SAMPLING PROBABILITY DISTRIBUTIONS

From conjugacy to Hamiltonian Monte Carlo

August 19, 2019

Jon Fintzi

Biostatistics Research Branch

National Institute of Allergy and Infectious Diseases

National Institutes of Health

Happy Monday-Funday!

Last time:

- Bayesian inference *always* starts with a model for the **joint distribution** of θ and y:.
$$\pi(\theta, y) = f(y|\theta)\pi(\theta) = \pi(\theta|y)m(y).$$
    - $\pi(\theta|y)$ is the **posterior distribution** of θ given y,
    - $f(y|\theta)$ is the **sampling distribution** for y given θ,
    - $\pi(\theta)$ is the **prior distribution** of θ,
    - $m(y)$ is the **marginal distribution** of y.
- **Bayes rule** yields the **posterior distribution**
$$\pi(\theta|y) = \frac{f(y,\theta)}{m(y)} = \frac{f(y|\theta)\pi(\theta)}{m(y)} \propto Likelihood \times Prior.$$
.
- All of the information used in the *update* to our prior is encoded in the **likelihood**,
$$L(\mathbf{y}|\theta) = \prod_{i=1}^{N} f(y_i|\theta).$$
    - *Likelihood principle*: implies proportional likelihoods encode equivalent updates for a single observer.
    - Two people can have different epistemic uncertainty (different priors).
    - The likelihood principle does not imply equivalent Bayesian inferences (corollary to Gelman, 2017).

Lecture 10 of Statistical Rethinking

Key takeaways:

- Bayes is all about the posterior distribution, not how you compute it.
- Sometimes, we can't get the posterior analytically, but we can approximate it by sampling.
- Samples also give us a way to approximate the distributions of complicated functionals of the posterior.
- Markov Chain Monte Carlo is one way to sample.
    - Metropolis/Metropolis-Hastings.
    - Hamiltonian Monte Carlo.

This Week

Iterations on Bayesian analysis of binomial data

- Motivating example — PREVAIL II Trial.
- Analysis with conjugate priors, beta-binomial model.
- Prior selection.
- Analysis with non-conjugate priors.
- First look at Stan if there's time.

Motivating Example — PREVAIL II Trial

Context:

- 2014–2016 Ebola virus disease (EVD) outbreak in Guinea, Liberia, and Sierra Leone.
- Over 28,000 suspected or confirmed cases and 11,000 fatalities.
- Urgent need to identify effective theraputics to reduce mortality.

Partnership for Research on Ebola Virus in Liberia (PREVAIL) II trial:

- Adaptive trial to determine the effectiveness of Zmapp, and possibly other agents, in reducing Ebola mortality.
- Primary endpoint: 28 day mortality on optimized standard of care (oSOC) vs. Zmapp + oSOC.
- 72 patients enrolled at sites in Liberia, Sierra Leone, Guinea, and the US.
  - Overall mortality: 21/71 died (30%),
  - SOC alone: 13/35 (37%),
  - Zmapp + SOC: 8/36 (22%).
- ~~Super-duper~~ Barely Bayesian design (Proschan, 2016).

Motivating Example — PREVAIL II Trial

Target of inference: $\pi(p_T, p_C | y_T, y_C)$, the posterior distributions for probability of death on treatment (T) and control (C).

- p_T, p_C: probabilities of 28 day mortality on T and C.
- y_T, y_C: # of deaths on T and C.
- N_T, N_C: # participants randomized to T and C.

Some questions of interest:

- Evidence for Zmapp + oSOC more effective than oSOC alone: $\Pr(p_T < p_C | y_T, y_C)$.
- Effectiveness of Zmapp + oSOC, effectiveness of oSOC alone: $\pi(p_T | y_T),\ \pi(p_C | y_C)$.

