Part 2:
One-parameter models

Bernoulli/binomial models

Return to $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Bin}(1, \theta)$. The sampling model/likelihood is

$$p(y_1, \ldots, y_n | \theta) = \theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}$$

When combined with a prior $p(\theta)$, Bayes' rule gives the posterior

$$p(\theta | y_1, \ldots, y_n) = \frac{\theta^{\sum y_i} (1 - \theta)^{n - \sum y_i} \times p(\theta)}{p(y_1, \ldots, y_n)}$$

- $\sum y_i$ is a sufficient statistic

The Binomial model

When $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Bin}(1, \theta)$, the sufficient statistic $Y = \sum_{i=1}^{n} Y_i$ has a $\text{Bin}(n, \theta)$ distribution

Having observed $\{Y = y\}$ our task is to obtain the posterior distribution of θ:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} = \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}p(\theta)}{p(y)}$$

$$= c(y)\theta^y(1-\theta)^{n-y}p(\theta)$$

where $c(y)$ is a function of y and not θ

A uniform prior

The parameter θ is some unknown number between 0 and 1

Suppose our prior information for θ is such that all subintervals of $[0,1]$ having the same length also have the same probability

$$P(a \leq \theta \leq b) = P(a + c \leq \theta \leq b + c),$$
$$\text{for } \ 0 \leq a < b < b + c \leq 1$$

This condition implies a uniform density for θ:

$$p(\theta) = 1 \qquad \text{for all } \ \theta \in [0,1]$$

Normalizing constant

Using the following result from calculus

$$\int_0^1 \theta^{a-1}(1-\theta)^{b-1}\, d\theta = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

where $\Gamma(n) = (n-1)!$ is evaluated in R as `gamma(n)`

we can find out what $c(y)$ is under the uniform prior

$$\int_0^1 p(\theta|y)\, d\theta = 1$$

$$\Rightarrow \;\; c(y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}$$

Beta posterior

So the posterior distribution is

$$\begin{aligned} p(\theta|y) &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^y(1-\theta)^{n-y} \\ &= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)}\theta^{(y+1)-1}(1-\theta)^{(n-y+1)-1} \\ &= \mathrm{Beta}(y+1, n-y+1) \end{aligned}$$

To wrap up:

- a uniform prior, plus
- a Bernoulli/Binomial likelihood (sampling model)
- gives a Beta posterior

Example: Happiness

Each female of age 65 or over in the 1998 General Social Survey was asked whether or not they were generally happy

Let $Y_i = 1$ if respondent i reported being generally happy, and $Y_i = 0$ otherwise, for $i = 1, \ldots, n = 129$ individuals

Since $129 \ll N$, the total size of the female senior citizen population, our joint beliefs about $Y_1, \ldots, Y_{129}$ are well approximated by (the sampling model)

- our beliefs about $\theta = \sum_{i=1}^{N} Y_i / N$
- the model that, conditional on θ, the Y_i's are IID Bernoulli RVs with expectation θ

Example: Happiness IID Bernoulli likelihood

This sampling model says that the probability for any potential outcome $\{y_1, \ldots, y_n\}$, conditional on θ, is given by

$$p(y_1, \ldots, y_{129}|\theta) = \theta^{\sum_{i=1}^{129} y_i}(1-\theta)^{129-\sum_{i=1}^{129} y_i}$$

The survey revealed that

- 118 individuals report being generally happy (91%)
- 11 individuals do not (9%)

