International Journal of Osteoarchaeology
Int. J. Osteoarchaeol. **26**: 255–265 (2016)
Published online 6 September 2014 in Wiley Online Library
(wileyonlinelibrary.com) **DOI**: 10.1002/oa.2416

Practical Significance: Ordinal Scale Data and Effect Size in Zooarchaeology

S. WOLVERTON,[a]* J. DOMBROSKY[a] AND **R. L. LYMAN[b]**

[a] *Department of Geography, University of North Texas, Denton, TX 76203, USA*
[b] *Department of Anthropology, University of Missouri, Columbia, MO 65211, USA*

ABSTRACT Quantitative analysis of zooarchaeological taxonomic abundances and skeletal part frequencies often relies on parametric techniques to test hypotheses. Data upon which such analyses are based are considered by some to be 'ordinal scale at best', meaning that non-parametric approaches may be better suited for addressing hypotheses. An important consideration is that archaeologists do not directly or randomly sample target populations of artefacts and faunal remains, which means that sampling error is not randomly generated. Thus, use of inferential statistics is potentially suspect. A solution to this problem is to rely on a weight of evidence research strategy and to limit analysis to descriptive statistics. Alternatively, if one chooses to use statistical inference, one should analyse effect size to determine practical significance of results and adopt conservative, robust inferential tests that require relatively few assumptions. Archaeologists may choose not to abandon statistical inference, but if so, they should temper how they use statistical tools. Copyright © 2014 John Wiley & Sons, Ltd.

Key words: zooarchaeology; hypothesis testing; effect size; practical significance; non-parametric statistics

Introduction

There was a quantitative revolution in archaeology that was embedded in the revolution of the new archaeology of the 1960s and 1970s (e.g., Watson *et al.*, 1971; see response from Thomas, 1978); statistical approaches were increasingly employed to retarget archaeological analysis away from the qualitative approaches relied upon heavily by culture historians (Vescelius, 1960; Binford, 1964; Clarke, 1968; Redman, 1974). Many archaeologists today approach their research questions by framing and testing statistical hypotheses. However, these analyses may take place without consideration of analytical scale or the relationship between samples and target populations.

Inferential statistical tests are commonly applied in quantitative zooarchaeological research. For example, such tests have been used to interpret shifts in taxonomic abundance over time (e.g., Broughton, 1994a, 1994b; Cannon, 2000; Nagaoka, 2002; Munro, 2004; Braje *et al.*, 2012), relationships between taxonomic richness and sample size (Grayson & Delpech, 1998; Cannon, 2001), taphonomic histories of zooarchaeological faunas (Grayson, 1984; Marean & Spencer, 1991; Lyman, 1994b; Marean & Frey, 1997; Lam *et al.*, 1998; Nagaoka *et al.*, 2008), biometric differences in size of skeletal remains (Wolverton *et al.*, 2008; Braje *et al.*, 2012), relationships between different types of quantitative units of taxonomic and skeletal element abundance (Stiner, 1991; Lyman, 1994a; Giovas, 2009; Domínguez-Rodrigo, 2012), and differences in mortality profiles (Klein, 1982; Lyman, 1987; Wolverton, 2006). There are many published examples in which zooarchaeologists mainly base their conclusions on interpretations of descriptive statistics (e.g., Stiner, 1990; Marean & Kim, 1998; Pickering *et al.*, 2003; Reitz, 2004; Steele, 2005; Lyman, 2010); nonetheless, inferential tests are applied to a diverse set of research problems in zooarchaeology.

Grayson (1979, 1984; Grayson & Frey, 2004) demonstrated that zooarchaeological data on taxonomic abundances and skeletal part frequencies are *at best ordinal scale* (see also Lyman, 2008). Ordinal scale data are those that are rank order or that should be interpreted as such, meaning that greater-than, less-than contrasts are interpretable, but magnitudes of difference should not be interpreted. For example, the number of identified specimens (NISP) of a particular taxon from a particular stratum in a particular faunal assemblage might be 100× greater than the NISP of the same taxon in a separate stratum, but this does

* Correspondence to: Steve Wolverton, Department of Geography, University of North Texas, Denton, TX 76203, USA.
e-mail: wolverton@unt.edu

Received 8 April 2014
Revised 19 June 2014
Accepted 28 July 2014

not necessarily mean that there was 100× more individual animals represented during the period represented by the first assemblage. The same applies to the minimum number of individuals (MNI) quantitative unit. Although NISP and MNI tallies of abundance are recorded at ratio scale, they should be interpreted, according to Grayson (1979, 1984), at ordinal scale or perhaps even more conservatively at nominal scale (presence–absence).

There are many reasons to adopt this conservative statistical approach. One reason is that *at ratio scale*, NISP is subject to the 'problem of interdependence', meaning that fragmentation of bones can lead to multiple tallies from the same bone, thus inflating sample size (Grayson, 1984). Use of MNI instead of NISP controls for the problem of interdependence; but, depending on how samples are aggregated (e.g., on the basis of excavation levels, strata, excavation units or combinations thereof), taxonomic and skeletal part abundance can vary greatly. Grayson (1979, 1984; see also Grayson & Frey, 2004) found, however, that in a large proportion of cases, *at ordinal scale*, taxonomic abundances derived from MNI and NISP correlate strongly to one another. Thus, if an analyst desires to avoid the effects of aggregation from MNI and interdependence on the basis of NISP, either measure can be used, but Grayson's warning that such data are *at best ordinal scale* must be heeded. Grayson's warning has important implications for the application of statistical inferential tests in zooarchaeology.

