Chapter 18
Reasoning with Zooarchaeological Counting Units and Statistics

The very idea of statistics plunges a substantial proportion of archaeology students either into a semi-comatose state or outright panic. I have been one of those students, for whom only repeated exposure to the subject has diminished the strength and duration of these reactions. But, just as one cannot discuss patterning in faunal assemblages without counting, one cannot adequately assess hunches about similarities or differences between and among archaeofaunas without employing statistically based comparisons. This chapter provides an overview of commonly used statistical tests in zooarchaeology. It starts with basics because, though thoroughly convinced of the value of the topics to be covered, basics are where I usually begin. Hopefully, this approach is useful to some readers, and math whiz readers are asked to literally overlook this back-to-basics approach. Chapter 18 opens by outlining what statistical tests ultimately tell us. It then discusses the statistical tests appropriate to the various zooarchaeological counting units of measure. It then outlines the respective strengths and weaknesses of zooarchaeological counting units, reporting on recent debates over which are the best measures of relative taxonomic and skeletal element abundances. It ends with a discussion of zooarchaeological counting units and statistics as tools that are variously useful or appropriate to different research problems. The reader might wish to review Chap. 10's section on NISP, MNI, MNE, and MAU, as these were introduced there.

18.1 Commonly Used Statistical Tests in Zooarchaeology

When comparing two or more things, one asks whether they are similar or different. Sometimes this may be straightforward to answer, as when the question is, are these five apples green, or are all of these stones human artifacts? As questions become more specific, such as, are all the apples the same shade of green, or are these artifacts projectile points, how to answer the questions becomes more complex. It may often involve units of measure, as with a color's hue and chroma characterized by

D. Gifford-Gonzalez, *An Introduction to Zooarchaeology*,
https://doi.org/10.1007/978-3-319-65682-3_18

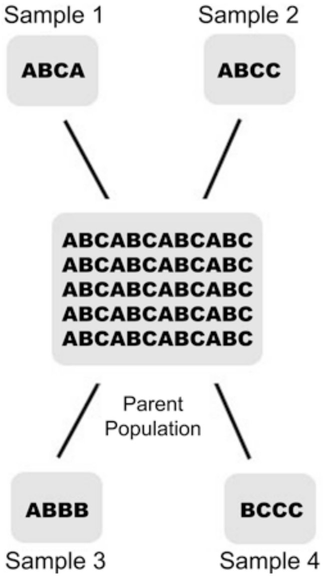


Fig. 18.1 A hypothetical parent population and four samples drawn randomly from it, showing the possibility that the samples may be differently representative of the parent population (Illustration by the author)

the Munsell® color system. In the case of the artifacts, measurements of length, width, and details of overall shape, as in stem length or angles of notching, if any, may be used to answer the question. With such artifacts, other questions may arise: is the weight of the individual artifacts within some rather narrow range of variation? Does there seem to be a preference for a relatively rare raw material for making the points? To answer these questions, we need is a set of standards to assist our judgments about similarity or difference.

18.1.1 Populations, Normal Distributions, and Samples

Probabilistic statistical tests evaluate such questions by calculating the likelihood that the specimens we are studying could have derived from the same *parent population*. Figure 18.1 shows a hypothetical parent population and four samples randomly drawn from it. Samples 1 and 2 have are quite similar to the parent population and to each other. They are therefore highly likely to have been drawn from the same parent population. By contrast, Samples 3 and 4 differ substantially from each other, and they might not be drawn from the same parent population, although it is possible. This sliding scale of likelihood underlies probability-based comparisons of samples.

In nearly all cases, statistical tests do not have a “real” parent population against which to compare samples. Instead, statistical inference assesses the *probability* of a parent population producing a sample. When two samples are compared, the test assesses the likelihood that the respective samples could have been drawn randomly from the same population of values. The *p*-value, is an estimate of the likelihood

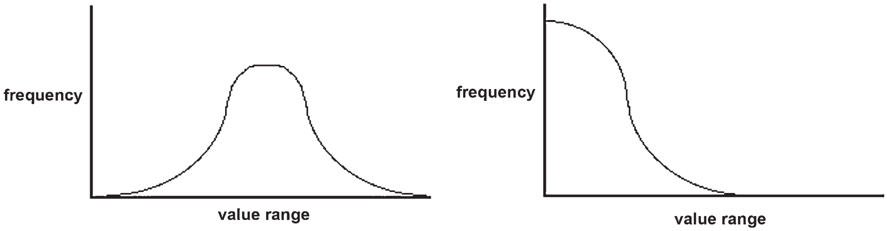


Fig. 18.2 **Left:** a normal curve, showing symmetrical distribution of values around a mode. **Right:** a Poisson distribution, showing a fall-off in values in one direction from the mode only (Illustration by the author)

that the first value is drawn from the same parent population as the second. Such *p*-values expressed as a decimal fraction of 1.0, such as: 0.50 0.05, 0.01, 0.001, and 0.0001 etc. A *p*-value of 0.70 indicates they are quite likely (70%) to have been drawn from the sample hypothetical parent population, whereas 0.001 indicates a much lower (1%) likelihood. A *p*-value of 0.05 value is usually considered to be on the cusp of statistical significance, but it involves some probability (5%) that the compared sets are drawn from the same population. Results of $p = 0.01$ are considered "significant," and $p = 0.001$ levels are termed "highly significant." Values >0.05 are not considered statistically significant, although many cases from biology and ecology show that strong, biologically relevant *trends* or tendencies are not always statistically significant.

Most statistical tests' derivation of *p*-values assume that a trait's values are randomly distributed in a *normal distribution* around a mode. This would be the case, say, with the tail length in a species of squirrel, where some are shorter, some longer, but most fall around a *mode*, or central tendency, where most data points lie (Fig. 18.2, left). Normal curves are symmetrical, having their highest point at the mean of the variable values, that is, the mode and mean values coincide, and terminate in zero frequency of occurrence of the variable of interest – in this case, lizard tail length – at either end of their distribution. This is called a two-tailed distribution – which is unrelated to squirrel tails, but rather to the tapering ends of the values.

However, some traits in nature, such as DNA mutation rates, and in archaeology, such as the sizes of flakes driven from a single core by percussion flaking, are not distributed in a normal curve, but rather in what amounts to half of a normal curve, with the mode falling off in only one direction of values. This is called a one-tailed or Poisson distribution (Fig. 18.2, right), which can have the mode to the left or the right. Common estimates of likelihood are based on the assumption of a normal curve one-tailed or two-tailed curve. However, not all data distributions follow "normal curve" models; some can be bimodal or show no mode, and tests assuming a normal distribution become less reliable when applied to these.

Not all statistical tests can be used with all types of variables. Understanding the differences among commonly used types of variables is important, especially because some common zooarchaeological quantification units incorporate basic assumptions that in effect disqualify them from certain statistical tests. The next section provides definitions of terms and tests commonly employed in zooarchaeological statistics. It then discusses those specific aspects of zooarchaeological counting units – NISP, MNE, MNI, MAU – that limit the types of tests that can be used. This does not mean these zooarchaeological measures are "defective," but rather that effectively using them requires some circumspection.

18.1.2 Discrete vs. Continuous Variables

Variables are basic units of counting that describe attributes of what we want to study. Variables could include coat color, height, depth, number of cut marks, and so forth. A given variable differs in its *values*. Coat color in cats, for example, may be black, white, orange, gray, brown, etc. Adult humans can have a wide range of heights. Lakes vary considerably in maximum depth and volumes of water. In using variables to compare two or more samples, it is necessary to understand how the variable "behaves" in terms of those values.

Two general types of variable may be distinguished: *discrete* and *continuous*. The numbers on the sides of playing dice are an example of a discrete variable. When someone throws dice, he or she can only come up with one of six values on each die. One never throws a $2^{1/2}$ or 5.67. Another example is DNA certain codons producing the expressed blood types of the ABO blood group: one is A, B, AB, or O. Discrete variables can be either finite (a definably limited number of values, such as the examples given above) or countably infinite, a more abstract possibility that does not concern archaeofaunal analysis. Thomas (1986) gave the example of a book of an infinite number of pages, but in which each page would nonetheless be a discrete, predictable member of a counted series, as in pages 101, 102, 103, etc.

Such discrete variables contrast with, for example the range of numerical possibilities in heat as measured by a thermometer. A Fahrenheit thermometer used to measure body temperature, could display a reading of 98.6° F, but also read 98.2°, 97.8°, 102.4°, and so on. If the thermometer had finer calibrations, gradations between temperature values could be even more precise. Temperature is thus a *continuous* variable. In the abstract, continuous variables can assume an infinite range of values. Common continuous variables include distance or miles per hour, since these could conceivably assume any value, though some may be more common than others.

18.1.3 Types of Variable Scales

Variables can be classified according to the *scales* of the categories, that is, whether and how continuous and discrete values are defined. These are:

1. **Nominal scale**: named categories for entities, such as left, right, red, yellow, *Homo sapiens*, *Bos taurus*, etc., are nominal scale variables. These require only

that the classificatory categories be *exhaustive and mutually exclusive*, that is, red can never be yellow, a human can never be a cow, etc.

