Bayesian Statistics: Beta-Binomial Model

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

December 3, 2008

Reference: The material in this note is taken from Lynch, S. M. (2007). *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists.* New York: Springer.

A few days prior to the 2004 presidential election, a poll was taken in Ohio. Of 1067 individuals interviewed, 556 individuals said they would vote for John Kerry and 511 individuals said they would vote for George W. Bush. Let K be the proportion of Kerry voters in Ohio. We want to know the probability that $K > 0.5$, meaning the probability that more Ohio voters will vote for Kerry than Bush.

Using Bayes' rule:

$$p(K|data) \propto p(data|K)\, p(K) \tag{1}$$

where $p(data|K)$ is the likelihood of the poll data given K and $p(K)$ is the prior probability distribution for K. Because the poll data is binary (1 = Kerry; 0 = Bush), the likelihood can be characterized as a binomial distribution with $x = 556$ "successes" (votes for Kerry) and $n - x = 511$ "failures" (votes for Bush), with $n = 1067$ total votes. Thus

$$p(data|K) \propto K^{556}\, (1 - K)^{511}. \tag{2}$$

What remains is to specify the prior distribution $p(K)$.

An appropriate prior distribution for an unknown proportion such as K is a beta distribution. The probability density function (pdf) of the beta distribution is:

$$p(K|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}\, K^{\alpha-1}\, (1 - K)^{\beta-1} \tag{3}$$

where $\Gamma(a)$ is the gamma function applied to a and $0 < K < 1$. (The gamma function is the generalization of the factorial to nonintegers. For integers, $\Gamma(a) = (a-1)!$. For nonintegers, $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$. Most software packages will compute this function, but it is often unnecessary in practice, because it tends to be part of the normalizing constant in most problems.) The parameters α and β can be thought of as the prior "successes" and "failures", respectively. This distribution looks similar to the binomial distribution. The key difference is that, whereas the random variable is x and the key parameter is K in the binomial distribution, the random variable is K and the parameters are α and β in the beta distribution.

How do we choose values for α and β? For the purposes of this example, three previous polls had been conducted. If we combine the results of these polls, then 942 individuals said they would vote for Kerry and 1008 individuals said they would vote for Bush. Thus, we set $\alpha = 942$ and $\beta = 1008$.

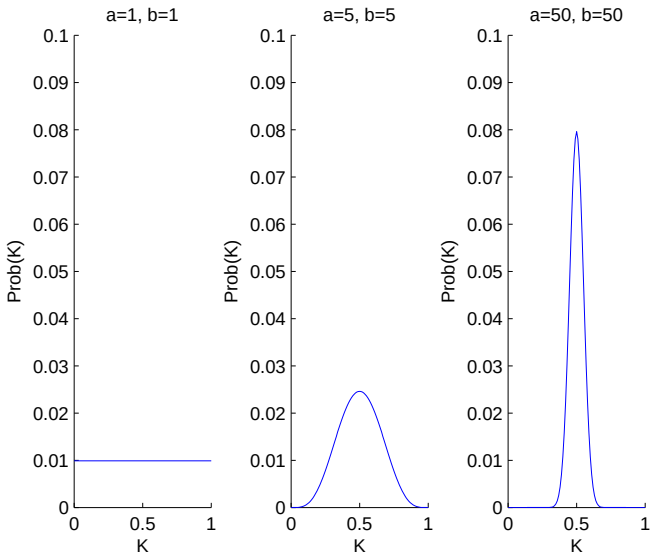

Figure 1: The beta distribution for three different values of α and β. See text for explanation.

After combining this prior with the likelihood, we get

$$p(K|\alpha, \beta, data) \propto K^{556}(1-K)^{511}K^{941}(1-K)^{1007} = K^{1497}(1-K)^{1518}. \quad (4)$$

Importantly, this posterior distribution is also a beta density, with $\alpha = 1498$ and $\beta = 1519$. This highlights the important concept of "conjugacy" in Bayesian statistics. When the prior and likelihood are of such a form that the posterior distribution follows the same form as the prior, the prior and likelihood are said to be conjugate.

To illustrate some of these ideas, Figure 1 plots the beta distribution for $(\alpha = 1, \beta = 1)$, $(\alpha = 5, \beta = 5)$, and $(\alpha = 50, \beta = 50)$. If we think of α and β as the number of successes and failures, the variance of the distribution decreases as the number of data items increases.

