UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Dipartimento di scienze economiche,
aziendali, matematiche e statistiche
“Bruno de Finetti”

Bayesian Statistics

Single parameter models

Francesco Pauli

A.A. 2018/19

Process of Bayesian data analysis

The steps of a Bayesian data analysis are

1. specify a full probability model: the joint distribution of observations and unknowns (parameter) or, which is the same, the prior and the likelihood
2. condition on the data: use Bayes theorem to obtain the posterior distribution
3. evaluate the fit of the model and the conclusions which the posterior implies.

Notation

The main characters

- we denote with Greek letters, typically θ, the parameter(s), unobservable quantities. θ can be a scalar or a vector.
- the observed data are denoted by y, if data are gathered on n units

$$y = (y_1, \ldots, y_n)$$

  where y_i can be a scalar or a vector (if more than one variable is observed on each unit).
  y can then be a scalar, a vector or a matrix.
- we will also use unknown but potentially observable quantities, that is, future observations, these will be denoted as $\tilde{y}$
- if covariates are available these will be denoted by x.

Model specification

Specifying a Bayesian model means specifying

- the distribution of y conditional on the parameter θ

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

  note that this is (proportional to) the likelihood
  $p(y|\theta) \propto L(\theta)(= L(\theta, y))$
- the prior distribution on θ

$$\theta \sim \pi(\theta)$$

Putting these together, we have specified the joint distribution of (y, θ)

$$p(y, \theta) = \pi(\theta)p(y|\theta)$$

and we can obtain the marginal distribution of y as

$$p(y) = \int_\Theta p(y, \theta)d\theta = \int_\Theta \pi(\theta)p(y|\theta)d\theta$$

Posterior distribution

Inference on θ will be based on the posterior distribution, which is derived through a straightforward application of Bayes theorem

$$\pi(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{p(y)}$$

The posterior distribution contains all the information on θ we have (from the data and prior to observing the data).

The work will have to do is to understand

- how to summarize the information in $\pi(\theta|y)$, to obtain for instance point and interval estimates or to perform hypotheses testing;
- how to explore the distribution, but for simple examples $p(y)$ is difficult to derive (impossible to derive analytically), so exploration of the posterior will be based on computational machinery (MCMC and other stuff) whose starting point is

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

Predictive distribution

We are sometimes interested on "unknown but potentially observable quantities" $\tilde{y}$ (think of prediction of y on new statistical units).

We assume that they behave like the data y, that is

$$\tilde{y}|\theta \sim p(\tilde{y}|\theta)$$

hence, unconditionally, the distribution of $\tilde{y}$ is

$$p(\tilde{y}) = \int_{\Theta} p(\tilde{y}|\theta)\pi(\theta)d\theta$$

the same as y. This is also called the *prior predictive distribution*

Predictive distribution

We are sometimes interested on "unknown but potentially observable quantities" $\tilde{y}$ (think of prediction of y on new statistical units).

After the data y have been observed, we can compute the *posterior predictive distribution*

$$\begin{aligned} p(\tilde{y}|y) &= \int_\Theta p(\tilde{y}, \theta|y) d\theta \\ &= \int_\Theta p(\tilde{y}|\theta, y)\pi(\theta|y) d\theta \\ &= \int_\Theta p(\tilde{y}|\theta)\pi(\theta|y) d\theta \end{aligned}$$

where we note that the conditional iid assumption implies that

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

Exchangeability

A common hypotheses in statistical inference is that observations are independent and identically distributed, that is, we consider n statistical units, collect $y_1, \ldots, y_n$ and assume these to be IID.

In Bayesian inference, where the inference process itself is fully probabilistic independence of observations would imply that we can not learn about future observations based on the past (since y_{n+1} would be independent of $y_1, \ldots, y_n$).

