9.1 A Single Coin From a Single Mint

Zongyi Liu

2023-06-19

9.1 A Single Coin From a Single Mint

In the previous chapter, we explored a case in which there were two parameters, and our goal was to estimate each parameter.

The bias θ of a coin determines the probability of getting a head, according to the Bernoulli distribution:

$$p(y|\theta) = bern(y|\theta) = \theta^y(1-\theta)^{1-y}$$

where $y = 1$ for the result "head" and y=0 for the result "tail". We assume independence across flips, so the joint probability of the particular $z = \sum_{i=1}^{N} y_i$ heads out of N flips is $\Pi_{i=1}^{N} p(y_i|\theta) = \theta^z(1-\theta)^{N-z}$

The prior distribution over the biases is denoted $p(\theta)$. For the present example, we sup- pose that the prior is a beta distribution. As was explained in Chapter 5, the beta distribution has two parameters, a and b, and is defined as $beta(\theta, a, b) = \theta^{a-1}(1-\theta)^{b-1}/B(a, b)$. If the mean of our prior belief is μ, and our confidence is reflected by a prior sample size of K, then the corresponding beta parameters are $a = \mu K$ and $b = (1-\mu)K$. For the purposes of the present example, we will treat K as a constant. Because the prior distribution depends on our choice of μ, the prior distribution is a function of μ and can be written

$$p(\theta|\mu)beta(\theta|\mu K, (1-\mu)K)$$

the magnitude of K is an expression of our prior certainty regarding the dependence of the bias on μ. When K is large, the distribution of θ is very narrowly loaded over μ. When K is small, the distribution of θ is very widely dispersed around μ. Thus, as K gets large, we are more and more certain about the form of the dependency of θ on μ.

Now we make the essential expansion of our scenario into the realm of hierarchical models. Instead of specifying a single particular value for μ, we think of μ as taking on many possible values (from 0 to 1), and we specify a probability distribution over those values. This distribution can be thought of as describing the uncertainty in our beliefs about the construction of the mint that manufactured the coin. When μ is large, the mint tends to produce coins with large biases, and when μ is small, the mint tends to produce coins with small biases. Our prior distribution over μ expresses what we believe about how mints are constructed. For the sake of making the example concrete, we suppose that the distribution on μ is again a beta distribution

