BEYOND CHRONOLOGY, USING BAYESIAN INFERENCE TO EVALUATE HYPOTHESES IN ARCHAEOLOGY

ERIK OTÁROLA-CASTILLO[1]*, MELISSA G. TORQUATO[1], JESSE WOLFHAGEN[1], MATTHEW E. HILL, JR.[2], AND CAITLIN E. BUCK[3]

[1]Department of Anthropology, Purdue University, West Lafayette, Indiana, USA

[2]Department of Anthropology, University of Iowa, Iowa City, Iowa, USA

[3]School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

Preprint of Manuscript accepted for publication by Advances in Archaeological Practice.

*Corresponding author email: eoc@purdue.edu

R Markdown version last compiled on Monday April 18 2022, 1:49:03 PM, EDT

INTRODUCTION

Many archaeologists are familiar with Bayesian statistics in the context of radiocarbon date calibration and chronology building. However, the Bayesian framework has broader applications beyond dating and chronology that are worthy of consideration by archaeologists. For example, many researchers in the natural and social sciences are using Bayesian statistics to evaluate how well observational or experimental data align with their hypotheses. For the most part, this use of Bayesian inference has not been applied to archaeology. Using a fictional zooarchaeological example, this paper provides a straightforward explanation of Bayesian inference and compares it to the more conventional null hypothesis significance testing (NHST). Although some have previously described and reviewed the application of these concepts elsewhere (Buck, Cavanagh, and Litton 1996; Buck 2001; Buck and Meson 2015; Otárola-Castillo and Torquato 2018; Otárola-Castillo, Torquato, and Buck 2022; Wolfhagen 2019, 2020), this work is focused on presenting replicable step-by-step examples of the Bayesian framework for evaluating and discerning among competing hypotheses.

Uncertainty and probability in archaeological applications

All data are uncertain. Measurements and observations are not exact, and their resulting values are variably imprecise. Archaeologists routinely use statistical quantities such as variance, standard deviation, and standard error, which rely on probability theory to describe this uncertainty. In their field and lab work, archaeologists regularly use equipment that relies on probabilistic descriptions of uncertainty. For example, the manufacturer of total stations, widely used to map archaeological sites, has stated accuracies of 2 mm plus an additional 2 mm per km, usually at the 1 sigma standard deviation level (e.g., Leica TS16). This is an example of a probability concept used to measure "random" uncertainty. In this case, assuming a "normal" probability distribution for the measurement error (although the manufacturer does not specify this), archaeologists should expect that 68% of the location of artifacts mapped by this instrument will have an error up to ± 2 mm, plus error related to increasing distance (and error due to atmospheric conditions, instrument stability, etc. (Walker and Awange 2020)). Similarly, the manufacturer's specification sheet for a typical Ohaus (Scout STX2202) portable digital scale claims to measure up to 2,200 grams, with an error of ± 0.02 g (1 sigma). Like the total stations, if we assume a normal error model, this means that the manufacturer certifies that 68% of all readings are within ±0.02 grams of the true reading under ideal circumstances. Similarly, after careful data collection and analyses, archaeologists also apply the concept of probability to test their hypotheses. These are formal statements that offer plausible explanations of the observed patterns of people or their environment in the past. Like the statements about field and laboratory instrument measurements, these hypotheses and their predictions also possess some degree of uncertainty

due to incomplete observation or knowledge. To formally quantify uncertainty about data and hypotheses, archaeologists frequently rely on specific probability models or probability functions (i.e., equations). The inputs of a probability function are observed or hypothesized values, and the outcomes are their probabilities ranging from zero to one, i.e., from least to most plausible. Archaeologists use this probabilistic system to test their hypotheses and describe the degree of uncertainty with which their hypotheses account for their current and likely future observations. Using a probabilistic approach gives archaeologists a powerful and systematic tool that makes it possible to interpret data and evaluate hypotheses. Below, we provide an overview of the central concepts of the two major probability paradigms to evaluate hypotheses: NHST and Bayesian inference. Whereas most scientists widely use NHST, the Bayesian approach is considered a modern data-driven learning system that has enjoyed increasing application in archaeology Buck and Meson (2015)

Null hypothesis significance testing

As the prevailing statistical framework in most sciences, NHST enables practitioners to use their data to evaluate hypotheses. This approach is rooted in the early 20th-century development of goodness of fit tests (R. A. Fisher 1922; Pearson 1900), experimental design, p-values (R. A. Fisher 1925, 1935), confidence intervals (CIs) and hypothesis testing (Neyman and Pearson 1933, 294). This methodology was introduced to archaeology in the mid-20th century (Binford 1964; Clarke 1968; Myers 1950; Spaulding 1953; Vescelius 1960). Applications of NHST in archaeology continue today, supported by new archaeology-specific statistical textbooks (Banning 2020; Baxter 2003; Carlson 2017; Drennan 2010; Fletcher and Lock 2005; McCall 2018; Shennan 1997).These textbooks provide detailed treatment of NHST and its procedures in the context of archaeology (for a multidisciplinary introductory textbook to NHST see for example (Diez, Cetinkaya-Rundel, and Barr 2019)).

In general, however, the NHST paradigm revolves around the concept of theoretically repeated sampling over the long term and the Central Limit Theorem (CLT; Diez, Cetinkaya-Rundel, and Barr (2019, 172)). The CLT informs NHST's approach to hypothesis description and evaluation. The theorem shows that given a large enough sample, in many cases, summary statistics (e.g., mean or standard deviation) will follow a normal distribution. For instance, after sampling the same population multiple times, the means of individual samples will be normally distributed. This phenomenon occurs often, even if the original variable was not normally distributed, this concept applies to many situations and data. The CLT further links sample statistics to their null distributions, such as the mean, through its standard error. According to the CLT, the standard error of a sample's mean estimates the standard deviation of the mean's null distribution. One may compute this quantity by dividing the sample's standard deviation by the sample size.

The CLT is helpful to archaeologists who often sample from a target population—a group of individuals, artifacts, events, measurements, or other phenomena that they wish to study. The aim is to use the sample to test a priori hypotheses about quantifiable characteristics of the sampled population. Statisticians refer to these characteristics as the population parameters. For example, a population's mean and standard deviation parameters represent its central tendency and variability, respectively. Sample statistics function as estimates of the population parameters and are thus also known as the parameter estimates. These statistics are used to test hypotheses about their respective population parameters. NHST requires archaeologists to state only two hypotheses: a null and an alternative hypothesis to evaluate. Null hypotheses are quantitative statements of "no difference" (difference = 0) between a hypothesized parameter value and its sample statistic, or between a sample statistic and its counterpart from another sample. Archaeologists often set up such null hypotheses to evaluate whether a sample statistic resulted from a population having the hypothesized parameter value (i.e., a one-sample test).Alternatively, they may wish to know if the statistics from two independent samples were drawn from the same population (i.e., a two-sample test). Alternative hypotheses are ordinarily simple statements negating the null hypothesis. Once archaeologists state the null and alternative hypotheses, they then sample the population, or "collect data," and calculate the sample statistics. We should point out that the NHST framework proceeds by assuming that the null hypothesis is true and then using the sample data, summarized by a statistic, to test that assumption. To do so, archaeologists use the sample statistic to define a test statistic (frequently the z-, t-, F-ratios, and chi-square values; e.g., Diez, Cetinkaya-Rundel, and Barr (2019);Thomas (1986); Drennan (2010, 177)) and calculate the probability that a value equal to or more extreme than the test statistic can occur under the assumption of the null hypothesis. The probability of the test statistic, or p-value, is often calculated with the help of probability distribution models, like the normal distribution. These probability models are also known as likelihood functions. The likelihood is a statistical function that describes the probability of the test statistic dependent on the hypothesized parameter values, e.g., those assumed by the null hypothesis. For instance, as we show in the fictitious example below, the normal likelihood function is used to compute the p-value of a z-ratio test statistic, assuming the null hypothesis is true. Using similar probability models, archaeologists conduct NHST and calculate quantities such as p-values and confidence intervals (CIs) to evaluate whether the test statistic rejects or fails to reject the null hypothesis. Confidence intervals (CIs) are grounded on the CLT's null distribution concept. Archaeologists often compute CIs in two contexts: 1) to conduct NHST, they calculate the CIs of a test statistic, and 2) to estimate the precision of a parameter estimate, they compute the CIs of a sample statistic. Generally, the CIs of either the test or sample statistic, are centered on their mean, represent their respective null distribution and are derived using their sample's standard error. Recall that the standard error of the either statistic is the standard deviation of its null distribution. For the sample

statistic, this distribution represents the range of plausible values within which one may find the true value of the population parameters. In the context of the test statistic, however, the CI is the range of possible within which the true difference, assumed by the null hypothesis, will be found.In other words, due to the CLT, ~68% of the test statistic's null distribution will capture the true value of the difference assumed to be zero by the null hypothesis. Likewise, in the case of a sample statistic, 68% of its null distribution will contain the true value of the population parameter. Alternatively, one may wish less uncertainty than 68% for the sample or test statistic. In this case, one may compute ranges similar to the standard error that capture the true parameter or difference values 95%-to-99% of the time-again, after theoretically repeated sampling. These ranges are the CIs, and we refer to them in terms of their percentage: e.g., as 95% or 99% CIs. In the context of NHST, archaeologists use CIs to reject or fail to reject a null hypothesis. If the value of no difference, 0, is within the test statistic's CI, then the null hypothesis fails to be rejected. However, if 0 is not within the test statistic's CI range, the null hypothesis is not supported by the data and is rejected in favor of the alternative. We offer one last note about the mechanics of CIs. It may seem tempting to interpret 95% CI as indicating that the true population parameter or difference has a 0.95 probability of being in the CI. Although somewhat confusing, however, the correct interpretation of the CI is that, based on repeated sampling over the long term, 95% of the CIs will contain the true population parameter or difference.

In addition to CIs, NHST uses p-values as an empirical signal of the plausibility of the test statistic assuming the null hypothesis is true. Archaeologists compute p-values by calculating the proportion of values in the null distribution that are equal to and more extreme than the sample's test statistic. Typically, test statistic values with a p-value less than or equal to a proportion of 0.05 (1 out 20 or 5%) are considered extreme. Archaeologists commonly judge whether to reject or fail to reject the null hypothesis using a p-value of 0.05 as a cut-off for rejection: the more extreme the data, the smaller the p-value.