So the probability of these data for a given value of θ is

$$p(y_1, \ldots, y_{129}|\theta) = \theta^{118}(1-\theta)^{11}$$

Example: Happiness Binomial likelihood

In the binomial formulation

$$Y = \sum_{i=1}^{n} Y_i \sim \mathrm{Bin}(n, \theta)$$

and our observed data is $y = 118$

So (now) the probability of these data for a given value of θ is

$$p(y|\theta) = \binom{129}{118} \theta^{118} (1 - \theta)^{11}$$

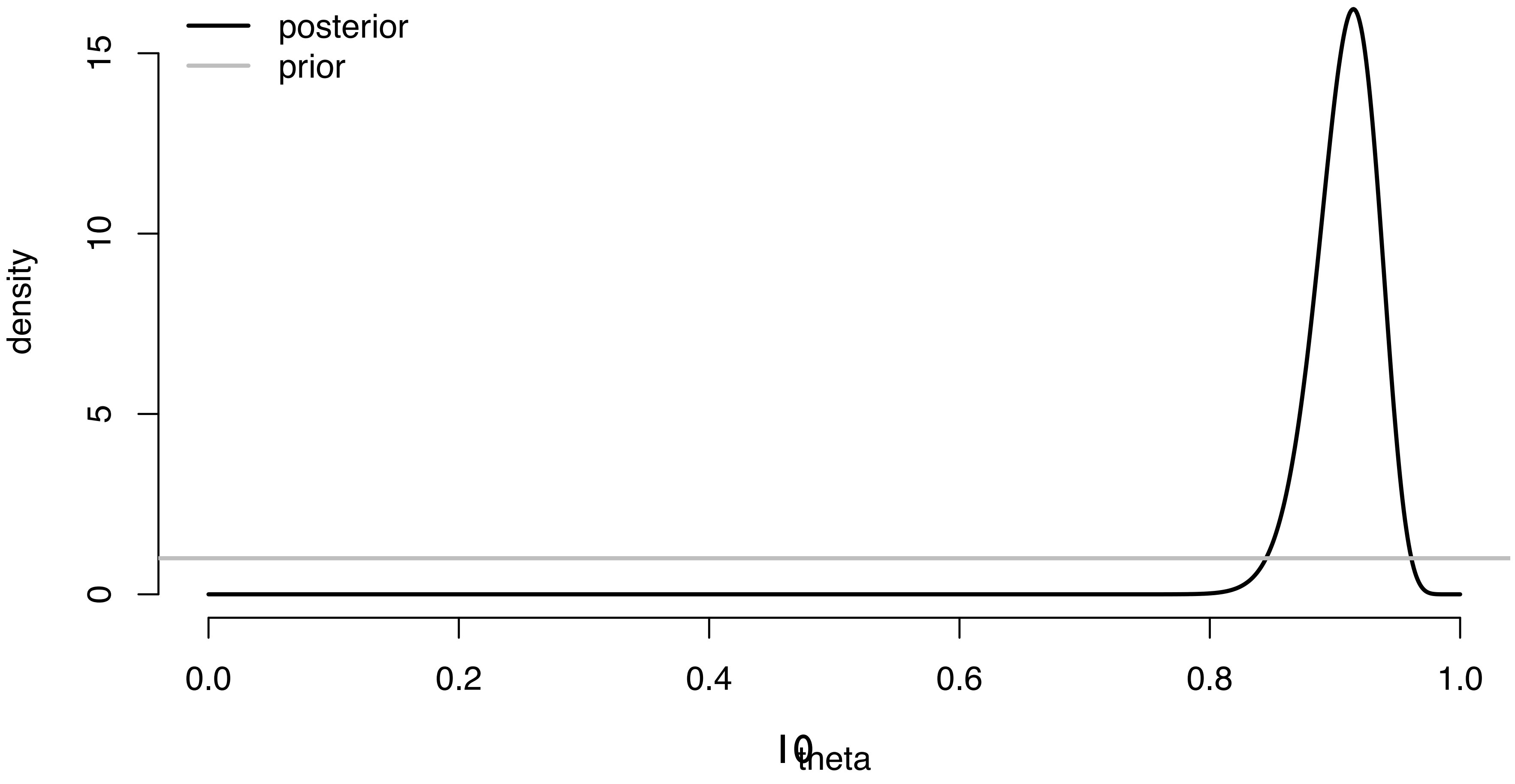
Example: Happiness Posterior

If we take a uniform prior for θ, expressing our ignorance, then we know the posterior is

$$p(\theta|y) = \text{Beta}(y+1, n-y+1)$$

For $n = 129$ and $y = 118$, this gives

$$\theta|\{Y = 118\} \sim \text{Beta}(119, 12)$$



Uniform is Beta

The uniform prior has $p(\theta) = 1$ for all $\theta \in [0, 1]$

This can be thought of as a Beta distribution with parameters $a = 1, b = 1$

$$p(\theta) = \frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\theta^{1-1}(1-\theta)^{1-1}$$

$$= \frac{1}{1 \times 1} 1 \times 1$$ (under the convention that $\Gamma(1) = 1$)

We saw that if $\left\{ \begin{array}{c} \theta \sim \text{Beta}(1,1) \\ Y|\theta \sim \text{Bin}(n,\theta) \end{array} \right\}$

General Beta prior

Suppose $\theta \sim \mathrm{Beta}(a,b)$ and $Y|\theta \sim \mathrm{Bin}(n,\theta)$

$$\begin{aligned} p(\theta|y) &\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1} \\ \Rightarrow \qquad \theta|y &\sim \mathrm{Beta}(a+y, b+n-y) \end{aligned}$$

In other words, the constant of proportionality must be:

$$c(n,y,a,b) = \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)}$$

How do I know?

- $p(\theta|y)$ (and the beta density) must integrate to 1

Posterior proportionality

We will use this trick over and over again!

I.e., if we recognize that the posterior distribution is proportional to a known probability density, then it must be identical to that density

But Careful! The constant of proportionality must be constant with respect to θ

Conjugacy

We have shown that a beta prior distribution and a binomial sampling model lead to a beta posterior distribution

To reflect this, we say that the class of beta priors is conjugate for the binomial sampling model

Formally, we say that a class $\mathcal{P}$ of prior distributions for θ is conjugate for a sampling model $p(y|\theta)$ if

$$p(\theta) \in \mathcal{P} \Rightarrow p(\theta|y) \in \mathcal{P}$$

Conjugate priors make posterior calculations easy, but might not actually represent our prior information

Posterior summaries

If we are interested in point-summaries of our posterior inference, then the (full) distribution gives us many points to choose from

E.g., if $\theta|\{Y = y\} \sim \mathrm{Beta}(a + y, b + n - y)$ then

$$\mathbb{E}\{\theta|y\} = \frac{a+y}{a+b+n}$$

$$\mathrm{mode}(\theta|y) = \frac{a+y-1}{a+b+n-2}$$

$$\mathrm{Var}[\theta|y] = \frac{\mathbb{E}\{\theta|y\}\mathbb{E}\{1-\theta|y\}}{a+b+n+1}$$

not including quantiles, etc.

