Introduction to Bayesian Statistics

John P. Burkett
burkett@uri.edu

September 9, 2009

Outline

Defining characteristics

Bayesian statistics could be defined as an approach to data analysis that focuses on conditional probabilities relating observed and unknown quantities. This approach is called Bayesian because some of its basic elements were first suggested by Thomas Bayes (c. 1702–1761).

Probabilities and their interpretation

Today, most statisticians, including Bayesians, regard probabilities as quantities that conform to a set of mathematical axioms such as those proposed in the 1930s by Kolmogorov. However, Bayesians, unlike frequentists, do not confine the application of probability theory to experiments that could be repeated indefinitely often under identical conditions. Bayesians argue that the axioms of probability theory also apply to rational degrees of belief concerning uncertain propositions including scientific hypothesis. Thus Bayesians can use probability distributions to characterize their beliefs about possible values for physical constants conditional on the available evidence. As evidence accumulates, these probability distributions may change. Bayesian statistics provides coherent means for updating probability distributions as new evidence arrives.

Steps in Bayesian analysis

Bayesian analysis of a problem usually involves at least two steps:

1. Specifying a "full probability model—joint probability distribution for all the observable and unobservable quantities in" the problem.
2. "Calculating and interpreting . . . the conditional probability distribution for the unobserved quantities of ultimate interest, given the observed data" (Gelman et al. 2004, p. 3)

After these steps, good Bayesians, like good statisticians of any persuasion, assess the fit of the model to the data and the reasonableness of the conclusions and then, if need be, respecify the model and redo the calculations.

Origins of Bayesian statistics

Bayes (1764) derived a conditional probability $p(\theta_1 < \theta < \theta_2 | k)$, where θ is the unobserved parameter of a binomial distribution, θ_1 and θ_2 are arbitrary bounds such that $0 \leq \theta_1 < \theta_2 \leq 1$, k is the observed number of successes in n trials, and the prior distribution on θ is uniform on $[0, 1]$.

Laplace (1774–1812) independently derived and generalized Bayes's results, applying them to astronomy and demographics.

de Morgan (1838) distinguished between "direct" and "inverse" probability. The former is a probability of an observable event conditional on a hypothesis and the latter (as developed by Bayes and Laplace) is a probability of the hypothesis conditional on the event.

Later development of Bayesian statistics

de Finetti (1937) analyzed bets, showing that a naive bettor may fall victim to a “Dutch book”—i.e., a combination of bets that assures loss of money. To avoid that fate, a bettor should try to make his probability assessments “coherent”—i.e., consistent with probability theory.

Cox (1961) showed that if degrees of belief, scaled to range from 0 to 1, satisfy two intuitively appealing rationality criteria they must also satisfy standard axioms of probability theory. Thus probabilities can be used to represent reasonable degrees of belief.

Rényi (1970) developed an axiomatic probability theory in which all probabilities are conditional.

Savage (1972) analyzed decision-making under uncertainty, showing that if decision makers satisfy seven rationality postulates, they will maximize expected utility, assessing subjective probabilities in a manner consistent with probability theory.

Recent developments in Bayesian computation

From sample information and prior information, Bayesians derive a posterior distribution for quantities of interest. When this distribution is hard to summarize by analytical methods, modern Bayesians often resort to sampling from it. An early and influential survey of sampling ("Monte Carlo") methods was provided by Gelfand and Smith (1990). Comprehensive recent treatments of computational issues include (in order of increasing difficulty) Albert (2009), Hoff (2009), and Carlin and Louis (2009). Many sampling methods have been implemented in R or specialized software such as BUG and JAGS.

Classification of Monte Carlo methods

Carlin and Louis (2009) organize sampling methods as follows:

- Noniterative methods
  - Direct
  - Indirect (importance sampling, rejection sampling, weighted bootstrap)
- Iterative (Markov chain) methods
  - Gibbs sampler
  - Metropolis-Hastings algorithm
  - Langevin-Hastings algorithm
  - Slice sampler

Outcomes, events, and probabilities I

The result of an experiment or process, described in as much detail as we need, is called an *outcome* (or in Lee's terminology, a *elementary event*). A set of outcomes is called an *event*. In the case of rolling a six-sided die, for example, the set $\{1, 3, 5\}$ is the event that the outcome is an odd number. An event is said to occur when any of its constituent outcomes occurs. Thus we say an odd number is rolled when 1, 3, or 5 is rolled. For any two events A and B, the statement "A or B occurs" is true if and only if the outcome is in the union of A and B (denoted $A \cup B$). The statement "A and B occur" is true if and only if the outcome is in the intersection of A and B (denoted $A \cap B$ or simply AB). For example, if $A = \{1, 2\}$ and $B = \{2, 3\}$, saying "A or B occurs" is equivalent to saying "the outcome is 1, 2, or 3 " and claiming "A and B occur" is equivalent to claiming "the outcome is 2."

