

Bayesian Estimation of Confidence Intervals: Lecture XXIV

Charles B. Moss

October 12, 2026

Mathematica Quickstart

- The homework assignment will ask for the construction of a Bayesian confidence interval. Please use Mathematica The example program is called BayesTriangle.nb.
- The first step is to use the prior $g(p) = 2(5-p)/9$ to derive the posterior distribution.

$$f[x, p] = \frac{6(1-p)^{1-x}(5-p)p^x \sin(\pi x)}{\pi(x-14)(x-1)x} \quad (1)$$

- One thing about this solution

$$\lim_{x \rightarrow 0} \frac{\sin(\pi x)}{x} = \pi \quad (2)$$

- Next, consider what it means if the coin toss is a “tails” – $x = 0$:

$$f[0, p] = \frac{3}{7} (1 - p) (5 - p) \quad (3)$$

- The expected value (or Bayesian estimator) of p is

$$E[p] = \int_0^1 \frac{3}{7} p (1 - p) (5 - p) dp = \frac{9}{28} \quad (4)$$

- In order to demonstrate how to solve for the lower bound of the confidence interval if $x = 0$ consider the integral

$$\int_0^{p^*} f[0, p] dp = \frac{3}{7} \int_0^{p^*} (1 - p) (5 - p) dp = 0.05 \quad (5)$$

Mathematica finds three possible solutions, one is $p^* = 0.0237$. The others are out of the range of p .

- 1 Bayesian Estimation
- 2 Bayesian Confidence Intervals

Bayesian Estimation

- Implicitly in our previous discussions about estimation, we adopted a classical viewpoint.
 - We had some process generating random observations.
 - This random process was a function of fixed, but unknown parameters.
 - We then designed procedures to estimate these unknown parameters based on observed data.

- Specifically, if we assumed that a random process such as students admitted to the University of Florida, generated heights. This height process can be characterized by a normal distribution.
 - We can estimate the parameters of this distribution using maximum likelihood.
 - The likelihood of a particular sample can be expressed as

$$L(X_1, X_2, \dots, X_n | \mu, \sigma^2) = \frac{1}{(2\pi)^{n/2} \sigma^n} \exp \left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right] \quad (6)$$

- Our estimates of μ and σ^2 are then based on the value of each parameter that maximizes the likelihood of drawing that sample.

- Turning this process around slightly, Bayesian analysis assumes that we can make some kind of probability statement about parameters before we start. The sample is then used to update our prior distribution.
 - First, assume that our prior beliefs about the distribution function can be expressed as a probability density function $\pi(\theta)$ where θ is the parameter we are interested in estimating.
 - Based on a sample (the likelihood function) we can update our knowledge of the distribution using Bayes rule

$$\pi(\theta|X) = \frac{L(X|\theta) \pi(\theta)}{\int_{-\infty}^{\infty} L(X|\theta) \pi(\theta) d\theta} \quad (7)$$

- Departing from the book's example, assume that we have a prior of a Bernoulli distribution. Our prior is that P in the Bernoulli distribution is distributed $B(\alpha, \beta)$.
 - The beta distribution is defined similar to the gamma distribution

$$f(P|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} P^{\alpha-1} (1-P)^{\beta-1} \quad (8)$$

$B(\alpha, \beta)$ is defined as

$$B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = \frac{\Gamma(\alpha) \Gamma(\beta)}{\Gamma(\alpha + \beta)} \quad (9)$$

- Thus, the beta distribution is defined as

$$f(P|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} P^{\alpha-1} (1 - P)^{\beta-1} \quad (10)$$

- Assume that we are interested in forming the posterior distribution after a single draw

$$\begin{aligned} \pi(P|X) &= \frac{P^X (1 - P)^{1-X} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} P^{\alpha-1} (1 - P)^{\beta-1}}{\int_0^1 P^X (1 - P)^{1-X} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} P^{\alpha-1} (1 - P)^{\beta-1} dP} \\ &= \frac{P^{X+\alpha-1} (1 - P)^{\beta-X}}{\int_0^1 P^{X+\alpha-1} (1 - P)^{\beta-X} dP} \end{aligned} \quad (11)$$

- Following the original specification of the beta function

$$\int_0^1 P^{X+\alpha-1} (1-P)^{\beta-X} dP = \int_0^1 P^{\alpha^*} (1-P)^{\beta^*-1} dp$$

where $\alpha^* = X + \alpha$ and $\beta^* = \beta - X + 1$

$$\Rightarrow \int_0^1 P^{X+\alpha-1} (1-P)^{\beta-X} dP = \frac{\Gamma(X+\alpha) \Gamma(\beta-X+1)}{\Gamma(\alpha+\beta+1)} \quad (12)$$

