Bayesian Principles

Ben Goodrich

February 12, 2019

Obligatory Disclosure

- Ben is an employee of Columbia University, which has received several research grants to develop Stan
- Ben is also a manager of GG Statistics LLC, which uses Stan for business purposes
- According to Columbia University policy, any such employee who has any equity stake in, a title (such as officer or director) with, or is expected to earn at least $\$5,000.00$ per year from a private company is required to disclose these facts in presentations

Krushcke Quotes

- Krushcke (2015, p.15): “[p]ossibilities, over which we allocate credibility, are parameter values in meaningful mathematical models”
- The possibilities are represented by Θ and the parameters are $\boldsymbol{\theta}$ in some model for the data-generating process for Y, which leads us to Bayes Rule

$$f\left(\boldsymbol{\theta}\right|\mathbf{y}) = \frac{f\left(\boldsymbol{\theta}\right) f\left(\mathbf{y}\right|\boldsymbol{\theta})}{f\left(\mathbf{y}\right)} = \frac{f\left(\boldsymbol{\theta}\right) f\left(\mathbf{y}\right|\boldsymbol{\theta})}{\int_{\Theta} f\left(\boldsymbol{\theta}\right) f\left(\mathbf{y}\right|\boldsymbol{\theta}) \, d\boldsymbol{\theta}} = \frac{f\left(\boldsymbol{\theta}, \mathbf{y}\right)}{\int_{\Theta} f\left(\boldsymbol{\theta}, \mathbf{y}\right) d\boldsymbol{\theta}}$$

- Krushcke (2015, p.15): “Bayesian inference is reallocation of credibility across possibilities” after observing new data
- Your prior beliefs are represented by the PDF $f\left(\boldsymbol{\theta}\right)$ and then you condition on the available data on $\mathbf{y}$ to map your prior beliefs into your posterior beliefs by multiplying the prior PDF by $\frac{L(\boldsymbol{\theta};\mathbf{y})}{f(\mathbf{y})}$ to obtain the posterior PDF $f\left(\boldsymbol{\theta}\right|\mathbf{y})$

Ex Ante Probability Density / Mass Function

A likelihood function is the same expression as a P{D,M}F with 3 distinctions:

1. For the PDF or PMF, $f\left(x\middle|\,\boldsymbol{\theta}\right)$, we think of X as a random variable and $\boldsymbol{\theta}$ as given, whereas we conceive of the likelihood function, $\mathcal{L}\left(\boldsymbol{\theta};x\right)$, to be a function of $\boldsymbol{\theta}$ evaluted at the OBSERVED data, x
   - As a consequence, $\int\limits_{-\infty}^{\infty} f\left(x\middle|\,\boldsymbol{\theta}\right)dx = 1$ or $\sum\limits_{i:x_i\in\Omega} f\left(x_i\middle|\,\boldsymbol{\theta}\right) = 1$ while $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\mathcal{L}\left(\boldsymbol{\theta};x\right)d\theta_1 d\theta_2\dots d\theta_K$ is positive but not 1
2. We often think of “the likelihood function” for N conditionally independent observations, so $\mathcal{L}\left(\boldsymbol{\theta};\mathbf{x}\right) = \prod_{n=1}^{N}\mathcal{L}\left(\boldsymbol{\theta};x_n\right)$
3. By “the likelihood function”, we often really mean the natural logrithm thereof, a.k.a. the log-likelihood function $\ell\left(\boldsymbol{\theta};\mathbf{x}\right) = \ln\mathcal{L}\left(\boldsymbol{\theta},\mathbf{x}\right) = \sum_{n=1}^{N}\ln\mathcal{L}\left(\boldsymbol{\theta};x_n\right)$

Bayesian Workflow

1. Classify concepts into Exogenous / Endogenous & Known / Unknowable
2. Draw from the prior predictive distribution of your generative model
    - Test that your software is working on generated data
    - Possibly pre-register your proposed analysis
3. Obtain new data, at least on the Endogenous Knowns
4. Use Bayes Rule to obtain the posterior distribution of the Unknowables given the Known data. If using Stan, this will often entail warning messages that you should fix.
5. Draw from the posterior predictive distribution and compare to the empirical distribution of the endogenous Knowns

- All of this can be iterated for multiple models that can then be compared using Bayesian techniques. You should build from simple models to more complex ones but do not let the posterior distribution of one model influence your prior distribution of another model for the SAME data