An important concern relates to the target population one seeks to study; at the onset of analysis, the zooarchaeologist has obtained a convenient sample that may or may not allow study of variables of interest (e.g., taxonomic abundances or skeletal part frequencies). The target population in many cases is the death assemblage produced by those who exploited an animal population in the past, or it may be the life assemblage of the prey animals that were exploited (Driver, 1992, 2011; Lyman, 2008; Wolverton, 2013). Although field excavation strategies may employ a random design, geographic space is sampled at archaeological sites (Hole, 1980 and various chapters in Mueller, 1975); populations of animals hunted, carcasses exploited, bones deposited or bones preserved for recovery during prehistory are not directly sampled. Collections of faunal remains are thus best considered to be fortuitous samples recovered from archaeological deposits. The taphonomic history of the assemblage places even greater contingency on the relationship between the convenient sample that the zooarchaeologist is studying and the target population of interest (references in Lyman, 1994b). As a result, Grayson (1979, 1984; Grayson & Frey, 2004) was correct; abundance data (whether concerning taxa or skeletal parts from a single taxon) are at best ordinal scale. Although this paper is framed with particular reference to zooarchaeology, the logic expressed by Grayson (1979, 1984) applies to quantitative analysis in other areas of archaeology (e.g., Jones *et al.*, 1983; Leonard & Jones 1989; Grayson & Cole, 1998) and paleontology (e.g., Krumbein, 1965).

An argument can be made that inferential statistics have no place in archaeology because the use of all inferential tests relies on the assumption that sampling of target populations is carried out in a manner to ensure that sampling error is randomly generated. This requirement of inferential statistics relies on an ability to directly sample the target population using a strategy that ensures random selection of individual elements of the population. As stated earlier, zooarchaeological data, although subsets of a taphonomically influenced population of accumulated, deposited and preserved animal remains, do not derive through random selection (although choice of excavation units may be random). Thus, the *p*-value used to ascertain whether or not an observed pattern, difference, trend or relationship is statistically significant will have an obscure meaning. In addition, the use of confidence intervals derived from probability distributions is subject to the same problem as interpretation of *p*-values. As a result, statistical inference in archaeology is deserving of more detailed consideration.

Statistical hypothesis testing

All inferential tests assess a statistical 'effect', such as a difference between sample means or a bivariate relationship in paired scores, in comparison with a null hypothesis, which states a condition of no effect. That is, if a non-random difference or relationship (effect) is observed, the null hypothesis of no effect can be rejected with some level of confidence and the alternative hypothesis that the effect is reliably present can be supported. The analyst sets a limit for the level of error or the possibility of incorrectly rejecting the null hypothesis that can be tolerated, which is the allowable level of type I error (e.g., alpha), traditionally set (by statisticians) at 0.05 or 0.01. The *p*-value for the test that has been run is the actual level of type I error for that analysis given the magnitude of difference or the strength of the relationship and the amount of sampling error. In random sampling, larger samples

result in less error because more of the population is represented. As error is reduced, the ability to detect even minor differences or weak relationships (trivial statistical effects) increases, which is termed test power. A probability distribution of one kind or another (e.g., the normal distribution) is used to set the conditions for type I error, and these distributions are the basis for statistical hypothesis testing. Such distributions predict the behaviour of sampling error only when such error is randomly distributed, which is why biological and behavioural scientists devote substantial investment to sampling design to ensure random selection (Zar, 1996).

Significance is the reliability of the result and can be worded as, 'if the same analysis was undertaken on another sample of the same size selected from the same population using the same random approach, a similar result should be observed.' It is helpful to contrast significance to 'effect size'. Cronk (2012:131) defines an *effect size measure* as a tool "that allows one to judge the relative importance of a difference or relationship by reporting the size of the difference". In contrasting significance to effect size, Cronk (2012:121, emphasis added) states the following: "statistical hypothesis testing [via significance] provides a way to tell the odds that differences are *real* [no matter how trivial], effect sizes provide a way to judge the *relative importance* of those differences".

We should be concerned about effect size because, as famous mathematician John Tukey (1991:100) stated, between two samples of phenomena measured on the same variable, A is almost always different than B "in some decimal place". He characterised questions asking about whether or not such differences exist as "foolish". With large samples and sensitive inferential tests, non-random effects can usually be detected. We should, therefore, be concerned about effect size (magnitude of the difference or association). In archaeology, in which representativeness of sampling is not controlled through random sampling but through other means, such as repetitive sampling and taphonomic analysis, *effect size is more important than significance* because of the latter's fundamental link to random sampling.

Effect size measures are often used to determine if sample size is adequate for providing sufficient test power to identify a non-random effect (significance) (Mumby, 2002). A researcher might obtain a highly significant/reliable result (low p-value) related to a relatively weak effect size (or a non-important/non-practical result). Conversely, a researcher might obtain a result with low significance (high p-value) with a large effect size measure (or an important/practical result). There is a tendency to overvalue significance and to ignore effect size (see early comment in archaeology by Thomas, 1978:233; see a recent discussion relevant to science by Gaudart *et al.*, 2014), which relates to singular focus on the role of the p-value in hypothesis testing (a multitude of examples could be cited here, but to do so would target case studies and authors; see Wolverton (2005) and Wolverton *et al*. (2008) for self-critical examples). We return to effect size and its interpretation in the Analyse effect size to determine 'practical significance' section later.

The p-value assists the analyst in determining if a null hypothesis can be rejected with assurance that randomly generated sampling error has a small influence on the observation being made (again, even if the effect is trivial). In archaeology, the p-value of statistical results has no clear meaning; the archaeologist must assume that samples under study were influenced randomly by the vagaries of time, such as differential preservation and the effects of site formation processes on spatial distributions of remains and artefacts. This observation is not new in archaeology and was recognised in discussions of cluster sampling related to the fact that uneven frequencies of artefacts might be encountered across randomly chosen excavation units or an uneven distribution of sites might be discovered in randomly chosen survey units (Redman, 1974; Mueller, 1975; Plog, 1976; Gamble, 1978). In addition to the fact that target populations are not directly sampled, the influence of past culture must also be assumed to be random; that is, a faunal assemblage must be assumed to have derived randomly from hunting of a past living animal population and to have passed randomly through a taphonomic history to the point in time at which the zooarchaeologist employs statistical inference. We see no clear way to test the randomness of these historical trajectories, but there is a virtual certainty that culture and taphonomy *have not converged to emulate random sampling*. Therefore, an important question is as follows: Should archaeologists be using inferential statistics to test hypotheses related to archaeological research questions? If the zooarchaeologist is confident that faunal samples represent target populations, then there may be value in statistical inference. However, significance is not the most important interpretative tool; there are two ways that archaeologists might refocus efforts to inquire about the meaning of results: (1) analyse effect size and (2) use conservative, robust approaches that treat data non-parametrically at ordinal scale.

