaks for Early Uruk. However, for Middle Uruk the really significant change is that Susa doubles in size and becomes far larger than any other settlement in the region. Settlements in what was tier 3 remain in the same size range (0–3 ha) but are now subdivided into tiers 3 and 4 on the basis of a minimum in the 1.5–2-ha range that would disappear if two sites in tier 3 were shifted to tier 4. Two sites that are ca. 10 ha in size might count as the real tier 2, whereas sites in the 5–7-ha range, previously tier 2, might instead be considered tier 3.

In other cases the statistics presented are puzzling, as in Flannery (2002, p. 426), where something called "T" is claimed to be statistically significant. I suspect "t" is meant. But I do not know how t-tests could be relevant for determining the number of peaks in a histogram. By inspection, this histogram suggests three, four, or even more peaks. Another troubling recent example is Spencer & Redmond (2006), thoughtful in many ways but too uncritical of the method of multiple peaks in a histogram.

Zipf (1949) observed that the sizes of sites in a region, if ranked from largest to smallest, often roughly follow the proportions 1, 1/2, 1/3, 1/4, etc., and, if so, their logarithms (whether expressed in base 10 or base e makes no difference) follow a descending straight line. This makes it easy to see deviations from Zipf's generalization. If the largest site is, by far, the largest, i.e., what is aptly called a primate center, the line will be concave. If several of the largest sites are of nearly the same size, suggesting that there is no single strong polity spanning the entire region, the line will be convex. Drennan & Peterson (2004) have made good use of this fact.

Notice that if there are multiple peaks in a plain histogram of site sizes, the logarithmic picture should be a stair-step line, that is, quite different from Zipf's generalization. You cannot have anything close to the Zipf pattern and have well-defined tiers.

I am not sure of the best statistic for testing the probability of various numbers of peaks in a given histogram. Perhaps kernel smoothing estimation, briefly discussed by Shennan (1997, pp. 29–30), would be useful, doing it by experimenting with different bin widths.

PROBLEMS WITH FREQUENTIST STATISTICS AND THE ADVANTAGES OF BAYESIAN METHODS

A traditional approach, inculcated in most statistics texts, is to frame a problem by formulating a hypothesis and calculating the probability of the observed result if the hypothesis is true or by estimating a confidence interval. We would like to estimate the population value for known data, but what we can actually do with traditional methods is to estimate the probability of data for a postulated population, though this is often not comprehended. In fact, it is difficult to wean students away from the idea that this is what their result is telling them. Also, this traditional method (often called "frequentist") does not provide any way of making formal use of any additional information. But in fact we often have considerable additional information, and in practice we are likely to somehow make use of it. For example, in a coin-tossing example, if we have prior reason to suspect the honesty of the tosser, then we are more likely to treat an outcome such as 8 heads in 10 tosses as good reason to be suspicious. But this informal ad hoc reasoning is what I call "folk-Bayesian" (perhaps coining a new expression), and we can often do better by making formal use of Bayes's rule. The notation for this rule uses a vertical slash to indicate that what is on the left side takes what is on the right side as given. For example "probability (I'll take an umbrella to work | I see it's raining)" is different from "probability (I'll take an umbrella to work | it's nice and sunny). Bayes's rule can then be stated as $P(H|D) = [P(D|H) \times P(H)]/P(D)$. $P(H|D)$ is called the posterior probability, the probability of hypothesis H, given that new data (D) are observed. P(H) is the prior probability, the probability of H before the new data are taken into account. It is based on various kinds of other available information. I like to think of P(H) as the "before" probability and $P(H|D)$ as the "after" probability, but the terms prior and posterior are too embedded in the literature to be changed. Notice that $P(D|H)$ is all that traditional frequentist methods allow you to compute. But $P(H|D)$ is what you usually want.

At the 2013 annual meeting of the Society for American Archaeology, the word "Bayes" occurs in the title and/or abstract of 9 papers out of more than 3,000 papers, about 0.3% of all papers. It appears that six of these nine deal only with radiocarbon dating, probably just plugging in a program whose rationale the authors may not understand very clearly. Only three appear to deal with Bayesian methods used for other purposes.