The common hypotheses in Bayesian inference is that observations are exchangeable meaning that the joint distribution of $(y_1, \ldots, y_n)$ is invariant to permutations of the indexes; in formulas

$$p(y_1, \ldots, y_n) = p(y_{i_1}, \ldots, y_{i_n})$$

for any $(i_1, \ldots, i_n)$ permutation of $1, \ldots, n$.

Exchangeability and conditional independence

We will usually specify the model assuming that

- $y_1, \ldots, y_n$ iid conditional on θ
- $\theta \sim \pi(\theta)$

this implies that $y_1, \ldots, y_n$ is exchangeable.

In fact consider the unconditional distribution

$$
\begin{aligned}
p(y_{i_1}, \ldots, y_{i_n}) &= \int p(y_{i_1}, \ldots, y_{i_n}|\theta)\pi(\theta)d\theta \\
&= \int \prod_{j=1}^{n} p(y_{i_j}|\theta)\pi(\theta)d\theta \\
&= \int \prod_{i=1}^{n} p(y_i|\theta)\pi(\theta)d\theta = \pi(y_1, \ldots, y_n)
\end{aligned}
$$

de Finetti's theorem

In the special case of $y_1, \ldots, y_n$ binary variables it can be shown that exchangeability is equivalent to conditional iid.

Theorem (de Finetti)

Let $Y_1, \ldots, Y_n$ be Bernoulli r.v., then they are exchangeable if and only if there exist a random variable θ valued in $[0, 1]$ such that

$$p(y_1, \ldots, y_n) = \int_0^1 \theta^{\sum y_i} (1-\theta)^{n-\sum y_i} dP(\theta)$$

An extension of this theorem to general random variables exist.

Exchangeability and conditional independence

So the following are equivalent

- $y_1, \ldots, y_n$ exchangeable
- $y_1, \ldots, y_n$ iid conditional on θ

in other words we assume that

> *observations are iid if we know the data generating mechanism*

since we do not know it the observations are not independent, on the contrary

> y_1 *gives informations on* y_2 **because** *it gives information on the data generating mechanism* θ*.*

Indice

Which example

We do not need complicated model and data to illustrate the application of Bayesian statistics.

In what follows we consider inference for discrete quantities (rather than parameters), framed as a Bayesian inference.

Inference about a genetic status: prior and model

Hemophilia is due to a recessive gene in the X-chromosome, that is, if X^* denotes an X-chromosome with the hemophilia gene,

- X^*X^* is a female with the disease
- X^*X is a female without the disease but with the gene
- X^*Y a male with the disease

Mary has

- an affected brother $\Rightarrow X^*Y$
- an unaffected mother $\Rightarrow XX^*$ or XX
- an unaffected father $\Rightarrow XY$

overall the mother must be XX^*.
Let $\theta = 1$ if Mary has the gene (is XX^*) and 0 otherwise (XX), then

based on the above information, **prior** *to any observation,*

$$P(\theta = 1) = 1/2$$

Inference about a genetic status: data and likelihood

Data consist of the status of Mary's two sons, who are not affected.

Let then y_i be an indicator equal to 1 if the i-th son is affected

$$P(y_i = 1|\theta) = \begin{cases} 0.5 & \text{if } \theta = 1 \\ 0 & \text{otherwise} \end{cases}$$

The likelihood function corresponding to Mary's two sons is

$$L(\theta) = P(y_1 = y_2 = 0|\theta) = \begin{cases} 0.25 & \text{if } \theta = 1 \\ 1 & \text{otherwise} \end{cases}$$

Inference about a genetic status: posterior

Prior and likelihood are combined to obtain the posterior, let $D = (y_1 = y_2 = 0)$,

$$P(\theta = 1|D) = \frac{P(D|\theta = 1)P(\theta = 1)}{P(D|\theta = 1)P(\theta = 1) + P(D|\theta = 0)P(\theta = 0)}$$
$$= \frac{0.25 \times 0.5}{0.25 \times 0.5 + 1 \times 0.5} = 0.20$$

with a discrete parameter it is useful also to express the above formula in terms of odds

$$\frac{\pi(\theta_1|y)}{\pi(\theta_2|y)} = \frac{p(y|\theta_1)}{p(y|\theta_2)} \frac{\pi(\theta_1)}{\pi(\theta_2)}$$

the posterior odds of θ_1 over θ_2 are given by the prior odds times the likelihood ratio

$$\frac{P(\theta = 1|D)}{P(\theta = 0|D)} = \frac{P(D|\theta = 1)}{P(D|\theta = 0)} \frac{P(\theta = 1)}{P(\theta = 0)}$$