Analyse effect size to determine 'practical significance'

The temptation to focus on the significance of statistical results should be countered with analysis of effect size of the observed test statistic (Kirk, 1996). Most statistical tests have related effect size measures (Table 1), which help the analyst step beyond *statistical significance* to *practical significance*. Regarding the difference between the two, Ellis (2010:3–4) states,

> A statistically significant result is one that is unlikely to be the result of chance. But a practically significant result is meaningful in the real world. It is quite possible, and unfortunately quite common, for a result to be statistically significant and trivial. It is also possible for a result to be statistically non-significant and important.

A clear example of the difference between statistical and practical significance that illustrates how effect size can aid in interpreting zooarchaeological results comes from Lyman (2004, 2008). In 2004, Lyman published a paper on the prehistoric abundance of elk (*Cervus elaphus*) in the American West. There had been debate as to whether or not elk were depleted by human hunters during prehistory or if they were common and abundant. The interpretation is important because wildlife biologist, Charles Kay (1987, 1990, 1994), asserted in previous publications that data from the American West support the interpretation that prehistoric hunter-gatherers overkilled large game during the Holocene. He argued on the basis of several forms of data that elk were never abundant in the Greater Yellowstone Ecosystem (GYE). Thus, it was argued that the modern 'elk problem' of too many elk in the GYE is a phenomenon of eradication of predators and the absence of human hunters in the system during the last 150 years. Wolf (*Canis lupus*) reintroduction in the 1990s and subsequent establishment of a wolf population has changed this dynamic (Cannon & Cannon, 2004; Joyce, 2012). Our point here is not to review the GYE wolf/elk debate; rather, it is to provide an example of statistical and practical significance related to a zooarchaeological case study. One can see in this example of applied zooarchaeology that whether or not there are significant shifts in elk abundance over time during the Holocene could have important implications for understanding the archaeology of hunter-gatherers in addition to wildlife management.

In his 2004 paper, Lyman did a series of analyses to examine elk abundance in the American West. His purpose was to consider data similar to those that Kay (1987, 1990) analysed but to do so with fewer smoothing effects of time-averaging and space-averaging that are caused by lumping together samples from broad areas and long periods. Lyman (2004) considered one subregion, Eastern Washington, and he compared results using site-by-site analyses compared with a single analysis that considered remains from 86 zooarchaeological assemblages together. The analysis in which assemblages were considered together is of particular interest here

Table 1. Effect size measures from some common inferential tests (adopted from Cohen, 1992; Cronk, 2012)

Test	H_0	Effect size measure	Criteria
Correlation coefficient	$r = 0$, rho = 0	r and rho	0.7 strong 0.5 moderate 0.3 weak
Coefficient of determination		r^2, rho^2	50% strong 25% moderate 10% weak
Independent *t*-test	$\mu_1 = \mu_2$	Cohen's d	0.8 strong 0.5 moderate 0.2 weak
One-way analysis of variance	$\mu_1 = \mu_2 = \ldots \mu_k$	η^2 (eta squared)	0.5 strong 0.3 moderate 0.1 weak
Mann–Whitney *U*-test	Median$_1$ = median$_2$	r = (calculated as z divided by the square root of n)	0.5 strong 0.3 moderate 0.1 weak
Kruskal–Wallis *H* test	Median$_1$ = median$_2$ = … median$_k$	η^2 (eta squared) based on rank-order data	0.5 strong 0.3 moderate 0.1 weak
X^2 test of independence and related tests	$f_{expected} = f_{observed}$	Phi (ø)/Cohen's w (calculated as the square root of X^2 divided by n)	0.5 strong 0.3 moderate 0.1 weak

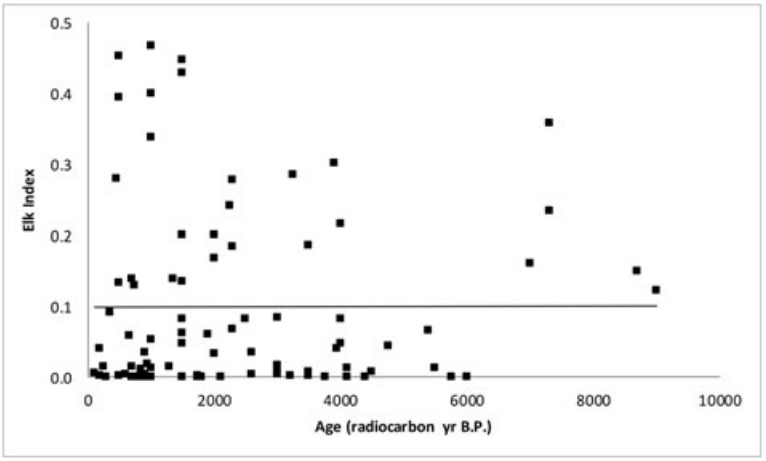

Figure 1. Bivariate scatterplot of the elk index ($\sum$ elk remains/$\sum$ elk + $\sum$ bison + $\sum$ deer + $\sum$ pronghorn + $\sum$ bighorn remains) and age of faunal assemblages from 86 sites in eastern Washington (data from Lyman, 2004, Table 8, after Figure 4).

because Lyman (2004:187–191) first analysed the data visually considering the best-fit regression line (Figure 1) and Pearson's *r* correlation and later redid the analysis with a variant of the χ^2 test of independence (Lyman, 2008: 200–209). The two tests provide strikingly different results.

