

# MÁS ALLÁ DE LA CRONOLOGÍA, USO DE LA INFERENCIA BAYESIANA PARA EVALUAR HIPÓTESIS EN ARQUEOLOGÍA

ERIK OTÁROLA-CASTILLO<sup>1\*</sup>, MELISSA G. TORQUATO<sup>1</sup>,  
JESSE WOLFHAGEN<sup>1</sup>, MATTHEW E. HILL, JR.<sup>2</sup>,  
AND CAITLIN E. BUCK<sup>3</sup>

<sup>1</sup>Department of Anthropology, Purdue University, West Lafayette, Indiana, USA

<sup>2</sup>Department of Anthropology, University of Iowa, Iowa City, Iowa, USA

<sup>3</sup>School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

Pre impresión del manuscrito aceptada para su publicación por Advances in Archaeological

\*Autor correspondiente email: eoc@purdue.edu

*Última versión de R Markdown compilada el Friday May 13 2022, 11:25:33 AM, EDT*

## RESUMEN

Los arqueólogos utilizan frecuentemente distribuciones de probabilidad y pruebas de significación de hipótesis nula (NHST por sus siglas en inglés) para evaluar qué tan bien se alinean los datos de estudios, excavaciones o experimentos con sus hipótesis sobre el pasado. La inferencia bayesiana es utilizada cada vez más como alternativa a NHST y, en arqueología, se aplica más comúnmente a la estimación de fechas de radiocarbono y la construcción de cronologías. Este artículo demuestra que las estadísticas bayesianas tienen aplicaciones más amplias. Comienza contrastando los marcos estadísticos NHST y bayesiano, antes de introducir y aplicar el Teorema de Bayes. Con la finalidad de guiar al lector a través de un análisis bayesiano elemental paso a paso, este estudio utiliza un conjunto ficticio de fauna arqueológica de un solo sitio. Posteriormente, el ejemplo ficticio se amplía para demostrar cómo pueden aplicarse los análisis bayesianos a datos con una variedad de propiedades, incorporando formalmente el conocimiento previo de los expertos en el proceso de evaluación de hipótesis.

**Keywords:** *Estadística bayesiana, modelos de probabilidad frecuencial, contraste de hipótesis, arqueostadística.*

## INTRODUCCIÓN

Muchos arqueólogos están familiarizados con las estadísticas bayesianas en el contexto de la calibración de fechas por radiocarbono y la construcción de cronologías. Sin embargo, el marco bayesiano tiene aplicaciones más amplias más allá de la datación y la cronología que merecen ser consideradas por los arqueólogos. Por ejemplo, muchos investigadores de las ciencias naturales y sociales utilizan las estadísticas bayesianas para evaluar qué tan bien se alinean los datos observacionales o experimentales con sus hipótesis. En su mayor parte, este uso de la inferencia bayesiana no se ha aplicado a la arqueología. Utilizando un ejemplo zooarqueológico ficticio, este documento proporciona una explicación directa de la inferencia bayesiana y la compara con la prueba de significación de hipótesis nula (NHST) más convencional. Aunque algunos han descrito y revisado previamente la aplicación de estos conceptos en otros lugares (por ejemplo Buck, Cavanagh, and Litten 1996; Buck 2001; Buck and Meson 2015; Otárola-Castillo and Torquato 2018; Otárola-Castillo, Torquato, and Buck 2022; Wolfhagen 2019, 2020), este trabajo está centrado en presentar ejemplos reproducibles paso a paso del marco bayesiano con la finalidad de evaluar y discernir entre hipótesis contrapuestas.

### **Incertidumbre y probabilidad en aplicaciones arqueológicas**

Todos los datos son inciertos. Las mediciones y observaciones no son exactas y sus valores resultantes son variablemente imprecisos. Los arqueólogos suelen utilizar cantidades estadísticas como la varianza, la desviación estándar y el error estándar, que están basados en la teoría de la probabilidad para describir esta incertidumbre. En su trabajo de campo y laboratorio, los arqueólogos emplean regularmente equipos basados en descripciones probabilísticas de incertidumbre. Por ejemplo, el fabricante de estaciones totales, ampliamente utilizadas para mapear sitios arqueológicos, ha establecido precisiones de 2 mm más 2 mm adicionales por km, generalmente en el nivel de desviación estándar de 1 sigma (por ejemplo, Leica TS16). Este es un ejemplo de un concepto de probabilidad utilizado para medir la incertidumbre «aleatoria». En este caso, suponiendo una distribución de probabilidad «normal» para el error de medición (aunque el fabricante no lo especifica), los arqueólogos deben esperar que el 68 % de las ubicaciones de los artefactos cartografiados por este instrumento tengan un error de hasta  $\pm$  2 mm, más el error relacionado con el incremento de la distancia (y el error debido a las condiciones atmosféricas, la estabilidad del instrumento, etc. (Walker and Awange 2020)). De forma similar, la hoja de especificaciones del fabricante para una báscula digital portátil típica de Ohaus (Scout STX2202) afirma medir hasta 2200 g, con un error de  $\pm$  0.02 g (1 sigma). Al igual que las estaciones totales, si asumimos un modelo de error normal, esto quiere decir que el fabricante certifica que el 68 % de todas las lecturas estarán dentro de  $\pm$  0.02 g de la lectura real en circunstancias ideales. De forma similar, después de una cuidadosa recopilación y análisis de datos, los arqueólogos también

aplican el concepto de probabilidad para probar sus hipótesis. Estas son declaraciones formales que ofrecen explicaciones plausibles de los patrones observados de las personas o su entorno en el pasado. Al igual que las afirmaciones sobre las mediciones de instrumentos de campo y laboratorio, estas hipótesis y sus predicciones también poseen cierto grado de incertidumbre debido a la observación o conocimiento incompletos. Para cuantificar de manera formal la incertidumbre sobre los datos y las hipótesis, los arqueólogos suelen confiar en modelos de probabilidad específicos o funciones de probabilidad (es decir, ecuaciones). Las entradas de una función de probabilidad son valores observados o hipotéticos, y los resultados son sus probabilidades que van de cero a uno, es decir, de menos probable a más probable. Son sus probabilidades que van de cero a uno, es decir, de menos probable a más probable. Los arqueólogos utilizan este sistema probabilístico para probar sus hipótesis y describir el grado de incertidumbre con el que sus hipótesis dan cuenta de las observaciones actuales y futuras probables. El uso de un planteamiento probabilístico ofrece a los arqueólogos una herramienta poderosa y sistemática que posibilita la interpretación de datos y evaluación de hipótesis. A continuación, proporcionamos una descripción general de los conceptos centrales de los dos principales paradigmas de probabilidad para evaluar hipótesis: NHST e inferencia bayesiana. Mientras que la mayoría de los científicos utilizan ampliamente NHST, el planteamiento bayesiano se considera un moderno sistema de aprendizaje basado en datos que ha gozado de una aplicación cada vez mayor en la arqueología Buck and Meson (2015)

### **Prueba de significación de la hipótesis nula**

Como marco estadístico predominante en la mayoría de las ciencias, NHST permite a los profesionales utilizar sus datos para evaluar hipótesis. Este planteamiento tiene sus raíces en el desarrollo a principios del siglo XX de las pruebas de bondad de ajuste (R. A. Fisher 1922; Pearson 1900), el diseño experimental, los valores de p (R. A. Fisher 1925, 1935), los intervalos de confianza (IC) y las pruebas de hipótesis (Neyman and Pearson 1933, 294). Esta metodología se introdujo en la arqueología a mediados del siglo XX (por ejemplo, Binford 1964; Clarke 1968; Myers 1950; Spaulding 1953; Vescelius 1960). Las aplicaciones de NHST en arqueología continúan hoy, respaldadas por nuevos libros de texto estadísticos específicos de arqueología (por ejemplo, Banning 2020; Baxter 2003; Carlson 2017; Drennan 2010; Fletcher and Lock 2005; McCall 2018; Shennan 1997). Estos libros de texto brindan un tratamiento detallado de NHST y sus procedimientos en el contexto de la arqueología (para un libro de texto de introducción multidisciplinario a NHST, consulte, por ejemplo, (Diez, Cetinkaya-Rundel, and Barr 2019)).