A Simple Model for Count Data

Binomial count model:

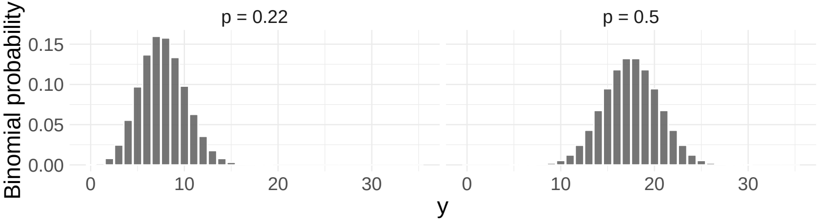
- Arises as a model for *independent* binary random variables (RVs), $Z_i \in \{0, 1\},\ i = 1, \ldots, N$, with *common success probability*, p.
- Let $Y = \sum_{i=1}^{N} Z_i$. The probability of seeing $Y = y$ successes in N trials is

$$\Pr(Y = y|p) = \binom{N}{y} p^y (1-p)^{N-y}. \qquad (1)$$
$$\propto p^y (1-p)^{N-y}$$

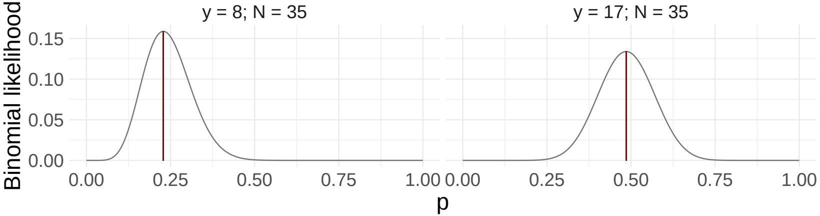
- For fixed y, we can view (1) as a function of p – this is the **likelihood function**.
- The maximum likelihood estimate (MLE), $\widehat{p} = y/N$, is the value of p under which the observed data are most likely (i.e., $\widehat{p}$ maximizes the likelihood).

A Simple Model for Count Data

Binomial distributions for two values of p



Binomial likelihoods for two datasets
Likelihoods in black, MLEs in red



Beta Distribution as a Prior for a Binomial Probability

Beta distribution

- If we though all values of p were equally likely, could take $p \sim \mathrm{Unif}(0, 1)$. In general, this is too restrictive.
- More flexible: $\theta \sim \mathrm{Beta}(a, b), \text{ with } a > 0, b > 0$, where

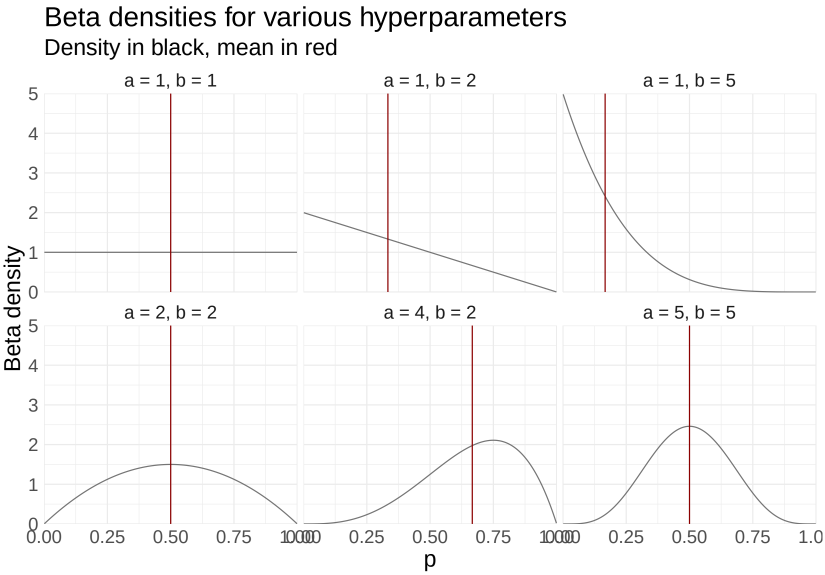
$$\begin{aligned} \pi(\theta|a, b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma b} p^{(a-1)}(1-p)^{b-1}, \\ &\propto p^{(a-1)}(1-p)^{b-1}, \end{aligned} \tag{2}$$

  for $0 < p < 1$ and where $\Gamma(\cdot)$ is the gamma function[1].
- $p \sim \mathrm{Unif}(0, 1)$ is equivalent to $p \sim \mathrm{Beta}(1, 1)$.
- Moments:

$$\begin{aligned} \mathrm{E}(p|a, b) &= \frac{a}{a+b}, \\ \mathrm{Var}(p|a, b) &= \frac{ab}{(a+b)^2(a+b+1)}. \end{aligned}$$

[1] $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\mathrm{d}t$, more on the Beta distribution here.

