# Introduction to Bayesian Inference

Rob W Rankin

Post-doc (Georgetown University),PhD (Murdoch University)

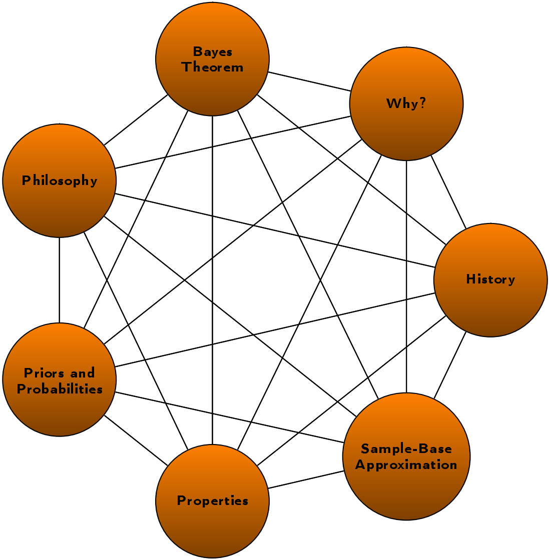
September 30, 2017

# Outline

# themes

see more themes in
/usr/share/texlive/texmf-dist/tex/latex/beamer/themes/theme

# outline



Bayes Theorem
Why?
Philosophy
History
Priors and Probabilities
Sample-Base Approximation
Properties

# Why?

Why are you interested in Bayesianism?

# Why? – some advantages

### Things Ecologist say. . .

- small sample sizes: exact inference
- missing data: easy to impute
- integrate other information
- derived quantities
- complex, hierarchical process models

multiple sources of variation (space,time)
“honest" epistemology

### Theoretical

- conditional on the observed data
- probabilistic statements
- evidential
- coherence, decision making
- good frequentist properties

– shrinkage, decision-theory

- model selection

# Why? – disadvantages

- what is probability (basing inference on something that doesn't exist!!!)?
- objective basis for science?
- misalignment: probability theory and human psychology
- bias (to prior)[1]
- language dependence

[1] Bayesian would claim that, if a prior exists, it would be irrational to believe in anything other than the posterior.

# History

## Neo Bayesian Revival (>1992)

Gelfand and Smith 1992 - Sampled based approximations of Bayesian posteriors [2]

## Revival (~1920s - )

Subjective Bayesian, Decision Theory

- Ramsey (1926)
- De Finetti (1937)
- Savage (1954)
- (Wald, 1939, 1954)

Hypothesis Testing, Logical/Objective Bayesism

- Jeffreys (1939)
- Jaynes (2003)

Hierarchical Bayesian

- Good (1953,65)

Relationship to Compact Coding Theory

- Rissasen (1978), Wallace (1968)

Prediction

# History - Frequentism's Ascendency

## Frequentist "lethal blow"[3]

Rallied against use of prior probabilities in statistical inference

- Ronald A. Fisher (1925,1935,. . . )

Maximum likelihood, significance tesing, ANOVA, sufficiency, randomized experiments

- Jerzy Neyman & Egon Pearson (1933)

Hypothesis testing, confidence intervals

## Falsificationism

- Karl Popper (1959,1963)

[3] S. Zabell 1989

# History - Hypotheses

## Frequentism

### Fisher's p-value

- continuous index of evidence against a hypothesis $H$
- NEVER prove a hypothesis $H$, only disprove

### Neyman-Pearson $\alpha$ and $\beta$

- long-run error rates of Type-I and Type-II
- bound Type-I at $\alpha \leq 0.05$ and hopefully maximize power $(1 - \beta)$ with high $n$ and most powerful tests
- never confirm a hypothesis: only act so as to "not be wrong too often"

## Bayesians

### Model probabilities

probabilistic confirmation of hypotheses

- \$p(H$_k$|$Y$) what is the probability of Hypothesis $H_k$ given the data?

### Bayes Factors

evidence in favour of one hypothesis over another

- $BF = \frac{p(Y|H_1)}{p(Y|H_2)}$.

find hypothesis that is more likely to be true

# History - Inverse Probability

from late 1700's to ~1920's: Method of Inverse Probability,the bread and butter of applied analyses

- Thomas Bayes (1778)
- Laplace (1774)

## Bayes Theorem

conditional probability

$$\overbrace{p(\theta|Y)}^{\text{posterior}} = \frac{\overbrace{p(\theta)}^{\text{prior}}\,\overbrace{\mathcal{L}(Y|\theta)}^{\text{likelihood}}}{\underbrace{f(Y)}_{\text{marginal likelihood}}} \tag{1}$$

- **Prior**: distribution of $\theta$ (before the data)
- **Likelihood**: joint probability density of the data (given $\theta$)
- **Posterior**: distribution of $\theta$ (given the $\theta$)
- $f(Y) \equiv$ marginal likelihood $\int f(Y|\theta)p(\theta)d\theta$