Pearson's *r* correlation indicates that there is no relationship as does visual inspection of Figure 1 ($r = 0.006$, $p > 0.9$), and "on this basis [Lyman] suggested that there is no evidence here that the abundance of elk relative to the abundance of all ungulates changed over the 10,000 years represented" (Lyman, 2008:205). However, following Cannon (2001), Lyman redid the analysis using Cochran's test of linear trends (also called χ^2_{trend} analysis) (Zar, 1996; Cannon, 2000; Lyman, 2008). Zooarchaeologists regularly use this type of chi-square test of independence (Cannon, 2001; Munro, 2004; Wolverton, 2005; Nagaoka, 2006; Broughton *et al.*, 2010), which subdivides the influence of a temporal trend from the effects of sample size across categories to determine if proportional abundances of taxa or skeletal parts change significantly over time. The χ^2_{trend} analysis for Lyman's data is highly significant ($\chi^2_{trend} = 112.96$, $p < 0.0001$), suggesting that elk abundance changed during the Holocene. Thus, contrary to what Lyman reported in 2004, humans may have had a long-term impact on elk populations if one accepts the χ^2_{trend} result over the Pearson's *r* result. Chi-square tests, however, are very powerful when combined tallies in categories (*N*) is large ($N = 20\,791$ for Lyman's analysis). Test power can be defined as the ability to reject the null hypothesis (here that no trend exists), and in chi-square tests, power (or sensitivity) increases with total sample size (*N*). Thus, because of the influence of high test power—related to a large sample across the represented assemblages—what appeared to be a non-significant relationship in 2004 was determined to be significant in 2008.

Assessment of effect size allows the zooarchaeologist to determine if the 2008 χ^2_{trend} result is meaningful. Effect size measures for chi-square tests take the square root of the χ^2 statistic divided by sample size. Measures such as this one, which is known as phi, for the χ^2 test of independence, are expressed as a continuum from 0 to 1. In the case of phi, a value of roughly 0.1 is considered weak, ranging to moderate effect size at 0.3 and large effect size at 0.5. The effect size for Lyman's χ^2_{trend} analysis is 0.07, which is very weak. We can safely conclude that though there is a non-random relationship between elk abundance and assemblage age, it is only a very minor non-random effect detected because of large aggregate sample size; that is, the result has no meaning in terms of practical significance and elk abundance (as concluded in 2004) did not change over time.

A more commonly recognised effect size measure is the coefficient of determination (r^2), which is affiliated with Pearson's *r* correlation. The coefficient of determination can be expressed as 'the percent of the variability in the dependent variable (y) that can be explained by a corresponding shift in the independent variable (x)'. Thus, the lower that r^2 is the less variability in *y* that can be explained by a shift in *x*. Low correlation coefficients (*r*) that are significant (with low *p*-values) will produce low r^2 values, meaning that although a statistically significant relationship exists, practical significance is low. For example, Wolverton (2005) reported a Spearman's rho of 0.54 with a *p*-value of <0.01 for the relationship between excavation depth and AMS-dated age of white-tailed deer (*Odocoileus virginianus*) bone specimens from Arnold Research Cave, a Holocene rockshelter fauna from central

Missouri. Spearman's rho is a non-parametric correlation of bivariate relation in ranks; thus, rho^2 is the percent of variability in the rank order of y that can be explained by a shift in rank order of x. Looking back with knowledge of effect size, although this relationship is non-random, only 25% of the rank-order variability in age of accelerator mass spectrometry (AMS)-dated bone specimens can be explained by a corresponding shift in excavation depth (which was pointed out in external review of that 2005 paper). Cochran's test of linear trends was then used to assess temporal changes in proportional faunal indices that highlight shifts in abundances of white-tailed deer and box turtle (*Terrepene* sp.) remains during the Holocene, and although highly significant trends were reported, sample sizes were medium ($N = 214$ for the turtle analysis, and $N = 797$ for the deer analysis) and effect size was not reported. Phi for the χ^2_{trend} statistic for the box turtle analysis is 0.28, which is moderate; it is also moderate for the analysis of proportional abundance of deer remains over time (0.25). These data were used to demonstrate that high-rank prey [either large bodied (deer) or slow-moving (turtles)] fluctuated with independent measures of Holocene climate change in the region. The moderate effect size indicates that the results have practical significance and that climate change and use of high-rank prey over time were correlated.

It is well known in zooarchaeology that patterns in proportional faunal abundances (whether skeletal part frequencies or taxonomic abundances) can be driven by substantial inter-assemblage differences in sample size (references cited by Grayson, 1984; Lyman, 2008). Cannon (2000, 2001) introduced Cochran's test of linear trends because it is a test that partitions the effect of shifts in proportions across time from those related to differences in sample size between the assemblages being analysed. Prior to his introduction of the test, researchers had been using Spearman's rho to do two separate analyses; the first determined whether or not sample size could be driving faunal abundances. The second, if the Spearman's rho showed no correlation with sample size, determined if there was a significant change in abundance between assemblages over time. Cochran's test of linear trends does not require two analyses, and Cannon (2001) found that its use was more likely to accurately detect temporal trends independent of sample size effects. Cochran's test of linear trends is a very useful test for analysing temporal zooarchaeological faunal data; we do not discourage its use. Rather, we are simply advocating that effect size measures be reported to assess the strength of the effects that are observed.

If the zooarchaeologist only considers the p-value, which may be very low and highly significant, it is tempting to conclude that an *important* shift or trend in faunal data has been observed. If the effect size is low, however, this interpretation would be incorrect. Virtually any non-random and minor effect will be detected reliably with large sample sizes using many inferential tests; thus, the zooarchaeologist should report effect size related to results. If the effect size is moderate or strong (following Table 1) and the result is significant, then the analyst has not only rejected the null hypothesis reliably but she/he has also observed an important trend. Most tests of difference and correlation, whether parametric or non-parametric, have associated effect size measures with clearly stated interpretive criteria (e.g., Cohen, 1988, 1992; Table 1). Zooarchaeologists should report effect size with p-values of inferential tests and use them to interpret the practical significance of results.

