Bayesian Analyses of Political Decision Making

Kumail Wasif, School of Public Affairs, American University
and Jeff Gill, Department of Government, American University

https://doi.org/10.1093/acrefore/9780190228637.013.1002
Published online: 27 October 2020

Summary

Bayes' theorem is a relatively simple equation but one of the most important mathematical principles discovered. It is a formalization of a basic cognitive process: updating expectations as new information is obtained. It was derived from the laws of conditional probability by Reverend Thomas Bayes and published posthumously in 1763. In the 21st century, it is used in academic fields ranging from computer science to social science.

The theorem's most prominent use is in statistical inference. In this regard, there are three essential tenets of Bayesian thought that distinguish it from standard approaches. First, any quantity that is not known as an absolute fact is treated probabilistically, meaning that a numerical probability or a probability distribution is assigned. Second, research questions and designs are based on prior knowledge and expressed as prior distributions. Finally, these prior distributions are updated by conditioning on new data through the use of Bayes' theorem to create a posterior distribution that is a compromise between prior and data knowledge.

This approach has a number of advantages, especially in social science. First, it gives researchers the probability of observing the parameter given the data, which is the inverse of the results from frequentist inference and more appropriate for social scientific data and parameters. Second, Bayesian approaches excel at estimating parameters for complex data structures and functional forms, and provide more information about these parameters compared to standard approaches. This is possible due to stochastic simulation techniques called Markov Chain Monte Carlo. Third, Bayesian approaches allow for the explicit incorporation of previous estimates through the use of the prior distribution. This provides a formal mechanism for incorporating previous estimates and a means of comparing potential results.

Bayes' theorem is also used in machine learning, which is a subset of computer science that focuses on algorithms that learn from data to make predictions. One such algorithm is the Naive Bayes Classifier, which uses Bayes' theorem to classify objects such as documents based on prior relationships. Bayesian networks can be seen as a complicated version of the Naive Classifier that maps, estimates, and predicts relationships in a network. It is useful for more complicated prediction problems. Lastly, the theorem has even been used by qualitative social scientists as a formal mechanism for stating and evaluating beliefs and updating knowledge.

Keywords: Bayes' theorem, statistical inference, machine learning, Markov Chain Monte Carlo, quantitative social science, qualitative Bayesian analysis, Bayesian networks, Naive Bayes Classifier, political decision making

Introduction

No discussion of systematic decision making is complete without including Bayes' theorem. Although this is a relatively simple equation, it is one of the most important mathematical principles discovered, and increasingly so with 21st century technology. It is a formalization of a basic cognitive process: updating expectations as new information is obtained. This, along with the versatility of the formula, makes it an exceptionally useful tool for scientists studying social phenomenon using a variety of methods. This article provides an overview of the primary uses of Bayes' theorem in the social sciences, which are an extremely active area of scientific research.

Bayesian approaches have enhanced conventional methods and created new ones. The discovery of Bayes' theorem dates back to the 18th century with the publication of Bayes's essay "An Essay Towards Solving a Problem in the Doctrine of Chances" (1763), although there is evidence that Laplace was working on the same principle simultaneously (1774, 1781). For a long time, this was just a basic staple of probability theory and Bayesian mathematical statistics. Due to explosive increases in computational technology in the early 21st century, Bayesian updating and inference are used as foundations of machine learning, artificial intelligence, and general decision theory.

There are three essential tenets of Bayesian thought that distinguish it from standard approaches. First, any quantity that is not known as an absolute fact is treated probabilistically, meaning that a numerical probability or a probability distribution is assigned. Second, research questions and designs are based on prior knowledge and expressed as prior distributions. Finally, these prior distributions are updated by conditioning on new data as they are observed to create a posterior distribution that is a compromise between prior and data knowledge. The posterior distribution is the key to Bayesian inference and Bayesian decision making since it includes all of the relevant information required to make some inference about an unknown quantity. The quality of the decision based on the posterior distribution is conditional on the level of information contained in that posterior: peaked forms imply high levels of knowledge and diffuse forms imply low levels of knowledge. This is a rigorous mathematical process.

Bayes' Theorem

Bayes's famous essay (1763) was found and published two years after his death by his friend Richard Price. A reverend and amateur mathematician, Bayes developed a probability theorem that relates the order of conditional probabilities, based on a uniform prior distribution. More formally, let A and B be two non-independent events. The law of conditional probability states that the probability of A occurring given that B is occurring is given by,

$$\mathrm{p}\Big(\mathrm{A}|\mathrm{B}\Big) = \frac{p\left(A \cap B\right)}{p\left(B\right)} \tag{1}$$

with $p(A \cap B)$ being the joint probability of both events A and B occurring (also denoted at $p(A, B)$) and $p(B)$ the probability that event B occurs. We thus get the probability of A occurring given that B is occurring in $p(A|B)$. For example the probability that the outcome of the roll of a single die being an even number is $p(A) = 3/6$, but conditional on the outcome being greater than 3, event B, is $p(A|B) = 2/3$. We can also get a conditional probability in the opposite direction, that is the probability of B occurring given that A is occurring, by a slight change in the formula,

$$p(B|A) = \frac{p(B \cap A)}{p(A)}, \tag{2}$$

which is a fundamentally different quantity than $P(A|B)$. However, the joint probability remains the same in both directions meaning that,

$$p\left(A \cap B\right) = p(B \cap A). \tag{3}$$

By simple algebraic manipulation of (1) and (2) using (3), we get the following result:

$$\begin{aligned} &p\left(A \cap B\right) = p\left(A|B\right)p\left(B\right) \qquad p\left(B \cap A\right) = p\left(B|A\right)p\left(A\right) \\ &p\left(A|B\right)p\left(B\right) = p\left(B|A\right)p\left(A\right) \\ &p\left(A|B\right) = \tfrac{p(B|A)p(A)}{p(B).} \end{aligned} \tag{4}$$

This result describes the relationship between reversed-order conditional probabilities showing how the probability of A given B relates to the probability of B given A.