$$p(\mu) = beta(\mu|A\_\mu, B\_\mu)$$

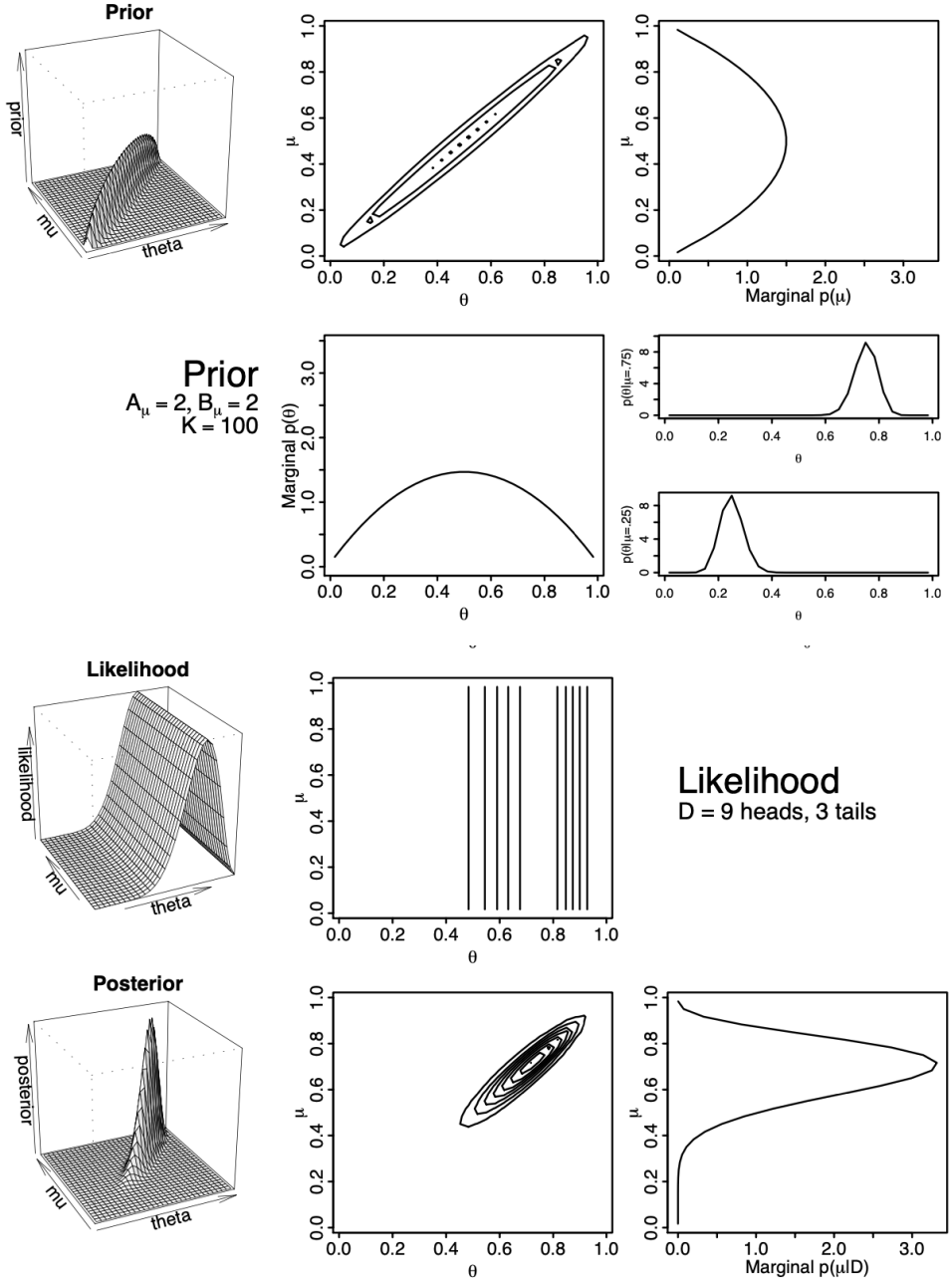
If we treat this situation as simply a case of two parameters, then Bayes' rule is merely $p(\theta, \mu|y) = p(y|\theta, \mu)p(\theta, \mu)/p(y)$. There are two aspects that are "special" about our present situation. First, the likelihood function does not involve μ, so $p(y|\theta, \mu)$ can be re-written as $p(y|\theta)$. Second, because by definition $p(\theta|\mu) = p(\theta, \mu)/p(\mu)$, the prior on the joint parameter space can be factored thus: $p(\theta, \mu) = p(\theta|\mu)p(\mu)$. Therefore, Bayes' rule for our current hierarchical model has the form

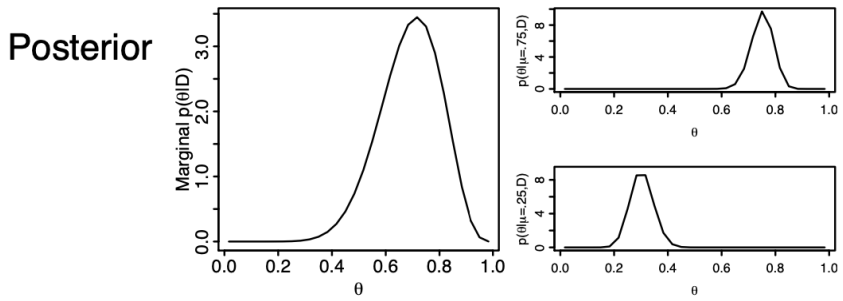
$$p(\theta, \mu|y) = p(y|\theta, \mu)p(\theta, \mu)/p(y) = p(y|\theta)p(\theta|\mu)p(\mu)/p(y)$$