Combining information

The posterior expectation $\mathbb{E}\{\theta|y\}$ is easily recognized as a combination of prior and data information:

$$\mathbb{E}\{\theta|y\} = \frac{a+b}{a+b+n} \times \left(\begin{array}{c}\text{prior}\\\text{expectation}\end{array}\right) + \frac{n}{a+b+n} \times \left(\begin{array}{c}\text{data}\\\text{average}\end{array}\right)$$

I.e., for this model and prior distribution, the posterior mean is a weighted average of the prior mean and the sample average with weights proportional to $a+b$ and n respectively

Prior sensitivity

This leads to the interpretation of a and b as "prior data":

$$a \approx \text{"prior number of 1's"}$$

$$b \approx \text{"prior number of 0's"}$$

$$a + b \approx \text{"prior sample size"}$$

If $n \gg a + b$, then it seems reasonable that most of our information should be coming from the data as opposed to the prior

Indeed: $\frac{a+b}{a+b+n} \approx 0$

$$\mathbb{E}\{\theta|y\} \approx \frac{y}{n} \quad \& \quad \text{Var}[\theta|y] \approx \frac{1}{n}\frac{y}{n}\left(1 - \frac{y}{n}\right)$$

Prediction

An important feature of Bayesian inference is the existence of a predictive distribution for new observations

Revert, for the moment, to our notation for Bernoulli data. Let $y_1, \ldots, y_n$, be the outcomes of a sample of n binary RVs

Let $\tilde{Y} \in \{0, 1\}$ be an additional outcome from the same population that has yet to be observed

The predictive distribution of $\tilde{Y}$ is the conditional distribution of $\tilde{Y}$ given $\{Y_1 = y_1, \ldots, Y_n = y_n\}$

Predictive distribution

For conditionally IID binary variables the predictive distribution can be derived from the distribution of $\tilde{Y}$ given θ and the posterior distribution of θ

$$p(\tilde{Y} = 1 | y_1, \ldots, y_n) = \mathbb{E}\{\theta | y_1, \ldots, y_n\}$$

$$= \frac{a + \sum_{i=1}^{N} y_i}{a + b + n}$$

$$p(\tilde{Y} = 0 | y_1, \ldots, y_n) = 1 - p(\tilde{Y} = 1 | y_1, \ldots, y_n)$$

$$\frac{b + n - \sum_{i=1}^{N} y_i}{a + b + n}$$

Two points about prediction

1. The predictive distribution does not depend upon any unknown quantities
    - if it did, we would not be able to use it to make predictions

2. The predictive distribution depends on our observed data
    - i.e. $\tilde{Y}$ is not independent of $Y_1, \ldots, Y_n$
    - this is because observing $Y_1, \ldots, Y_n$ gives information about θ, which in turn gives information about $\tilde{Y}$

Example: Posterior predictive

Consider a uniform prior, or $\mathrm{Beta}(1,1)$, using $Y = \sum_{i=1}^{n} y_i$

$$P(\tilde{Y} = 1 | Y = y) = \mathbb{E}\{\theta | Y = y\} = \frac{2}{2+n}\frac{1}{2} + \frac{n}{2+n}\frac{y}{n}$$

but $$\mathrm{mode}(\theta | Y = y) = \frac{y}{n}$$

Does this discrepancy between these two posterior summaries of our information make sense?

Consider the case in which $Y = 0$, for which

$$\mathrm{mode}(\theta | Y = 0) = 0$$

but $$P(\tilde{Y} = 1 | Y = 0) = 1/(2+n)$$

Confidence regions

An interval $[l(y), u(y)]$, based on the observed data $Y = y$, has 95% Bayesian coverage for θ if

$$P(l(y) < \theta < u(y)|Y = y) = 0.95$$

The interpretation of this interval is that it describes your information about the true value of θ after you have observed $Y = y$

Such intervals are typically called credible intervals, to distinguish them from frequentist confidence intervals which is an interval that describes a region (based upon $\hat{\theta}$) wherein the true θ_0 lies 95% of the time

Both, confusingly, use the acronym CI

Quantile-based (Bayesian) CI

Perhaps the easiest way to obtain a credible interval is to use the posterior quantiles

To make a $100 \times (1 - \alpha)$% quantile-based CI, find numbers $\theta_{\alpha/2} < \theta_{1-\alpha/2}$ such that

(1) $$P(\theta < \theta_{\alpha/2} | Y = y) = \alpha/2$$

(2) $$P(\theta > \theta_{1-\alpha/2} | Y = y) = \alpha/2$$

The numbers $\theta_{\alpha/2}, \theta_{1-\alpha/2}$ are the $\alpha/2$ and $1 - \alpha/2$ posterior quantiles of θ

Example: Binomial sampling and uniform prior

Suppose out of $n = 10$ conditionally independent draws of a binary random variable we observe $Y = 2$ ones