It is not immediately clear how this simple manipulation of conditional probability becomes an engine of modern inference. Consider $p\left(A\right)$ as unconditional prior knowledge on some parameter of interest. Here A is an unknown parameter affecting some data generation process and is therefore treated probabilistically under essential Bayesian tenet number one. Now think about $p\left(B|A\right)$ as the likelihood of seeing some data B given the parameter A that affects this process. In a very basic case, the Poisson intensity parameter λ affects the data X that can be drawn given a Poisson probability mass function (PMF):

$p(X|\lambda) = \exp[-\lambda]\lambda^X/X!$. Using Bayes' theorem we can produce the posterior distribution of A conditional on B, $p(A|B)$, which updates our knowledge about this parameter by incorporating more information assuming that this information is relevant and correct.

Notice that $p(B)$ was left out of this discussion so far. There is both a philosophical and practical reason for this. First, note that we "saw" B, meaning that it happened: some data were observed. Therefore, this is now longer a random quantity and therefore no longer has a distribution, much less an unconditional one: to Bayesians the probability of observing data that has been observed is 1. So in practical terms the purpose of the quantity $p(B)$ is simply to make sure that the posterior distribution is proper: it sums or integrates to 1 (depending on the level of measurement). Furthermore, the posterior distribution can be normalized later to meet this standard condition, which is why Bayes' law is often shown in proportional terms:

$$p(A|B) \propto p(B|A)p(A) \tag{5}$$

This version reduces the complexity of analytical calculations, which was important before modern Bayesian computing. Note that there are uses for $p(B)$ in other settings such as Bayesian model comparison, but in the discussion here this is not germane.

An Illustrative Example

This can be illustrated with a contrived example. Suppose we know that the probability of Ann going to a party is $p(A) = 0.50$, but then we receive new information that Ann going to the party is influenced by her friend Ben going. The probability of Ben going is known to be $p(B) = .70$. We also know that the probability of Ben going if Ann is going is $p(BA) = 0.80$. All these probabilities are derived from past social experience. We can use Bayes' theorem to update our expectations, that is, to calculate the probability of Ann going now that we know Ben is going $p(A|B) = p(B|A)p(A)/P(B) = 0.80 \times 0.50/0.80 = 0.57$. This is an increased probability compared to our prior one.

It should be noted that had the probability of Ben going been the same as his probability of going when Ann was going (i.e., $p(B) = p(B|A)$) then the numerator and denominator would cancel each other out, making our prior the posterior. The probability of Ann going would be unchanged because both events appear independent. However, if the probability of Ben going if Ann were not was greater than his probability of going when Ann was going (i.e., $p(B) > p(B|A)$) then Ann's probability of going would be decreased in the posterior. If this were the case, it raises the question: why is Ben more likely to go when Ann is not going? It could be because Ben is avoiding Ann or Ann is avoiding Ben or they are both avoiding each other or some other reason or just coincidence. We cannot know the real social story from this information alone.

This highlights an important caveat about conditional probabilities treated in this way: they are based on past observations. They cannot provide a causal explanation for these scenarios, since they are purely inferential. Since they imply a sequential relationship between the events, it is tempting to assume the relationship is causal. But the relationship is not

necessarily causal or even sequential. Conditional probabilities signify only that we assume the probability being conditioned on is a fixed quantity. As with other models in social science, causality has to be justified through a different process (Pearl, 2001).

Statistical Inference

Bayes' theorem is an extremely useful tool in basic probability theory, but its most powerful use is in applied statistics. This section explains the role of Bayes' theorem in inferential statistics.

Setting Up the Modeling Process

In the social example used to explain the basics of the theorem, we wanted to find the probability that our friend was going to the party, which was a discrete event out of two possible events (i.e., going or not going). Now we want to find the probability that the true parameter, which is one out of a potentially infinite number of parameters that are possibly not discrete, has a certain range or value. Consequently, instead of the probability of one potential outcome, we need the entire probability distribution of potential outcomes. For discrete variables, this is called the probability mass function (PMF) and for continuous variables it is called the probability density function (PDF). These distributions tell us the probability that our true parameter is equal to some value (for discrete parameters) or a range of values (for continuous parameters). They have multiplicative properties that allow them to be used effectively as terms in Bayes' theorem.

The general purpose of statistical inference is to use probability statements and data to estimate parameters of interest. The classic frequentist approach relies on an unending stream of independent and identically distributed data (IID) to make hypothesis decisions based on the assumption that parameters are fixed by nature and all uncertainty resided in the data. Thus the quantity of interest is $p(\boldsymbol{X}|\theta)$, the probability of the data given fixed parameter θ (Diaconis & Skyrms, 2019). This is at odds with the Bayesian posterior described previously, which in statistical notation is $p(\theta|\boldsymbol{X})$. Therefore the prior distribution of interest is $p(\theta)$, and the likelihood function is $p(\boldsymbol{X}|\theta)$. The latter is frequently labeled as $L(\theta|\boldsymbol{X})$, which looks like a reverse of the condition but is just notational since the likelihood is simply the joint distribution of the observed data, $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ given dependency on a parameter:

$$L\left(\theta|\boldsymbol{X}\right) = p\left(X_1|\theta\right) \times p\left(X_2|\theta\right) \times \ldots \times p\left(X_n|\theta\right). \tag{6}$$

Given the Bayesian treatment of observed data, $p(\mathbf{X}) = 1$, the canonical statement of Bayesian inference is given by:

$$p\left(\theta|\boldsymbol{X}\right) \propto p\left(\theta\right) L(\theta|\boldsymbol{X}). \tag{7}$$

This form of the expression also highlights the fact that, like in all Bayesian inference, the likelihood function here is stipulated in exactly the same way as any other likelihood-based statistical model, by assigning a parametric form for the data and plugging in the observed data. The maximum likelihood estimation process replaces the bounded notion of probability with the unbounded notion of likelihood. Fisher (1925) and others provide the standard and convenient inferential treatment of the likelihood function, which is to find the multidimensional mode of this function, which is usually log-concave to the x-axis, and to treat curvature around this mode as a measure of uncertainty. Standard forms of the probability mass/density function are used to guarantee a unimodal function concave to the x-axis. These procedures are incredibly well-established and well-accepted. See Gill and Torres (2019) for a recent detailed exposition.

Comments on the Prior Distribution

The first quantity required for Bayesian inference is the prior distribution. This is the presumed distribution of the parameter of interest before the data for the current analysis are collected and analyzed. It is based on previous knowledge about the effect of interest. As routine as this sounds it actually has quite a history of controversy. Influential early statisticians objected to the prior on two grounds: a perceived personalistic subjectivity, and the common stipulation of uniform priors at the time.