The broader scientific community has become increasingly critical of NHST (Gelman 2018, 2006; Vidgen and Yasseri 2016). Statisticians have strongly pointed out the arbitrariness of the 0.05 p-value threshold for statistical significance (Cowgill 1977; Valeggia and Fernández-Duque 2022; Wasserstein and Lazar 2019). Some argue that inadequate statistical training may lead researchers to misunderstand p-values, (Hubbard 2011; McShane and Gal 2015). One consequence of not fully understanding the concept of p-values, for instance, is that some researchers confuse practical significance, or relevance, for statistical significance. In particular, it is possible for effects that are practically negligible, irrelevant or uninteresting to result in small p-values (Aarts, Winkens, and Den Akker 2012; Johnson 1999; Kramer, Veile, and Otárola-Castillo 2016; McCall 2018; Wolverton, Dombrosky, and Lyman 2016). In one case, while investigating the effects of sibling competition on the growth patterns of Maya children, Kramer et al. (2016) found that the effects of

family size on child growth were statistically significant but "of little consequence to early childhood health or fitness." Here, interpreting the 0.05 p-value cutoff would have led to incorrect conclusions. In other cases, some researchers have confused p-values for the type-I error rate, . The p-value is the probability that the test statistic may occur under the null hypothesis; is the probability of rejecting the null hypothesis when it is true (Hubbard 2011). Historically, these two statistical quantities belong to competing NHST philosophies (R. A. Fisher 1925; Neyman and Pearson 1933). Neyman and Pearson developed the concept of Type-1 error in the context of designing infinitely repeatable experiments, wherein defines the probability that an analysis will fail to find a difference between two hypotheses when there is a genuine difference. Fisher's p-value, by contrast, empirically estimates if a specific set of observations fit a specified null hypothesis. These two quantities have completely different theoretical underpinnings and relationships to actual observations. For example, is unrelated to observations, and the p-value is not influenced by the alternative hypotheses under consideration. Typical NHST practice, unfortunately, can lead researchers to directly associate the two concepts, complicating efforts to provide reasonable definitions and interpretations (Hubbard and Bayarri 2003). The misuse of p-values and statistical significance, due to either misunderstanding (Thiese, Arnold, and Walker 2015) or intention (Chuard et al. 2019; Head et al. 2015), can lead to the so-called scientific replication crisis (Ioannidis 2005), which is beginning to reach archaeological science (Bayliss and Marshall 2019; Marwick 2017; McPherron et al. 2021)

Even accounting for these nuances, the interpretation of NHST concepts such as p-values, statistical significance, hypothesis testing, and CIs is not entirely straightforward. Statements about sample statistics—standard errors and CIs—are based on hypothetical repeated sampling, which is difficult to conceive of in non-experimental situations or, as in archaeology, where true replication is hard or even impossible to achieve. In terms of evaluation, although most researchers might generally understand how to interpret a significant p-value in the context of rejecting a null hypothesis, the meaning of a non-significant p-value may cause confusion. This confusion might be exacerbated by the fact that NHST has no mechanism for "accepting" or "verifying" a null hypothesis. This critical misunderstanding of NHST may lead some to interpret a non-significant p-value as acceptance of their null hypothesis rather than "failing to reject it" (Greenland et al. 2016). However, knowledge production in the NHST paradigm is centered on rejecting null hypotheses, rather than accepting the null or alternative hypotheses. To be fair, the NHST language is confusing. For example, stating that a null hypothesis failed to be rejected is a triple negative, meaning that "the hypothesis of no difference was not not-accepted." Such convoluted language embedded in NHST obfuscates the relationship between the p-value, null and alternative hypotheses. Moreover, the role of the alternative hypothesis and its connection to the p-value are also unclear and often incorrectly interpreted (Cohen 1994; Benjamin and Berger 2019). As a result, inference using traditional NHST statistics can

be difficult, especially when a study is wishes to discern among multiple working hypotheses (Chamberlin 1965; Gelman, Hill, and Yajima 2012), for example, when two or more hypotheses fail to be rejected. In theory, such hypotheses are consistent with the data. However, ranking multiple unrejected null hypotheses is difficult, if not impossible. One way to rank them may be to use the hypotheses' p-values. After all, the p-value is a continuous metric mediating hypothesis rejection and failure-to-reject. However, statisticians discourage this procedure (Hubbard and Lindsay 2008; McShane et al. 2019) because the magnitude of the p-value does not reflect the weight of evidence of one hypothesis over another. Consequently, traditional NHST does not offer a straightforward procedure for further comparing "unrejected" null hypotheses.

Bayesian statistics

Bayesian inference offers an alternative approach with several advantages over NHST. First, Bayesian statistics enables scientists to use data to assign probabilities to their parameter estimates and hypotheses, facilitating a more straightforward comparison of competing hypotheses. Second, while NHST uses only new data to inferences, a Bayesian framework allows both new data and existing information to be combined. As we detail below, this characteristic more closely resembles scientists' decision making processes and is likely one of the key reasons scientists, including anthropologists and archaeologists, are increasingly adopting the Bayesian inference to evaluate their hypotheses.

Bayes' theorem derives its name from the Reverend Thomas (1763), an English Presbyterian minister and mathematician who researched problems in probability that involved conditional and prior probabilities (defined below). However, it was not until the late 1900s that the Bayesian approach to statistical inference was popularized in science (Bellhouse 2004).Although archaeologists notably began adopting Bayesian statistics to assess hypotheses in the 1990s (Buck, Cavanagh, and Litton 1996; Cowgill 1993), earlier applications can be found scattered throughout the archaeological literature beginning in the 1970s (Doran et al. 1975; D. C. Fisher 1987; Freeman 1976; Thomas 1986; Salmon 1982). Today, scientists, including anthropologists and archaeologists who find this approach advantageous, are increasingly applying Bayesian statistics to evaluate their hypotheses with data (Gelman et al. 2020; McElreath 2020; Naylor and Smith 1988; Otárola-Castillo and Torquato 2018).

One advantage of Bayesian inference is that it enables expert, or prior, information about hypotheses to be incorporated into a statistical analysis. As we show in our example below, the prior knowledge of an archaeologist or collection of archaeologists and other experts can be very valuable, especially in archaeology, as "we depend very much on prior information to help us in evaluating the degree of plausibility in a new problem" (Jaynes 2003, 6). Formally, including previous experience or expert information into statistical analyses to "update" one's state of knowledge is a natural learning process and improves the inferences made

by NHST (Cowgill 2001). To accomplish this, practitioners of Bayesian inference convert prior knowledge into prior probabilities and use them and their distributions as part of statistical analyses. Once analysts determine their prior probability distributions, as with NHSt, they can observe new data to test their hypothesis (or hypotheses). In this context, the liklihood of the data is combined with (or weighted by) the prior to give Bayesian posterior probability. The posterior is the probability of the hypothesis given the observed data's likelihood and prior knowledge (Buck, Cavanagh, and Litton 1996). As we discuss in more detail below, the Bayesian process is particularly helpful in situations where only small amounts of data are obtained, as is often the case in archaeology.

In simple cases, determining the posterior and its distribution is relatively straightforward. However, the calculus underlying more complex cases is impossible to solve without the application of novel simulation methods. In particular, the Markov Chain Monte Carlo (MCMC) algorithms has facilitated progress of Bayesian analyses. MCMC is a combination of Monte Carlo sampling and Markov Chains. Monte Carlo sampling is used to estimate difficult to compute quantities from the unknown distribution of an observed random variable. Markov Chains are a stochastic series of events associated with one another, where the probability of a new event is dependent only on the state of the last event. Together, these characteristics of Monte Carlo sampling and Markov Chains are essential to find the posterior probability distribution of complex problems. Today, variations on the original MCMC algorithms (Metropolis et al. 1953), such as the Metropolis-Hastings, Gibbs, and Hamiltonian , and other methods, are now in widespread use, facilitating broad application of the Bayesian paradigm (Dunson and Johndrow 2020; Gilks, Richardson, and Spiegelhalter 1995; Howson and Urbach 2006; Robert and Casella 2011).

To further contextualize the application of Bayesian statistics, we provide a fictional example that illustrates how one can use this probabilistic framework to solve an idealized archaeological research problem. To do this, we choose to use a parable[1] rather than a real case study in order to avoid the complexities of site formation processes and sampling bias. The contrived, fictional example in this parable also helps focus attention on specific aspects of Bayesian inference, which we feel are most instructive. The parable of the "Monico Culture and the Bayesian Archaeologist" demonstrates how inferences can be made using data and prior information about a hypothesis, how to evaluate the uncertainty surrounding a hypothesis, why this approach seems less ambiguous than NHST, and thus, why it is becoming increasingly popular.

[1]This example was inspired by the creative works similar to Neil Thompson's (1972) The Mysterious Fall of the Nacirema, Kent Flannery's The Early Mesoamerican Village (1976) and The Golden Marshalltown (1982), and John Shea's Uwasi Valley Tales from "Prehistoric Stone Tools of Eastern Africa: A Guide" (2020).

THE MONICO CULTURE: A SIMPLIFIED APPLICATION OF BAYESIAN STATISTICS

The Bayesian archaeologist and the Monico culture

The "Monico culture" is a fictitious group of people who might have lived between the ethnographic present and long ago across multiple environmental settings and sociocultural contexts worldwide. The imaginary archaeological record of the Monico is well known. In general, their material culture reflects patterns of foraging, farming, and pastoralist economies. Monico sociocultural dynamics are broad. They range from egalitarian practices exhibited at highly mobile camps to a greater social complexity derived from more permanent settlements. Some Monico experts argue that later Monico settlements show evidence of intensive food production, trade of exotic goods, and a highly centralized political organization administered by an increasingly hierarchical elite.

A famous Bayesian archaeologist, an authority on the Monico, has excavated a post-contact period site associated with this culture. Excavation work at the site, named Monico-1, has yielded an impressive faunal assemblage among the widely diverse material culture. The archaeofauna is composed of two species of animals: "dog" and "coyote." Individual animals of both species are represented by complete skeletons. Consequently, in this report the archaeologist uses the term "individual" to refer to complete dogs or coyotes. Likewise, when the archaeologist mentions "the number of" dogs or coyotes, they mean a count of complete individuals of the representative species. So far, the archaeologist has identified 100 such individuals and assigned them to their respective species. Based on the observations, the assemblage is composed of 71 dogs and 29 coyotes (Figure 2).

However, the archaeologist has also excavated a bone fragment that is difficult to identify. The archaeologist wishes to know the most probable species to which this fragment belongs.

The archaeologist defines "probability" as the relative frequency or proportion of times that an event occurs. On the basis of the data alone, the probability (P) of dog remains in the assemblage is:

$$P(Dog) = \frac{71}{100} = 0.71.$$

whereas the probability of coyote remains is:

$$P(Coyote) = \frac{29}{100} = 0.29.$$

Given these probabilities, it is reasonable for the archaeologist to believe that the unidentifiable bone

Figure 1: A reconstruction of the fictitious Monico archaeological culture, from the Monico-1 site (see text below).