Generative Models

- In order to draw from the prior predictive distribution, you have to have a model that you can simulate from
- Without a generative model, you cannot update your beliefs with Bayes Rule

Concept	Known	Unknowable
Exogenous	sizes, predictors, prior modes / medians / etc.	parameters
Endogenous	outcomes	intermediates, predictions, utility

- **Endogenous Known**: Count of people (Y) in residences
- **Exogenous Unknowable** : Expected number (μ) of people in a residence
- **Exogenous Known**: Number of residences, prior mode and mean for μ
- **Endogenous Unknowable** : Predicted number in a future residence
- Still need to choose distributions for μ and $Y \mid \mu$

Prior Predictive Distribution

- Drawing from the prior predictive distribution of a generative model is a good way to check that the (prior) probability distribution for the exogenous unknowables is generates reasonable looking outcomes
- Drawing from the prior predictive distribution is also a good way to confirm that your software is working well before you try it with observed data
- Since our outcome is a count of the number of people in a residence, we can use the Poisson distribution with unknowable parameter μ
- How would we write the log-probability of observing N Poisson independent random variables with expectation μ?
- $\prod_{n=1}^N \frac{e^{-\mu}\mu^{y_n}}{y_n!} = e^{-N\mu}\mu^{\sum_{n=1}^N y_n} \prod_{n=1}^N \frac{1}{y_n!}$ so $\ln \prod_{n=1}^N \frac{e^{-\mu}\mu^{y_n}}{y_n!} = -N\mu + \left(\sum_{n=1}^N y_n\right) \ln \mu - \sum_{n=1}^N \ln(y_n!)$
- But we need a prior distribution to draw realizations of $\widetilde{\mu}$ from

Gamma Distribution

- If $\Theta = \mathbb{R}_+$ and $a > 0$, the PDF of the “unit” Gamma distribution is

$$f\left(\theta \middle| a\right) = \frac{1}{\Gamma\left(a\right)}\theta^{a-1}e^{-\theta}$$

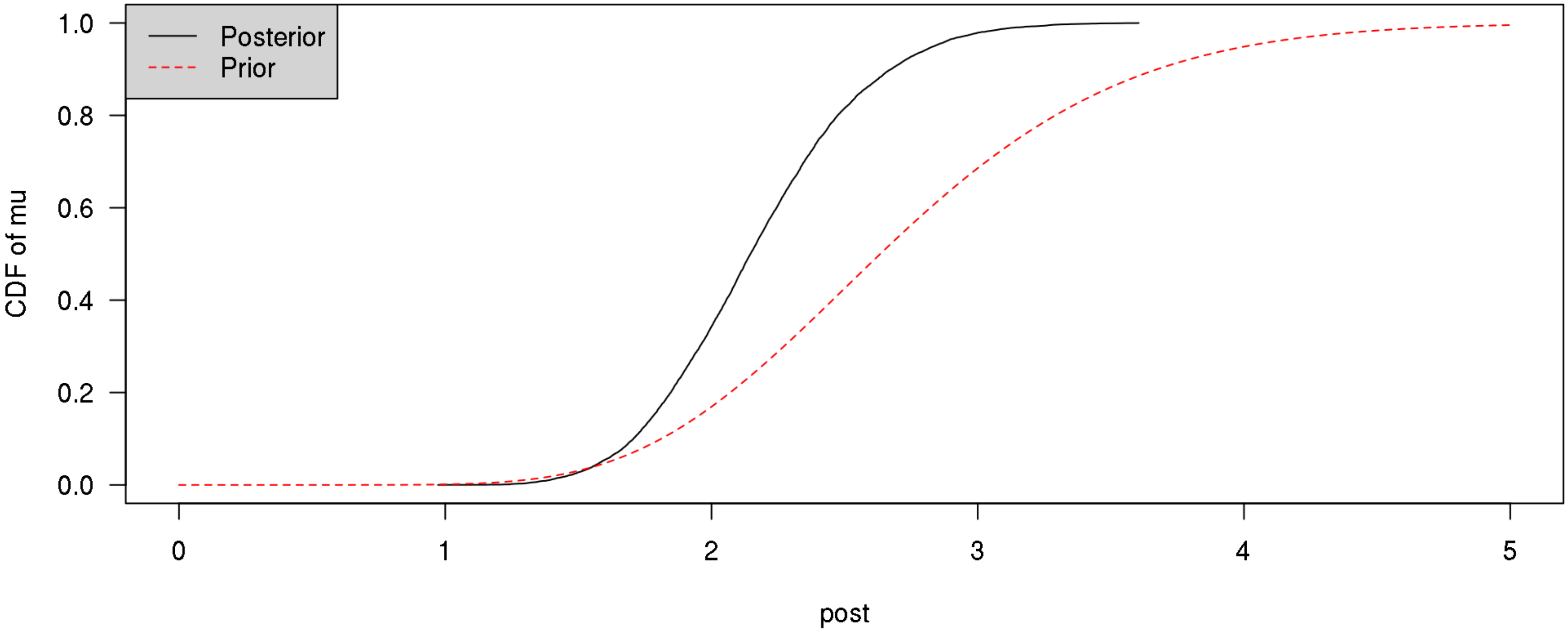
- What is an expression for $\Gamma\left(a\right)$?
- Suppose $\mu = \frac{\theta}{b}$ with $b > 0$. What is the PDF of μ?
- $\theta = b\mu$ so $\frac{\partial\theta}{\partial\mu} = b$ and $f\left(\mu \middle| a, b\right) = \frac{b}{\Gamma(a)}\left(b\mu\right)^{a-1}e^{-b\mu} = \frac{b^a}{\Gamma(a)}\mu^{a-1}e^{-b\mu}$
- This is the PDF of the (general) Gamma distribution with shape $a > 0$ and rate $b > 0$, but sometimes you will see it with scale $\frac{1}{b}$
- $\mathbb{E}\mu = \frac{a}{b}$ and the mode is $M = \frac{a-1}{b}$, provided $a > 1$
- What should M and $\mathbb{E}\mu$ be for the residences example? What are a and b?

