Single-parameter Models

Chapter 2.4-2.9, BDA3

A prior distribution must be chosen to perform Bayesian inference.

How do we decide what the prior distribution for θ should be?

Gelman et al. (2013) suggest two basic interpretations for prior distributions:

- **Population** interpretation: the prior distribution represents a population of possible parameter values from which the θ of current interest has been drawn.
- **State of knowledge** interpretation: the prior distribution describes our knowledge and uncertainty about θ as if its value were a random realization from the prior distribution.

Approaches to choosing a prior distribution can be partitioned in various ways:

Subjective vs Objective

Subjective prior: The prior distribution is chosen based on your beliefs about the behavior of θ.

Objective prior: The prior distribution is chosen in a systematic way to minimize personal bias in the result.

Informative vs Noninformative

Informative prior: The prior distribution is specifically chosen based on available information or knowledge to help make the posterior distribution more precise.

Noninformative prior: The prior distribution is intentionally chosen to be vague so that it plays only a minimal role in the posterior distribution.

- Let the data “speak for themselves”.

Proper vs improper

Proper prior: The prior distribution is a valid statistical distribution and does not depend on the data.

Improper prior: The prior is NOT a valid statistical distribution (usually it does not integrate to 1).

Caution: An improper prior can sometimes lead to situations where the posterior doesn't integrate to a finite constant! However, sometimes it is fine, e.g., let $p(\sigma^2) = 1/\sigma^2$ be the prior density for σ^2 when the data distribution of $y|\sigma^2 \sim N(\theta, \sigma^2)$ with θ assumed to be known.

Conjugate vs non-conjugate

Conjugate prior: The posterior distribution has the same parametric form as the prior distribution.

Non-conjugate prior: Any prior that’s not a conjugate prior.

The prior distribution should include all plausible values of θ, though it doesn’t need to be centered around the true value since the data will far outweigh any reasonable prior.

The parameters of the prior distribution are known as **hyperparameters**.

Definition of conjugate prior

Let F be a family of distributions containing the data distribution, $p(y|\theta)$, and H be a family of distributions containing the prior distribution for θ, $p(\theta)$.

- $N(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, $\sigma^2 > 0$, is a family of distributions.
- A $N(0,1)$ a member of that family.

Distributional family H is conjugate for distributional family F if $p(y|\theta) \in F$ and $p(\theta) \in H$ ensures $p(\theta|y) \in H$.

- e.g., $p(y|\theta) \in$ Binomial and $p(\theta) \in$ Beta implies $p(\theta|y) \in$ Beta.

We are most interested in *natural* conjugate prior families, which arise by taking H to be the set of all densities having the same functional form as the likelihood.

Typically, only data distributions from exponential families have conjugate prior distributions.

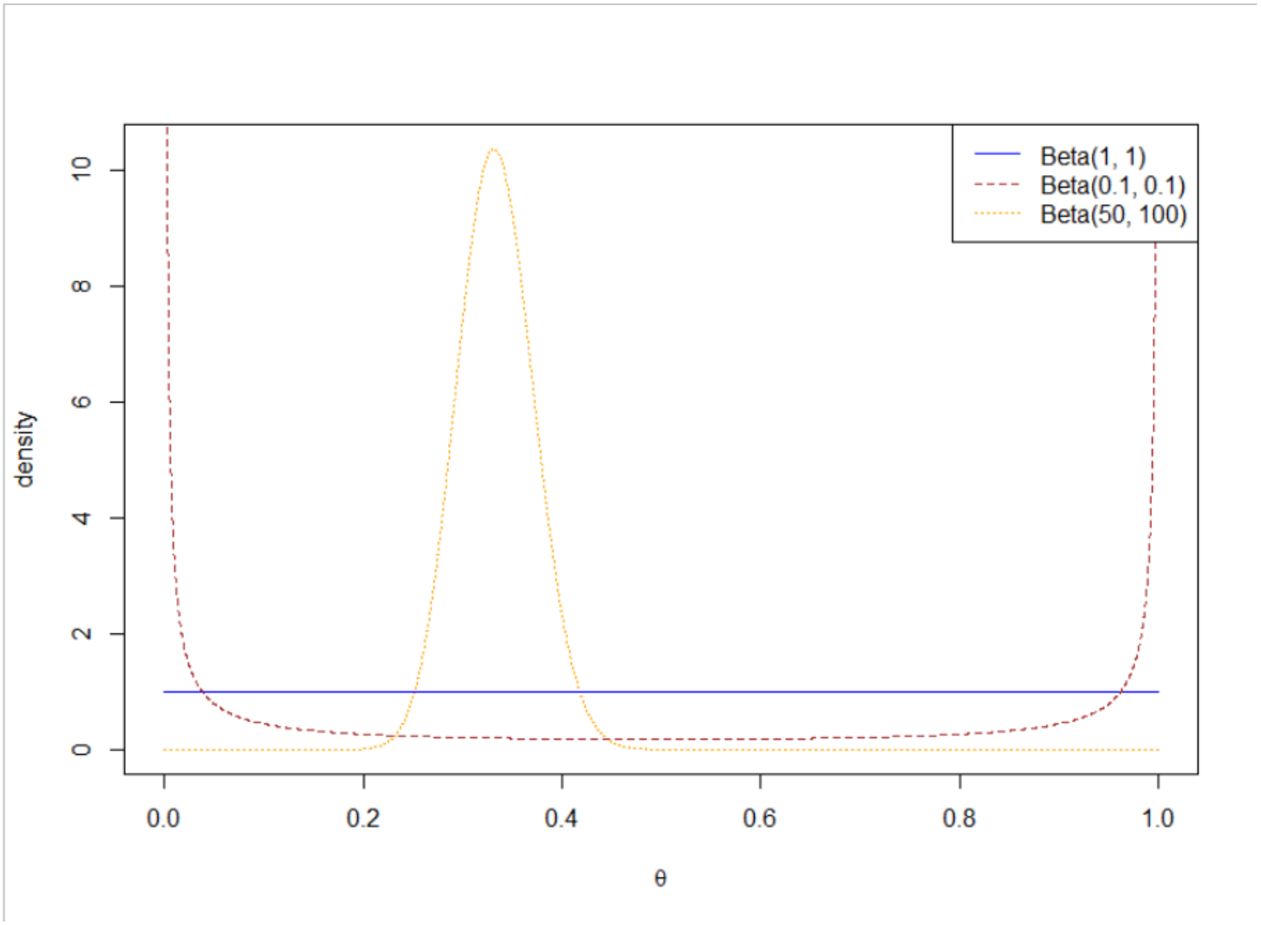
Advantages:

