# Bayesian Hierarchical Modeling in JAGS

Marcel Niklaus

July 16, 2015

# Not content of this talk

- Bayesian Hypothesis test
- Model comparison
- Too many details

# Why Bayes: Because it's what you want to know

- We are actually interested in the probability of a model given data
- p-values are based on probability of (unobserved) data given model (conceptually hard)

# Why Bayes: Because it's what you want to know

- We are actually interested in the probability of a model given data
- p-values are based on probability of (unobserved) data given model (conceptually hard)
- Surprise exercise; think of a situation you are interested in data given model!

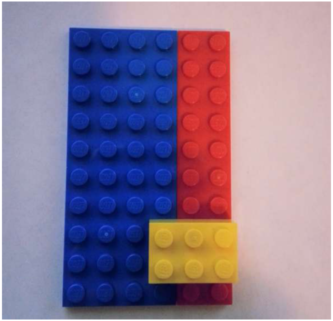
# Why Bayes: Because it's what you want to know

- We are actually interested in the probability of a model given data
- p-values are based on probability of (unobserved) data given model (conceptually hard)
- Surprise exercise; think of a situation you are interested in data given model!
- Bayes rule helps you to get from $p(data|model)$ to $p(model|data)$

# Bayes versus Likelihood approach

- Maximum likelihood assumption: There is a true fixed value of $\theta$. We maximize the likelihood to estimate it with a certain uncertainty (SE, CI) based on sampling.
- Bayesian way: $\theta$ is a random variable. It has a fixed value, but we reflect our uncertainty about it.
- We summarize the posterior distribution

# Formal Bayes Theorem



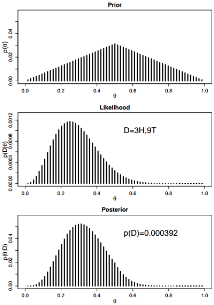
- Bayes rule is rooted in conditional probability (and is uncontroversial).
- $P(Red|Yellow) = P(Yellow|Red) * \frac{P(Red)}{P(Yellow)}$
- $P(Red|Yellow) = (4/20) * (20/60)/(6/60)$
- $2/3 = 1/5 * 1/3 * 10$

# Formal Bayes Theorem (only slide with many formulas)

- $P(R|Y) = P(Y|R) * \frac{P(R)}{P(Y)}$
- Yellow = data, red = hypothesis $\theta$
- $p(\theta|data) = \frac{p(data|\theta) * p(\theta)}{p(data)}$
- $posterior = \frac{likelihood * prior}{marginal\ \ likelihood}$
- $p(\theta|D) \propto p(D|\theta) * p(\theta)$
- Posterior is the likelihood weighted by the prior
- Bayes Theorem tells us how to rationally revise prior beliefs in light of the data to yield posterior beliefs.

# General Principles of Bayesian Analysis

- Uncertainty (of parameter estimate) is quantified by probability (distributions)
- Observed data is used to update **prior** information to yield **posterior** information
- Bayesian workflow: set prior beliefs $\Rightarrow$ get data $\Rightarrow$ update prior beliefs $\Rightarrow$ summarize posterior beliefs



# Let's play a game: Odd-Even Guess

http://87.106.45.173:
3838/felix/BayesLessons/BayesianLesson1.Rmd

- 6 Trials of Odd-Even Guess )

# Prior

Pick your Prior: What's your ability?

- $\theta$: the rate with which you guess correctly
- $\theta = 1$: you guess correctly all the time
- $\theta = 0$: you are terrible at this game
- $\theta = 0.5$: equal odds
- Uncertainty (of parameter estimate) is quantified by probability (distributions)
- We assume our beliefs can be represented by a beta distribution
- A beta distribution is constraint to lie within 0-1 (perfect for proportion)

Pick your Prior: What's your ability?

- $\theta \sim Beta(1,1)$: Uninformed prior: Do you think (before seeing the data) that it is equally likely that $\theta = 0.01$ and $\theta = 0.5$?
- $\theta \sim Beta(3,3)$ : Slightly informed prior

We are now ready to play: Go!