Beta Distribution as a Prior for a Binomial Probability



Posterior Derivation

In the Beta-Binomial hierarchy, concentrate only on terms that involve θ.

$$\begin{aligned}
\pi(p|y) &\propto L(y|p)\pi(p), \\
&= p^{y}(1-p)^{N-y} \times p^{a-1}(1-p)^{b-1}, \\
&= p^{y+a-1}(1-p)^{N-y+b-1}, \\
&= p^{\widetilde{a}-1}(1-p)^{\widetilde{b}-1},
\end{aligned}$$

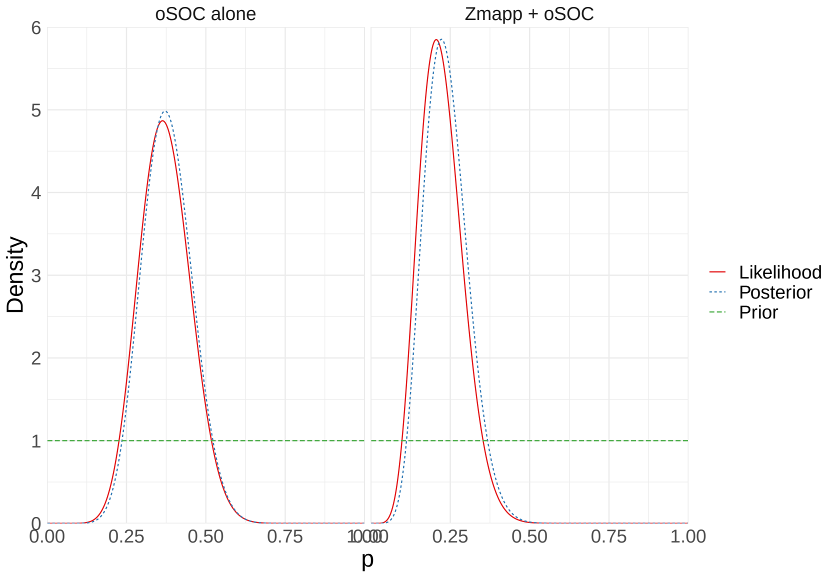
where $\widetilde{a} = y + a$ and $\widetilde{b} = N - y + b$.

- The posterior takes the form of a $\mathrm{Beta}(\widetilde{a}, \widetilde{b})$!
- We say the prior is *conjugate* when the posterior is of the same form as the prior.
- Fun fact: all exponential family distributions have conjugate priors!

PREVAIL II Posterior Distributions

- Priors: $p_T \sim \mathrm{Beta}(1,1)$ and $p_C \sim \mathrm{Beta}(1,1)$.
- Data: $y_T = 8$ and $y_C = 13$, with $N_T = 36$ and $N_C = 35$.
- Posteriors: $p_T|y_T \sim \mathrm{Beta}(9,29)$ and $p_C|y_C \sim \mathrm{Beta}(14,23)$.
  - Posterior medians (95% Credible Intervals):
    - Zmapp + oSOC, $p_T|y_T$ 0.23 (0.12, 0.38),
    - oSOC alone, $p_C|y_C$: 0.38 (0.23, 0.54).
    - Risk difference, $p_T - p_C \mid y_T, y_C$: -0.14 (-0.34, 0.06).
    - Risk ratio, $p_T/p_C \mid y_T, y_C$: 0.62 (0.29, 1.24).
    - Odds ratio, $[(p_T/(1-p_T)) \;/\; (p_C/(1-p_C))] \mid y_T, y_C : 0.50(0.18, 1.36)$
    - $\Pr(p_T < p_C|y_T, y_C) \approx 0.91$.