- Closed-form solution for the posterior distribution.
- Simplifies computations.
- Often a “good enough” approximation of your real beliefs.
- Can often be interpreted as “additional data” in the model.

Disadvantages:

- Can be unrealistic.
- May be impossible for complex models.

Conjugate priors may be sufficiently flexible to be useful in your analysis!



Conjugate prior examples

Beta-Binomial conjugate pair

Data distribution: $y|\theta \sim \text{Bin}(n, \theta)$

Prior distribution: $\theta \sim \text{Beta}(\alpha, \beta)$

Posterior distribution: $\theta|y \sim \text{Beta}(y + \alpha, n - y + \beta)$

Posterior predictive distribution: $P(\tilde{y} = 1|y) = \frac{y+\alpha}{n+\alpha+\beta}$.

Example: Placenta previa

Placenta previa is a condition in which the placenta of an unborn child is implanted very low in the uterus, obstructing the child from a normal vaginal delivery. An early study concerning the sex of placenta previa births in Germany found that of a total of 980 births, 437 were female. How strong is the evidence that the proportion of female births in the population of placenta previa births is less than 0.485 (the proportion in the general population)?

Analysis using a uniform prior

Assuming a Uniform(0, 1) prior distribution for θ, and a binomial sampling model for $y|\theta$, then the posterior distribution for θ is Beta$(y + 1, n - y + 1)$= Beta$(438, 544)$.

The mean is 438/(438 + 544) = .446.

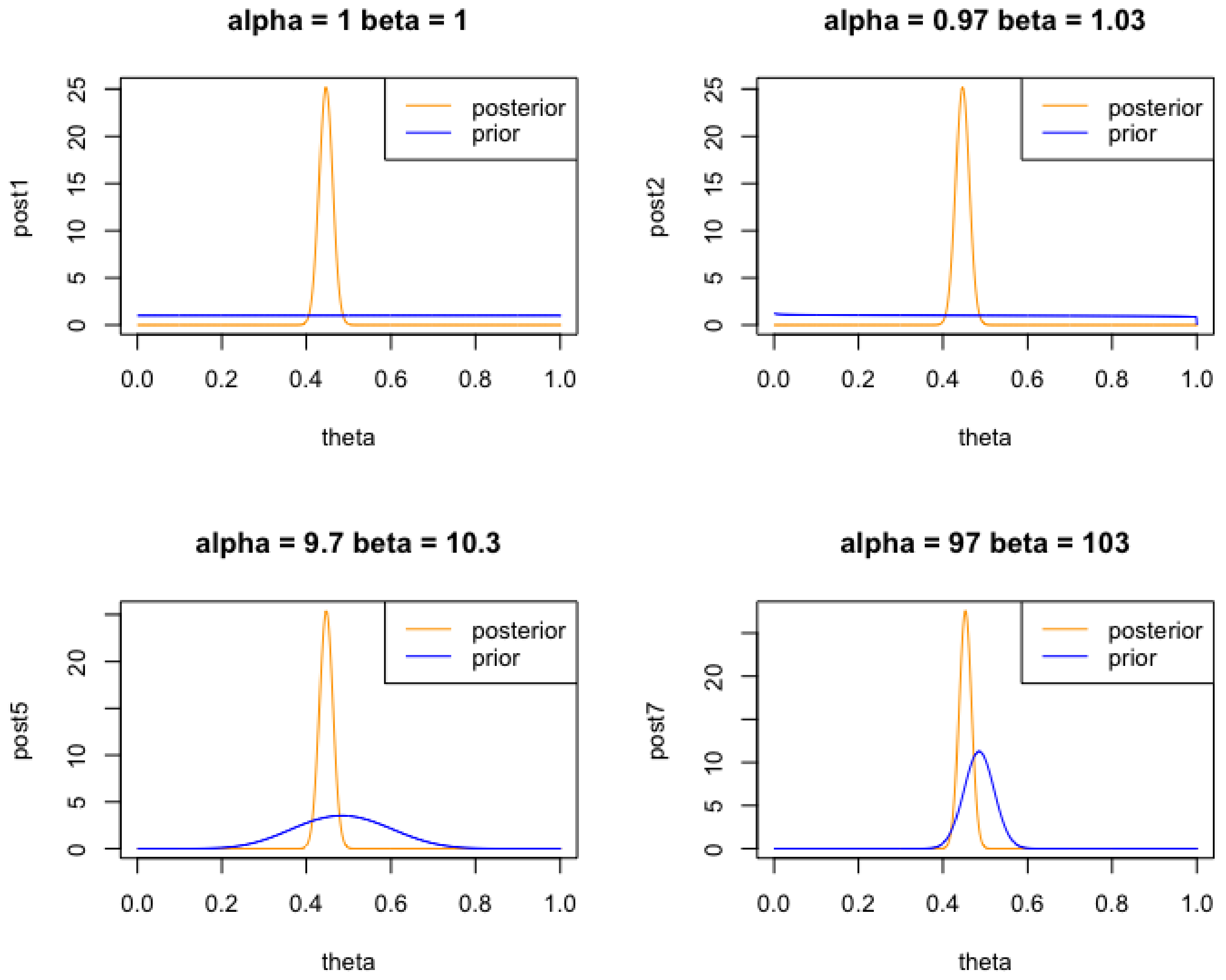
The 95% central credible interval is [.415, .477]

```
> qbeta(c(.025, .975), 438, 544)
[1] 0.4150655 0.4771998
```