...or more commonly,

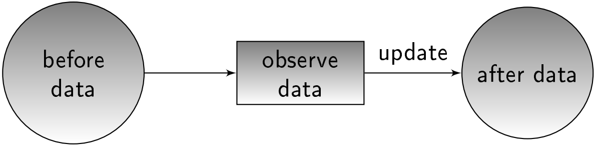
$$p(\theta|Y) \propto f(Y|\theta)p(\theta)$$

where...

$$\begin{aligned}
p(\theta) &\equiv \text{prior information (before the data)} \\
f(Y|\theta) &\equiv \text{ likelihood (information from the data)} \\
p(\theta|Y) &\equiv \text{distibuion of } \theta \text{ after the data} \\
\cancel{f(Y)}\ & \cancel{\text{marginal likelihood}}\text{(often ignore)}
\end{aligned} \tag{2}$$

- posterior is a mixture of information in prior and likelihood

# Bayesian Conditionalization




Prior Likelihood Posterior

# What's in a Posterior

## Mixture of information

- $\mathcal{L}(Y|\theta)$: Likelihood, specified by model. Similar between Bayesian and non-Bayesian analyses[4]
- $p(\theta)$ . . . where do they come from?

## How to specify priors (HUGE topic)

- a previous posterior distribution
- elicitation from experts, previous studies
- Priors as degrees-of-beliefs: **Subjectivist/personalist** Bayesians}
- Default prior and reference priors: **Objective/logical Bayesians**
- adhoc

[4] Frequentists reserve the term likelihood for a function of $\theta$ for fixed y, whereas Bayesians consider "joint probability density of the data" given $\theta$.

# Posterior Inference (for estimation $\theta$)

- Probabilistic statements about abstract quantity ($\theta$) (*only* Bayesians can do)
- Posterior probability necesarily depends on a *prior*

"to make an Omelette, you must crack a few eggs" (Savage)

## The joy of Posterior Inference

can make statements like...

- what is the probability that $\theta > 0$?
- what is the most probable value of $\theta$? (MAP)
- what is the expected value of $\theta$? (posterior mean)
- what is a *high probability region* of $\theta$ (Q% credibility interval)

# Posterior Inference (estimation example)

## Example 1:

- men's height, n=20 observations.
- $y_i \sim \mathcal{N}(175, 10^2)$

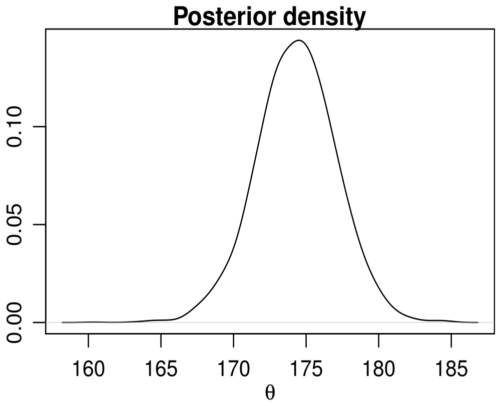
```
y <- c(183.46, 182.32, 178.31, 181.36, 165.12, 185.68, 170.47,
178.11, 174.86, 182.03, 180.09, 172.88, 177.94, 177.26, 182.58,
171, 173.74, 177.78, 180.02, 163.05)
```

- estimate $\theta = [\mu, \sigma^2]$: mean population height and variance
- priors: $p(\mu) = \mathcal{N}(0, 90^2)$, $p(\sigma^2) = \mathcal{IG}(0.1, 0.1)$
- specify a likelihood: $\mathcal{L}(\mathbf{y}|\mu, \sigma^2) = \prod_i^n \mathcal{N}(y_i; \mu, \sigma^2)$

*now run a Gibbs sampler to approximate the posterior* $p(\mu, \sigma^2|\mathbf{y})$ ...

# Posterior density

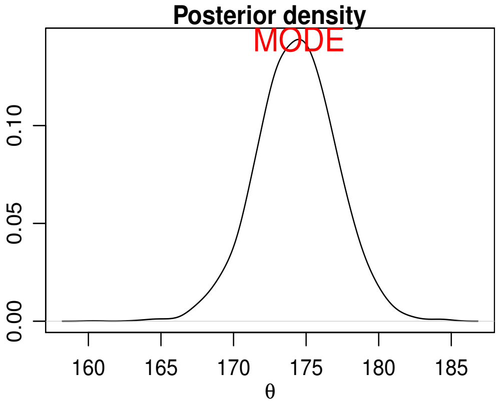
- IS a probability distribution




- easy to interpret

# Posterior density

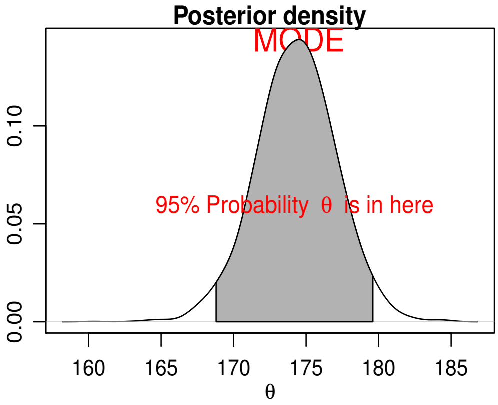
- IS a probability distribution




- Posterior mode: most probable value
- Posterior mean $\mathbb{E}[\theta] = \int p(\theta|Y)\theta d\theta$: expected value