Figure 2: The Bayesian Archaeologist and their crew excavate the Monico-1 site.

specimen is more likely to be from a dog. However, the archaeologist is skeptical. Moreover, as a Monico scholar, the archaeologist possesses ethnographic details on the Monico people's behavior, particularly on their eating taboos. Historical accounts reveal that the Monico once maintained hunting dogs in their villages to hunt coyotes. Because the Monico's traditional subsistence base was dependent mainly on coyote hunting, dogs developed special relationships with their owners. Consequently, the Monico came to treat their dogs respectfully, as they would other people.

Oral histories passed down over generations have documented that dogs were thought to be a close sibling of people. Notably, the Monico culture is known to have had taboos against killing or eating dogs. However, oral histories have also revealed that the Monico did eat dogs during times of resource scarcity. With this additional or "prior" information, the archaeologist decides to observe the skeletons more closely to check for the presence of butchery marks (i.e., cut marks) on the dog remains. The archaeologist tabulates this additional information on the recovered bones under two butchery conditions: 1) butchery marks are present, and 2) butchery marks are absent. Table 1 shows the frequencies of butchery marks on the skeletons of each species.

Table 1: Frequencies of individual animals and observed butchery marks at Monico 1. Note that while most of the butchery marks are on coyote bones, 9 of the 71 dog bones also show signs of butchery.

	Individuals of Each Species		
	Coyote	Dogs	Total_Butchery_marks
Butchery Marks			
Present	23	9	32
Absent	6	62	68
Total individuals	**29**	**71**	**100**

To convert these data into a probability table, the archaeologist standardizes (or divides) all of the values by the total number of observations (100 in this case). The inner cells (dark font, light shading) in Table 2 provide the probabilities of butchery marks and species occurring together, or jointly, and are thus known as joint probabilities.

Table 2: Joint probabilities of individual animals and observed butchery marks, which describe the probability of identifying a species and observing butchery marks on the bones of that species; for example, P(Coyote and Butchery mark present) is 0.23, or 23%.

	Species Identified		
	P.Coyote.	P.Dogs.	Marginal_Butchery_marks
Butchery Marks			
P(Present)	0.23	0.09	0.32
P(Absent)	0.06	0.62	0.68
Marginal Species	**0.29**	**0.71**	**Total = 1**

The values in the right and bottom margins of Table 2 are suitably named "marginal probabilities". These represent the presence and absence of butchery marks (on the right) and the species identified (bottom). The marginal totals are the total probabilities of each subsetted space (species or butchery mark). By definition, all probabilities lie in the range of 0 to 1, and the total sum of the marginal rows or columns (i.e., the sum over all marginal outcomes) must be 1. At this point, the archaeologist focuses on the unidentifiable bone specimen and finds several butchery marks on it. The archaeologist can use this additional information to gain an inferential advantage by accounting for, or conditioning on, the presence of butchery marks—a process called "conditioning". The archaeologist conditions the species identified on the presence or absence of butchery marks. This procedure is otherwise known as subsetting or stratifying the variable "species identified" by the "presence" or "absence" of butchery marks. Naturally, the archaeologist asks, "What is the probability that the unidentifiable bone specimen is from a dog compared to the probability that it is from a coyote, given that butchery marks are present on the bones of an individual?" The archaeologist observed 32 animals from Monico-1 with butchery marks present. Of those, butchery marks were present on 9 dogs and 23 coyotes. The archaeologist can thus calculate the probabilities of the individual belonging to one species or the other, given that butchery marks are present (statisticians use the "|" symbol below to mean "given that" and to signify that conditioning is taking place). For a dog, the probability is:

$$P(\text{Dog} \mid \text{Butchery mark present}) = \frac{9}{32} = 0.28.$$

whereas the probability that an individual with butchery marks belongs to the coyote species is:

$$P(\text{Coyote} \mid \text{Butchery mark present}) = \frac{23}{32} = 0.72.$$

Therefore, after observing butchery marks on the individual (unidentified) bone, the archaeologist can

state that the probability is 0.72 that it came from a coyote. In other words, the archaeologist is 72% certain that the bone was part of a coyote. A few days later, a local newspaper reporter became aware of an ongoing archaeological excavation at another Monico village site nearby, named Monico-2. Sources reveal to the reporter that the excavators there are also recovering faunal remains. Because the archaeologist is a well-known expert on the Monico's eating habits, the reporter contacts the archaeologist and communicates the fact that the new faunal assemblage at Monico-2 is wholly composed of remains from dog species. Even though the investigators at Monico-2 have not yet conducted a thorough faunal analysis, the reporter asks the archaeologist how likely it is that the Monico were butchering and eating dogs at the new site. By now, the archaeologist has estimated the probabilities of finding butchery marks associated with each animal species based on experience at the Monico-1 village. To make a probabilistic inference about behavior at the new site, the archaeologist conditions on the "species identified" instead of on the "presence of butchery marks." Out of the 71 dogs identified, the archaeologist observed 9 with butchery marks and 62 without. This means that, based on the evidence from Monico-1, the probability of finding evidence of butchery on dogs is:

$$P(\text{ Butchery mark present } | \text{ Dog}) = \frac{9}{71} = 0.13.$$

whereas the probability of no butchery evidence on dogs is:

$$P(\text{ Butchery mark present } | \text{ Coyote}) = \frac{62}{71} = 0.87.$$

After a moment's thought, the archaeologist tells the reporter that (based on knowledge from Monico-1) the probability of the dog bones from Monico-2 having resulted from dietary activities is relatively low at around 13%. This calculation draws on Bayes' theorem, as well as the information regarding the Monico's relationship with their dogs and the butchery practices at Monico-1.

WHAT IS BAYES' THEOREM

Bayes' theorem is the algebraic formalization of the probabilistic table work that we conducted in the previous section using a discrete event. The theorem is most useful when a conditional probability statement is known and one wishes to obtain its inverse conditional statement. For example, from the previous model, we know that P(Butchery mark present | Dog) = 0.13. If we wish to know the inverse conditional statement P(Dog | Butchery mark present), we can calculate the inverse conditional statement using:

$$P(\text{Dog} \mid \text{Butchery mark present}) = \frac{P(\text{Butchery mark present} \mid \text{Dog}) \times P(\text{Dog})}{P(\text{Butchery mark present})}$$

Tables 1 and 2 provide the necessary values to plug into this expression so that:

$$P(\text{Dog} \mid \text{Butchery mark present}) = \frac{\left(\frac{0.09}{0.71}\right) \times 0.71}{0.32} = 0.28.$$

When generalized, the algorithm applied here is known as Bayes' theorem. It is usually exemplified by considering two related events: A and B. Simply put, Bayes' theorem states that:

$$P(A|B) = \frac{P\left(B|A\right) \times P(A)}{P(B)}$$

In this case, to obtain the conditional probability of A given B, P(A|B), one needs to divide the joint probability of A and B, P(A and B), by the marginal probability of B, P(B). The product of P(B|A) and P(A) is the joint probability, P(A and B). The formula then generalizes to:

$$P(A/B) = \frac{P\left(A \text{ and } B\right)}{P(B)}$$

,

where the joint probability is divided by the marginal P(B). Statisticians call P(A|B) the posterior probability of A given B; P(B|A) the inverse conditional (or likelihood) of B given A; and P(A) the prior probability of A.

The Bayesian archaeologist continued

After a few days, the reporter acquires more information from the continued excavations at the Monico-2 village. The frequencies are described in Table 3 and joint probabilities and 4 below. The reporter is quite excited to inform the archaeologist that excavators had recovered 10 dogs, all but one of which had butchery marks on them. By contrast, the archaeologists at the Monico-2 site had recovered only one coyote that exhibited butchery marks on the remains. The researchers at Monico-2 used an appropriate NHST test statistic, the one-sided z-test for proportions (Diez, Cetinkaya-Rundel, and Barr 2019: 194-197), with continuity correction, to test whether the observed dog butchery rate (9/10) was statistically significantly greater than 50%- the default null hypothesis in this test. The Monico-2 archaeologists rejected the null hypothesis with a p-value <0.05 (z-ratio = 2.21, mean = 5, sdev = 1.58, p = 0.013). Becasue of the small sample size, they also conducted a one-sided binomial test, which yielded results in line with the z-test results (successes = 9, trials = 10, p = 0.01074). Based on these statistically significant results, the monico-2

archaeologists told the reporter that the majority of dogs were butchered at the site. Moreover, according to the reporter, the archaeologists also suggested that the evidence and results of the atatistical analysis indicated that the people at Monico-2 village included dogs as an important part of their diet. In light of this evidence, the reporter begins to question the ethnographic record on the dietary taboos of the Monico.

Table 3: Frequencies of individual animals and observed butchery marks from the Monico-2 village. Note the small total number of individuals and the particularly tiny sample of coyote individuals.

	Individuals of Each Species		
	Coyote	Dogs	Total_Butchery_marks
Butchery Marks			
Present	1	9	10
Absent	0	1	1
Total Individuals	**1**	**10**	**Total = 1**

Table 4: Joint probabilities of individual animals and observed butchery marks from the Monico-2 village. Note the larger proportion of dog bones with butchery marks when compared to the sample from Monico-1.

	Species Identified		
	P.Coyote.	P.Dogs.	Marginal_Butchery_marks
Butchery Marks			
P(Present)	0.09	0.82	0.91
P(Absent)	0.00	0.09	0.09
Marginal Species	**0.09**	**0.91**	**Total = 1**

The archaeologist explains that the posterior probabilities of dog and coyote butchery drawn from the (much larger) Monico-1 faunal assemblage have become new "prior" information on the probabilities that Monico villagers butchered dogs and coyotes. These quantities can be represented by:

$$P(\text{Butchery mark present} \mid \text{Dog})_{\text{Monico -1}} = \frac{9}{71} = 0.13,$$

and

$$P(\text{Butchery mark present} \mid \text{Coyote})_{\text{Monico -1}} = \frac{23}{29} = 0.79,$$

The archaeologist's knowledge about the degree to which the Monico-1 villagers butchered dogs and coyotes can be updated in a new iteration of Bayes' theorem which includes the data from Monico-2. To ac-

count for the archaeological context from which the calculations derive, the archaeologist adds the subscripts Monico-1 and Monico-2 to the equation terms, as follows:

$$
\begin{aligned}
P(\text{Butchery mark present} \mid \text{Dog})_{\text{Monico -2}} &= \frac{P(\text{Dog} \mid \text{Butchery})_{\text{Monico -2}} \times P(\text{Dog} \mid \text{Butchery})_{\text{Monico -1}}}{P(\text{Dog})_{\text{Monico -2}}} \\
&= \frac{\frac{0.82}{0.91} \times 0.13}{0.91} \\
&= 0.13.
\end{aligned}
$$

Adding in the dog data from Monico-2 causes the probability of dog butchery to decrease slightly (from 0.127 to 0.126, but rounded to 0.13). The same operation can be conducted using the prior from the first excavation and the new coyote data:

$$
\begin{aligned}
P(\text{Butchery mark present} \mid \text{Coyote})_{\text{Monico -2}} &= \frac{P(\text{Coyote} \mid \text{Butchery})_{\text{Monico -2}} \times P(\text{Coyote} \mid \text{Butchery})_{\text{Monico -1}}}{P(\text{Coyote})_{\text{Monico -2}}} \\
&= \frac{\frac{0.09}{0.91} \times 0.79}{0.09} \\
&= 0.87.
\end{aligned}
$$

In this case, after updating the data, the new posterior probability of coyote butchery is also higher (changing even more from the prior probability than in the case of dogs). The archaeologist explains this to the reporter. Furthermore, the archaeologist urges caution given that the data and resulting probabilities from the original site were derived from a sample of 100 individuals, whereas the current selection represents a total of only 11. Although the probability calculations are correct, it would be prudent to wait for more data, as the excavation at Monico-2 is ongoing. However, the archaeologist's Bayesian analysis suggests that, at this point, we should not expect butchery marks on any newly discovered dogs at the Monico-2 site.

