# Approximating a posterior two ways

**10 total points**

Christopher Cahill

15 September 2023

### NOTE: date above is the due date

#### Background

Your friend is a biologist who spends their time flying around in a Bell-206 helicopter staring out at the western Hudson Bay lowlands in Nunavut, Canada. They stare out the window for hours at the vast, blinding whiteness and usually see nothing and often feel queasy because of motion sickness. However, on some days things go well and they see polar bears. When they do, they fly down and dart the adults and collar them. If there are any cubs around, they catch them and take measurements (and usually a lot of cute pictures).

Over the last six years your friend has weighed a total of 37 cubs (see `cubs.rds`), and they now want to determine whether or not cubs in the region are underweight relative to cubs elsewhere in the Canadian Arctic. Thus, they need to estimate the mean weight of the 37 cubs and the associated uncertainty around this mean weight. Their graduate supervisor is rather opinionated, and feels strongly that they need to use a highly flexible gamma distribution to estimate mean cub weight and also that ideally they should combine this with prior knowledge from the supervisor's previous studies. Your friend googled Bayesian statistics and the gamma distribution and saw the equations (below) and they wanted to die. Instead of choosing death, they decide to ask you for help as they know you are a hyper-nerd.

The gamma distribution is often written as

If $\alpha \in \mathbb{R}^+$and $\beta \in \mathbb{R}^+$, then for $y \in \mathbb{R}^+$,

$$\text{Gamma}(y \mid \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y).$$

Here, $\alpha$ and $\beta$ are the shape and rate parameters of the Gamma distribution, respectively, while $y$ is data. The fancy R with the plus sign (i.e., $\mathbb{R}^+$) just means $\alpha$ and $\beta$ are real values (as opposed to integers) and must be positive (i.e., this is what the plus sign means). The fancy $\in$ symbol means that the element on the left of the symbol is in the set on the right

of the symbol. The final horrible math symbol that needs explaining is the $\Gamma$, which is a gamma function. We don't need to worry about the gamma function too much other than to know that it is a mathematical tool for taking complex numbers and breaking them down into simpler things that we can use when we estimate statistical models.

Your friend needs to do the following things to satisfy their grumpy supervisor:

1. Rather than develop a model that estimates $\alpha$ and $\beta$ directly, they should use a reparameterization of the standard gamma probability distribution function. In particular, they need to build their gamma model in terms of the mean cub weight $\mu$ and coefficient of variation $cv$ in cub weights rather than $\alpha$ and $\beta$, which will make it easier to put priors on these parameters. To do this, you look on the wikipedia page and see that the mean and variance of the gamma distribution are nicely written out for you as $E(y) = \frac{\alpha}{\beta}$ and $Var(y) = \frac{\alpha}{\beta^2}$, respectively. However, you will need to do the math to get the $cv$ of the gamma distribution. (2 points).

2. Next, because this problem only has two parameters, the supervisor wants your friend to approximate the best estimates of the $\mu$ and $cv$ parameters using a grid search approximation. The supervisor thinks this will help your friend better understand Bayesian statistics and make sure the estimates you get from any Markov Chain Monte Carlo (MCMC) routines later on are reasonable. Conduct a grid search approximation of the posterior distribution for $\mu$ and $cv$, and report the best posterior estimates you find for $\mu$, $cv$, and the log of the posterior. Your grid search should consider 1000 values for each parameter. The grid of parameter values for $\mu$ should range from 300 to 900 while the grid for $cv$ should range from 0.01 to 0.5. Additionally, the priors from a previous study the supervisor conducted are $\mu \sim \text{N}(\mu = 635, \sigma^2 = 900^2)$ and $cv \sim \text{N}(\mu = 0.2, \sigma^2 = 1^2)$. (4 points).

3. Repeat the analysis using MCMC in Stan and by writing your own `.stan` file to do this. For your final solution, run four independent chains and use `iter_warmup` = `iter_sampling` = 1000. Check the model diagnostics, and record any warnings you receive. Plot chains and posterior histograms of the relevant parameters in the model. Compare the posterior median and 95% credible intervals to the best parameter estimates from the grid approximation above, recognizing they almost certainly will not match up perfectly. (4 points).

Provide a brief write-up of your findings and all math, R, and Stan code required to conduct these analyses.

4. BONUS (0 points): If your friend measured a new bear cub (see Figure 1 below) that was 598 g and wanted to know whether this cub was underweight relative to other cubs in the region, could you determine the probabliity that the bear is underweight using the MCMC sample? Hint: you need to generate a posterior predictive distribution for mean cub weight.

Figure 1: photo credit: Andy Derocher