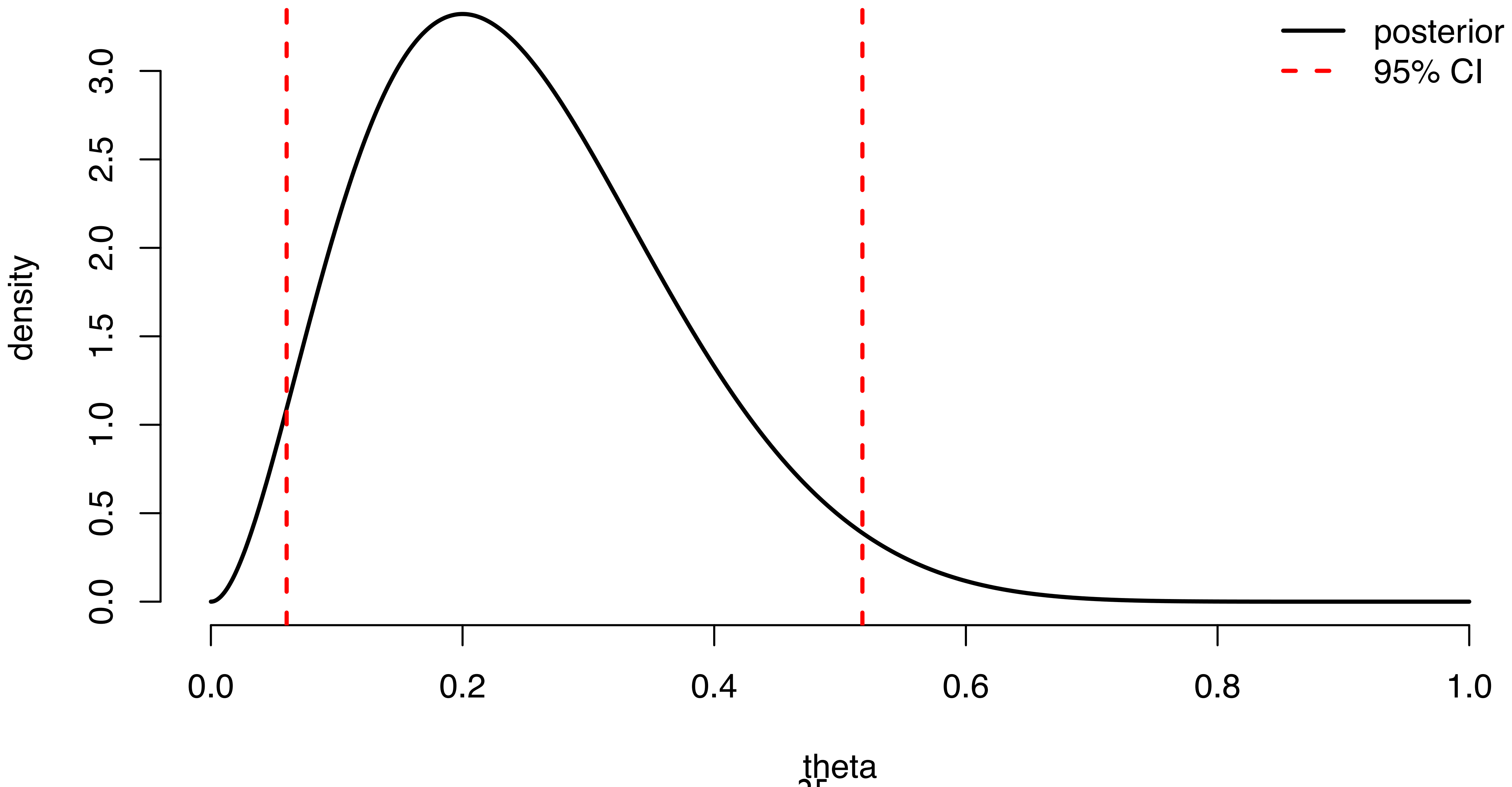
Using a uniform prior distribution for θ, the posterior distribution is $\theta|\{Y = 2\} \sim \mathrm{Beta}(1 + 2, 1 + 8)$

A 95% CI can be obtained from the 0.025 and 0.975 quantiles of this beta distribution

These quantiles are 0.06 and 0.52, respectively, and so the posterior probability that $\theta \in [0.06, 0.52]$ is 95%

Example: a quantile-based CI drawback

Notice that there are θ-values outside the quantile-based CI that have higher probability [density] than some points inside the interval



Example:
Estimating the probability of a female birth

The proportion of births that are female has long been a topic of interest both scientifically and to the lay public

E.g., Laplace, perhaps the first "Bayesian", analyzed Parisian birth data from 1745 with a uniform prior and a binomial sampling model

241,945 girls, and 251,527 boys

Example: female birth posterior

The posterior distribution for θ is $(n = 493472)$

$$\theta|\{Y = 249145\} \sim \mathrm{Beta}(241945 + 1, 251527 + 1)$$

Now, we can use the posterior CDF to calculate

$$P(\theta > 0.5|Y = 241945) =$$

```
pbeta(0.5,241945+1, 251527+1,
      lower.tail=FALSE)
```

$$= 1.15 \times 10^{-42}$$

The Poisson model

Some measurements, such as a person’s number of children or number of friends, have values that are whole numbers

In these cases the sample space is $\mathcal{Y} = \{0, 1, 2, \ldots\}$

Perhaps the simplest probability model on $\mathcal{Y}$ is the Poisson model

A RV Y has a Poisson distribution with mean θ if

$$P(Y = y|\theta) = \theta^y \frac{e^{-\theta}}{y!} \quad \text{for} \;\; y \in \{0, 1, 2, \ldots\}$$

IID Sampling model

If we take $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Pois}(\theta)$ then the joint PDF is:

$$\begin{aligned} P(Y_1 = y_1, \ldots, Y_n = y_n|\theta) &= \prod_{i=1}^{n} p(y_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{y!}\theta^{y_i} e^{-\theta} \\ &= c(y_1, \ldots, y_n)\theta^{\sum y_i} e^{-n\theta} \end{aligned}$$

Note that $\sum_{i=1}^{n} Y_i \sim \text{Pois}(n\theta)$

Conjugate prior

Recall that a class of priors is conjugate for a sampling model if the posterior is also in that class

For the Poisson model, our posterior distribution for θ has the following form

$$\begin{aligned} p(\theta|y_1,\ldots,y_m) &\propto p(y_1,\ldots,y_n|\theta)p(\theta) \\ &\propto \theta^{\sum y_i} e^{-n\theta} p(\theta) \end{aligned}$$

This means that whatever our conjugate class of densities is, it will have to include terms like $\theta^{c_1} e^{-c_2\theta}$ for numbers c_1 and c_2

Conjugate gamma prior

The simplest class of probability distributions matching this description is the gamma family

$$p(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} \qquad \text{for} \;\; \theta, a, b > 0$$

Some properties include

$$\mathbb{E}\{\theta\} = \frac{a}{b} \qquad \mathrm{Var}[\theta] = \frac{a}{b^2}$$

$$\mathrm{mode}(\theta) = \begin{cases} (a-1)/b & \text{if } a > 1 \\ 0 & \text{if } a \leq 1 \end{cases}$$

We shall denote this family as $\mathrm{G}(a, b)$

Posterior distribution

Suppose $Y_1, \dots Y_n | \theta \overset{\text{iid}}{\sim} \text{Pois}(\theta)$ and $\theta \sim \text{G}(a, b)$

Then

$$p(\theta | y_1, \dots, y_n) \propto \theta^{a + \sum y_i - 1} e^{-(b+n)\theta}$$

This is evidently a gamma distribution, and we have confirmed the conjugacy of the gamma family for the Poisson sampling model

$$\left\{ \begin{array}{r} \theta \sim G(a, b) \\ Y_1, \dots, Y_n \sim \text{Pois}(\theta) \end{array} \right\} \quad \Rightarrow$$

$$\{\theta | Y_1, \dots, Y_n\} \sim G\left(a + \sum_{i=1}^{n} Y_i, b + n\right)$$

Combining information

Estimation and prediction proceed in a manner similar to that in the binomial model

The posterior expectation of θ is a convex combination of the prior expectation and the sample average

$$\mathbb{E}\{\theta|y_1, \ldots, y_n\} = \frac{b}{b+n}\frac{a}{b} + \frac{n}{b+n}\frac{\sum y_i}{n}$$

- b is the number of prior observations
- a is the sum of the counts from b prior observations

For large n, the information in the data dominates

$$n \gg b \Rightarrow \mathbb{E}\{\theta|y_1, \ldots y_n\} \approx \bar{y},\ \mathrm{Var}[\theta|y_1, \ldots, y_n] \approx \frac{\bar{y}}{n}$$

Posterior predictive

Predictions about additional data can be obtained with the posterior predictive distribution

$$
\begin{aligned}
&p(\tilde{y}|y_1, \ldots y_n) \\
&\quad = \frac{(b+n)^{a+\sum y_i}}{\Gamma(\tilde{y}+1)\Gamma(a+\sum y_i)} \int_0^\infty \theta^{a+\sum y_i+\tilde{y}+1} e^{-(b+n+1)\theta}\, d\theta \\
&\quad = \frac{(b+n)^{a+\sum y_i}}{\Gamma(\tilde{y}+1)\Gamma(a+\sum y_i)} \left(\frac{b+n}{b+n+1}\right)^{a+\sum y_i} \left(\frac{1}{b+n+1}\right)
\end{aligned}
$$

for $\tilde{y} \in \{0, 1, 2, \ldots, \}$

Posterior predictive

This is a negative binomial distribution with parameters $(a + \sum y_i, b + n)$ for which

$$\begin{aligned}
\mathbb{E}\{\tilde{Y}|y_1, \ldots, y_n\} &= \frac{a + \sum y_i}{b + n} \\
&= \mathbb{E}\{\theta|y_1, \ldots, y_n\} \\
\mathrm{Var}[\tilde{Y}|y_1, \ldots, y_n] &= \frac{a + \sum y_i}{b + n} \frac{b + n + 1}{b + n} \\
&= \mathrm{Var}[\theta|y_1, \ldots, y_n] \times (b + n + 1) \\
&= \mathbb{E}\{\theta|y_1, \ldots, y_n\} \frac{b + n + 1}{b + n}
\end{aligned}$$

Example: birth rates and education

Over the course of the 1990s the General Societal Survey gathered data on the educational attainment and number of children of 155 women who were 40 years of age at the time of their participation in the survey

These women were in their 20s in the 1970s, a period of historically low fertility rates in the United States

In this example we will compare the women with college degrees to those without in terms of their numbers of children

Example: sampling model(s)

Let $Y_{1,1}, \ldots, Y_{n_1,1}$ denote the numbers of children for the n_1 women without college degrees and $Y_{1,2}, \ldots, Y_{n_2,2}$ be the data for the n_2 women with degrees

For this example, we will use the following sampling models

$$Y_{1,1}, \ldots, Y_{n_1,1} \overset{\text{iid}}{\sim} \text{Pois}(\theta_1)$$

$$Y_{1,2}, \ldots, Y_{n_2,2} \overset{\text{iid}}{\sim} \text{Pois}(\theta_2)$$

The (sufficient) data are

no bachelors: $n_1 = 111,\ \sum_{i=1}^{n_1} y_{i,1} = 217,\ \bar{y}_1 = 1.95$

bachelors: $n_2 = 44,\ \sum_{i=1}^{n_2} y_{i,2} = 66,\ \bar{y}_2 = 1.50$

Example: prior(s) and posterior(s)

In the case were

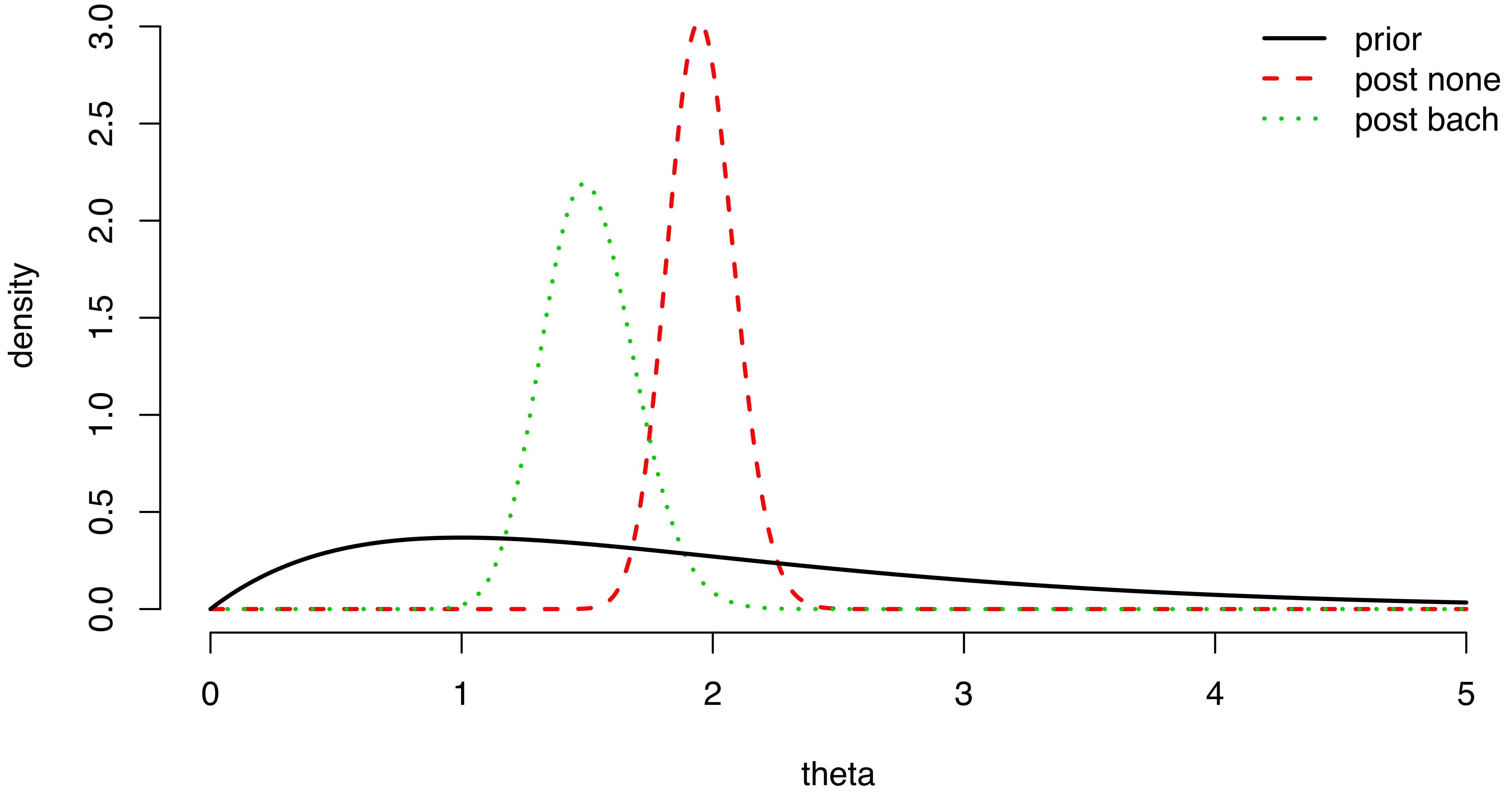
$$\{\theta_1, \theta_2\} \overset{\text{iid}}{\sim} G(a = 2, b = 1)$$

we have the following posterior distributions:

$$\theta_1 | \{n_1 = 111, \sum Y_{i,1} = 217\} \sim G(2 + 217, 1 + 111)$$
$$\theta_2 | \{n_1 = 44, \sum Y_{i,2} = 66\} \sim G(2 + 66, 1 + 44)$$