A probability can be viewed as a relationship between a known or hypothesized event and an uncertain one. In Lee's notation,

Outcomes, events, and probabilities II

$P(E|H)$ is the probability of event E conditional on the occurance of event H. Most Bayesians interpret $P(E|H)$ as a measure of an individual's belief in E given certainty about H.
Following Rényi (1970), Lee stipulates 4 postulates or axioms about probability:

- P1 $P(E|H) \geq 0$ for all E, H.
- P2 $P(H|H) = 1$ for all H.
- P3 $P(E \cup F|H) = P(E|H) + P(F|H)$ when $EFH = \emptyset$.
- P4 $P(E|FH)P(F|H) = P(EF|H)$.

Postulate P3 is one form of the addition law for probabilities. This particular form is called finite additivity. Sometimes probability theorists prefer to use a stronger form of the postulate, called countable additivity.

The postulate of countable additivity (P3*) asserts that

$$\text{P3*} \quad P\left(\bigcup_{n=1}^{\infty} E_n | H\right) = \sum_{n=1}^{\infty} P(E_n|H)$$

whenever $E_1, E_2, E_3, \ldots$ are exclusive events given H.

Axiom P4 is often called the multiplication law of probability.

"Events E and F are said to be *independent* given H if $P(EF|H) = P(E|H)P(F|H)$" (Lee 2004, p. 6). Noticing that the left side of this equation is the same as the right side of P4, we may deduce that independence implies that $P(E|FH)P(F|H) = P(E|H)P(F|H)$. Thus idenpendene implies that when $P(F|H) > 0$, $P(E|FH) = P(E|H)$. In other words, conditioning on FH is no different that conditioning on H alone.

Consequences of the axioms I

When all the probabilities of interest are conditional on the same event H, we may get tired of writing it. Omitting it yields as simpler notation. Thus we could rewrite P4 as

$$P(E|F)P(F) = P(EF). \tag{1}$$

The order of E and F here is arbitrary. We could equally well write

$$P(F|E)P(E) = P(EF). \tag{2}$$

Noting that the right sides of equations (1) and (2) are the same, we can equate the left sides, getting

$$P(E|F)P(F) = P(F|E)P(E) \tag{3}$$

Dividing by both sides by P(E), we get

$$P(F|E) = \frac{P(E|F)P(F)}{P(E)} \tag{4}$$

Consequences of the axioms II

Equation (4) is one simple form of Bayes's theorem.

Let (F_n) be a sequence of exclusive and exhaustive events. Then for any event E we may write

$$E = \bigcup_n (EF_n) \tag{5}$$

Based on P3* and equation 5 we can write

$$P(E) = \sum_n P(EF_n). \tag{6}$$

Based on P4 we can write

$$P(EF_n) = P(E|F_n)P(F_n) \tag{7}$$

From equations 6 and 7, we get

$$P(E) = \sum_n P(E|F_n)P(F_n) \tag{8}$$

Consequences of the axioms III

Equation 8 is called the generalized addition law or the law of the extension of the conversation.
Substituting the right side of equantion 8 for $P(E)$ in equation 4, we get a new and particularly useful form of Bayes theorem:

$$P(F_n|E) = \frac{P(E|F_n)P(F_n)}{\sum_n P(E|F_n)P(F_n)} \tag{9}$$

References I

Albert, J. (2009). *Bayesian Computation with R*. Springer, New York, second edition.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418.

Carlin, B. P. and Louis, T. A. (2009). *Bayesian Methods for Data Analysis*. CRC Press, Boca Raton, third edition.

Cox, R. T. (1961). *The Algebra of Probable Inference*. Johns Hopkins Press.

de Morgan, A. (1838). *An Essay on Probabilities and Their Application to Life Contingencies and Insurance Offices*. Longman, Orme, Brown, Green, & Longmans, London.

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

References II

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall, London, second edition.

Hoff, P. D. (2009). *A First Course in Bayesian Statistical Methods*. Springer, New York.

Lee, P. M. (2004). *Bayesian Statistics: An Introduction*. Arnold, London, third edition.

Rényi, A. (1970). *Foundations of Probability*. Holden-Day, San Francisco.

Savage, L. J. (1972). *The Foundations of Statistics*. Dover, New York.