Use ordinal scale and/or categorical statistics

Given the discipline-wide inability to randomly sample target populations, it is tempting to wonder why we use inferential statistics at all. If the p-value is potentially meaningless, then our tests of difference and correlation are essentially elegant descriptions of samples of data for which data quality and representativeness are unknown. This fact is not easily brushed aside; use of inferential statistics should be embedded in a weight of evidence approach and be accompanied by detailed taphonomic research. One alternative is to restrict analysis to descriptive statistics; however, descriptive statistics still require an assumption of representativeness. In addition, should inferential tests be employed, conservative approaches are mandatory and effect size is important. The normality assumption in statistics (descriptive or inferential) relates to the central limit theorem, which concerns how the distribution of randomly generated error behaves at large ($n \geq 30$) sample sizes. Given that sample size and representativeness are of critical concern in zooarchaeology (e.g., Gamble, 1978; Lyman & Ames, 2004), at a minimum, the normality assumption should be avoided. Regarding inference, parametric tests are more powerful because they take advantage of the normality assumption; a more powerful and sensitive test is more likely to detect a significant statistical effect. However, so-called gunning for significance is not the goal of hypothesis testing. If one wishes to extend beyond description, a moderately conservative approach given the constraints of archaeological sampling requires that

the analyst make it more difficult to reject the null hypothesis by using less powerful inferential tests, such as non-parametric and categorical inferential tests (given the caveat of the influence of large sample size discussed earlier).

It is helpful to think of the use of non-parametric and categorical statistical tests in terms of *robustness* and *resistance*. Robustness refers to the insensitivity of statistical techniques to small deviations from the assumptions about the distribution of sampling error (e.g., normality) (Huber & Ronchetti, 2009:2). Resistant statistical tests are those in which the test statistic is insensitive to small changes in the sample but not the underlying assumptions about the distribution of the population (which is robustness). It is crucial to understand that non-parametric statistics cannot be considered robust in all instances (Huber & Ronchetti, 2009:6). However, given what is known about the zooarchaeological record, non-parametric statistics will often be the more robust choice compared with parametric approaches. It is the convergence of two (aforementioned) observations that makes parametric inference testing in zooarchaeology less robust: (1) zooarchaeological assemblages are fortuitous samples from populations that were not sampled directly (thus, the distribution of sampling error is unknown); and (2) faunal assemblages represent a conglomeration of diverse taphonomic histories. Non-parametric (ordinal scale) and categorical statistical tests are also considered to be more resistant because they are based on ranks or counts in categories (Lanzante, 1996). Resistant statistical procedures should be utilised because they are less affected by outliers. An example of a resistant measure of central tendency is the median, which is less sensitive to the influence of outliers than is the mean. The median is the score at the middle rank; thus, it is ordinal scale.

Even the most basic zooarchaeological data, tallies, can be interpreted very differently at ordinal versus ratio scales. For example, there has been a lengthy debate in zooarchaeology during the last two decades concerning whether or not midshaft fragments of long bones of ungulates should be refit prior to tallying skeletal part frequencies (Marean & Kim, 1998; Stiner, 2002; Pickering *et al.*, 2003; Cleghorn & Marean, 2004; Grayson & Frey, 2004; Marean *et al.*, 2004). The logic in support of refitting long-bone shaft fragments is the previously mentioned problem of interdependence using NISP; multiple specimens from the same element could be counted as NISP and thus inflate skeletal part frequencies for a particular body part and/or taxon (Marean & Kim, 1998).

Pickering *et al.* (2003) demonstrate using data from an actualistic study by Capaldo (1995) that there is an influence of carnivore ravaging on frequencies of proximal-end, distal-end and midshaft portions of long bones of bovids of three distinctive body sizes *at ratio scale* (represented in Table 2). The study mimicked discard of ungulate remains from Hominid sites in East Africa, and remains were ravaged by carnivores. Long-bone portion frequencies were compared between pre-ravaged and ravaged assemblages. In all cases, the NISP of midshaft specimens increased substantially and frequencies of ends decreased with ravaging. The

Table 2. Number of identified specimens (NISP) of pre-carnivore-ravaged and carnivore-ravaged bone fragments and the difference between ratio scale and ordinal scale interpretation (adopted from Table 3 in Pickering et al. (2003))

Size class	Limb portion	Pre-ravaged NISP	Pre-ravaged NISP rank	Ravaged NISP	Ravaged NISP Rank	Source
Size class 1						
	Proximal epiphysis	64	2	9	2	Capaldo, 1995
	Shaft	138	1	253	1	
	Distal epiphysis	62	3	5	3	
Size class 2						
	Proximal epiphysis	101	2	27	2	Capaldo, 1995
	Shaft	418	1	780	1	
	Distal epiphysis	96	3	14	3	
Size class 3						
	Proximal epiphysis	71	2	22	2	Capaldo, 1995
	Shaft	306	1	577	1	
	Distal epiphysis	66	3	11	3	
Size class 3						
	Proximal epiphysis	63	3	19	3	Domínguez-Rodrigo (from Pickering *et al.*, 2003)
	Shaft	181	1	260	1	
	Distal epiphysis	86	2	27	2	

authors concluded that use of NISP without refitting is flawed because of these shifts in skeletal part frequencies at ratio scale. However, if Grayson's approach is followed and data are considered at ordinal scale, which is (again) appropriate with zooarchaeological data given the contingencies preventing direct, representative sampling of target populations, a much different interpretation emerges. First, for each type of bovid, there is no change in the rank-order abundance of proximal-end, distal-end and midshaft portions despite the ratio-scale reduction in ends and increase in shaft fragments (Table 2). Second, the abundance of all portions, regardless of bovid size between the 12 pre-ravaged and ravaged assemblages is strongly correlated at ratio scale ($r = 0.99$, $p < 0.05$) and ordinal scale (rho = 0.92, $p < 0.05$), indicating that NISP of portions prior to ravaging drives abundance after ravaging. We do not wish to make a recommendation as to whether or not zooarchaeologists should refit midshaft fragments; rather, our point, much like with Lyman's study of elk abundance, is that a different interpretation emerges with a more robust, resistant, yet less powerful statistical approach (ordinal scale instead of ratio scale). We have used one portion of Pickering *et al.*'s (2003) analysis here to clarify this point; for a more complete history of the midshaft refitting debate, see Marean *et al.* (2004).

