

## Bayesian Inference

Bayesian inference is a collection of statistical methods which are based on Bayes' formula.

Statistical inference is the procedure of drawing conclusions about a population or process based on a sample. Characteristics of a population are known as *parameters*. The distinctive aspect of Bayesian inference is that both parameters and sample data are treated as random quantities, while other approaches regard the parameters non-random. An advantage of the Bayesian approach is that all inferences can be based on probability calculations, whereas non-Bayesian inference often involves subtleties and complexities. One disadvantage of the Bayesian approach is that it requires both a *likelihood function* which defines the random process that generates the data, and a *prior probability distribution* for the parameters. The prior distribution is usually based on a subjective choice, which has been a source of criticism of the Bayesian methodology. From the likelihood and the prior, Bayes' formula gives a *posterior distribution* for the parameters, and all inferences are based on this.

**Bayes' formula:** There are two interpretations of the probability of an event  $A$ , denoted  $P(A)$ : (1) the long run proportion of times that the event  $A$  occurs upon repeated sampling; (2) a subjective belief in how likely it is that the event  $A$  will occur. If  $A$  and  $B$  are two events, and  $P(B) > 0$ , then the *conditional probability of  $A$  given  $B$*  is  $P(A|B) = P(AB)/P(B)$  where  $AB$  denotes the event that both  $A$  and  $B$  occur. The frequency interpretation of  $P(A|B)$  is the long run proportion of times that  $A$  occurs when we restrict attention to outcomes where  $B$  has occurred. The subjective probability interpretation is that  $P(A|B)$  represents the updated belief of how likely it is that  $A$  will occur if we

know  $B$  has occurred. The simplest version of Bayes' formula is  $P(B|A) = P(A|B)P(B)/(P(A|B)P(B) + P(A|\sim B)P(\sim B))$  where  $\sim B$  denotes the complementary event to  $B$ , i.e. the event that  $B$  does not occur. Thus, starting with the conditional probabilities  $P(A|B)$ ,  $P(A|\sim B)$ , and the unconditional probability  $P(B)$  ( $P(\sim B) = 1 - P(B)$  by the laws of probability), we can obtain  $P(B|A)$ . Most applications require a more advanced version of Bayes' formula.

Consider the "experiment" of flipping a coin. The mathematical model for the coin flip applies to many other problems, such as survey sampling when the subjects are asked to give a "yes" or "no" response. Let  $\theta$  denote the probability of heads on a single flip, which we assume is the same for all flips. If we also assume that the flips are statistically independent given  $\theta$  (i.e. the outcome of one flip is not predictable from other flips), then the probability model for the process is determined by  $\theta$  and the number of flips. Note that  $\theta$  can be any number between 0 and 1. Let the random variable  $X$  be the number of heads in  $n$  flips. Then the probability that  $X$  takes a value  $k$  is given by