Sin embargo, en general, el paradigma NHST gira en torno al concepto de muestreo repetido teóricamente a largo plazo y el Teorema central del límite (TCL; Diez, Cetinkaya-Rundel, and Barr (2019, 172)). El TCL informa el planteamiento del NHST para la descripción y evaluación de hipótesis. El teorema muestra que

dada una muestra lo suficientemente grande, en muchos casos, las estadísticas de resumen (por ejemplo, media o desviación estándar) seguirán una distribución normal. Por ejemplo, después de muestrear la misma población varias veces, las medias de las muestras individuales se distribuirán normalmente. Esta distribución se conoce como muestreo o distribución «nula» de la estadística. Como este fenómeno ocurre frecuentemente, incluso si la variable original no tenía una distribución normal, este concepto se aplica a muchas situaciones y datos. El TCL vincula además las estadísticas de muestra con sus distribuciones nulas, como la media, a través de su error estándar. Según la TCL, el error estándar de la media de una muestra estima la desviación estándar de la distribución nula de la media. Esta cantidad puede calcularse dividiendo la desviación estándar de la muestra por el tamaño de la muestra.

El TCL es útil para los arqueólogos que a menudo toman muestras de una población objetivo: un grupo de individuos, artefactos, eventos, mediciones u otros fenómenos que desean estudiar. El objetivo es usar la muestra para probar hipótesis a priori sobre características cuantificables de la población muestreada. Los estadísticos se refieren a estas características como los parámetros de la población. Por ejemplo, los parámetros de media y desviación estándar de una población representan su tendencia central y variabilidad, respectivamente. Las estadísticas de muestra funcionan como estimaciones de los parámetros de la población y, por lo tanto, también se conocen como estimaciones de parámetros. Estas estadísticas se usar con la finalidad de probar hipótesis sobre sus respectivos parámetros de población. NHST requiere que los arqueólogos establezcan únicamente dos hipótesis: una hipótesis nula y una alternativa a evaluar. Las hipótesis nulas son afirmaciones cuantitativas de «ninguna diferencia» (diferencia = 0) entre un valor de parámetro hipotético y su estadística de muestra, o entre una estadística de muestra y su contraparte de otra muestra. Los arqueólogos a menudo establecen tales hipótesis nulas para evaluar si una muestra estadística resultó de una población que tiene el valor del parámetro hipotético (es decir, una prueba de una muestra). Alternativamente, pueden desear saber si las estadísticas de dos muestras independientes se extrajeron de la misma población (es decir, una prueba de dos muestras). Las hipótesis alternativas son declaraciones ordinariamente simples que niegan la hipótesis nula. Una vez que los arqueólogos establecen las hipótesis nula y alternativa, toman muestras de la población o «recolectan datos» y calculan las estadísticas de la muestra. Tenemos que señalar que el marco NHST procede asumiendo que la hipótesis nula es verdadera y posteriormente utiliza los datos de la muestra, resumidos por una estadística, para probar esa suposición. Para hacerlo, los arqueólogos utilizan la estadística de muestra para definir una estadística de prueba (con frecuencia, los valores z-, t-, ratios F- y los valores de chi-cuadrado; por ejemplo, Diez, Cetinkaya-Rundel, and Barr (2019);Thomas (1986); Drennan (2010, 177)) y calculan la probabilidad de que un valor igual o más extremo que el estadístico de prueba pueda ocurrir bajo el supuesto de la hipótesis nula. La probabilidad de la estadística de prueba, o valor p, a menudo se calcula con la ayuda de modelos de distribución

de probabilidad, como la distribución normal. Estos modelos de probabilidad son conocidos también como funciones de verosimilitud. La verosimilitud es una función estadística que describe la probabilidad de la estadística de prueba que depende de los valores de los parámetros hipotéticos, por ejemplo, los asumidos por la hipótesis nula. Por ejemplo, como mostramos en el ejemplo ficticio a continuación, la función de verosimilitud normal es utilizada para calcular el valor  $p$  de una estadística de prueba de ratio  $z$ , suponiendo que la hipótesis nula es verdadera. Utilizando modelos de probabilidad similares, los arqueólogos realizan NHST y calculan cantidades tales como valores  $p$  e intervalos de confianza (IC) para evaluar si la estadística de prueba rechaza o no rechaza la hipótesis nula. Los IC están basados en el concepto de distribución nula de TCL. Los arqueólogos a menudo calculan los IC en dos contextos: 1) para realizar NHST, calculan los IC de una estadística de prueba, y 2) para estimar la precisión de la estimación de un parámetro, calculan los IC de una estadística de muestra. En general, los IC de la estadística de prueba o de muestra se centran en su media, representan su distribución nula respectiva y se derivan usando el error estándar de su muestra. Recuerde que el error estándar de cualquier estadística es la desviación estándar de su distribución nula. Para la estadística muestral, esta distribución representa el rango de valores plausibles dentro de los cuales puede encontrarse el verdadero valor de los parámetros de la población. Sin embargo, en el contexto de la estadística de prueba, el IC es el rango de valores posibles dentro del cual se encontrará la verdadera diferencia, asumida por la hipótesis nula. Dicho de otro modo, debido al TCL, ~68 % de la distribución nula de la estadística de prueba capturará el verdadero valor de la diferencia, que la hipótesis nula supone que es cero. Asimismo, en el caso de un estadístico muestral, el 68 % de su distribución nula contendrá el verdadero valor del parámetro poblacional. Alternativamente, se puede desear una incertidumbre inferior al 68 % para la muestra o la estadística de prueba. En este caso, pueden calcularse rangos similares al error estándar que capturan el parámetro real o los valores de diferencia entre el 95 % y el 99 % de las veces, nuevamente, después de un muestreo teórico repetido. Estos rangos son los IC, y nos referimos a ellos en términos de su porcentaje: por ejemplo, como 95 % o 99 % IC. En el contexto de NHST, los arqueólogos usan los IC de la estadística de prueba para rechazar o no una hipótesis nula. Si el valor de ninguna diferencia, 0, está dentro del IC de la estadística de prueba, entonces la hipótesis nula no se rechaza. Sin embargo, si 0 no está dentro del rango de IC de la estadística de prueba, los datos no respaldan la hipótesis nula y se rechaza a favor de la alternativa. Ofrecemos una última nota sobre la mecánica de los IC. Puede parecer tentador interpretar el IC del 95 % como una indicación de que el verdadero parámetro o diferencia de la población tiene una probabilidad de 0.95 de estar en el IC. Aunque algo confuso, sin embargo, la interpretación correcta del IC es que, con base en un muestreo repetido a largo plazo, el 95 % de los IC contendrá el verdadero parámetro o diferencia de la población.

Además de los IC, NHST utiliza valores  $p$  como una señal empírica de la plausibilidad de la estadística de

prueba, asumiendo que la hipótesis nula es verdadera. Los arqueólogos calculan los valores p calculando la proporción de valores en la distribución nula igual y más extrema que el estadístico de prueba de la muestra. Por lo general, los valores estadísticos de prueba con un valor p menor o igual a una proporción de 0.05 (1 de 20 o 5 %) son considerados extremos. Es común que los arqueólogos juzguen si rechazar o no la hipótesis nula utilizando un valor p de 0.05 como límite para el rechazo: cuanto más extremos son los datos, menor es el valor p.

La comunidad científica en general se ha vuelto cada vez más crítica con el NHST (por ejemplo: Gelman 2018, 2006; Vidgen and Yasseri 2016). Los estadísticos han señalado enérgicamente la arbitrariedad del umbral del valor p de 0.05 para la significación estadística (Cowgill 1977; Valeggia and Fernández-Duque 2022; Wasserstein and Lazar 2019). Algunos argumentan que una formación estadística no adecuada puede llevar a los investigadores a malinterpretar los valores p (Hubbard 2011; McShane and Gal 2015). Una consecuencia de no comprender completamente el concepto de valores p, por ejemplo, es que algunos investigadores confunden el significado práctico, o relevancia, con el significado estadístico. En particular, es posible que los efectos que son prácticamente insignificantes, irrelevantes o poco interesantes den como resultado valores p pequeños (por ejemplo, Aarts, Winkens, and Den Akker 2012; Johnson 1999; Kramer, Veile, and Otárola-Castillo 2016; McCall 2018; Wolverton, Dombrosky, and Lyman 2016). En un caso, mientras investigaban los efectos de la competencia entre hermanos en los patrones de crecimiento de los niños mayas, Kramer et al. (2016) encontró que los efectos del tamaño de la familia en el crecimiento de los niños eran estadísticamente significativos, pero «de poca importancia para la salud o el estado físico de la primera infancia». Aquí, interpretar el límite del valor p de 0.05 como demográficamente importante habría llevado a conclusiones incorrectas. En otros casos, los investigadores han confundido los valores p con la tasa de error de tipo I, . El valor p es la probabilidad de que la estadística de prueba pueda ocurrir bajo la hipótesis nula; es la probabilidad de rechazar la hipótesis nula cuando es verdadera (Hubbard 2011). Históricamente, estas dos cantidades estadísticas pertenecen a filosofías NHST en competencia (R. A. Fisher 1925; Neyman and Pearson 1933). Neyman y Pearson desarrollaron el concepto de error tipo 1 en el contexto del diseño de experimentos infinitamente repetibles, donde define la probabilidad de que un análisis no encuentre una diferencia entre dos hipótesis cuando hay una diferencia genuina. Contrariamente, el valor p de Fisher, estima empíricamente si un conjunto específico de observaciones se ajusta a una hipótesis nula específica. Estas dos cantidades tienen fundamentos teóricos y relaciones completamente distintas con las observaciones reales. Por ejemplo, no está relacionado con las observaciones y el valor p no está influenciado por las hipótesis alternativas bajo consideración. Desafortunadamente, la práctica típica de NHST puede llevar a los investigadores a asociar directamente los dos conceptos, complicando los esfuerzos para proporcionar definiciones e interpretaciones razonables (Hubbard and Bayarri 2003). El mal uso de los valores de p y la significación estadística, debido

a un malentendido (por ejemplo, Thiese, Arnold, and Walker 2015) o intención (Chuard et al. 2019; Head et al. 2015), puede dar lugar a la llamada crisis de replicación científica (Ioannidis 2005), que empieza a alcanzar a la ciencia arqueológica (Bayliss and Marshall 2019; Marwick 2017; McPherron et al. 2021)