In archaeology, robust approaches that require relatively few assumptions and resistant statistics that are not as influenced by outliers represent conservative approaches for statistical description and inference. In terms of basic inference, nearly all parametric tests, such as *t*-tests and correlation analyses, have sibling non-parametric tests that avoid the normality assumption by addressing data at ordinal scale (Figure 2). Data may also be considered conservatively using categorical tests of proportions (Figure 2). However, it must be acknowledged that avoiding more powerful parametric tests does not allow the archaeologist to escape from the general problem of being unable to probabilistically sample target populations of interest.

Statistical Approaches

Description

Parametric (Ratio scale)	Non-parametric (Ordinal scale)
Central Tendency	
Mean	Median, Mode
Dispersion	
Variance, Standard deviation, Coefficient of variation	Interquartile range, Range

Inference

Parametric (Ratio scale)	Non-parametric (Ordinal scale)
Independent samples t test	Mann-Whitney U test
Dependent samples t test	Wilcoxon rank sum test
One-way ANOVA	Kruskal-Wallis H test
Pearson's r correlation	Spearman's rho correlation
	Categorical
	Chi-square goodness of fit
	Chi-square test of independence
	Cochran's test of linear trends

Figure 2. Examples of related parametric (ratio scale) and non-parametric (ordinal scale) descriptive statistics and inferential tests.

Pandora's box

Archaeologists have opened Pandora's box because statistical inference now assumes a large role in research (as evident by the numerous statistical textbooks authored by archaeologists over the past several decades). Given that neither a purposeful nor a probabilistic sampling design during excavation can ensure random sampling of target populations, should the box be closed? Or should we simply ignore these problems of statistical inference and move on wearing comfortable disciplinary blinders? Perhaps statistical inference should be abandoned in archaeology, but if we are to employ traditional inferential statistical approaches in archaeology, the box cannot be closed. How can representativeness of a sample be assessed? This may depend on the type of research question being addressed, but there are some promising avenues of inquiry. These are merely suggested here and are not covered in great detail.

First, the writing of taphonomic histories in zooarchaeology (analysis of site formation processes in general for archaeology) is critically important because without full consideration of the processes that influence assemblages of remains leading up to analysis, all processes that influence the sample are unknown. Second, analytical approaches such as nestedness and the use of sampling to redundancy may help us understand representativeness of taxonomic assemblages of past ecological communities (Lyman, 2008; Peacock, 2012). Isotopic chemistry approaches may also be useful in this regard (Peacock *et al.*, 2012). Similarly, the bauplan of different types of animals provides an expectation for analyses of skeletal part frequencies (references in Lyman, 1994b). Third, within a locality or region, a weight of evidence approach that relies on several related proxy variables, such as taxonomic abundance data, prey demographic data, independent environmental records and related measures can increase confidence in the validity of observed statistical effects (e.g., Stiner, 1994; Broughton *et al.*, 2010). There are many ways that the archaeologist can increase sample size or compare multiple related assemblages to determine if they register the same changes and thus inferentially monitor the same cause (e.g., climate change) (Findley, 1964; Grayson, 1981). Zooarchaeologists should test hypotheses using weight of evidence, conservative ordinal scale approaches and interpretation of effect size, rather than emphasising statistical significance because random sampling of target populations in most cases is not possible.

Conclusion

Our brief but critical examination of sampling in zooarchaeology and how it relates to the use of inferential statistics has identified problems but also solutions. We surmise that archaeologists tend to focus on the significance of results, and we suggest that effect size criteria for those results should also be assessed and are more important. Our commentary might be taken to mean that we should not use inferential statistics in archaeology/zooarchaeology. Indeed, a conservative solution is to simply use non-parametric descriptive statistics instead of inferential tests. Statisticians did not have in mind the types of sampling problems archaeologists face when fundamental statistical tests were developed. The origins of these tests relate to datasets that derive from analyses in which sampling design can be highly controlled and random generation of error can be ensured. In archaeology, we embrace an important trade-off; we exchange the intrigue and mystery of studying the past for reduced control of sampling and data quality. It is our contention that we be transparent about this exchange when we employ and interpret inferential statistical approaches.

Acknowledgements

The authors thank Jon Driver and two anonymous reviewers for helpful review comments.

References

Binford LR. 1964. A consideration of archaeological research design. *American Antiquity* **29**: 425–441.

Braje TJ, Rick TC, Erlandson JM. 2012. Rockfish in the Longview: Applied archaeology and conservation of Pacific red snapper (Genus Sebastes) in Southern California. In Applied Zooarchaeology and Conservation Biology, S Wolverton, RL Lyman (eds.). University of Arizona Press: Tucson; 157–178.

Broughton JM. 1994a. Declines in mammalian foraging efficiency during the late Holocene, San Francisco Bay, California. *Journal of Anthropological Archaeology* **13**: 371–401.

Broughton JM. 1994b. Late Holocene resource intensification in the Sacramento Valley, California: The vertebrate evidence. *Journal of Archaeological Science* **21**: 501–514.

Broughton J, Cannon M, Bartelink E. 2010. Evolutionary ecology, resource depression, and niche construction theory: Applications to central California hunter-gatherers and Mimbres-Mogollon agriculturalists. *Journal of Archaeological Method and Theory* **17**: 371–421.

Capaldo SD. 1995. Inferring Hominid and carnivore behavior from dual-patterned archaeological assemblages. PhD Dissertation. Rutgers University.