The posterior distribution, the distribution of P after the value of the observation is known, is then

$$\pi(P|X) = \frac{\Gamma(\alpha+\beta+1)}{\Gamma(X+\alpha) \Gamma(\beta-X+1)} P^{X+\alpha-1} (1-P)^{\beta-X} \quad (13)$$

- The Bayesian estimate of P is then the value that minimizes a loss function. Several loss functions can be used, but we will focus on the quadratic loss function consistent with mean square errors

$$\begin{aligned}\min_P E \left[(\hat{P} - P)^2 \right] &\Rightarrow \frac{\partial E \left[(\hat{P} - P)^2 \right]}{\partial \hat{P}} = 2E \left[\hat{P} - P \right] = 0 \\ &\Rightarrow \hat{P} = E[P]\end{aligned}\tag{14}$$

- Taking the expectation of the posterior distribution yields

$$\begin{aligned} E[P] &= \int_0^1 \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(X + \alpha) \Gamma(\beta - X + 1)} P^{X+\alpha} (1 - P)^{\beta-X} dP \\ &= \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(X + \alpha) \Gamma(\beta - X + 1)} \int_0^1 P^{X+\alpha} (1 - P)^{\beta-X} dP \end{aligned} \quad (15)$$

- As before, we solve the integral by creating $\alpha^* = \alpha + X + 1$ and $\beta^* = \beta - X + 1$. The integral then becomes

$$\begin{aligned}\int_0^1 P^{\alpha^*-1} (1-P)^{\beta^*-1} dP &= \frac{\Gamma(\alpha^*) \Gamma(\beta^*)}{\Gamma(\alpha^* + \beta^*)} \\ &= \frac{\Gamma(\alpha + X + 1) \Gamma(\beta - X + 1)}{\Gamma(\alpha + \beta + 2)}\end{aligned}\tag{16}$$

Hence,

$$E[P] = \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + \beta + 2)} \frac{\Gamma(\alpha + X + 1)}{\Gamma(\alpha + X)} \frac{\Gamma(\beta - X + 1)}{\Gamma(\beta - X + 1)}\tag{17}$$

- Which can be simplified using the fact

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \quad (18)$$

Therefore

$$\begin{aligned} \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(\alpha + \beta + 2)} \frac{\Gamma(\alpha + X + 1)}{\Gamma(\alpha + X)} &= \frac{\Gamma(\alpha + \beta + 1)}{(\alpha + \beta + 1) \Gamma(\alpha + \beta + 1)} \\ &\times \frac{(\alpha + X) \Gamma(\alpha + X)}{\Gamma(\alpha + X)} \\ &= \frac{(\alpha + X)}{(\alpha + \beta + 1)} \end{aligned} \quad (19)$$

- To make this estimation process operational, assume that we have a prior distribution with parameters $\alpha = \beta = 1.4968$ that yields a beta distribution with a mean P of 0.5 and a variance of the estimate of 0.0625. Next assume that we flip a coin and it comes up heads ($X = 1$). The new estimate of P becomes 0.6252. If, on the other hand, the outcome is a tail ($X = 0$) the new estimate of P is 0.3747.
- Extending the results to n Bernoulli trials yields

$$\pi(P|X) = \frac{\Gamma(\alpha + \beta + n)}{\Gamma(\alpha + Y) \Gamma(\beta - Y + n)} P^{Y+\alpha-1} (1 - P)^{\beta-Y+n-1} \quad (20)$$

- where Y is the sum of individual X s or the number of heads in the sample. The estimated value of P then becomes

$$\hat{P} = \frac{Y + \alpha}{\alpha + \beta + n} \quad (21)$$

- Going back to the example in the last lecture, in the first draw $Y = 15$ and $n = 50$. This yields an estimated value of P of 0.3112. This value compares with the maximum likelihood estimate of 0.3000. Since the maximum likelihood estimator in this case is unbiased, the results imply that the Bayesian estimator is biased.

Bayesian Confidence Intervals

- Apart from providing an alternative procedure for estimation, the Bayesian approach provides a direct procedure for the formulation of parameter confidence intervals.
- Returning to the simple case of a single coin toss, the probability density function of the estimator becomes

$$\pi(P|X) = \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(X + \alpha) \Gamma(\beta - X + 1)} P^{X+\alpha-1} (1 - P)^{\beta-X} \quad (22)$$

- As previously discussed, we know that given $\alpha = \beta = 1.4968$ and a head, the Bayesian estimator of P is 0.6252. However, using the posterior distribution function, we can also compute the probability that the value of P is less than 0.5 given a head

$$\begin{aligned}
 P[p < 0.5] &= \int_0^{0.5} \frac{\Gamma(\alpha + \beta + 1)}{\Gamma(X + \alpha) \Gamma(\beta - X + 1)} P^{X+\alpha-1} (1 - P)^{\beta-X} dP \\
 &= 0.2976.
 \end{aligned}
 \tag{23}$$

Hence, we have a very formal statement of confidence intervals.