Bayes's rule, as such, is logically impeccable and accepted by all statisticians. If nobody questions Bayes's rule in the abstract, why has it not been universally adopted? I think a main reason is that mathematically expressing P(H), your additional information, is often very difficult. Sometimes prior information can be captured pretty well by something called a beta distribution, a bell-shaped curve that looks something like normal (Gaussian) distribution, but is not quite the same. Iversen (1984) describes its use. It is rather simple and has sometimes been used by archaeologists. But it does not always capture prior information well. Kruschke (2014) describes other methods, including Gibbs sampling, an example of a Markov chain Monte Carlo (MCMC) process, which is much used in improving radiocarbon dating by Bayesian methods. Many of these methods have become feasible only with recent improvements in computer power.

If the prior information is very vague, the improved estimate is based mainly on the sample data, and it amounts to little more than legitimizing the common misinterpretation of frequentist methods. Drennan (2009) does not mention Bayesian methods. Shennan (1997, p. 48) mentions them and considers the Bayesian approach in many ways more attractive in philosophical terms than the frequentist approach but finds it difficult to apply in practical terms, and he discusses only frequentist methods. Buck et al. (1996) covers a range of Bayesian methods but is mathematically more demanding than is Kruschke (2014). McGrayne (2011) is an enthusiastic journalistic account of the history of Bayesian approaches, but it contains no math and does not tell you how to carry

out Bayesian analyses. Efron (2013) is an excellent brief assessment of Bayesian ideas. He suggests using Bayesian analysis in the presence of genuine prior information and advises caution when invoking uninformative priors. "In the . . . [latter] case, Bayesian calculations cannot be uncritically accepted and should be checked by other methods, which usually means frequentistically." The bottom line, then, is that Bayes is not much help if you do not have good prior information. But if you do have good prior information, Bayes is a powerful way to make the most of it.

Probably the best way to learn about Bayes, at least for the present, is to begin with an introductory book on descriptive and frequentist statistics, such as Shennan (1997) or Drennan (2009), and then look into an introductory book on Bayesian methods, such as Kruschke (2014), which is somewhat advanced but includes many computer programs. Iversen (1984) can be useful, though it covers only a limited range of methods.

BAYESIAN RADIOCARBON DATING

In a well-stratified excavation, where there are layers with little or no evidence of redeposition from lower layers or intrusion from higher layers, it is reasonable to assume that the true date of any radiocarbon specimen in a particular layer is not earlier than the true date of any specimen in a lower layer and is not later than the true date of any specimen in a higher layer. Or it can be useful just to be confident that there is very little or no intrusion from later layers, even if redeposition and heirlooms cannot be ruled out. Bayesian methods can be used to combine this information with the radiocarbon readings to provide narrower and better-founded dates for the layers than are provided by the radiocarbon values alone. Bayesian statistical packages to improve chronologies based on a rational combination of calibrated[5] radiocarbon determinations and stratigraphy have been in use for some time in Great Britain (Bayliss 2009) and, after a slow start (e.g., Zeidler et al. 1998), are beginning to catch on rapidly in Latin America (e.g., Beramendi-Orosco et al. 2009, Inomata et al. 2014, Cowgill 2015b, Overholtzer 2015), Polynesia (e.g., Allen & Morrison 2013, Athens et al. 2014), and elsewhere.

A very attractive feature is that results can be shown in pictures that you can readily grasp even if you are not adept at numbers. But it is not clear how well the rationale behind these packages is understood. Furthermore, they can be no better than the quality of the radiocarbon specimens, and a fairly large number of related specimens are needed for the methods to work well, preferably more than about 20, I'd guess. It is important to budget for many high-quality radiocarbon specimens from good contexts, and it is important for reviewers of grant proposals to recognize this. Tighter chronologies do not just refine details; they open up new possibilities in detecting short-term events.