Incluso teniendo en cuenta estos matices, la interpretación de los conceptos del NHST, como los valores p, la significación estadística, las pruebas de hipótesis y los IC, no es del todo sencilla. Las declaraciones sobre las estadísticas de la muestra (errores estándar e IC) están basados en un muestreo repetido hipotético, que es difícil de concebir en situaciones no experimentales o, como en arqueología, donde la replicación real es difícil o incluso imposible de lograr. En términos de evaluación, aunque la mayoría de los investigadores en general pueden entender cómo interpretar un valor p significativo en el contexto de rechazar una hipótesis nula, el significado de un valor p no significativo puede causar confusión. Esta confusión podría verse exacerbada por el hecho de que no existe un mecanismo para «aceptar» o «verificar» una hipótesis nula. Este malentendido crítico del NHST puede llevar a algunos a interpretar un valor p no significativo como una aceptación de su hipótesis nula en vez de rechazarla (Greenland et al. 2016). Sin embargo, la producción de conocimiento en el paradigma NHST se centra en rechazar las hipótesis nulas, en vez de aceptar las hipótesis nulas o alternativas. Para ser justos, el lenguaje del NHST es confuso. Por ejemplo, afirmar que una hipótesis nula no pudo ser rechazada es un triple negativo, lo que significa que «la hipótesis de no diferencia no fue no aceptada». Tal lenguaje intrincado incluido en NHST ofusca la relación entre las hipótesis de valor p, nula y alternativa. Además, el papel de la hipótesis alternativa y su conexión con el valor p tampoco están claros y, a menudo, se interpretan incorrectamente (Cohen 1994; Benjamin and Berger 2019). Como resultado, la inferencia utilizando estadísticas NHST tradicionales puede ser difícil, especialmente cuando un estudio desea discernir entre múltiples hipótesis de trabajo (por ejemplo, Chamberlin 1965; Gelman, Hill, and Yajima 2012), por ejemplo, cuando dos o más hipótesis no logran ser rechazadas. En teoría, tales hipótesis son consistentes con los datos. Sin embargo, clasificar múltiples hipótesis nulas no rechazadas es difícil, si no imposible. Una forma de clasificarlos puede ser utilizar los valores p de las hipótesis. Después de todo, el valor p es una métrica continua que media el rechazo de hipótesis y la falta de rechazo. Sin embargo, los estadísticos desaconsejan este procedimiento (Hubbard and Lindsay 2008; McShane et al. 2019) porque la magnitud del valor p no refleja el peso de la evidencia de una hipótesis sobre otra. En consecuencia, el NHST tradicional no ofrece un procedimiento sencillo para comparar más hipótesis nulas «no rechazadas».

### **Estadísticas bayesianas**

La inferencia bayesiana ofrece un planteamiento alternativo con varias ventajas sobre NHST. En primer lugar, las estadísticas bayesianas permiten a los científicos utilizar datos para asignar probabilidades a las estimaciones de sus parámetros e hipótesis, lo que facilita una comparación más directa de las hipótesis con-

trapuestas. En segundo lugar, mientras que el NHST utiliza solamente datos nuevos para hacer inferencias, un marco bayesiano permite combinar tanto datos nuevos como información existente. Como detallamos a continuación, esta característica se parece más a los procesos de toma de decisiones de los científicos y es probablemente una de las razones clave por las que los científicos, incluidos los antropólogos y arqueólogos, están adoptando cada vez más la inferencia bayesiana para evaluar sus hipótesis.

El teorema de Bayes deriva su nombre del reverendo Thomas Bayes (1763), un ministro presbiteriano inglés y matemático que investigó problemas de probabilidad que involucraban probabilidades condicionales y previas (definidas a continuación). Sin embargo, no fue hasta finales de 1900 que el planteamiento bayesiano de la inferencia estadística se popularizó en la ciencia (Bellhouse 2004). Aunque los arqueólogos comenzaron notablemente a adoptar estadísticas bayesianas para evaluar hipótesis en la década de 1990 (por ejemplo, Buck, Cavanagh, and Litton 1996; Cowgill 1993), pueden encontrarse aplicaciones anteriores dispersas en la literatura arqueológica a partir de la década de 1970 (Doran et al. 1975; D. C. Fisher 1987; Freeman 1976; Thomas 1986; Salmon 1982). Hoy en día, los científicos, incluidos los antropólogos y arqueólogos que encuentran ventajoso este planteamiento, están aplicando cada vez más las estadísticas bayesianas para evaluar sus hipótesis con datos (Gelman et al. 2020; McElreath 2020; Naylor and Smith 1988; Otárola-Castillo and Torquato 2018).

Una ventaja de la inferencia bayesiana es que permite incorporar información previa o experta sobre hipótesis en los análisis estadísticos. Como mostramos en nuestro ejemplo a continuación, el conocimiento previo de un arqueólogo o un conjunto de arqueólogos y otros expertos puede ser muy valioso ya que «dependemos mucho de la información previa para ayudarnos a evaluar el grado de plausibilidad en un nuevo problema» (Jaynes 2003, 6). Incluir formalmente experiencia previa o información de expertos en análisis estadísticos para «actualizar» el estado de conocimiento de uno es un proceso de aprendizaje natural y mejora las inferencias hechas por NHST (Cowgill 2001). Para lograr esto, los practicantes de la inferencia bayesiana convierten el conocimiento previo en probabilidades previas y las utilizan junto con sus distribuciones como parte de los análisis estadísticos. Una vez que los analistas determinan sus distribuciones de probabilidad previas, como con el NHST, pueden observar nuevos datos para probar su hipótesis (o hipótesis). En este contexto, la verosimilitud de los datos se combina con (o se pondera por) la previa para dar la probabilidad posterior bayesiana. La posterior es la probabilidad de la hipótesis dada la verosimilitud de los datos observados y el conocimiento previo (Buck, Cavanagh, and Litton 1996). Como comentamos más detalladamente a continuación, el proceso bayesiano es particularmente útil en situaciones en las que solamente se obtienen pequeñas cantidades de datos, como suele ser el caso en arqueología.

En casos simples, determinar el posterior y su distribución es relativamente sencillo. Sin embargo, el cálculo subyacente a casos más complejos es imposible de resolver sin la aplicación de nuevos métodos de

simulación. En particular, los algoritmos Markov Chain Monte Carlo (MCMC) han facilitado el progreso en los análisis bayesianos. La simulación MCMC es una combinación de muestreo de Monte Carlo y cadenas de Markov. El muestreo de Monte Carlo se utiliza para estimar cantidades difíciles de calcular a partir de la distribución desconocida de una variable aleatoria observada. Las cadenas de Markov son una serie estocástica de eventos asociados entre sí, donde la probabilidad de un nuevo evento depende únicamente del estado del último evento. Juntas, estas características del muestreo de Monte Carlo y las cadenas de Markov son esenciales para encontrar la distribución de probabilidad posterior de problemas complejos. Hoy en día, las variaciones del algoritmo MCMC original (Metropolis et al. 1953), como Metropolis-Hastings, Gibbs, hamiltoniano y otros métodos, ahora se utilizan ampliamente, lo que facilita una amplia aplicación del paradigma bayesiano (por ejemplo, Dunson and Johndrow 2020; Gilks, Richardson, and Spiegelhalter 1995; Howson and Urbach 2006; Robert and Casella 2011).