# Likelihood

The likelihood is the workhorse of Bayesian inference. It represents the data part.

- What's the likelihood we observe the data (y wins given n trials) given the parameter $\theta$ (for each)
- Probability density of the data, considered as a function of $\theta$
- The likelihood of a hypothesis conditions on the data as if they are fixed while allowing the hypotheses (our parameter $\theta$ to vary
- Binomial likelihood: $L(p|n,y) = \binom{n}{y} p^y (1-p)^{n-y}$
- `http://shiny.stat.calpoly.edu/MLE_Binomial/`

# Posterior

After playing

- The posterior distribution summarizes our state of uncertainty about the true value of $\theta$ after having observed the game.

# Markov chain Monte Carlo

- Posterior distribution = Prior distribution $*$ Likelihood density
- Analytical calculation of posterior only possible for simple models (high-dimensionality integration problem)
- Draw samples from posterior and summarize the distribution of those samples (it works!)
- MCMC: algorithm that whose draws are dependent on previous draw. This chain will converge to the posterior.
- Metropolis-Hastings algorithm and the Gibbs sampler (google it!)
- JAGS WinBugs and STAN can do this for you
- Algorithm to approximate an unknown distribution

# JAGS Basics

- $\leftarrow$ is equal to: $y \leftarrow a + b$
- $\sim$ is distributed as..
- Binomial likelihood: $y \sim dbin(\theta, nAttempts)$
- Normal likelihood: $y \sim dnorm(mean, precision)$
- $precision = 1/Var$
- Loop: for (i in 1:3){bla[i]=1+i}
- bla = 2,3,4, bla[2] = 3
- order of commands is irrelevant

# Soccer Shootout

Let's estimate soccer players' ability to score a penalty in a world cup!

- Open Soccer.R in your R console
- Set your working directory (line 8)
- Jags model: ShootoutAbility.txt
- What is their ability $\theta$, [0-1]?
- $Y\theta$ Prior: remember the game
- Credible interval: $\theta$ lies between lower and upper bound with a probability of 95%.

Convergence: Has the MCMC Gibbs sampler converged on posterior (after starting from random values)

- Trace plot: "fat, hairy caterpillar"
- Autocorrelation plot: "Chains should forget previous visits with time": drop off quickly: thinning
- Gelman-Rubin-Brooks diagnostic: Between and within chain variance: 1 indicates convergence (F-value); less than 1.2. See gelman.diag [CODA]
- Geweke test of non-stationarity; Heidelberger-Welch test etc.

# Soccer Shootout: Afrika vs. Europa and America

- Jags model: ShootoutAbilityDifference.txt
- theta[Cindex[i]] can either be theta[1], or theta[2]
- estimates separate parameters for Africa and Europe & America

# Critiques

Bayesians use unjustified priors

- Frequentists use uninformative priors

Subjective priors dominate the results

- data overwhelm the prior with enough n

And also...

- Bayes factors are consistent: With large N, Bayes statistics will tell you if null hypothesis is true

# Hierarchical Modeling

- With nested data (e.g. data for participants is organized on more than 1 level), a multilevel model is appropriate.
- The classic: students grouped in classes, which nest in school districts, which in turn nest in states.
- Many names: Mixed effects modeling, multi-level modeling... Bayesian hierarchical modeling gets rid of these confusions.
- All Bayesian Models are hierarchical because every parameter has a prior.
- All parameters in bayesian models are random effects

# Easy IQ example: IQ measurements

Ignore the grouping variable: 1 $\mu$ for all:

- Scenario: We measure some IQs
- $y \sim N(\mu, \tau)$
- $\mu \sim N(100, 0.004)$
- $\tau \sim Gamma(0.001, 0.001)$

# Intercept varies by group

- Scenario: We measure some IQs
- $y \sim N(\mu[G], \tau)$
- for each Group G :$\mu[G] \sim N(100, 0.004)$
- $\tau \sim Gamma(0.001, 0.001)$
- This estimates G means