Posterior means, modes and 95% quantile-based CIs for θ_1 and θ_2 can be obtained from their gamma posterior distributions

Example: prior(s) and posterior(s)

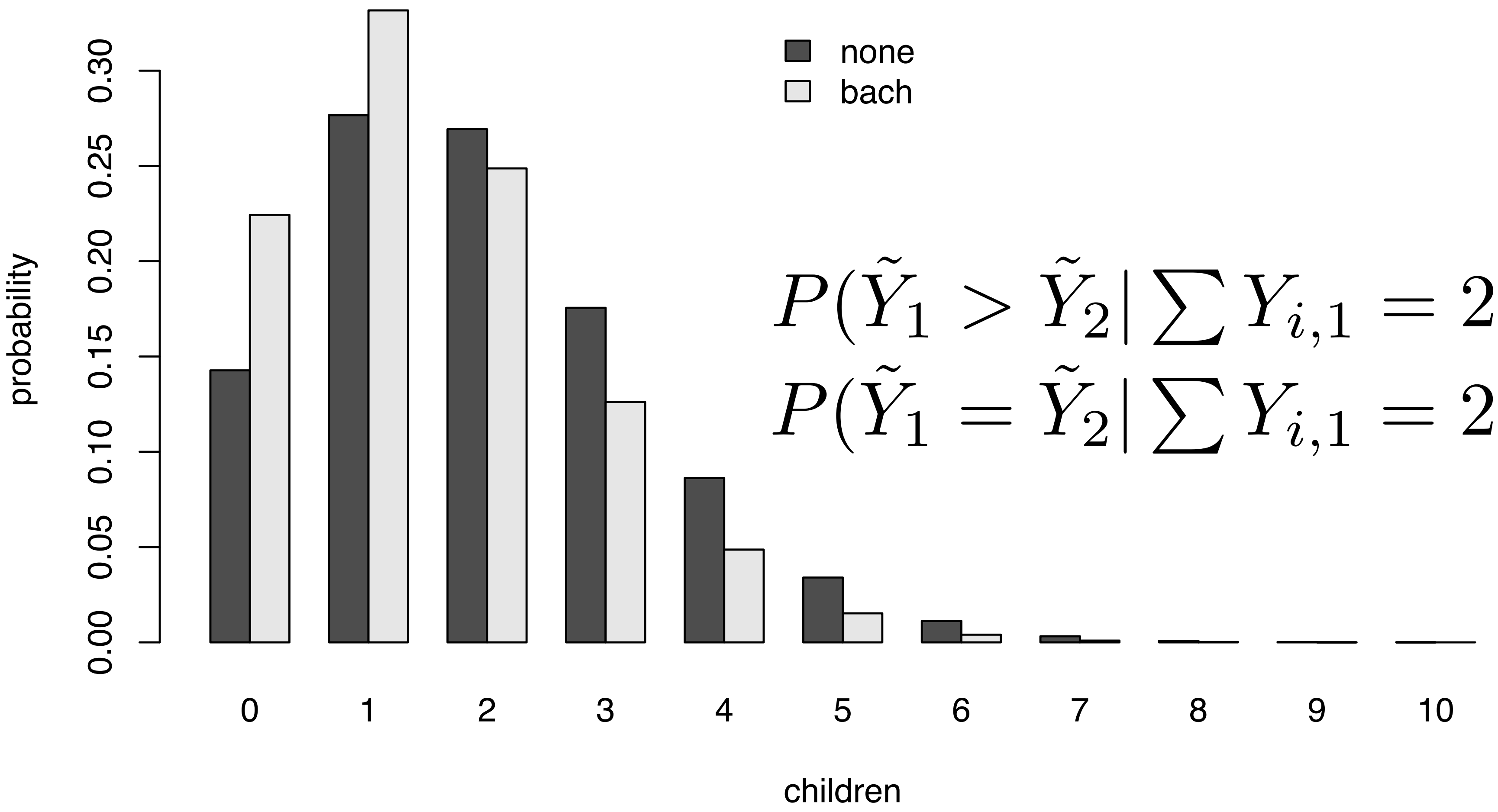


The posterior(s) provide substantial evidence that $\theta_1 > \theta_2$. E.g.,

$$P(\theta_1 > \theta_2 | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66) = 0.97$$

Example: posterior predictive(s)

Consider a randomly sampled individual from each population. To what extent do we expect the uneducated one to have more children than the other?



$$P(\tilde{Y}_1 > \tilde{Y}_2 | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66) = 0.48$$

$$P(\tilde{Y}_1 = \tilde{Y}_2 | \sum Y_{i,1} = 217, \sum Y_{i,2} = 66) = 0.22$$

The distinction between $\{\theta_1 > \theta_2\}$ and $\{\tilde{Y}_1 > \tilde{Y}_2\}$ is important: strong evidence of a difference between two populations does not mean the distance is large

Non-informative priors

- Priors can be difficult to construct
- There has long been a desire for prior distributions that can be guaranteed to play a minimal role in the posterior distribution
- Such distributions are sometimes called “reference priors”, and described as “vague”, “flat”, “diffuse”, or non-informative
- The rationale for using non-informative priors is often said to be to “let the data speak for themselves”, so that inferences are unaffected by the information external to the current data