Today prior distributions range from very informative descriptions of previous research in the field to purposefully vague forms that reflect little or no prior knowledge about the effect in question. This level of information depends on the volume and reliability of previous studies on the topic of interest. This previous research need not be quantitative; it can be qualitative, narrative, or intuitive knowledge. But all informed prior distributions should be defended and researchers should show the impact of their prior distribution relative to another prior distribution and/or some benchmark.

Given the challenges of using informed priors, and the pointed discussions about their use, the overwhelming majority of Bayesian modelers in the social sciences use uninformed or minimally informed priors. These can be in the form of a uniform distribution, which means there is a constant probability that θ takes some value between a and b. It can also be in the form of a normal distribution with an average of 0 and high variance. A popular approach is to specify what are called conjugate priors that have mathematically convenient distributional forms that flow through to the posterior distribution (Gill & Witko, 2013). However, there are statistical limits to the specification of priors, including the need to have the same support over the range of the parameter: no density in substantively illogical regions, and no illogical parameter values such as negative variance.

Comments on the Posterior Distribution

We multiply the prior distribution with the likelihood function to produce the posterior distribution (since, as noted, the denominator can be ignored for now), which represents the most informed set of knowledge about the phenomenon of interest because it is the most updated version available. This process converts our prior knowledge into our posterior knowledge through conditioning with the likelihood function. Thus the posterior distribution is

a compromise between prior information and likelihood information. When the amount of data going into the likelihood function is large, it has more influence that pulls the posterior closer to its location. Conversely, when the amount of data is small and the prior distribution carries little information, the resulting posterior is closer to this prior. These countervailing effects are called shrinkage, and we can measure how much the likelihood function shrinks some statistic, say the posterior mode, toward its mode, away from the prior distribution.

The posterior distribution provides more information about the effect of interest than the typical summary from likelihood analysis. The latter provides a single point-estimate and a measure of curvature around it, whereas the former provides the distributional summary, which allows us to extract many more quantities of interest including the mode, mean, and median. In addition, the posterior distribution can also be treated as a new prior distribution if additional data are later observed. In this way the parameter of interest is updated and knowledge is accumulated over time. Suppose $p(\theta|\boldsymbol{X}_1)$ is the posterior from a model with data $\boldsymbol{X}_1$. If later a new data set, $\boldsymbol{X}_2$ is observed, we can use $p(\theta|\boldsymbol{X}_1)$ as the prior for the second update. Then the formula for our new posterior distribution is $p(\theta|\boldsymbol{X}_2) \propto L(\theta|\boldsymbol{X}_2)\, p(\theta|\boldsymbol{X}_1)$. Interestingly, this is the same final posterior distribution we would get if $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ arrived together and we created a posterior distribution using the combined data set.

Summarizing Posterior Results

The results from a Bayesian model are presented as an informative summary of the posterior distribution. This is typically done with distributional qualities and probability statements. We can give the probability that the parameter of interest is greater or less than zero or any other substantively relevant constant or that it occupies some region of support. Such statements are a function of posterior mean and variance, which means they represent the central tendency and uncertainty about it.

Bayesian posterior distributions can easily be described with a mean and standard deviation that correspond to the coefficient point-estimate and standard error, respectively, in conventional models. Bayesian models have a superior analogue to the confidence interval called credible interval. These are constructed in exactly the same way as the confidence interval: a point estimate plus/minus some critical value times the standard error. It has the intuitive interpretation that is often mistakenly given to confidence intervals: the probability that some effect exists between these two bounds (Gill, 1999).

Bayesian modelers can also test hypotheses. They are typically less concerned with rejecting a null model, though it is common to evaluate estimated regression effects relative to the zero point. More often they seek to compare two or more possible plausible models (Raftery, 1995). This is typically done through the use of the Bayes Factor (Kass & Raftery, 1995) or the Deviance Information Criterion (Spiegelhalter, Best, Carlin, & Van Der Linde, 2002) which is a Bayesian oriented version of the Akaike Information Criterion (Akaike, 1976) designed to balance model fit and covariate parsimony.

Computing Results

Integration of a joint probability statement of many dimensions is required to create a regression table that describes each coefficient effect in marginal (individual) terms. This can be exceedingly hard, if not impossible, to do with pen-and-paper calculations and even standard mode-finding software for complicated models. This is why Evans (1994) referred to mid-20th-century Bayesians as “unmarried marriage counselors” who could tell others how to do inference but could not do it themselves. This changed with the introduction of Bayesian stochastic simulation into the general statistical literature by Gelfand and Smith (1990).

This technique is called Markov Chain Monte Carlo (MCMC). It entails having a Markov chain wander through the sample space preferring areas in proportion to the underlying posterior probabilities. A useful analogy is to think of the Markov chain as a robot and our posterior distribution as a geographic region that we are interested in exploring, say Denali National Park in Alaska. Our robot will walk through the park, reporting the elevation values of where it visits. The walk is random in the sense that its every move is conditional only on where the robot is at the current state and the desirability of its next proposed position based on elevation. The robot has no memory of its previous path, only an algorithm that dictates its next step. This is referred to as the Markovian aspect of the chain. The algorithm attempts to balance the amount of time the robot spends exploring the high elevation areas it discovers with the amount of time it spends looking for high elevation areas. Conversely, a maximum likelihood estimation robot would go to the highest elevation in Denali National Park, report the elevation, and never leave.

Each step of the chain is a multidimensional position. In terms of our analogy, at each step the robot reports not just the height but the vegetation, wind speed, moisture, temperature, and so forth. It explores the elevated areas in each of these dimensions. Marginalizing the joint posterior is looking at the history of each dimension individually. Each gives us a row of the regression table, which is just a summary of a particular coefficient estimate. Since MCMC if used correctly is thorough in its exploration of the distribution, and not designed to just locate the highest peak like the maximum likelihood algorithm, it gives us a far more comprehensive profile of the distribution, and is less likely to get stalled at a local maximum point. However, this typically requires more computational power, and the researcher has to ensure that the chain has achieved convergence, which means it has located the high density areas, and has had enough time to explore them. With the proliferation of fast computers and diagnostic tools, these tasks are relatively easy. For a step-by-step guide to this process with required code, and a deeper dive into these algorithms, see Gill and Witko (2013).