Inference about a genetic status: predictive distributions

Prior to the observations the predictive distribution is

$$\begin{aligned} P(y_1 = 1) &= P(y_1 = 1|\theta = 1)P(\theta = 1) + P(y_1 = 1|\theta = 0)P(\theta = 0) \\ &= 0.5 \times 0.5 + 0 \times 0.5 = 0.25 \end{aligned}$$

Given the data the posterior predictive is

$$\begin{aligned} P(\tilde{y}_3 = 1|D) &= P(\tilde{y}_3 = 1|\theta = 1, D)P(\theta = 1|D) + P(\tilde{y}_3 = 1|\theta = 0, D)P(\theta = 0|D) \\ &= P(\tilde{y}_3 = 1|\theta = 1)P(\theta = 1|D) + P(\tilde{y}_3 = 1|\theta = 1)P(\theta = 0|D) \\ &= 0.5 \times 0.2 + 0 \times 0.8 = 0.1 \end{aligned}$$

Inference about a genetic status: update

Suppose a third son is born and he is not affected, that is we have a new observation $y_3 = 0$, in order to obtain the new posterior distribution we can use the old posterior $P(\theta = 1|D)$ as a prior and update it based on the likelihood $P(y_i = 0|\theta)$

$$P(\theta = 1|D, y_3 = 0) = \frac{P(y_3 = 0|\theta = 1)P(\theta = 1|D)}{P(y_3 = 0|\theta = 1)P(\theta = 1|D) + P(y_3 = 0|\theta = 0)P(\theta = 0|D)}$$
$$= \frac{0.5 \times 0.2}{0.5 \times 0.2 + 1 \times 0.8} = 0.111$$

A similar mechanism works with the odds

$$\frac{P(\theta = 1|D, y_3 = 0)}{P(\theta = 0|D, y_3 = 0)} = \frac{P(y_3 = 0|\theta = 1)}{P(y_3 = 0|\theta = 0)} \frac{P(\theta = 1|D)}{P(\theta = 0|D)}$$
$$\frac{1}{8} = \frac{1}{2} \frac{1}{4}$$

The same result is obtained by starting from the prior and considering the data $D' = (y_1 = y_2 = y_3 = 0)$.

Indice

Data

We observe

$$y_1, \ldots, y_n$$

where $y_i \in \{0, 1\}$, and, conditional on θ the $y_1, \ldots, y_n$ are

- independent: $y_1, \ldots, y_n$ independent conditional on θ
- identically distributed: $P(y_i = 1|\theta) = \theta \ \ \forall i$.

Equivalently, we could say that $y_1, \ldots, y_n$ are exchangeable:

$$p(y_1, \ldots, y_n) = p(y_{i_1}, \ldots, y_{i_n}) \text{ for any permutation } i_1, \ldots, i_n \text{ of } 1, \ldots, n$$

(note that here $p(y_1, \ldots, y_n) = P(Y_1 = y_1 \wedge Y_2 = y_2 \wedge \ldots \wedge Y_n = y_n)$).

Likelihood

By virtue of exchangeability, data can be summarized by the number of successes

$$y = \sum_{i=1}^{n} y_i$$

which, conditional on θ (and on n), has a binomial distribution

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

(note that here $p(y|\theta) = P(Y = y|\theta)$).

Inference for a probability

Although a very simple model, it has relevant applications. Also, it was dealt with by many of the first scholars working in probability.

In fact, it was the motivating example to develop Bayesian statistics both for T. Bayes and for Laplace.

The former considered it in an abstract context, the latter had the aim of estimating the probability of a female birth.

Posterior distribution

Let us assume, for the moment without discussion, a uniform prior on θ

$$\pi(\theta) = I_{[0,1]}(\theta)$$

then

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

and

$$\pi(\theta|y) \propto \theta^y (1-\theta)^{n-y}$$

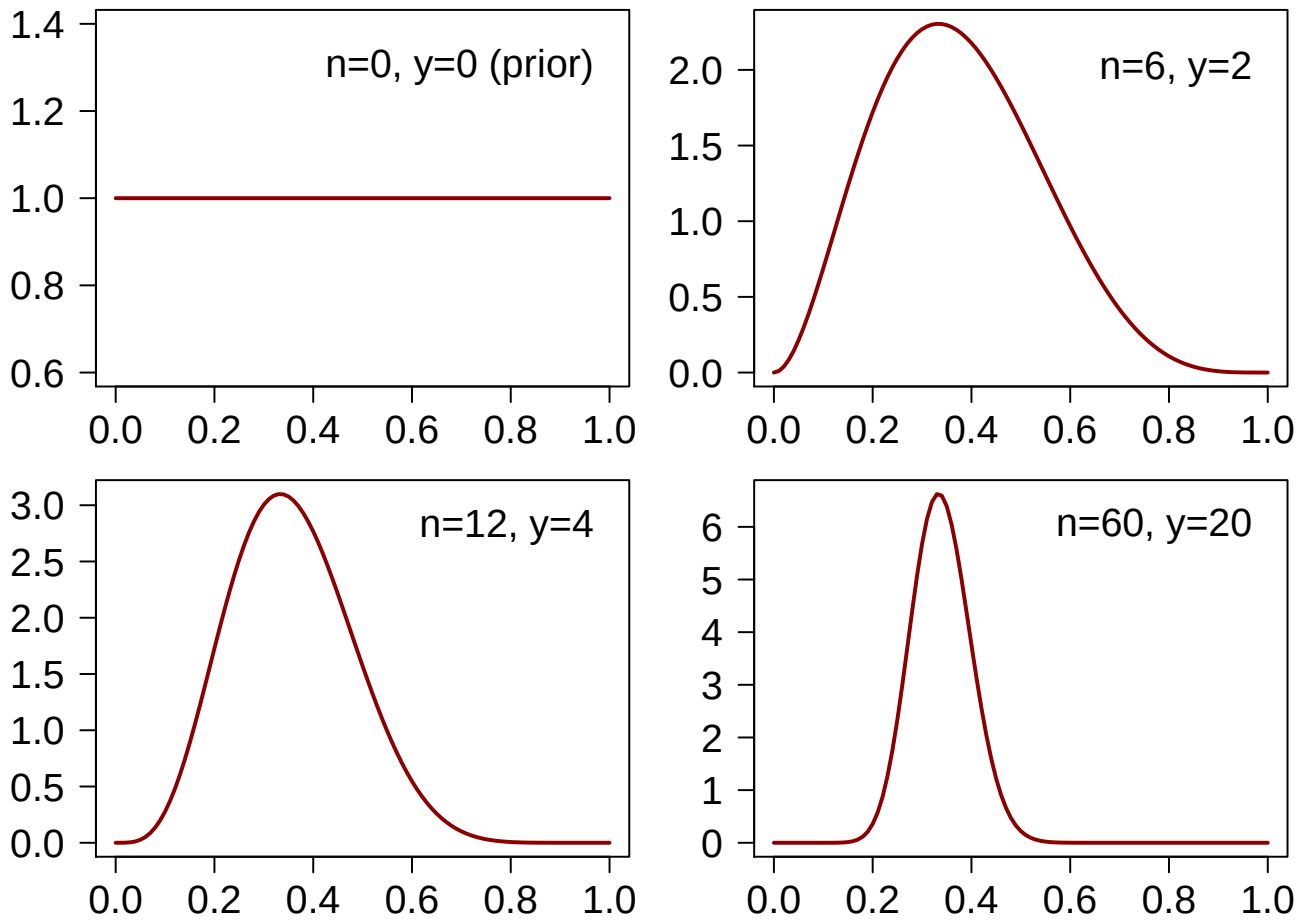
where we recognize the kernel of a **Beta distribution** with parameters $y+1$ and $n-y+1$, so

$$\pi(\theta|y) = \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \theta^y (1-\theta)^{n-y}.$$

and can also write

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

Posterior distributions



Laplace example, revisited

Laplace observed 241 945 females and 251 527 males, that is if

$$\theta = \text{probability of a female birth}$$

he had

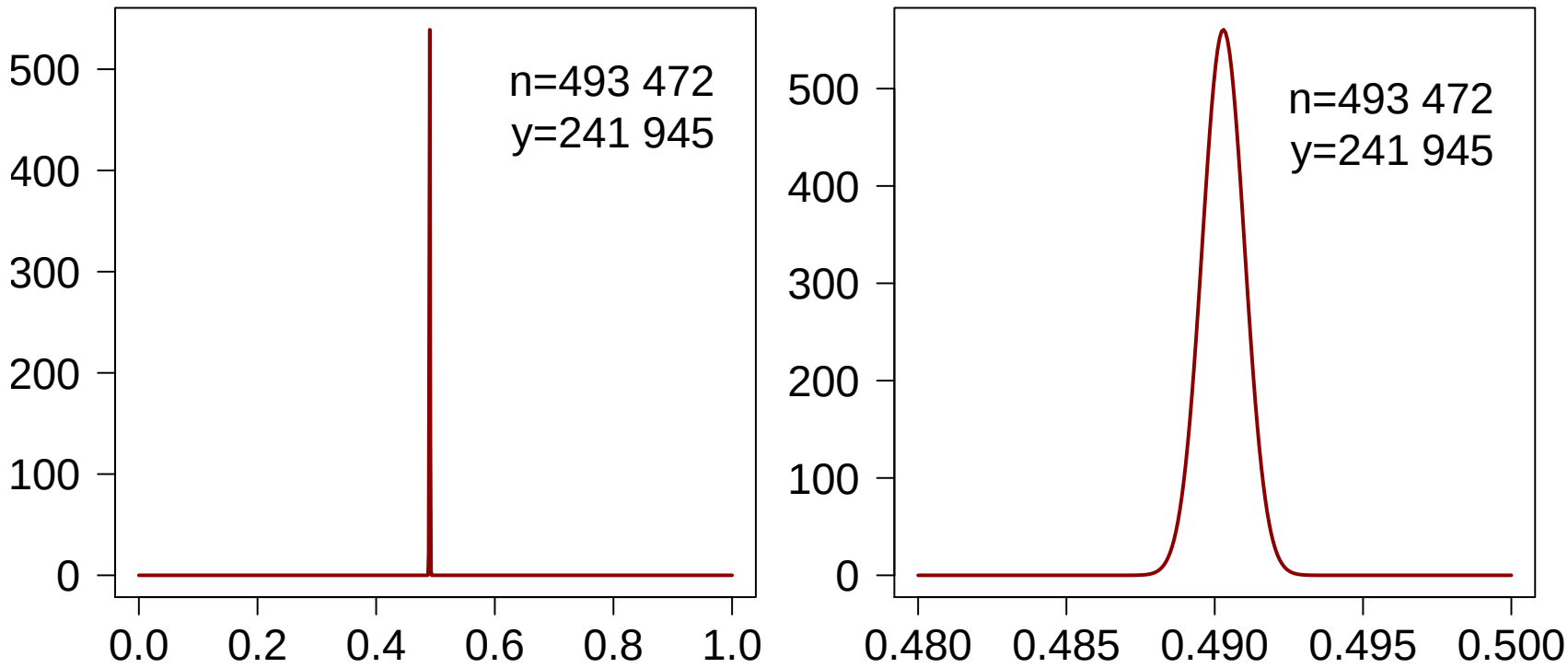
$$n = 241\,945 + 251\,527 = 493\,472; \quad y = 241\,945$$