Current/Recent Approaches to Bayesian Analysis

- As a starting point, consider a sample of five observations with a normal likelihood function with parameters $\hat{\mu}$ and $\hat{\sigma}^2$. Next, we construct a prior. First, we assume that the $\hat{\mu}$ is distributed $N[\mu, \sigma^2]$. Second we assume that $\hat{\sigma}^2$ is distributed $\Gamma[\alpha, \beta]$.

$$\ln(L) = k_L - \frac{5}{2} \ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{k=1}^5 (x_k - \hat{\mu})^2$$
$$\ln(f_N) = k_N - \frac{1}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} (\hat{\mu} - \mu)^2 \quad (24)$$

$$\ln(f_\Gamma) = -\ln(\Gamma[\alpha]) - \alpha \ln(\beta) + (\alpha - 1) \ln(\hat{\sigma}^2) - \frac{\hat{\sigma}^2}{\beta}$$

- Maybe to overstate the case, we can write the posterior distribution as $\mathbf{L} \times f_N \times f_\gamma \Rightarrow \ln(L) + \ln(f_N) + \ln(f_\Gamma)$

$$\begin{aligned} \ln(g(\theta)) = & -\frac{5}{2} \ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2} \sum_{k=1}^5 (x_k - \hat{\mu})^2 - \\ & \frac{1}{2} \ln(\sigma^2) \frac{1}{2} (\hat{\mu} - \mu)^2 - \ln(\Gamma(\alpha)) - \alpha \ln(\beta) + \\ & (\alpha - 1) \ln(\hat{\sigma}^2) - \frac{\hat{\sigma}^2}{\beta} \end{aligned} \quad (25)$$

- Taking a brief aside, consider the concept of simulated annealing - integration by simulation.
 - Suppose that you want to integrate the function $u(x) = \ln(x)$ over all possible outcomes for where $x \sim \Gamma(\alpha, \beta)$ where $\alpha = 0.75$ and $\beta = 0.25$.
 - One approach would be to draw a large number of x s from the gamma distribution and take the average of the results

$$E[u(x)] = \frac{1}{N} \sum_{i=1}^N u(x_k) \text{ where } x_i \sim \Gamma[\alpha, \beta]. \quad (26)$$

- The current approach to Bayesian estimation is similar, except we use a sequence commonly referred to as Markov Chain Monte Carlo.

- Mathematically returning to the results in Equation 25 we start with some guess or initial value of θ_0 and draw θ_1
- Basically, we start with a vector of parameter θ_k and we want to draw the next set of parameters θ_{k+1}

$$\begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \\ \mu \\ \sigma^2 \\ \alpha \\ \beta \end{bmatrix}_i \rightarrow \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \\ \mu \\ \sigma^2 \\ \alpha \\ \beta \end{bmatrix}_{i+1} \quad (27)$$

- Kind of the hidden step in modern applied Bayes in the sampling distribution.
- The point is that the distribution implied by Equation 25 is hard to work with. Hence, following the path we have worked with several times, suppose that we use a second-order Taylor series expansion of the posterior function

$$\ln(g(\theta_{i+1})) \approx \ln(g(\theta_i)) + \frac{1}{2}(\theta_{i+1} - \theta_i)' \frac{\partial^2 \ln(g(\theta))}{\partial \theta \partial \theta'} \bigg|_{\theta=\theta_k} (\theta_{k+1} - \theta_k) \quad (28)$$

$$H(\theta) = -\frac{\partial^2 \ln(g(\theta))}{\partial \theta \partial \theta'}, \quad \theta_{k+1} \sim N[\theta_k, H(\theta_k)]$$

- The current approach to Bayesian estimation is similar, except we use a sequence commonly referred to as Markov Chain Monte Carlo. Mathematically we start with some guess or initial value of θ_0 and draw θ_1

$$\theta_1 = \theta_0 + \text{Chol}[H(\theta_0)] \epsilon_k, \epsilon_k \sim N[0, I] \quad (29)$$

where $\text{Chol}[\cdot]$ denotes the Cholesky Decomposition of the matrix H and ϵ_k is a vector of standard normal random variables.