Analysis using different conjugate prior distributions

Since the Beta(α, β) distribution is conjuguate for the binomial sampling model, we will see the effect of varying the hyperparameters α and β on the posterior distribution $p(\theta|y)$.

Parameters of Beta(α, β) prior		Summaries of posterior distribution	
α	β	Posterior median	95% credible interval for θ
1	1	0.446	[0.415, 0.477]
0.97	1.03	0.446	[0.415, 0.477]
2.425	2.575	0.446	[0.415, 0.477]
4.85	5.15	0.446	[0.415, 0.477]
9.7	10.3	0.447	[0.416, 0.478]
48.5	51.5	0.450	[0.420, 0.479]
97	103	0.453	[0.424, 0.481]



alpha = 1 beta = 1

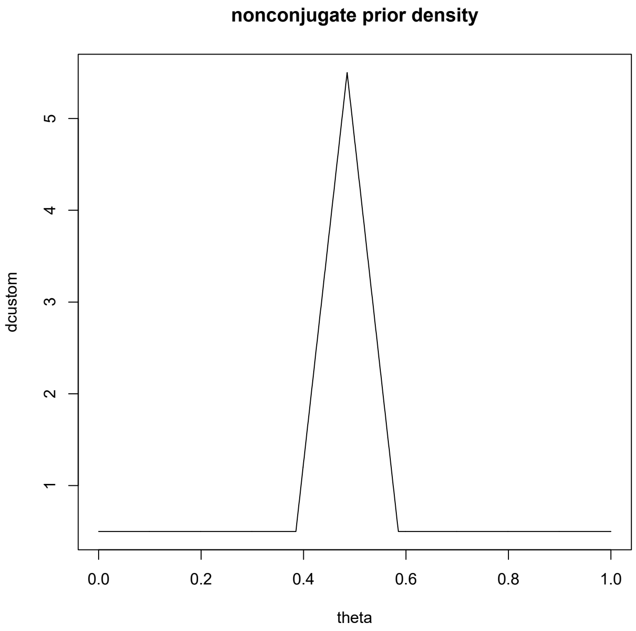
alpha = 0.97 beta = 1.03

alpha = 9.7 beta = 10.3

alpha = 97 beta = 103

Analysis using a nonconjugate prior distribution

Suppose we had a prior distribution such that 40% of the probability mass is outside the interval [0.385, 0.585], with a symmetric peak between.

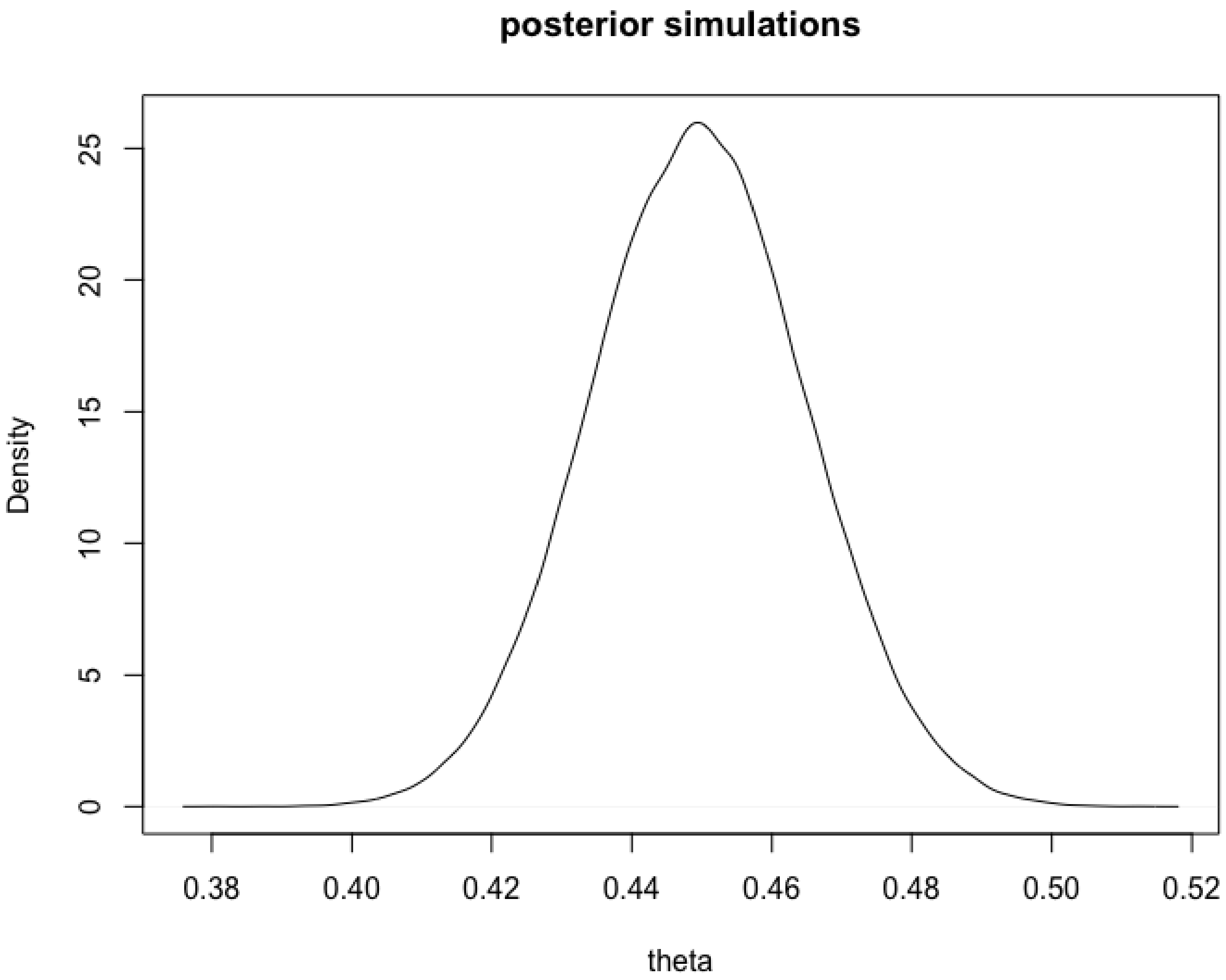


To obtain a sample from the approximate posterior distribution:

1. Create a sequence of gridded values for θ, e.g., (0.000, 0.001, …, 1.000)
2. Evaluate the likelihood function and prior distribution on the grid and multiply them together to get an approximate unnormalized posterior density.
3. Sample from the grid with weights equal to the unnormalized posterior density.

or

4. Create an approximate inverse cdf function
5. Use the inverse cdf method to generate a sample from the approximate posterior distribution.


posterior simulations
Density
0
5
10
15
20
25
theta
0.38
0.40
0.42
0.44
0.46
0.48
0.50
0.52

Simulated mean = 0.449
Simulated median = .449
95% central credible interval = [0.419, 0.481]

Normal-normal conjugate pair for the mean

A normal prior for the mean parameter θ is conjugate to the normal data distribution if the variance σ^2 is known.

Data distribution: $y_i|\theta,\sigma^2 \sim N(\theta,\sigma^2)$, so

$$p(y_i|\theta,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma^2}(y_i-\theta)^2\right),$$

with $\theta \in (-\infty,\infty)$ and $\sigma^2 > 0$.

The normal distribution is often a good approximation of the data distribution and the posterior distribution.

Prior distribution: $\theta \sim N(\mu_0, \tau_0^2)$, so that

$$p(\theta) \propto \exp\left(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2\right).$$

The **precision** is the inverse of the variance and plays an important role in manipulating normal distributions.

Posterior distribution:

$$\begin{aligned} p(\theta|y) &\propto p(\theta)p(y|\theta) \\ &\propto \exp\left(-\frac{1}{2\tau_0^2}(\theta-\mu_0)^2\right)\prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma^2}(y_i-\theta)^2\right) \\ &\propto \exp\left(-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right) \end{aligned}$$

with

$$\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\ \bar{y}}{\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}} \text{ and } \frac{1}{\tau_n^2} = \frac{1}{\tau_0^2}+\frac{n}{\sigma^2}.$$

The posterior mean μ_n is a weighted average of the prior mean and the observed sample mean, with weights proportional to the precisions.

- The larger the variance, the smaller the precision, and the less weight that component ($\bar{y}$ or μ_0) has in the posterior mean.
- This is intuitive because if a component has greater variability associated with it, it should play a less prominent role in affecting our posterior belief.
- As the sample size increases (with everything else held constant), our posterior belief becomes largely determined by σ^2 and $\bar{y}$.

Posterior predictive distribution

Notice that

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y},\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta,y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &\propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right)\exp\left(-\frac{1}{2\tau_n^2}(\theta-\mu_n)^2\right)d\theta \end{aligned}$$

The product in the integrand is a quadratic function of $(\tilde{y}, \theta)$, so the joint distribution is a normal distribution. Hence, the marginal distribution of $\tilde{y}$ will be a normal distribution.

Using properties of the mean and variance:

$$E(\tilde{y}|y) = E(E(\tilde{y}|\theta, y)|y) = E(\theta|y) = \mu_n.$$

$$\begin{aligned}\text{var}(\tilde{y}|y) &= E(\text{var}(\tilde{y}|\theta, y)|y) + \text{var}(E(\tilde{y}|\theta, y)|y) \\ &= E(\sigma^2|y) + \text{var}(\theta|y) = \sigma^2 + \tau_n^2.\end{aligned}$$

The posterior predictive mean is simply the mean of the posterior distribution and the posterior predictive variance is the sum of the posterior variance and the variance of the data distribution.

Poisson-gamma conjugate pair

The Poisson model arises naturally in the study of data taking counts, e.g., epidemiology, where the incidence of diseases is studied.

Data distribution: $y_1, y_2, \dots, y_n|\theta \overset{i.i.d.}{\sim} \text{Poisson}(\theta)$, with

$$p(y_i|\theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!} I_{\{0,1,2,\dots\}}(y_i).$$

Prior: $\theta \sim \text{Gamma}(\alpha, \beta)$, with

$$p(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} I_{(0,\infty)}(\theta),$$

for shape parameter $\alpha > 0$, and scale parameter $\beta > 0$.

Posterior distribution, $p(\theta|y)$:

Extended Poisson-Gamma conjugate pair

An **extended Poisson model** is $y_i|\theta, x_i \sim \text{Poisson}(x_i\theta)$, where x_i is the value of an explanatory variable for observation i.

- In epidemiology, x is known as the exposure and θ the rate.

Data distribution: $y_i|\theta, x_i \overset{\text{indep.}}{\sim} \text{Poisson}(x_i\theta)$, with joint density

$$p(y|\theta) = \theta^{\sum_{i=1}^n y_i} \exp\left(-\theta \sum_{i=1}^n x_i\right) \prod \frac{1}{y_i!}.$$

Prior: $\theta \sim \text{Gamma}(\alpha, \beta)$
Posterior distribution:

$$\theta|y \sim \text{Gamma}\left(\alpha + \sum_{i=1}^n y_i, \beta + \sum_{i=1}^n x_i\right).$$

Example: Asthma (single observation)

Out of a population of 200,000 people, 3 persons died of asthma last year. Thus, a crude estimate of asthma mortality rate in the city is 1.5 cases per 100,000 persons per year.

The data distribution of y, the number of deaths in a city of 200,000 in one year, may be expressed as Poisson$(2.0 \times \theta)$, where θ represents the true underlying mortality rate in the city (in cases per 100,000 persons per year).

Our observation $y = 3$ with an exposure rate $x = 2.0$.

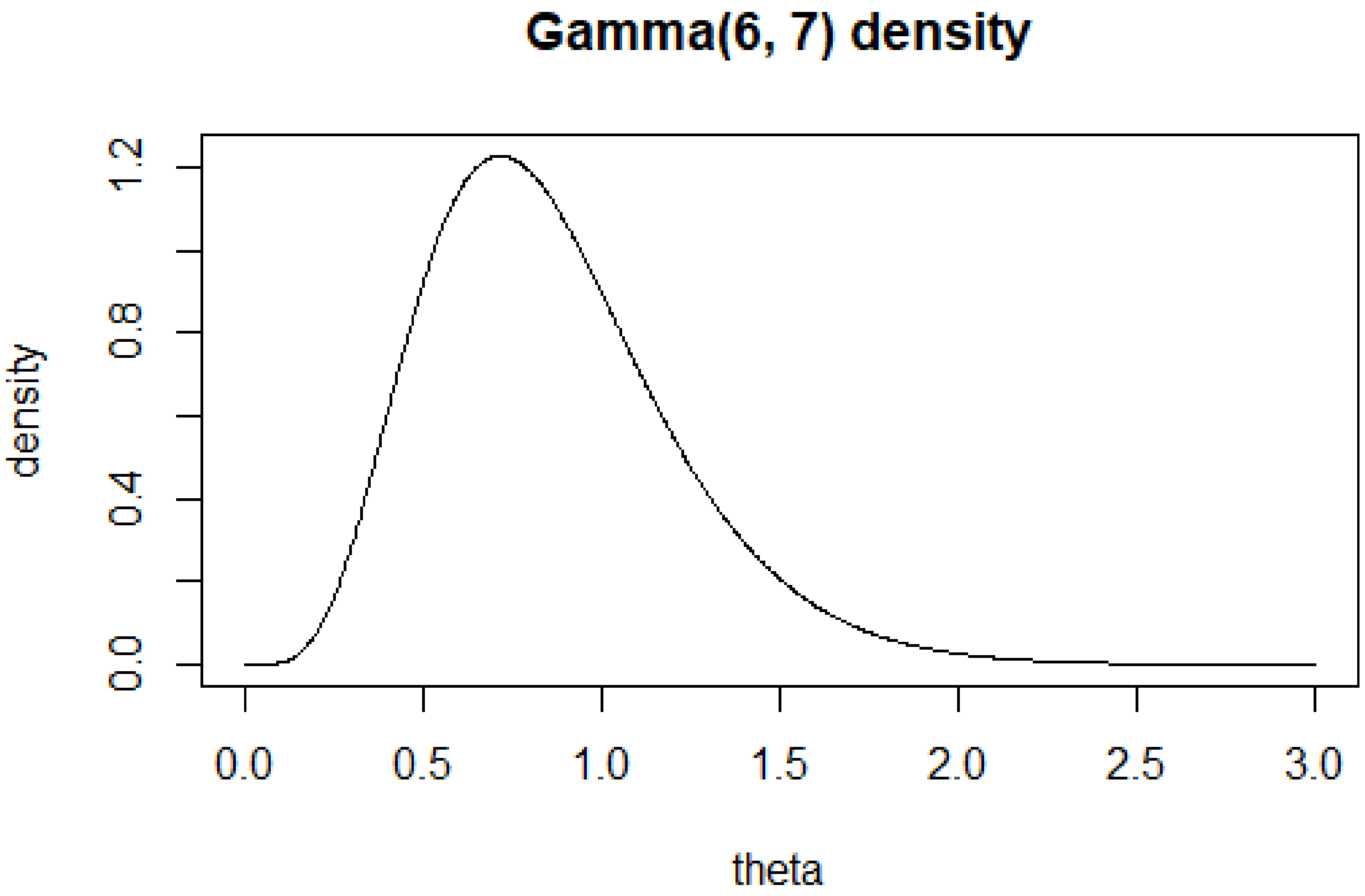
In the west, asthma mortality rates are around 0.6, and nearly always less than 1.5.

