Introduction

- Start with a probability distribution $f(\mathbf{y}|\boldsymbol{\theta})$ for the data $\mathbf{y} = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, and add a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a vector of hyperparameters

Introduction

- Start with a probability distribution $f(\mathbf{y}|\boldsymbol{\theta})$ for the data $\mathbf{y} = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, and add a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a vector of hyperparameters
- Inference for $\boldsymbol{\theta}$ is based on its posterior distribution,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\eta}) &= \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\eta})}{p(\mathbf{y}|\boldsymbol{\eta})} = \frac{p(\mathbf{y}, \boldsymbol{\theta}|\boldsymbol{\eta})}{\int p(\mathbf{y}, \mathbf{u}|\boldsymbol{\eta})\, d\mathbf{u}} \\ &= \frac{f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})}{\int f(\mathbf{y}|\mathbf{u})p(\mathbf{u}|\boldsymbol{\eta})\, d\mathbf{u}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})}{m(\mathbf{y}|\boldsymbol{\eta})} . \end{aligned}$$

Introduction

- Start with a probability distribution $f(\mathbf{y}|\boldsymbol{\theta})$ for the data $\mathbf{y} = (y_1, \ldots, y_n)$ given a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, and add a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, where $\boldsymbol{\eta}$ is a vector of hyperparameters
- Inference for $\boldsymbol{\theta}$ is based on its posterior distribution,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\eta}) &= \frac{p(\mathbf{y},\boldsymbol{\theta}|\boldsymbol{\eta})}{p(\mathbf{y}|\boldsymbol{\eta})} = \frac{p(\mathbf{y},\boldsymbol{\theta}|\boldsymbol{\eta})}{\int p(\mathbf{y},\mathbf{u}|\boldsymbol{\eta})\, d\mathbf{u}} \\ &= \frac{f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})}{\int f(\mathbf{y}|\mathbf{u})p(\mathbf{u}|\boldsymbol{\eta})\, d\mathbf{u}} = \frac{f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})}{m(\mathbf{y}|\boldsymbol{\eta})}\,. \end{aligned}$$

- We refer to this formula as *Bayes' Theorem*. Note its similarity to the definition of conditional probability,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Example 2.1

- Consider the normal (Gaussian) likelihood, $f(y|\theta) = N(y|\theta, \sigma^2)$, $y \in \Re,\ \theta \in \Re$, and $\sigma > 0$ known. Take $p(\theta|\boldsymbol{\eta}) = N(\theta|\mu, \tau^2)$, where $\mu \in \Re$ and $\tau > 0$ are known hyperparameters, so that $\boldsymbol{\eta} = (\mu, \tau)$. Then

$$p(\theta|y) = N\left(\theta \,\Big|\, \frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2},\ \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right) .$$

Example 2.1

- Consider the normal (Gaussian) likelihood, $f(y|\theta) = N(y|\theta, \sigma^2)$, $y \in \Re,\ \theta \in \Re$, and $\sigma > 0$ known. Take $p(\theta|\boldsymbol{\eta}) = N(\theta|\mu, \tau^2)$, where $\mu \in \Re$ and $\tau > 0$ are known hyperparameters, so that $\boldsymbol{\eta} = (\mu, \tau)$. Then

$$p(\theta|y) = N\left(\theta \,\middle|\, \frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right) .$$

- Write $B = \frac{\sigma^2}{\sigma^2+\tau^2}$, and note that $0 < B < 1$. Then:

Example 2.1

- Consider the normal (Gaussian) likelihood, $f(y|\theta) = N(y|\theta, \sigma^2)$, $y \in \Re,\ \theta \in \Re$, and $\sigma > 0$ known. Take $p(\theta|\boldsymbol{\eta}) = N(\theta|\mu, \tau^2)$, where $\mu \in \Re$ and $\tau > 0$ are known hyperparameters, so that $\boldsymbol{\eta} = (\mu, \tau)$. Then

$$p(\theta|y) = N\left(\theta \,\middle|\, \frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}\,,\ \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right) .$$

- Write $B = \frac{\sigma^2}{\sigma^2+\tau^2}$, and note that $0 < B < 1$. Then:
  - $E(\theta|y) = B\mu + (1-B)y$, a weighted average of the prior mean and the observed data value, with weights determined sensibly by the variances.

Example 2.1

- Consider the normal (Gaussian) likelihood, $f(y|\theta) = N(y|\theta, \sigma^2)$, $y \in \Re$, $\theta \in \Re$, and $\sigma > 0$ known. Take $p(\theta|\boldsymbol{\eta}) = N(\theta|\mu, \tau^2)$, where $\mu \in \Re$ and $\tau > 0$ are known hyperparameters, so that $\boldsymbol{\eta} = (\mu, \tau)$. Then

$$p(\theta|y) = N\left(\theta \,\middle|\, \frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right) .$$

- Write $B = \frac{\sigma^2}{\sigma^2+\tau^2}$, and note that $0 < B < 1$. Then:
  - $E(\theta|y) = B\mu + (1-B)y$, a weighted average of the prior mean and the observed data value, with weights determined sensibly by the variances.
  - $Var(\theta|y) = B\tau^2 \equiv (1-B)\sigma^2$, smaller than τ^2 and σ^2.

Example 2.1

- Consider the normal (Gaussian) likelihood, $f(y|\theta) = N(y|\theta, \sigma^2)$, $y \in \Re$, $\theta \in \Re$, and $\sigma > 0$ known. Take $p(\theta|\boldsymbol{\eta}) = N(\theta|\mu, \tau^2)$, where $\mu \in \Re$ and $\tau > 0$ are known hyperparameters, so that $\boldsymbol{\eta} = (\mu, \tau)$. Then

$$p(\theta|y) = N\left(\theta \,\middle|\, \frac{\sigma^2\mu + \tau^2 y}{\sigma^2 + \tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + \tau^2}\right) .$$

- Write $B = \frac{\sigma^2}{\sigma^2+\tau^2}$, and note that $0 < B < 1$. Then:
  - $E(\theta|y) = B\mu + (1 - B)y$, a weighted average of the prior mean and the observed data value, with weights determined sensibly by the variances.
  - $Var(\theta|y) = B\tau^2 \equiv (1 - B)\sigma^2$, smaller than τ^2 and σ^2.
  - Precision (which is like "information") is additive: $Var^{-1}(\theta|y) = Var^{-1}(\theta) + Var^{-1}(y|\theta)$.

Sufficiency still helps

- Lemma: If $S(\mathbf{y})$ is sufficient for $\boldsymbol{\theta}$, then $p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}|s)$, so we may work with s instead of the entire dataset $\mathbf{y}$.

Sufficiency still helps

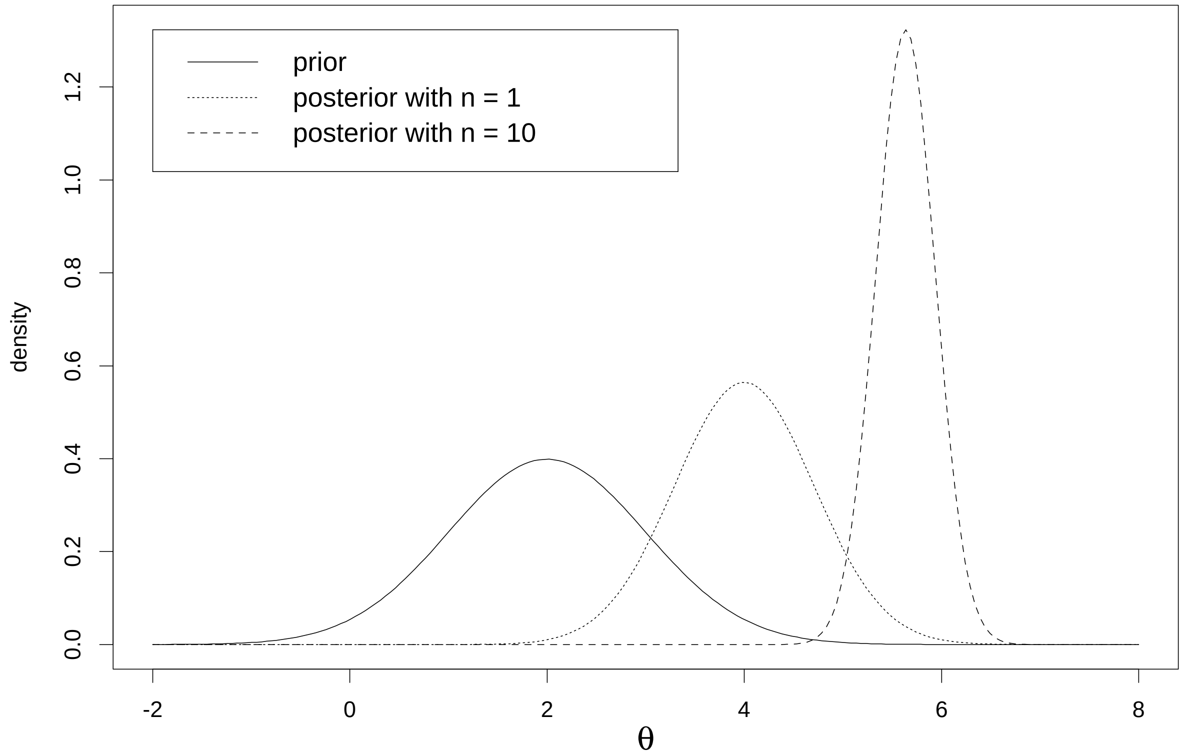
- Lemma: If $S(\mathbf{y})$ is sufficient for $\boldsymbol{\theta}$, then $p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}|s)$, so we may work with s instead of the entire dataset $\mathbf{y}$.
- Example 2.2: Consider again the normal/normal model where we now have an independent sample of size n from $f(\mathbf{y}|\theta)$. Since $S(\mathbf{y}) = \bar{y}$ is sufficient for θ, we have that $p(\theta|\mathbf{y}) = p(\theta|\bar{y})$.