Para contextualizar todavía más la aplicación de las estadísticas bayesianas, proporcionamos un ejemplo ficticio que ilustra cómo puede utilizarse este marco probabilístico para resolver un problema de investigación arqueológica idealizado. Para hacer esto, elegimos usar una parábola<sup>1</sup> en vez de un estudio de caso real para evitar las complejidades de los procesos de formación de sitios y el sesgo de muestreo. El ejemplo inventado y ficticio de esta parábola también ayuda a centrar la atención en aspectos específicos de la inferencia bayesiana, que creemos que son muy instructivos. La parábola de la «Cultura de Monico y el arqueólogo bayesiano» demuestra cómo se pueden hacer inferencias utilizando datos e información previa sobre una hipótesis, cómo evaluar la incertidumbre que rodea a una hipótesis, por qué este plantenamiento parece menos ambiguo que el NHST y, por lo tanto, por qué se está volviendo cada vez más popular.

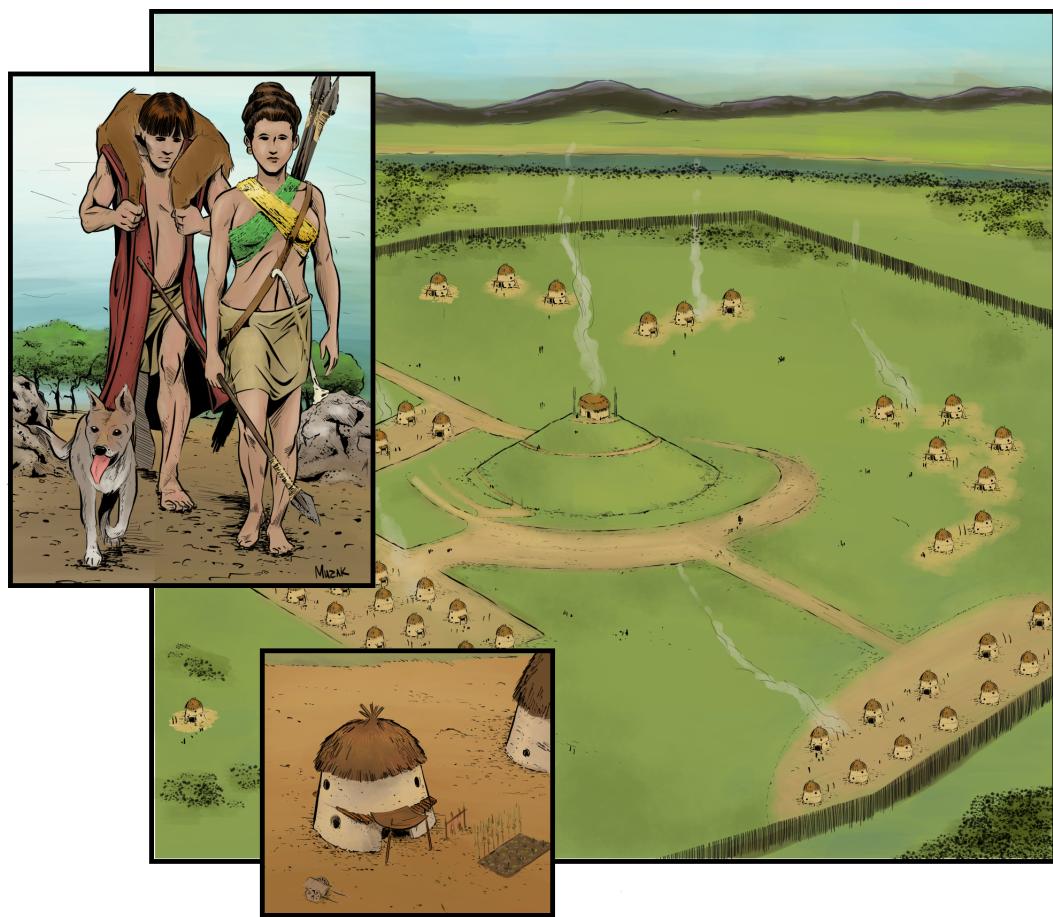
## LA CULTURA MONICO: UNA APLICACIÓN SIMPLIFICADA DE LAS ESTADÍSTICAS BAYESIANAS

### El Arqueólogo Bayesiano y la Cultura Monico

La «cultura Monico» es un grupo ficticio de personas que podrían haber vivido entre el presente etnográfico y hace mucho tiempo en múltiples entornos ambientales y contextos socioculturales en todo el mundo. El registro arqueológico imaginario del Monico es bien conocido. En general, su cultura material refleja patrones de economías de forrajeo, agricultura y pastoreo. Las dinámicas socioculturales de Monico son amplias. Van desde prácticas igualitarias exhibidas en campamentos altamente móviles hasta una mayor complejidad social derivada de asentamientos más permanentes. Algunos expertos de Monico argumentan

<sup>1</sup>Este ejemplo se inspiró en obras creativas similares a Neil Thompson's (1972) *The Mysterious Fall of the Nacirema*, Kent Flannery's *The Early Mesoamerican Village* (1976) y *The Golden Marshalltown* (1982), y John Shea's *Uwasi Valley Tales de «Prehistoric Stone Tools of Eastern Africa: A Guide»* (2020).

que los asentamientos posteriores de Monico muestran evidencia de producción intensiva de alimentos, comercio de productos exóticos y una organización política altamente centralizada administrada por una élite cada vez más jerárquica. (Figura 1).



**Figure 1:** Una reconstrucción de la cultura arqueológica ficticia de Monico, del sitio Monico-1 (ver texto a continuación).

Un famoso arqueólogo bayesiano, una autoridad en Monico, ha excavado un sitio del período posterior al contacto asociado con esta cultura. El trabajo de excavación en el sitio, llamado Monico-1, ha producido un conjunto de fauna impresionante entre la cultura material ampliamente diversa. La arqueofauna está compuesta por dos especies de animales: «perro» y «coyote». Los animales individuales de ambas especies están representados por esqueletos completos. En consecuencia, en este informe, el arqueólogo utiliza el término «individuo» para referirse a perros o coyotes completos. Asimismo, cuando el arqueólogo menciona «el número de» perros o coyotes, se refiere a un conteo de individuos completos de la respectiva especie. Hasta el momento, el arqueólogo ha identificado 100 de estos individuos y los ha asignado a sus respectivas

especies. Con base en las observaciones, el conjunto está compuesto por 71 perros y 29 coyotes (Figura 2).



**Figure 2:** El arqueólogo bayesiano y su equipo excavan el sitio Monico-1.

Sin embargo, el arqueólogo también ha excavado un fragmento de hueso que es difícil de identificar. El arqueólogo quiere saber la especie más probable a la que pertenece este fragmento.

El arqueólogo define «probabilidad» como la frecuencia relativa o proporción de veces que ocurre un evento. Sobre la base únicamente de los datos, la probabilidad ( $P$ ) de que el perro permanezca en el conjunto es:

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

mientras que la probabilidad de que quede un coyote es:

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

Dadas estas probabilidades, es razonable que el arqueólogo crea que es más probable que el espécimen óseo no identificable sea de un perro. Sin embargo, el arqueólogo se muestra escéptico. Además, como

estudiosos de Monico, el arqueólogo tiene detalles etnográficos sobre el comportamiento de la gente de Monico, particularmente sobre sus tabúes alimenticios. Los relatos históricos revelan que los Monico alguna vez mantuvieron perros de caza en sus aldeas para cazar coyotes. Debido a que la base de subsistencia tradicional de Monico dependía de la caza del coyote, los perros desarrollaron relaciones especiales con sus dueños. En consecuencia, los Monico llegaron a tratar a sus perros con respeto, como lo harían con otras personas.

Las historias verbales transmitidas de generación en generación han documentado que se pensaba que los perros eran hermanos cercanos de las personas. En particular, se sabe que la cultura Monico tenía tabúes contra matar o comer perros. Sin embargo, las historias verbales también han revelado que los Monico comían perros en tiempos de grave escasez de alimentos. Con esta información adicional o «previa», el arqueólogo decide observar los esqueletos más de cerca para comprobar la presencia de marcas de matanza (es decir, marcas de corte) en los restos del perro. El arqueólogo tabula esta información adicional sobre los huesos recuperados bajo dos condiciones de matanza: 1) las marcas de matanza están presentes y 2) las marcas de matanza están ausentes. La Tabla 1 muestra las frecuencias de marcas de matanza en los esqueletos de cada especie.

**Table 1:** Frecuencias de animales individuales y marcas de matanza observadas en Monico1. Tenga en cuenta que, si bien la mayoría de las marcas de matanza están en huesos de coyote, 9 de los 71 huesos de perro también muestran signos de matanza.