OTHER BAYES APPLICATIONS IN ARCHAEOLOGY

Buck et al. (1996) discuss a number of other applications of Bayesian methods to archaeological problems. Confronted with the problem of small sherd samples from many of the surface collection tracts at Teotihuacan, in the early 1980s I enlisted the aid of statistician Herman Chernoff, who devised an empirical Bayesian method of making the small-sample-size problem less severe. Miriam Chernoff applied this method to a sample of the more than 5,000 tracts in an unpublished paper at a Society for American Archaeology meeting (M. Chernoff 1982). Robertson (1999, 2001)

[5]It would be good if "calibration" were used to refer strictly to corrections made only because ambient radiocarbon changes over the centuries and if some other term were used for Bayesian estimates.

subsequently applied this and other methods to the entire city. Ortman et al. (2007) use an empirical Bayesian approach in a complex effort to make good use of data acquired by inconsistent means by many archaeologists over a long interval in a part of the southwestern United States. There are other Bayesian applications in archaeology, but as yet they are few.

SUMMARY

In recent decades I have found myself increasingly working on topics that rarely required any complicated formal mathematics. Since 1964 I have worked at the great ancient city of Teotihuacan in central Mexico, which flourished between ca. 100 BCE and 600 CE; during much of that time, the city covered about 8 square miles, with a likely population of more than 80,000 (Cowgill 2015a). I have also worked quite a bit on demography, which appealed to me in part because it is mildly quantitative and especially because it is directly relevant to urgent contemporary problems. In 1975 I published a paper that was influential for some years (Cowgill 1975b). However, I was unable to get later demographic efforts published. Much of my work has been on early urbanism, especially the rise and fall of ancient complex societies (Yoffee & Cowgill 1988) and migration (Cowgill 2013b). I am also interested in improving archaeological chronologies (Cowgill 2015b).

In this article, I have mainly stressed simple topics that are unavoidable for everyone who just wants to do good archaeology, regardless of your theoretical and methodological persuasions. My most important advice is, in general, to use estimation and confidence intervals rather than null hypothesis testing and, if you have good prior information, to try Bayesian methods.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

For their thoughtful responses to my questions about the status of quantitative teaching in their institutions I thank Jeffrey Altschul, Richard Blanton, Richard Burger, Geoffrey Braswell, David Carballo, John Clark, Robert D. Drennan, Dan Healan, John Hodgson, Rosemary Joyce, Keith Kintigh, Richard Lesure, and Robert Whallon. I also thank Robert Drennan for useful comments on statistics for settlement patterns. Bradley Efron steered me to Andrew Gelman, who in turn steered me to the Kruschke book.

LITERATURE CITED

Allen MS, Morrison AE. 2013. Modelling site formation dynamics: geoarchaeological, chronometric and statistical approaches to a stratified rockshelter sequence, Polynesia. *J. Archaeol. Sci.* 40:4560–75

Athens JS, Rieth TM, Dye TS. 2014. A paleoenvironmental and archaeological model-based age estimate for the colonization of Hawai'i. *Am. Antiq.* 79(1):144–55

Baxter M. 2003. *Statistics in Archaeology*. New York: Oxford Univ. Press

Baxter M. 2015. *Exploratory Multivariate Analysis in Archaeology*. Edinburgh: Edinburgh Univ. Press. 2nd ed.

Bayliss A. 2009. Rolling out revolution: using radiocarbon dating in archaeology. *Radiocarbon* 51(1):123–47

Beramendi-Orosco L, González-Hernández G, Urrutia-Fucugauchi J, Manzanilla LR, Soler-Arechalde AM, et al. 2009. High-resolution chronology for the Mesoamerican urban center of Teotihuacan derived from Bayesian statistics of radiocarbon and archaeological data. *Quat. Res.* 71:99–107

Brainerd GW. 1951. The place of chronological ordering in archaeological analysis. *Am. Antiquity* 16:301–13

Brumfiel EM. 1995. Heterarchy and the analysis of complex societies: comments. In *Heterarchy and the Analysis of Complex Societies*, ed. RM Ehrenreich, CL Crumley, JE Levy, pp. 125–31. Arlington, VA: Am. Anthropol. Assoc.