Sufficiency still helps

- Lemma: If $S(\mathbf{y})$ is sufficient for $\boldsymbol{\theta}$, then $p(\boldsymbol{\theta}|\mathbf{y}) = p(\boldsymbol{\theta}|s)$, so we may work with s instead of the entire dataset $\mathbf{y}$.
- Example 2.2: Consider again the normal/normal model where we now have an independent sample of size n from $f(\mathbf{y}|\theta)$. Since $S(\mathbf{y}) = \bar{y}$ is sufficient for θ, we have that $p(\theta|\mathbf{y}) = p(\theta|\bar{y})$.
- But since we know that $f(\bar{y}|\theta) = N(\theta, \sigma^2/n)$, previous slide implies that

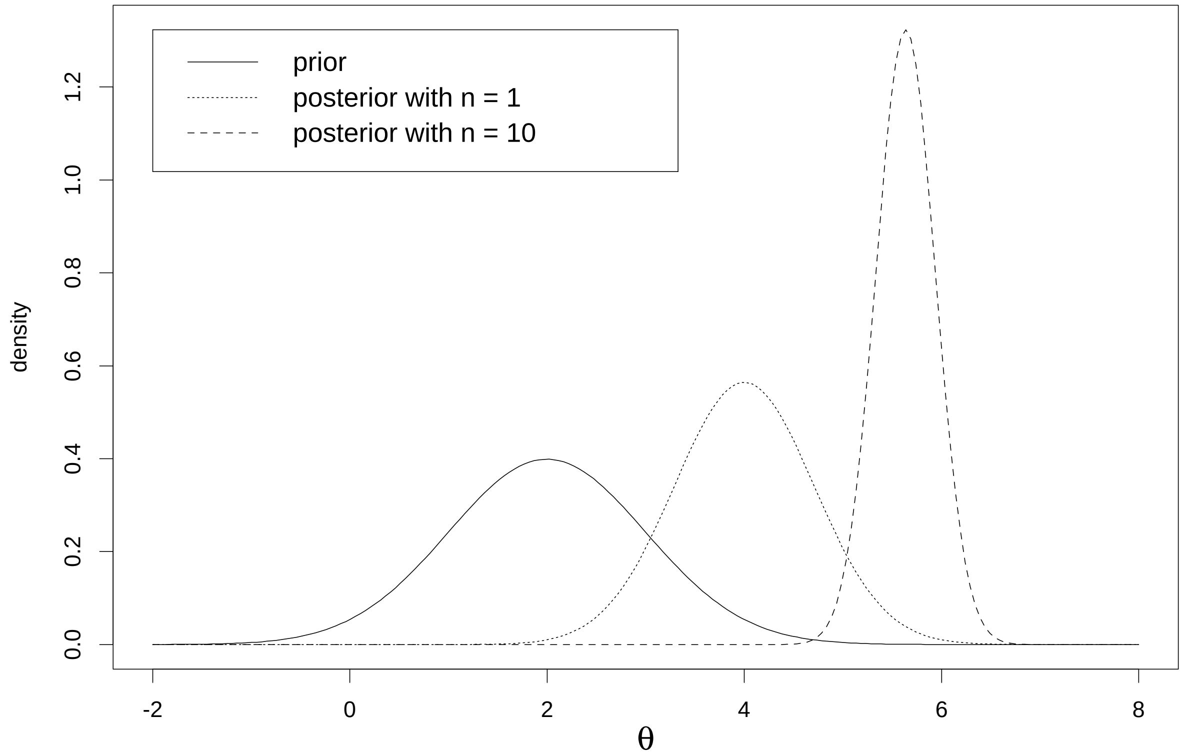
$$
\begin{aligned}
p(\theta|\bar{y}) &= N\left(\theta \,\middle|\, \frac{(\sigma^2/n)\mu + \tau^2\bar{y}}{(\sigma^2/n) + \tau^2}, \frac{(\sigma^2/n)\tau^2}{(\sigma^2/n) + \tau^2}\right) \\
&= N\left(\theta \,\middle|\, \frac{\sigma^2\mu + n\tau^2\bar{y}}{\sigma^2 + n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2 + n\tau^2}\right) .
\end{aligned}
$$

Example: $\mu = 2, \bar{y} = 6, \tau = \sigma = 1$



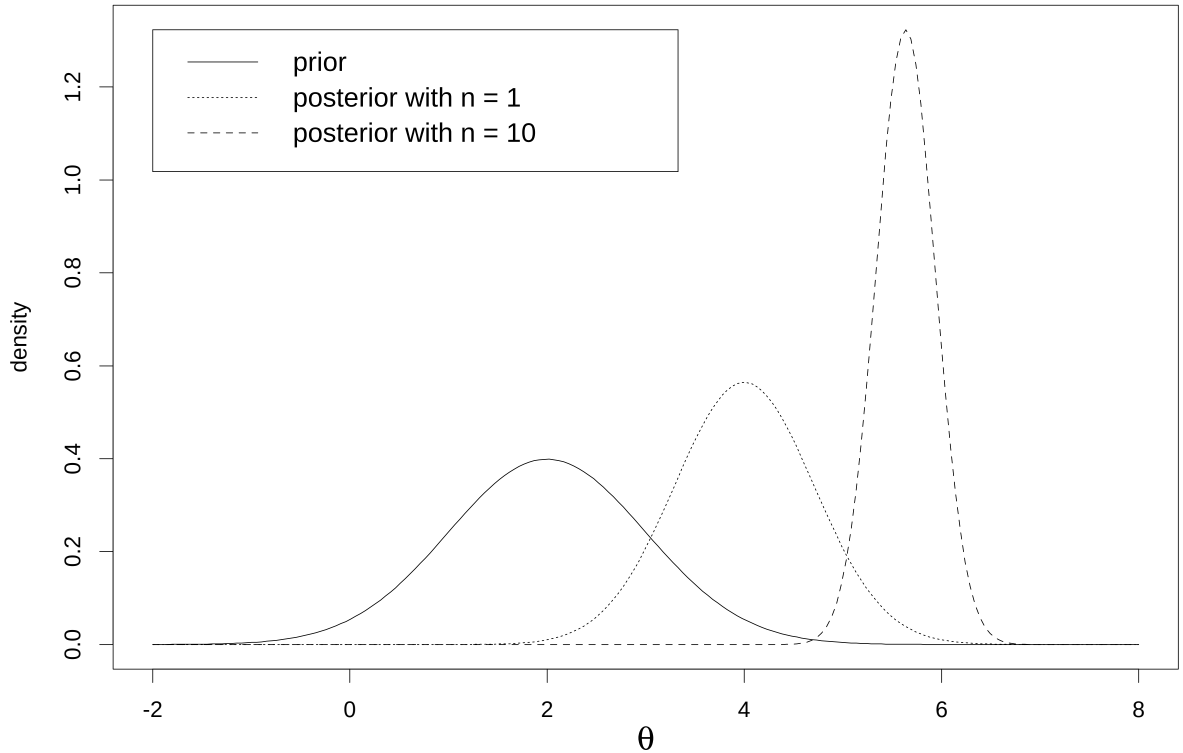
- When $n = 1$ the prior and likelihood receive equal weight, so the posterior mean is $4 = \frac{2+6}{2}$.

Example: $\mu = 2, \bar{y} = 6, \tau = \sigma = 1$



- When $n = 1$ the prior and likelihood receive equal weight, so the posterior mean is $4 = \frac{2+6}{2}$.
- When $n = 10$ the data dominate the prior, resulting in a posterior mean much closer to $\bar{y}$.

Example: $\mu = 2, \bar{y} = 6, \tau = \sigma = 1$



- When $n = 1$ the prior and likelihood receive equal weight, so the posterior mean is $4 = \frac{2+6}{2}$.
- When $n = 10$ the data dominate the prior, resulting in a posterior mean much closer to $\bar{y}$.
- The posterior variance also shrinks as n gets larger; the posterior collapses to a point mass on $\bar{y}$ as $n \to \infty$.

Three-stage Bayesian model

- If we are unsure as to the proper value of the hyperparameter η, the natural Bayesian solution would be to quantify this uncertainty in a third-stage distribution, sometimes called a hyperprior.

Three-stage Bayesian model

- If we are unsure as to the proper value of the hyperparameter $\boldsymbol{\eta}$, the natural Bayesian solution would be to quantify this uncertainty in a third-stage distribution, sometimes called a hyperprior.
- Denoting this distribution by $h(\boldsymbol{\eta})$, the desired posterior for $\boldsymbol{\theta}$ is now obtained by marginalizing over $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$:

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{y}) &= \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{\int p(\mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\eta}) \, d\boldsymbol{\eta}}{\int \int p(\mathbf{y}, \mathbf{u}, \boldsymbol{\eta}) \, d\boldsymbol{\eta} \, d\mathbf{u}} \\
&= \frac{\int f(\mathbf{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\eta}) h(\boldsymbol{\eta}) \, d\boldsymbol{\eta}}{\int \int f(\mathbf{y}|\mathbf{u}) p(\mathbf{u}|\boldsymbol{\eta}) h(\boldsymbol{\eta}) \, d\boldsymbol{\eta} \, d\mathbf{u}} \, .
\end{aligned}
$$

Hierarchical modeling

- The hyperprior for $\boldsymbol{\eta}$ might itself depend on a collection of unknown parameters $\boldsymbol{\lambda}$, resulting in a generalization of our three-stage model to one having a third-stage prior $h(\boldsymbol{\eta}|\boldsymbol{\lambda})$ and a fourth-stage hyperprior $g(\boldsymbol{\lambda})$...

Hierarchical modeling

- The hyperprior for $\boldsymbol{\eta}$ might itself depend on a collection of unknown parameters $\boldsymbol{\lambda}$, resulting in a generalization of our three-stage model to one having a third-stage prior $h(\boldsymbol{\eta}|\boldsymbol{\lambda})$ and a fourth-stage hyperprior $g(\boldsymbol{\lambda})$...
- This enterprise of specifying a model over several levels is called hierarchical modeling, which is often helpful when the data are nested:

Hierarchical modeling

- The hyperprior for $\boldsymbol{\eta}$ might itself depend on a collection of unknown parameters $\boldsymbol{\lambda}$, resulting in a generalization of our three-stage model to one having a third-stage prior $h(\boldsymbol{\eta}|\boldsymbol{\lambda})$ and a fourth-stage hyperprior $g(\boldsymbol{\lambda})$...
- This enterprise of specifying a model over several levels is called hierarchical modeling, which is often helpful when the data are nested:
- **Example:** Test scores Y_{ijk} for student k in classroom j of school i:

$$\begin{aligned} Y_{ijk}|\theta_{ij} &\sim N(\theta_{ij}, \sigma^2) \\ \theta_{ij}|\mu_i &\sim N(\mu_i, \tau^2) \\ \mu_i|\lambda &\sim N(\lambda, \kappa^2) \end{aligned}$$

Adding $p(\lambda)$ and possibly $p(\sigma^2, \tau^2, \kappa^2)$ completes the specification!

Prediction

- Returning to two-level models, we often write

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ ,$$

  since the likelihood may be multiplied by any constant (or any function of $\mathbf{y}$ alone) without altering $p(\boldsymbol{\theta}|\mathbf{y})$.

Prediction

- Returning to two-level models, we often write

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ ,$$

  since the likelihood may be multiplied by any constant (or any function of $\mathbf{y}$ alone) without altering $p(\boldsymbol{\theta}|\mathbf{y})$.

- If y_{n+1} is a future observation, independent of $\mathbf{y}$ given $\boldsymbol{\theta}$, then the predictive distribution for y_{n+1} is

$$p(y_{n+1}|\mathbf{y}) = \int f(y_{n+1}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ ,$$

  thanks to the conditional independence of y_{n+1} and $\mathbf{y}$.

Prediction

- Returning to two-level models, we often write

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \ ,$$

since the likelihood may be multiplied by any constant (or any function of $\mathbf{y}$ alone) without altering $p(\boldsymbol{\theta}|\mathbf{y})$.

- If y_{n+1} is a future observation, independent of $\mathbf{y}$ given $\boldsymbol{\theta}$, then the predictive distribution for y_{n+1} is

$$p(y_{n+1}|\mathbf{y}) = \int f(y_{n+1}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ ,$$

thanks to the conditional independence of y_{n+1} and $\mathbf{y}$.

- The naive frequentist would use $f(y_{n+1}|\widehat{\boldsymbol{\theta}})$ here, which is correct only for large n (i.e., when $p(\boldsymbol{\theta}|\mathbf{y})$ is a point mass at $\widehat{\boldsymbol{\theta}}$).

Prior Distributions

- Suppose we require a prior distribution for

  $\theta =$ true proportion of U.S. men who are HIV-positive.

Prior Distributions

- Suppose we require a prior distribution for

  $\theta =$ true proportion of U.S. men who are HIV-positive.

- We cannot appeal to the usual long-term frequency notion of probability – it is not possible to even imagine "running the HIV epidemic over again" and reobserving θ. Here θ is random only because it is unknown to us.

Prior Distributions

- Suppose we require a prior distribution for

  $\theta =$ true proportion of U.S. men who are HIV-positive.

- We cannot appeal to the usual long-term frequency notion of probability – it is not possible to even imagine “running the HIV epidemic over again” and reobserving θ. Here θ is random only because it is unknown to us.
- Bayesian analysis is predicated on such a belief in subjective probability and its quantification in a prior distribution $p(\theta)$. But:

Prior Distributions

- Suppose we require a prior distribution for

  $\theta =$ true proportion of U.S. men who are HIV-positive.

- We cannot appeal to the usual long-term frequency notion of probability – it is not possible to even imagine "running the HIV epidemic over again" and reobserving θ. Here θ is random only because it is unknown to us.
- Bayesian analysis is predicated on such a belief in subjective probability and its quantification in a prior distribution $p(\theta)$. But:
  - How to create such a prior?

Prior Distributions

- Suppose we require a prior distribution for

  $\theta =$ true proportion of U.S. men who are HIV-positive.

- We cannot appeal to the usual long-term frequency notion of probability – it is not possible to even imagine "running the HIV epidemic over again" and reobserving θ. Here θ is random only because it is unknown to us.
- Bayesian analysis is predicated on such a belief in subjective probability and its quantification in a prior distribution $p(\theta)$. But:
  - How to create such a prior?
  - Are "objective" choices available?

Elicited Priors

- Histogram approach: Assign probability masses to the "possible" values in such a way that their sum is 1, and their relative contributions reflect the experimenter's prior beliefs as closely as possible.

Elicited Priors

- Histogram approach: Assign probability masses to the "possible" values in such a way that their sum is 1, and their relative contributions reflect the experimenter's prior beliefs as closely as possible.
  - BUT: Awkward for continuous or unbounded θ.

Elicited Priors

- Histogram approach: Assign probability masses to the "possible" values in such a way that their sum is 1, and their relative contributions reflect the experimenter's prior beliefs as closely as possible.
  - BUT: Awkward for continuous or unbounded θ.
- Matching a functional form: Assume that the prior belongs to a parametric distributional family $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, choosing $\boldsymbol{\eta}$ so that the result matches the elicitee's true prior beliefs as nearly as possible.

Elicited Priors

- Histogram approach: Assign probability masses to the "possible" values in such a way that their sum is 1, and their relative contributions reflect the experimenter's prior beliefs as closely as possible.
  - BUT: Awkward for continuous or unbounded θ.
- Matching a functional form: Assume that the prior belongs to a parametric distributional family $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, choosing $\boldsymbol{\eta}$ so that the result matches the elicitee's true prior beliefs as nearly as possible.
  - This approach limits the effort required of the elicitee, and also overcomes the finite support problem inherent in the histogram approach...

Elicited Priors

- Histogram approach: Assign probability masses to the "possible" values in such a way that their sum is 1, and their relative contributions reflect the experimenter's prior beliefs as closely as possible.
  - BUT: Awkward for continuous or unbounded θ.
- Matching a functional form: Assume that the prior belongs to a parametric distributional family $p(\boldsymbol{\theta}|\boldsymbol{\eta})$, choosing $\boldsymbol{\eta}$ so that the result matches the elicitee's true prior beliefs as nearly as possible.
  - This approach limits the effort required of the elicitee, and also overcomes the finite support problem inherent in the histogram approach...
  - BUT: it may not be possible for the elicitee to "shoehorn" his or her prior beliefs into any of the standard parametric forms.

Conjugate Priors

- Defined as one that leads to a posterior distribution belonging to the same distributional family as the prior.

Conjugate Priors

- Defined as one that leads to a posterior distribution belonging to the same distributional family as the prior.
- Example 2.5: Suppose that X is distributed Poisson(θ), so that

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!},\ x \in \{0, 1, 2, \ldots\},\ \theta > 0.$$

Conjugate Priors

- Defined as one that leads to a posterior distribution belonging to the same distributional family as the prior.
- Example 2.5: Suppose that X is distributed Poisson(θ), so that

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}, \; x \in \{0, 1, 2, \ldots\}, \; \theta > 0.$$

- A reasonably flexible prior for θ having support on the positive real line is the $Gamma(\alpha, \beta)$ distribution,

$$p(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \; \theta > 0, \alpha > 0, \; \beta > 0,$$

Conjugate Priors

- The posterior is then

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)p(\theta) \\ &\propto \left(e^{-\theta}\theta^{x}\right)\left(\theta^{\alpha-1}e^{-\theta/\beta}\right) \\ &= \theta^{x+\alpha-1}e^{-\theta(1+1/\beta)}\ . \end{aligned}$$

Conjugate Priors

- The posterior is then

$$\begin{aligned} p(\theta|x) &\propto f(x|\theta)p(\theta) \\ &\propto \left(e^{-\theta}\theta^{x}\right)\left(\theta^{\alpha-1}e^{-\theta/\beta}\right) \\ &= \theta^{x+\alpha-1}e^{-\theta(1+1/\beta)}\ . \end{aligned}$$

- But this form is proportional to a $Gamma(\alpha', \beta')$, where

$$\alpha' = x + \alpha \ \text{ and } \ \beta' = (1 + 1/\beta)^{-1}.$$

Since this is the only function proportional to our form that integrates to 1 and density functions uniquely determine distributions, $p(\theta|x)$ must indeed be $Gamma(\alpha', \beta')$, and the gamma is the conjugate family for the Poisson likelihood.

Notes on conjugate priors

- Can often guess the conjugate prior by looking at the likelihood as a function of θ, instead of x.

Notes on conjugate priors

- Can often guess the conjugate prior by looking at the likelihood as a function of θ, instead of x.
- In higher dimensions, priors that are conditionally conjugate are often available (and helpful).

Notes on conjugate priors

- Can often guess the conjugate prior by looking at the likelihood as a function of θ, instead of x.
- In higher dimensions, priors that are conditionally conjugate are often available (and helpful).
- a finite mixture of conjugate priors may be sufficiently flexible (allowing multimodality, heavier tails, etc.) while still enabling simplified posterior calculations.

Noninformative Prior

– is one that does not favor one θ value over another

- Examples:

Noninformative Prior

– is one that does not favor one $\boldsymbol{\theta}$ value over another

- Examples:
  - $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_n\} \Rightarrow p(\theta_i) = 1/n, \ i = 1, \ldots, n$

Noninformative Prior

– is one that does not favor one $\boldsymbol{\theta}$ value over another

- Examples:
  - $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_n\} \Rightarrow p(\theta_i) = 1/n,\ i = 1, \ldots, n$
  - $\boldsymbol{\Theta} = [a, b],\ -\infty < a < b < \infty$
    $\Rightarrow p(\theta) = 1/(b - a),\ a < \theta < b$

Noninformative Prior

– is one that does not favor one $\boldsymbol{\theta}$ value over another

- Examples:
  - $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_n\} \Rightarrow p(\theta_i) = 1/n,\ i = 1, \ldots, n$
  - $\boldsymbol{\Theta} = [a, b],\ -\infty < a < b < \infty$
    $\Rightarrow p(\theta) = 1/(b-a),\ a < \theta < b$
  - $\boldsymbol{\Theta} = (-\infty, \infty) \Rightarrow p(\theta) = c,\ \text{any } c > 0$

Noninformative Prior

– is one that does not favor one $\boldsymbol{\theta}$ value over another

- Examples:
  - $\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_n\} \Rightarrow p(\theta_i) = 1/n,\ i = 1, \ldots, n$
  - $\boldsymbol{\Theta} = [a, b],\ -\infty < a < b < \infty$
    $\Rightarrow p(\theta) = 1/(b-a),\ a < \theta < b$
  - $\boldsymbol{\Theta} = (-\infty, \infty) \Rightarrow p(\theta) = c,\ \text{any } c > 0$

    This is an improper prior (does not integrate to 1), but its use can still be legitimate if $\int f(\mathbf{x}|\theta) d\theta = K < \infty$, since then

$$p(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta) \cdot c}{\int f(\mathbf{x}|\theta) \cdot c\, d\theta} = \frac{f(\mathbf{x}|\theta)}{K},$$

so the posterior is just the renormalized likelihood!

Jeffreys Prior

- another noninformative prior, given in the univariate case by

$$p(\theta) = [I(\theta)]^{1/2} \ ,$$

where $I(\theta)$ is the expected Fisher information in the model, namely

$$I(\theta) = -E_{\mathbf{x}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)\right] \ .$$

Jeffreys Prior

- another noninformative prior, given in the univariate case by

$$p(\theta) = [I(\theta)]^{1/2} \; ,$$

  where $I(\theta)$ is the expected Fisher information in the model, namely

$$I(\theta) = -E_{\mathbf{x}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{x}|\theta)\right] \; .$$

- Unlike the uniform, the Jeffreys prior is invariant to 1-1 transformations. That is, computing the Jeffreys prior for some 1-1 transformation $\gamma = g(\theta)$ directly produces the same answer as computing the Jeffreys prior for θ and subsequently performing the usual Jacobian transformation to the γ scale (see p.54, problem 7).

Other Noninformative Priors

- When $f(x|\theta) = f(x-\theta)$ (location parameter family),

$$p(\theta) = 1, \ \theta \in \Re$$

is invariant under location transformations ($Y = X + c$).

Other Noninformative Priors

- When $f(x|\theta) = f(x-\theta)$ (location parameter family),

$$p(\theta) = 1,\ \theta \in \Re$$

  is invariant under location transformations ($Y = X + c$).

- When $f(x|\sigma) = \frac{1}{\sigma} f(\frac{x}{\sigma})$, $\sigma > 0$ (scale parameter family),

$$p(\sigma) = \frac{1}{\sigma},\ \sigma > 0$$

  is invariant under scale transformations ($Y = cX,\ c > 0$).

Other Noninformative Priors

- When $f(x|\theta) = f(x - \theta)$ (location parameter family),

$$p(\theta) = 1,\ \theta \in \Re$$

is invariant under location transformations ($Y = X + c$).

- When $f(x|\sigma) = \frac{1}{\sigma} f(\frac{x}{\sigma})$, $\sigma > 0$ (scale parameter family),

$$p(\sigma) = \frac{1}{\sigma},\ \sigma > 0$$

is invariant under scale transformations ($Y = cX,\ c > 0$).

- When $f(x|\theta, \sigma) = \frac{1}{\sigma} f(\frac{x-\theta}{\sigma})$ (location-scale family), prior "independence" suggests

$$p(\theta, \sigma) = \frac{1}{\sigma},\ \theta \in \Re,\ \sigma > 0\ .$$

Bayesian Inference: Point Estimation

- Easy! Simply choose an appropriate distributional summary: posterior mean, median, or mode.

Bayesian Inference: Point Estimation

- Easy! Simply choose an appropriate distributional summary: posterior mean, median, or mode.
- Mode is often easiest to compute (no integration), but is often least representative of “middle”, especially for one-tailed distributions.

Bayesian Inference: Point Estimation

- Easy! Simply choose an appropriate distributional summary: posterior mean, median, or mode.
- Mode is often easiest to compute (no integration), but is often least representative of “middle”, especially for one-tailed distributions.
- Mean has the opposite property, tending to "chase" heavy tails (just like the sample mean $\bar{X}$)

Bayesian Inference: Point Estimation

- Easy! Simply choose an appropriate distributional summary: posterior mean, median, or mode.
- Mode is often easiest to compute (no integration), but is often least representative of “middle”, especially for one-tailed distributions.
- Mean has the opposite property, tending to "chase" heavy tails (just like the sample mean $\bar{X}$)
- Median is probably the best compromise overall, though can be awkward to compute, since it is the solution θ^{median} to

$$\int_{-\infty}^{\theta^{median}} p(\theta|x)\, d\theta = \frac{1}{2}\ .$$

Example: The General Linear Model

- Let $\mathbf{Y}$ be an $n \times 1$ data vector, X an $n \times p$ matrix of covariates, and adopt the likelihood and prior structure,

$$\mathbf{Y}|\boldsymbol{\beta} \sim N_n\left(X\boldsymbol{\beta}, \Sigma\right) \text{ and } \boldsymbol{\beta} \sim N_p\left(A\boldsymbol{\alpha}, V\right)$$

Example: The General Linear Model

- Let $\mathbf{Y}$ be an $n \times 1$ data vector, X an $n \times p$ matrix of covariates, and adopt the likelihood and prior structure,

$$\mathbf{Y}|\boldsymbol{\beta} \sim N_n\left(X\boldsymbol{\beta}, \Sigma\right) \text{ and } \boldsymbol{\beta} \sim N_p\left(A\boldsymbol{\alpha}, V\right)$$

- Then the posterior distribution of $\boldsymbol{\beta}|\mathbf{Y}$ is

$$\boldsymbol{\beta}|Y \sim N\left(D\mathbf{d}, D\right), \text{ where}$$

$$D^{-1} = X^T\Sigma^{-1}X + V^{-1} \text{ and } \mathbf{d} = X^T\Sigma^{-1}\mathbf{Y} + V^{-1}A\boldsymbol{\alpha}.$$

Example: The General Linear Model

- Let $\mathbf{Y}$ be an $n \times 1$ data vector, X an $n \times p$ matrix of covariates, and adopt the likelihood and prior structure,

$$\mathbf{Y}|\boldsymbol{\beta} \sim N_n\left(X\boldsymbol{\beta}, \Sigma\right) \text{ and } \boldsymbol{\beta} \sim N_p\left(A\boldsymbol{\alpha}, V\right)$$

- Then the posterior distribution of $\boldsymbol{\beta}|\mathbf{Y}$ is

$$\boldsymbol{\beta}|Y \sim N\left(D\mathbf{d}, D\right), \text{ where}$$

$$D^{-1} = X^T\Sigma^{-1}X + V^{-1} \text{ and } \mathbf{d} = X^T\Sigma^{-1}\mathbf{Y} + V^{-1}A\boldsymbol{\alpha}.$$

- $V^{-1} = 0$ delivers a "flat" prior; if $\Sigma = \sigma^2 I_p$, we get

$$\boldsymbol{\beta}|Y \sim N\left(\hat{\boldsymbol{\beta}}\,,\, \sigma^2(X'X)^{-1}\right), \text{ where}$$

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} \Longleftrightarrow \text{usual likelihood approach!}$$

Bayesian Inference: Interval Estimation

- The Bayesian analogue of a frequentist CI is referred to as a credible set: a $100 \times (1-\alpha)$% credible set for $\boldsymbol{\theta}$ is a subset C of $\boldsymbol{\Theta}$ such that

$$1-\alpha \leq P(C|\mathbf{y}) = \int_C p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta} \ .$$

Bayesian Inference: Interval Estimation

- The Bayesian analogue of a frequentist CI is referred to as a credible set: a $100 \times (1-\alpha)$% credible set for $\boldsymbol{\theta}$ is a subset C of $\boldsymbol{\Theta}$ such that

$$1-\alpha \leq P(C|\mathbf{y}) = \int_C p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \; .$$

- In continuous settings, we can obtain coverage exactly $1-\alpha$ at minimum size via the highest posterior density (HPD) credible set,

$$C = \{\boldsymbol{\theta} \in \boldsymbol{\Theta} : \; p(\boldsymbol{\theta}|\mathbf{y}) \geq k(\alpha)\} \; ,$$

where $k(\alpha)$ is the largest constant such that

$$P(C|\mathbf{y}) \geq 1-\alpha \; .$$

Interval Estimation (cont'd)

- Simpler alternative: the equal-tail set, which takes the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$.

Interval Estimation (cont'd)

- Simpler alternative: the equal-tail set, which takes the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$.
- Specifically, consider q_L and q_U, the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$:

$$\int_{-\infty}^{q_L} p(\theta|\mathbf{y})d\theta = \alpha/2 \text{ and } \int_{q_U}^{\infty} p(\theta|\mathbf{y})d\theta = 1-\alpha/2 \,.$$

  Then clearly $P(q_L < \theta < q_U|\mathbf{y}) = 1-\alpha$; our confidence that θ lies in (q_L, q_U) is $100 \times (1-\alpha)$%. Thus this interval is a $100 \times (1-\alpha)$% credible set ("Bayesian CI") for θ.

Interval Estimation (cont'd)

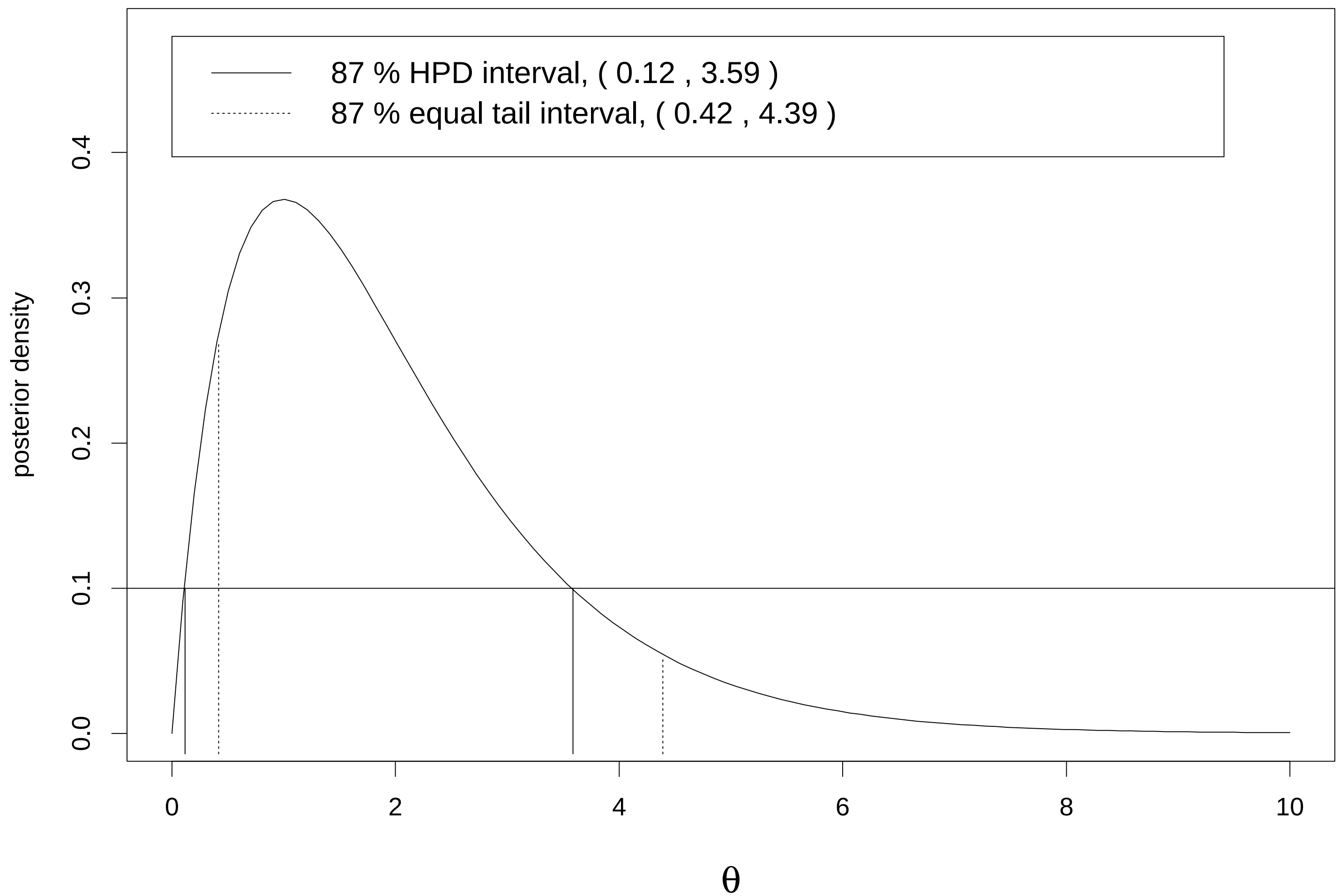
- Simpler alternative: the equal-tail set, which takes the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$.
- Specifically, consider q_L and q_U, the $\alpha/2$- and $(1-\alpha/2)$-quantiles of $p(\theta|\mathbf{y})$:

$$\int_{-\infty}^{q_L} p(\theta|\mathbf{y})d\theta = \alpha/2 \;\text{ and }\; \int_{q_U}^{\infty} p(\theta|\mathbf{y})d\theta = 1-\alpha/2 \,.$$

  Then clearly $P(q_L < \theta < q_U|\mathbf{y}) = 1-\alpha$; our confidence that θ lies in (q_L, q_U) is $100 \times (1-\alpha)$%. Thus this interval is a $100 \times (1-\alpha)$% credible set ("Bayesian CI") for θ.
- This interval is relatively easy to compute, and enjoys a direct interpretation ("The probability that θ lies in (q_L, q_U) is $(1-\alpha)$") that the frequentist interval does not.

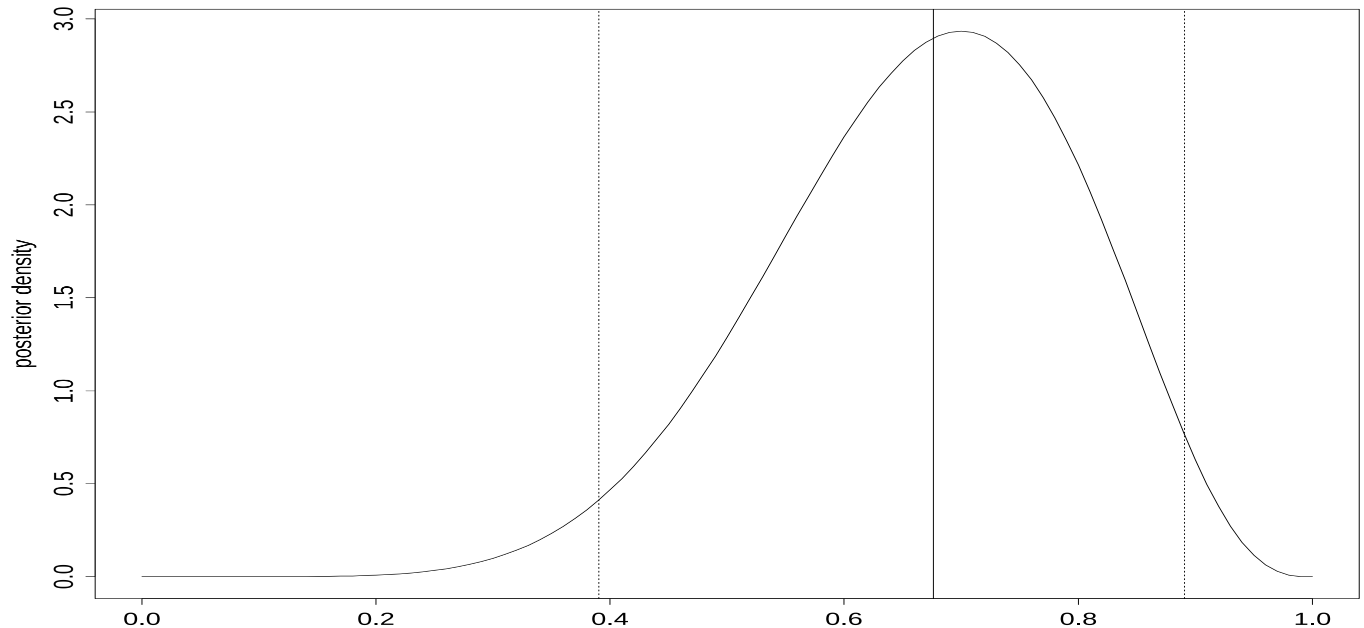
Interval Estimation: Example

Using a $Gamma(2, 1)$ posterior distribution and $k(\alpha) = 0.1$:

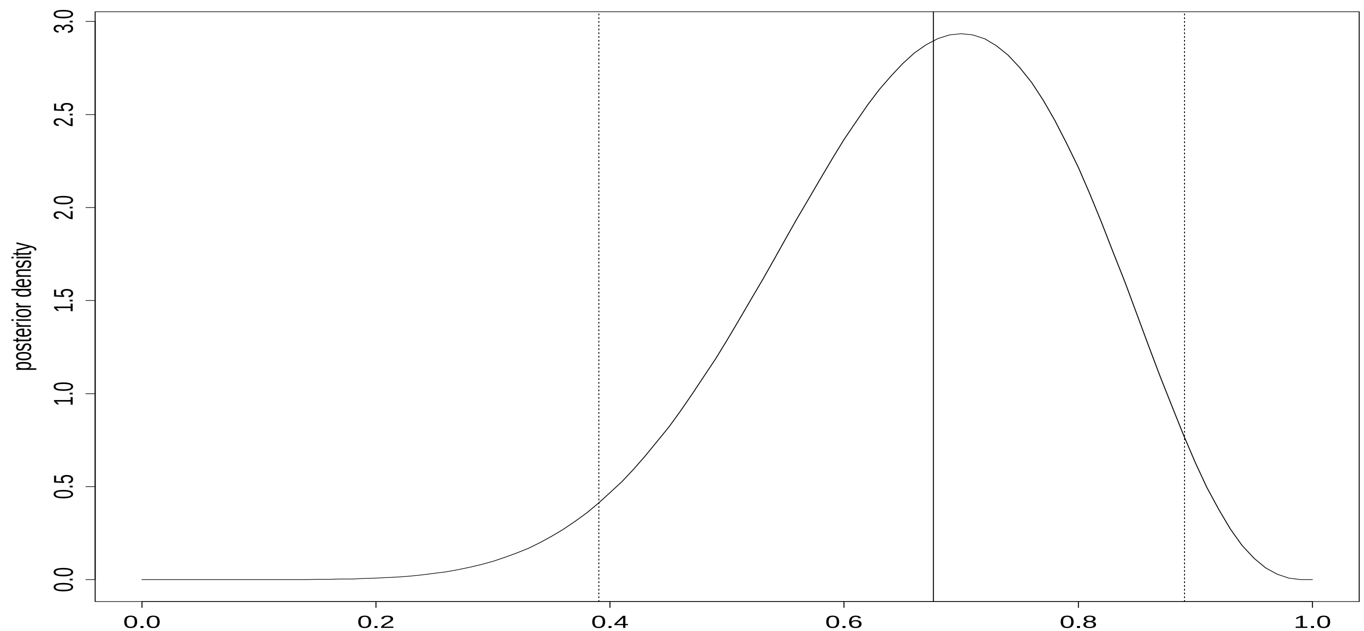


Equal tail interval is a bit wider, but easier to compute (just two gamma quantiles), and also transformation invariant.

Ex: $Y \sim Bin(10, \theta)$, $\theta \sim U(0, 1)$, $y_{obs} = 7$



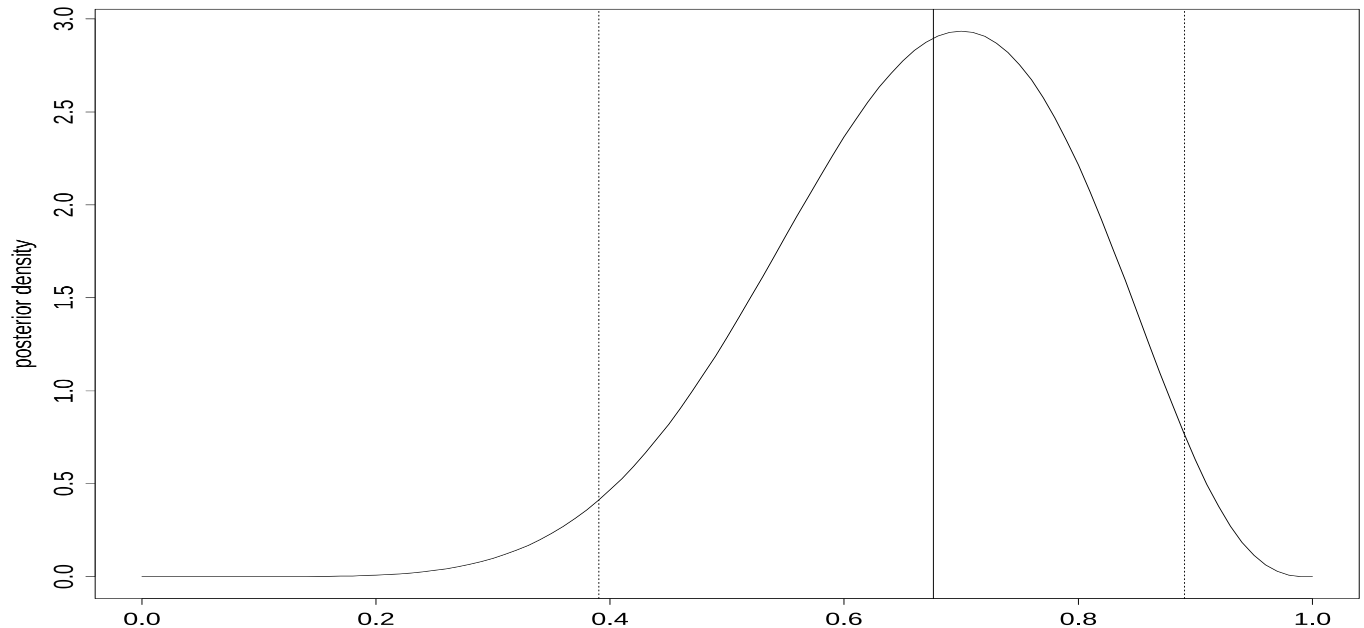
Ex: $Y \sim Bin(10, \theta)$, $\theta \sim U(0, 1)$, $y_{obs} = 7$



Plot $Beta(y_{obs} + 1, n - y_{obs} + 1) = Beta(8, 4)$ posterior in R/S:

```
> theta <- seq(from=0, to=1, length=101)
> yobs <- 7; n <- 10
> plot(theta, dbeta(theta, yobs+1, n-yobs+1), type="l")
```

Ex: $Y \sim Bin(10, \theta)$, $\theta \sim U(0, 1)$, $y_{obs} = 7$



Plot $Beta(y_{obs} + 1, n - y_{obs} + 1) = Beta(8, 4)$ posterior in R/S:

```
> theta <- seq(from=0, to=1, length=101)
> yobs <- 7; n <- 10
> plot(theta, dbeta(theta, yobs+1, n-yobs+1), type="l")
```

Add 95% equal-tail Bayesian CI (dotted vertical lines):

```
> abline(v=qbeta(.5, yobs+1, n-yobs+1))
> abline(v=qbeta(c(.025, .975), yobs+1, n-yobs+1), lty=2)
```

Bayesian hypothesis testing

- Classical approach bases accept/reject decision on

  p-value $= P\{T(\mathbf{Y})$ more "extreme" than $T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\}$,

  where "extremeness" is in the direction of H_A

Bayesian hypothesis testing

- Classical approach bases accept/reject decision on

  p-value $= P\{T(\mathbf{Y})$ more "extreme" than $T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\}$,

  where "extremeness" is in the direction of H_A
- Several *troubles* with this approach:

Bayesian hypothesis testing

- Classical approach bases accept/reject decision on

  p-value $= P\{T(\mathbf{Y})$ more "extreme" than $T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\}$,

  where "extremeness" is in the direction of H_A

- Several *troubles* with this approach:
  - hypotheses must be nested

Bayesian hypothesis testing

- Classical approach bases accept/reject decision on

  p-value $= P\{T(\mathbf{Y})$ more "extreme" than $T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\}$,

  where "extremeness" is in the direction of H_A
- Several *troubles* with this approach:
  - hypotheses must be nested
  - p-value can only offer evidence against the null

Bayesian hypothesis testing

- Classical approach bases accept/reject decision on

  p-value $= P\{T(\mathbf{Y})$ more "extreme" than $T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\}$,

  where "extremeness" is in the direction of H_A
- Several *troubles* with this approach:
  - hypotheses must be nested
  - p-value can only offer evidence against the null
  - p-value is not the "probability that H_0 is true" (but is often erroneously interpreted this way)

Bayesian hypothesis testing

- Classical approach bases accept/reject decision on

  p-value $= P\{T(\mathbf{Y})$ more "extreme" than $T(\mathbf{y}_{obs})|\boldsymbol{\theta}, H_0\}$,

  where "extremeness" is in the direction of H_A
- Several *troubles* with this approach:
  - hypotheses must be nested
  - p-value can only offer evidence against the null
  - p-value is not the "probability that H_0 is true" (but is often erroneously interpreted this way)
  - As a result of the dependence on "more extreme" $T(\mathbf{Y})$ values, two experiments with different designs but identical likelihoods could result in different p-values, violating the Likelihood Principle!

