Bayesian statistics – Nanjing Forestry University

Lecture 3

- Wei Wu, The University of Southern Mississippi

- December, 2016

Acknowledgement: SESYNC Bayesian modeling for socio-environmental data workshop

THE UNIVERSITY OF SOUTHERN MISSISSIPPI

GULF COAST RESEARCH LABORATORY

Bayes theorem

$$\underbrace{[\theta|y]}_{\text{Posterior}} = \frac{\overbrace{[y|\theta]}^{\text{Likelihood}} \overbrace{[\theta]}^{\text{Prior}}}{\underbrace{\int [y|\theta][\theta]d\theta}_{\text{Marginal}}}$$

- Likelihood: Links unobserved θ to observed y.

Bayes theorem

$$\underbrace{[\theta|y]}_{\text{Posterior}} = \frac{\overbrace{[y|\theta]}^{\text{Likelihood}} \overbrace{[\theta]}^{\text{Prior}}}{\underbrace{\int [y|\theta][\theta]d\theta}_{\text{Marginal}}}$$

- Likelihood: Links unobserved θ to observed y.
- Prior distribution: What is already known about θ.

Bayes theorem

$$\underbrace{[\theta|y]}_{\text{Posterior}} = \frac{\overbrace{[y|\theta]}^{\text{Likelihood}} \; \overbrace{[\theta]}^{\text{Prior}}}{\underbrace{\int [y|\theta][\theta]d\theta}_{\text{Marginal}}}$$

- Likelihood: Links unobserved θ to observed y.
- Prior distribution: What is already known about θ.
- Marginal distribution: the area under the joint distribution curve. Serves to normalize the curve with respect to θ.

Bayes theorem

$$\underbrace{[\theta|y]}_{\text{Posterior}} = \frac{\overbrace{[y|\theta]}^{\text{Likelihood}}\ \overbrace{[\theta]}^{\text{Prior}}}{\underbrace{\int [y|\theta][\theta]d\theta}_{\text{Marginal}}}$$

- Likelihood: Links unobserved θ to observed y.
- Prior distribution: What is already known about θ.
- Marginal distribution: the area under the joint distribution curve. Serves to normalize the curve with respect to θ.
- Posterior distribution: a true PDF.

Prior distributions

Prior distributions can be informative, reflecting knowledge gained in previous research or they can be vague, reflecting a lack of information about θ before data are collected.

Prior distributions

- In special cases the posterior, $[\theta|y]$, has the same form as the prior, $[\theta]$.

Prior distributions

- In special cases the posterior, [θ|y], has the same form as the prior, [θ].
- In these cases, the prior and the posterior are said to be conjugate.

Prior distributions

- In special cases the posterior, [θ|y], has the same form as the prior, [θ].
- In these cases, the prior and the posterior are said to be conjugate.
- Conjugacy is important for two reasons:

Prior distributions

- In special cases the posterior, [θ|y], has the same form as the prior, [θ].
- In these cases, the prior and the posterior are said to be conjugate.
- Conjugacy is important for two reasons:

- Conjugacy in simple cases minimizes computational work and, in more complicated cases, allow us to break down calculations into management units.

Prior distributions

- In special cases the posterior, [θ|y], has the same form as the prior, [θ].
- In these cases, the prior and the posterior are said to be conjugate.
- Conjugacy is important for two reasons:

- Conjugacy in simple cases minimizes computational work and, in more complicated cases, allow us to break down calculations into management units.
- Conjugacy plays an important role in Markov Chain Monte Carlo procedures.

Conjugate priors

Table A.3: Table of conjugate distributions

Likelihood	Prior distribution	Posterior distribution
$y_i \sim \text{binomial}\,(n, \phi)$	$\phi \sim \text{beta}\,(\alpha, \beta)$	$\phi \sim \text{beta}\,(\sum y_i + \alpha, n - \sum y_i + \beta)$
$y_i \sim \text{Bernoulli}(\phi)$	$\phi \sim \text{beta}\,(\alpha, \beta)$	$\phi \sim \text{beta}\,(\sum_{i=1}^n y_i + \alpha, \sum_{i=1}^n (1 - y_i) + \beta)$
$y_i \sim \text{Poisson}\,(\lambda)$	$\lambda \sim \text{gamma}\,(\alpha, \beta)$	$\lambda \sim \text{gamma}\,(\alpha + \sum_{i=1}^n y_i,\ \beta + n)$
$y_i \sim \text{normal}(\mu, \sigma^2)$ σ^2 is known.	$\mu \sim \text{normal}\,(\mu_0, \sigma_0^2)$	$\mu \sim \text{normal}\left(\frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n y_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$
$y_i \sim \text{normal}(\mu, \sigma^2)$ μ is known.	$\sigma^2 \sim$ inverse gamma(α, β)	$\sigma^2 \sim$ inverse gamma $\left(\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n (y_i - \mu)^2}{2}\right)$
$y_i \sim \text{lognormal}(\mu, \sigma^2)$, μ is known	$\sigma^2 \sim$ inverse gamma(α, β) ,	$\sigma^2 \sim$ inverse gamma $\left(n/2 + \alpha, \frac{(\log(y_i) - \mu)^2}{2} + \beta\right)$
$y_i \sim \text{lognormal}\,(\mu, \sigma^2)$ σ^2 is known	$\mu \sim \text{normal}\,(\mu_0, \sigma_0^2)$	$\mu \sim \text{normal}\left(\frac{\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n \log y_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)}, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right)$

Just a bit more on priors

- There is no such thing as a truly non-informative prior – only those that influences the posterior more (or less) than others.