2. **Ordinal scale**: "ordinal" implies a ranking, or ordering. This type of variable involves an ordering of discrete categories into *a meaningful sequence of classes ranked along a continuum*. However, the distance between two or more such categories is either unknown or undefined. For example, one can say A > B > C, but not stipulate by how much, or even whether, the interval between each variable is of the same magnitude. By the same token, "sedentary" is less mobile than "transhumant" to an unspecified degree, species A is less arboreal than species B, etc.
3. **Interval scale:** variables of this type possess all the properties of ordinal scale variables except that they also possess equal distances (*intervals*) between each variable category. Thermometers and calendars are interval scale. Interval scale devices do not necessarily begin their calibration at zero. For example, thermometers, which have zero as a value, have an interval scale that goes below zero. Because the intervals are stipulated in a quantifiable way, these scales *can be mathematically manipulated through addition and subtraction*.
4. **Ratio scale:** these variables are based on scales in which *the starting point is fixed* rather than arbitrarily defined. Such scales quantitatively express the relationship between physical properties, such as miles traveled per hour, number of inhabitants per hectare, number of beta particle emissions per 24-hour period, etc. In all cases, *zero is a fixed point in relation to the scale established*, whereas in the interval scales, only the space between points need be specified.

The distinctions outlined above are significant because certain statistical tests can only be applied in certain variable scales and not to others. Choosing the statistical test appropriate to the variables used in zooarchaeology depends on understanding the nature of its basic counting units and other associated values, such as nutritional indices and bone-mineral density indices, commonly used in the field.

18.1.4 Parametric and Nonparametric Statistics

Parametric variables fulfill certain statistical criteria, and tests applied to them assume that these criteria are met. Paraphrasing and reducing Siegel and Castellan (1988), parametric variables require the following:

1. Observations must be *independent*. That is, selection of any one case from a population in a sample must not bias the chances of any other case being selected.
2. The observations must have been randomly drawn from a *normally distributed population* (Fig. 18.2), with predictable implications for computation of their means.
3. Populations to be compared must have either the *same variance* or a known *ratio of variances*
4. Variables must be measured at *least in an interval scale*, so that normal arithmetic operations can be performed.

Tests that operate on these assumptions include comparisons of means and the variation around them. Some, such as the Student's t-test, allow greater leeway on these conditions; because the distributions need only be approximately normal, variables could exhibit largely independent errors, etc. However, even Student's t-test requires that the variables be ratio or interval scale (Siegel and Castellan 1988). As should be clear from examples given above, many variables of interest to archaeologists may not satisfy all or any of these requirements. Therefore, other ways of assessing whether two samples are similar or different are needed.

Nonparametric statistics are designed to be used in cases where not all the parameters of the variables of interest listed above are fully known, such as whether a variable has a normal distribution and if a standard deviation of the mean can be calculated. These tests can be used under *any one* of the following conditions:

The variables may be nominal, ordinal, and interval scale.

or

The variables are *not distributed randomly* but in a distributional pattern that is unspecified, i.e., not necessarily a normal distribution.

Since many variables studied by archaeologists, and specifically zooarchaeologists, have an unknown distribution, this last aspect of nonparametric statistics is especially important. Nonparametric statistical tests appropriate to each type of variable scale are summarized below.

1. Nominal scale variables may be subject to parametric statistical tests, such as the Chi-Square test or Fisher's Exact test, but analytic tactics such as regression analysis are not appropriate
2. Ordinal scale variables may be subject to the tests listed above for nominal scale variables and also to rank-order correlation coefficients, such as Wilcoxen Two-sample test, Kendall's tau, Spearman's rho, and the Kolmogorov-Smirnov test. Kendall's tau and Spearman's rho are often used in assessing the degree to which relative nutritional utility or bone mineral density indices predict element frequencies in an assemblage. The Kolmogorov-Smirnov test compares cumulative frequencies commonly run on ordered percentages, such as is the case with mortality profiles and survivorship curves. As with nominal-scale variables, regression analysis is not appropriate.
3. For interval scale variables, rank-order correlation coefficient tests, such as Kendall's tau and Spearman's rho, are appropriate, as are linear regression and correlation.
4. Ratio scale variables are amenable to all mathematical operations and thus to all parametric tests.

18.1.5 Correlation Coefficient Analysis

The correlation coefficient, r, expresses how two interval or ratio scale variables, x and y, are related to one another (co-related). Variables are often displayed on a bivariate plot (Fig. 18.3a–c), as are similar variables in regression analysis, but

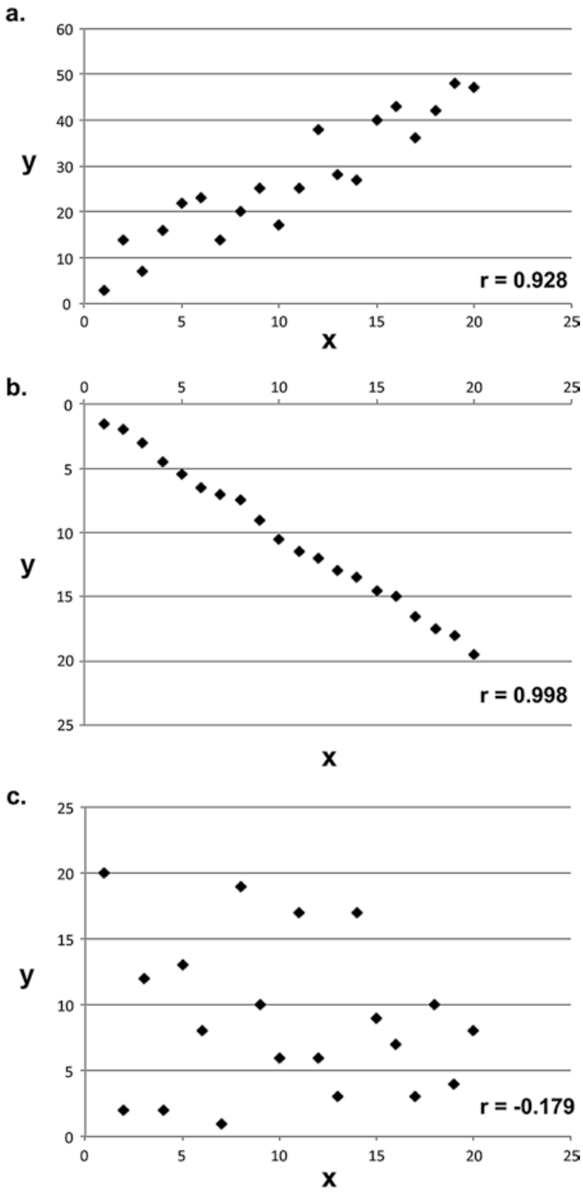


Fig. 18.3 Three cases in which the value of x relates to the value of y: **a.** a strong *positive* correlation exists between the values of x and y; **b**. a strong *negative* correlation, where increasing values in x correlate with decreasing values in y; **c**. a case in which no consistent relation exists between the two variables. Pearson's r values provided in each example. (Illustration by the author)

several key differences between these two approaches exist. First, correlation coefficients do not assume a causal relation between the two variables, as would regression analysis. In correlation coefficient analysis, x and y variables are theoretically interchangeable, whereas, for the purposes of regression analysis, values of x is assumed to determine values of y (see below). Second, correlation coefficient analysis does not assume that the variables have a normal distribution, whereas assumption of a normal (Gaussian) distribution is basic to regression analysis. Third, only regression analysis "fits a line" to an x-y variable distribution (see below).

Correlation coefficients can range from 1.0, reflecting perfect positive correlation, to 0.0, reflecting no correlation between the two variables examined. Correlations can be positive, with both values increasing together in a positive direction (e.g. Fig. 18.3a), or negative, with one value decreasing as the other increases (Fig. 18.3b). A negative correlation such as that shown in Fig. 18.3b always involves one positive variable; if both variables moved in a numerically negative direction, this would produce a positive correlation. It is worth reiterating that correlation does not equal causation.

18.1.6 Regression (Coefficient of Determination) Analysis

Regression analysis explores whether, given multiple cases, a given variable, x, determines the value of a second variable, y. For the purposes of exploration, x is assumed to be the *independent (determining) variable* while y is the *dependent (or determined) variable*. Regression analysis can explore simple linear patterns of relationship, such as those shown for total body weight vs. skeletal weight by Reitz and Wing (2008:64–70), or relationships in which the variables covary in a more curvilinear relationship, such as crown heights of molars as they wear over time (Chap. 4). More complex forms of regression analysis use multiple variables simultaneously. Regression analysis requires that both variables be interval or ratio scale variables. One cannot regress ordinal or nominal scale data; nonparametric statistical tests work for those scales.

Regression analysis has been used in zooarchaeology on variables as bone size in relation body length, practices largely drawn from established in wildlife management research (cf. Reitz and Wing 2008:186–187). In relationships such as that of bone size and body length, zooarchaeologists use regression coefficients of determination, derived from contemporary measurements of both variables, to estimate the value of one from value of the other when they estimate body length from an animal's bone length.