	Individuos de cada especie		Total_de_marcas_de_matanza
	Coyote	Perro	
<b>Marcas de matanza</b>			
Presente	23	9	32
Ausente	6	62	68
<b>Total de individuos</b>	<b>29</b>	<b>71</b>	<b>100</b>

Para convertir estos datos en una tabla de probabilidad, el arqueólogo estandariza (o divide) todos los valores por el número total de observaciones (100 en este caso). Las celdas internas (fuente oscura, sombreado claro) en la Tabla 2 proporcionan las probabilidades de que las marcas de matanza y las especies ocurran juntas o en conjunto, lo que se conoce como probabilidades conjuntas.

**Table 2:** Probabilidades conjuntas de animales individuales y marcas de matanza observadas que describen la probabilidad de identificar una especie y observar marcas de matanza en los huesos de esa especie por ejemplo, P(Coyote and Butchery mark present) es 0.23, o 23%.

	Species Identified		Marcas_marginales_de_matanza
	P.Coyote.	P.Perro.	
<b>Marcas de matanza</b>			
P(Presente)	0.23	0.09	0.32
P(Ausente)	0.06	0.62	0.68
<b>Especies marginales</b>	<b>0.29</b>	<b>0.71</b>	<b>Total = 1</b>

Los valores en los márgenes derecho e inferior de la Tabla 2 se denominan adecuadamente probabilidades marginales. Estos representan la presencia y ausencia de marcas de matanza (a la derecha) y la especie identificada (abajo). Los totales marginales son las probabilidades totales de cada espacio subdividido (especie o marca de matanza). Por definición, todas las probabilidades se encuentran en el rango de 0 a 1, y la suma total de las filas o columnas marginales (es decir, la suma de todos los resultados marginales) debe ser 1. En este punto, el arqueólogo se enfoca en el espécimen de hueso no identificable y encuentra varias marcas de matanza en él. El arqueólogo puede utilizar esta información adicional para obtener una ventaja inferencial al tener en cuenta o condicionar la presencia de marcas de matanza, un proceso llamado condicionamiento. El arqueólogo condiciona la especie identificada a la presencia o ausencia de marcas de matanza. Este procedimiento también se conoce como subconjunto o estratificación de la variable «especies identificadas» por la presencia o ausencia de marcas de matanza. Naturalmente, el arqueólogo pregunta: «¿Cuál es la probabilidad de que el espécimen óseo no identificable sea de un perro en comparación con la probabilidad de que sea de un coyote, dado que hay marcas de matanza en los huesos del individuo?». El arqueólogo observó 32 animales de Monico-1 con marcas de matanza presentes. De esos, las marcas de matanza estaban presentes en 9 perros y 23 coyotes. El arqueólogo puede así calcular las probabilidades de que el individuo pertenezca a una especie u otra, dado que hay marcas de matanza (los estadísticos usan el símbolo «|» para referirse a «dado que» y para indicar que se está produciendo un condicionamiento). Para un perro, la probabilidad es:

$$P(\text{Perro} | \text{Marca de matanza presente}) = \frac{9}{32} = 0.28.$$

mientras que la probabilidad de que un individuo con marcas de matanza pertenezca a la especie coyote es:

$$P(\text{Coyote} \mid \text{Marca de matanza presente}) = \frac{23}{32} = 0.72.$$

Por lo tanto, después de observar las marcas de matanza en el hueso individual (no identificado), el arqueólogo puede afirmar que la probabilidad de que provenga de un coyote es de 0.72. En otras palabras, tienen un 72 % de certeza de que el hueso provino de un coyote. Unos días después, un reportero de un periódico local se enteró de una excavación arqueológica en curso en otro sitio cercano de la aldea de Monico, llamado Monico-2. Fuentes revelan al reportero que las excavadoras allí también están recuperando restos de fauna. Debido a que el arqueólogo es un conocido experto en los hábitos alimenticios de Monico, el reportero se pone en contacto con el arqueólogo y le comunica el hecho de que el nuevo conjunto de fauna en Monico-2 está compuesto en su totalidad por restos de especies de perros. Aunque los investigadores de Monico-2 todavía no han realizado un análisis completo de la fauna, el reportero le pregunta al arqueólogo qué tan probable es que los Monico estuvieran matando y comiendo perros en el nuevo sitio. Por ahora, el arqueólogo ha estimado las probabilidades de encontrar marcas de matanzas asociadas con cada especie animal en base a la experiencia en el pueblo de Monico-1. Para hacer una inferencia probabilística sobre el comportamiento en el nuevo sitio, el arqueólogo condiciona las «especies identificadas» en lugar de la «presencia de marcas de matanza». De los 71 perros identificados en Monico-1, el arqueólogo observó 9 con marcas de matanza y 62 sin ellas. Esto significa que, según la evidencia de Monico-1, la probabilidad de encontrar evidencia de matanza de perros es:

$$P(\text{Marca de matanza presente} \mid \text{Perro}) = \frac{9}{71} = 0.13.$$

mientras que la probabilidad de que no haya evidencia de matanza en perros es:

$$P(\text{Marca de matanza ausente} \mid \text{Perro}) = \frac{62}{71} = 0.87.$$

Después de pensar un momento, el arqueólogo le dice al reportero que (basado en el conocimiento de Monico-1) la probabilidad de que los huesos de perro de Monico-2 hayan resultado de actividades dietéticas es relativamente baja, alrededor del 13 %. Este cálculo se basa en el teorema de Bayes, así como en la información sobre la relación de Monico con sus perros y las prácticas de matanza en Monico-1.

## QUE ES EL TEOREMA DE BAYES

El teorema de Bayes es la formalización algebraica del trabajo de tabla probabilística que realizamos en la sección anterior utilizando un evento discreto. El teorema es más útil cuando se conoce un enunciado

de probabilidad condicional y se desea obtener su enunciado condicional inverso. Por ejemplo, del modelo anterior, sabemos que  $P(\text{Marca de matanza presente} | \text{Perro}) = 0.13$ . Si deseamos conocer el enunciado condicional inverso  $P(\text{Perro} | \text{Marca de matanza presente})$ , podemos calcularlo usando:

$$P(\text{Perro} | \text{Marca de matanza presente}) = \frac{P(\text{Butchery mark present} | \text{Dog}) \times P(\text{Dog})}{P(\text{Butchery mark present})}$$

Las tablas 1 y 2 proporcionan los valores necesarios para sustituir esta expresión de manera que:

$$P(\text{Perro} | \text{Marca de matanza presente}) = \frac{\left(\frac{0.09}{0.71}\right) \times 0.71}{0.32} = 0.28.$$

Cuando se generaliza, el algoritmo aplicado aquí es conocido como teorema de Bayes. Por lo general, se exemplifica considerando dos eventos relacionados: A y B. En pocas palabras, el teorema de Bayes establece que:

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

En este caso, para obtener la probabilidad condicional de A dado B,  $P(A|B)$ , hay que dividir la probabilidad conjunta de A y B,  $P(A \text{ y } B)$ , por la probabilidad marginal de B,  $P(B)$ . El producto de  $P(B|A)$  y  $P(A)$  es la probabilidad conjunta,  $P(A \text{ y } B)$ . La fórmula entonces se generaliza a:

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

,

Donde la probabilidad conjunta se divide por el  $P(B)$  marginal. Los estadísticos llaman a  $P(A|B)$  la probabilidad posterior de A dado B;  $P(B|A)$  el condicional inverso (o verosimilitud) de B dado A; y  $P(A)$  la probabilidad previa de A.

### **El arqueólogo bayesiano continuó**

Unos días después, el reportero adquiere más información de las continuas excavaciones en el pueblo de Monico-2. Las frecuencias y probabilidades conjuntas se describen en las Tablas 3 y 4 a continuación. El reportero está muy emocionado de informar al arqueólogo que los excavadores habían recuperado 10 perros, todos menos uno tenían marcas de matanza. Por el contrario, los arqueólogos del sitio Monico-2 habían recuperado solamente un coyote que exhibía marcas de matanza en los restos. Los investigadores de Monico-2 utilizaron una estadística de prueba NHST apropiada, la prueba z unilateral para proporciones (Diez, Cetinkaya-Rundel, and Barr 2019: 194-197), con corrección de continuidad, para probar si la tasa de