Advantages Over Frequentist Inference

The first advantage was mentioned in the beginning of the section. The results we get from frequentist inference describe the probability of observing the data given the parameter: $p(\boldsymbol{X}|\theta)$. This is the appropriate assumption if we are dealing with data that is independent and identically distributed (IID) and arrives constantly. This is the type of data we would expect from a physics experiment or factory production process. The data are independently produced and constantly incoming, and our job is to determine the underlying parameter, such as a cosmic constant or quality measure. However, this assumption is harder to justify

for social science data, the production of which usually does not resemble a telemetric or conveyor belt process. Social scientists generally do not get data as a constant stream. Instead, they may be given a cross-national time series data set from the years 1947–2018. In this case, it is useful to think of the data as a fixed quantity and the job of the researcher being to determine the parameters given the data.

The assumption we make regarding the parameters is arguably more important. Physicists and factory engineers can reasonably assume the existence of underlying natural or mechanical constants that they are tasked with accurately measuring. But assuming that there are similar constants underlying social phenomena is riskier and has not been borne out by decades of research. We know that social phenomena are characterized by a stronger degree of contingency and interdependence. This not does mean that it is pointless to measure parameters in social science but that it is best not to think of them as fixed quantities. The Bayesian approach to statistical inference, which gives us the probability of identifying the parameter given the data: $p(\theta|\boldsymbol{X})$, is the best conceptual framework for social science data. For this reason, Bayesian results such as the credible interval have a more intuitive and straightforward interpretation than frequentist analogues.

The second advantage is that other approaches either struggle to provide estimates or simply cannot provide estimates for settings with realistically large and complex data due to high dimensionality, complex functional forms, or identifiability, but MCMC approaches work exceptionally well in such situations. We have seen how MCMC exploration of the posterior distribution provides far more information about the parameters than likelihoodist and frequentist approaches. The empirical description of the distribution not only allows for the estimation (marginalization) of previously inestimable models but also gives additional inferential information about posterior parameters of interest that can be used for model checking, model comparison, enhanced specifications, and other purposes. For a detailed explanation of these procedures with examples, see Gill and Heuberger (2019).

The third advantage is that Bayesian inference allows for the explicit incorporation of previous estimates of the parameters through the prior distribution. Systemically building on previous knowledge is arguably the essence of the scientific method. But we acknowledge that informed priors are not always available or desirable. In such cases, we can use an appropriate uniform prior, which would produce a posterior equal to the maximum likelihood estimate (which is asymptotically equal to the posterior given any prior). By doing this, we will get all the other advantages of the Bayesian approach. Moreover, even in cases where informed priors are unavailable, priors can be used to test sensitivity to produce insight on how the effect changes with varying assumptions. For example, we can show that posteriors from two dramatically different priors are consistent, to emphasize the overwhelming influence of the data. Or conversely that such posteriors are significantly different, which would emphasize the relatively weak influence of the data.

Bayesian Qualitative Analysis

We noted how priors in Bayesian inference can be used to incorporate qualitative information about the parameters. Bayes' theorem can also be used more fully and directly for qualitative analysis. There is an active and growing community of researchers that use it for this purpose (Beach & Pedersen, 2019; Bennett, 2014; Fairfield & Charman, 2019; Humphreys & Jacobs,

2015; Zaks, 2017). The rationale for this is similar to that of standard empirical Bayesian statistical inference: it encourages the explicit statement of beliefs and assumptions, and allows for systematized incorporation of previous knowledge. This framework is a useful tool for qualitative analysts concerned about transparency. The focus here on the explicit incorporation of beliefs relates to an important debate about the formalization of knowledge and assumptions in social science (Evans, Handley, Over, & Perham, 2002; Gill & Walker, 2005; McKenzie, 1994; Steffey, 1992). This approach holds the promise of providing a transparent, rigorous, and systematic form of qualitative research. In this section, we derive the odds ratio version of the theorem, which is often used for qualitative analysis, and explain its use.

Odds Ratio

Some advocates of Bayesian qualitative analysis have encouraged the use of the odds ratio form of Bayes' theorem (Bennett, 2014; Fairfield & Charman, 2019). Odds are an alternative way of making statements about probability. They are the ratio of something happening (e.g., Ann going to the party) to something not happening (e.g., Ann not going). In contrast, probability is the ratio of something happening (e.g., Ann going) to everything that could happen (e.g., Ann going and not going). This relationship is given by $odds = p/(1-p)$ where p is the probability of the event happening. In other words, the odds of Ann going is equal to the probability of Ann going divided by the probability of Ann not going. Using this, we can derive the odds ratio form of the Bayes theorem from the form discussed in the previous section:

$$p\left(A|B\right) = \frac{p\left(B|A\right)p\left(A\right)}{p\left(B\right)}, \tag{8}$$

and therefore:

$$\begin{aligned} \frac{p(A|B)}{p(\neg A|B)} &= \frac{\frac{p(B|A)p(A)}{p(B)}}{\frac{p(B|\neg A)p(\neg A)}{p(B)}} \\ &= \frac{p(B|A)p(A)}{p(B|\neg)p(\neg A)} \\ &= \frac{p(B|A)}{p(B|\neg A)} \bullet \frac{p(A)}{p(\neg A)}. \end{aligned} \tag{9}$$

Rearranging this last line gives us:

$$\textbf{Posterior odds} = \textbf{Relative Likelihoods} \times \textbf{Prior odds}, \tag{10}$$

which now looks more like the original expression of Bayes' theorem.

In the context of Bayesian qualitative analysis, the prior odds represent ratio of the probabilities that our hypothesis is true and not true. The relative likelihoods represent the ratio of the probabilities of observing some evidence if our hypothesis is true and of observing the same piece of evidence if the hypothesis is untrue. These true and untrue hypotheses can be seen as alternative hypotheses if they are mutually exclusive and exhaustive. In such a case, we can convert the posterior odds into the posterior probability for the true hypothesis by normalizing the odds. This version of Bayes' theorem allows us to incorporate probabilities that can be understood and estimated relatively easily. For this reason, it has found use in medical research and diagnosis (Trafimow, 2015). Moreover, in contrast to the original Bayes' theorem, it allows us to explicitly incorporate opposing hypotheses as noted. For a detailed guide to this approach in social science, including its use in process tracing, see Bennett (2014), Humphreys and Jacobs (2015), Zaks (2017), Beach and Pedersen (2019), and Fairfield and Charman (2019).