LINKING BAYES' THEOREM TO DATA AND HYPOTHESES

The Monico case study provides a tangible example of the different components of a Bayesian analysis, including estimating an event's probability and the probability of one event given another (using currently available data), along with the key concepts of likelihood, prior and posterior probabilities, and how to update one's knowledge using the previous Bayesian posterior as the new prior. Although the procedure exemplified here is specific to archaeological count data, Bayes' theorem is very general and can be useful for a wide variety of data and data-generating processes. This section generalizes Bayes' theorem to a variety

of other scientific scenarios.

We stated earlier that Bayesian statistics uses the data in hand (D) to assign probabilities to statements or hypotheses (H) about a population. The statement $P(H|D)$, i.e., the probability of the hypothesis given the data, formalizes this relationship. In our example of the Monico sites, the archaeologist was trying to calculate the probability that the Monico people butchered dogs and coyotes (the hypothesis) given the number of cut marks on their bones (the data in hand). To operationalize this statement in the context of data and hypotheses, Bayes' theorem functions as follows:

$$P(H|D) = \frac{P\left(D|H\right) \times P(H)}{P(D)}$$

where $P(H|D)$ is the **posterior probability** of the hypothesis given the data; $P(D|H)$ is the probability of the data given the hypothesis (or the **likelihood**) of the observed data; $P(H)$ is the **prior probability** of the hypothesis (before the data were collected); and $P(D)$ is the probability of the data in hand (out of all possible values of the data). Alternatively, generalizing and using more modern statistical vernacular, this operation can be expressed as:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{P(\text{data})}$$

In this manner, Bayesian statistics offers an alternative statistical framework for updating and evaluating hypotheses through a mechanism that obtains *a posteriori* information about the posterior of interest based upon the data, a statistical model (expressed as a likelihood), and appropriately formulated prior information. In other words, with an explicit statement of our prior information, a clearly defined statistical model, and a desire to update our understanding, Bayes' theorem provides us with a probabilistic framework for making interpretations.

In addition to the coherent and explicit nature of the framework, there is another attractive feature of the Bayesian paradigm, namely that it allows us to learn from experience. Priors enable the explicit contextualization of previous knowledge or beliefs about the topic under investigation (Buck, Cavanagh, and Litton 1996; Cowgill 1993). Using previous knowledge should be a natural tendency for archaeologists. As Buck, et al. (1996) discusses, archaeologists apply previous knowledge often, for example, when inferring the function of newly discovered artifacts by using their association to artifacts and features that have already been discovered. Similarly, the archaeologist in our example was able to contextualize the data from the Monico-2 site based on Monico-1 observations. Few other interpretive frameworks offer a clear structure for updating beliefs in the light of new information, and yet this is such an important part of most intuitive approaches to learning about the world in which we live. Moreover, today's posterior information (based on

current data and prior information) is in a suitable form to become the prior for further work if and when more data become available.

From inferences about discrete points to data distributions

Thus far, the example has shown how Bayesian inference can be applied to hypotheses defined by statements about discrete events. In the fictitious example above, the hypotheses were represented by statements about whether the observed faunal remains were the result of butchery. The observed data assigned probabilities to each hypothesis, thus indicating the amount or degree of belief in the hypothesis. These data were discrete events from only two sites. Yet, in reality, although the population of the proportion of butchered dog bones are the outcomes of the same behavioral process (butchery), these values are likely to vary from site to site.

Consequently, many archaeologists might wish to compare their single-site data to the universe of known sites. In this case, the hypotheses to be evaluated are characterized by the values of the **parameters** of a probability model. Although we mentioned this earlier, at this point it is worth recalling that such parameters describe certain characteristics of a sample or population. For archaeologists, the most common parameters are those that measure central tendency, such as the mean or median. Bayesian inference can be conducted using other parameters, as well as the full distribution of the posterior, data and prior information. These are usually represented by probability models. Likely the most well-known such model is the normal probability model, in which the probability distribution has a symmetrical, bell-shape around a single mean value. When (sample) data and associated models of probability are involved, it is conventional to use the Roman symbol x to represent the observed (or sample) data and the Greek symbol θ *(theta)* to represent the parameter (or multiple set of parameters) of the model of the population that we are trying to learn about. Given x and a model with parameter(s) θ, we can re-couch Bayes' theorem and its three components—the *likelihood*, the *prior*, and the *posterior*—in the context of data distributions and their probability models.

The *likelihood* is a statistical function, or a mathematical expression, that associates individual data quantities with their respective probability values. Its form is determined by the specific probability model being used, but, in general terms, it is represented by $P(x|\theta)$, i.e., the probability distribution of the newly observed data conditioned on the parameter(s). Consequently, the likelihood is the probability of observing particular data values given some specific (or hypothesized) values of the unknown parameters. Therefore, this is a formal statement of the relationship between the parameters about which we want to learn and the data we collect.

The *prior* is also a function and can be represented by $P(\theta)$. It is a statement of what we know about the probability distribution of the parameter(s) before new data are collected. In simple terms, we can think of this as the probability we attach to observing specified values of the unknown parameters before we observed

the data. This is a formal statement of our knowledge prior to collecting the latest data.

The *posterior* is what we want to obtain: a combination of the information contained in the new data, the likelihood and the prior. The posterior is represented by $P(\theta|x)$. As presented in the previous section, this is the probability of the hypothesis given the data, or $P(H|D)$. It is the probability distribution of the model's parameter(s) conditioned on the data. In simple terms, we can think of this as the probability we attach to specified or hypothetical values of the unknown parameters after observing the data. In this context, we can express Bayes' theorem as:

$$P(\theta|x) = \frac{P\left(x|\theta\right) \times P(\theta)}{P(x)}$$

The Bayesian archaeologist and the uncertainty of hypotheses

As described above, the Bayesian inference about Monico-2 given to the reporter was based only on the Monico-2 data and the archaeologist's prior expert experience with Monico-1. However, if the archaeologist wants to give the reporter the best possible estimate, they could use all available evidence, including the Monico-2 data, their expert knowledge and information from other archaeological sites. To do this, the archaeologist reviews the published literature and identifies additional information on the proportion of dogs with butchery marks recovered from 38 previously excavated Monico sites. The archaeologist then seeks to investigate the variability of dog butchery behavior as evidenced by the proportion of dogs with butchery marks at each Monico site, with a view to obtaining a probabilistic prior statement about the theta parameter, θ (the proportion of dogs with butchery marks).

Table 5 illustrates the distribution of θ values across the frequency and proportions of sites. The table shows that out of the 38 sites, 20 have reported having between 0% and 5% of dogs showing evidence of butchery marks. Twelve sites have between 6% and 15% of dogs showing evidence of butchery marks, while another four sites report values for θ between 16% and 35%. Meanwhile, another two archaeological sites report that θ ranges from 36% to 75%. There are no sites with more than 75% of dog remains showing evidence of butchery.

Table 5: Estimates of the proportion of dog remains with butchery marks (θ) and the distribution of the proportion of the total number of sites with evidence of butchery marks on dog bones (prior probabilities)

The proportion of dog remains with butchery marks (θ)	Number of sites with (θ) evidence of butchery on dog bones	Proportion of total number of sites with (θ) evidence of butchery on dog bones (prior probability)
0-0.05	20	0.53

The proportion of dog remains with butchery marks (θ)	Number of sites with (θ) evidence of butchery on dog bones	Proportion of total number of sites with (θ) evidence of butchery on dog bones (prior probability)
0.05-0.15	12	0.32
0.15-0.25	3	0.08
0.25-0.35	1	0.03
0.35-0.45	1	0.03
0.45-0.55	0	0.00
0.55-0.65	0	0.00
0.65-0.75	1	0.03
0.75-0.85	0	0.00
0.85-0.95	0	0.00
0.95-1	0	0.00

To begin, the archaeologist speaks with other experts about nutrition, the archaeology of food, and Monico archaeology and ethnography. Based on their scientific knowledge, they hypothesize that, to consider dogs as having made a substantial food contribution at a Monico site, there would need to be evidence of butchery marks on at least 50% of individual dogs. "So," the archaeologist thinks, "my first hypothesis, H_1, is that the value of θ should be at least 50%, or 0.5, for any specific Monico site. What is the probability of this hypothesis being correct for Monico-2 based on the data I have and my prior knowledge?"

The Monico-2 site sample indicated that, out of 10 individual dog bones, 9 had butchery marks on them (so, $\theta = 0.9$). The archaeologist wants to use prior knowledge including the information from the literature review to understand the variability of θ at Monico village sites.

The archaeologist first records the dog butchery proportions (θ), from the 38 sites found in the literature. To summarize these data, in Table 5 (column 1), they group the θ values into equal intervals in increments of 0.10 (10%, except the first interval, which is smaller). They also record the number of sites reporting θ values in each interval (column 2). The archaeologist then calculates the prior probabilities for each θ-interval by dividing the number in each cell of column 2 of Table 5 by the total number of sites i.e., 38. In this way, the third column of Table 5 reports the proportion of sites within each θ-interval. This frequency distribution also serves as the prior distribution of θ values.

The archaeologist then plots the distribution of the proportion of dogs butchered at Monico sites (Table 5) in order to visualize the resulting prior knowledge that can be derived from this dataset (Figure 3).