matanza de perros observada (9/10) era estadísticamente significativamente superior al 50 %, la hipótesis nula predeterminada en esta prueba. Los arqueólogos de Monico-2 rechazaron la hipótesis nula con un valor de p < 0.05 (ratio z = 2.21, media = sdev = 1.58, p = 0.013). Debido al pequeño tamaño de la muestra, también realizaron una prueba binomial unilateral, que arrojó resultados en línea con los resultados de la prueba z (éxitos = 9, intentos = 10, p = 0.01074). Con base en estos resultados estadísticamente significativos, los arqueólogos de Monico-2 le dijeron al reportero que la mayoría de los perros fueron sacrificados en el sitio. Además, según el reportero, los arqueólogos también sugirieron que la evidencia y los resultados del análisis estadístico indicaron que la gente de la aldea Monico-2 incluía perros como una parte importante de su dieta. A la luz de esta evidencia, el reportero comienza a cuestionar el registro etnográfico sobre los tabúes dietéticos de los Monico. El arqueólogo de Monico-1 echa un vistazo rápido a las tablas, hace algunos cálculos y sostiene que la probabilidad de que los aldeanos de Monico-2 hayan sacrificado a sus perros ahora es incluso menor, especialmente en comparación con la nueva probabilidad de matanza de coyotes, que es ligeramente superior. Sin embargo, el arqueólogo insiste en esperar una muestra más grande antes de sacar conclusiones firmes. Incrédulo, el reportero pide una explicación de por qué el arqueólogo cuestiona las significativas pruebas de hipótesis nulas realizadas por los arqueólogos de Monico-2. El arqueólogo mira al reportero y dice: «Bueno, los procedimientos del NHST como la prueba z únicamente consideran datos nuevos. Desafortunadamente, estos métodos no tienen en cuenta toda la información disponible, nueva y previa, sobre la subsistencia de Monico. Personalmente -prosigue el arqueólogo- intento no formar mis opiniones basándome únicamente en nuevos datos. Más bien, utilizo nuevos datos con el fin de actualizar mis opiniones existentes hechas con conocimiento previo, por ejemplo, del sitio Monico-1». Luego, el arqueólogo guía al reportero a través de las tablas y comienza a explicar cómo hacen su inferencia usando el teorema de Bayes.

**Table 3:** Frecuencias de animales individuales y marcas de matanza observadas en el pueblo de Monico-2. Tenga en cuenta el pequeño número de individuos y la muestra particularmente pequeña de coyotes.

	Individuos de cada especie		Total_de_marca_de_matanza
	Coyote	Perro	
<b>Marcas de matanza</b>			
Presente	1	9	10
Ausente	0	1	1
<b>Total de Individuos</b>	<b>1</b>	<b>10</b>	<b>Total = 1</b>

**Table 4:** Probabilidades conjuntas de animales individuales y marcas de matanza observadas en la aldea Monico-2. Tenga en cuenta la mayor proporción de huesos de perro con marcas de matanza en comparación con la muestra de Monico-1.

	Species Identified		Marcas_marginales_de_matanza
	P.Coyote.	P.Perro.	
<b>Marcas de matanza</b>			
P(Presente)	0.09	0.82	0.91
P(Ausente)	0.00	0.09	0.09
<b>Especies marginales</b>	<b>0.09</b>	<b>0.91</b>	<b>Total = 1</b>

El arqueólogo explica que las probabilidades posteriores de la matanza de perros y coyotes extraídas del conjunto de fauna (mucho más grande) Monico-1 se han convertido en nueva información «previa» sobre las probabilidades de que los aldeanos de Monico mataran perros y coyotes. Estas cantidades pueden ser representadas por:

$$P(\text{Marca de matanza presente} \mid \text{Perro})_{\text{Monico -1}} = \frac{9}{71} = 0.13,$$

y

$$P(\text{Marca de matanza presente} \mid \text{Coyote})_{\text{Monico -1}} = \frac{23}{29} = 0.79,$$

El conocimiento del arqueólogo sobre el grado en que los aldeanos de Monico-1 sacrificaron perros y coyotes puede actualizarse en una nueva iteración del teorema de Bayes que incluye los datos de Monico-2. Para dar cuenta del contexto arqueológico del que derivan los cálculos, el arqueólogo añade los subíndices  $-1$  y  $-2$  a los términos de la ecuación, de la siguiente manera:

$$\begin{aligned} P(\text{Marca de matanza presente} \mid \text{Perro})_{\text{Monico -2}} &= \frac{P(\text{Perro} \mid \text{Matanza})_{\text{Monico -2}} \times P(\text{Perro} \mid \text{Matanza})_{\text{Monico -1}}}{P(\text{Perro})_{\text{Monico -2}}} \\ &= \frac{\frac{0.82}{0.91} \times 0.13}{0.91} \\ &= 0.13. \end{aligned}$$

Añadir los datos de perros de Monico-2 hace que la probabilidad de matar perros disminuya ligeramente (de 0.127 a 0.126, pero redondeada a 0.13). Puede utilizarse la misma operación utilizando los datos anteriores de la primera excavación y los nuevos coyotes:

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

En este caso, después de actualizar los datos, la nueva probabilidad posterior de la matanza del coyote también es mayor (cambiando aún más respecto a la probabilidad previa que en el caso de los perros). El arqueólogo le explica esto al reportero. Además, el arqueólogo insta a la cautela dado que los datos y las probabilidades resultantes del sitio original se derivaron de una muestra de 100 individuos, mientras que la selección actual representa un total de solamente 11. Aunque los cálculos de probabilidad son correctos, sería prudente esperar más datos, ya que la excavación en Monico-2 está en curso. Sin embargo, el análisis bayesiano del arqueólogo sugiere que, en este punto, no deberíamos esperar marcas de matanza en ningún perro recién descubierto en el sitio Monico-2.

## VINCULACIÓN DEL TEOREMA DE BAYES A DATOS E HIPÓTESIS

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:

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

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  $\theta$  (*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)  $\theta$ , 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:

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

### 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,  $\theta$  (the proportion of dogs with butchery marks).

Table 5 illustrates the distribution of  $\theta$  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  $\theta$  between 16% and 35%. Meanwhile, another two archaeological sites report that  $\theta$  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 ( $\theta$ ) 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 ( $\theta$ )	Number of sites with ( $\theta$ ) evidence of butchery on dog bones	Proportion of total number of sites with ( $\theta$ ) evidence of butchery on dog bones (prior probability)
0-0.05	20	0.53
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

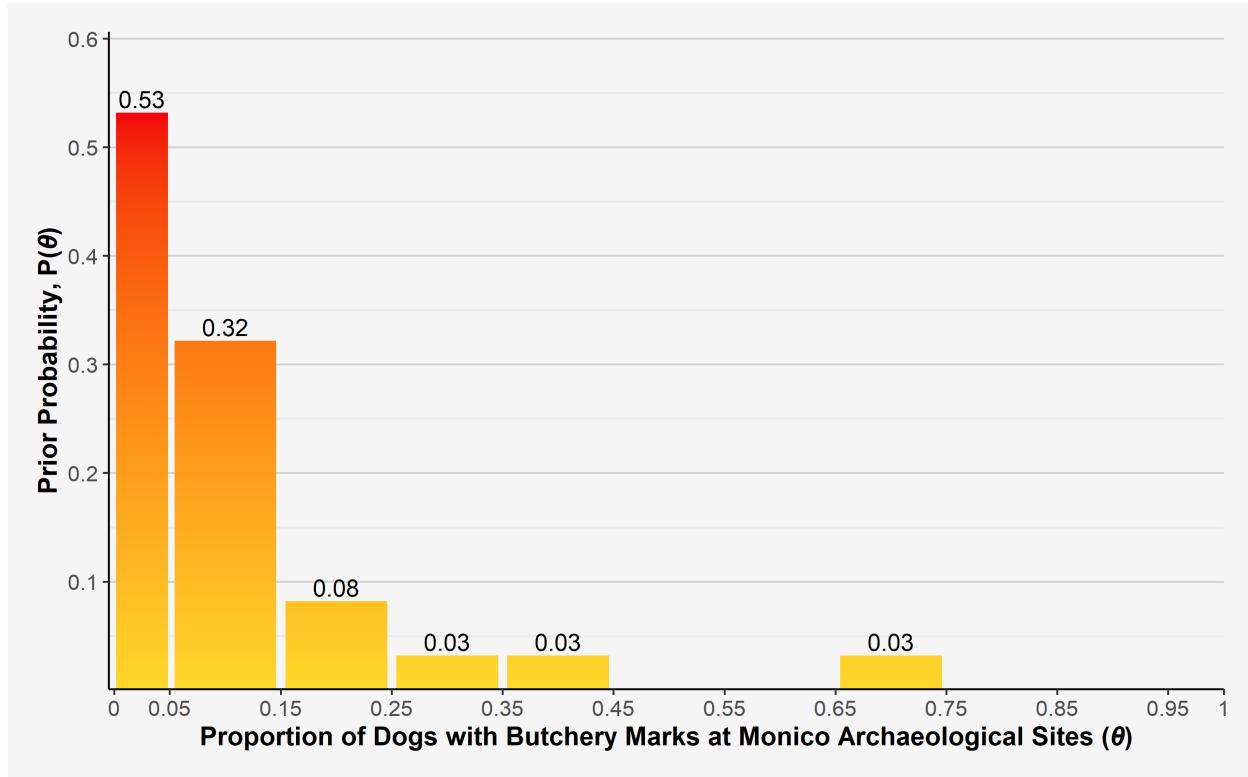
The proportion of dog remains with butchery marks ( $\theta$ )	Number of sites with ( $\theta$ ) evidence of butchery on dog bones	Proportion of total number of sites with ( $\theta$ ) evidence of butchery on dog bones (prior probability)
0.75-0.85	0	0.00
0.85-0.95	0	0.00
0.95-1	0	0.00

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  $\theta$  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  $\theta$  at Monico village sites.