- Hence, the toughest part of this process is deriving $H(\theta)$.
This is hard, but not impossible.

$$\frac{\partial^2 \ln(g(\theta))}{\partial \theta \partial \theta'} = \frac{\partial^2 \ln(L)}{\partial \theta \partial \theta'} + \frac{\partial^2 \ln(f_N)}{\partial \theta \partial \theta'} + \frac{\partial^2 f_{\Gamma}}{\partial \theta \partial \theta'}. \quad (30)$$

- To put a little flesh on the discussion, consider the first set of second derivatives (i.e, with respect to the likelihood function)

$$\begin{aligned} \frac{\partial \ln(L)}{\partial \hat{\mu}} &= \frac{1}{\hat{\sigma}^2} \sum_{k=1}^5 (x_k - \hat{\mu}) \\ \frac{\ln(L)}{\partial \hat{\sigma}^2} &= -\frac{5}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2} \sum_{k=1}^N (x_k - \hat{\mu})^2 \end{aligned} \quad (31)$$

- This gives us four Hessian terms

$$\begin{aligned}
 H_{L,1,1} &= \frac{\partial^2 \ln(L)}{\partial \hat{\mu} \partial \hat{\mu}} = -\frac{1}{\hat{\sigma}^2} \\
 H_{L,1,2} &= \frac{\partial^2 \ln(L)}{\partial \hat{\sigma}^2 \partial \hat{\mu}} = -\frac{1}{(\hat{\sigma}^2)^2} \sum_{k=1}^5 (x_k - \hat{\mu}) \\
 H_{L,2,2} &= \frac{\partial^2 \ln(L)}{\partial \hat{\sigma}^2 \partial \hat{\sigma}^2} = \frac{5}{2(\hat{\sigma}^2)^2} - \frac{1}{(\hat{\sigma}^2)^3} \sum_{k=1}^5 (x_k - \hat{\mu})^2
 \end{aligned} \tag{32}$$

Software such as BUGS or Stan use numerical routines to compute the Hessian matrix.

Improved Markov Chain Monte Carlo Algorithms

- **Metropolis-Hastings:** The first alteration involves making a rule for whether you want to take a step or not.
 - For example, suppose that your current solution is θ_k which has a likelihood function of $g(\theta_k)$. Given this point, you draw a new point θ_{k+1} which has the likelihood value $g(\theta_{k+1})$.
 - The question is whether this value is better - is $g(\theta_{k+1})$ as good as your original point?

$$\rho = \frac{g(\theta_{k+1})}{g(\theta_k)}. \quad (33)$$

One approach is to decide whether or not to go the new location based on a random uniform draw u . If $u \leq \rho$, you go to the new location.

- **Gibbs Sampler:** The Gibbs sampler proposes a process that concentrates the step by imposing sequential information.

Gibbs Sampler

- Suppose that we have three parameters θ_1 , θ_2 , and θ_3 . Mathematically, it is possible to draw all three parameters at the same time. An alternative is to draw the numbers in sequence. Focusing on Equation 29

$$\begin{bmatrix} \theta_{k+1,1} \\ \theta_{k+1,2} \\ \theta_{k+1,3} \end{bmatrix} = \begin{bmatrix} \theta_{k,1} \\ \theta_{k,2} \\ \theta_{k,3} \end{bmatrix} + \text{Chol}[H(\theta_k)] \begin{bmatrix} \epsilon_{k,1} \\ \epsilon_{k,2} \\ \epsilon_{k,3} \end{bmatrix} \quad (34)$$

The Gibbs sampler involving drawing the first theta by itself using based on information from the last draw

$$\theta_{k+1,1} = \theta_{k,1} + \text{Chol}[H(\theta_{k,1}, \theta_{k,2}, \theta_{k,3})] \begin{bmatrix} \epsilon_{k,1} \\ 0 \\ 0 \end{bmatrix} \quad (35)$$

- Given the information from the first draw, the second parameter is then drawn conditional on this information

$$\theta_{k+1,2} = \theta_{k,2} + \text{Chol}[H(\theta_{k+1,1}, \theta_{k,2}, \theta_{k,3})] \begin{bmatrix} \epsilon_{k,1} \\ \epsilon_{k,2} \\ 0 \end{bmatrix}. \quad (36)$$

Finally, the last parameter is drawn based on information from the first two

$$\theta_{k+1,3} = \theta_{k,3} + \text{Chol}[H(\theta_{k+1,1}, \theta_{k+1,2}, \theta_{k,3})] \begin{bmatrix} \epsilon_{k,1} \\ \epsilon_{k,2} \\ \epsilon_{k,3} \end{bmatrix} \quad (37)$$

Readings

- Moss 9.1 – 9.2