Buck C, Cavanagh WG, Litton CD. 1996. *Bayesian Approach to Interpreting Archaeological Data*. Chichester, UK: Wiley

Carballo DM, Pluckhahn T. 2007. Transportation corridors and political evolution in highland Mesoamerica: settlement analyses incorporating GIS for Northern Tlaxcala, Mexico. *J. Anthropol. Archaeol.* 26:607–29

Chernoff H. 1973. The use of faces to represent points in k-dimensional space graphically. *J. Am. Stat. Assoc.* 68:361–68

Chernoff M. 1982. *Empirical Bayes Estimation of ceramic proportions at Teotihuacan*. Presented at Annu. Meet. Soc. Am. Archaeol., Minneapolis, MN

Cowgill GL. 1964. Statistics and sense: more on the Purum case. *Am. Anthropol.* 66:1358–65

Cowgill GL. 1968. Computer analysis of archaeological data from Teotihuacan, Mexico. In *New Perspectives in Archeology*, ed. SR Binford, LR Binford, pp. 143–50. Chicago: Aldine

Cowgill GL. 1970. Some sampling and reliability problems in archaeology. In *Archéologie et Calculateurs*, ed. JC Gardin, pp. 161–75. Paris: CNRS

Cowgill GL. 1972. Models, methods, and techniques for seriation. In *Models in Archaeology*, ed. D Clarke, pp. 381–421. London: Methuen

Cowgill GL. 1975a. A selection of samplers: comments on archaeo-statistics. In *Sampling in Archaeology*, ed. J Muller, pp. 258–74. Tucson: Univ. Ariz. Press

Cowgill GL. 1975b. On causes and consequences of ancient and modern population changes. *Am. Anthropol.* 77:505–25

Cowgill GL. 1977. The trouble with significance tests and what we can do about it. *Am. Antiq.* 42(3):350–68

Cowgill GL. 1989. The concept of diversity in archaeological theory. In *Quantifying Diversity in Archaeology*, ed. R Leonard, G Jones, pp. 131–41. Cambridge, UK: Cambridge Univ. Press

Cowgill GL. 1990a. Artifact classification and archaeological purposes. In *Mathematics and Information Science in Archaeology*, ed. A Voorrips, pp. 61–78. Bonn, Ger.: HOLOS Verlag

Cowgill GL. 1990b. Toward refining concepts of full-coverage survey. In *The Archaeology of Regions: A Case for Full-Coverage Survey*, ed. S Fish, S Kowalewski, pp. 249–59. Washington, DC: Smithson. Inst. Press

Cowgill GL. 2005. Things to remember about statistics (whatever else you forget). *SAA Archaeol. Rec.* 5(4):35

Cowgill GL. 2006. Using numerous cases to extract valid information from noisy surface data at Teotihuacan. In *Managing Archaeological Data: Essays in Honor of Sylvia W. Gaines*, ed. JL Hantman, R Most, pp. 147–54. Tempe: Ariz. State Univ. Anthropol. Res. Pap. 57

Cowgill GL. 2008. How I got to where I am now: one thing after another, a (mostly) linear narrative. *Anc. Mesoam.* 19(2):165–73

Cowgill GL. 2013a. Conversation with William L. Rathje and Michael Shanks. In *Archaeology in the Making: Conversations through a Discipline*, ed. WL Rathje, M Shanks, C Witmore, pp. 185–203. New York: Routledge

Cowgill GL. 2013b. Possible migrations and shifting identities in the Central Mexican Epiclassic. *Anc. Mesoam.* 24(1):131–49

Cowgill GL. 2015a. *Ancient Teotihuacan: Early Urbanism in Central Mexico*. New York: Cambridge Univ. Press

Cowgill GL. 2015b. We need better chronologies: progress in getting them. *Latin Am. Antiq.* 26:26–29

Cowgill GL, Altschul JA, Sload R. 1984. Spatial analysis of Teotihuacan: a Mesoamerican metropolis. In *Intrasite Spatial Analysis in Archaeology*, ed. H Hietala, pp. 154–95. Cambridge, UK: Cambridge Univ. Press

Drennan RD. 2009. *Statistics for Archaeologists: A Common Sense Approach*. New York: Springer. 2nd ed.