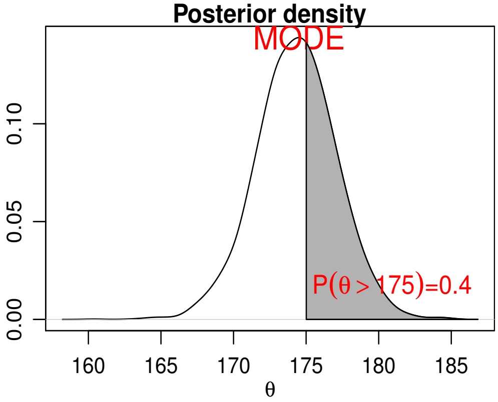
# Posterior density

- IS a probability distribution




- Posterior mode: most probable value
- Posterior mean $\mathbb{E}[\theta] = \int p(\theta|Y)\theta d\theta$: expected value
- 95%CI of $\theta$

- IS a probability distribution




- Posterior mode: most probable value
- Posterior mean $\mathbb{E}[\theta] = \int p(\theta|Y)\theta d\theta$: expected value
- What is the probability that $\theta > X$? Area of $p(\theta|Y) > X$

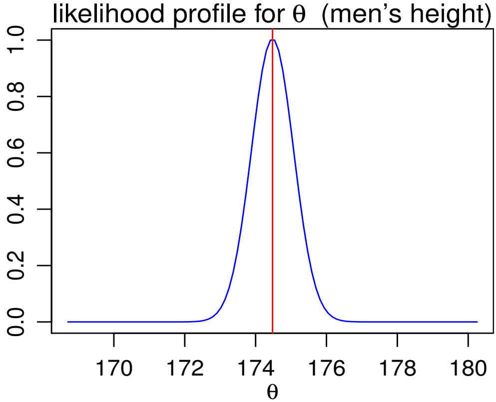
Bayesian vs. frequentist estimates: compare posteriors to maximum-likelihood method

### method of maximum likelihood

- Choose $\theta$ such that we *maximize* the likelihood ($\mathcal{L}$) of seeing **y**
- interpretation "It would be very (un)likely to see the data that I saw, if the value of $\theta$ were X"
- Most common method among Frequentists (single model estimation)
- $\hat{\theta}_{\mathsf{MLE}}$ is NOT the "most probabilty value of $\theta$
- optimality: unbaised, efficient [5]

[5] but see shrinkage estimators for high-dimensional problems

# Posterior Estimation vs Maximum Likelihood Estimation




- frequentist point estimates: `glm(y~1)`

| MLE | se | 95CI-low | 95CI-hi |
|---|---|---|---|
| 172.3 | 1.48 | 169.4 | 175.17 |

- compare to (approximate)[6] Posterior descriptive statistics

| $E[\theta]$ | SD | 95CI-low | 95CI-hi |
|---|---|---|---|
| 172.21 | 1.51 | 169.21 | 175.16 |

nearly the same

[6] ... to Monte Carlo approximation of the posterior

# Posterior Inference (estimation example 2)

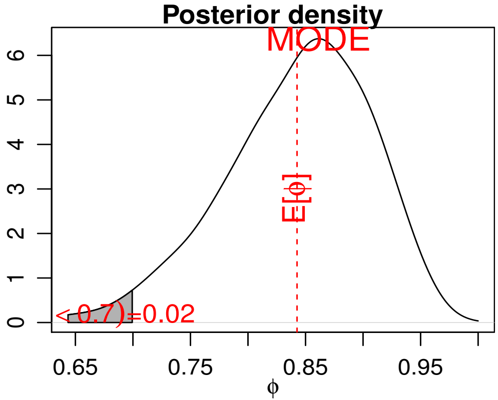
## Example 2:

- survival $[0 = \text{died}, 1 = \text{survived}]$, $n = 30$ observations.
- $s_i \sim \text{Bern}(0.9)$

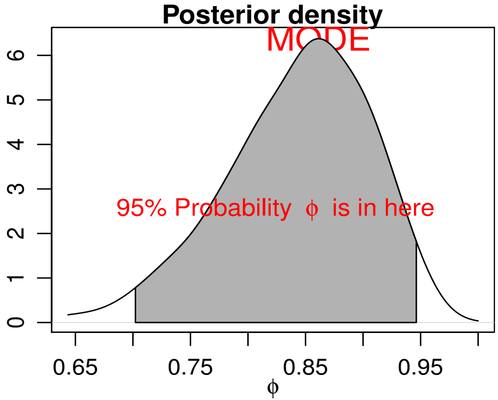
```
s <-
c(1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,0,1,0,1,1,1,1,1,1,1,1,1,1,1,1)
```

- estimate $\theta = [\phi]$: mean population survival
- priors: $p(\phi) = \text{Beta}(1, 1)$
- specify a likelihood: $\mathcal{L}(\mathbf{s}|\phi, n_s) = \prod_s^n \text{Bern}(s_i; \phi, n_s)$

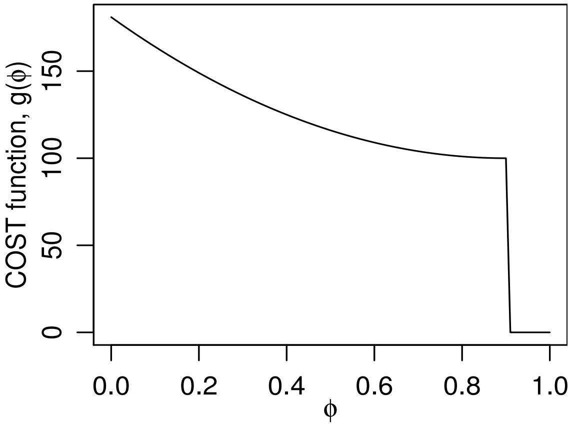
now run a Gibbs sampler to approximate the posterior $p(\phi|\mathbf{s})$


Posterior density
MODE
E[ϕ]
<0.7)=0.02
0
1
2
3
4
5
6
0.65
0.75
0.85
0.95
ϕ

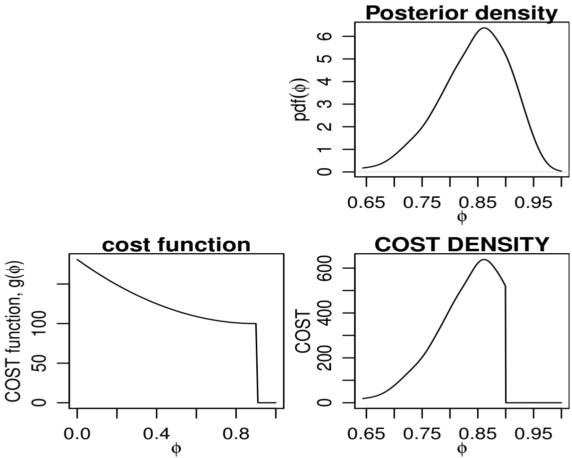
# Posterior density



What if you have a "cost function" $g(phi)$? e.g., cost of conservation action conditional on the estimated values of $\phi$?



# Posterior inference: cost functions




$\mathbb{E}[\text{COST}] = \int_0^1 g(\phi)p(\phi|s)d\phi$

- full cost including all uncertainty in $\phi$

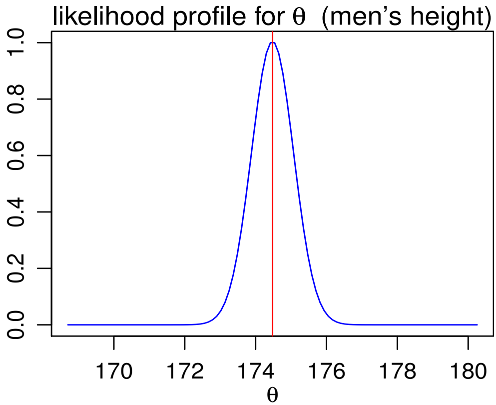
# Posterior Estimation vs Maximum Likelihood Estimation

How do Bayesian posterior estimates compared to more-familiar (frequentist) point-estimates based on maximum likelihood?

## method of maximum likelihood

- Most common method among Frequentists (single model estimation)
- "It would be very (un)likely to have seen the data that I saw, if the value of $\theta$ were X"
- Choose $\theta$: that which *maximize's* the likelihood of seeing $\mathbf{y}$
- $\hat{\theta}_{\mathrm{MLE}}$ is NOT the "most probabilty value of $\theta$
- optimality properties: unbaised, efficient [7]

[7] but see shrinkage estimators for high-dimensional problems




- frequentist point estimates: `glm(y~1)`

| MLE | se | 95CI-low | 95CI-hi |
|---|---|---|---|
| 172.3 | 1.48 | 169.4 | 175.17 |

- compare to (approximate)[8] Posterior descriptive statistics

| $E[\theta]$ | SD | 95CI-low | 95CI-hi |
|---|---|---|---|
| 172.21 | 1.51 | 169.21 | 175.16 |

nearly the same

[8] ... Monte Carlo ... of the posterior

# Posterior Estimation vs Maximum Likelihood Estimation

## for $n$ getting LARGE, and for WEAK priors

- Posterior Mode $\theta_{\text{MAP}} \rightarrow \hat{\theta}_{\text{MLE}}$
- Posterior Confidence Intervals $\rightarrow$ Confidence Intervals

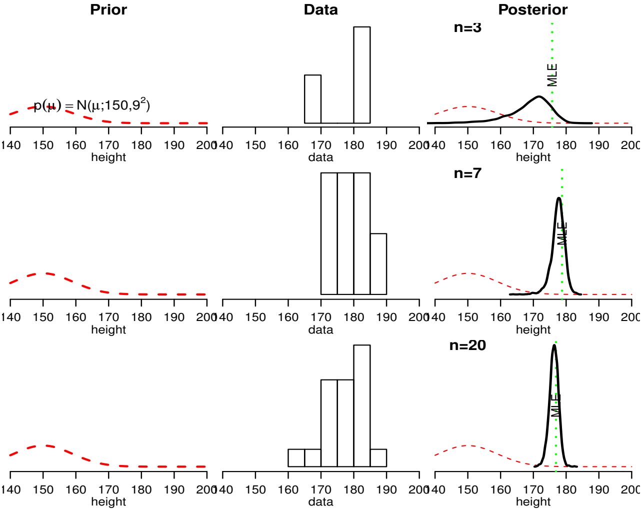
## for low $n$ and/or for STRONG priors

- shrinkage: $\theta \rightarrow$ Prior expectation.
- Posterior mean $\bar{\theta}$ is "biased" towards the priors

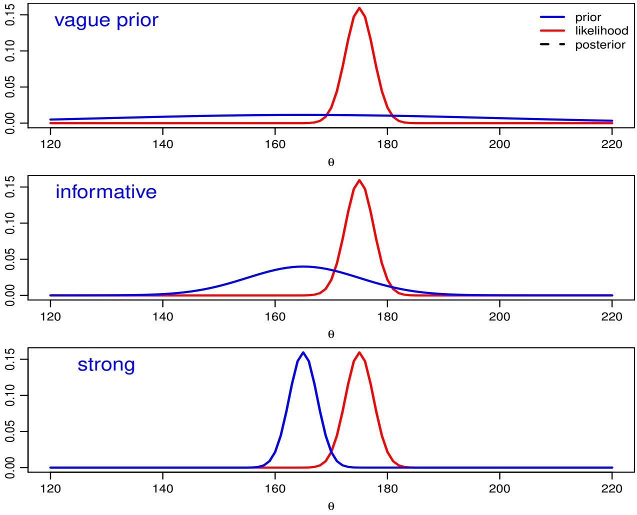
## role of priors (from an estimation perspective)

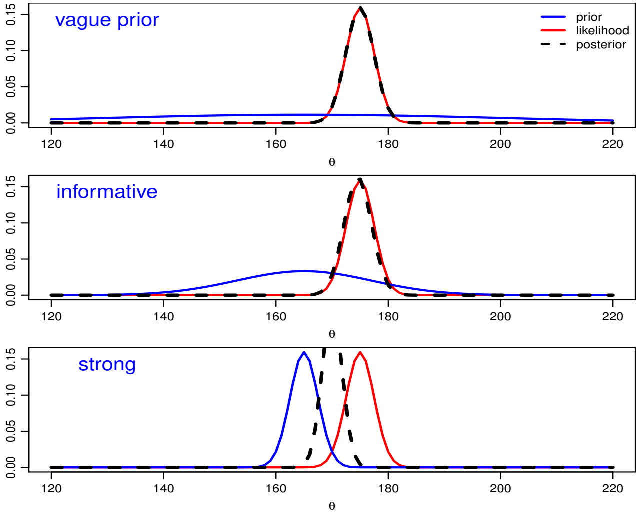
- Priors retard/accerlate rate of convergence of $\bar{\theta} \rightarrow$ truth
- At low samples-sizes, "sensible" priors induce *shrinkage* and have better estimation properties than MLEs
- Key POINTS: you must be a master of prior distributions.

# Posteriors and Sample Size



# Posteriors and Prior information




vague prior
prior
likelihood
posterior
0.00
0.05
0.10
0.15
120
140
160
180
200
220
θ
informative
strong

### Most Biologists are reluctant Bayesians

- **Frequentist** vs. **Bayesian**: often desire that point estimates are identical between posteriors and MLEs
- but, only for: i) weak priors, and ii) large-samples sizes
- key point: **Be a Master of Priors!**

# Priors and Philosophy of Probabilities

### Subjective Personalist Bayesians

Probabilities are your "degree of belief"

- priors: prior beliefs
- posteriors: bring your beliefs into alignment with posterior
- decision making

### Objective Logical Bayesians

Probabilities are continuous extension of Aristolean logic, deductive

- Probabilities capture "degree of truth"
- Priors: non-informative, set by default (Jeffrey's Priors, reference priors, language-invariant priors)

e.g., $p(\phi) = \text{Beta}(0.5, 0.5)$ (Jeffrey's Prior)

- Elicit priors from previous studies (posterior becomes new prior)

# Priors and Philosophy of Probabilities

## Instrumentalist

priors useful for good estimation properties

- shrinkage, efficiency

## Frequentist

Principal principle: “probabilities ($p$(Event)) should align with long-run frequencies of Event"

- probabilities do not exist in reality

## Other

- Quantum mechanics
- Propensities (Karl Popper)

# Probability Distributions in JAGS/BUGS

- you must express your prior information probabilitistically

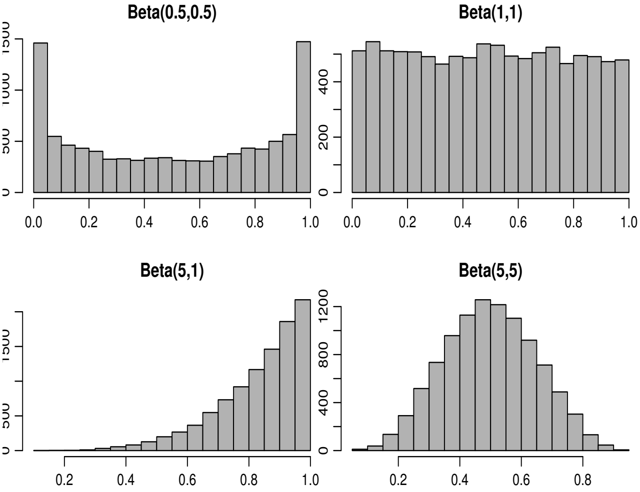
## Know the distributions and their parameters (JAGS Manual)

| Name | Usage | Density | Lower | Upper |
|---|---|---|---|---|
| Beta | `dbeta(a,b)` <br> $a > 0,\ b > 0$ | $\frac{x^{a-1}(1-x)^{b-1}}{\beta(a,b)}$ | 0 | 1 |
| Chi-square | `dchisqr(k)` <br> $k > 0$ | $\frac{x^{\frac{k}{2}-1}\exp(-x/2)}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})}$ | 0 | |
| Double exponential | `ddexp(mu,tau)` <br> $\tau > 0$ | $\tau\exp(-\tau\lvert x-\mu\rvert)/2$ | | |
| Exponential | `dexp(lambda)` <br> $\lambda > 0$ | $\lambda\exp(-\lambda x)$ | 0 | |
| F | `df(n,m)` <br> $n > 0,\ m > 0$ | $\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})}\left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1}\left\{1+\frac{nx}{m}\right\}^{-\frac{(n+m)}{2}}$ | 0 | |
| Gamma | `dgamma(r, lambda)` <br> $\lambda > 0,\ r > 0$ | $\frac{\lambda^r x^{r-1}\exp(-\lambda x)}{\Gamma(r)}$ | 0 | |
| Generalized gamma | `dgen.gamma(r,lambda,b)` <br> $\lambda > 0,\ b > 0,\ r > 0$ | $\frac{b\lambda^{br}x^{br-1}\exp\{-(\lambda x)^b\}}{\Gamma(r)}$ | 0 | |
| Logistic | `dlogis(mu, tau)` <br> $\tau > 0$ | $\frac{\tau\exp\{(x-\mu)\tau\}}{[1+\exp\{(x-\mu)\tau\}]^2}$ | | |
| Log-normal | `dlnorm(mu,tau)` <br> $\tau > 0$ | $\left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} x^{-1}\exp\left\{-\tau(\log(x)-\mu)^2/2\right\}$ | 0 | |
| Noncentral Chi-squre | `dnchisqr(k, delta)` <br> $k > 0, \delta \geq 0$ | $\sum_{r=0}^{\infty}\frac{\exp(-\frac{\delta}{2})(\frac{\delta}{2})^r}{r!}\frac{x^{(k/2+r-1)}\exp(-\frac{x}{2})}{2^{(k/2+r)}\Gamma(\frac{k}{2}+r)}$ | 0 | |
| Normal | `dnorm(mu,tau)` <br> $\tau > 0$ | $\left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}}\exp\{-\tau(x-\mu)^2/2\}$ | | |
| Pareto | `dpar(alpha, c)` <br> $\alpha > 0,\ c > 0$ | $\alpha c^{\alpha}x^{-(\alpha+1)}$ | $c$ | |

- easy to learn in R

e.g., $r \sim \text{Beta}(a, b)$

```
r <- rbeta(10000, 0.5, 0.5)
```



# Sample-based inference

## Posteriors

often no 'analytical' solution to $(\theta|Y)$

## Solution: Sampling

- it is a Probability Distribution!!!
- find a way to sample from posterior
- with enough samples: mean(samples) = Posterior Expectation

assuming $\theta_j \sim p(\theta|y)$ for $j = 1, \ldots, J$

$$\begin{array}{lll}
\text{Expected Value} & = \int \theta p(\theta|y) d\theta & \approx \frac{1}{J} \sum_j^J \theta_j \\
\text{Standard Error}(\theta) & = SE(\theta) & \approx SD(\theta_j) \\
\text{Probability } \theta > 0 & = \int \mathbb{I}[\theta > 0] p(\theta|y) d\theta & \approx \frac{1}{J} \sum_j^J \mathbb{I}[\theta_j > 0]
\end{array}$$

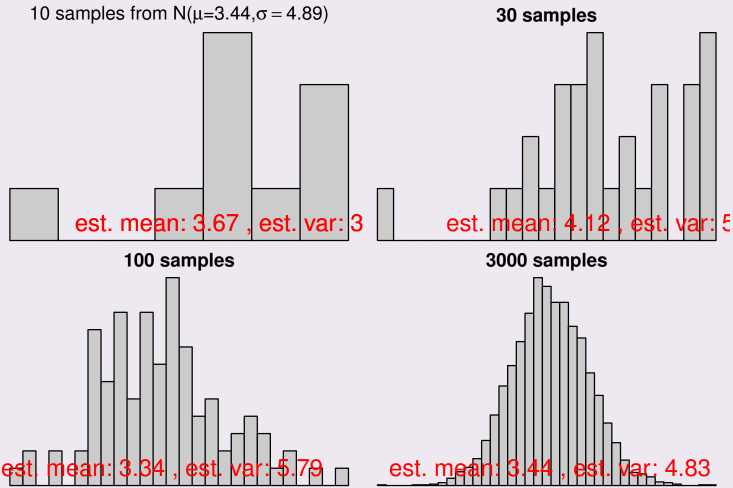
## Sampling Algorithms

MCMC; Gibbs Sampling; Metropolis-Hastings; Slice-Sampling; Importance Sampling; "Conjugate Priors";

## Approximate the joint-posterior distribution"

### example: estimate mean and variance of $\theta$

$\theta_{\text{true}} = 3.44$; $\text{Var}(\theta)_{\text{true}} = 4.89$



# Gibbs Sampling

break-down joint posterior into (simpler) conditional distributions

- difficult: sampling $P(\beta_0, \beta_1, \beta_2, \sigma^2 | Y)$
- easy: sampling $P(\beta_0, \beta_1, \beta_2, | \sigma^2, Y)$ then $P(\sigma^2 | \beta_0, \beta_1, \beta_2, Y)$ then repeat

approximates the joint posterior

## algorithm

- initialize: $\beta_0^{(0)}, \beta_1^{(0)}, \beta_2^{(0)}, \sigma^{2(0)}$

$$\begin{aligned} \{\beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}\} &\sim P(\beta | \sigma^{2(0)}, Y) \\ \sigma^{2(1)} &\sim P(\sigma^2 | \beta_0^{(1)}, \beta_1^{(1)}, \beta_2^{(1)}, Y) \\ \{\beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}\} &\sim P(\beta | \sigma^{2(1)}, Y) \\ \sigma^{2(2)} &\sim P(\sigma^2 | \beta_0^{(2)}, \beta_1^{(2)}, \beta_2^{(2)}, Y) \end{aligned} \tag{3}$$