Bayesian hypothesis testing (cont'd)

- Bayesian approach: Select the model with the largest posterior probability, $P(M_i|\mathbf{y}) = p(\mathbf{y}|M_i)p(M_i)/p(\mathbf{y})$,

$$\text{where} \quad p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \ .$$

Bayesian hypothesis testing (cont'd)

- Bayesian approach: Select the model with the largest posterior probability, $P(M_i|\mathbf{y}) = p(\mathbf{y}|M_i)p(M_i)/p(\mathbf{y})$,

  where $$p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \ .$$

- For two models, the quantity commonly used to summarize these results is the Bayes factor,

  $$BF = \frac{P(M_1|\mathbf{y})/P(M_2|\mathbf{y})}{P(M_1)/P(M_2)} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \ ,$$

  i.e., the likelihood ratio if both hypotheses are simple

Bayesian hypothesis testing (cont'd)

- Bayesian approach: Select the model with the largest posterior probability, $P(M_i|\mathbf{y}) = p(\mathbf{y}|M_i)p(M_i)/p(\mathbf{y})$,

$$\text{where} \quad p(\mathbf{y}|M_i) = \int f(\mathbf{y}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \ .$$

- For two models, the quantity commonly used to summarize these results is the Bayes factor,

$$BF = \frac{P(M_1|\mathbf{y})/P(M_2|\mathbf{y})}{P(M_1)/P(M_2)} = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)} \ ,$$

  i.e., the likelihood ratio if both hypotheses are simple

- Problem: If $\pi_i(\boldsymbol{\theta}_i)$ is improper, then $p(\mathbf{y}|M_i)$ necessarily is as well $\Longrightarrow$ BF is not well-defined!...

Bayesian hypothesis testing (cont'd)

When the BF is not well-defined, several alternatives:

- Modify the definition of BF: partial Bayes factor, fractional Bayes factor (text, pp.41-42)

Bayesian hypothesis testing (cont'd)

When the BF is not well-defined, several alternatives:

- Modify the definition of BF: partial Bayes factor, fractional Bayes factor (text, pp.41-42)
- Switch to the conditional predictive distribution,

$$f(y_i|\mathbf{y}_{(i)}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{(i)})} = \int f(y_i|\boldsymbol{\theta}, \mathbf{y}_{(i)}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \ ,$$

  which will be proper if $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ is. Assess model fit via plots or a suitable summary (say, $\prod_{i=1}^{n} f(y_i|\mathbf{y}_{(i)})$).

Bayesian hypothesis testing (cont'd)

When the BF is not well-defined, several alternatives:

- Modify the definition of BF: partial Bayes factor, fractional Bayes factor (text, pp.41-42)
- Switch to the conditional predictive distribution,

$$f(y_i|\mathbf{y}_{(i)}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{(i)})} = \int f(y_i|\boldsymbol{\theta}, \mathbf{y}_{(i)}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \ ,$$

  which will be proper if $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ is. Assess model fit via plots or a suitable summary (say, $\prod_{i=1}^{n} f(y_i|\mathbf{y}_{(i)})$).

- Penalized likelihood criteria: the Akaike information criterion (AIC), Bayesian information criterion (BIC), or Deviance information criterion (DIC).

Bayesian hypothesis testing (cont'd)

When the BF is not well-defined, several alternatives:

- Modify the definition of BF: partial Bayes factor, fractional Bayes factor (text, pp.41-42)
- Switch to the conditional predictive distribution,

$$f(y_i|\mathbf{y}_{(i)}) = \frac{f(\mathbf{y})}{f(\mathbf{y}_{(i)})} = \int f(y_i|\boldsymbol{\theta}, \mathbf{y}_{(i)}) p(\boldsymbol{\theta}|\mathbf{y}_{(i)}) d\boldsymbol{\theta} \ ,$$

  which will be proper if $p(\boldsymbol{\theta}|\mathbf{y}_{(i)})$ is. Assess model fit via plots or a suitable summary (say, $\prod_{i=1}^{n} f(y_i|\mathbf{y}_{(i)})$).
- Penalized likelihood criteria: the Akaike information criterion (AIC), Bayesian information criterion (BIC), or Deviance information criterion (DIC).
- IOU on all this – Chapter 6!

Example: Consumer preference data

- Suppose 16 taste testers compare two types of ground beef patty (one stored in a deep freeze, the other in a less expensive freezer). The food chain is interested in whether storage in the higher-quality freezer translates into a "substantial improvement in taste."

Example: Consumer preference data

- Suppose 16 taste testers compare two types of ground beef patty (one stored in a deep freeze, the other in a less expensive freezer). The food chain is interested in whether storage in the higher-quality freezer translates into a "substantial improvement in taste."
- Experiment: In a test kitchen, the patties are defrosted and prepared by a single chef/statistician, who randomizes the order in which the patties are served in double-blind fashion.

Example: Consumer preference data

- Suppose 16 taste testers compare two types of ground beef patty (one stored in a deep freeze, the other in a less expensive freezer). The food chain is interested in whether storage in the higher-quality freezer translates into a "substantial improvement in taste."
- Experiment: In a test kitchen, the patties are defrosted and prepared by a single chef/statistician, who randomizes the order in which the patties are served in double-blind fashion.
- Result: 13 of the 16 testers state a preference for the more expensive patty.

Example: Consumer preference data

- Likelihood: Let

$$\theta = \text{prob. consumers prefer more expensive patty}$$

$$Y_i = \begin{cases} 1 & \text{if tester } i \text{ prefers more expensive patty} \\ 0 & \text{otherwise} \end{cases}$$

Example: Consumer preference data

- Likelihood: Let

$$\begin{aligned}\theta &= \text{prob. consumers prefer more expensive patty}\\ Y_i &= \begin{cases} 1 & \text{if tester } i \text{ prefers more expensive patty} \\ 0 & \text{otherwise} \end{cases}\end{aligned}$$

- Assuming independent testers and constant θ, then if $X = \sum_{i=1}^{16} Y_i$, we have $X|\theta \sim Binomial(16, \theta)$,

$$f(x|\theta) = \binom{16}{x} \theta^x (1-\theta)^{16-x} .$$

Example: Consumer preference data

- Likelihood: Let

$$\begin{aligned} \theta &= \text{prob. consumers prefer more expensive patty} \\ Y_i &= \begin{cases} 1 & \text{if tester } i \text{ prefers more expensive patty} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

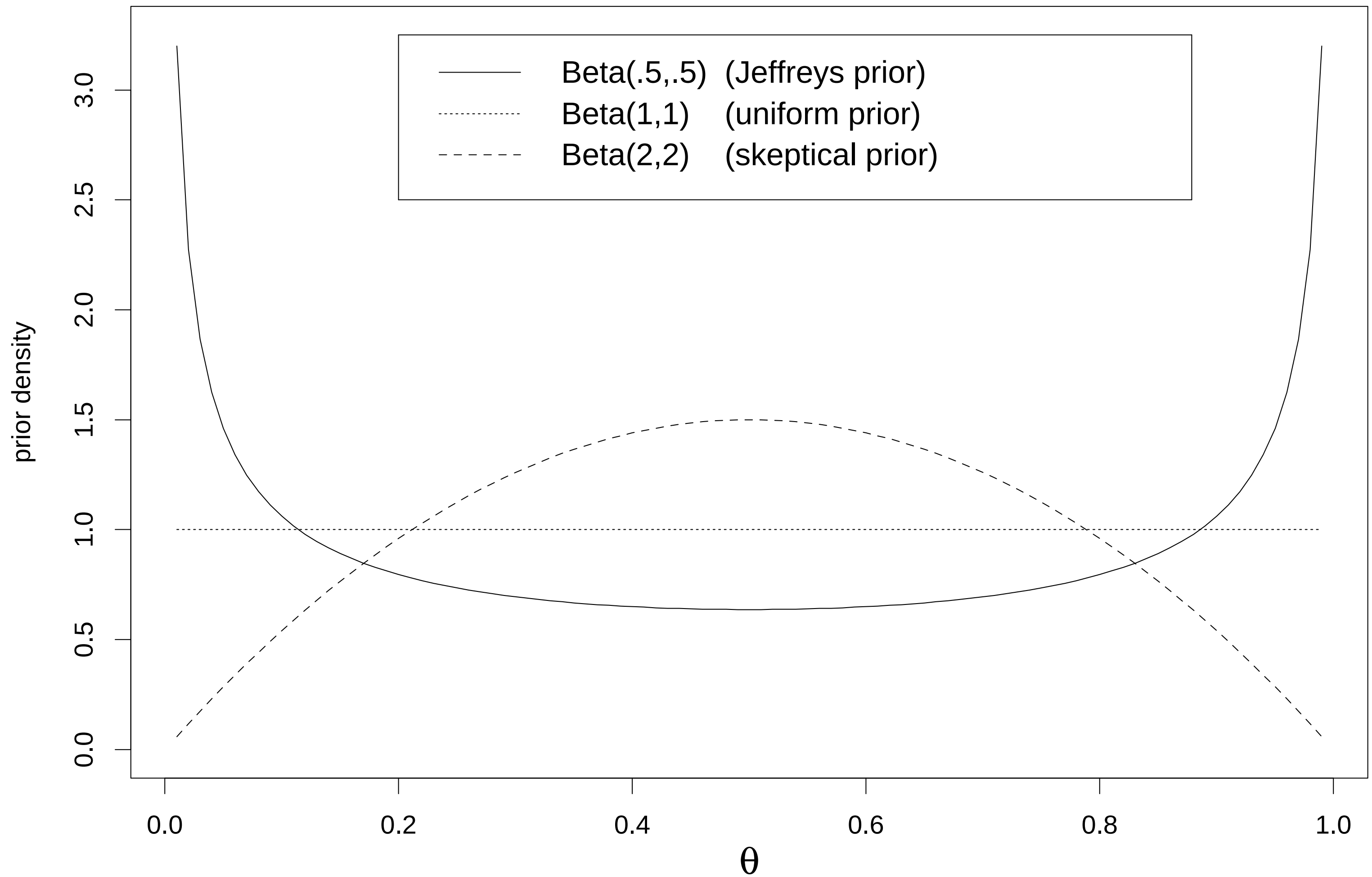
- Assuming independent testers and constant θ, then if $X = \sum_{i=1}^{16} Y_i$, we have $X|\theta \sim Binomial(16, \theta)$,