PREVAIL II Posterior Distributions



Posterior Mean and Likelihood-Prior Interaction

- Recall the mean of a Beta(a, b) is $a/(a+b)$.
- The posterior mean of a Beta$(y+a, N-y+b)$ is therefore

$$
\begin{aligned}
\mathrm{E}(p|y) &= \frac{y+a}{N+a+b} \\
&= \frac{y}{N+a+b} + \frac{a}{N+a+b} \\
&= \frac{y}{N} \times \frac{N}{N+a+b} + \frac{a}{a+b} \times \frac{a+b}{N+a+b} \\
&= \mathrm{MLE} \times \mathrm{W} + \mathrm{PriorMean} \times (1-\mathrm{W}),
\end{aligned}
$$

  where the *weight* W is $\mathrm{W} = \frac{N}{N+a+b}$.
- As N increases, the weight tends to 1, so that the posterior mean gets closer to the MLE.
- Notice that the uniform prior $a = b = 1$ gives a posterior mean of $\mathrm{E}(p|y) = \frac{y+1}{N+2}$.

Choosing Prior Hyperparameters

How to specify hyperparameters a and b?

- *Suggestion #1:* Use information about prior mean prior “sample size.”
    - Prior mean: m_{prior} = a/(a+b)$.
    - Recall, $\mathrm{E}(p|y) = \frac{y+a}{N+a+b}$, so the denominator is like the posterior sample size, $\implies N_{prior} = a + b.$.
    - Solve for a and b via

$$a = N_{prior} \times m_{prior},$$
$$b = N_{prior} \times (1 - m_{prior}).$$

    - *Intuition*: view a and b as pseudo-observations of successes and failures.
- *Suggestion #2:* Choose a and b by specifying *two quantiles* for p associated with prior probabilities.
    - e.g., $\Pr(p < 0.2) = 0.1 and \Pr(p > 0.6) = 0.1$.
    - Can find values of a and b numerically.
    - In more complicated models, simulate.

How to Specify Priors in General?

Theme: What aspects of my model do I know something about? How do I encode that knowledge?

- **Containment:** Does my prior predictive distribution produce realistic datasets?
- **Caveat:** People who don't interrogate and justify their priors deserve what's coming to them.
  - Table of priors with references.
  - Prior predictive checks.
  - Sensitivity analyses.

Issues with Uniformity

We might think that if we have little prior opinion about a parameter then we can simply assign a uniform prior, i.e. a prior $p(\theta) \propto \text{constant}.$

There are two problems with this strategy:

- We can't be uniform on all scales since, if $\phi = g(\theta)$:

$$\underbrace{p_\phi(\phi)}_{\text{Prior for } \phi} = \underbrace{p_\theta(g^{-1}(\phi))}_{\text{Prior for } \theta} \times \underbrace{\left|\frac{d\theta}{d\phi}\right|}_{\text{Jacobian}}$$

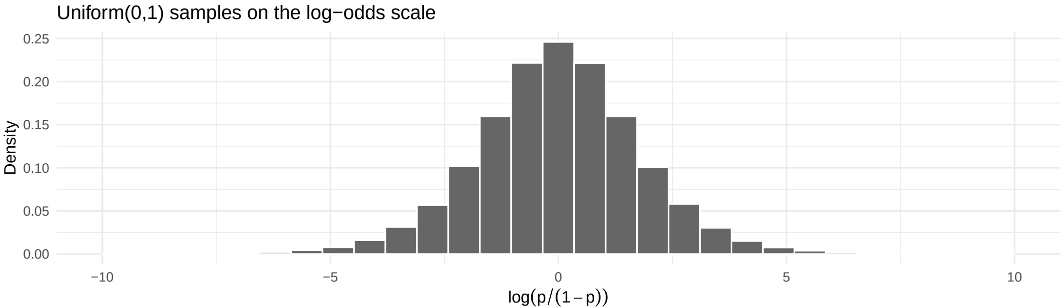
  and so if $g(\cdot)$ is a nonlinear function, the Jacobian will be a function of ϕ and hence not uniform (more on this in a bit).
- If the parameter is not on a finite range, an improper distribution will result (that is, the form will not integrate to 1). This can lead to all kinds of paradoxes (see e.g., Dawid, 1973).
- And importantly, improper priors are non-generative $\implies$ cannot interrogate their predictive distribution.