Matching the Prior Predictive Distribution in Stan

```
functions { /* saved as residences_rng in R's working directory */
  vector residences_rng(int S, real M, real Emu, int[] y) { // y is a 1D integer array
    real b = 1 / (Emu - M); real a = Emu * b;
    int sum_y = sum(y); // must declare everything, such as ...
    vector[S] post;
    int s = 1;
    if (a < 1 || b < 0) reject("M and Emu are inconsistent"); // before evaluating anything
    while (s <= S) {
      real mu_tilde = gamma_rng(a, b); // but get to make new declarations after any opening {
      int y_tilde[size(y)]; // holds prior predictive distribution for entire sample
      for (n in 1:size(y)) y_tilde[n] = poisson_rng(mu_tilde);
      if (sum_y == sum(y_tilde)) {
        post[s] = mu_tilde;
        s += 1;
      }
    }
    return sort_asc(post);
  }
}
```

Posterior vs. Prior CDF

```
rstan::expose_stan_functions("residences_rng.stan")
M <- 2.5; Emu <- 2.7; b <- 1 / (Emu - M); a <- Emu * b
y <- c(3, 1, 0, 2, 2, 4, 1, 3, 2, 1); S <- 10000; post <- residences_rng(S, M, Emu, y)
plot(post, (1:S) / S, type = "l", xlim = c(0, 5), ylab = "CDF of mu")
curve(pgamma(mu, a, b), lty = 2, col = 2, add = TRUE, xname = "mu")
legend("topleft", legend = c("Posterior", "Prior"), lty = 1:2, col = 1:2, bg = "lightgrey")
```



Posterior Kernel in Stan

```
functions { /* saved as kernel.stan in R's working directory */
  vector kernel(vector mu, real M, real Emu, int[] y) {
    real b = 1 / (Emu - M);
    real a = Emu * b;
    int K = rows(mu);
    vector[K] log_mu = log(mu);
    vector[K] log_prior =     (a - 1) * log_mu - mu * b;       // omits normalizing constant
    vector[K] log_likelihood = sum(y) * log_mu - mu * size(y); // omits sum of log-factorials
    if (a < 1 || b < 0) reject("M and Emu are inconsistent");
    return exp(log_prior + log_likelihood);
  }
}
```

```
rstan::expose_stan_functions("kernel.stan")
(denom <- integrate(kernel, lower = 0, upper = Inf, M = M, Emu = Emu, y = y)$value)
```

```
## [1] 0.0002772902
```