9.1.1 Posterior via Grid Approximation

When the parameter(s) and hyperparameter(s) extend over a finite domain, and there are not too many of them, then we can approximate the posterior via grid approximation. In our present situation, we have the parameter θ and hyperparameter μ that both have finite domains, namely the interval [0, 1]. Therefore a grid approximation is tractable and the distributions can be readily graphed.

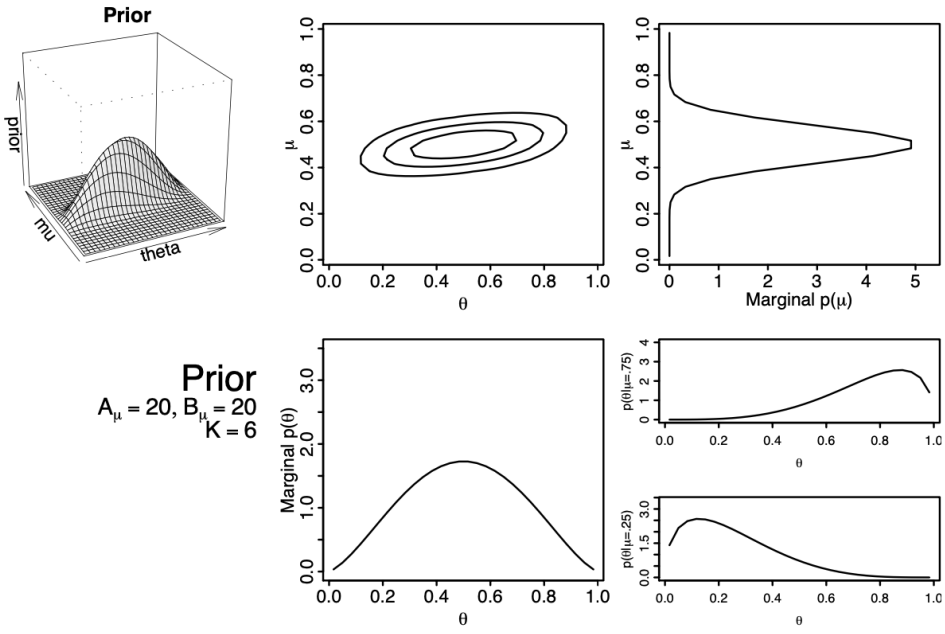
Figure 9.2 shows an example in which the hyperprior distribution has the form of a beta distribution as in Equation 9.3, with $A_\mu = 2$ and $B_\mu = 2$, i.e., $p(\mu) = beta(\mu|2, 2)$. This hyperprior expresses the belief that the mint's μ value is near .5, but there is large uncertainty.