Drennan RD, Peterson CE. 2004. Comparing archaeological settlement systems with rank-size graphs: a measure of shape and statistical confidence. *J. Archaeol. Sci.* 31:533–49

Drennan RD, Peterson CE. 2008. Centralized communities, population, and social complexity after sedentarization. In *The Neolithic Demographic Transition and Its Consequences*, ed. JP Bocquet-Appel, O Bar-Yosef, pp. 359–86. New York: Springer

Efron B. 2013. Bayes' theorem in the 21st century. *Science* 340:1177–78

Flannery KV. 1998. The ground plans of archaic states. In *Archaic States*, ed. GM Feinman, J Marcus, pp. 15–57. Santa Fe, NM: Sch. Am. Res. Press
Flannery KV. 2002. The origins of the village revisited: from nuclear to extended households. *Am. Antiq.* 67(3):417–33
Heilen M, Altschul JH. 2013. The accuracy and adequacy of in-field artifact analysis. *Adv. Archaeol. Pract.* Nov:121–38
Hodder I, Orton C. 1976. *Spatial Analysis in Archaeology*. London: Cambridge Univ. Press
Huff D. 1954. *How to Lie With Statistics*. New York: Norton
Inomata T, Ortiz R, Arroyo B, Robinson EJ. 2014. Chronological revision of Preclassic Kaminaljuyú, Guatemala: implications for social processes in the Southern Maya Area. *Latin Am. Antiq.* 25(4):337–408
Iversen GR. 1984. *Bayesian Statistical Inference*. Beverley Hills, CA: Sage
Kruschke JK. 2014. *Doing Bayesian Data Analysis*. New York: Elsevier. 2nd ed.
Lokaj J, Chiarelli JA, Cowgill GL. 1986. *The Reliability of Surface Collecting at Teotihuacan*. Presented at Annu. Meet. Soc. Am. Archaeol., New Orleans
McGrayne SB. 2011. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. New Haven, CT: Yale Univ. Press
Milne AA. 1926. *Winnie-the-Pooh*. London: Methuen
Ortman SG, Varien MD, Gripp TL. 2007. Empirical Bayesian methods for archaeological survey data: an application from the Mesa Verde Region. *Am. Antiq.* 72(2):241–72
Ossa AE. 2013. Using network expectations to identify multiple exchange systems: a case study from Postclassic Sauce and its Hinterland in Veracruz, Mexico. *J. Anthropol. Archaeol.* 32:415–32
Overholtzer L. 2015. Agency, practice, and chronological context: a Bayesian approach to household chronologies. *J. Anthropol. Archaeol.* 37:37–47
Peterson CE, Drennan RD. 2005. Communities, settlements, sites, and surveys: regional-scale analysis of prehistoric human interaction. *Am. Antiq.* 70(1):5–30
Robertson IG. 1999. Spatial and multivariate analysis, random sampling error, and analytical noise: empirical Bayesian methods at Teotihuacan, Mexico. *Am. Antiq.* 64:137–52
Robertson IG. 2001. *Mapping the social landscape of an early urban center: socio-spatial variation in Teotihuacan*. PhD Diss., Dep. Anthropol., Ariz. State Univ., Tempe
Robinson WS. 1951. A method for chronologically ordering archaeological deposits. *Am. Antiq.* 16:293–301
Schiffer MB. 1976. *Behavioral Archeology*. New York: Academic
Shennan SJ. 1997. *Quantifying Archaeology*. Iowa City: Univ. Iowa Press
Spaulding AC. 1953. Statistical techniques for the discovery of artifact types. *Am. Antiq.* 18(4):305–13
Spencer CS, Redmond EM. 2006. Resistance strategies and early state formation in Oaxaca, Mexico. In *Intermediate Elites in Pre-Columbian States and Empires*, ed. CM Elson, RA Covey, pp. 21–43. Tucson: Univ. Ariz. Press
Tufte ER. 1983. *The Visual Display of Quantitative Data*. Cheshire, CT: Graphics Press
Willey GR. 1953. *Prehistoric Settlement Patterns in the Virú Valley, Peru*. Washington, DC: Smithson. Inst.
Yates RDS. 1997. The city state in ancient China. In *The Archaeology of City-States: Cross Cultural Approaches*, ed. DL Nichols, TH Charlton, pp. 71–90. Washington, DC: Smithson. Inst.
Yoffee N, Cowgill GL, eds. 1988. *The Collapse of Ancient States and Civilizations*. Tucson: Univ. Ariz. Press
Zeidler JA, Buck CE, Litton CD. 1998. Integration of archaeological phase information and radiocarbon results from the Jama River Valley, Ecuador: a Bayesian approach. *Latin Am. Antiq.* 9(2):160–79
Zipf GK. 1949. *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley