

# ESSENTIALS OF BAYESIAN MODELING

Course overview and introduction to Bayesian statistics

August 12, 2019

**Jon Fintzi**

Biostatistics Research Branch

National Institute of Allergy and Infectious Diseases

National Institutes of Health

## Overview

- Introduction to Bayesian inference
- Linear and generalized linear models
- Hierarchical models
- Prior selection and model parameterization
- Model selection
- Model criticism, diagnostics, and visualization
- Markov chain Monte Carlo
- Bayesian workflow

In other words, what are Bayesian models? How can we use them to do cool science things? How can we interrogate their limitations?

**Website:** [https://github.com/fintzij/BRB\\_Bayes\\_course](https://github.com/fintzij/BRB_Bayes_course)

# Hi everybody!

## Materials

- Lectures: Richard McElreath's Statistical Rethinking (StaRt), Winter 2019.
- Software: leaning towards raw **Stan**, but **brms/RStanArm** also an option.
- Additional books, papers, and case studies linked on website.

## Plan

- At home: watch ~2 lecture videos (approx. 2hrs/wk, 1 hour at 2x speed).
- When we meet: summarize, supplement with examples, case studies, and additional material.

**VERY IMPORTANT:** I want this to be useful. Tell me if I'm going too fast/slow, can explain something more clearly, or there's something you to cover. Don't tell me if my jokes are lame.

# This week - “Basics” of Bayesian inference

Lectures 1 and 2 of StaRt touched on the following topics:

- Statistical procedures for interpreting natural phenomena.
  - Models as golems.
  - *Core idea*: Count all the ways data can happen, according to assumptions. Assumptions with more ways that are consistent with data are more plausible.
- Small world (model) vs. large world (real world).
- Workflow: design model (globe) → collect then condition on data (throw it around, compute posterior) → evaluate (sample to summarize, simulate to critique).

# This week - “Basics” of Bayesian inference

Lectures 1 and 2 of StaRt touched on the following topics:

- Statistical procedures for interpreting natural phenomena.
  - Models as golems.
  - *Core idea*: Count all the ways data can happen, according to assumptions. Assumptions with more ways that are consistent with data are more plausible.
- Small world (model) vs. large world (real world).
- Workflow: design model (globe) → collect then condition on data (throw it around, compute posterior) → evaluate (sample to summarize, simulate to critique).

I want to talk more about

- Bayesian inference and how it differs from frequentist inference.
- What is a Bayesian model?

# Quantifying uncertainty

Dicing with the unknown (T. O'Hagan, 2004):

- Two kinds of uncertainty: *aleatory* and *epistemic*.
  - Aleatory: due to randomness, e.g., outcomes of rolling dice.
  - Epistemic: uncertainty about things one could know, but doesn't in practice, e.g., disease risk factors for participants in a clinical trial.
  - *N.B.* Two people may have different epistemic uncertainty about the same question.
- Two definitions of probability: *frequency* and *degree of belief*.
  - Frequency of occurrence under infinite replication, describes aleatory uncertainty.
  - For Bayesians, probability represents the degree of belief about a proposition and may describe both aleatory and epistemic uncertainty.
- Implications:
  - p-values and CIs are statements about aleatory uncertainty.
  - Bayesians quantify uncertainty using probability distributions conditioned on the data.
  - "Bayesian statistics is about the statistician, for whatever reason they may have, guessing or estimating the distribution of the next outcome" (Walker, 2013).
  - Bayesians quantify uncertainty about parameters,  $\theta$ , given data,  $y$ , in the *posterior*,  $\pi(\theta|y)$ .

Some notation:

- $\theta$ : unobserved parameter, e.g.,  $\theta = \Pr(Y = \text{heads})$ .
- $y$ : observed data, e.g.,  $y \in \{\text{heads}, \text{tails}\}$ .
- $\tilde{y}$ : unknown but possibly observable quantities, e.g., future data.

Bayesian inference *always* starts with a model for the *joint* probability distribution of  $\theta$  and  $y$ :

$$\pi(\theta, y) = f(y|\theta)\pi(\theta).$$

- $f(y|\theta)$  is the *sampling distribution* for  $y$  given  $\theta$ .
- $\pi(\theta)$  is the prior distribution of  $\theta$ .