Just a bit more on priors

- There is no such thing as a truly non-informative prior – only those that influences the posterior more (or less) than others.
- Informative priors can be justified and useful.

Just a bit more on priors

- There is no such thing as a truly non-informative prior – only those that influences the posterior more (or less) than others.
- Informative priors can be justified and useful.
- “Non-informative:, vague, or flat priors are provisional starting points for a Bayesian analysis.

Example: Childhood asthma and PM

- You are studying the relationship between childhood asthma and industrial airborne PM.
- In one school, 17 of 80 students have been hospitalized for asthma-related issues.

What distributions would you choose for the likelihood and prior?
How would you draw the DAG?

Picking distributions

Prior: $([\phi])$

- continuous quantity ranging from 0 to 1

Likelihood: $([\mathbf{y}|\phi])$

- count data: $y = 17$ successes, given $n = 80$ trials

Posterior: $([\phi|\mathbf{y}])$

- Is there a conjugate to use?

Picking distributions

Prior: $([\phi])$

- continuous quantity ranging from 0 to 1.
- Uniform beta with parameters, $\alpha_{prior} = 1, \beta_{prior} = 1$
- $\phi \sim \text{beta}(1, 1)$

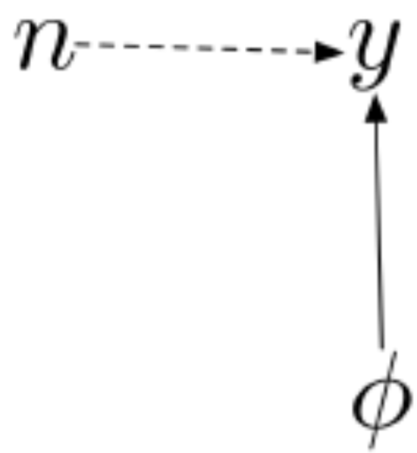
Likelihood: $([\boldsymbol{y}|\phi])$

- binomial distribution with $y = 17$ successes, given $n = 80$ trials

Posterior: $([\phi|\boldsymbol{y}])$

- Using the beta-binomial conjugate prior relationship
- $\phi \sim \text{beta}(\alpha_{post}, \beta_{post})$

Drawing the DAG



Writing out the full posterior

$$\text{beta}(\alpha_{post}, \beta_{post}) = \frac{\text{binomial}(y|\phi, n)\,\text{beta}(\alpha_{prior}, \beta_{prior})}{[y]}$$

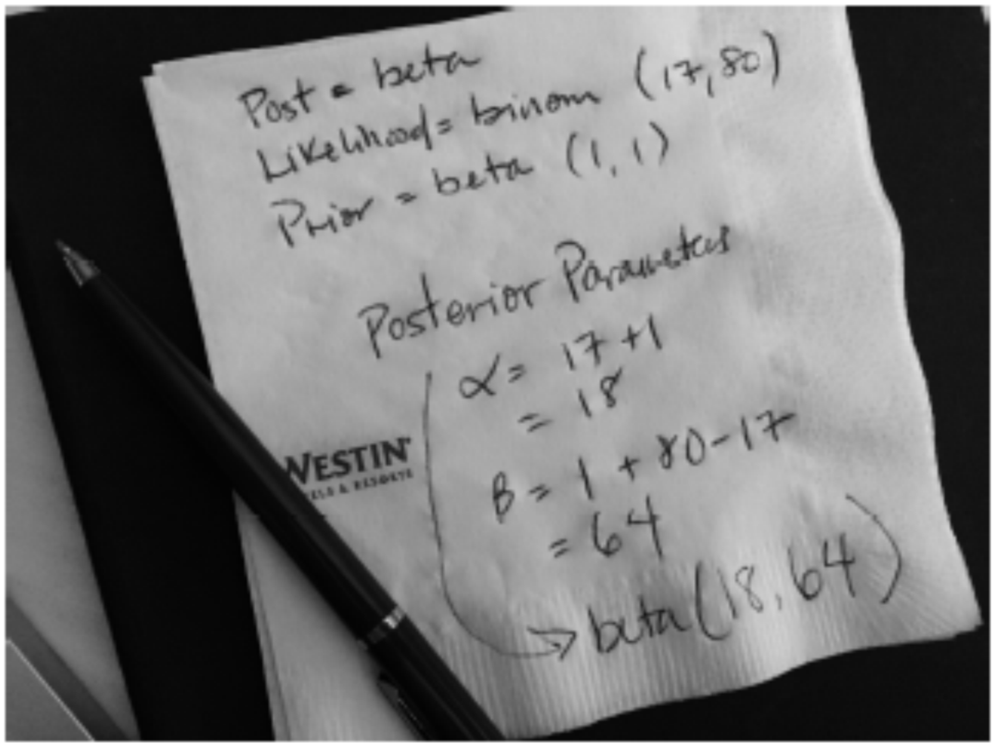
Posterior distribution parameters

$$\alpha_{post} = \Sigma_{y_i} + \alpha_{prior}$$
$$\beta_{post} = n - \Sigma_{y_i} + \beta_{prior}$$

This means you can

This means you can

Literaly calculate the parameters of the posterior distribution on the back of a hotel napkin.