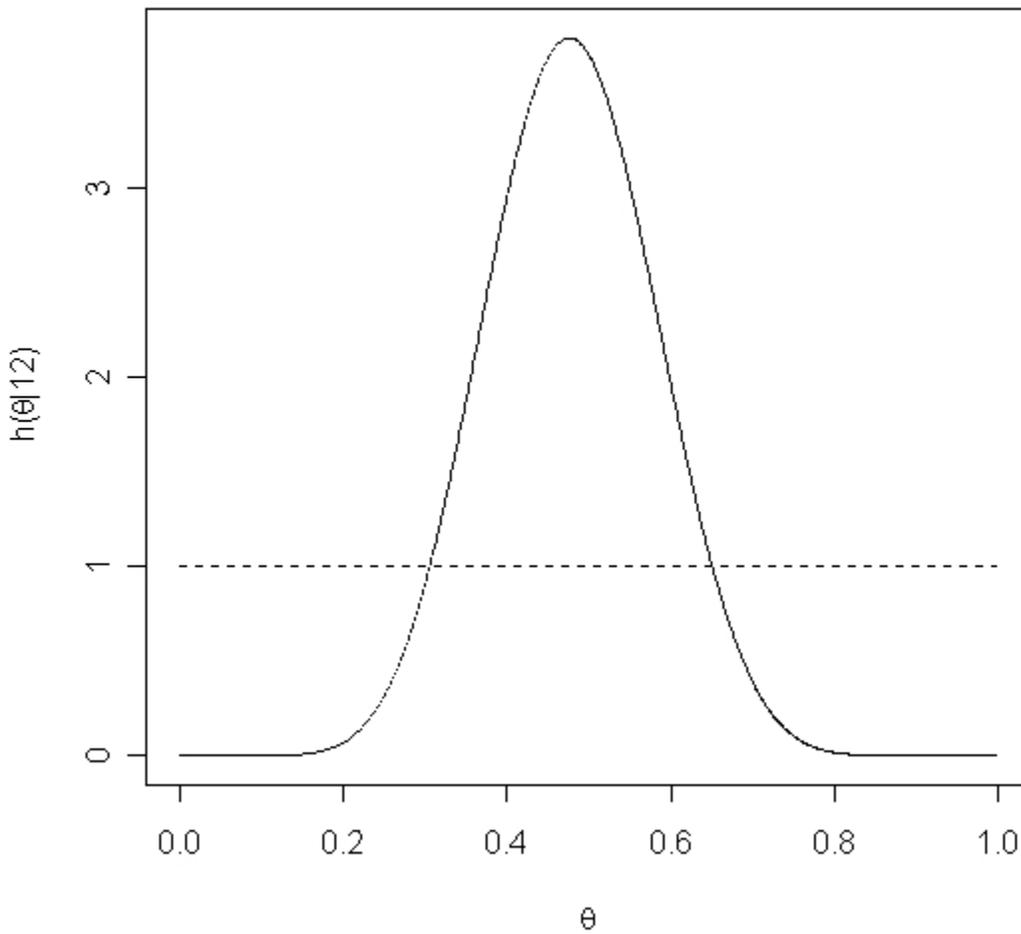
$$P(X=k|\theta) = C_{n,k} \theta^k (1-\theta)^{n-k}, \quad k = 0, 1, \dots, n.$$

$C_{n,k}$  is a binomial coefficient whose exact form is not needed. This probability distribution is called the *binomial distribution*. We will denote  $P(X=k|\theta)$  by  $f(k|\theta)$ , and when we substitute the observed number of heads for  $k$ , it gives the *likelihood function*.

To complete the Bayesian model we specify a prior distribution for the unknown parameter  $\theta$ . If we have no belief that one value of  $\theta$  is more likely than another, then a natural choice for the prior is the uniform distribution on the interval of numbers from 0 to 1. This distribution has a probability density function  $g(\theta)$  which is 1 for  $0 \leq \theta \leq 1$  and otherwise equals 0, which means that  $P(a \leq \theta \leq b) = b-a$  for  $0 \leq a < b \leq 1$ .

The posterior density of  $\theta$  given  $X=x$  is given by a version of Bayes' formula:  $h(\theta|x) = K(x)f(x|\theta)g(\theta)$ , where  $K(x)^{-1} = \int f(x|\theta)g(\theta)d\theta$  is the the area under the curve  $f(x|\theta)g(\theta)$  when  $x$  is fixed at the observed value.

A quarter was flipped  $n=25$  times and  $x=12$  heads were observed. The plot of the posterior density  $h(\theta|12)$  is shown in Figure 1. This represents our updated beliefs about  $\theta$  after observing 12 heads in 25 coin flips. For example, there is little chance that  $\theta \geq .8$ ; in fact,  $P(\theta \geq .8 | X=12) = 0.000135$ , whereas according to the prior distribution,  $P(\theta \geq .8) = 0.200000$ .



**Figure 1:** Posterior density for the heads probability  $\theta$  given 12 heads in 25 coin flips. The dotted line shows the prior density.