Cannon KP, Cannon MB. 2004. Zooarchaeology and wildlife management in the greater Yellowstone

ecosystem. In Zooarchaeology and Conservation Biology, RL Lyman, KP Cannon (eds.). The University of Utah Press: Salt Lake City; 45–60.
Cannon MD. 2000. Large mammal relative abundance in pithouse and pueblo period archaeofaunas from southwestern New Mexico: Resource depression among the Mimbres-Mogollon?. *Journal of Anthropological Archaeology* **19**: 317–347.
Cannon MD. 2001. Archaeofaunal relative abundance, sample size, and statistical methods. *Journal of Archaeological Science* **28**: 185–195.
Clarke DL. 1968. Analytical Archeology. Methuen: London.
Cleghorn N, Marean CW. 2004. Distinguishing selective transport and *in situ* attrition: A critical review of analytical approaches. *Journal of Taphonomy* **2**: 43–67.
Cohen J. 1988. Statistical Power Analysis for the Behavioral Sciences. L. Erlbaum Associates: Hillsdale, NJ.
Cohen J. 1992. A power primer. *Psychological Bulletin* **112**: 155–159.
Cronk BC. 2012. How to Use SPSS (12th Edition). Pyrczak Publishing: Glendale, CA.
Domínguez-Rodrigo M. 2012. Critical review of the MNI (minimum number of individuals) as a zooarchaeological unit of quantification. *Archaeological and Anthropological Sciences* **4**: 47–59.
Driver JC. 1992. Identification, classification and zooarchaeology. *Circaea* **9**: 35–47.
Driver JC. 2011. Identification, classification and zooarchaeology (featured reprint and invited comments). *Ethnobiology Letters* **2**: 19–39.
Ellis PD. 2010. The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge University Press: New York.
Findley JS. 1964. Paleoecologic reconstruction: Vertebrate limitations. In The Reconstruction of Past Environments, JJ Hester, J Schoenwetter (eds.). Fort Burgwin Research Center: Taos, NM; 23–25.
Gamble C. 1978. Optimising information from studies of faunal remains. In Sampling in Contemporary British Archaeology, JF Cherry, C Gamble, S Shennan (eds.). British Archaeological Reports British Series vol 50: Oxford; 321–353.
Gaudart J, Huiart L, Milligan PJ, Thiebaut R, Giorgi R. 2014. Reproducibility issues in science, is P value really the only answer? *Proceedings of the National Academy of Science* **111**: E1934.
Giovas CM. 2009. The shell game: Analytic problems in archaeological mollusc quantification. *Journal of Archaeological Science* **36**: 1557–1564.
Grayson DK. 1979. On the quantification of vertebrate archaeofaunas. *Advances in Archaeological Method and Theory* **2**: 199–237.
Grayson DK. 1981. A critical view of the use of archaeological vertebrates in paleoenvironmental reconstruction. *Journal of Ethnobiology* **1**: 28–38.
Grayson DK. 1984. Quantitative Zooarchaeology: Topics in the Analysis of Archaeological Faunas. Academic Press: Orlando, FL.
Grayson DK, Cole SC. 1998. Stone tool assemblage richness during the Middle and Early Upper Palaeolithic in France. *Journal of Archaeological Science* **25**: 927–938.
Grayson DK, Delpech F. 1998. Changing diet breadth in the early Upper Palaeolithic of southwestern France. *Journal of Archaeological Science* **25**: 1119–1129.
Grayson DK, Frey CJ. 2004. Measuring skeletal part representation in archaeological faunas. *Journal of Taphonomy* **2**: 27–42.
Hole BL. 1980. Sampling in archaeology: A critique. *Annual Review of Anthropology* **9**: 217–234.
Huber PJ, Ronchetti EM. 2009. Robust Statistics. Wiley: Hoboken, NJ.
Jones GT, Grayson DK, Beck C. 1983. Artifact class richness and sample size in archaeological surface assemblages. In Lulu Linear Punctated: Essays in Honor of George Irving Quimby, RC Dunnell, DK Grayson (eds.). University of Michigan Anthropological Papers No. 72: Ann Arbor, MI; 55–73.
Joyce M. 2012. Constructing nature: Art, conservation, and applied zooarchaeology. *Journal of Ethnobiology* **32**: 246–264.
Kay CE. 1987. Too many elk in Yellowstone? *Western Wildlands* **13**: 39–41.
Kay CE. 1990. Yellowstone's northern elk herd: A critical evaluation of the "natural regulation" paradigm. Unpublished doctoral dissertation. Utah State University: Logan.
Kay CE. 1994. Aboriginal overkill. *Human Nature* **5**: 359–398.
Kirk RE. 1996. Practical significance: A concept whose time has come. *Educational and Psychological Measurement* **56**: 746–759.
Klein RG. 1982. Age (mortality) profiles as a means of distinguishing hunted species from scavenged ones in Stone Age archeological sites. *Paleobiology* **8**: 151–158.
Krumbein WC. 1965. Sampling in paleontology. In Handbook of Paleontological Techniques, B Kummel, D Raup (eds.). WH Freeman and Company: San Francisco, CA; 137–150.
Lam YM, Chen X, Marean CW, Frey CJ. 1998. Bone density and long bone representation in archaeological faunas: Comparing results from CT and photon densitometry. *Journal of Archaeological Science* **25**: 559–570.
Lanzante JR. 1996. Resistant, robust and non-parametric techniques for the analysis of climate data: Theory and examples, including applications to historical radiosonde station data. *International Journal of Climatology* **16**: 1197–1226.
Leonard RD, Jones GT (eds.). 1989. Quantifying Diversity in Archaeology. Cambridge University Press: Cambridge.
Lyman RL. 1987. On analysis of vertebrate mortality profiles: sample size, mortality type, and hunting pressure. *American Antiquity* **52**: 125–142.
Lyman RL. 1994a. Quantitative units and terminology in zooarchaeology. *American Antiquity* **59**: 36–71.
Lyman RL. 1994b. Vertebrate Taphonomy. Cambridge University Press: Cambridge.
Lyman RL. 2004. Aboriginal overkill in the intermountain west of North America: Zooarchaeological tests and implications. *Human Nature* **15**: 169–208.
Lyman RL. 2008. Quantitative Paleozoology. Cambridge University Press: Cambridge.
Lyman RL. 2010. Taphonomy, pathology, and paleoecology of the terminal Pleistocene Marmes Rockshelter (45FR50)