Posterior distribution parameters

```
aPrior <- 1
bPrior <- 1
y <- 17
n <- 80
aPost <- aPrior + y
aPost
```

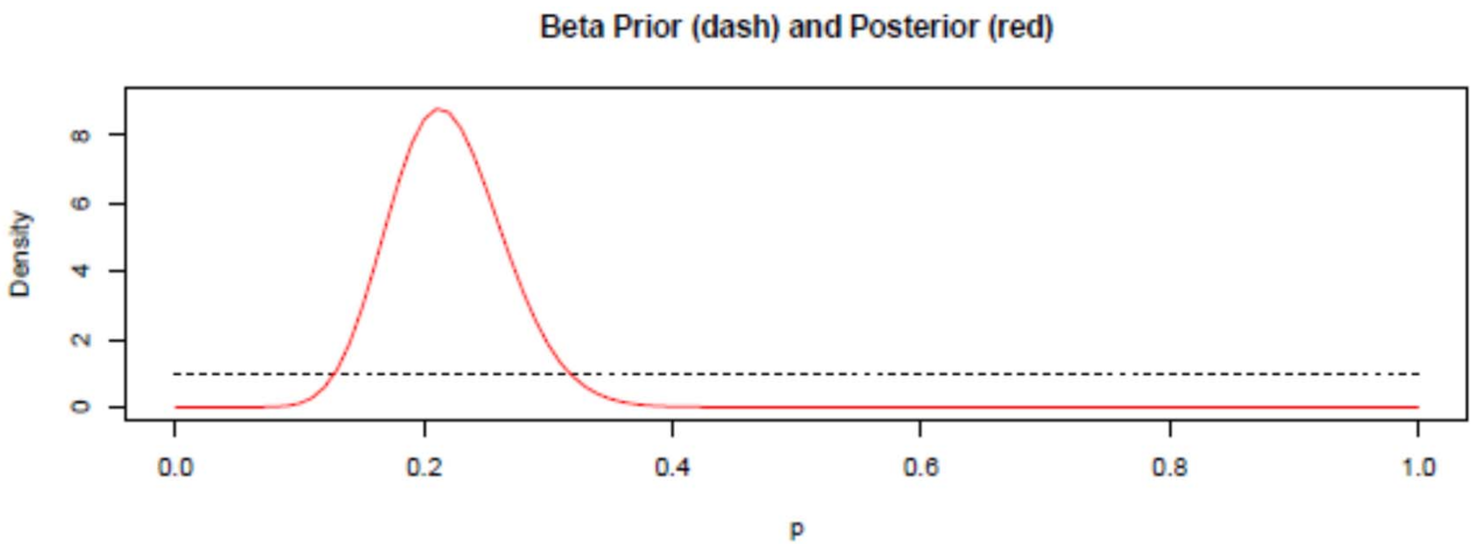
```
## [1] 18
```

```
bPost <- bPrior + n - y
bPost
```

```
## [1] 64
```

$\phi \sim \text{beta}(18, 64)$

Plotting the prior and posterior



Between which quantiles does ϕ lie with probability 0.95?

```
## [1] 0.1373419 0.3146275
```

What if, instead, the new study showed a much lower incidence of hospitalization, rather than a closer finding (e.g. 3 of 75 hospitalized)?

```r
aPrior <- 18
bPrior <- 64
y <- 3
n <- 75
aPost <- aPrior + y
aPost
```

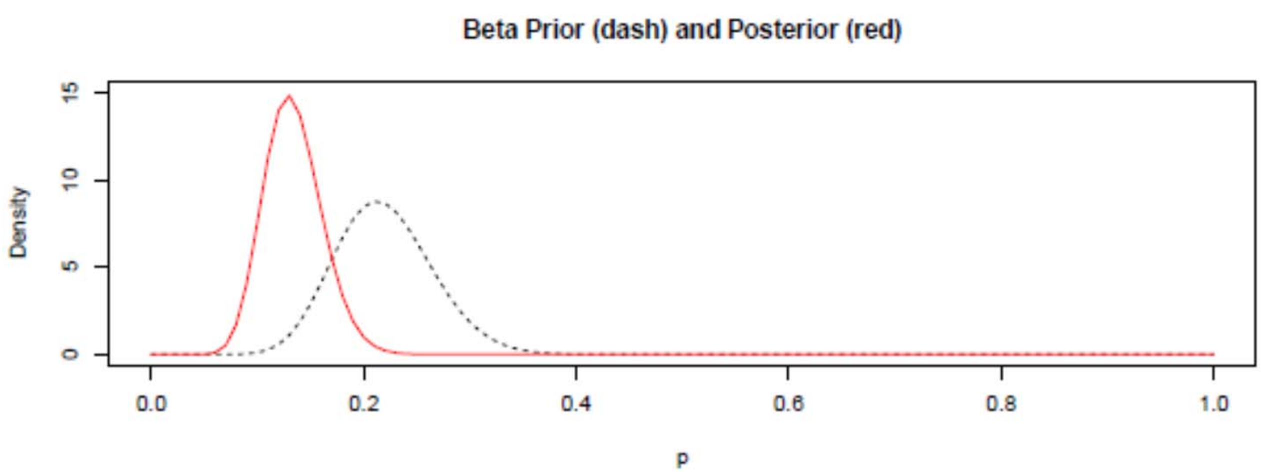
```
## [1] 21
```

```r
bPost <- bPrior + n - y
bPost
```

```
## [1] 136
```

$\phi \sim \text{beta}(21, 136)$

Plotting the prior and posterior



Between which quantiles does ϕ lie with probability 0.95?

```
## [1] 0.08529956 0.19103790
```

What if, instead, the new study showed a much higher incidence of hospitalization, rather than a closer finding (e.g. 80 of 80 hospitalized)?

```r
aPrior <- 18
bPrior <- 64
y <- 80
n <- 80
aPost <- aPrior + y
aPost
```