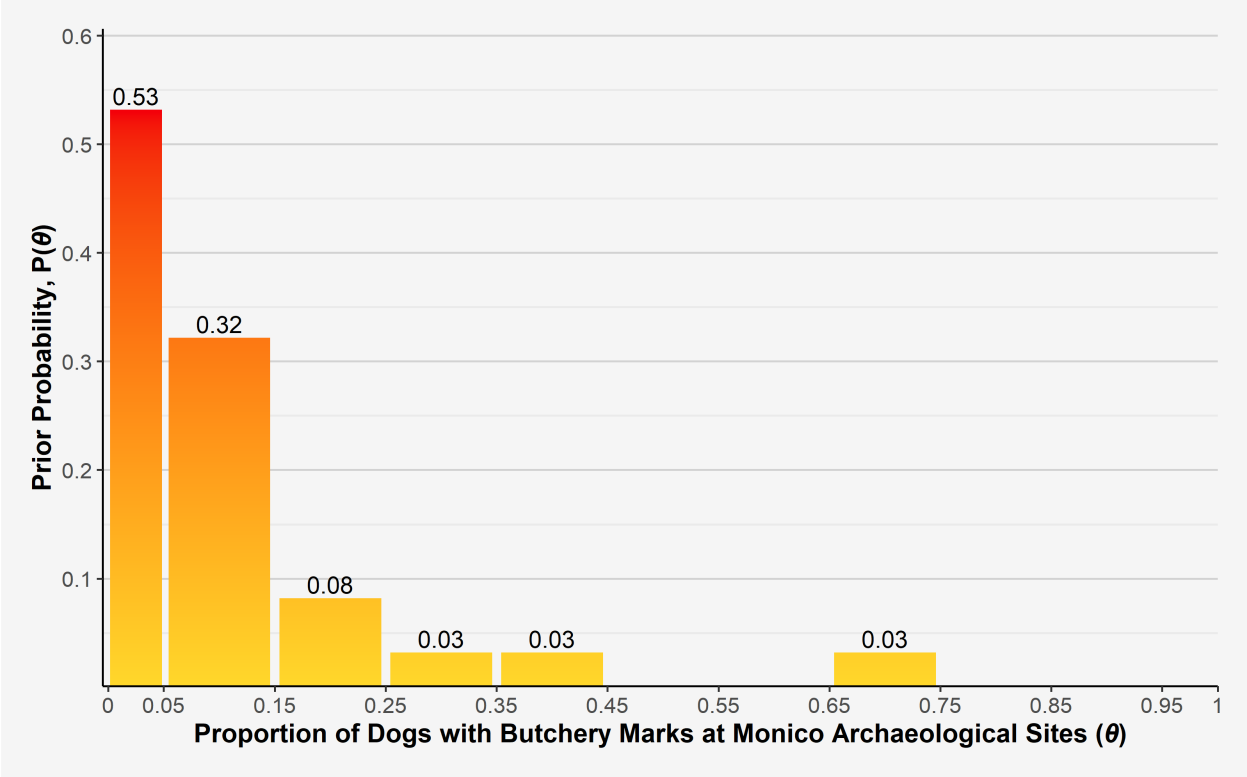

Figure 3: Simple representation of the distribution of the archaeologist's prior probabilities of the estimates of θ (theta), the proportion of dogs with butchery marks at Monico archaeological sites (from Table 5). Note that small values of *heta* have a higher prior probability than larger ones.

Recall that in the Bayesian framework, one needs the likelihood ($P(x|\theta)$), the probability of the data ($P(x)$), and the prior probability of the hypothesis ($P(\theta)$) to compute the posterior probability of the hypothesis that $\theta > 0.50$, given the data ($P(\theta > 0.5|x)$). Figure 3 illustrates the prior probability, $P(\theta)$, for different θ values.

Note that in contrast to the single-event values in the previous examples above, the components of Bayes' theorem in this case are distributions of values. Applying Bayesian statistics in such situations provides a particular advantage because the framework enables archaeologists to evaluate the probability of a hypothesis and the associated uncertainty. Thus, to continue with the Bayesian analysis of the Monico-2 data in light of the prior knowledge from the 38 sites (represented in Figure 1), the archaeologist needs a model to represent the probability of the data, x, and associated parameter(s), θ, in order to compute the likelihood, $P(x|\theta)$, and the probability of the data, $P(x)$.

The likelihood

To compute the probability of the Monico-2 data given the hypothesis, the archaeologist needs a function that can represent the likelihood, $P(x|\theta)$, of these data, x, given the parameter of interest, θ. Archaeologists frequently employ a probability function termed the "binomial" model to calculate the likelihood of data composed of binary observations, i.e., observations expressed as 1/0, yes/no, true/false, or present/absent. In this case, the binomial model is appropriate for observations indicating the presence or absence of butchery marks on individual dog skeletons, as in the Monico-2 data. As such, the archaeologist wants to compute the likelihood that 9 out of 10 dog skeletons from this site exhibited butchery marks on them.

Mathematically, the binomial model is expressed by:

$$P(x|\theta) = \binom{N}{k} \times \theta^k \times (1-\theta)^{N-k}$$

The symbols k and N represent the data: k is the number of dogs observed with butchery marks, while N is the total dogs observed. The model's parameter, θ, in this example represents the proportion of dogs with butchery marks out of all dogs observed at Monico-2.

The archaeologist uses the parameter estimate method called *maximum likelihood* (ML) to determine the most likely value of θ that would have generated the data. ML asks, under the binomial model, which value of θ is most likely to lead to the data observed? In this case, the archaeologist's binomial data are $k = 9$ dogs with butchery marks and $N = 10$ total dogs. ML evaluates which value of the θ parameter maximizes $P(x|\theta)$, the likelihood, over a systematic range of quantities between 0 and 1.

To estimate the most likely value of θ, the archaeologist assumes that the probability of observing

each butchered dog is independent of the others, making the probability of observing 9 butchered dogs θ^9. Conversely, the probability of observing a single unbutchered dog is $(1-\theta)^{(10-9)}$, and the probability of both 9 butchered dogs and 1 butchered dog occurring is $\theta^9 \times (1-\theta)^{(10-9)}$. However, to compute the likelihood of the data, the archaeologist also needs to account for the number of different ways that the 9 observations of dogs with butchery marks, k, can occur in the sequence of 10 dog observations, N.

The binomial model does this by computing $\binom{N}{k}$, known as the *binomial coefficient* (read "N *choose* k"). In this case, if positive identifications of butchery marks on dogs are represented by 1s and no butchery marks are 0s, the binomial coefficient computes how many unordered sets could have resulted in nine 1s and one 0: for example $x = \{0,1,1,1,1,1,1,1,1,1\}$, $\{1,1,1,1,0,1,1,1,1,1\}$, $\{1,1,1,1,0,1,1,1,1,1\}$, or $\{1,1,1,1,1,1,1,1,1,0\}$[2], … etc. The binomial coefficient is shorthand, and may be calculated using the following equation:

$$\binom{N}{k} = \frac{N!}{k! \times (N-k)!}$$

where ! is the factorial function that yields the product of an integer and all the integers below it. In our case, $N = 10$ and $k = 9$, and so:

$$N! = 10 \times 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 3628800$$

$$k! = 9 \times 8 \times 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 362880$$

and $(N-k)! = (10-9)! = 1! = 1$.

Therefore,

$$\binom{N}{k} = \frac{N!}{k! \times (N-k)!} = \frac{10!}{9! \times (10-9)!} = \frac{3,628,800}{362,880 \times 1} = 10.$$

Once $\binom{N}{k}$ has been computed, the archaeologist may continue to estimate the likelihood value of a given quantity of θ by calculating:

$$P(N,k|\theta) = 10 \times \theta^9 \times (1-\theta)^{(10-9)}$$

across the range of θ values from 0 to 1 to find the likelihood distribution of the data and, thus, the value of θ that maximizes the likelihood function. This approach is illustrated in Figure 4, from which

[2]Not all sets are enumerated here, but this example should enable the reader to imagine how this can occur in a total of 10 unique ways. Although in this case the solution is quite simple, in other applications, the solution might not be as obvious, e.g., the number of ways five successes can occur in 10 tries, i.e., $\binom{10}{5} = 252$.

the archaeologist learns that the ML estimate of θ (given the Monico-2 data) is 0.9; in other words, the observations at Monico-2 are most likely if the proportion of dogs butchered across Monico-2 (θ) is 0.9 (or 90%).

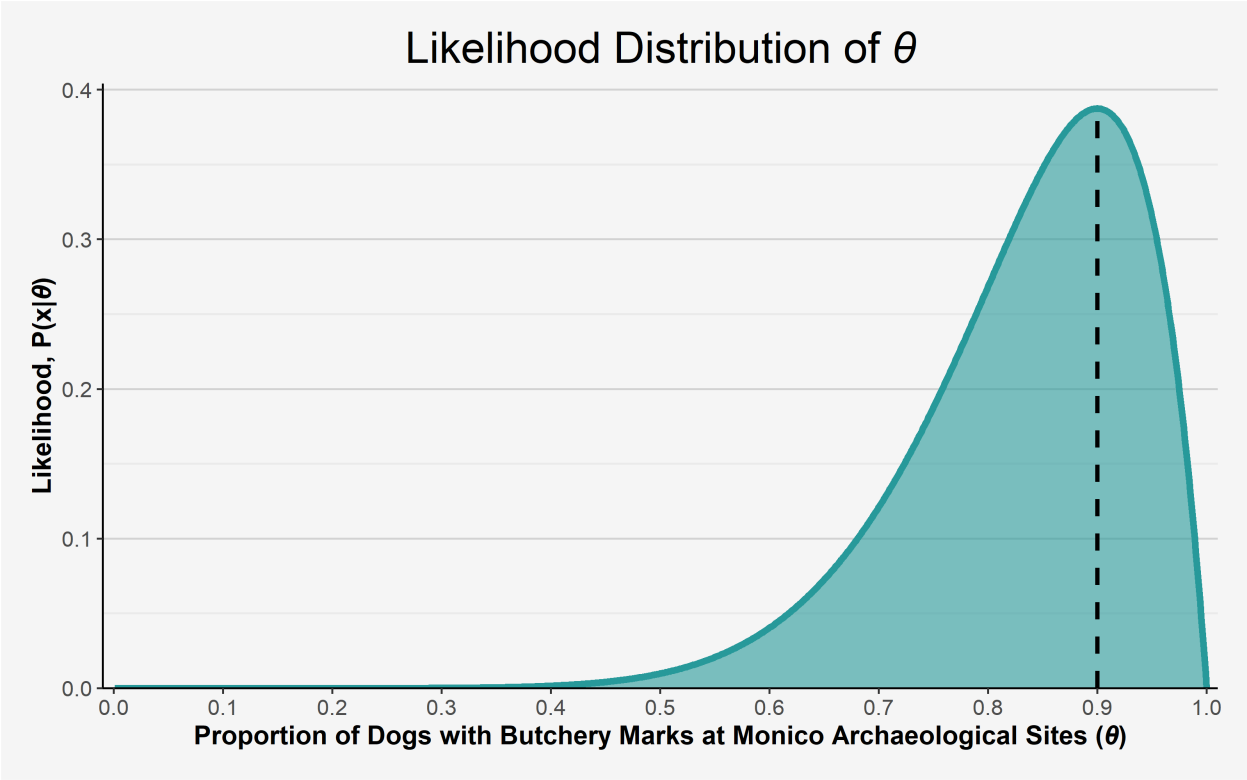

Figure 4: Distribution of standardized likelihood values corresponding to variable quantities of θ (theta) across the 0, 1 range. Dashed red line indicates the value of θ that maximizes the likelihood of the data. This is known as the Maximum Likelihood estimate of θ