Are Priors Really Uniform?

- In the binomial example, $p \sim Unif(0,1)$ seems a natural choice.
- But suppose we are going to model on the logistic scale so that

$$\phi = \log\left(\frac{\theta}{1-\theta}\right)$$

is a quantity of interest. -A uniform prior on θ produces the very non-uniform distribution on ϕ. -Not being uniform on all scales is not a problem, and is correct probabilistically, but one should be aware of this characteristic.

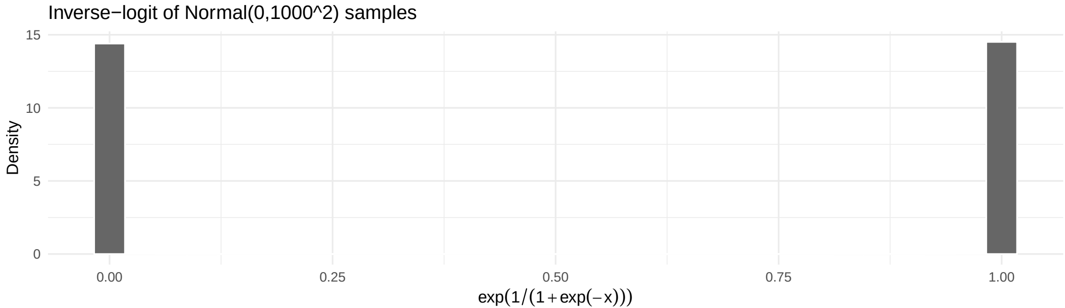


Are Priors Really Uniform?

- In the binomial example, $p \sim Unif(0,1)$ seems a natural choice.
- But suppose we are going to model on the logistic scale so that

$$\phi = \log\left(\frac{\theta}{1-\theta}\right)$$

is a quantity of interest. -A uniform prior on θ produces the very non-uniform distribution on ϕ. -Not being uniform on all scales is not a problem, and is correct probabilistically, but one should be aware of this characteristic.



Non-Conjugate Priors

Suppose we want to model mortality on the log-odds scale, $\theta = \log(p/(1-p))$.

Bayesian inference *always* starts with a model for the **joint distribution** of θ and y.

- The parameter in our model is θ.
- Lose conjugacy, no closed form for the posterior, now we rely on MCMC.
- Our MCMC targets the posterior $\pi(\theta|y) \propto \pi(\theta, y) = L(y|\theta)\pi(\theta)$.
- If our prior is on the log-odds of death, we have no problems. It does not matter that $L(y|\theta) = Binomial(N, 1/(1+exp(-\theta)))$.
- If our prior is on the probability of death but our model is defined in terms of the log-odds, we must include a Jacobian adjustment.

Critical: We must never lose sight of how our model is defined.

For more on this, see case studies here and here study.

Why Non-Conjugate Priors?

- Information encoded naturally on other scales.
- More flexible/natural representation using other types of distributions.
- Hierachical information.
- Compuational considerations.
- Induce particular features in the posterior, e.g., sparsity.

Next week

Linear regression. Watch lecture 3 (SmaRt).

We'll talk about:

- Bayesian linear regression.
- Weekly informative priors.

References

P.A. Dawid, M. Stone, and J.V. Zidek. “Marginalization paradoxes in Bayesian and structural inference.” *Journal of the Royal Statistical Society: Series B (Methodological)* 35.2 (1973): 189-213.

A. Gelman, D.A. Simpson, and M. Betancourt. “The prior can often only be understood in the context of the likelihood.” *Entropy* 19.10 (2017): 555.

The PREVAIL II Writing Group and Multi-National PREVAIL II Study Team. “A randomized, controlled trial of ZMapp for Ebola virus infection.” *The New England Journal of Medicine* 375.15 (2016): 1448.

M.A. Proschan, L.E. Dodd, and D. Price. “Statistical considerations for a trial of Ebola virus disease therapeutics.” *Clinical Trials* 13.1 (2016): 39-48.