Bayesian Decision Making

Statistical decision making (decision theory) generalizes hypothesis testing to include the cost of alternatives, and in particular the cost of making the wrong decision. It starts with defining a real-valued loss function in $-\infty : \infty$ or $-\infty : 0$ that gives a penalty for picking one possibly wrong alternative over the other alternatives. These alternatives can be hypothesis test conclusions or just decisions in general like purchasing a stock over a range of alternative stocks in a financial decision. This requires a stated decision rule that stipulates how losses are assumed to occur, such as minimizing the maximum possible loss or minimizing squared distance from the correct estimate. The efficacy of a decision rule is the risk for each viable decision possible (Leamer, 1979; Shao, 1989).

Given data $\mathbf{X}$ conditional on the parameter θ through $p\left(\boldsymbol{X}|\theta\right)$, a function of $\mathbf{X}$ is the decision rule, defined by the set D that provides a single decision (sometimes called an action):

$$d\left(\mathbf{X}\right) = A,\ \ A \in A \tag{11}$$

where A is the allowable class of actions for this problem. Here each possible estimate of θ or range of θ is mapped to a specific action A. A decision rule leads to a loss function where $\mathbf{X}$ is observed and our decision rule is $d(\mathbf{X})$. The loss function incorporates these as inputs:

$$\mathit{L}(\mathrm{A}, \mathrm{d}\left(\mathbf{X}\right)), \tag{12}$$

and maps the decision, A from the decision rule $d(\mathbf{X})$, to a specified penalty (DeGroot, 1970). As the language implies, smaller losses are preferred.

Specific losses are mapped to specific actions, but it may be more useful to have a single summary statistic that gives an overall quality measure for a decision rule. If $p(\theta|\boldsymbol{X}, A)$ is the posterior distribution for θ produced from data $\mathbf{X}$ and action A, the Bayesian expected loss (posterior expected loss) over the full set of possible actions is given by:

$$E_{\pi}\left[\mathbb{L}\left(A, d\left(\boldsymbol{X}\right)\right)\right] = \int_{\Theta} \mathbb{L}\left(A, d\left(\boldsymbol{X}\right)\right) dF_{\pi}\left(\theta\right) \tag{13}$$

This clearly averages over the loss in θ conditional on an observed $\mathbf{X}$ through the integration giving the average loss resulting from making one of the defined decisions given by the decision function. Notice that this is a theoretically driven Bayesian definition of loss since it is an average over the uncertainty in the posterior distribution of θ not over the already observed $\mathbf{X}$. Some specific loss functions are given in Gill (2014):

- squared error loss: $\mathbb{L}\left(A, d\left(\mathbf{X}\right)\right) = \left(\theta - \theta_A\right)^2$.
- absolute error loss: $\mathbb{L}\left(A, d\left(\mathbf{X}\right)\right) = |\theta - \theta_A|$.
- $0-1$ loss for discrete variables: $\mathbb{L}\left(A, d\left(\mathbf{X}\right)\right) = 0$ if $\theta = \theta_A$, and 1 otherwise.
- interval loss: $\mathbb{L}\left(A, d\left(\mathbf{X}\right)\right) = 0$ if $CI_{1-\propto}\left[\theta_A\right]$ covers θ, and 1 otherwise.

We often want an estimator from a decision rule that minimizes (13) for every possible sample $\mathbf{X}$, which is called optimal. Limiting the choice of decision rule to the optimal one is called a Bayes Rule, and is given by:

$$\widehat{R}_B\left(\theta,\ d\left(\mathbf{X}\right)\right) = \inf_{\theta_A} \int_{\Theta} R_F\left(A,\ d\left(\mathbf{X}\right)\right) dF_{\pi}\left(\theta\right). \tag{14}$$

It is common to set the derivative of this expression equal to zero and solve for the resulting minimum value of θ_A. For squared error loss and mean estimation the Bayes rule estimate is the posterior mean, $E_{f(\theta|x)}\left[\theta\right]$, and for absolute error loss, $\mathbb{L}\left(A, d\left(\mathbf{X}\right)\right) = |\theta - \theta_A|$, the Bayes rule estimate is the posterior median.

To see how this works in practice consider estimation of the mean, μ, from an assumed normally distributed sample $\mathbf{X}$, where the data variance, σ^2, is known. We first stipulate a conjugate normal prior: $\mathrm{N}(m, s^2)$ with assigned hyperprior values m and s. This is a classic setup that analytically produces the posterior distribution:

$$\mu|\sigma^2, \mathbf{X}\, \alpha\, \exp\left[-\frac{1}{2}\left(\frac{1}{s^2} + \frac{n}{\sigma^2}\right)\left(\mu - \frac{\left(\frac{m}{s^2} + \frac{n\bar{x}}{\sigma^2}\right)}{\left(\frac{1}{s^2} + \frac{n}{\sigma^2}\right)}\right)^2\right]. \tag{15}$$

The corresponding Bayesian expected loss is:

$$R_B\left(\theta, d\left(X\right)\right) = \int_{\Theta} R_F\left(A, d\left(\mathbf{X}\right)\right) dF_{\pi}\left(\theta\right)$$
$$= \int_{\mu} \left(\mu - \frac{n\overline{x} + ms}{n+s}\right)^2 \left(2\pi\frac{\sigma^2}{n+s}\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2/(n)}\left(\mu - \frac{n\overline{x}+m}{n}\right)^2\right] d\mu. \tag{16}$$

Making the change of variable according to $\delta_{\mu} = \mu - \widehat{\mu} = \mu - (n\overline{x} + m)/(n+)$ (Gill, 2014), we get the simplified form where:

$$R_B\left(\theta, d\left(\mathbf{X}\right)\right) = E\left[{\delta^2}_{\mu}\right], \; = \; Var[\delta_{\mu}] \; + \left(E[\delta_{\mu}]\right)^2 \; = \; Var[\mu] - 0 = \frac{\sigma^2}{n.} \tag{17}$$

Here the variance of δ_{μ} equals the variance μ by assumption since $\widehat{\mu}$ consists of all constants, so $E[\delta_{\mu}] = 0$.

If we know σ^2 then this reduces to a very simple calculation for the Bayesian expected loss.