Jeffreys' invariance principle

One approach that is sometimes used to define a non-informative prior distribution was introduced by Jeffreys (1946), based on considering one-to-one transformations of the parameter

Jeffreys' general principle is that any rule for determining the prior density should yield an equivalent posterior if applied to the transformed parameter

Naturally, one must take the sampling model into account in order to study this invariance

The Jeffreys' prior

Jeffreys' principle leads to defining the non-informative prior as $p(\theta) \propto [i(\theta)]^{1/2}$, where $i(\theta)$ is the Fisher information for θ

Recall that

$$i(\theta) = \mathbb{E}_\theta \left\{ \left(\frac{d}{d\theta} \log p(y|\theta) \right)^2 \right\} = -\mathbb{E}_\theta \left\{ \frac{d^2}{d\theta^2} \log p(y|\theta) \right\}$$

A prior so-constructed is called the Jeffreys' prior for the sampling model $p(y|\theta)$

Example:
Jeffreys' prior for the binomial sampling model

Consider the binomial distribution $Y \sim \mathrm{Bin}(n, \theta)$ which has log-likelihood

$$\log p(y|\theta) = c + y \log \theta + (n - y) \log(1 - \theta)$$

Therefore

$$i(\theta) = -\mathbb{E}_\theta \left\{ \frac{d^2}{d\theta^2} \log p(y|\theta) \right\} = \frac{n}{\theta(1 - \theta)}$$

So the Jeffreys' prior density is then

$$p(\theta) \propto \theta^{-1/2}(1 - \theta)^{-1/2}$$

which is the same as $\theta \sim \mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$

Example: non-informative?

This $\theta \sim \mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ Jeffreys' prior is non-informative in some sense, but not in all senses

The posterior expectation would give $a + b = 1$ sample(s) worth of weight to the prior mean $\frac{a}{a+b} = \frac{1}{2}$

This Jeffreys' prior "less informative" than the uniform prior which gives a weight of 2 samples to the prior mean (also 1)

In this sense, the "least informative" prior is

$$\theta \sim \mathrm{Beta}(0, 0)$$

But is this a valid prior distribution?

Propriety of priors

We call a prior density $p(\theta)$ proper if it does not depend on the data and is integrable, i.e.,

$$\int_\Theta p(\theta)\, d\theta < \infty$$

Proper priors lead to valid joint probability models $p(\theta, y)$ and proper posteriors ($\int_\Theta p(\theta|y)\, d\theta < \infty$)

Improper priors do not provide valid joint probability models, but they can lead to proper posterior by proceeding with the "Bayesian algebra"

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Example: continued

The $\theta \sim \mathrm{Beta}(0,0)$ prior is improper since

$$\int_0^1 \theta^{-1}(1-\theta)^{-1}\,d\theta = \infty$$

However, the $\mathrm{Beta}(0+y, 0+n-y)$ posterior that results is proper as long as $y \neq 0, n$

A $\theta \sim \mathrm{Beta}(\epsilon, \epsilon)$ prior, for $0 < \epsilon \ll 1$ is always proper, and (in the limit) non-informative

In practice, the difference between these alternatives $(\mathrm{Beta}(0,0), \mathrm{Beta}(\frac{1}{2}, \frac{1}{2}), \mathrm{Beta}(1,1))$ is small; all three allow the data (likelihood) to dominate in the posterior

Example:
Jeffreys' prior for the Poisson sampling model

Consider the Poisson distribution $Y \sim \mathrm{Pois}(\theta)$ which has log-likelihood

$$\log p(y|\theta) = c + y \log \theta - \theta$$

Therefore

$$i(\theta) = -\mathbb{E}_\theta \left\{ \frac{d^2}{d\theta^2} \log p(y|\theta) \right\} = \frac{1}{\theta}$$

So the Jeffreys' prior "density" is then

$$p(\theta) \propto \theta^{-1/2}$$

Example: Improper Jeffreys' prior

Since $\int_0^\infty \theta^{-1/2}\, d\theta = \infty$ this prior is improper

However, we can interpret is as a $G(\frac{1}{2}, 0)$, i.e., within the conjugate gamma family, to see that the posterior is

$$G\left(\frac{1}{2} + \sum_{i=1}^{n} y_i, 0 + n\right)$$

This is proper as long as we have observed one data point, i.e., $n \geq 1$, since

$$\int_0^\infty \theta^\alpha e^{-\beta\theta} < \infty \quad \text{for all } \alpha > 0,\ \beta \geq 1$$

Example: non-informative?

Again, this $G(\frac{1}{2}, 0)$ Jeffreys' prior is non-informative in some sense, but not in all senses

The (improper) prior $p(\theta) \propto \theta^{-1}$, or $\theta \sim G(0,0)$, is in some sense equivalent since it gives the same weight $(b=0)$ to the prior mean (although different)

It can be shown that, in the limit as the prior parameters $(a,b) \rightarrow (0,0)$ the posterior expectation approaches the MLE

- $p(\theta) \propto \theta^{-1}$ is a typical default for this reason

As before, the practical difference between these choices is small

Notes of caution

The search for non-informative priors has several problems, including

- it can be misguided: if the likelihood (data) is truly dominant, the choice among relatively flat priors cannot matter
- for many problems there is no clear choice

Bottom line: If so few data are available that the choice of non-informative prior matters, then one should put relevant information into the prior