[3]

The prior

Much like using the binomial probability model to obtain the likelihood distribution of the Monico-2 data, the archaeologist can also use another probability model to express $P(\theta)$, the probability distribution of θ, also known as the prior. In this case, the archaeologist needs a probability function that models the distribution of θ, the proportion of dogs with butchery marks across the 38 sites observed before the excavation of Monico-2. Statisticians frequently use the Beta probability function to model the distribution of proportions like θ. The mathematical expression of the Beta model is:

[3]It should be noted here that while the likelihood renders values in the 0-1 scale, it is not necessarily a probability function that adds up (integrates) to 1. To plot the likelihood on the same scale as the prior and the posterior distributions, all distributions have been normalized (rescaled) to sum to 1

$$P(H) = P(\theta) = \theta^{a-1} \times (1-\theta)^{b-1}$$

The shape of the Beta model is thus controlled by two parameters, a and b, which in turn control key summary statistics such as the model's mean and variance. Unlike with the likelihood model, the archaeologist in this case wants to find a distribution of θ that quantitatively describes their prior knowledge. To do this, the Beta parameters can be adjusted to fit the shape of the prior data distribution in Figure 3. Through a visual best fit, the archaeologist estimates that the values $a = 1.5$ and $b = 16$ result in a probability distribution that resembles that of the prior knowledge about θ (i.e., the shape shown in Figure 3). Thus, the distribution of the probability,

$$P(H) = P(\theta) = \theta^{(1.5-1)} \times (1-\theta)^{(16-1)}$$

across all θ values between 0 and 1 is illustrated in Figure 5.

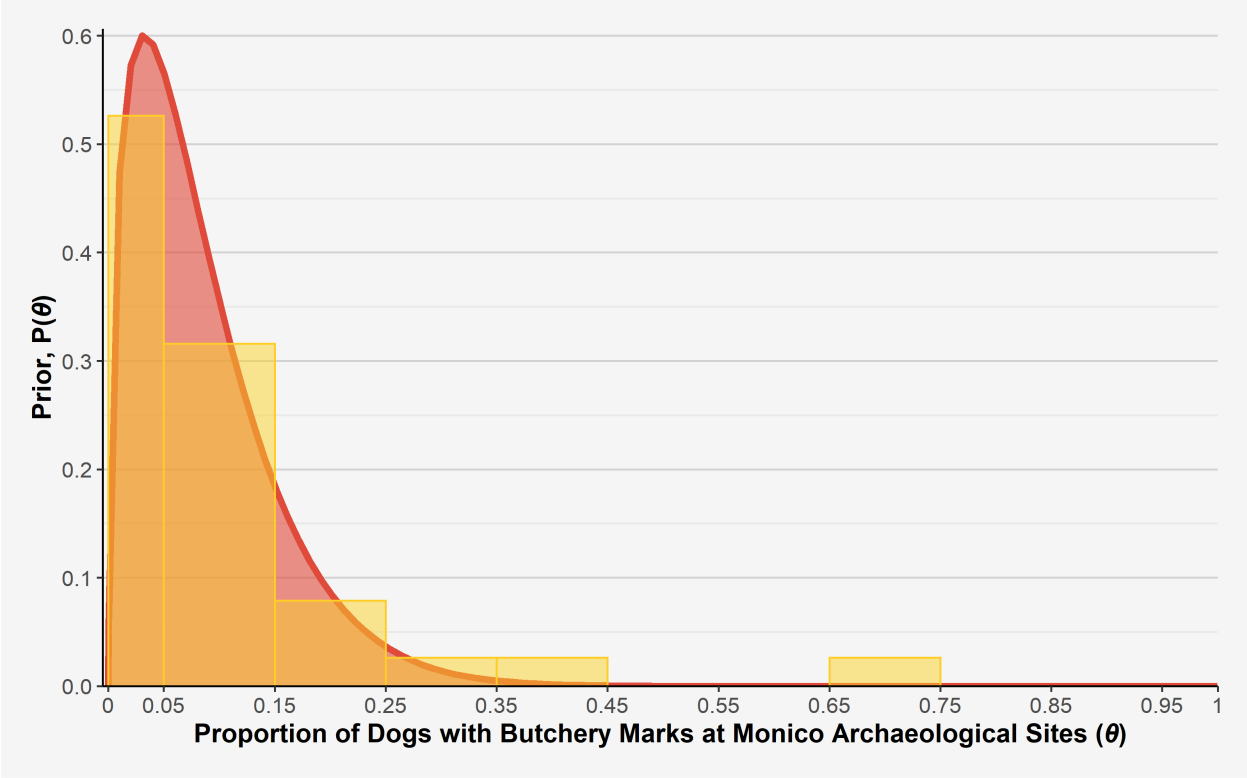

Figure 5: Standardized Beta probability model, with parameters $a = 1.5$, and $b = 16$, representing the archaeologist's prior probabilities depicted in Figure 1. Note the similarity in shpae, and in particular the location of the mode and range of values, to Figure 1.

The posterior

The archaeologist is aware that statisticians frequently use the binomial and beta distributions in the context of Bayesian analyses because they work well together for modelling the likelihood and prior probability distributions, respectively, simplifying the calculations needed to compute the posterior. Such convenient pairs of probability models are known as *conjugates*. As a result of the modelling choices made, the archaeologist may algebraically combine the binomial likelihood data with the parameters of the beta prior distribution to produce a posterior beta distribution represented by:

$$P(H|D) = P(\theta|x) = \theta^{(k_{likelihood}+a_{prior}-1)} \times (1-\theta)^{(N_{likelihood}-k_{likelihood}+b_{prior}-1)}$$

$$P(\theta|x) = \theta^{(9+1.5-1)} \times (1-\theta)^{(10-0+16-1)}$$

They thus generate values of $P(x|\theta)$, the likelihood, and $P(\theta)$, the prior probabilities, to calculate $P(\theta|x)$, the posterior probability distribution across a fine grid of θ values in the 0, 1 interval (1,000 values between 0, and 1). These are illustrated in Figure 6.

The archaeologist then focuses on $P(\theta|x)$, the posterior distribution. The posterior will help them to make inferences about the probability of θ and its surrounding uncertainty (Figure 6). The archaeologist can visually represent the estimate of θ (the expected proportion of dogs with butchery marks at Monico-2, based on the observed data and prior knowledge from the 38 other Monico archaeological sites) and the 90% uncertainty range of its estimate with a graph of the sort in Figure 7.

Unlike the NHST framework, the Bayesian posterior probability enables the archaeologist to assign probabilities to hypotheses about parameter values. In this case, the hypothesis is that the value of θ, the proportion of dogs butchered at Monico archaeological sites, is greater than 0.5 (50%, Table 6). The values shown in Table 6 are inferences resulting from calculations made using the posterior distribution. The archaeologist computed the probability that θ is greater than 0.5 (top left-most value in the table), and the values of θ at the 5th, 50th, and 95th probability percentiles. Recall, earlier, the archaeologist in conjunction with other scientists proposed that cut marks would have to appear on at least 50% (or 0.5) of the dog remains at Monico sites in order to consider dogs as "an important food contribution". However, Table 6 shows that the value of θ only has a 10% chance of being greater than 50%. Therefore, the inference that dogs were a substantial part of the Monico diet is not highly probable. For example, the archaeologist thinks, "if a meteorologist told me that there was a 10% chance of rain today, I would not carry an umbrella."

Importantly, the uncertainty around the value of θ can also be expressed as a probability interval. In

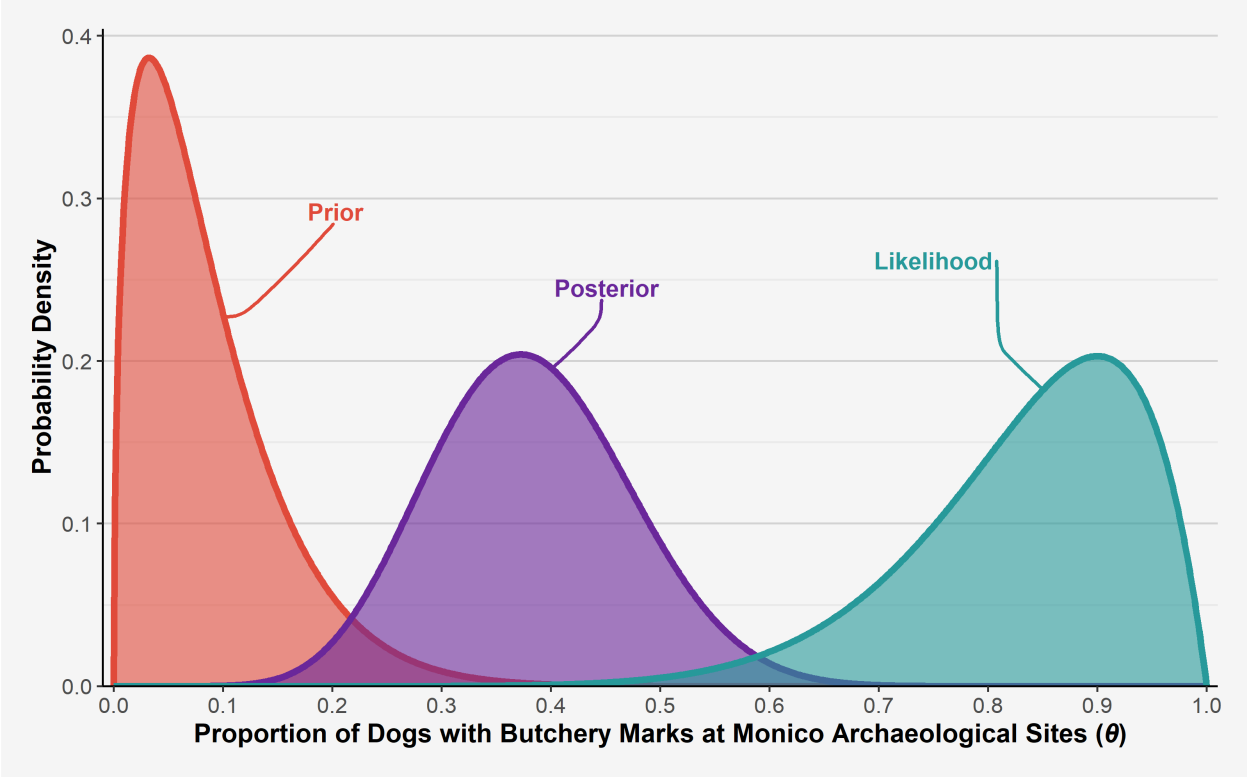

Figure 6: Distributions of the archaeologist's prior probabilities, the likelihood of the data, and the posterior probabilities across a fine grid of values of theta. All probability densities are standardized by a normalizing constant.