Bayesian Machine Learning

The Bayesian concepts and procedures discussed in this article have been applied in various applications of statistics: econometrics, psychometrics, biostatistics, and machine learning. In all these cases, Bayesian tools provide the advantages discussed in the section on statistical inference. Machine learning is a subfield of computer science but its techniques are increasingly borrowed by social scientists (Cranmer & Desmarais, 2017). It uses algorithms and statistical models to learn from data, usually with the objective of prediction. Unsurprisingly, Bayesian tools are useful in this regard.

A simple and popular machine learning algorithm is called the Naive Bayes Classifier. It is a relatively straightforward application of Bayes' theorem. Consider the problem of predicting civil war. Imagine we have the following data in the form of dichotomous variables for a group of countries: did it experience civil war (y), is it a poor country (x_1), does it have difficult terrain (x_2), does it have a youth bulge (x_3), does it have strong institutions (x_4), and does it have democracy (x_5). This is an illustrative simplification of predictive models in the civil war literature. According to this example, Bayes' theorem can be written as:

$$P\left(y|\boldsymbol{X}\right) = \frac{P\left(\boldsymbol{X}|y\right)P\left(y\right)}{P\left(\boldsymbol{X}\right)}$$

where $\boldsymbol{X}$ is given as

$$\boldsymbol{X} = (x_1, x_2 x_3, x_4, x_5)\,.$$

Substituting for $\mathbf{X}$ and expanding using the chain rule gives:

$$P(y|x_1, x_2 x_3, x_4, x_5) = \frac{P(x_1|y)\,P(x_2|y)\,P(x_3|y)\,P(x_4|y)\,P(x_5|y)\,P(y)}{P(\boldsymbol{X})}.$$

Following this step, all the terms in the numerator can be acquired easily from the data set, and the denominator can be considered to be "1" or else ignored as it is a constant. The equation can be written as the following proportionality:

$$P(y|x_1, x_2 x_3, x_4, x_5) \propto P(x_1|y)\,P(x_2|y)\,P(x_3|y)\,P(x_4|y)\,P(x_5|y)\,P(y)\,.$$

After we have the conditional probabilities, we can make predictions about whether a new country with a given $\mathbf{X}$ will experience civil war. There are two important assumptions here: First, that the predictors are independent, which is to say that, for example, the strength of a country's institutions are not determined by it being a democracy. This is generally a questionable assumption for social science data. Second, that all predictors have an equal effect on the outcome, which is also a questionable assumption. However, the Naive Bayes Classifier has been successfully used in applications such as spam detection and document classification. It performs well with categorical input variables; it can be used for continuous or count variables provided we make the appropriate distributional assumptions and use the appropriate formulas for conditional probability. Here, it serves as an illustration of the use of Bayes' theorem in machine learning.

Bayesian Networks for Decision Making

In this section decision making with Bayesian networks is briefly introduced as a machine learning tool. In its most basic form a Bayesian network is a *directed acyclic graph* (DAG), which maps relationships between variables as conditional probabilities. (Pearl, Glymour, & Jewell, 2016). These conditional probabilities represent assumed causal relationships. The direction of arrows in the DAG represent the direction of causality. Generally, a Bayesian network is specified by an expert with knowledge of the data generating process. It is then used to infer specific relationships in the network and update them in light of new information. It can even be used to specify the form of the network when the task is too complex for humans, which is an approach in machine learning.

More formally, a Bayesian network is defined by $B = \{N, A, \Theta\}$, where $N = \{n_1, n_2, \ldots\}$ is a set of domain variables of interest (e.g., data set provided values), $A = \{a_1, a_2, \ldots\}$ are probabilistic edges between nodes, and $\Theta = \{\theta_1, \theta_2, \ldots\}$ are the nodes of interest. This setup means that a Bayesian network can be configured to give the conditional probability of a specified node given known or assigned values of the other nodes. Often this is the probability

that the node of interest belongs to one of a set of alternative groups making the Bayesian network a classifier in the machine learning sense highlighting the relationships between data set attributes.

In order to produce learning in a Bayesian network we need to first learn the graphical structure of the network and then learn (estimate) the parameters of this network. If we know or can assume the structure of the network then the work is very straightforward. The simplest approach in this regard is the Naive Bayes network, which posits a single parental node λ and a specified number of descendant nodes $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\}$ as shown in the example in Figure 1.

This is not a recent approach and a lot of work has been done to improve and extend this basic structure. One appealing feature is that the structure is set in advance, meaning that only the parameter

Figure 1. Example of a Naive Bayes network.

estimation part is necessary. The main cost, however, is the accompanying assumption that the descendant nodes are independent of each other, although this can be slightly violated without a serious cost. Like linear regression this is an old warhorse that remains useful and is robust to violations of assumptions. There are four basic types of causal relationships represented by conditional probabilities in this context: (a) A directly causes $B : p(B|A)$, (b) A causes C through $B : p(B|A)$ and $p(C|B)$, (c) A is the common cause of B and $C : p(B|A)$ and $p(C|A)$, (d) A and B commonly effect $C : p(C|A, B)$. Consider the following as an example of a Bayesian network: Amy going to the party influences the probability of Bob going: $p(B|A)$. The probability of Bob going influences the probabilities of Chen and Dan going: $p(C|B)$ and $p(D|B)$. The probabilities of Chen and Dan going influences the probability of Eve going: $p(E|C, D)$. Suppose we have all these probabilities from past experience. In this example, the individuals are represented by nodes and their influence on each other is represented by probabilistic edges. If we consider only the relationship between Bob, Chen, and Dan, it would be the most basic form of the Naive Bayes network.