**Statistical Inference:** There are three general problems in statistical inference. The simplest is *point estimation*: what is our best guess for the true value of the unknown parameter  $\theta$ ? One natural approach is to select the highest point of the posterior density, which is the *posterior mode*. In this example, the posterior mode is  $\theta_{Mode} = x/n = 12/25 = 0.48$ . The posterior mode here is also the *maximum likelihood estimate*, which is the estimate most non-Bayesian statisticians would use for this problem. The maximum likelihood estimate would not be the same as the posterior mode if we had used a different prior. The generally preferred Bayesian point estimate is the *posterior mean*:  $\theta_{Mean} = \int \theta h(\theta|x) d\theta = (x+1)/(n+2) = 13/27 = 0.4815$ , almost the same as  $\theta_{Mode}$  here.

The second general problem in statistical inference is interval estimation. We would like to find two numbers  $a < b$  such that  $P(a < \theta < b | X=12)$  is large, say 0.95. Using a computer package one finds that  $P(0.30 < \theta < 0.67 | X=12) = 0.950$ . The interval  $0.30 < \theta < 0.67$  is known as a *95% credibility interval*. A non-Bayesian 95% confidence interval is  $0.28 < \theta < 0.68$ , which is very similar, but the interpretation depends on the subtle notion of “confidence.”

The third general statistical inference problem is hypothesis testing: we wish to determine if the observed data support or lend doubt to a statement about the parameter. The Bayesian approach is to calculate the posterior probability that the hypothesis is true. Depending on the value of this posterior probability, we may conclude the hypothesis is likely to be true, likely to be false, or

the result is inconclusive. In our example we may ask if our coin is biased against heads, i.e. is  $\theta < 0.50$ ? We find  $P(\theta < 0.50 | X=12) = 0.58$ . This probability is not particularly large or small, so we conclude that there is not evidence for a bias for (or against) heads.

Certain problems can arise in Bayesian hypothesis testing. For example, it is natural to ask whether the coin is fair, i.e. does  $\theta = 0.50$ ? Because  $\theta$  is a continuous random variable,  $P(\theta = 0.50 | X=12) = 0$ . One can perform an analysis using a prior that allows  $P(\theta = 0.50) > 0$ , but the conclusions will depend on the prior. A non-Bayesian approach would not reject the hypothesis  $\theta = 0.50$  since there is no evidence against it (in fact,  $\theta = 0.50$  is in the credible interval).

This coin flip example illustrates the fundamental aspects of Bayesian inference, and some of its pros and cons. L. J. Savage (1954) posited a simple set of axioms and argued that all statistical inferences should logically be Bayesian. However, most practical applications of statistics tend to be non-Bayesian. There has been more usage of Bayesian statistics since about 1990 because of increasing computing power and the development of algorithms for approximating posterior distributions.

**Technical Notes:** All computations were performed with the R statistical package, which is available for free from the website <http://cran.r-project.org/>. The prior and posterior in the example belong to the family of Beta distributions, and the R functions `dbeta`, `pbeta`, and `qbeta` were used in the calculations.

**Bibliography**

Berger, James O., and Bernardo, José M. 1992, On the Development of the Reference Prior Method, in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O . Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, 35-60.

Berger, James O. and Sellke, Thomas. 1987, Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence *Journal of the American Statistical Association*, 82: 112-122.

Gelman, Andrew, Carlin, John B., Stern, Hal S., and Rubin, Donald B. 2004, *Bayesian Data Analysis, 2<sup>nd</sup> edition*. Chapman & Hall/CRC.

O'Hagan, Anthony, and Forster, Jonathan. 2004, *Bayesian Inference*. Volume 2b of Kendall's Advanced Theory of Statistics, second edition. Arnold, London.

Savage, Leonard J. 1954. *The Foundations of Statistics*. Wiley, New York. Second edition 1972, Courier Dover Publications.

Dennis D. Cox

Professor of Statistics

Rice University

Houston, TX 77005