Analytical Posterior PDF

- Gamma prior PDF is again $f\left(\mu \middle| a, b\right) = \frac{b^a}{\Gamma(a)} \mu^{a-1} e^{-b\mu}$
- Poisson PMF for N observations is again $f\left(y_1, \ldots, y_n \middle| \mu\right) = e^{-N\mu} \mu^{\sum_{n=1}^N y_n} \prod_{n=1}^N \frac{1}{y_n!}$
- Posterior PDF, $f\left(\mu \middle| a, b, y_1, \ldots, y_n\right)$, is proportional to their product:

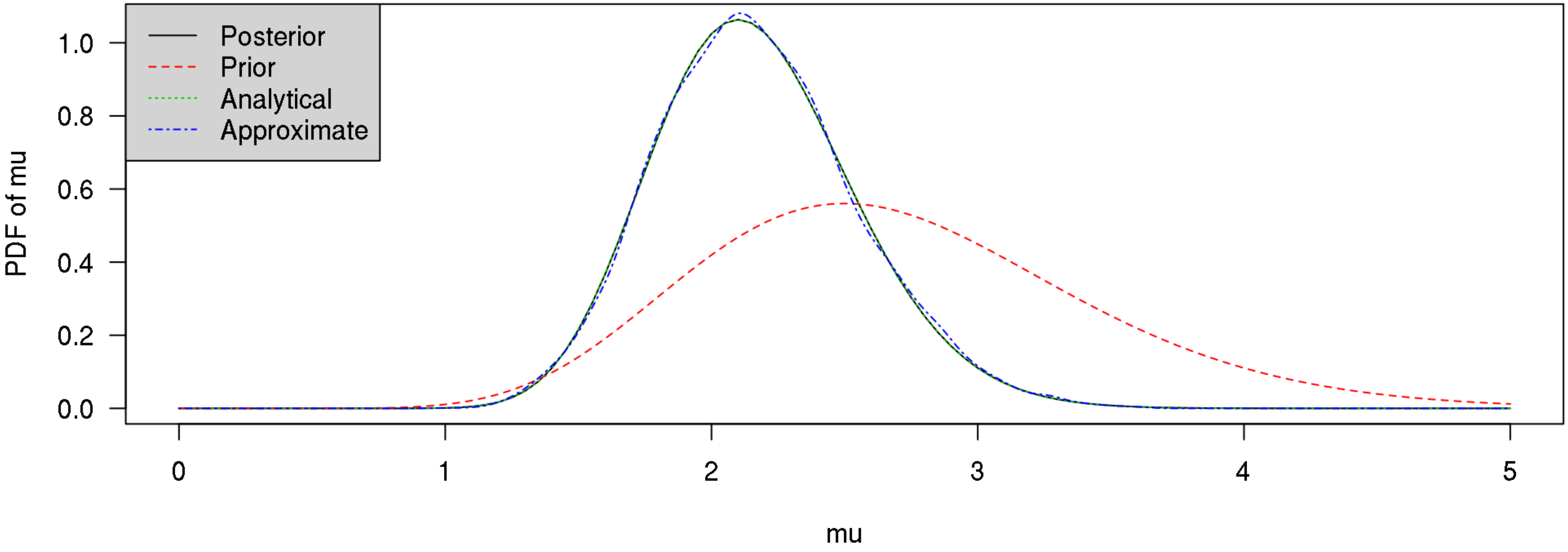
$$\mu^{a-1} e^{-b\mu} \mu^{\sum_{n=1}^N y_n} e^{-N\mu} = \mu^{a-1+\sum_{n=1}^N y_n} e^{-(b+N)\mu} = \mu^{a^*-1} e^{-b^*\mu},$$

where $a^* = a + \sum_{n=1}^N y_n$ and $b^* = b + N$

- Ergo, the posterior has a Gamma kernel and the normalizing constant is $\frac{(b^*)^{a^*}}{\Gamma(a^*)}$

Posterior vs. Prior PDF

```
curve(kernel(mu, M, Emu, y) / denom, from = 0, to = 5, ylab = "PDF of mu", xname = "mu")
curve(dgamma(mu, a, b), lty = 2, col = 2, add = TRUE, xname = "mu")
curve(dgamma(mu, a + sum(y), b + length(y)), lty = 3, col = 3, add = TRUE, xname = "mu")
lines(density(post, from = 0, to = 5), lty = 4, col = 4)
legend("topleft", legend = c("Posterior", "Prior", "Analytical", "Approximate"),
       lty = 1:4, col = 1:4, bg = "lightgrey")
```


Posterior
Prior
Analytical
Approximate
PDF of mu
mu