Letting $\theta \sim \text{Gamma}(3, 5)$ for the prior density matches these facts well since the $E(\theta) = \frac{3}{5} = 0.6$ and $\Pr(\theta \leq 1.5) = 0.98$.

The posterior distribution is

$$\theta|y \sim \text{Gamma}(3 + 3, 5 + 2) = \text{Gamma}(6, 7).$$

The mean of the posterior is $\frac{6}{7} = 0.86$, a substantial shrinkage toward the prior (since there aren't a lot of data).

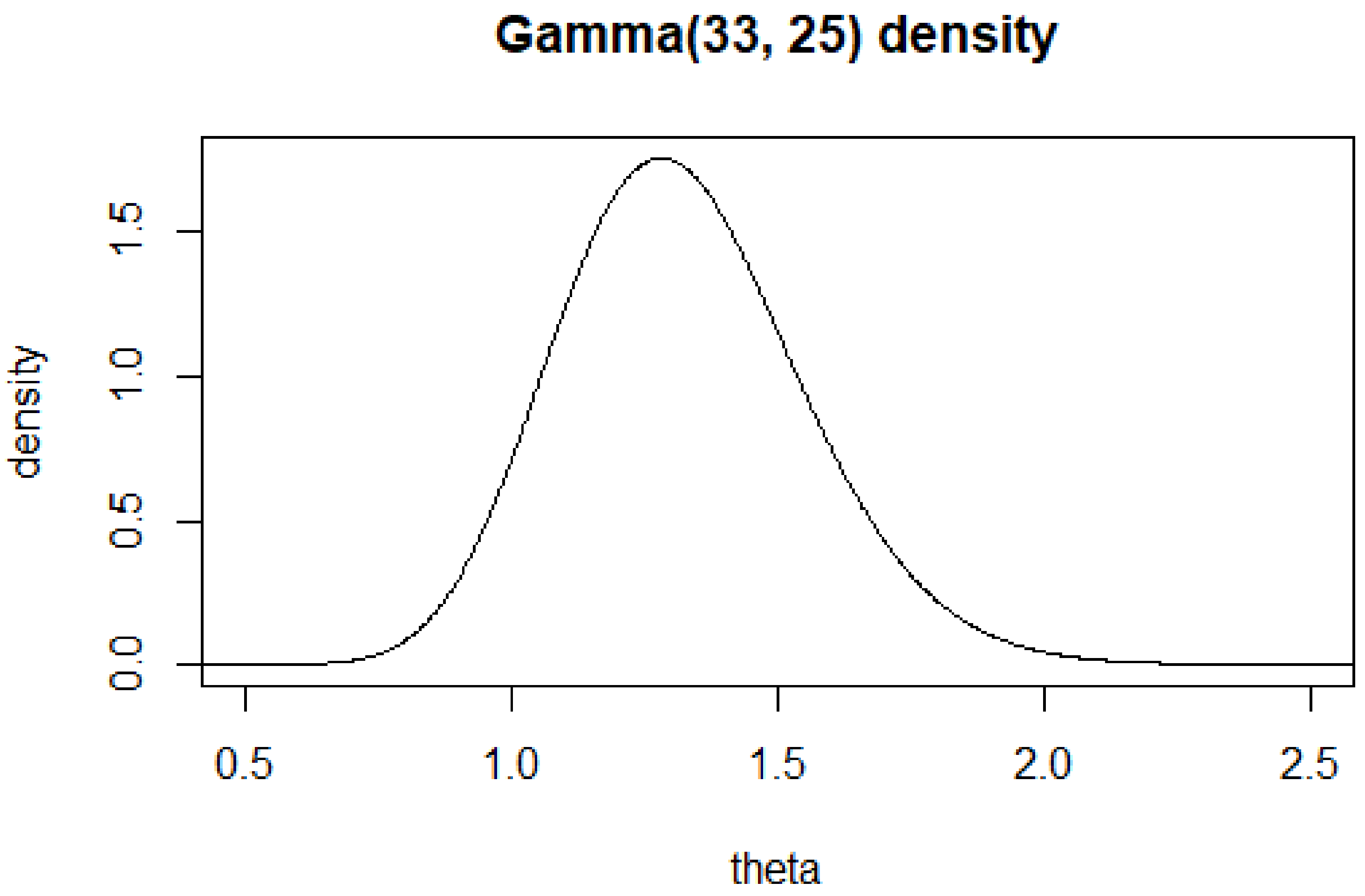

Gamma(6, 7) density
density
0.0
0.4
0.8
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
theta

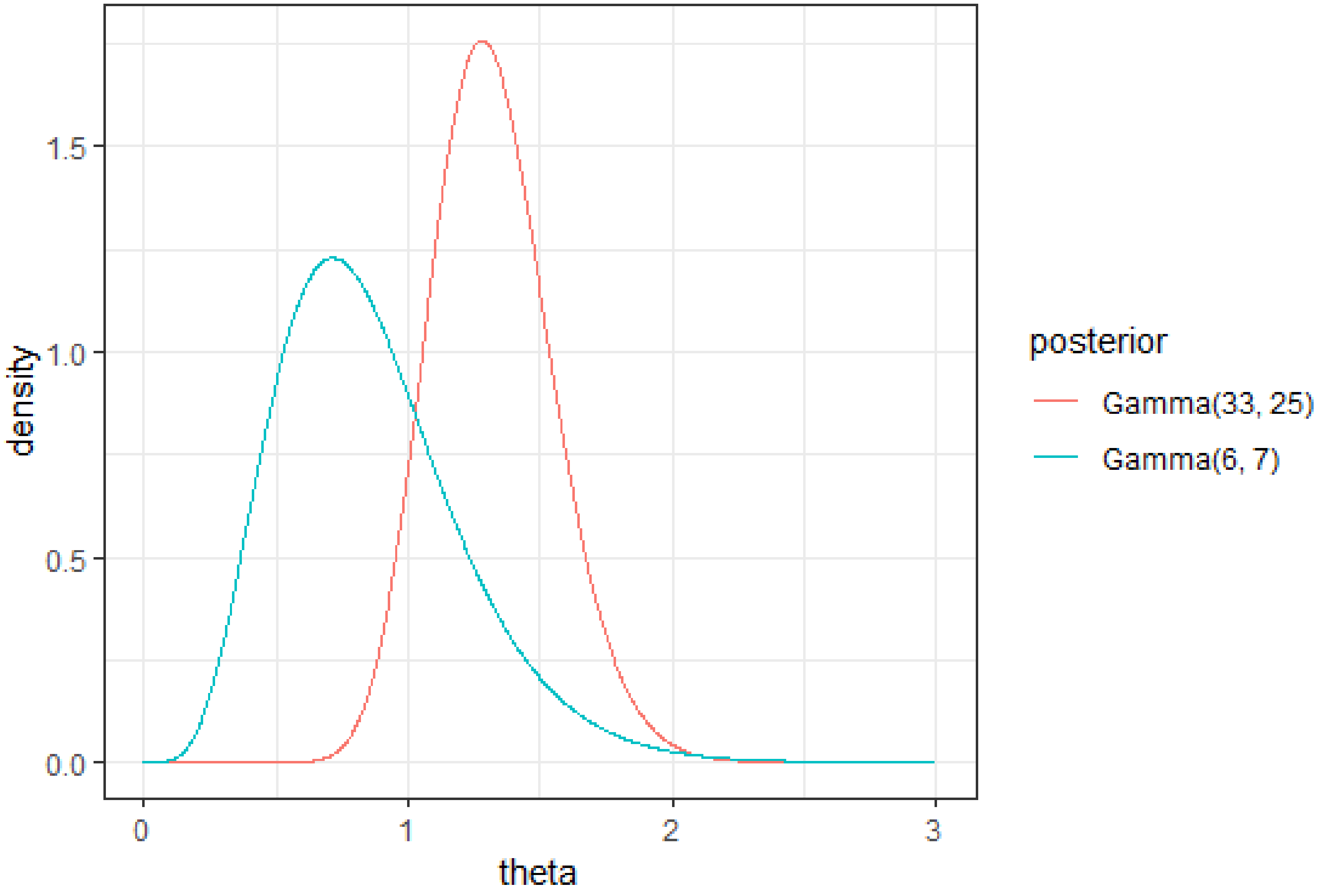
Example: Asthma with additional data

Suppose that we have ten years of data and that we have 30 deaths over the 10 years. Assuming a constant population, then the posterior distribution for θ is

$$p(\theta|y) \sim \text{Gamma}(3 + 30, 5 + 10 \times 2.0) = \text{Gamma}(33, 25).$$

The posterior mean is now 33/25 = 1.32 and the posterior probability that θ exceeds 1.0 is 0.93.


Gamma(33, 25) density
density
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
2.5
theta


1.5
1.0
0.5
0.0
density
0
1
2
3
theta
posterior
Gamma(33, 25)
Gamma(6, 7)

Exponential-gamma conjugate pair

The exponential model is commonly used to model waiting times and other continuous, positive, real-valued random variables (usually measured on a time scale).

Data distribution: $y_1, y_2, \ldots, y_n|\theta \overset{i.i.d.}{\sim}$ Expon(θ) with rate θ and mean $1/\theta$, with

$$p(y_i|\theta) = \theta \exp(-y_i\theta)\, I_{(0,\infty)}(y_i).$$

Note: The Expon(θ) is a special case of the Gamma(α, β) with $(\alpha, \beta) = (1, \theta)$.

Prior distribution: $\theta \sim \text{Gamma}(\alpha, \beta)$, with

$$p(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta} I_{(0,\infty)}(\theta),$$

for shape parameter $\alpha > 0$, and scale parameter $\beta > 0$.

Posterior distribution, $p(\theta|y)$:

Note: $p(y|\theta)$ is proportional to a Gamma$(n+1, n\bar{y})$ density. A Gamma(α, β) prior for θ can be thought of as $\alpha - 1$ additional exponential observations with total waiting time β.

Improper prior distributions

An improper prior distribution is not really a distribution; it's simply a function of θ that may not result in a valid posterior distribution, but they may work.

Example: Improper prior that works

Data distribution: $y_i|\theta,\sigma^2 \overset{i.i.d.}{\sim} N(\theta,\sigma^2)$

Prior distribution: $p(\theta) \propto c$ for all $\theta \in (-\infty,\infty)$

Posterior distribution: $\theta|y \sim N(\bar{y},\sigma^2/n)$

Objective prior distributions

When prior distributions have no population basis or are not based on genuine prior information, there is typically a desire to have a prior distribution that plays only a minimal role in the posterior distribution.

These distributions are known as vague, flat, diffuse, **noninformative**, or **objective**.

The idea behind these types of priors is to let the data "speak for themselves".

Jeffreys' prior

Harold Jeffreys proposed a rule often used to define a noninformative prior based on the prior distribution being invariant to transformation.

Let $\phi = h(\theta)$, where $h(\cdot)$ is a one-to-one function.