The example in Fig. 18.4 suggests that the variables involve a causal relation, but regression analysis itself does not *prove* causation. Zooarchaeologists as well as other scientists who use regression analysis may have a high coefficient of determination, but they still must investigate *how* these variables relate in functional terms, and the underlying causes of their relationship. For example, in 18.4, in this relationship, the number of cats is the determinative variable and volume of cat poop is the determined variable, however, despite our intuitive grasp of the functional relationships involved in this simplistic example, a scientific approach would ask that these be specified by direct observation, to confidently proceed assuming that the former has a causal relation to the latter.

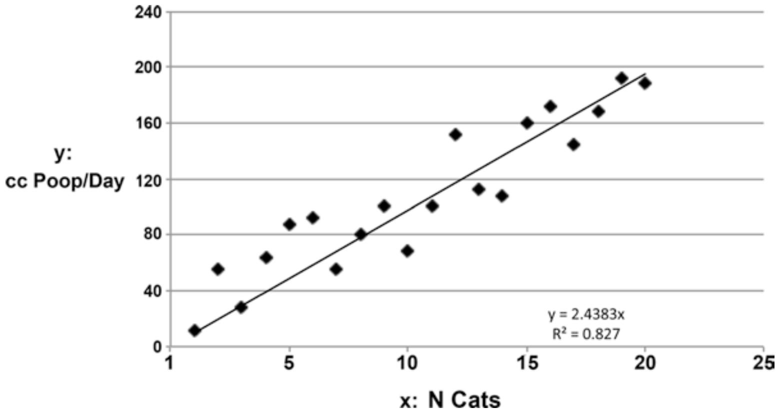


Fig. 18.4 A plot showing a distribution of variables with regression line, with formula for the intercept and R^2 (Illustration by the author)

18.1.7 Simple Linear Regression

In simple linear regression, a scatter of individual points of x/y values plotted, which is readily done with many computer applications, including Excel® and other database applications. Often, simply visually inspecting a plot will indicate whether a strong relationship exists between these variables; this can be expressed quantitatively as r, the correlation coefficient. Next, to specify how well (precisely) x predicts y, one multiplies r by itself, resulting in r^2. If there is a perfect correlation r = 1.0, then r^2 will also be 1.0. However, many strong r statistics are more like r = 0.8, in which case, r^2 is 0.64, meaning that 64% of the variability of y, the determined variable, is explained by the regression model. The r^2 is also called the *coefficient of determination*: the higher the r^2, the more of the variation in the data is accounted for by the regression line. Recall that 1.0 would represent perfect determination, so how good is the r^2 of 0.64? This is done by assessing the probability that the results were not associated with the action of x in the regression model. The *p*-value is calculated assuming a normal curve. The lower the *p*-value, the more likely the action of x on y is the cause of the patterning.

Another way of visualizing how well x accounts for the variation in y is to present the data with two lines running parallel to the regression line, showing the 0.05 confidence levels of estimates predicted by the regression line, that is, two *standard errors* on either side of the average in a two-tailed normal distribution. The same estimate error range can be shown for individual points, each with their own standard errors of their mean (e.g. Fig. 18.5). The standard error quantifies the *precision* of the mean, or average, value in a sample. This is in turn derived from th*e standard deviation*, the degree to which individual values in a sample differ from the sample mean. The standard deviation ("± value") is the sum of the squares of the difference between each value in the sample and the sample mean, divided by the total number

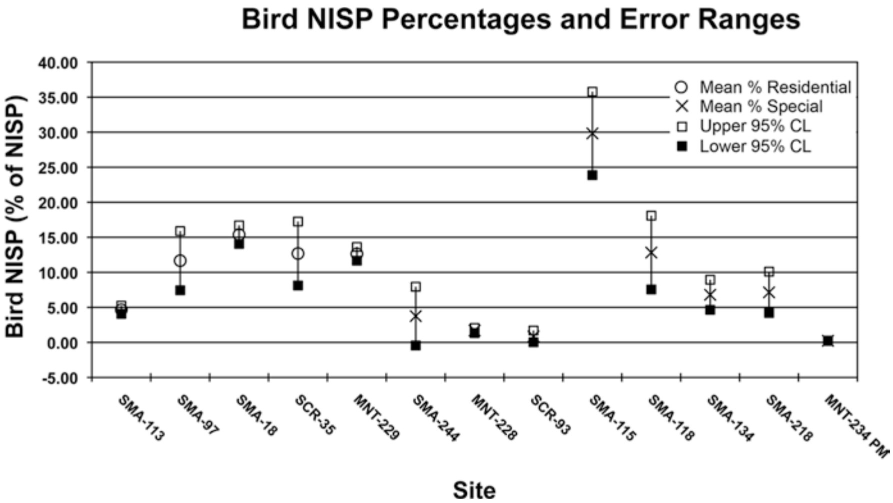


Fig. 18.5 A plot showing the 0.05 confidence levels (2 standard errors) of mean bird NISP frequencies at 13 archaeological sites in southern San Mateo, Santa Cruz, and northern Monterey counties. Longer-term residential sites are marked with an open circle, while short-term residential and special purpose sites means are marked with an X (From Gifford-Gonzalez et al. 2013:307, Fig. 2, produced by R. Cuthrell, used with permission of Taylor and Francis and Society for California Archaeology)

of sample values minus 1, with the square root taken of the resulting number. A cogent discussion of "why N-1?" can be found in Motulsky (1995-2015).

Again assuming that the parent population follows a normal distribution, one standard error accounts for about 67% the variation around a mean, while the two standard error space accounts for about 95% of the variation around a mean. Graphically, the more space between the regression line and the 0.05 confidence lines, the more highly variable the values in the sample are.

18.1.8 Spearman's Rho and Kendall's Tau, Nonparametric Correlation Coefficients

Chapters 20 and 21 outline how, during the 1990s, zooarchaeologists debated whether nutritionally motivated human selectivity or bone durability has more influence on creating the structure of certain archaeofaunal samples. Zooarchaeologists Grayson and Lyman recognized that both nutritional value indices (e.g. MGUI) and bone mineral density indices were ordinal scale variables, in which each index value was a ratio, but that the distance between any two skeletal elements' values was not the even interval required by parametric statistics such as regression analysis. They therefore advocated using one of two nonparametric tests of the correlation of element frequencies against these indices, Spearman's rho (or R) and Kendall's tau (or T), rather than regression, which requires ratio scale variables. Both Spearman's rho and Kendall's tau basically assess how well x (in this case the index being studied) predicts the frequencies of the osteological elements, up to now usually expressed as %MAU or %NISP.

Statisticians consider that Kendall's tau and Spearman's rho have similar basic assumptions and statistical power (Siegel and Castellan 1988). Yet Kendall's tau has a very different computational formula than Spearman's rho because it expresses *probabilities*, rather than the *amount of variability explained*, as does Spearman's rho. The Kendall's tau product represents the difference between the probability that the two compared variables are in the same order and the probability that they are in a different order. Spearman's rho computes its outcome much like the simplest regression correlation coefficient, the Pearson product moment correlation coefficient: its outcome expresses the proportion of variability accounted for by the correlation. However, Spearman's rho is computed from ranks rather than interval or ratio scale variables (StatSoft 2013). Spearman's rho assumes that the variables assessed are at least ordinal scale, so individual observations can be ranked into two ordered series. For example, in any assemblage, one can list element frequencies in a rank from highest to lowest NISP. Likewise, bone mineral density (BMD) per cubic cm for various elements can be arranged from the highest BMD values to the lowest.

18.1.9 The Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov test is a nonparametric test that compares the cumulative distributions of two datasets to assess whether they differ significantly. The Kolmogorov-Smirnov test makes no assumption regarding the normal distribution of the data but depends on a continuous distribution of the variable graphed Siegel and Castellan 1988). Classically, it is used to compare a curve of unknown distribution of the parent population against a cumulative curve produced by a normal distribution. However, it can also be used to compare two sample curves. Technically, it is a goodness-of-fit test between two curves. It produces a D-statistic that expresses the degree of discrepancy between the two distributions and hence whether the hypothesis that the two curves are similar – or, that the "observed" of the one matches the "expected" of the other – is supported. In zooarchaeology and other areas of archaeology, this test has normally been applied to cumulative percentage frequencies of values of a variable (Siegel and Castellan 1988). The Kolmogorov-Smirnov test has been used by Klein to compare mortality profiles, which plots cumulative frequencies of ages at death, or mortality profiles (Chap. 22).

The balance of this chapter explores some of the confounding effects of using NISP, MNI and other derived measures as the basic variables of comparison in zooarchaeology.

18.2 Problems with NISP, MNI, MNE, and MAU

Whether they are interested in shifts in species abundances with climate change over time, or in comparing the relative frequencies of long bones to axial elements from two sites, zooarchaeologists are dealing with measures of relative abundances. Even before deciding on appropriate tests for assessing similarity and difference between or among samples, a researcher must choose the counting unit to be used to make the comparisons. The fundamental question for someone working with archaeofaunas is whether one's choice in counting unit affects the patterning perceptible in quantitative data. This section will draw upon Grayson's (1984) classic assessment of these issues, in his book, *Quantitative zooarchaeology*, as well as Lyman's (2008) *Quantitative paleozoology*, and other discussions of statistical methods in zooarchaeology e.g. (Marshall and Pilgram 1993; Morin et al. 2017a; Ringrose 1993; Pilgram and Marshall 1995), including experimental research on the performance of various counting units under controlled circumstances where taxonomic and skeletal element "input" was known (Morin et al. 2017a).