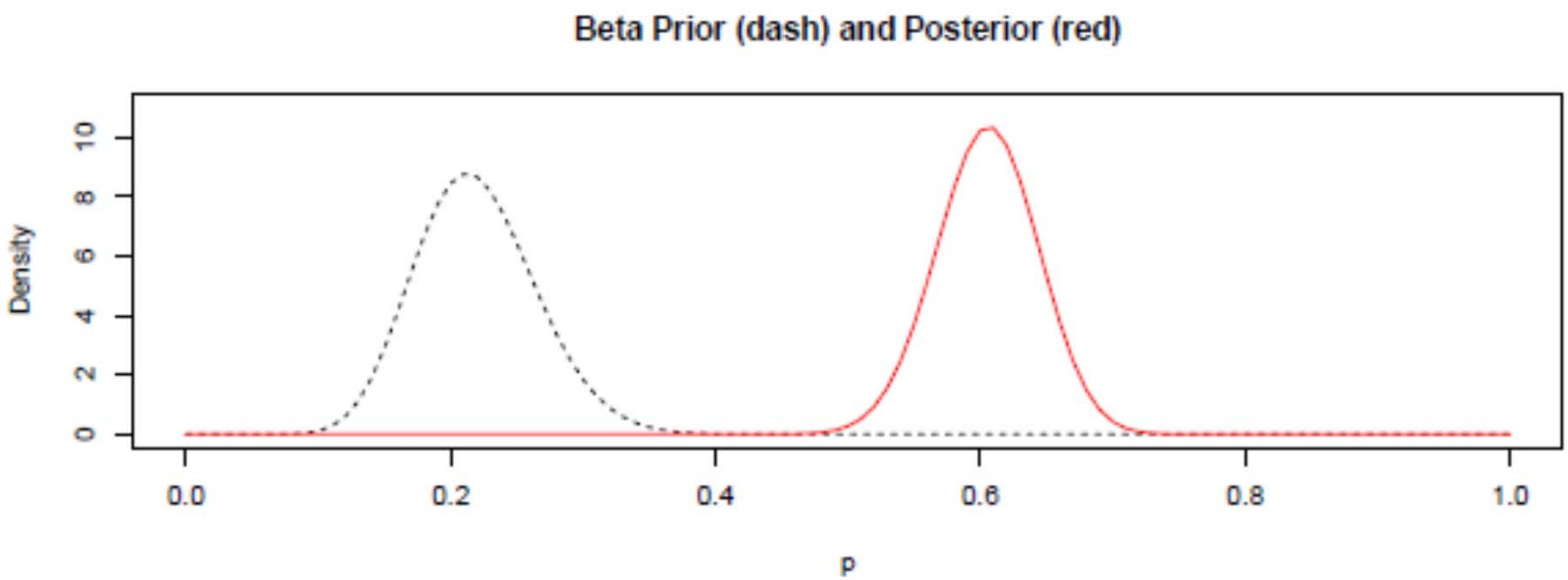
```
## [1] 98
```

```r
bPost <- bPrior + n - y
bPost
```

```
## [1] 64
```