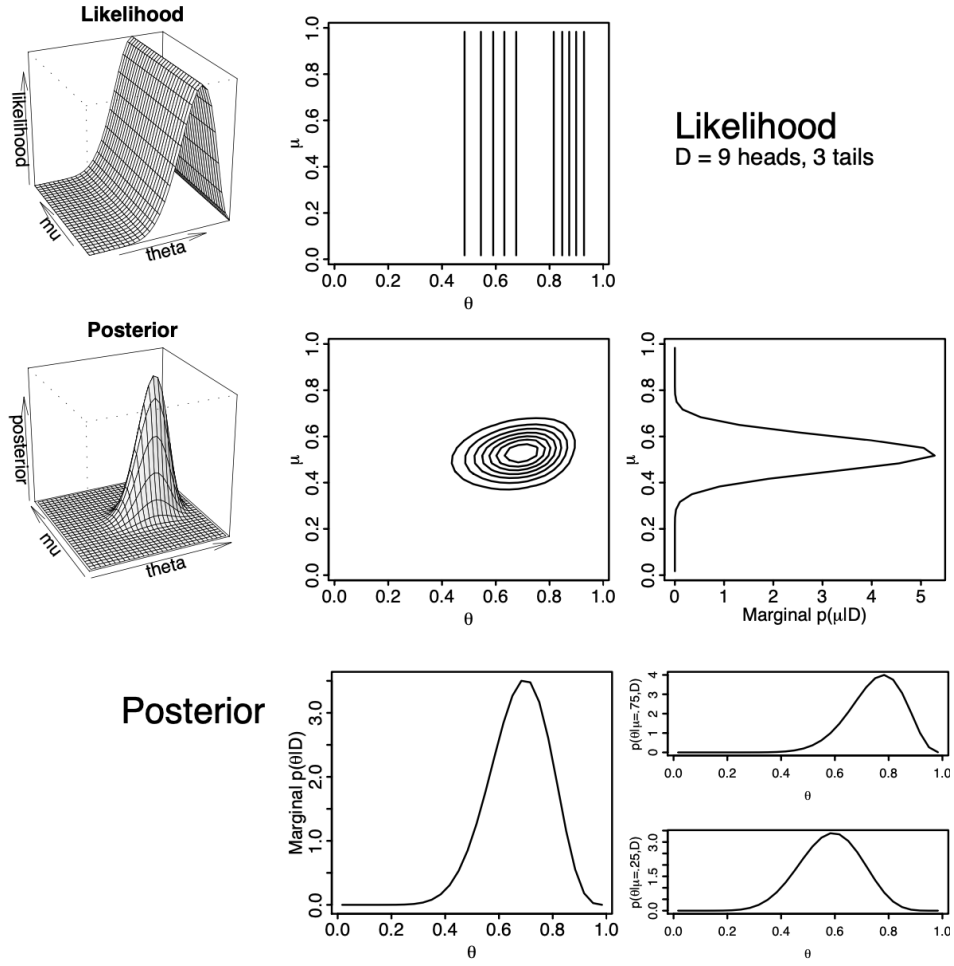




The prior distribution on θ, or more precisely, the prior distribution regarding the dependency of θ on μ, is expressed by another beta distribution, as in Equation 9.2 with K = 100, whereby $p(\theta|\mu) = beta(\theta, \mu 100, (1-\mu)100)$. The prior expresses a high degree of certainty that a mint with hyperparameter μ generates coins that have a bias θ close to μ. Two cases of this conditional distribution are shown in the right panel of the second row of Figure 9.2. The upper graph within that panel shows $p(\theta|\mu = 0.75)$, and the lower graph in that panel shows $p(\theta|\mu = 0.25)$. You can see that the conditional distributions are fairly tightly centered on 0.75 and 0.25, respectively.

A contrasting case, when there is high certainty on the prior regarding μ, but low certainty on the prior regarding the dependence of θ on μ. Fig- ure 9.3 illustrates such a case, where $p(\mu) = beta(\mu|20, 20)$ and $p(\theta|\mu) = beta(\theta|\mu 6, (1-\mu)6)$. The top row, right panel, shows that $p(\mu)$ is sharply peaked over $\mu = .5$, but the conditional distributions $p(\theta|\mu)$ are very broad (second row, right panel). The same data as for Figure 9.2 are used here, so the likelihood graphs look the same in the two figures.





We can now compare prior and the posterior within Figure 9.3. The distribution over μ hardly changes, because it began with high certainty. The distributions $p(\theta|\mu, D)$ are very different from their priors, however. This is because they began with low certainty, so the data can have a big impact on these distributions. In this case, the data suggest that θ depends on μ in a rather different way than we initially suspected.

In summary, the data influence

1. our beliefs about the hyperparameter
2. our beliefs about the dependence of the parameter on the hyperparameter.