The focus in many of the uses of Bayesian networks is on causal inference as implied by the estimation of relationships. How do we find the probability that Amy, Chen, and Eve went to the party and that Bob and Dan did not? We calculate the joint probability: $p(A, \neg B, C, \neg D, E)$. We do this by repeatedly applying the definition of conditional probability (Lewicki, 2007):

$$
\begin{aligned}
p\left(A, \neg B, C, \neg D, E\right) &= p\left(E|A, \neg B, C, \neg D\right) p\left(A, \neg B, C, \neg D\right) \\
&= p\left(E|C, \neg D, A, \neg B\right) p\left(A, \neg B, C, \neg D\right) \\
&= p\left(E|C, \neg D\right) p\left(A, \neg B, C, \neg D\right) \\
&= p\left(E|C, \neg D\right) p\left(\neg D|A, \neg B, C\right) p\left(A \neg B, C\right) \\
&= p\left(E|C, \neg D\right) p\left(\neg D|\neg B\right) p\left(A, \neg B, C\right) \\
&= p\left(E|C, \neg D\right) p\left(\neg D|\neg B\right) p(C|A, \neg B) \left(A, \neg B\right) \\
&= p\left(E|C, \neg D\right) p\left(\neg D|\neg B\right) p(C|\neg B) \left(A, \neg B\right) \\
&= p\left(E|C, \neg D\right) p\left(\neg D|\neg B\right) p\left(C|\neg B\right) \left(\neg B|A\right) p\left(A\right).
\end{aligned}
$$

By specifying all the conditional probabilities, we have also specified the joint probability. This means we can calculate any joint probability in principle. Moreover, with the joint probability, we can calculate any conditional probability using Bayes' theorem. These procedures are very useful but, as is apparent, can get very mathematically and computationally expensive with increasing size and complexity of the model. For this reason, oftentimes the MCMC approaches discussed are used to integrate joint probability distributions. For a detailed treatment of Bayesian networks in social science, see Whitney, White, Walsh, Dalton, and Brothers (2011).

Recent Bayesian Work in Political Science

Bayesian work is increasingly popular in empirical political science because it helps solve specific data challenges in a more direct and principled way than do competing approaches.

Over the last few decades, any sense of controversy or skepticism has been swept aside. With a wide range of available computational tools and approaches, estimation challenges are now manageable, even under the most difficult data circumstances.

Hollenbach, Montgomery, and Crespo-Tenorio (2019) are concerned with estimating causal effects in instrumental variable models with a dichotomous treatment and a dichotomous outcome. Here non-Bayesian models lead to all kinds of difficulties such as poor parameter estimates and no measures of uncertainty in the estimated main causal effects. Conversely, a Bayesian approach directly provides posterior distributions, which fully describe the effects and their uncertainty. Bisbee (2019) gives a modified version of multilevel regression and post-stratification (MRP) built on the nonparametric regularization method Bayesian additive regression trees (BART). The application is to public opinion at different levels of geographic aggregation. In this case a more directly Bayesian version of MRP provides better estimates of political quantities of interest. In political science data, it is not uncommon to have right-censored outcomes classified as failures because of measurement errors. This is a serious problem that leads to biased estimators in model estimation. Bagozzi, Joo, Kim, and Mukherjee (2019) deal with the problem with a modified Bayesian survival model. Their Bayesian split population survival estimator gives two simultaneous equations such that outcome misclassification and survival are both accounted for. Their approach relies heavily on Bayesian computational methods. Another data problem that political scientists deal with is the estimation of latent variables such as the ideological positions of candidates and parties. Standard spatial econometric approaches often fail here, so Juhl (2019) builds on the political

science literature for Bayesian dynamic item response models to more accurately estimate parties' ideological positions, spatial interactions, and all of the required posterior uncertainty.

An area that deserves special discussion is the analysis of text as data in political science. The last decade has seen an immense amount of attention to political texts that are produced by governments, parties, and politicians. This is an area that non-Bayesian modeling struggled to address since more traditional methods such as the EM algorithm assume that estimates are known with certainty. Grimmer (2010) shows that Bayesian models estimated with variational approximation methods provide the entire posterior distribution for U.S. senators' policy agendas from recorded text. In a follow-on work, Grimmer and Stewart (2013) use a Naive Bayes unsupervised classifier to learn the relationship between words and functional political categories. Barbera (2015) considers text that come from politicians posting on Twitter to understand clustering in the network sense. His Bayesian spatial model treats ideology as a latent variable such that a large sample of both political elite and public Twitter users' policy preferences in the United States and five European countries can be estimated. The scope of the data analysis is vast and is made available in realistic time with Bayesian stochastic simulations tools such as a Hamiltonian Monte Carlo sampling algorithm and Metropolis-Hasting MCMC procedures. Another active area in political science is the application of Bayesian Latent Dirichlet Allocation (LDA) models to text analysis. For example, Isoaho, Gritsenko, and Mäkelä notice that regular LDA for topic models approaches assumes that topics appear in political documents independently of each other, and this is typically not true as governments, parties, and candidates respond to each other's communications. Their review piece discusses alternatives and extensions.

Future Bayesian work in political science will likely continue to exploit rapid increases in computational ability, both in hardware and algorithms. Compute-intensive work includes Dirichlet process prior models, big data (however we choose to define that), more demanding text analysis requirements, and high-dimensional latent variable problems. The probabilistic basis of all Bayesian inference and the associated computational tools means that this will continue to be a very dynamic methodological paradigm in political science.

Conclusion

This article introduced Bayes' theorem from first principles of probability. This remarkably simple expression belies a deep principle for updating knowledge that underlies Bayesian statistical inference and decision making. The article then described how it fits into an inference structure that processes prior information to posterior information by conditioning on relevant data when they come in. This led to a presentation of qualitative Bayesian decision making that incorporates human judgements. Finally, Bayesian decision analysis was described from a technical standpoint following a discussion of prominent Bayesian techniques in machine learning. The purpose is to generate more interest in Bayesian methods, motivating readers to continue exploring these key ideas and their modern use in the 21st century.

The Reverend Thomas Bayes could not have imagined all the modern uses of his theorem. It is used for such varied tasks as detecting spam emails, professional sports betting, weather forecasting, predicting armed conflict, statistical inference, and beating humans at games

(Silver, 2012). In this article, we have discussed, accessibly but fairly rigorously, the basics of its most salient applications in social science. Inevitably, there are theoretical challenges and limitations to Bayes' theorem, most notably the difficulty of incorporating non-probabilistic knowledge including causality (Pearl, 2001). However, we have obviously not discovered all mathematical principles, which are nothing but formalized general insights about reality. If Bayes, an 18th-century reverend, could uncover such a useful insight with a little creative thinking, it is possible that researchers today, perhaps inspired by Bayes' theorem, will do the same, thereby enhancing our grasp of the natural and social world.

References

Akaike, H. (1976). Canonical correlation analysis of time series and the use of an information criterion. In *Mathematics in science and engineering* (Vol. 126, pp. 27–96). Elsevier.