$\phi \sim \text{beta}(98, 64)$

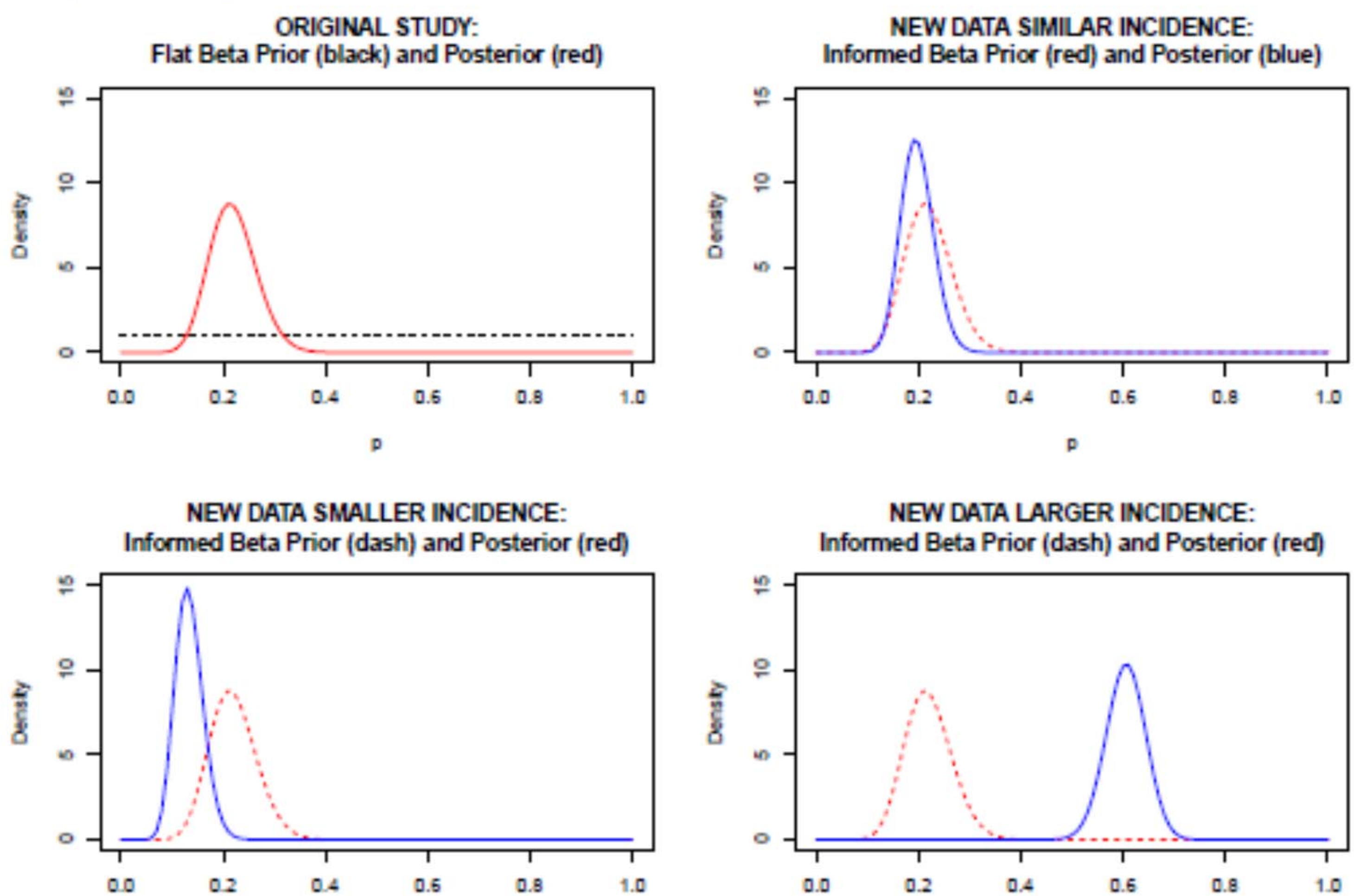
Plotting the prior and posterior



Between which quantiles does ϕ lie with probability 0.95?

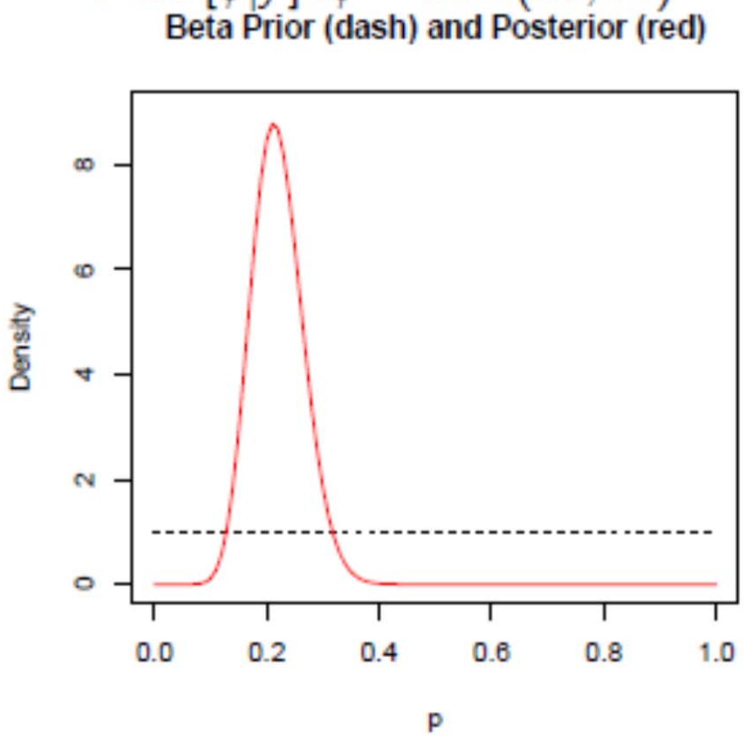
```
## [1] 0.5287686 0.6786532
```

Exploring the role of new data (prior and data influence posteriors)

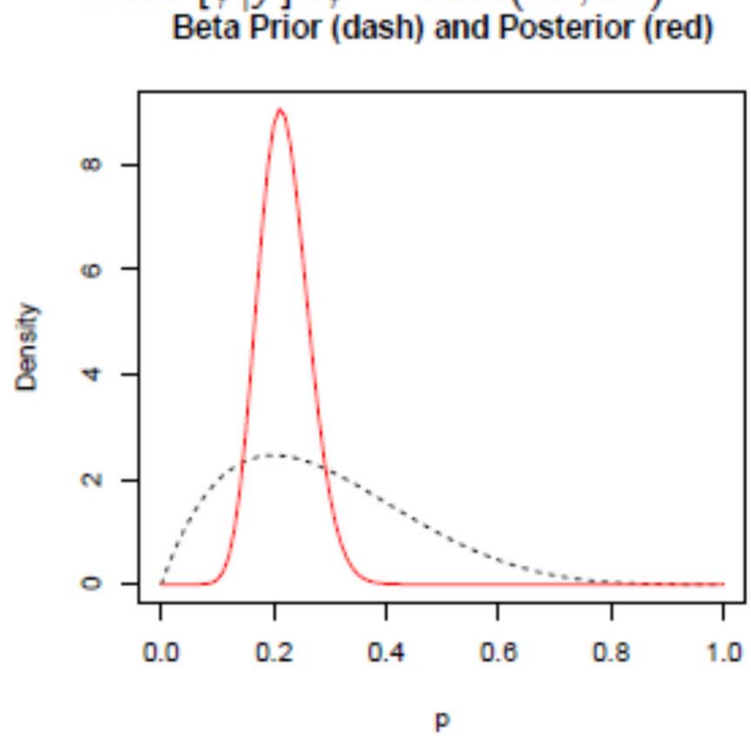


Comparing priors (data: 17 out of 80 hospitalized)

Prior $[\phi]$: $\phi \sim \text{beta}(1,1)$
Post $[\phi|y]$:$\phi \sim \text{beta}(18,64)$



Prior $[\phi]$: $\phi \sim \text{beta}(2,5)$
Post $[\phi|y]$:$\phi \sim \text{beta}(19,68)$



Now consider the following data

School: Hospitalizations/Total students

School 1: 17/80
School 2: 17/75
School 3: 19/100
School 4: 10/55
School 5: 33/111

Now consider the following data

For a childhood asthma hospitalization model that, use the new data across schools, what is the DAG?