$$f(x|\theta) = \binom{16}{x} \theta^x (1-\theta)^{16-x} \; .$$

- The beta distribution offers a conjugate family, since

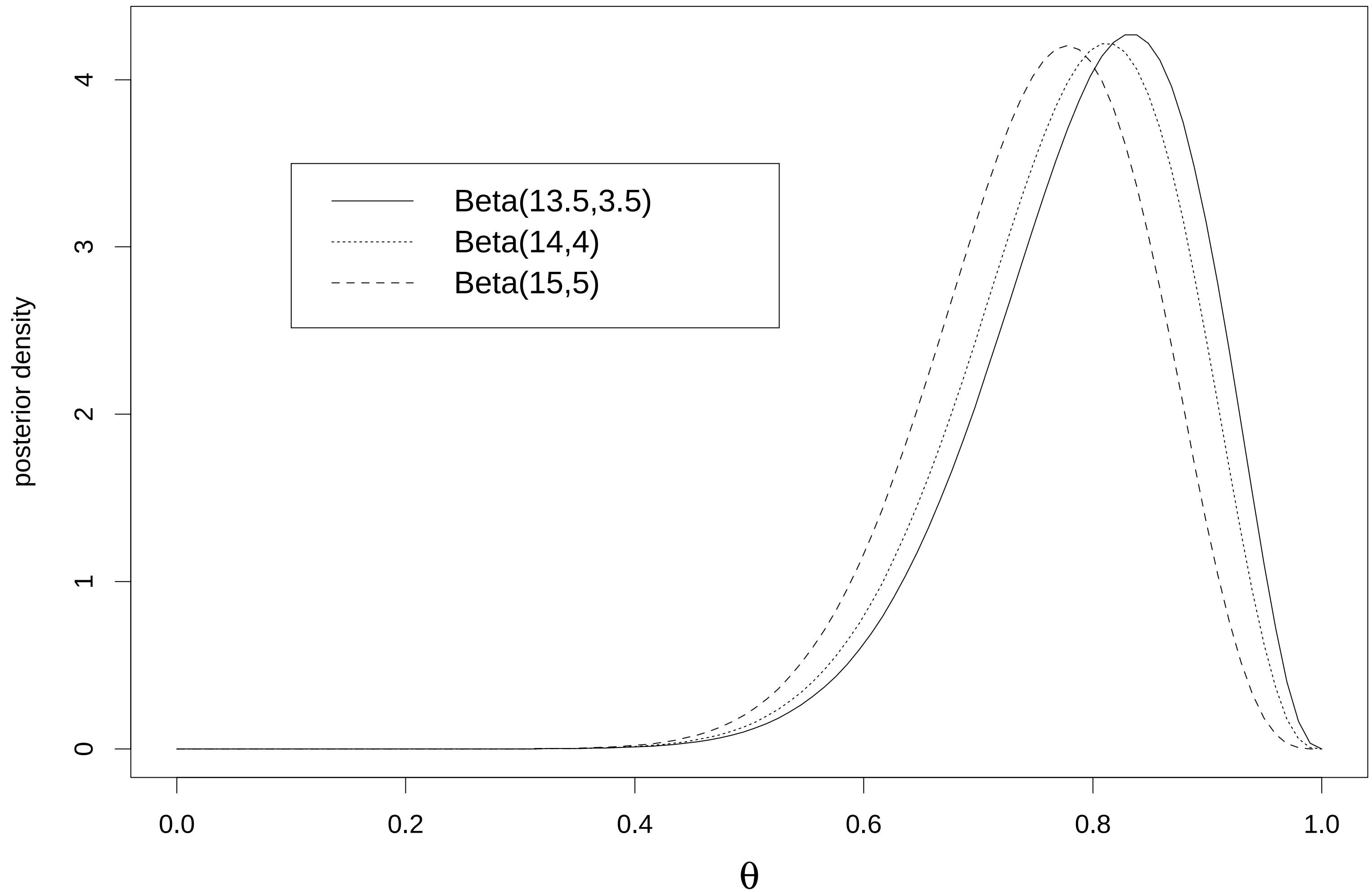
$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \; .$$

Three "minimally informative" priors



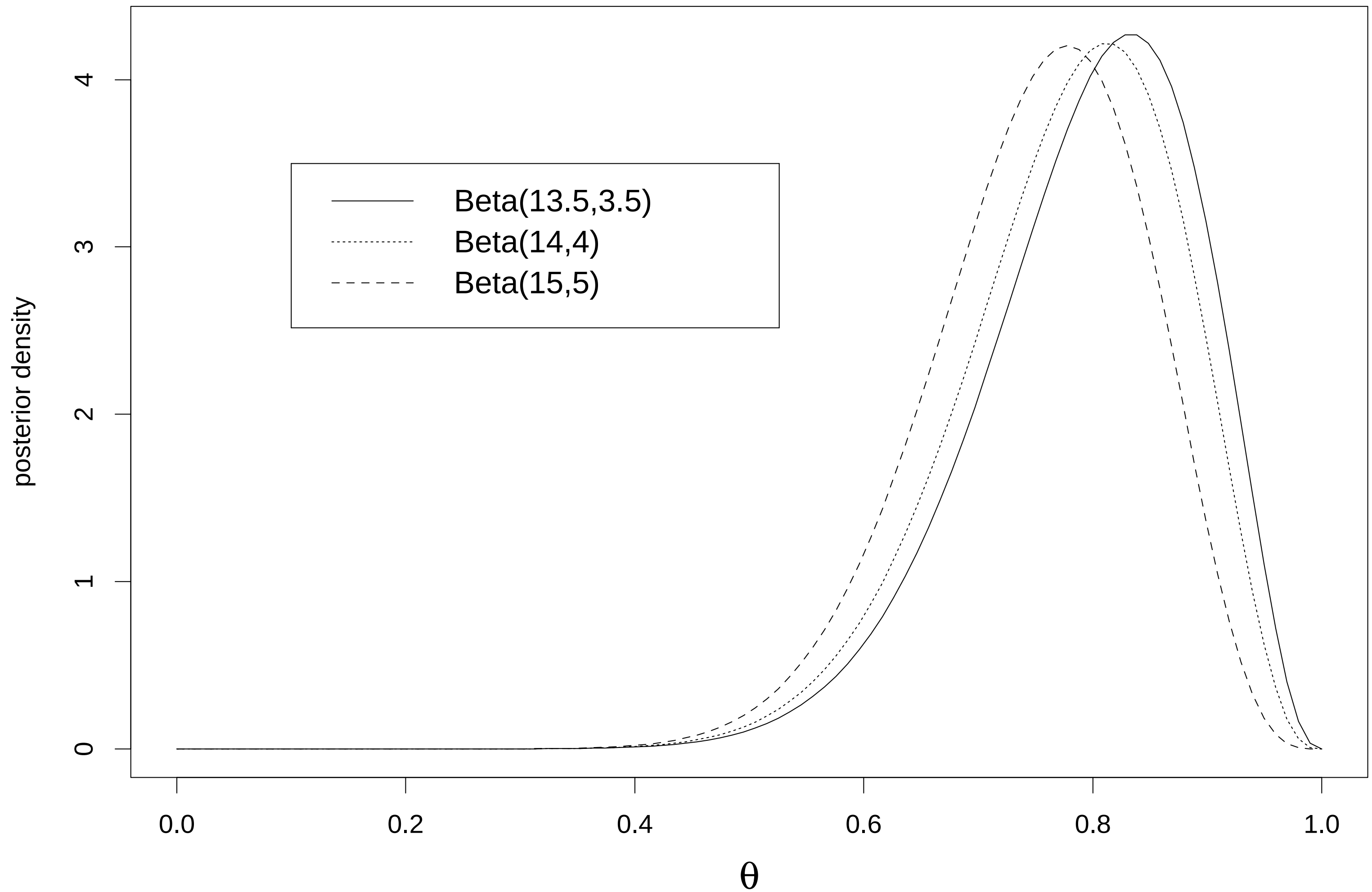
The posterior is then $Beta(x + \alpha, 16 - x + \beta)$...

Three corresponding posteriors



- Note ordering of posteriors; consistent with priors.

Three corresponding posteriors



- Note ordering of posteriors; consistent with priors.
- All three produce 95% equal-tail credible intervals that exclude 0.5 $\Rightarrow$ there is an improvement in taste.

Posterior summaries

Prior distribution	Posterior quantile .025	.500	.975	$P(\theta > .6 \mid x)$
$Beta(.5, .5)$	0.579	0.806	0.944	0.964
$Beta(1, 1)$	0.566	0.788	0.932	0.954
$Beta(2, 2)$	0.544	0.758	0.909	0.930

Posterior summaries

Prior distribution	Posterior quantile .025	.500	.975	$P(\theta > .6 \mid x)$
$Beta(.5, .5)$	0.579	0.806	0.944	0.964
$Beta(1, 1)$	0.566	0.788	0.932	0.954
$Beta(2, 2)$	0.544	0.758	0.909	0.930

- Suppose we define "*substantial* improvement in taste" as $\theta \geq 0.6$. Then under the uniform prior, the Bayes factor in favor of $M_1 : \theta \geq 0.6$ over $M_2 : \theta < 0.6$ is

$$BF = \frac{0.954/0.046}{0.4/0.6} = 31.1 \; ,$$

or fairly strong evidence (adjusted odds about 30:1) in favor of a substantial improvement in taste.