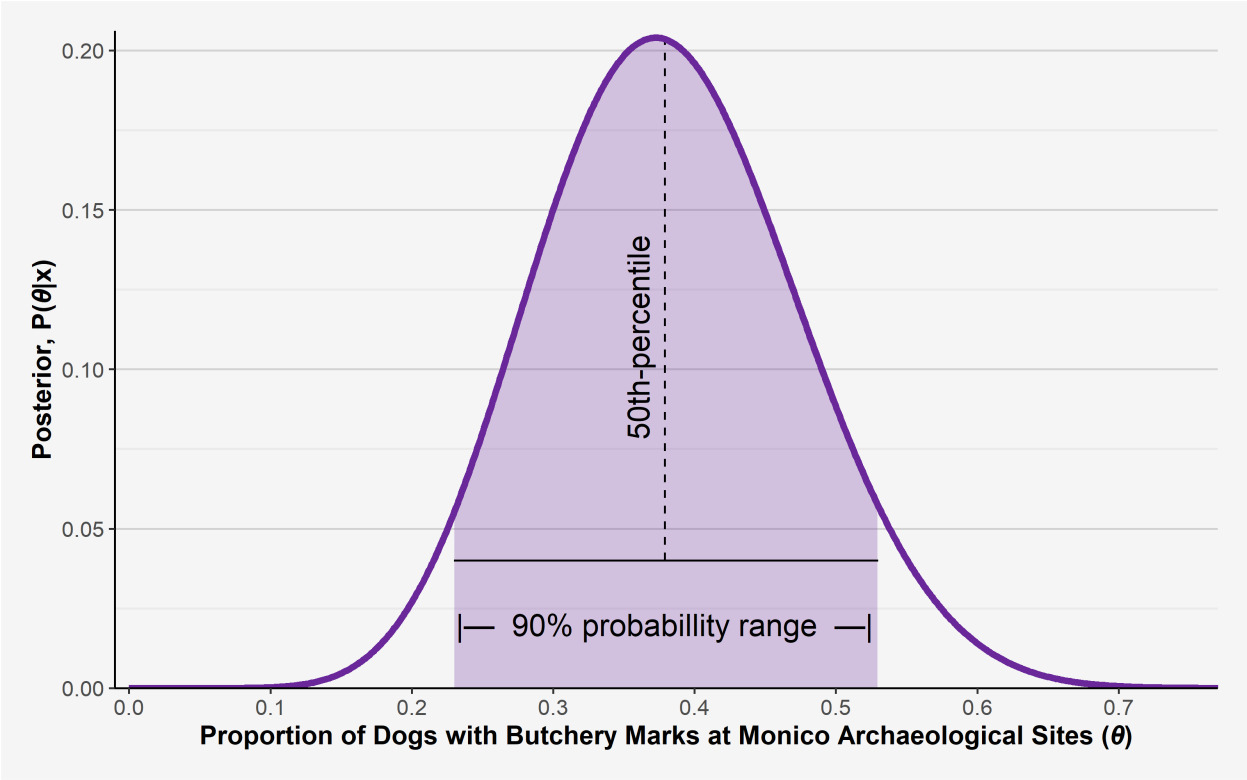

Figure 7: Posterior probability distribution with the blue dotted line showing the 50th-percentile estimate (0.38). The solid red line depicts the 90% probability density interval ($0.23 - -0.53$).

the Bayesian framework, these probability intervals are known as the *highest probability density intervals* (HPDIs) and differ from NHST's confidence intervals (CIs). One of the most important differences is that the interpretation of the HPDI is much more straightforward. The HPDI is the probability of the parameter given the data, whereas the CI is not a probability about the value of the parameter estimate. In the case of θ, Figure 7 tells the archaeologist that there is quite a lot of uncertainty around the true value of θ. For example, the median, or 50^{th} percentile, estimate of θ is 0.38, meaning that, once the available prior information from the literature and the Monico–2 data are incorporated, it is most likely that the Monico-2 occupants included dogs in their diet 38% of the time. However, the 90% HPDI spans 0.23 to 0.53 (23% to 53%), meaning that, based on our prior information and current data, there is a 90% chance that θ is between these values and only a 10% chance that it is larger or smaller than these limits. Although the variation in θ reaches over 50%, it does so only slightly and again is not very probable. These results mean that the archaeologist is very uncertain about the occupants' proclivity to butcher dogs (presumably) for dietary purposes at Monico-2, especially considering the small sample size and the fact that the current Monico-2 data differ quite markedly from those found at other sites.

Table 6: Inferences about θ from posterior probability distribution

Posterior Median	90% HPDI	$P(\theta > 0.5)$
0.38	0.23-0.53	0.1

CONCLUSIONS

Bayesian inference has advantages for archaeologists that extend well beyond the realm of radiocarbon calibration and chronological modeling. The NHST framework has served archaeologists well for many years, but has limitations. Unfortunately, NHST bases inference on new data alone due to its inherent structure. Its language and assumptions can be convoluted and confusing, and the approach cannot be used to directly compare multiple working hypotheses. Bayesian inference overcomes many of these problems for archaeologists. In many ways, archaeologists often think through problems using a Bayesian framework without knowing they are doing so and without using a formal probabilistic framework. Like the Bayesian archaeologist in our parable, most archaeologists do not form inferences about the past using new data isolated from the existing body of knowledge. Instead, we continually update our prior knowledge with new evidence to make decisions, form opinions and generate conclusions. The advantage of Bayesian inference over NHST is that it affords archaeologists 1) a more natural toolkit to learn from data, 2) straightforward language to make hypotheses quantifiable, explicit and transparent, and 3) the ability to use probability for

comparing multiple hypotheses and conducting further evaluation.

Consequently, *the Bayesian approach represents a paradigm shift in archaeological inference.* Bayesian statistics offers a coherent inferential framework that explicitly outlines the way in which one's prior information is updated with new data to produce the current state of knowledge. The process helps to evaluate the degree to which current and new evidence support hypotheses. This may be conducted iteratively until there is a desirable amount of confidence (or lack thereof) in the accuracy of a hypothesis. In this context, the Bayesian framework resembles a learning process not unlike scientific investigation. For example, archaeologists continually update their knowledge and degree of belief in hypotheses using new information gathered through multiple data collection methods, including excavation, survey, experimental, laboratory and other analytical activities.

An increasing number of archaeologists are using Bayesian statistics to calibrate radiocarbon dates, build chronologies and evaluate their hypotheses about the past. The popularity of chronology-related Bayesian software has made Bayesian inference in that context a simple operation, meaning that most users will find the software easy to operate without a basic understanding of the logic of Bayesian inference and its three fundamental components: the likelihood, the prior and the posterior. Moreover, without such fundamental understanding, the analytical power of Bayesian statistics, beyond chronology construction, may not be obvious, thus slowing rather than enhancing more general adoption.

To mitigate this problem, this paper highlights how archaeologists may use Bayesian inference to approach complex questions through a simple fictional example. This approach allows archaeologists to evaluate, compare and update their hypotheses directly, using the weight of evidence and a straightforward process. We consider this one of the most significant impacts of the Bayesian paradigm. In addition, Bayesian inference requires archaeologists to become cognizant of and transparent about prior and current information for statistical analyses within a probabilistic structure. The framework explicitly incorporates all information (prior and current) to enable a more comprehensive understanding of a problem.

As a result, applications of this method are conducive to replication, allowing them to be improved upon by other archaeological scientists. In this light, Bayesian inference dovetails with ongoing efforts to promote open science methods and open data in archaeological research. This context encourages researchers to outline the entire logical process that underlies their results. Due to its advantages, we believe that Bayesian inference is well-positioned to become a standard approach to evaluating quantitative hypotheses in archaeology.

ACKNOWLEDGMENTS

EOC thanks Deb Nichols, John Watanabe, Bob Kelly, and the Dartmouth Coach.

REFERENCES CITED

Aarts, Sil, BjöRn Winkens, and Marjan van Den Akker. 2012. "The Insignificance of Statistical Significance." *European Journal of General Practice* 18 (1): 50–52. https://doi.org/10.3109/13814788.2011.618222.

Banning, Edward B. 2020. *The Archaeologist's Laboratory: The Analysis of Archaeological Evidence*. 2nd ed. New York: Springer International Publishing. https://doi.org/10.1007/978-3-030-47992-3.

Baxter, Michael John. 2003. *Statistics in Archaeology*. Arnold.

Bayes, Thomas. 1763. "An Essay Towards Solving a Problem in the Doctrine of Chances." *Philosophical Transactions* 53: 370–418.

Bayliss, Alex, and Peter Marshall. 2019. "Confessions of a Serial Polygamist: The Reality of Radiocarbon Reproducibility in Archaeological Samples." *Radiocarbon* 61 (5): 1143–58. https://doi.org/10.1017/RDC.2019.55.

Bellhouse, David R. 2004. "The Reverend Thomas Bayes, FRS: A Biography to Celebrate the Tercentenary of His Birth." *Statistical Science* 19 (1): 3–43.

Benjamin, Daniel J, and James O Berger. 2019. "Three Recommendations for Improving the Use of p-Values." *The American Statistician* 73 (sup1): 186–91.

Binford, Lewis R. 1964. "A Consideration of Archaeological Research Design." *American Antiquity*, 425–41.

Buck, Caitlin E. 2001. *Applications of the Bayesian Statistical Paradigm*.

Buck, Caitlin E, William G Cavanagh, and Cliff D Litton. 1996. *Bayesian Approach to Interpreting Archaeological Data*. New York: Wiley.

Buck, Caitlin E, and Bo Meson. 2015. "On Being a Good Bayesian." *World Archaeology* 47 (4): 567–84. https://doi.org/10.1080/00438243.2015.1053977.

Carlson, David L. 2017. *Quantitative Methods in Archaeology Using R*. Cambridge, UK/New York: Cambridge University Press.