- repeat 1000's or 1000000 's times

# BUGS to the rescue

Previously, Bayesian analysis demanded custom-coding MCMC algorithms

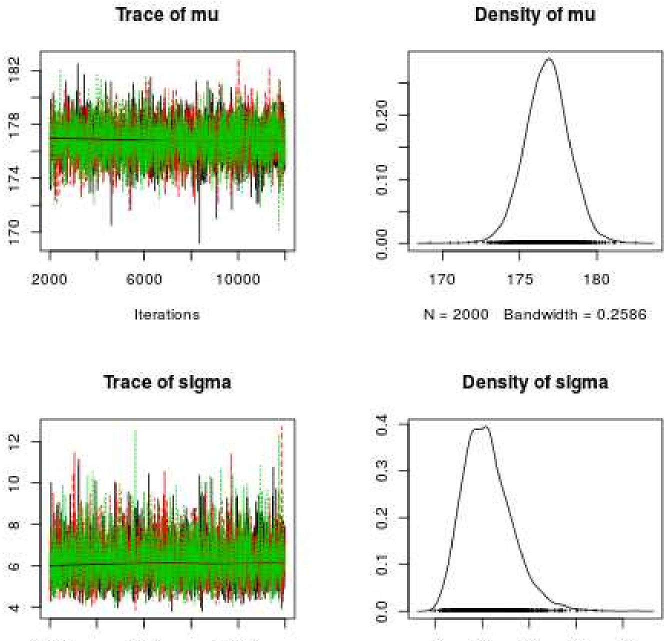
### WinBUGS & OpenBUGS & JAGS

automatically use appropriate sampling techniques; so we don't have to worry

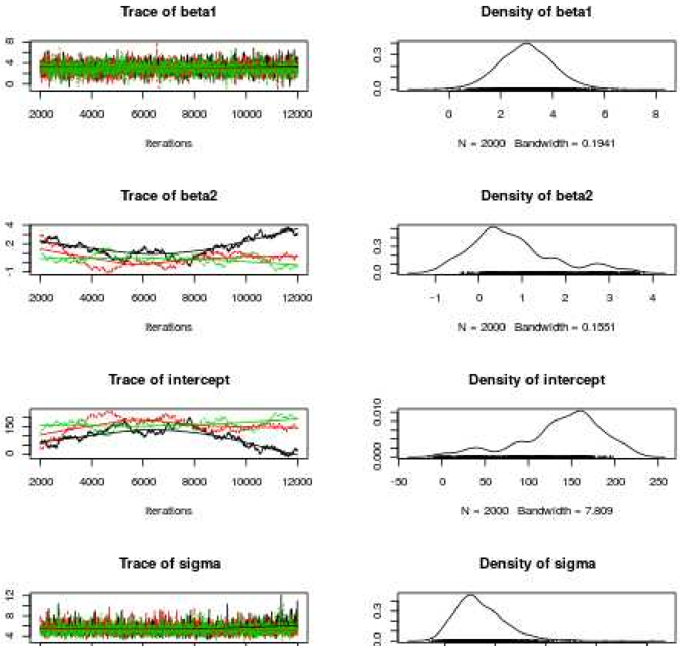
### BUT you must: Monitor the MCMC!

- give reasonable **initial values**
- ensure **convergence**: no trend; independent chains give same answer
- ensure adequate **mixing**: independent samples

# MCMC: Good mixing



# MCMC: Poor convergence



## MCMC parameters in JAGS

- `n.chains`: num. of MCMC chains; more is better
- `n.adapt`: discard first samples; let algorithm 'adapt'
- `n.burn`: discard extra samples; allow algorithm to reach stationary distribution
- `n.iter`: total number of sample; more is better
- `thin`: take every $k^{th}$ iteration for a sample; decorrelates one sample from the next; higher is better
- total samples: number of samples to approximate your Posterior; target at least 2000 to 5000

# MCMC: what to do with bad mixing

- run longer chains
- ensure long enough adaption phase
- misspecified priors
- bad initial values?

# Bayesian Analysis Example

Time to open up R and JAGS

- go to website: `colugos.blogspot.com`

## 'JAGS: Just Another Gibbs Sampler'

Uses BUGS-like syntax (similar to OpenBUGS, WinBUGS)

- `rjags` Package: R friendly JAGS interface
- easy easy easy Bayesian *estimation*
- not so easy for *model selection*

Don't worry about 'samplers': JAGS does the hard work

- specify **likelihood** (how the data arose) and the **priors** example model: height of 20 Australian `y <- c(183.46, 182.32, 178.31, 181.36, 165.12, 185.68, 170.47, 178.11, 174.86, 182.03, 180.09, 172.88, 177.94, 177.26, 182.58, 171, 173.74, 177.78, 180.02, 163.05)`

- lets estimate the mean height (mu) and the dispersion (sigma)

JAGS we estimate the 'precision' (tau): $\tau = \frac{1}{\sigma^2}$

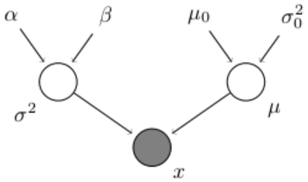



Figure: Prof Mike Jordan lecture notes

# Bayesian Analysis Example 1

- open up R and `rjags`
- download and open the R file:

Jags model syntax: specify priors and likelihood

```
model.txt<-'model{
 # Normal priors on mean height
 mu0 <- 100
 sigma0 <- 35
 tau0 <- pow(sigma0,-2)
 mu ~ dnorm(mu0,tau0) T(0,) # truncated normal
 # Gamma prior on precision
 alpha0 <- 0.1
 beta0 <- 0.1
 tau ~ dgamma(alpha0,beta0)
 # Likelihood: how the data arose
 for(i in 1:length(y)){
   y[i] ~ dnorm(mu,tau) T(0,) # truncated normal
 }
 sigma <- pow(tau,-0.5)
}'
```