Grayson has a long history of research archaeofaunas of the western United States and has written extensively on quantification of faunal data (Grayson 1978, 1979, 1981). Much of Grayson's earlier research focused on the paleobiogeography and diachronic changes in relative abundances of taxa in the Great Basin of western North America, using NISP or percentages of NISP of different species. *Quantitative zooarchaeology* (Grayson 1984) focused on the relationship between analytical methods and perceived patterning in archaeofaunal data. By experimentally manipulating data from his own and others' analyses, Grayson explored the possible effects of sample size and choices in archaeofaunal sample subdivision in relation to the quantitative units chosen on species abundance data, arguing that some "patterning" could simply be produced by analytic choices in combination with specific quantitative measures.

18.2.1 Problems with NISP

NISP is the most basic statistic in zooarchaeology. It is "primary data" in the sense that it is the total of identifiable (and, as used by some researchers, less identifiable) specimens that can be counted. As a raw count, NISP appears to behave as a continuous, interval scale variable, with increments of one between each value. Moreover, NISP is "set" to begin at zero, so it has at least one property of a ratio scale variable. It appears therefore to be potentially amenable to parametric statistics.

However, when NISP is used to estimate the proportional representation of various taxa in an archaeofauna, several biologically intrinsic properties and postmortem taphonomic processes can bias relative taxonomic proportions. These include:

1. The variable numbers of skeletal elements in bodies of different taxa.

2. Human taphonomic factors, including
   (a) differential effects of butchery and transport on bodies of different-sized taxa and
   (b) differential intensities of breakage on skeletal elements of taxa with different within-bone nutrient levels.
3. Non-human taphonomic biases in taxonomic representation through differential destruction among elements with different durability.
4. Collection biases (screen size, visual inspection, etc.).

Moreover, Grayson (1984) notes that we can normally assume that it is highly likely that more than one skeletal element of a given taxon in an archaeofauna actually derives from the same individual, and therefore NISP cannot guarantee the *specimen independence* required by nearly all the statistical tests. Given this, some have opted to use MNI to correct for these problems. To these general problems, one should add a consideration of the Morin et al. (2017a, b) assessment of NISP's performance as an estimator in an experimentally controlled situation (see Sect. 18.3).

18.2.2 Problems with MNI

Because it depends on the most numerous unique element of a given taxon, MNI would seem to resolve the problems created by specimen interdependence, differing numbers of bones in various vertebrates' bodies, differential recovery methods, and some differential transport and processing. While MNI may help offset the effects of modest differences in transport and processing of different taxa, but this has logical limits. For example, MNI cannot "correct" for the absence of elements that were reduced into small, unidentifiable scrap through processing, nor can it testify to the presence of flesh from animals, such as flensed whales, whose bones were never transported to a site.

Despite MNI's efficacy at coping with some problems inherent to NISP, Chap. 10 foreshadowed problems with MNI that affect it as a measure of relative element and taxonomic abundance, and thus as an effective means of comparing samples within or among sites. These can be divided into four areas:

1. Effects of different sample aggregation strategies on MNI.
2. Faulty assumptions of specimen independence for MNI in stratified sites.
3. Effects of differences in sample size (expressed as NISP) on MNI.
4. Problematic assumptions about MNI in relation to carcass utilization.

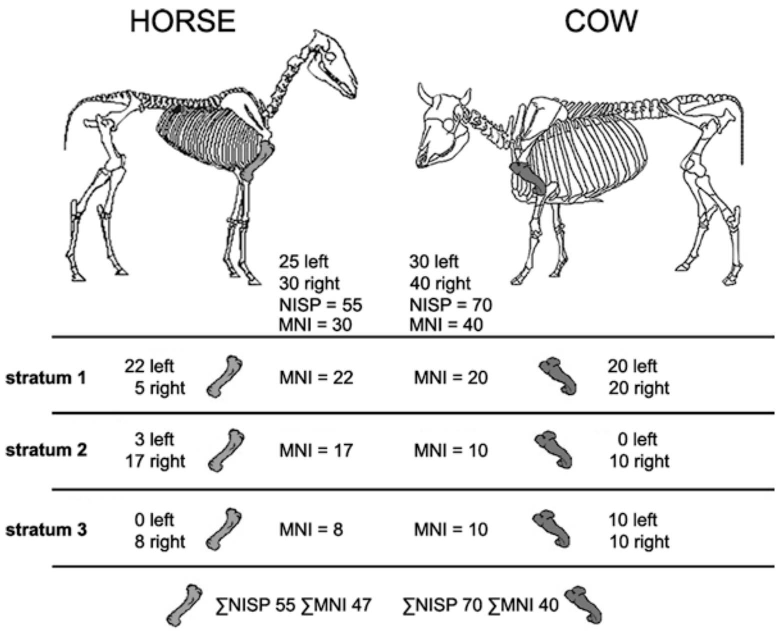
The next sections discuss each of these in turn.

18.2.2.1 Effects of Aggregation Strategies on MNI

Grayson (1984) demonstrated that MNI is sensitive to the relative fineness with which a site is subdivided, or "aggregation effects." Aggregation refers to the strategy of grouping specimens in a site's archaeofauna into sub-assemblages, based on various criteria. For stratified sites, this commonly involves aggregating the archaeofauna into stratigraphic samples according to matrix lithology or artifactually distinctive strata. An example is Klein's (1978, 1979, 1981) aggregation of small lithologic unit samples from Elands Bay Cave and Klasies River Mouth Cave 1, South Africa, into larger Middle Stone Age aggregates, based on the overall stoneworking traditions associated with those strata. This allowed Klein to compare age structures in these faunal samples with those of similarly aggregated Later Stone Age samples from these and other South African sites. Another common aggregation practice in archaeology is treating house-floor or house-compound assemblages from a site as subunits and comparing trash-pit samples independently, as did Crader (1984, 1989) for Thomas Jefferson's main residence and slave quarters at Monticello, Virginia.

These practices are often logical from an archaeological viewpoint, but they can produce unintended zooarchaeological consequences, by inflating MNI estimates for various taxa in an unpredictable manner. Figure 18.6 shows data originally presented by Grayson (1984:31) in a different format, illustrating a hypothetical example in which more finely subdividing an assemblage produces higher total MNI figures for the site. This results because the most abundant (i.e. MNI-diagnosing) elements of each taxon are *not evenly distributed* through the entire stratigraphic section. Rather, they are distributed in highs and lows that manifest as varying MNI estimates for different layers into which a stratigraphic section – and the associated archaeofauna – has been subdivided. The high and lows of horse specimen abundances do not track identically with those of cattle because the entry of one taxon into a human campsite or settlement normally doesn't depend on the presence of another taxon. We might even expect independence in the spatial distributions of two taxa to be the case, as with a wild mammal species versus a domestic one. Subdivision of a stratigraphic section can thus produce differing effects on MNI of one taxon in relation to another. This example shows that the patterning perceived in the MNI data, and in the relative abundances of the two taxa, may depend not on some "real" variations in relative abundances, but rather on the units into which a stratigraphic section is subdivided – or aggregated. It follows that the story a zooarchaeologist could tell from these relative abundance data could differ as well.

Especially with stratified sites, whether patterning in archaeofaunal data is just an artifact of how the site was subdivided is of special concern. Archaeologists aim to excavate in rather small stratigraphic slices to maintain maximum control over specimen provenience and, they hope, maximum temporal resolution. In the United States, it has been common practice to excavate in arbitrary 10 cm levels rather than according to sedimentological units, while in continental Europe, very small sedimentological deposits are often dug as discrete units, but some are later aggregated according to their artifactual contents or other criteria. In either case, dividing or



stratum 1	22 left 5 right	MNI = 22	MNI = 20	20 left 20 right
stratum 2	3 left 17 right	MNI = 17	MNI = 10	0 left 10 right
stratum 3	0 left 8 right	MNI = 8	MNI = 10	10 left 10 right

∑NISP 55 ∑MNI 47 ∑NISP 70 ∑MNI 40

Fig. 18.6 Hypothetical case in which differing aggregation effects emerge through stratigraphic subdivision of a sample. NISP and MNI at the top are based on treating the entire sample as a single aggregate. The subdivided sample produces different total MNI for one of the two species (horse), and therefore a different relative proportion of horse to cattle, compared to the undivided one. (Illustration by author, using data from Grayson 1984:31, Table 2.2)

lumping levels and archaeofaunal samples as opposed to keeping them distinct can impact derived MNI estimates. Criteria for choosing how to subdivide a stratified site may have little to do with the archaeofaunal remains themselves, but zooarchaeologists must be aware that these decisions ultimately can have a major influence on MNI. Not only does this problem affect MNI as a relative abundance statistic, but it also affects other measures of abundance such as Minimum Number of Elements (MNE) and Minimum Animal Units (MAU). Moreover, Grayson (1984:45–48) notes that aggregation effects will impact estimates of other qualities of a sample, such as percent survival of various elements, if these are based on ratios of MNI values.