The archaeologist first records the dog butchery proportions ( $\theta$ ), from the 38 sites found in the literature. To summarize these data, in Table 5 (column 1), they group the  $\theta$  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  $\theta$  values in each interval (column 2). The archaeologist then calculates the prior probabilities for each  $\theta$ -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  $\theta$ -interval. This frequency distribution also serves as the prior distribution of  $\theta$  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  $\theta$  (theta), the proportion of dogs with butchery marks at Monico archaeological sites (from Table 5). Note that small values of  $\theta$  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  $\theta$  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 3), the archaeologist needs a model to represent the probability of the data,  $x$ , and associated parameter(s),  $\theta$ , 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,  $\theta$ . 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,  $\theta$ , 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  $\theta$  that would have generated the data. ML asks, under the binomial model, which value of  $\theta$  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  $\theta$  parameter maximizes  $P(x|\theta)$ , the likelihood, over a systematic range of quantities between 0 and 1.

To estimate the most likely value of  $\theta$ , the archaeologist assumes that the probability of observing each butchered dog is independent of the others, making the probability of observing 9 butchered dogs  $\theta^9$ . 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  $\{1, 1, 1, 1, 1, 1, 1, 1, 1, 0\}$ <sup>2</sup>, ... etc. The binomial coefficient is shorthand, and may be calculated using the following equation:

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

---

<sup>2</sup>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.,  $\binom{10}{5} = 252$ .

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

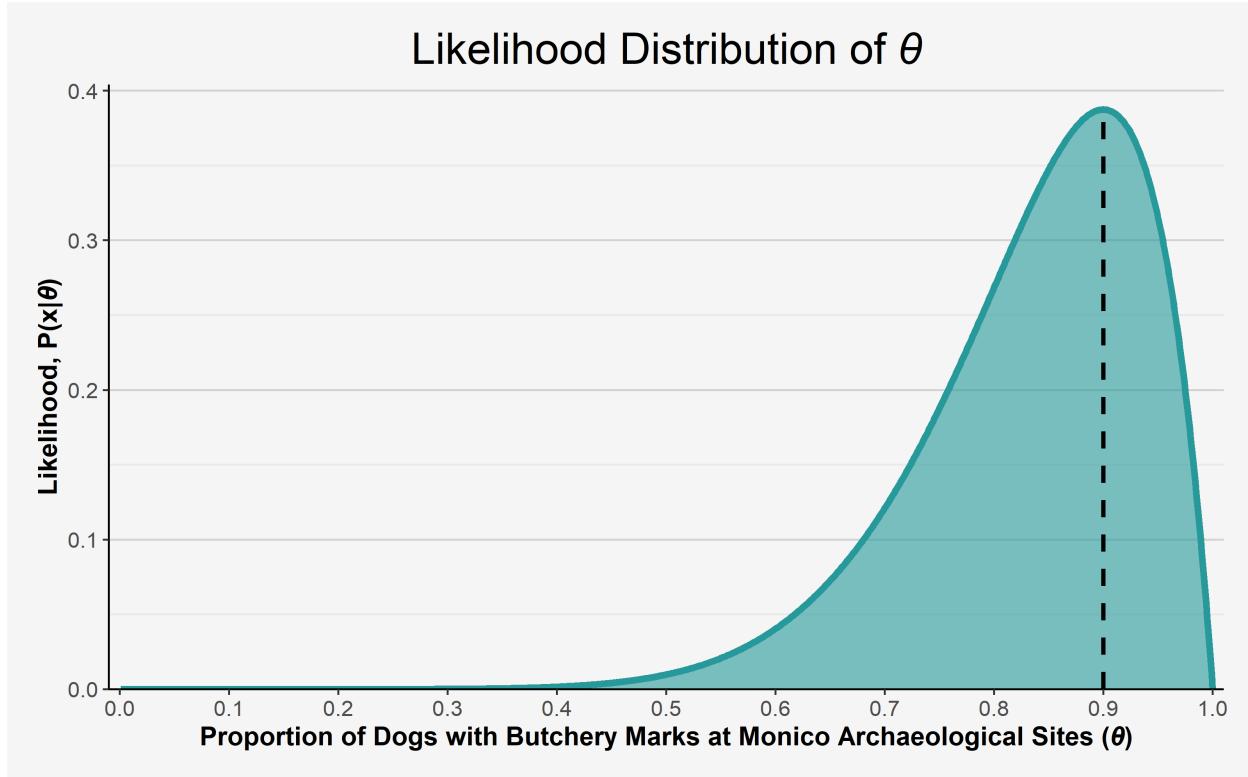
Therefore,

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

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

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

across the range of  $\theta$  values from 0 to 1 to find the likelihood distribution of the data and, thus, the value of  $\theta$  that maximizes the likelihood function. This approach is illustrated in Figure 4, from which the archaeologist learns that the ML estimate of  $\theta$  (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 ( $\theta$ ) is 0.9 (or 90%).



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

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  $\theta$ , also known as the prior. In this case, the archaeologist needs a probability function that models the distribution of  $\theta$ , 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  $\theta$ . The mathematical expression of the Beta model is:

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

The shape of the Beta model is thus controlled by two parameters,  $a$  and  $b$ , which in turn control

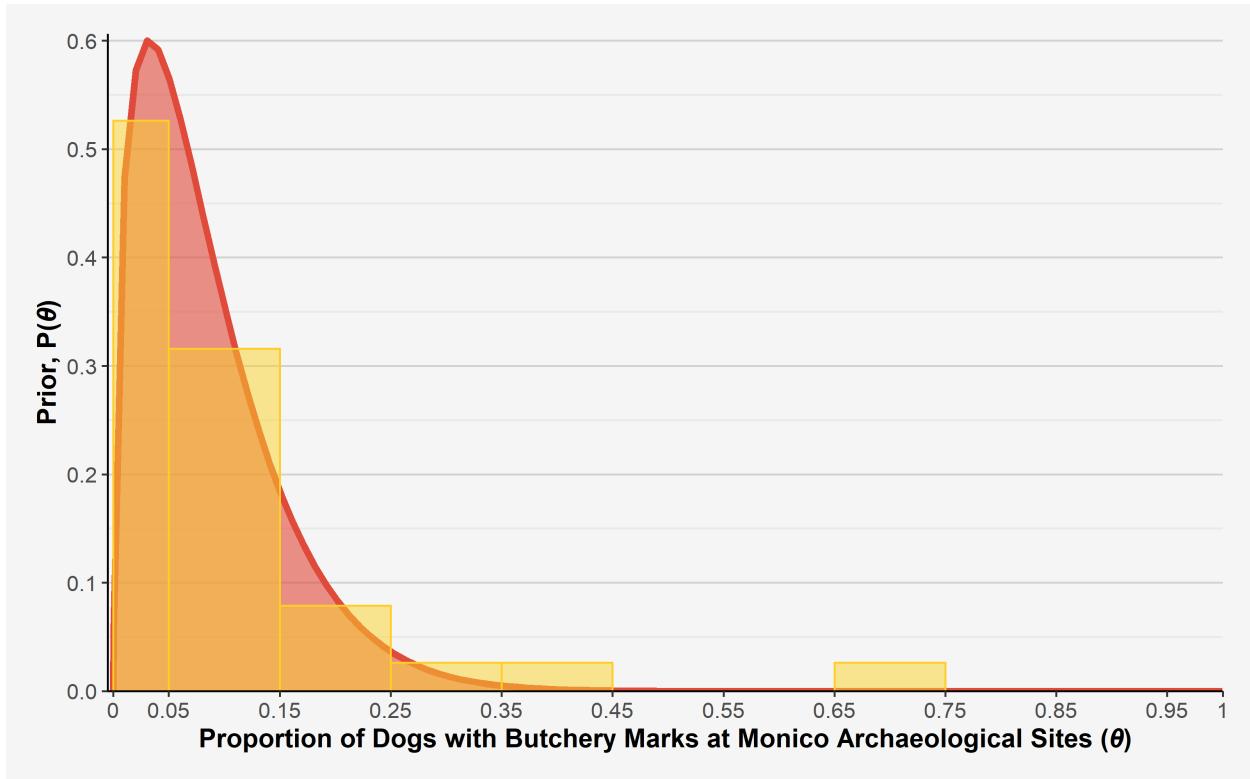
---

<sup>3</sup>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

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  $\theta$  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  $\theta$  (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  $\theta$  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