# Hierarchical Model

- Instead of assuming a completely different mean for each Group G, we assume that they are drawn from a common Normal distribution
- "Random" means we draw the values from a normal distribution
- $y \sim N(\mu[G], \tau)$
- for each Group G: $\mu[G] \sim N(Hypermean, HyperSD)$
- $Hypermean \sim N(100, 0.0044)$
- $HyperSD \sim Gamma(0.001, 0.001)$

# Linear Mixed Models

- Linear Regression model
- Mixed because it includes coefficients that vary over group (random: participants, items) and some that don't (fixed, treatment group).
- Random Effects: zero mean restriction $N(0, \omega))$
- Random intercepts model: random intercept, fixed slope
- Random intercepts and slope model: random intercept, random slope

In Bayesian Statistics, all parameters are random, i.e. drawn from an overarching distribution!

# Fixed Effects Model: Math grade and IQ

Ignore the grouping variable

- $y \sim N(\mu, \tau)$
- $\mu <- \ x0 + x1 * math$

# Random Intercept

- $y \sim N(\mu, \tau)$
- $\mu < -x0[G] + x1 * math$
- for all groups j: $x0[j] \sim N(hypermean, hypersd)$
- $hypermean \sim N(100, 0.004)$
- $hypersd \sim Gamma(0.001, 0.001)$
- $x1 \sim N(0, 0.001)$

# Random Intercept: 2

- $y \sim N(\mu, \tau)$
- $\mu < -x0 + u0[G] + x1 * math$
- for all groups g: $u0[g] \sim N(0, hypersd)$
- $x0 \sim Uniform(-\infty, \infty)$
- $x1 \sim Uniform(-\infty, \infty)$
- $hypersd \sim Uniform(-\infty, \infty)$

# Random Intercept and Slope

- $y \sim N(\mu, \tau)$
- $mu < -x0[G] + x1[G] * math$
- for all groups g: $x0[g] \sim N(hypermeanintercept, hypersdint)$
- for all groups g: $x1[g] \sim N(hypermeanslope, hypersdslope)$
- $hypermeanintercept \sim N(100, 0.004)$
- $hypermeanslope \sim N(0, 0.001)$
- $hypersdint \sim Gamma(0.001, 0.001)$
- $hypersdslope \sim Gamma(0.001, 0.001)$

# Variance-Covariance Matrix

- if random slopes and random intercepts may not be independent
- Google cholesky decomposition!

# Multi-level Models

- $y \sim N(\mu, \tau)$
- $\mu \sim N(\mu2, \tau2)$
- $\mu2 = z0 + z1 * level2covariate$
- random intercept and slopes can be given if there are even more levels
- $\mu2 = z0[level2Groups] + z1 * level2covariate$

# Simple linear regression

London school match-exam tests and London Reading Test scores

- Open Exam_ple.R in your R console
- Set your working directory (line 7)
- Jags model: ExamSimple.txt

# Random intercept model

- b0[school[i]]
- for all schools: $b0[j] \sim dnorm(school.b0, school.tau)$
- $school.b0 \sim dnorm(0, 0.0001)$
- all j school means are drawn from school.b0

# Random intercept and slope model

- $b0[school[i]] + b1[school[i]] * lrt[i]$
- for all schools: $b0[j] \sim dnorm(school.b0, school.tau)$
- for all schools: $b1[j] \sim dnorm(school.b1, school.tau.b1)$
- school.b0 $\sim$ dnorm(0,0.0001)
- school.b1 $\sim$ dnorm(0,0.0001)
- all j intercepts and slopes are drawn from a Normal with estimated hyperparameters

# Multi-Level Models

Covariates that capture the way groups vary are included

- Level 1: Regression for London Reading Test
- Level 2: Group Level with covariates: Regression for entry score
- b0[j] $\sim$ c0+c1$*$ entry[j]

# Multi-Level Models

Covariates that capture the way groups vary are included

- c0: intercept
- c1: effect of entry (L2) on intercept
- d0: effect of LRT
- d1: effect of entry on LRT effect