Hope exists for revealing and resolving this problem in a given dataset. Grayson proposes two strategies for deciding whether the vertical distribution of faunal remains in a site is susceptible to such aggregation effects. The first involves monitoring the distribution of most abundant elements (MAE) for all taxa across the finest possible subdivisions of the stratigraphic section (Grayson 1984:33). If MAE distributions vary from level to level, then aggregation effects are likely to skew relative abundance data. Another approach is to calculate MNI for taxa first in the site sample as a whole, then for major stratigraphic units, and finally for the finest stratigraphic subdivisions possible (Grayson 1984:34). If MNI values vary from one aggregation scheme to the next, then aggregation effects are a potential problem in "reading" the MNI statistics.

Grayson notes that the same problems apply to using MNI in intrasite comparisons of single-component sites, if faunal remains from various features are aggregated in different ways -- for example, treating all trash pits in a site as a single sample vs. comparing Trash Pit A, B, and C. It is therefore critical to assess such intrasite or within-level datasets for the possibility of aggregation effects and to use MNI with utmost caution – or not at all – when aggregation effects appear in the recommended evaluations.

Grayson points out that NISP does not suffer from aggregation effects, and therefore it can be used in situations in which aggregation effects appear to be affecting MNI statistics. He further contends that for subdivided stratified sites, NISP's costs in specimen interdependence are outweighed by its benefits in avoiding aggregation effects. Grayson contends that MNI in stratigraphic sites often does not actually escape specimen interdependence (see 18.2.2.2 *Effects of Specimen Interdependence on MNI in Stratified Sites*).

Recall that Grayson's criticisms apply *only to the use of MNI with sites, and with sampling tactics, in which aggregation effects could operate*. This problem does not affect MNI-based comparisons between sites, if each site is treated as a single sample and specimen interdependence is not a reasonable issue. For example, I might use MNI figures to compare taxonomic abundances of the Prolonged Drift archaeofauna (Gifford et al. 1980) to those of Ngamuriak (Marshall 1990), around 200 km away, since neither sample was subdivided, and it is reasonable to assume specimen independence. Grayson does question the applicability of MNI statistics for intersite comparisons, but they are based on the relation of MNI to sample size (expressed as NISP), to be discussed below.

18.2.2.2 Effects of Specimen Interdependence on MNI in Stratified Sites

MNI seemingly offers an escape from specimen interdependence, but in stratified sites, we have no assurance that the elements of a taxon drawn from one level *necessarily* exclude elements from the same individuals in another level. This is due to site formation and excavation recovery tactics and element distributions across strata.

Strata, even those defined by sedimentary differences, are arbitrary subdivisions imposed by excavators, which may transect elements of one skeleton that do not "map" congruently onto the sediments. Skeletal elements at the top of a thick stratigraphic unit are often spatially closer to elements in the next stratigraphic unit than they are to elements at the lower end of their own stratum. Simply for that reason, they could derive from a single vertebrate individual with skeletal elements in two strata. Even elements recovered from lithologically distinct stratigraphic units, are not *necessarily* from different animals. Processes that form small-scale, sedimento-

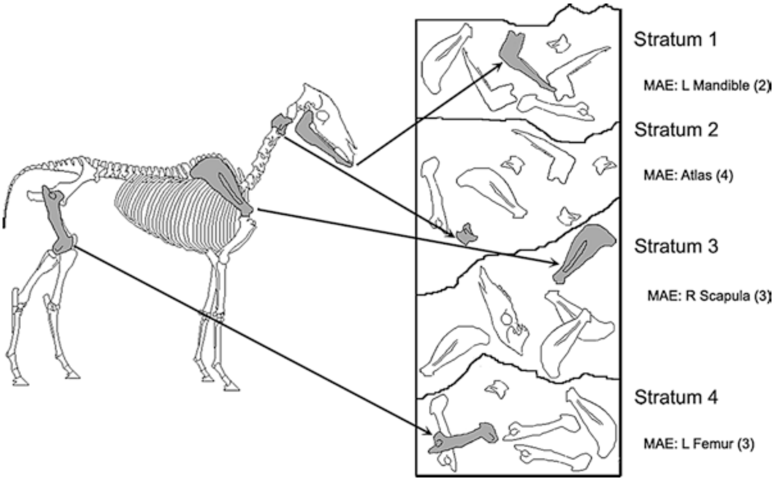


Fig. 18.7 A hypothetical case in which elements of one individual horse (shaded) occur in several strata: a mandible, an axis, a scapula, and a femur. Each of these elements is counted as one of the most abundant elements (MAE) in reckoning the MNI for each layer, thereby introducing specimen interdependence into the MNI statistic (Illustration by the author, inspired by discussion in Grayson 1984)

logically distinct deposits operate independently of human activities, and elements of one carcass theoretically could end up in laterally varied depositional zones – a puddle in a depression, a gravelly high spot – each contemporaneously creating its own depositional matrix.

Moreover, refit studies of bone fragments, ceramics, and lithics show specimens may join together across as much as a meter's depth, transgressing major sedimentological distinct strata lacking rodent or insect disturbance (Villa 1982; Villa and Courtin 1983; Cahen and Moeyersons 1977; Todd and Stanford 1987). This suggests that, like pieces of a broken pot or flakes from a single core, bones of an individual skeleton may appear in more than one stratum.

Because MNI relies on the most abundant element (MAE) in each level, and because different elements may be used to calculate MNI from level to level, a possibility exists that a single animal's skeletal elements could be counted in different levels. Figure 18.7 illustrates a case of specimen interdependence in which the MAE of equids in several strata of a site each incorporate an element derived from one individual.

Summing up, evidence suggests that specimen independence should not be assumed, but rather be demonstrated, for faunal remains in stratified sites. What we know of site formation processes supports Grayson's assertion that, unless one is comparing the same element and side across stratigraphic levels, one should not assume the elements derived from different individuals.

18.2.2.3 Effects of Differences in Sample Size (NISP) on MNI

The preceding sections stipulated that *intersite* comparisons of relative abundance using MNI are not affected by aggregation effects, so long as the respective sites' faunas can be treated as integral samples with no specimen interdependence. Likewise, specimen interdependence problems can be presumed not to affect most sites. However, intersite comparisons using MNI is affected by another factor: the dependence of MNI values at a site on the size of the NISP that generated them. In other words, there is a relation between NISP and MNI that may affect comparisons of substantially different-sized assemblages.

Several researchers, starting with French zooarchaeologist Pierre Ducos (1968), noted that a systematic relationship exists between the size of an archaeofaunal sample (NISP) and the MNI statistic. Ducos compared these data for several early food-producing sites in the Levant and found that the logarithms ($\log_{10}$) of NISP related to the logarithms ($\log_{10}$) of MNI in a consistent, linear fashion. One important finding of Ducos' exploration was that MNI figures tended to overemphasize the abundance of rare (low NISP) species relative to more abundant (high NISP) ones. Ducos' work was essentially replicated by Casteel (1977), a zooarchaeologist, and by Holtzman (1979), a paleontologist, in a related discussion of the likelihood of drawing new unique individuals with increasing sample size.

The situation with rare specimens is as follows: if one has a single element of, say, a rabbit from a site, its MNI is 1. Given the nature of most archaeological bone deposits, it is highly unlikely that drawing a second rabbit specimen will yield a specimen that is not only a rabbit but also the same element and side, thus increasing the MNI to two. Therefore, the relationship of NISP:MNI that started as 1:1 now changed to 2:1, and it's probable that one would have to retrieve many more rabbit specimens from a deposit before finding two elements with which to calculate a new MNI of 2. Let us say that the hundredth specimen was the same as the first, producing NISP of 100 and MNI estimate of 2, or a ratio of 50:1. From this, one can get a sense that very rare elements of one species thus carry relatively more weight when "translated" into MNI than more numerous specimens.

Unlike Casteel, Grayson (1984:53) did not believe that a one-size-fits-all formula for this relationship held, but that the slope and curve must be empirically derived for any specific archaeofaunal sample of NISP ≥1000, using the power function $\text{MNI} = a(\text{NISP})^b$, where a and the exponent are derived from the best fit regression line. Grayson further argues that the relationship between NISP and MNI is essentially curvilinear (hyperbolic) in most faunas examined, with the precise shape of the curve varying from site to site, but notes that it can be linear under certain conditions.

The fact that NISP and MNI varies from assemblage to assemblage probably stems in part from carcass processing strategies and in part from the variable number of osteological elements in different taxa (see Lyman 2015). For example, if people regularly killed larger prey animals far from their home camp, they might only bring selected elements back to their base camps. This would lower the ratio of NISP:MNI, because only a small range of elements drawn from whole carcasses would be likely to enter the site and be discarded there, but those that *were* transported would be systematically selected and hence likely to produce higher MNI (Chaps. 19 and 20). By contrast, pastoralists killing their livestock at their home encampment would probably produce higher NISP:MNI ratios because each kill would contribute more elements to a residential faunal assemblage than would hunters' selectively transported body segments (Fig. 15.2).