Chamberlin, Thomas Chrowder. 1965. "The Method of Multiple Working Hypotheses." *Science* 148 (3671): 754–59. http://www.jstor.org/stable/1716334.

Chuard, Pierre J. C., Milan Vrtílek, Megan L. Head, and Michael D. Jennions. 2019. "Evidence That Nonsignificant Results Are Sometimes Preferred: Reverse P-Hacking or Selective Reporting?" *PLOS Biology* 17 (1): e3000127. https://doi.org/10.1371/journal.pbio.3000127.

Clarke, David L. 1968. *Analytical Archaeology*. London: Methuen.

Cohen, Jacob. 1994. "The Earth Is Round ($p < .05$)." *American Psychologist* 49 (12): 997–1003. https://doi.org/10.1037/0003-066X.49.12.997.

Cowgill, George L. 1977. "Trouble with Significance Tests and What We Can Do About It." *American*

Antiquity 42 (3): 350–68. <Go to ISI>://A1977DQ60500004.

———. 1993. "Distinguished Lecture in Archeology: Beyond Criticizing New Archeology." *American Anthropologist* 95 (3): 551–73. http://www.jstor.org/stable/679650.

———. 2001. "Past, Present, and Future of Quantitative Methods in United States Archaeology." In *Computing Archaeology for Understanding the Past. CAA 2000. Computer Applications and Quantitative Methods in Archaeology*, edited by Z Stančič and T Veljanovski, 35–40. Oxford, UK: Archaeopress.

Diez, David, Mine Cetinkaya-Rundel, and Christopher D Barr. 2019. *OpenIntro Statistics*. openintro.org/os.

Doran, James Edward, Jim Doran, Frank E Hodson, and Frank Roy Hodson. 1975. *Mathematics and Computers in Archaeology*. Harvard University Press.

Drennan, Robert D. 2010. *Statistics for Archaeologists*. Springer.

Dunson, David B, and James E Johndrow. 2020. "The Hastings Algorithm at Fifty." *Biometrika* 107 (1): 1–23.

Fisher, Daniel C. 1987. "Mastodont Procurement by Paleoindians of the Great Lakes Region: Hunting or Scavenging?" In *The Evolution of Human Hunting*, 309–421. Springer.

Fisher, Ronald Aylmer. 1922. "On the Interpretation of X2 from Contingency Tables, and the Calculation of P." *Journal of the Royal Statistical Society* 85 (1): 87–94. https://doi.org/10.2307/2340521.

———. 1925. *Statistical Methods for Research Workers*. Edinburgh/London: Oliver; Boyd.

———. 1935. *The Design of Experiments*. Edinburgh: Oliver; Boyd.

Fletcher, Mike, and Gary R Lock. 2005. *Digging Numbers: Elementary Statistics for Archaeologists*. Oxford, UK: Oxford Press.

Freeman, Peter R. 1976. "A Bayesian Analysis of the Megalithic Yard." *Journal of the Royal Statistical Society: Series A (General)* 139 (1): 20–35. https://doi.org/https://doi.org/10.2307/2344382.

Gelman, Andrew. 2006. "Multilevel (Hierarchical) Modeling: What It Can and Cannot Do." *Technometrics* 48 (3): 432–35. https://doi.org/10.1198/004017005000000661.

———. 2018. "The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do about It." *Personality and Social Psychology Bulletin* 44 (1): 16–23.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2020. *Bayesian Data Analysis*. Chapman; Hall/CRC press.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness* 5 (2): 189–211. https://doi.org/10.1080/19345747.2011.618213.

Gilks, Walter R, Sylvia Richardson, and David Spiegelhalter. 1995. *Markov Chain Monte Carlo in Practice*. Chapman; Hall/CRC Press.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations." *European Journal of Epidemiology* 31 (4): 337–50. https://doi.org/10.1007/s10654-016-0149-3.

Head, Megan L., Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. 2015. "The Extent and Consequences of P-Hacking in Science." *PLOS Biology* 13 (3): e1002106. https://doi.org/10.1371/journal.pbio.1002106.

Howson, Colin, and Peter Urbach. 2006. *Scientific Reasoning: The Bayesian Approach*. Open Court Publishing.

Hubbard, Raymond. 2011. "The Widespread Misinterpretation of p-Values as Error Probabilities." *Journal of Applied Statistics* 38 (11): 2617–26. https://doi.org/10.1080/02664763.2011.567245.

Hubbard, Raymond, and M. J. Bayarri. 2003. "Confusion Over Measures of Evidence (p's) Versus Errors ('s) in Classical Statistical Testing." *The American Statistician* 57 (3): 171–78. https://doi.org/10.1198/0003130031856.

Hubbard, Raymond, and R. Murray Lindsay. 2008. "Why P Values Are Not a Useful Measure of Evidence in Statistical Significance Testing." *Theory & Psychology* 18 (1): 69–88. https://doi.org/10.1177/0959354307086923.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings Are False." *PLOS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124.

Jaynes, Edwin T. 2003. *Probability Theory: The Logic of Science*. Cambridge University Press.

Johnson, Douglas H. 1999. "The Insignificance of Statistical Significance Testing." *The Journal of Wildlife Management* 63 (3): 763–72. https://doi.org/10.2307/3802789.

Kramer, Karen L, Amanda Veile, and Erik Otárola-Castillo. 2016. "Sibling Competition & Growth Tradeoffs. Biological Vs. Statistical Significance." *PloS One* 11 (3): e0150126.

Marwick, Ben. 2017. "Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation." *Journal of Archaeological Method and Theory* 24 (2): 424–50.

McCall, Grant S. 2018. *Strategies for Quantitative Research: Archaeology by Numbers*. Routledge.

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC press.

McPherron, Shannon P., Will Archer, Erik R. Otárola-Castillo, Melissa G. Torquato, and Trevor L. Keevil. 2021. "Machine Learning, Bootstrapping, Null Models, and Why We Are Still Not 100% Sure Which Bone Surface Modifications Were Made by Crocodiles." *Journal of Human Evolution*, 103071. https://doi.org/https://doi.org/10.1016/j.jhevol.2021.103071.

McShane, Blakeley B., and David Gal. 2015. "Blinding Us to the Obvious? The Effect of Statistical Training on the Evaluation of Evidence." *Management Science* 62 (6): 1707–18. https://doi.org/10.1287/mnsc.2015.2212.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett. 2019. "Abandon Statistical Significance." *The American Statistician* 73 (sup1): 235–45. https://doi.org/10.1080/00031305.2018.1527253.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92. https://doi.org/10.1063/1.1699114.

Myers, OH. 1950. *Some Applications of Statistics to Archaeology*. Cairo: Serv. Antiq. Egypte.

Naylor, John C, and Adrian FM Smith. 1988. "An Archaeological Inference Problem." *Journal of the American Statistical Association* 83 (403): 588–95.

Neyman, Jerzy, and Egon Sharpe Pearson. 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231: 289–337.

Otárola-Castillo, Erik, and Melissa G. Torquato. 2018. "Bayesian Statistics in Archaeology." *Annual Review of Anthropology* 47 (1): 435–53. https://doi.org/10.1146/annurev-anthro-102317-045834.

Otárola-Castillo, Erik, Melissa G Torquato, and Cailtin E Buck. 2022. "The Bayesian Inferential Paradigm in Archaeology." In *Handbook of Archaeological Sciences*, edited by M. Pollard, R. A. Armitage, and C. M. Makarewicz, 2nd ed. Wiley.

Otárola-Castillo, Erik, Melissa G. Torquato, Hannah C. Hawkins, Emma James, Jacob A. Harris, Curtis W. Marean, Shannon P. McPherron, and Jessica C. Thompson. 2018. "Differentiating Between Cutting Actions on Bone Using 3d Geometric Morphometrics and Bayesian Analyses with Implications to Human Evolution." *Journal of Archaeological Science* 89: 56–67. https://doi.org/https://doi.org/10.1016/j.jas.2017.10.004.

Pearson, Karl. 1900. "X. On the Criterion That a Given System of Deviations from the Probable in the Case of a Correlated System of Variables Is Such That It Can Be Reasonably Supposed to Have Arisen from Random Sampling." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50 (302): 157–75.

Robert, Christian, and George Casella. 2011. "A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data." *Statistical Science*, 102–15.

Salmon, Merrilee H. 1982. "Philosophy and Archaeology."

Shennan, Stephen. 1997. *Quantifying Archaeology*. University of Iowa Press.

Spaulding, Albert C. 1953. "Statistical Techniques for the Discovery of Artifact Types." *American Antiquity* 18 (4): 305–13.

Thiese, Matthew S, Zachary C Arnold, and Skyler D Walker. 2015. "The Misuse and Abuse of Statistics in Biomedical Research." *Biochemia Medica: Biochemia Medica* 25 (1): 5–11.

Thomas, David Hurst. 1986. "Reguring Anthropology: First Principles of Probability and Statistics." In, 515–24. Long Grove, IL: Waveland Press, Inc. https://search.alexanderstreet.com/view/work/bibliographic_entity%7Cdocument%7C1680968.

Valeggia, Claudia R., and Eduardo Fernández-Duque. 2022. "Moving Biological Anthropology Research Beyond $p < 0.05$." *American Journal of Biological Anthropology* n/a (n/a): 1–3. https://doi.org/https://doi.org/10.1002/ajpa.24444.

Vescelius, Gary S. 1960. *Archaeological Sampling: A Problem of Statistical Inference*. Essays in the Science of Culture in Honor of Leslie White. New York: Thomas Y. Crowell Company.

Vidgen, Bertie, and Taha Yasseri. 2016. "P-Values: Misunderstood and Misused." *Frontiers in Physics* 4: 6.

Walker, John, and Joseph Awange. 2020. "Total Station: Measurements and Computations." In *Surveying for Civil and Mine Engineers: Acquire the Skills in Weeks*, edited by John Walker and Joseph Awange, 77–99. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-45803-4_4.

Wasserstein, Allen L. Schirm, Ronald L., and Nicole A. Lazar. 2019. "Moving to a World Beyond "$p < 0.05$"." *The American Statistician* 73: 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Wolfhagen, Jesse. 2019. "Rethinking Human-Cattle Interactions at Çatalhöyük (Turkey) Through Bayesian Analysis of Cattle Biometry and Behavior." Thesis.

———. 2020. "Re-Examining the Use of the LSI Technique in Zooarchaeology." *Journal of Archaeological Science* 123: 105254. https://doi.org/https://doi.org/10.1016/j.jas.2020.105254.

Wolverton, S., J. Dombrosky, and R. L. Lyman. 2016. "Practical Significance: Ordinal Scale Data and Effect Size in Zooarchaeology." *International Journal of Osteoarchaeology* 26 (2): 255–65. https://doi.org/https://doi.org/10.1002/oa.2416.