"big elk" (*Cervus elaphus*), southeastern Washington State, USA. *Canadian Journal of Earth Sciences* **47**: 1367–1382.
Lyman RL, Ames KM. 2004. Sampling to redundancy in zooarchaeology: Lessons from the Portland Basin, northwestern Oregon and southwestern Washington. *Journal of Ethnobiology* **24**: 329–346.
Marean CW, Domínguez-Rodrigo M, Pickering TR. 2004. Skeletal element equifinality in zooarchaeology begins with method: the evolution and status of the "shaft critique." *Journal of Taphonomy* **2**: 69–98.
Marean CW, Frey CJ. 1997. Animal bones from caves to cities: Reverse utility curves as methodological artifacts. *American Antiquity* **62**: 698–711.
Marean CW, Kim SY. 1998. Mousterian large-mammal remains from Kobeh Cave behavioral implications for neanderthals and early modern humans. *Current Anthropology* **39**: 79–114.
Marean CW, Spencer LM. 1991. Impact of carnivore ravaging on zooarchaeological measures of element abundance. *American Antiquity* **56**: 645–658.
Mueller JW (ed.). 1975. Sampling in Archaeology. University of Arizona Press: Tucson.
Mumby PJ. 2002. Statistical power of non-parametric tests: A quick guide for designing sampling strategies. *Marine Pollution Bulletin* **44**: 85–87.
Nagaoka L. 2002. The effects of resource depression on foraging efficiency, diet breadth, and patch use in southern New Zealand. *Journal of Anthropological Archaeology* **21**: 419–442.
Nagaoka L. 2006. Prehistoric seal carcass exploitation at the Shag Mouth site, New Zealand. *Journal of Archaeological Science* **33**: 1474–1481.
Nagaoka L, Wolverton S, Fullerton B. 2008. Taphonomic analysis of the Twilight Beach seals. In Islands of Inquiry: Colonisation, Seafaring and the Archaeology of Maritime Landscapes, G Clark, F Leach, S O'Connor (eds.). ANUE Press: Canberra; 475–498.
Munro ND. 2004. Zooarchaeological measures of hunting pressure and occupation intensity in the Natufian: Implications for agricultural origins. *Current Anthropology* **45**: 5–34.
Peacock E. 2012. Archaeological freshwater mussel remains and their use in the conservation of imperiled fauna. In Conservation Biology and Applied Zooarchaeology, S Wolverton, RL Lyman (eds.). The University of Arizona Press: Tucson; 42–67.
Peacock E, Randklev CR, Wolverton S, Palmer RA, Zaleski S. 2012. The "cultural filter," human transport of mussel shell, and the applied potential of zooarchaeological data. *Ecological Applications* **22**: 1446–1459.
Pickering TR, Marean CW, Domínguez-Rodrigo M. 2003. Importance of limb bone shaft fragments in zooarchaeology: A response to "On in situ attrition and vertebrate body part profiles" (2002), by M. C. Stiner. *Journal of Archaeological Science* **30**: 1469–1482.
Plog S. 1976. Relative efficiencies of sampling techniques for archaeological surveys. In The Early Mesoamerican Village, KV Flannery (ed.). Academic Press: New York; 136–158.
Redman CL. 1974. Archeological Sampling Strategies. Addison-Wesley Module in Anthropology No. 55: New York.
Reitz EJ. 2004. "Fishing down the food web": A case study from St. Augustine, Florida, USA. *American Antiquity* **69**: 63–83.
Steele TE. 2005. Comparing methods for analysing mortality profiles in zooarchaeological and palaeontological samples. *International Journal of Osteoarchaeology* **15**: 404–420.
Stiner MC. 1990. The use of mortality patterns in archaeological studies of hominid predatory adaptations. *Journal of Anthropological Archaeology* **9**: 305–351.
Stiner MC. 1991. Food procurement and transport by human and non-human predators. *Journal of Archaeological Science* **18**: 455–482.
Stiner MC. 1994. Honor Among Thieves: A Zooarchaeological Study of Neandertal Ecology. Princeton University Press: Princeton, NJ.
Stiner MC. 2002. On in situ attrition and vertebrate body part profiles. *Journal of Archaeological Science* **29**: 979–991.
Thomas DH. 1978. The awful truth about statistics in archaeology. *American Antiquity* **43**: 231–244.
Tukey JW. 1991. The philosophy of multiple comparisons. *Statistical Science* **6**: 100–116.
Vescelius GS. 1960. Archaeological sampling: A problem of statistical inference. In Essays in the Science of Culture in Honor of Leslie A. White, GE Dole, RL Carneiro (eds.). Thomas Y. Crowell Company: New York; 457–470.
Watson PJ, LeBlanc SA, Redman CL. 1971. Explanation in Archaeology: An Explicitly Scientific Approach. Columbia University Press: New York.
Wolverton S. 2005. The effects of the hypsithermal on prehistoric foraging efficiency in Missouri. *American Antiquity* **70**: 91–106.
Wolverton S. 2013. Data quality in zooarchaeological faunal identification. *Journal of Archaeological Method and Theory* **20**: 381–396.
Wolverton S, Nagaoka L, Densmore J, Fullerton B. 2008. White-tailed deer harvest pressure & within-bone nutrient exploitation during the mid-to late Holocene in southeast Texas. *Before Farming* **2**: 1–23.
Wolverton S. 2006. Natural-trap ursid mortality and the Kurtén response. *Journal of Human Evolution* **50**: 540–551.
Zar JH. 1996. Biostatistical Analysis. Prentice Hall: Englewood Cliffs, NJ.