The density of ϕ may be obtained through the formula

$$p_\phi(\phi) = p_\theta(\theta)\left|\frac{d\theta}{d\phi}\right| = p_\theta(h^{-1}(\phi))\left|\frac{dh^{-1}(\phi)}{d\phi}\right|.$$

Jeffreys' prior is $p(\theta) = [J(\theta)]^{1/2}$, where $J(\theta)$ is the **Fisher information** for θ,

$$J(\theta) = E\left[\left(\frac{d \ln p(y|\theta)}{d\theta}\right)^2 |\theta\right] = -E\left[\left(\frac{d^2 \ln p(y|\theta)}{d\theta^2}\right) |\theta\right].$$

Jeffreys chose this rule so that

$$\sqrt{J(\phi)} = \sqrt{J(\theta)}\left|\frac{d\theta}{d\phi}\right|,$$

i.e., the prior for ϕ, $\sqrt{J(\phi)}$, can be obtained by directly computing the Fisher's information for ϕ or finding $\sqrt{J(\theta)}$ and transforming the associated distribution using the change-of-variable formula.

Jeffreys' principle can be extended to multiparameter models.

- Results vary based on whether you multiply the Jeffreys' priors for the parameters independently or use the square root of the determinant of the Fisher's information matrix.
    - E.g., $\sqrt{J(\theta_1)J(\theta_2)} \neq \sqrt{\det\ J(\theta)}$ for $\theta = (\theta_1, \theta_2)$.
- In high-dimensional parameter spaces, hierarchical models are often favored over noninformative priors.

Reference priors

Noninformative priors are related to the notion of a **reference** prior.

Reference priors can be:

- A conventional or “default” prior.
- A prior that expresses ignorance of θ in a formal sense.
- A prior that has as little impact as possible on the posterior inference[1], or alternatively, it maximizes the expected difference between the prior and posterior distributions.

[1] See Berger, J.O., Bernardo, J.M., and Sun. D. (2009). The Formal Definition of Reference Priors. *The Annals of Statistics*. 37(2).

Maximum entropy prior

The maximum entropy prior seeks to find the prior that maximizes the entropy of the prior distribution while satisfying a few key characteristics.

The entropy is defined as $-E[\log p(\theta)] = -\int \log\ p(\theta)p(\theta)d\theta$.

e.g., you might try to find the prior with $E(\theta) = 0.5$ that maximizes the entropy.

This is tough because you must consider not just different parameter values, but different distributional families.

Empirical Bayes

The Empirical Bayes approach to choosing a prior distribution uses the data to select the parameters of the prior distribution.

e.g., if γ denotes the hyperparameters of the prior distribution, then an Empirical Bayes approach might be to choose γ such that

$$\hat{\gamma} = \mathrm{argmax}_{\gamma} \int p(y|\theta)p(\theta|\gamma)\, d\theta.$$

- This is the marginal maximum likelihood estimator for the hyperparameter.

Pros of Empirical Bayes:

- It's an objective method for choosing the prior.
- Typically provides a prior that will behave well computationally.

Cons of Empirical Bayes:

- Uses the data twice.
- Overly optimistic results since the uncertainty in γ is ignored.

Cons of noninformative priors

- Why use a vague prior if you have useful information?
- The noninformative prior may not be scale invariant.

Pros of noninformative priors:

- Provides an approach to choosing a prior when you have a lot of information!

Jeffreys' prior examples

A probability distribution $p(y|\theta)$ belongs to an **exponential family** if it has the form

$$p(y|\theta) = f(y)g(\theta)\exp\big(\phi(\theta)^T t(y)\big),$$

where $\phi(\theta)$ and $t(y)$ are vectors with the same length as θ.

Assuming θ is a vector of length p:

- $t(y) = \big(t_1(y), \dots, t_p(y)\big)$ is the vector of sufficient statistics.
- $\phi(\theta) = \big(\phi_1(\theta), \dots, \phi_p(\theta)\big)$ is the vector of natural parameters.

Example: Natural parameters

Consider a set of i.i.d $N(\mu, \sigma^2)$ random variables. Determine the natural parameters $\phi(\theta)$ and the sufficient statistics $t(y)$ for this distribution.

If $y_1, \ldots, y_n$ are i.i.d., then the Fisher's information for the n observations is $nJ(\theta)$, were $J(\theta)$ is the Fisher's information for a single observation.

Example: Jeffreys prior for Binomial data distribution

If $y|\theta \sim \text{Binomial}(n,\theta)$, the log-likelihood is

$$\begin{aligned}\ln\ p(y|\theta) &= \ln\left(\binom{n}{y}\theta^y(1-\theta)^{n-y}\right)\\ &= \text{constant} + y\ln(\theta) + (n-y)\ln(1-\theta)\,.\end{aligned}$$

From this,

$$\frac{d^2\ln p(y|\theta)}{d\theta^2} = -\frac{n\theta^2 + y - 2\theta y}{(1-\theta)^2\theta^2},$$

Since $E(y|\theta) = n\theta$ and $E(cX) = cE(X)$ for a random variable X and constant c, we have

$$J(\theta) = \frac{n\theta^2 + n\theta - 2n\theta^2}{(1-\theta)^2\theta^2} = \frac{n\theta(1-\theta)}{(1-\theta)^2\theta^2} = \frac{n}{\theta(1-\theta)}$$

and

$$p(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2} = \text{Beta}(1/2\,, 1/2).$$

Example: Jeffreys prior for a normal distribution

Determine Jeffreys' prior distribution for μ when the data are i.i.d. $N(\mu, \sigma^2)$, with σ^2 fixed.