In sum, statistical tests comparing MNI give equal weight to taxa with NISP:MNI ratios of 1:1 as to those represented by NISP:MNI of 100:1 or even 1000:1. Grayson (1984) advises using MNI with a good deal of caution in such comparisons. No *a priori* method exists for deciding how close ratios must before they are "similar" enough to compare, nor how far away their values must be before they are "different." Is 50:1 different enough from 40:1 that we shouldn't compare MNI with those different ratios? What about 30:1?

Grayson suggests that, because the NISP:MNI problems are compounded with the aggregation effects in stratified sites, one should use NISP instead of MNI for comparing relative abundances. His argument goes as follows: MNI statistics derived from varying-sized NISP are problematic entities for statistical comparisons. Plots of MNI, or of ratios of MNI to NISP against NISP for all sites studied show a consistent relationship between the two variables (Fig. 18.8). Therefore, the information contained in MNI is also inherent to NISP. Since NISP presents none of the size-dependence (nor aggregation effect) problems that MNI does, why not just use NISP? This assertion, and experimental findings on the performance of NISP and other measures will be dealt with in 18.4 What Do You Want to Do, What Tool Do You Need? later in this chapter.

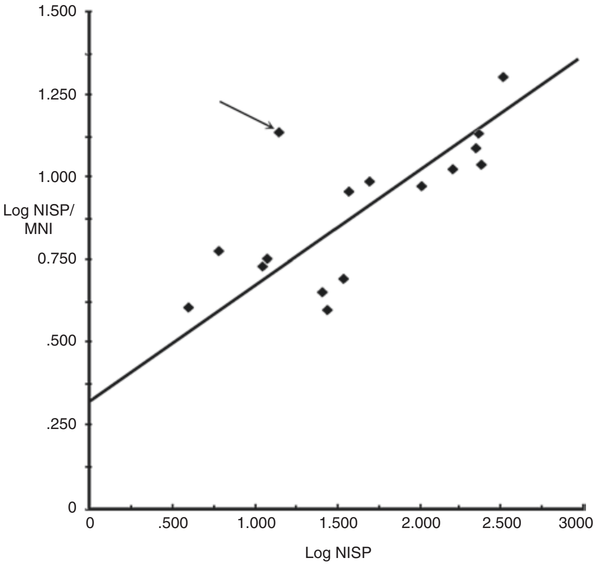


Fig. 18.8 The relationship between MNI and NISP, expressed as $\log_{10}$, for the site of Prolonged Drift. Arrow indicates an outlier in the ratios plotted, a subject further explored in Chap. 22. (Redrawn by the author from Grayson (1984:71, Fig. 2.17), in turn based on data from Gifford et al. 1980)

18.2.2.4 Problematic Assumptions About MNI in Relation to Carcass Utilization

The final problem regarding MNI is of a different order than those outlined above. This has to do with underlying assumptions about the meaning of MNI figures in terms of carcass utilization. Binford (1978) pointed out that researchers have often worked with MNI on the assumption that it represents the minimum number of entire animals once present at a site. Meat weights, for example, are often derived by multiplying the MNI by the average meat yield of an entire carcass of that species. Based in part on his observations of Nunamiut butchery patterns and in part on his preoccupation with possible scavenging strategies in early hominins, Binford questioned this assumption. He contended that the count of elements at a site can, at most, be taken to represent the number of *body segments* present. Lyman (1979) advanced the same criticism of MNI, arguing that the appropriate unit of analysis was not MNI but a "butchery unit" based on the bones actually analyzed and indications of actual carcass subdivision patterns based on butchery marks.

Parallel issues exist with Minimum Number of Elements and Minimum Animal Units as measures of abundance.

18.3 Problems with MAU and MNE

Grayson (1984) criticized the MAU statistic on two counts. First, because it is derived from another unit of counting, the Minimum Number of Elements (MNE), it is liable to estimation problems inherent in the MNE. He concluded that, as a derived estimator of abundance, MAU is sensitive to the same aggregation effects as MNI. Therefore, it should be used with the same caution as MNI in stratified sites and in single-component sites with distinct bone-bearing features. Second, MAU assumes specimen independence from sample to sample, so the same *inter*dependence problems outlined for MNI could also afflict MAU estimates in stratified situations (Grayson 1984). For better or worse (Chap. 10), Binford's %MAU statistic dominated much of the literature on nutritional utility for a decade, if only because archaeologists wanted to compare their data to those published by Binford. Readers of that literature should consider Grayson's critiques and recommendations for evaluating problematic aspects of these estimates.

In the 1990s and 2000s, many assessments of skeletal element abundances shifted from MAU to MNE (Marean et al. 2001; Marean and Spencer 1991; Domínguez-Rodrigo 1997; Pickering et al. 2008; Pickering et al. 2006). Uses of both MAU and MNE in zooarchaeology generally parallel an increasing concern with prey handling, including butchery, selective transport, culinary processing, and varieties of distribution of body segments (see also Lyman 2008:240–241). This trend supplemented zooarchaeology's earlier and continued use of NISP and MNI to characterize the taxonomic composition of archaeofaunas, which enabled analyses of environmental change, prey species choice, and, by extension, discussions of different hominin species' predatory capabilities.

Despite this trend, two sets of scholars questioned whether MNE is essential to assessing element frequencies. Grayson and Frey (2004) and Lyman (2008:214–249), engaged in extended, statistically buttressed discussions of whether the MNE statistic is in fact preferable to NISP when assessing the relative abundance of elements. This discussion is quite similar to Grayson's earlier (1984) discussion of problems inherent in MNI estimates as the basis for taxonomic abundances. Just as MNI corrects for disparities in element abundances across different taxa but is subject to sample-size and aggregation effects, so, too, does MNE correct for differences in degree of fragmentation across different elements of a given taxon, but it is subject to the same sample-size and aggregation effects as MNI.

Grayson and Frey (2004) used three large site archaeofaunas analyzed by others Elandsfontein, South Africa (Klein and Cruz-Uribe 1991), Kobeh Cave, Iran (Marean and Kim 1998), and Rond-Du-Barry, France (Costamagno 1999) to explore this issue. They showed that a "normed NISP" ("the skeletal part NISP values divided by the number of times the relevant part occurs in the skeleton of the animal involved," Grayson and Frey 2004:29) correlates so highly with both MNE and MNI ($r \geq 0.90$) that, regardless of whether one wants to examine species abundance or selective transport of elements, it should work equally well as these other, NISP-derived measures. Lyman (2008) agreed with Grayson and Frey's contentions regarding MNE, and used more examples of the tight relationship between NISP and MNE, to advocate for using NISP rather than MNE to assess skeletal element abundances by taxon. Lyman (2008:250–354) also engaged more with MNE as an index of fragmentation, when used in a ratio with NISP, which will be explored below. Moreover, methods for calculating MNE lack standardization, as pointed out by Lyman (1994) and Marean et al. (2001).

Insights from blind test experiments by Morin et al. (2017a, 2017b) have shed further light on which of these measures of element and taxonomic abundance offer estimates closest to actual known assemblage composition. Morin et al.'s (2017a) experimental program used a known input of fragmented , primarily of red deer/North American elk, to assess both the replicability and accuracy of element and taxonomic identification among analysts, as well as the relative performance of zooarchaeology's three main quantitative measures of abundance: NISP, MNI, and MNE. Their research reported not only on overall experimental results but also on the relative success of region-by-region identifications (e.g. long bone epiphyses versus diaphyses) and explored the possible sources of poorer outcomes. Chapter 9 reported on Morin et al.'s inter-analyst performance in identifying large, fragmentary assemblages produced by marrow extraction and by bone grease production. This section focuses on two other experimental results: first, the overall success of NISP, MNI, and MNE in estimations that approach those of the original experimental "inputs" and second, variations in successful element frequency estimates for various segments of the ruminant bodies in relation to the actual frequencies of elements and individuals in the experimental set.

Of special interest is the performance of NISP, MNI, and MNE respectively in identifying fragmentary long bone diaphyses to skeletal element. Some zooarchaeologists working with Pleistocene archaeofaunas developed strong arguments for identifying fragmentary long bone shafts and methods to do so (Chap. 21).

Their motivations stemmed from naturalistic and experimental observations of hyena bone processing. These indicated that long bone diaphyses – among the most durable components of the mammal skeleton – remained after hyenas had consumed epiphyses (Chap. 12, see also Chap. 17). Hyenas were considered to be proxies for other large carnivores, such as wolves, inhabiting much of Pleistocene Eurasia and Africa during the span of human evolution. Thus, the reliability of identifications based on diaphyseal fragments emerged as a particular concern to those working where large carnivore impacts on bone assemblages were a likely taphonomic factor.

The three blind test participants in the Morin et al. experiment outlined in Chap. 9 were asked to calculate NISP and MNI according to an instruction guide, which also asked them to record specific landmarks and regions of long bones for reckoning MNI (see Morin et al. 2017a:899-900 for details). MNE values were derived from summing MNI for left and right sides. Thus, all three participants followed the same estimation procedures.