Full posterior

$$
\begin{aligned}
y_i &\sim \text{binomial}(n_i, \phi_i) \\
\phi_i &\sim \text{beta}(\alpha, \beta) \\
\alpha &\sim \text{uniform}(0, 500) \\
\beta &\sim \text{uniform}(0, 500) \\
[\boldsymbol{\phi}, \alpha, \beta | \mathbf{y}] &\propto \prod_{i=1}^{5} [y_i | \phi_i, n_i][\phi_i | \alpha, \beta][\alpha][\beta]
\end{aligned}
$$

This is a hierarchical model because parameter, ϕ, is on both sides of the conditioning symbol.

Things to remember

- This is no such a thing as a noninformative prior, but certain priors influence the posterior distribution more than others.

Things to remember

- This is no such a thing as a noninformative prior, but certain priors influence the posterior distribution more than others.
- Informative priors, when properly justified, can be useful.

Things to remember

- This is no such a thing as a noninformative prior, but certain priors influence the posterior distribution more than others.
- Informative priors, when properly justified, can be useful.
- Strong data overwhelms a prior.

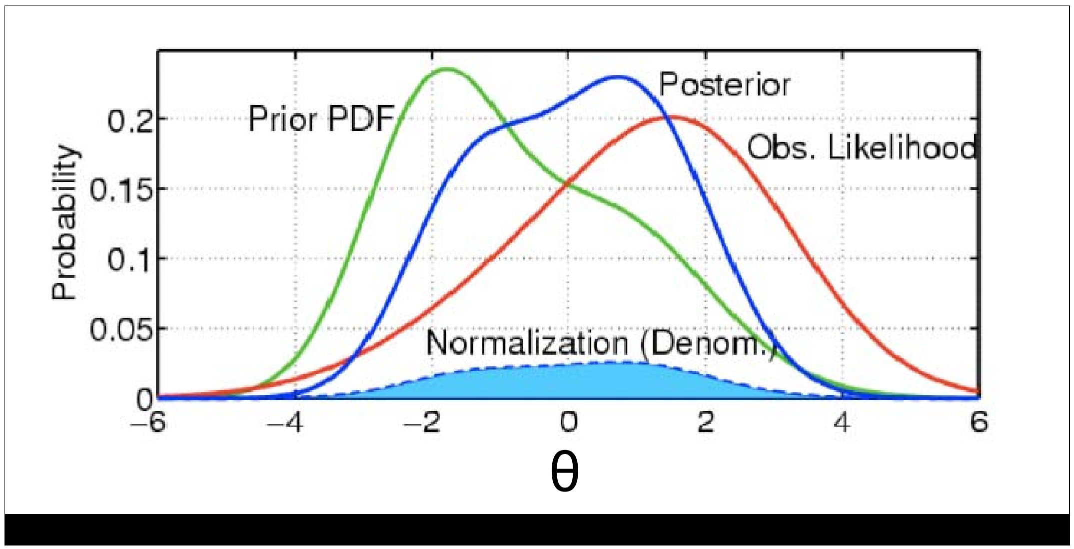
Role of priors in science

- Priors represent our current knowledge (or lack of current knowledge), which is updated with data.

Role of priors in science

- Priors represent our current knowledge (or lack of current knowledge), which is updated with data.
- We encourage you to think of vague priors as a provisional starting point.

The posterior distribution represents a balance between the information contained in the likelihood and the information contained in the prior distribution.



An informative prior influences the posterior distribution. A vague prior exerts minimal influence.

Influence of data and prior information

$$\text{beta}(\varphi|y) = \frac{\text{binomial}(y|\varphi, n)\,\text{beta}(\varphi|\alpha_{prior}, \beta_{prior})}{[y]}$$

$$\alpha_{posterior} = \alpha_{prior} + y$$
$$\beta_{posterior} = \beta_{prior} + n - y$$

Influence of data and prior information

$$\text{gamma}(\lambda \mid \mathbf{y}) = \frac{\prod_{i=1}^{4} \text{Poisson}(y_i \mid \lambda)\,\text{gamma}(\lambda \mid \alpha_{prior}, \beta_{prior})}{[\mathbf{y}]}$$

$$\alpha_{posterior} = \alpha_{prior} + \sum_{i=1}^{4} y_i$$

$$\beta_{posterior} = \beta_{prior} + n$$

Roadmap

- Informative priors
- Vague priors
  - Scaling
  - Potential problems
  - Non-linear transformations of "vague" priors
  - Guidance

Why use informative priors?

A natural tool for synthesis and updating

- Speed convergence
- Reduce problems with identifyability
- Allows estimation of quantities that would otherwise be inestimable
- Reduces problems with sensitivity to transformation

They are a great tool! Why would you not use them?

Why are they not used more often?