Figure 2 shows the posterior distribution for K in a scenario in which a coin is flipped and lands either heads-up or tails-up. (The horizontal axis shows a value, and the vertical axis shows the probability assigned to that value by the posterior distribution. Actually, the probabilities have been linearly scaled so that the largest probability is always equal to 1.) The prior distribution is a beta distribution with $\alpha = 5$ and $\beta = 5$ (4 prior heads, 4 prior tails). The different graphs correspond to different numbers of trials (where a trial is a coin flip). Note that the upper left graph (0 coin flips) shows the prior distribution. As the number of trials increases, the variance of the posterior distribution decreases.

Figure 3 is identical to Figure 2 except that the prior distribution is a beta distribution with $\alpha = 20$ and $\beta = 5$. With small sample sizes, the mean of the posterior distribution is a compromise between the mean of the prior distribution and the mean of the data. As sample sizes increase, the mean of the posterior distribution is closer to the mean of the data, and the variance of the posterior distribution shrinks.

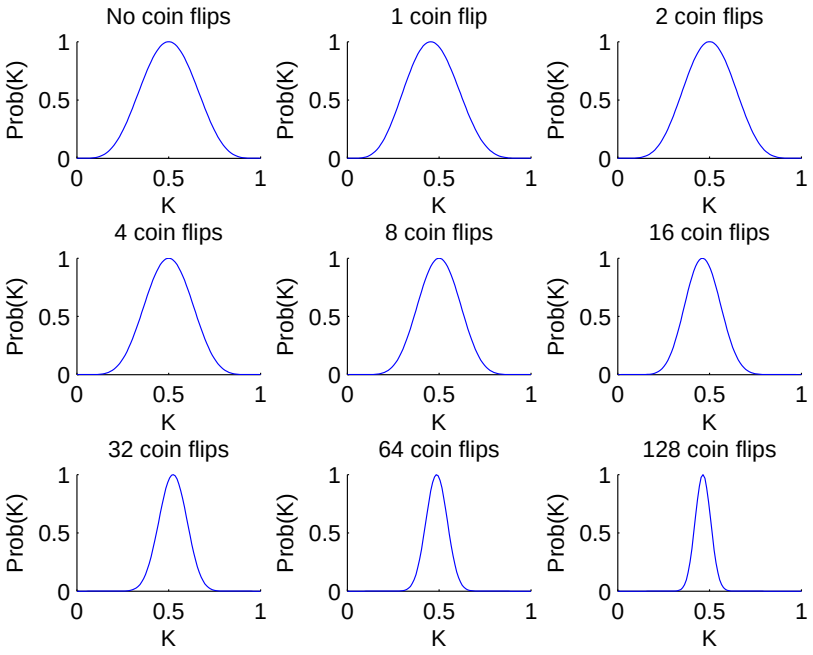

Figure 2: The posterior distribution for K. See text for explanation.

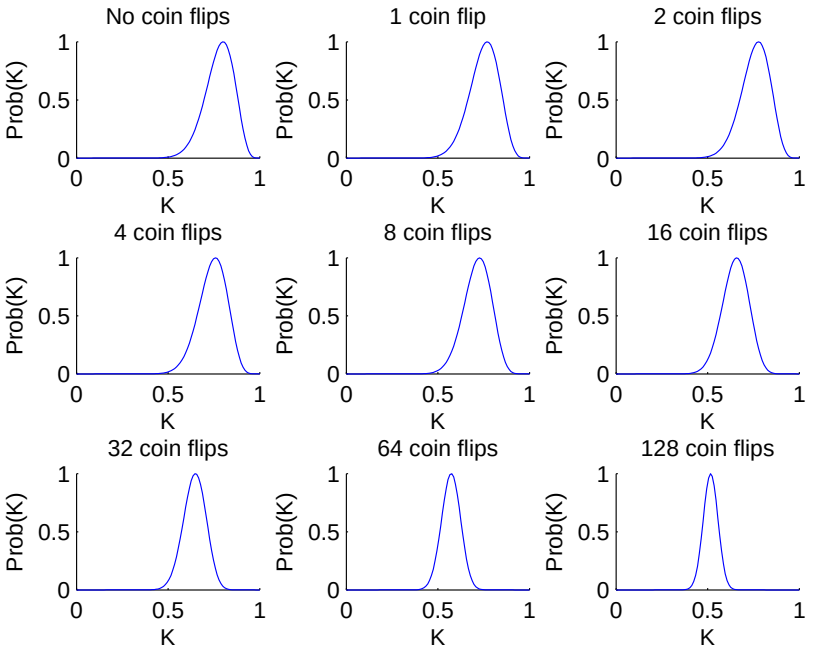

Figure 3: The posterior distribution for K. See text for explanation.