Bagozzi, B. E., Joo, M. M., Kim, B., & Mukherjee, B. (2019). A Bayesian split population survival model for duration data with misclassified failure events. *Political Analysis*, *27*(4), 415–434.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76–91.

Bayes, T. (1763). LII: An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Beach, D., & Pedersen, R. B. (2019). *Process-tracing methods: Foundations and guidelines*. Ann Arbor: University of Michigan Press.

Bennett, A. (2014). Process tracing with Bayes: Moving beyond the criteria of necessity and sufficiency. *Qualitative and Multimethod Research*, *12*(1), 46–51.

Bisbee, J. (2019). BARP: Improving Mister P using Bayesian additive regression trees. *American Political Science Review*, *113*(4), 1060–1065.

Cranmer, S. J., & Desmarais, B. A. (2017). What can we learn from predictive modeling? *Political Analysis*, *25*(2), 145–166.

DeGroot, M. H. (1970). *Optimal statistical decisions*. New York, NY: McGraw-Hill.

Diaconis, P., & Skyrms, B. (2019). *Ten great ideas about chance*. Princeton, NJ: Princeton University Press.

Evans, J. St. B. T., Handley S. J., Over, D. E., & Perham, N. (2002). Background beliefs in Bayesian inference. *Memory & Cognition*, *30*(2), 179–190.

Evans, S. (1994). Discussion of the paper by Spiegelhalter, Freedman and Parmar. *Journal of the Royal Statistical Society-Series*, *A 157*.

Fairfield, T., & Charman, A. (2019). A dialogue with the data: The Bayesian foundations of iterative research in qualitative social science. *Perspectives on Politics*, *17*(1), 154–167.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, *22*(5), 700–725.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409.

Gill, J. (1999). The insignificance of null hypothesis significance testing. *Political Research Quarterly*, *52*(3), 647–674.

Gill, J. (2014). *Bayesian methods for the social and behavioral sciences*. Boca Raton, FL: CRC.

Gill, J., & Heuberger, S. (2019). Bayesian modeling and inference: A postmodern perspective. In L. C. Curini, J. Franzese, & J. Robert (Eds.), *Handbook of Research Methods in Political Science & International Relations*. SAGE.

Gill, J., & Torres, M. (2019). *Generalized linear models: A unified approach* (2nd ed.). Los Angeles, CA: SAGE.

Gill, J., & Walker, L. D. (2005). Elicited priors for Bayesian model specifications in political science research. *The Journal of Politics*, *67*(3), 841–872.

Gill, J., & Witko, C. (2013). Bayesian analytical methods: A methodological prescription for public administration. *Journal of Public Administration Research and Theory*, *23*(2), 457–494.

Grimmer, J. (2010). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*, 1–35.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Hollenbach, F. M., Montgomery, J. M., & Crespo-Tenorio, A. (2019). Bayesian versus maximum likelihood estimation of treatment effects in bivariate probit instrumental variable models. *Political Science Research and Methods*, *7*(3), 651–659.

Humphreys, M., & Jacobs, A. M. (2015). Mixing methods: A Bayesian approach. *American Political Science Review*, *109*(4), 653–673.

Isoaho, K., Gritsenko, D., & Mäkelä, E. (2019). Topic modeling and text analysis for qualitative policy research. *Policy Studies Journal*.

Juhl, S. (2019). Measurement uncertainty in spatial models: A Bayesian dynamic measurement model. *Political Analysis*, *27*(3), 302–319.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par le évènemens. *Mémoires de Mathématque et Physique, Presentés à l'Académie Royale des Sciences, par divers Savans & lûs dans ses Assemblées, Tome Sixiéme*, *66*, 621–656.

Laplace, P. S. (1781). Mémoire sur la probabilités. *Mémoires de l'Académie Royale des Sciences de Paris*, *1778*, 227–332.

Leamer, E. E. (1979). Information criteria for choice of regression models: A comment. *Econometrica*, *47*, 507–510.

Lewicki, M. S. (2007). *Bayesian Networks <https://www.cs.cmu.edu/afs/cs/academic/class/15381-s07/www/slides/032707bayesNets1.pdf>*.

McKenzie, C. R. M. (1994). The accuracy of intuitive judgment strategies: Covariation assessment and Bayesian inference. *Cognitive Psychology*, *26*(3), 209–239.

Neapolitan, R., & Jiang, X. (2016). The Bayesian network story *<https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199607617.001.0001/oxfordhb-9780199607617-e-31>*. In A. Hájek & C. Hitchcock (Eds.), *The Oxford handbook of probability and philosophy*.

Pearl, J. (2001). Bayesianism and causality, or, why I am only a half-Bayesian. In D. Corfield & J. Williamson (Eds.), *Foundations of Bayesianism*. *Applied Logic Series* (Vol 24, pp. 19–36). Dordrecht: Springer.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. Chichester, UK: John Wiley & Sons.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–164.

Shao, J. (1989). Monte Carlo approximations in Bayesian decision theory. *Journal of the American Statistical Association*, *84*, 727–732.

Silver, N. (2012). *The signal and the noise: Why so many predictions fail—But some don't*. New York, NY: Penguin.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 583–639.

Steffey, D. (1992). Hierarchical Bayesian modeling with elicited prior information. *Communications in Statistics-Theory and Methods*, *21*(3), 799–821.

Trafimow, D. (2015). The benefits of applying Bayes' theorem in medicine. *American Research Journal of Humanities and Social Sciences*, *1*, 14–23.

Whitney, P., White, A., Walsh, S., Dalton, A., & Brothers, A. (2011). Bayesian networks for social modeling. In J. Salerno, S. J. Yang, D. Nau, & S. K. Chai (Eds.), *Social Computing, behavioral-cultural modeling and prediction. SBP 2011. Lecture Notes in computer science* (Vol. 6589, pp. 227–235). Berlin, Heidelberg: Springer.

Zaks, S. (2017). Relationships among rivals (RAR): A framework for analyzing contending hypotheses in process tracing. *Political Analysis*, *25*(3), 344–362.

Related Articles

Process-Tracing Methods in Social Science

Information Aggregation in Political Decision Making

Dynamic Process Tracing Methods in the Study of Political Decision Making

Motivated Reasoning and Political Decision Making