Outcomes of the experiment included some that contradicted longstanding assumptions. The much higher levels of fragmentation of the bone grease production set were expected to lead to lower rates of accurate specimen identification (NISP) in this sample versus the marrow cracking assemblage (Cannon 2013; Marshall and Pilgram 1993). Instead, results were comparable. Morin et al. chalk this up to the fact that bone grease manufacture produced more epiphyseal fragments, which were more identifiable.

Statistical assessment of estimation results showed that MNE was straightforwardly replicated among the analysts *and* generally provided more accurate estimates of original relative skeletal abundances than did NISP. Morin et al. explored the sources of NISP's poorer performance by assessing the effects of differential fragmentation rates among each long bone, discussing the implications of this inter-element variation for specimen identifiability. Different long bone elements were shown to break consistently into divergent numbers of fragments (Morin et al. 2017a:925, Table 18). This was interpreted as a major factor responsible for differing levels of long bone identifiability. However, the relative identifiability of various elements in the marrow cracking assemblage differed from those in the bone grease extraction assemblages, arguing against using an element-specific constant, or correction factor, for fragmentation rates (Morin et al. 2017a:925).

In sum, the performance of both NISP and MNE in estimating original skeletal element abundances *across analysts* was generally quite good, with MNE being generally better and more easily replicated among analysts than was NISP. However, the performance of *any* measure in abundance estimates using long bone fragments, where NISP did especially poorly, a special concern for those concerned about carnivore attrition in archaeofaunal assemblages.

Morin et al. (2017a:927) stress that their results should not be taken to suggest that MNE is a cure-all for quantifying relative skeletal abundances, for the reasons already cited by Grayson (1984) and Lyman (2008). They stress that MNE is susceptible to the same problems as MNI, highlighting its potential to inflate the importance of rare specimens in relation to numerous ones. The authors conclude with a call for greater standardization in recording MNE, MNI, and even NISP, as these will continue to be used as important measures of abundance in their application to different problems.

In a second article, Morin et al. (2017b) explored MNE's tendency to inflate the importance of low-NISP elements across 58 Middle and Upper Palaeolithic archaeofaunas from Western Europe. Their approach differed from that of Grayson and Frey (2004) and of Lyman (2008) because they monitored the relationships of the NISP to MNE across 24 classes of skeletal parts. NISP-MNE relationships were explored by comparing coefficients of determination in linear and curvilinear (power function) regression analysis. A better fit with the latter implies that the two measures do not increase commensurately as NISP increases. They found that in very small, NISP≤50 samples, power functions better described the relationship of NISP to MNE, whereas the higher the NISP, the more linear was the relationship, although some elements sustained a strong curvilinear relationship. Morin et al.'s (2017b) details of high- and low-scaling exponents for different elements and their observations on the behavior of different skeletal elements in the NISP-MNE relationship bears careful attention by analysts because these create expectations for the "behavior" of different fragmented long bones.

Finally, Morin et al. (2017b) introduced an alternative measure of element abundance: Number of Distinct Elements (NDE). This is a tally of the number of times at least 50% of a stipulated diagnostic landmark occurs in a sample, with each occurrence scored as "1." Landmarks for each element for cervid and bovid long bones are listed and illustrated, with some supplemental instructions for estimation, in the article (Morin et al. 2017b: Table 4, Figs. 9–11). The 87 NDE landmarks are not an exhaustive list of osteological landmarks but rather those that the researchers, all of whom have extensive experience with Pleistocene archaeofaunas, often encounter in these samples. They assessed the accuracy of the NDE using the same experimental marrow cracking and bone grease extraction assemblages as was used in their earlier study. Rank order correlation coefficients between element abundance estimates based on NDE tallies and the actual element abundances of the inputs were very strong (Morin et al. 2017b: Table 5), although, again weaker for long bone regions than for other elements (Morin et al. 2017b: Table 6). They state, "These observations suggest that the NDE is as robust as MNE for estimating skeletal, and possibly, taxonomic abundances." (Morin et al. 2017b: 956). Advantages enumerated for this method are:

1. NDE counts are more easily calculated than MNE counts.
2. The measure is inherently more standardized than the MNE method.
3. NDE values are expected to increase linearly with NISP sample size.
4. The NDE approach does not suffer from the aggregation problems.

The authors add that the NDE approach is similar to that developed recently to calculate mollusk abundances in archaeomalacological samples (Harris et al. 2015; Mason et al. 1998).

Given my own present inclination to opt for using landmarks in element identification and quantification (Chap. 10), I recommend that interested researchers road test this method as a measure of skeletal element and taxonomic abundance. As the authors state, one "sacrifices" the NDE method requires is excluding identifiable fragments that lack one of the small number of landmarks listed for each larger element. However, this does not mean that researchers must discard such specimens from their datasets, nor that such pieces would not be useful in zooarchaeological analyses *other than measuring abundances*. For example, if one were interested in evidence for handling of deer elements, one could use all specimens to assess cutting edge and percussion damage to skeletal elements, so as not to exclude specimens that testified to consistent patterns in the placement of hammerstone impacts, chops, cuts, and so forth. The tradeoff here is that one may gain a quickly recorded method of quantification that is not subject to aggregation and sample size effects in the same ways as MNI and MNE, and is more reliable in its abundance estimates than NISP appears to be. The NDE approach could be used by cultural heritage/resource zooarchaeologists without much additional work, thereby adding valuable, comparable data on species abundances, and even element abundances, for sites that they must work with efficiently and swiftly. One concern about the NDE parallels that raised by Faith and Gordon in (2007, Chap. 21) regarding Marean and Cleghorn's recommendation to restrict inter-assemblage comparisons to high-BMD specimens only. Faith and Gordon noted that this might reduce some sample sizes so much that they are liable to Type I and Type II errors. This is one of the reasons it might be good to "road test" NDE with such issues in mind.

18.4 What Do You Want to Do, What Tools Do You Need?

Confronted with the problems of NISP on the one hand, and those of MNI and MNE on the other, plus the prospect of learning how - or if - to use NDE, a prospective zooarchaeologist may be tempted to consider a new career in Medieval French literature. But before rushing out to buy a copy of *Larousse Etymologique*, the analyst should at least consider some other factors – that is, act "thoughtfully."

For any given sample, whether NISP for different taxa is likely to have differed from the outset, due to differing skeletal element counts, or to have been impacted differently by taphonomic processes, is empirically investigable. One can, for example, start simply by comparing the skeletal element counts of species one knows are in a sample. One can learn how many skeletal elements make up a rock cod skeleton versus the counts for a sea otter skeleton. The cod and the otter also diverge in the ways their remains respond to taphonomic processes. A landmark-based system may well be the best way to estimate their relative abundances in an archaeofauna sample containing both taxa, while still keeping an eye on other measures such as NISP. An analyst comparing mammals and birds also will need to consider differential bone count and taphonomic effects on counting units (Bartosiewicz and Gál 2007; Lyman 2015). In the African Neolithic archaeofaunas I have analyzed, nearly all specimens came from one zoological family, the Bovidae.

The number of skeletal elements in a bovid body is the same, regardless of the species, and though some bovids are the size of terrier dogs, while others weigh a metric ton, the responses of those elements to stresses pre- and postmortem will fall along a spectrum rather than radically diverge. In such cases, I can assess whether NISP would work as well as MNI or MNE in comparisons of element abundance, if I did not go back and score NDE for the specimens.

One must also consider intertaxonomic differences in human processing effects on all abundance statistics. For example, the archaeofaunas used by Grayson and Frey, as well as by Lyman, in their explorations of the efficacy of NISP are dominated by ungulate remains, which come in different sizes but which people handled in remarkably similar ways. I presently work with coastal California archaeofaunas that combine deer, rabbits, and seals in the mammal component. People intensively fragmented bones of deer long bones to a modal size of 2 cm, slightly damaged rabbit bones, probably as a by-product of processing these much less robust skeletons, but seldom broke anything except the crania of the seals. Seal and sea lion bones are very densely packed with bone tissue that serves as “diving ballast” for their blubber-insulated – and therefore buoyant – bodies. Their long bone marrow cavities are absent or so small as not to repay the work of opening them, especially since their blubber offers greater returns in fat. In such samples, NISP seriously overrepresents deer relative to pinnipeds, and producing MNE estimates presents disparate challenges. However, fragmentation is empirically investigable by a number of analytic procedures, as will be addressed in Chap. 20. In terms of reckoning relative element abundances, I believe a landmark-based system in fact would make explicit the approach I already use in poring through 1–2 cm deer specimens for element-distinctive features.

18.4.1 Units of Measure and Research Goals

To sum up, one’s research goals influence the statistical tools appropriate for the job. As Lyman (1994:44) put it, “We must be clear about the target population the properties of which we wish to infer, and, thus we must consider how the quantitative units we use are related to those properties.” Nearly all of Grayson’s and Lyman’s significant research on North American archaeofaunas has concerned changes in taxonomic abundances over time. They considered that NISP, for all its failings, was a suitable tool for monitoring such shifts, a decision they might wish to reconsider in light of recent experimental research (Morin et al. 2017a, 2017b). However, if a zooarchaeologist is focusing on household culinary processing and consumption patterns, one may wish to use other quantitative tools, such as MNE and fragmentation indices (Chap. 21), taking into consideration all the cautions outlined in this chapter and Chap. 10 In all cases, Lyman’s (2008:221) advice to those using any form of quantification holds: “We must be explicit about how we count, whether we count NISP, MNE, MNI, or any other measure.”