**Bayes rule** yields the *posterior density*

$$\pi(\theta|y) = \frac{f(y, \theta)}{\pi(y)} = \frac{f(y|\theta)\pi(\theta)}{\pi(y)},$$

where  $\pi(y) = \int \pi(y, \theta) d\theta = \int f(y|\theta)\pi(\theta) d\theta$ .



At various points, we will be interested in the following distributions:

- *Prior distribution*:  $\pi(\theta)$ .
- *Sampling distribution*:  $f(y|\theta)$ .
- *Joint distribution*:  $\pi(y, \theta) = f(y|\theta)\pi(\theta)$ .
- *Marginal distribution*:  $\pi(y) = \int \pi(y, \theta) d\theta = \int f(y|\theta)\pi(\theta) d\theta$ .
- *Posterior distribution*:  $\pi(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\pi(y)} \propto f(y|\theta)\pi(\theta)$ .
- *Posterior predictive*:  $f(\tilde{y}|y) = \int f(y|\theta)\pi(\theta|y) d\theta$ .

The term  $\pi(y) = \int f(y|\theta)\pi(\theta)d\theta$  is a normalizing constant.

- Sometimes solvable analytically, e.g., conjugate priors.
- Can be difficult to evaluate without conjugacy, especially in high dimensions.
- Markov chain Monte Carlo (next time)  $\implies$  don't evaluate explicitly:

$$\pi(\theta|y) \propto f(y|\theta)\pi(\theta)$$

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

# Why Bayes?

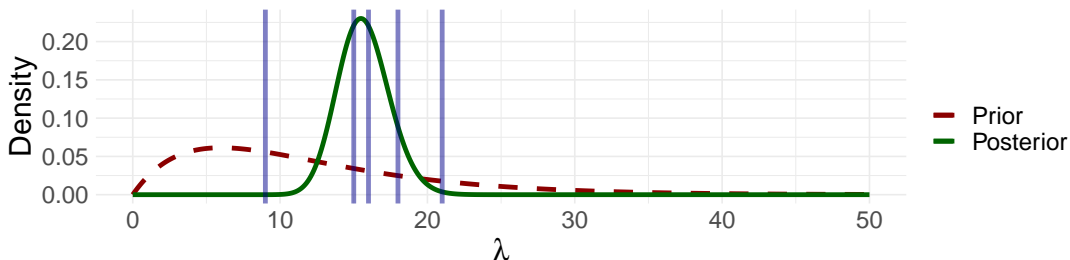
- Incorporate prior information.
- Sample to summarize: all possible inferences computable from the posterior.
- Flexibility: specify prior + likelihood, turn the crank.
- Self-consistent framework for handling missing data.
- Diagnostics.

## Example

Let  $Y$  = number of earthquakes of magnitude  $>4.0$  per year in southern California.

- Model:  $Y|\lambda \sim \text{Poisson}(\lambda)$ , where  $\lambda$  = rate of earthquake occurrence.
- Prior:  $\lambda \sim \text{Gamma}(\alpha, \beta) \implies E(\lambda) = \alpha/\beta$ ,  $SD(\lambda) = \sqrt{\alpha}/\beta$ . Set  $\alpha, \beta$  s.t.,  $E(\lambda) = 12$ , and  $SD(\lambda) \approx 8.5$ .
- Data: 5 years, observe 21, 9, 15, 16, and 18 earthquakes w/magnitude  $> 4.0$  in each of the years.
- Posterior:  $\lambda|y \sim \text{Gamma}(\alpha + \sum_{i=1}^5 y_i, \beta + 5)$ .

Gamma(  $\alpha = 2$ ,  $\beta = 1/6$ ) prior for  $\lambda$



Skipping ahead a bit. Watch lecture 10 on MCMC (SmaRt).

We'll talk about:

- Sampling from probability distributions,
- Markov chain Monte Carlo,
- Stan.

# REFERENCES

T. O'Hagan "Dicing with the unknown." *Significance* 1.3 (2004): 132-133.

S.G. Walker "Bayesian inference with misspecified models." *Journal of Statistical Planning and Inference* 143.10 (2013): 1621-1633.

This lecture also borrowed material from Vladimir Minin's [MCMC for infectious diseases](#) short course, and from Aki Vehtari's [Bayesian data analysis](#) course. These are both fantastic resources and you should check them out.