- Cultural–"All studies stand alone."
- Current texts mostly use vague priors (including ours!)
- Hard work!
- Worries about "excessive subjectivity"

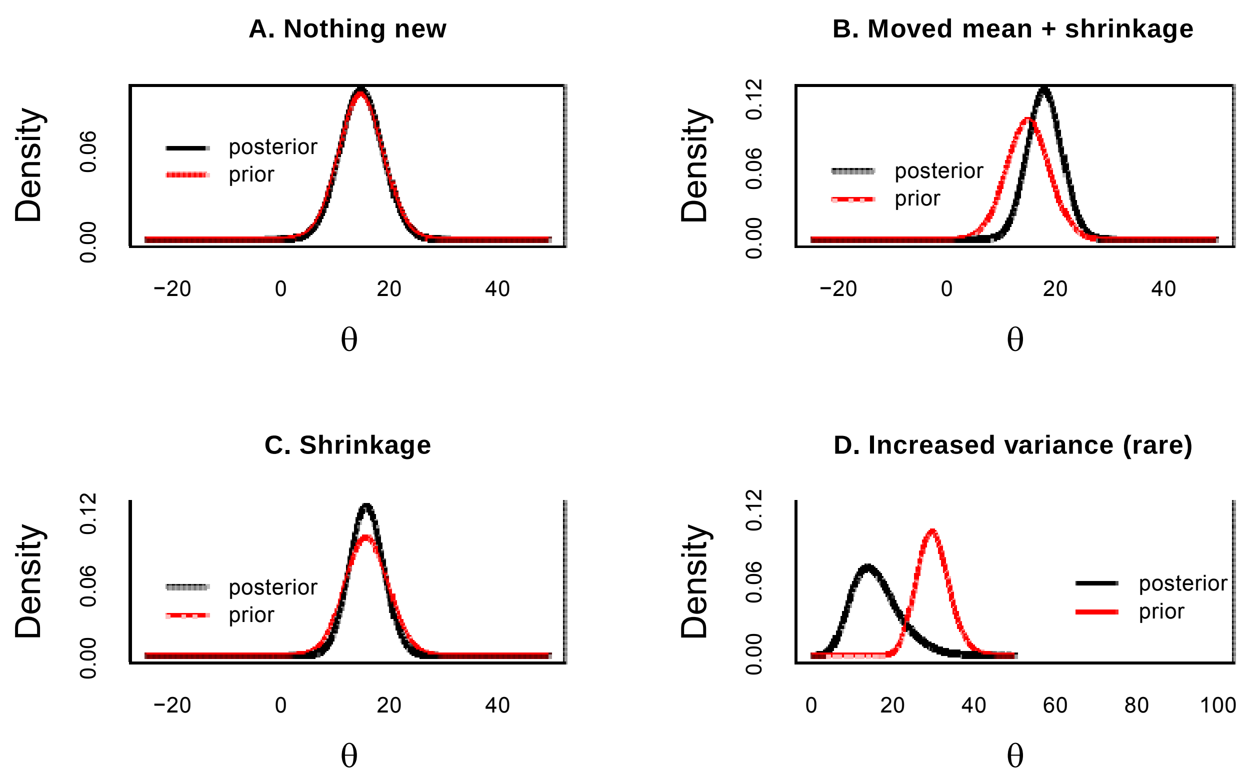
How to develop?

- Strong scholarship is the basis of strong priors.
- Often need to use moment matching to convert means and standard deviations into parameters for priors.
- Pilot studies
- In biology, allometric relationships are a great source of informative priors on all sorts of parameters.[1]
- Build deterministic models with parameters with biological or socio-ecological definitions.
- Think about rescaling the data.

---

[1] See Peters. 1983. The Ecological Implications of Body Size. Cambridge University Press, Cambridge, U.K. and Pennycuick, C. J. 1992. Newton Rules Biology. Oxford University Press, Oxford U.K.

Interpreting posteriors relative to priors



A vague prior is a distribution with a range of uncertainty that is clearly wider than the range of reasonable values for the parameter (Gelman and Hill 2007:347).

Also called: diffuse, flat, automatic, nonsubjective, locally uniform, objective, and, incorrectly, “non-informative.”

Vague priors are *provisional* in two ways:

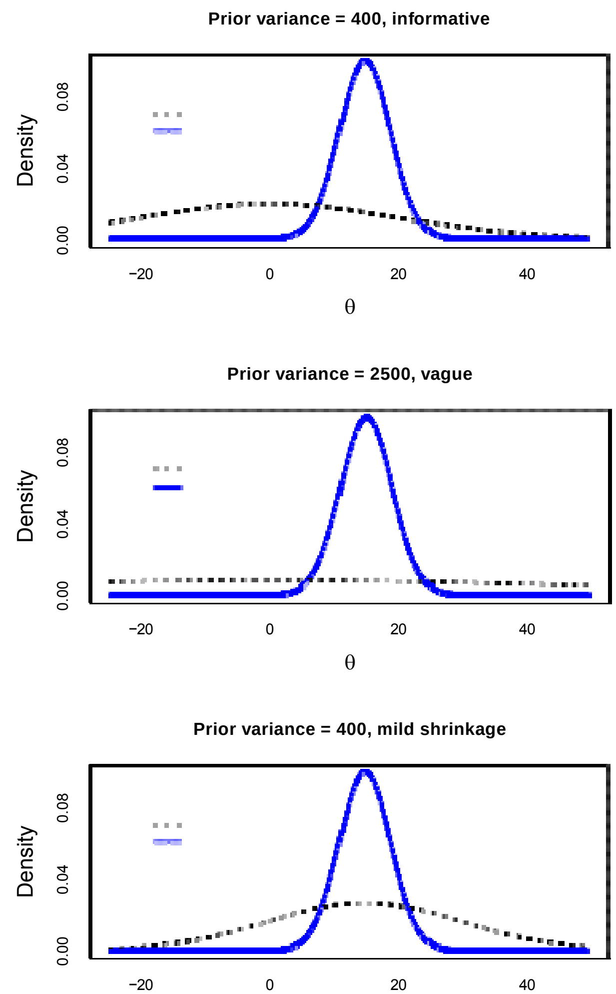
1. Operationally provisional: We try one. Does the output make sense? Are the posteriors sensitive to changes in parameters? Are there values in the posterior that are simply unreasonable? We may need to try another type of prior.
2. Strategically provisional: We use vague priors until we can get informative ones, which we prefer to use.