References

Bartosiewicz, L., & Gál, E. (2007). Sample size and taxonomic richness in mammalian and avian bone assemblages from archaeological sites. *Archeometriai Műhely, 1*, 37–44.

Binford, L. R. (1978). *Nunamiut ethnoarchaeology*. New York: Academic Press.

Cahen, D., & Moeyersons, J. (1977). Subsurface movements of stone artefacts and their implications for the prehistory of Central Africa. *Nature, 266*, 812–815.

Cannon, M. D. (2013). NISP, bone fragmentation, and the measurement of taxonomic abundance. *Journal of Archaeological Method and Theory, 20*(3), 397–419.

Casteel, R. W. (1977). Characterization of faunal assemblages and the minimum number of individuals determined from paired elements: Continuing problems in archaeology. *Journal of Archaeological Science, 4*(2), 125–134.

Costamagno, S. (1999). *Stratégies de chasse et fonction des sites au Magdalénien dans le sud de la France*. Talence: Université Bordeaux I.

Crader, D. C. (1984). The Zooarchaeology of the Storehouse and the Dry Well at Monticello. *American Antiquity, 49*(3), 542–558.

Crader, D. C. (1989). Faunal remains from slave quarter sites at Monticello, Charlottesville, Virginia. *Archaeozoologia, 3*(1–2), 229–236.

Domínguez-Rodrigo, M. (1997). Meat-eating by early hominids at the FLK 22 *Zinjanthropus* site, Olduvai Gorge (Tanzania): An experimental approach using cut-mark data. *Journal of Human Evolution, 33*(6), 669–690.

Ducos, P. (1968). L'origine des animaux domestiques en Palestine. *Publications de l'Institut de Préhistoire l'Université de Bordeaux, Mémoire 6*.

Faith, J. T., & Gordon, A. D. (2007). Skeletal element abundances in archaeofaunal assemblages: Economic utility, sample size, and assessment of carcass transport strategies. *Journal of Archaeological Science, 34*(6), 872–882.

Gifford, D. P., Isaac, G. L., & Nelson, C. M. (1980). Evidence for predation and pastoralism at prolonged drift, a pastoral Neolithic site in Kenya. *Azania, 15*, 57–108.

Gifford-Gonzalez, D., Boone, C. M., & Reid, R. E. (2013). The fauna from Quiroste: Insights into indigenous foodways, culture, and land modification. *California Archaeology, 5*(2), 291–317.

Grayson, D. K. (1978). Minimum numbers and sample size in vertebrate faunal analysis. *American Antiquity, 43*(1), 53–65.

Grayson, D. K. (1979). On the quantification of vertebrate archaeofaunas. In M. B. Schiffer. In *Advances in archaeological method and theory* (Vol. 2, pp. 199–237). New York: Academic Press.

Grayson, D. K. (1981). The effects of sample size on some derived measures in vertebrate faunal analysis. *Journal of Archaeological Science, 8*(1), 77–88.

Grayson, D. K. (1984). *Quantitative Zooarchaeology. Topics in the analysis of archaeological faunas*. New York: Academic Press.

Grayson, D. K., & Frey, C. J. (2004). Measuring skeletal part representation in archaeological faunas. *Journal of Taphonomy, 2*(1), 27–42.

Harris, M., Weisler, M., & Faulkner, P. (2015). A refined protocol for calculating MNI in archaeological molluscan shell assemblages: A Marshall Islands case study. *Journal of Archaeological Science, 57*, 168–179.

Holtzman, R. C. (1979). Maximum likelihood estimation of fossil assemblage composition. *Paleobiology, 5*(2), 77–89.

Klein, R. G. (1978). Stone age predation on large African bovids. *Journal of Archaeological Science, 5*(3), 195–217.

Klein, R. G. (1979). Stone age exploitation of animals in southern Africa: Middle Stone Age people living in southern Africa more than 30,000 years ago exploited local animals less effectively than the Later Stone Age people who succeeded them. *American Scientist, 67*(2), 151–160.

Klein, R. G. (1981). Stone age predation on small African bovids. *South African Archaeological Bulletin, 36*(134), 55–65.

Klein, R. G., & Cruz-Uribe, K. (1991). The bovids from Elandsfontein, South Africa, and their implications for the age, paleoenvironment, and origins of the site. *The African Archaeological Review, 9*, 21–79.

Lyman, R. L. (1979). Available meat from faunal remains: A consideration of techniques. *American Antiquity, 44*(3), 536–546.

Lyman, R. L. (1994). Quantitative units and terminology in zooarchaeology. *American Antiquity, 59*(1), 36–71.

Lyman, R. L. (2008). *Quantitative Paleozoology*. Cambridge: Cambridge University Press.

Lyman, R. L. (2015). On the variable relationship between NISP and NTAXA in bird remains and in mammal remains. *Journal of Archaeological Science, 53*, 291–296.

Marean, C. W., Abe, Y., Nilssen, P. J., & Stone, E. C. (2001). Estimating the minimum number of skeletal elements (MNE) in zooarchaeology: A review and a new image-analysis GIS approach. *American Antiquity, 66*(2), 333–348.

Marean, C. W., & Kim, S. Y. (1998). Mousterian large-mammal remains from Kobeh Cave behavioral implications for Neanderthals and early modern humans. *Current Anthropology, 38*(S1), S79–S113.

Marean, C. W., & Spencer, L. M. (1991). Impact of carnivore ravaging on zooarchaeological measures of element abundance. *American Antiquity, 56*(4), 645–658.

Marshall, F. B. (1990). Cattle herds and caprine flocks. In P. T. Robertshaw (Ed.), *Early pastoralists of south-western Kenya* (pp. 205–260). Nairobi: British Institute in Eastern Africa.

Marshall, F. B., & Pilgram, T. (1993). NISP vs. MNI in quantification of body-part representation. *American Antiquity, 58*(2), 261–269.

Mason, R. D., Peterson, M. L., & Tiffany, J. A. (1998). Weighing vs. counting: Measurement reliability and the California school of midden analysis. *American Antiquity, 63*(2), 303–324.

Morin, E., Ready, E., Boileau, A., Beauval, C., & Coumont, M.-P. (2017a). Problems of identification and quantification in archaeozoological analysis, part I: Insights from a blind test. *Journal of Archaeological Method and Theory, 24*, 886–937. https://doi.org/10.1007/s10816-016-9300-4.

Morin, E., Ready, E., Boileau, A., Beauval, C., & Coumont, M.-P. (2017b). Problems of identification and quantification in archaeozoological analysis, part II: Presentation of an alternative counting method. *Journal of Archaeological Method and Theory, 23*, 938–973. https://doi.org/10.1007/s10816-016-9301-3.

Motulsky, H. J. (1995–2015). Computing the SD, GraphPad statistics guide. http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_computing_the_sd.htm Accessed 2017.

Pickering, T. R., Egeland, C. P., Domínguez-Rodrigo, M., Brain, C. K., & Schnell, A. G. (2008). Testing the "shift in the balance of power" hypothesis at Swartkrans, South Africa: Hominid cave use and subsistence behavior in the early Pleistocene. *Journal of Anthropological Archaeology, 27*(1), 30–45.

Pickering, T. R., Egeland, C. P., Schnell, A. G., Osborne, D. L., & Enk, J. (2006). Success in identification of experimentally fragmented limb bone shafts: Implications for estimates of skeletal element abundance in archaeofaunas. *Journal of Taphonomy, 4*(2), 97–108.

Pilgram, T., & Marshall, F. B. (1995). Bone counts and statisticians: A reply to Ringrose. *Journal of Archaeological Science, 22*(1), 93–97.

Reitz, E. J., & Wing, E. S. (2008). *Zooarchaeology* (2nd ed.). Cambridge: Cambridge University Press.

Ringrose, T. J. (1993). Bone counts and statistics: A critique. *Journal of Archaeological Science, 20*(2), 121–157.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). New York: McGraw Hill.

StatSoft (2013). How to analyze data with low quality or small samples, nonparametric statistics. http://www.statsoft.com/Textbook/Nonparametric-Statistics Accessed 2013, 2017.

Thomas, D. H. (1986). *Refiguring anthropology: First principles of probability and statistics*. Prospect Heights: Waveland Press.

Todd, L. C., & Stanford, D. (1987). Application of conjoined bone data to site structural studies. In J. L. Hofman & J. G. Enloe (Eds.), *Piecing together the past: Applications of refitting studies in archaeology*, *British Archaeological Reports, International Series* (Vol. 578). Oxford: Tempus Reparatum.

Villa, P. (1982). Conjoinable pieces and site formation processes. *American Antiquity, 47*(2), 276–290.

Villa, P., & Courtin, J. (1983). The interpretation of stratified sites: A view from underground. *Journal of Archaeological Science, 10*(3), 267–281.