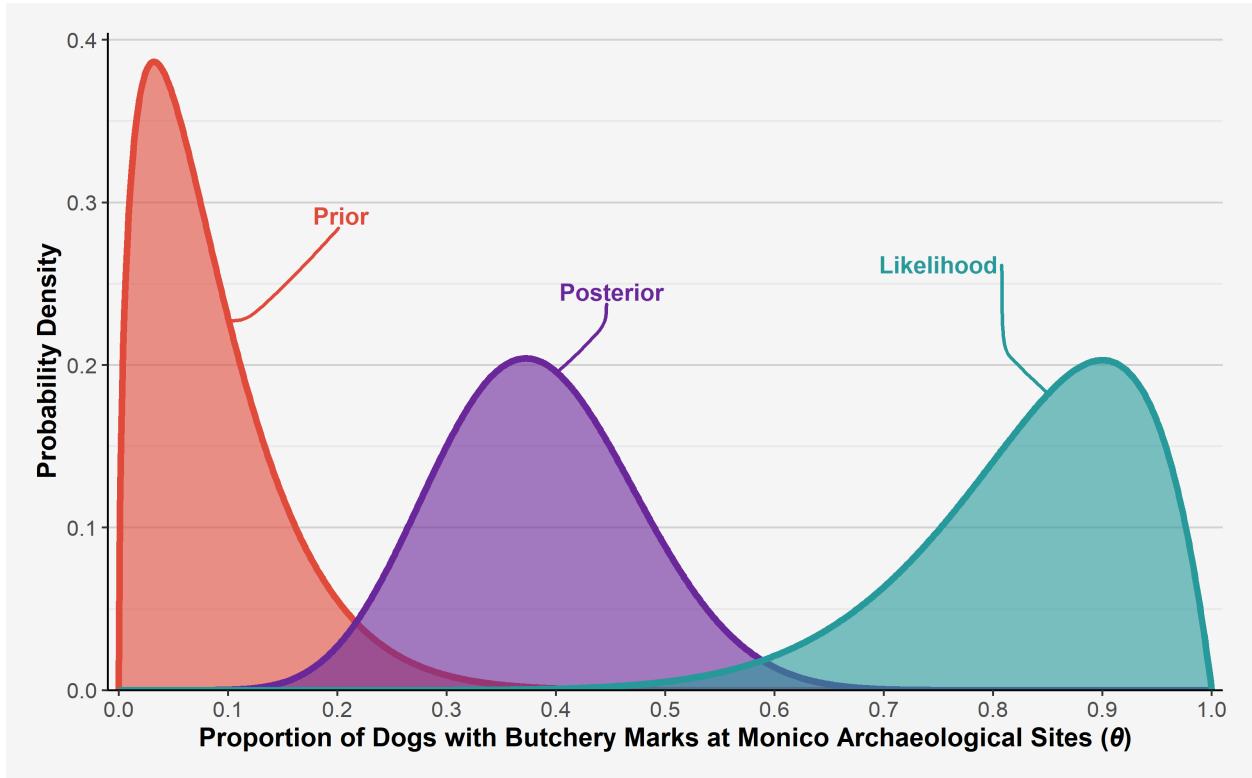
The archaeologist is aware that statisticians frequently use the binomial and beta distributions in the context of Bayesian analyses because they work well together for modelling the likelihood and prior probability distributions, respectively, simplifying the calculations needed to compute the posterior. 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:

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

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

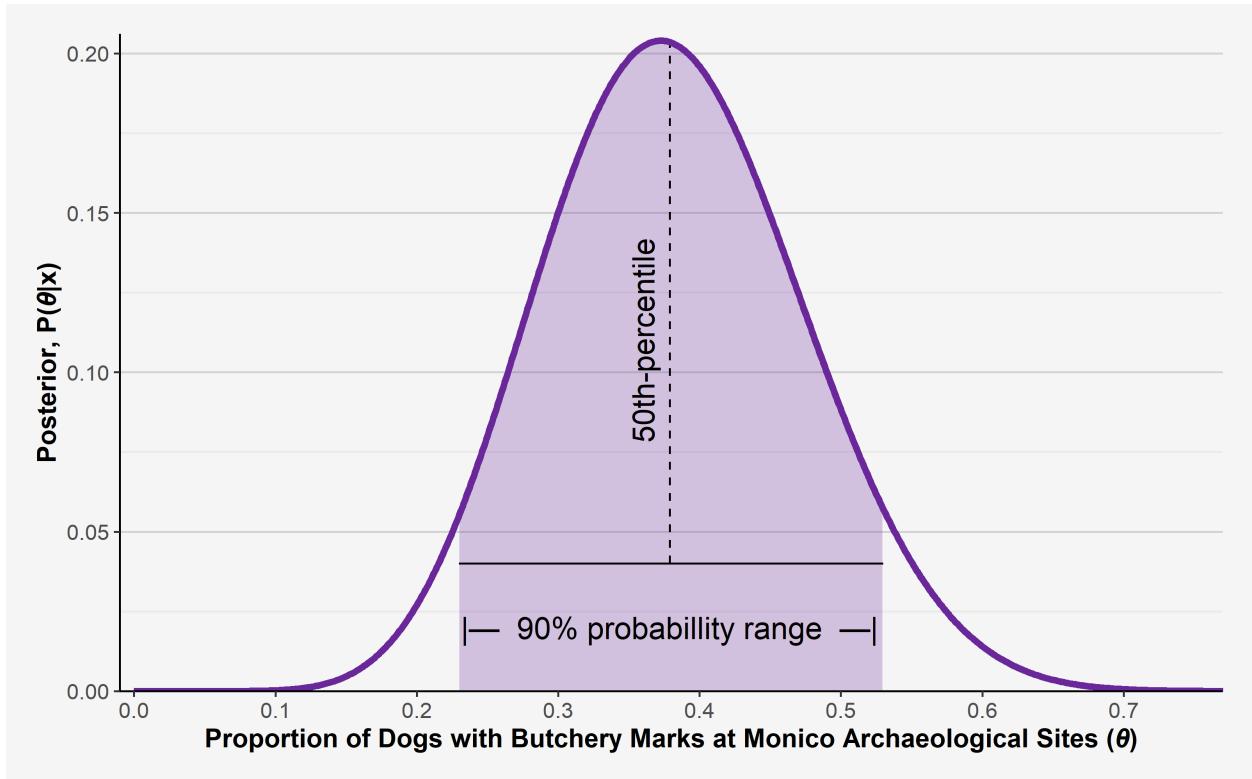
They thus generate values of  $P(x|\theta)$ , the likelihood, and  $P(\theta)$ , the prior probabilities, to calculate  $P(\theta|x)$ , the posterior probability distribution across a fine grid of  $\theta$  values in the 0, 1 interval (1,000 values between 0, and 1). These are illustrated in Figure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



**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.

The archaeologist then focuses on  $P(\theta|x)$ , the posterior distribution. The posterior will help them to make inferences about the probability of  $\theta$  and its surrounding uncertainty (Figure 6). The archaeologist can visually represent the estimate of  $\theta$  (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.



**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).

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  $\theta$ , 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  $\theta$  is greater than 0.5 (top left-most value in the table), and the values of  $\theta$  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  $\theta$  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  $\theta$  can also be expressed as a probability interval. In 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  $\theta$ , Figure 7 tells the archaeologist that there is quite a lot of uncertainty around the true value of  $\theta$ . For example, the median, or 50<sup>th</sup> percentile, estimate of  $\theta$  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  $\theta$  is between these values and only a 10% chance that it is larger or smaller than these limits. Although the variation in  $\theta$  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  $\theta$  from posterior probability distribution

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

## CONCLUSIONES

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.

## **AGRADECIMIENTOS**

This work did not require permits. EOC is thankful to Deb Nichols, John Watanabe, Sophie Nichols-Watanabe, Robert (Bob) L. Kelly, and the Dartmouth Coach for inspiring and facilitating the development of some concepts in this paper. In addition, Amanda Veile, Mike Shott, Eduardo Fernandez-Duque, and two anonymous reviewers provided constructive comments on earlier drafts of this manuscript. Warren Muzak's (<http://www.warrenmuzak.com/>) stunning illustrations allowed the fictional Monico culture to come to life. Finally, Sarah Herr and the AAP editorial team have been delightful and made the submission and peer-review process fantastic.

## 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. %3CGo%20to%20ISI%3E://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](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 Caitlin 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](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 Yasser. 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](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>.