Scaling

Vague priors need to be scaled properly.

Suppose you specify a prior on a parameter, $\theta \sim \text{normal}(\mu = 0, \sigma^2 = 1000)$. Will this prior influence the posterior distribution?

Scaling vague priors (dashed – prior)



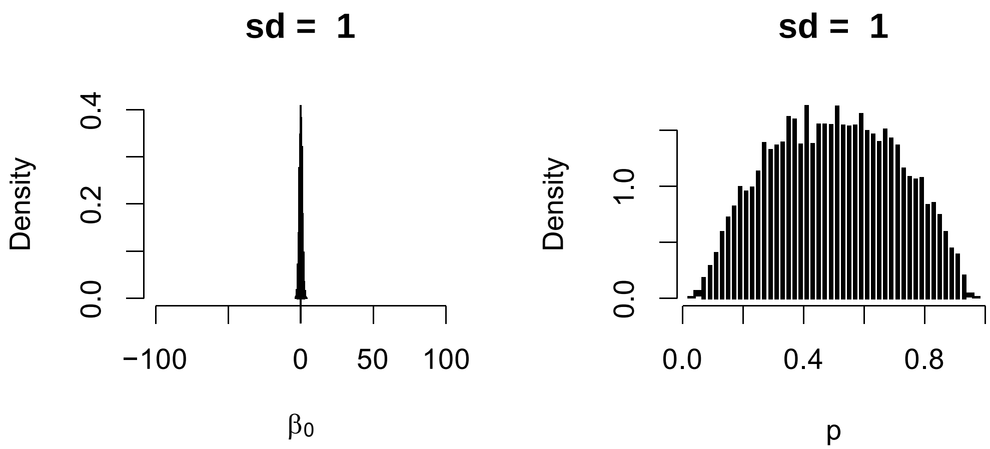
Problems with excessively vague priors

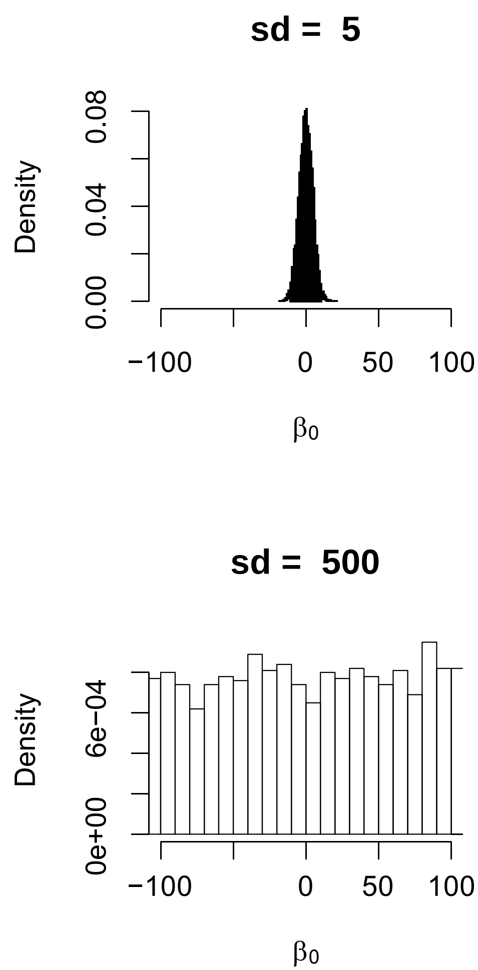
- Computational: failure to converge, slicer errors, failure to calculate log density, etc.
- Cause pathological behavior in posterior distribution, i.e, values are included that are unreasonable.
- Sensitivity: changes in parameters of "vague" priors meaningfully changes the posterior.
- Non-linear functions of parameters with vague priors have informative priors.

"Priors" on nonlinear functions of parameters

$$p_i = g(\beta, x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$$[\beta|\mathbf{y}] \propto \prod_{i=1}^{n} \text{Bernoulli}(y_i|g(\beta, x_i)) \times \text{normal}(\beta_0|0, 10000)\text{normal}(\beta_1|0, 10000)$$




sd = 5
Density
0.00
0.04
0.08
−100
0
50
100
β0
sd = 5
Density
0
4
8
0.0
0.4
0.8
p
sd = 500
Density
0e+00
6e−04
−100
0
50
100
β0
sd = 500
Density
0
10
20
0.0
0.4
0.8
p

More guidance

Choose vague priors thoughtfully, particularly when data are limited.

Use uniform priors on σ for group level models when there are 4 or more groups. Consider half-Cauchy priors with scale parameter set at reasonable estimate of σ with fewer groups.

Know that priors that are vague for parameters can influence non-linear functions of parameters. This influence can be minimized if vague priors are centered in the vicinity of the central tendency of the posteriors of parameters.

Always use informative priors when you can. You know more than you think you do.

References for this lecture

Seaman III, J. W. and Seaman Jr., J. W. and Stamey, J. D. 2012 Hidden dangers of specifying noninformative priors, The American Statistician 66, 77-84 (2012)

Hobbs and Hooten 2015, Section 5.4