hence the posterior distribution for θ is a Beta$(241\,946, 251\,528)$ and

$$P(\theta \geq 0.5|y) \approx 1.15 \times 10^{-42}$$

We ought to appreciate the fact that to get to this number Laplace had to develop appropriate approximations, it is not immediate even today (R may give 0 depending on how the problem is formulated due to machine precision).

Posterior distribution for Laplace



Prediction

Consider a new observation $\tilde{y}$, which behaves like the y_i, that is

- is independent of $y_1, \ldots, y_n$ conditional on θ
- $P(y_i = 1|\theta) = \theta$

then the prior predictive distribution is

$$P(\tilde{y} = 1) = \int_0^1 \theta \pi(\theta) d\theta = \int_0^1 \theta d\theta = E(\theta) = 1/2$$

while the posterior predictive distribution is

$$P(\tilde{y} = 1|y) = \int_0^1 \theta \pi(\theta|y) d\theta = E(\theta|y) = \frac{y+1}{n+2}$$

[Detail]

$$
\begin{aligned}
P(\tilde{y}=1|y) &= \int_0^1 \theta \pi(\theta|y) d\theta \\
&= \int_0^1 \theta \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \theta^y (1-\theta)^{n-y} d\theta \\
&= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \int_0^1 \theta^{y+1} (1-\theta)^{n-y} d\theta \\
&= \frac{\Gamma(n+2)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(y+2)\Gamma(n-y+1)}{\Gamma(n+3)} \\
&= \frac{y+1}{n+2}
\end{aligned}
$$

Detail: prior predictive distribution for y

Consider the distribution of y prior to observing the data (based on the uniform prior on θ)

$$
\begin{aligned}
p(y) &= \int_0^1 p(y,\theta)\pi(\theta)d\theta \\
&= \int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}d\theta \\
&= \binom{n}{y}\int_0^1 \theta^y(1-\theta)^{n-y}d\theta \\
&= \binom{n}{y}\frac{\Gamma(y+1)\Gamma(n-y+1)}{\Gamma(n+2)} \\
&= \frac{1}{n+1}
\end{aligned}
$$

Summaries of the posterior distribution

The posterior is the result of the inference on θ, it is relevant to summarize the information it contains, this can be done in the usual ways in which we summarize a probability distribution, so we summarize the position with

- the mean
- the median
- the mode

and the variability with the

- variance (standard deviation)

Posterior mean and mode for the binomial model

The posterior mean is

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

which is a compromise between the observed proportion y/n and the prior mean $1/2$ (more on this later).

The posterior mode is

$$\text{Mode}(\theta|y) = \frac{y}{n}$$

which is the maximum likelihood estimator.

Posterior variance for the binomial model

The posterior variance is

$$V(\theta|y) = \frac{(y+1)(n-y+1)}{(n+2)^2(n+3)}$$

which is less readable, we notice that it has n^3 at the denominator and n^2 at the numerator.

We may compute the average over y

$$\begin{aligned} E(V(\theta|y)) &= \frac{1}{(n+2)^2(n+3)} E((y+1)(n-y+1)) \\ &= \frac{1}{(n+2)^2(n+3)} E(ny+n-y^2+1) \\ &= \frac{1}{(n+2)^2(n+3)} (n^2/6+5n/6+1) \end{aligned}$$

Posterior intervals

Another common way to summarize the posterior conveying uncertainty is to use intervals of a given posterior probability, say $100(1-\alpha)\%$, this is any interval $[\theta_L, \theta_U]$ such that

$$P(\theta_L \leq \theta \leq \theta_U | y) = 1 - \alpha$$

(This is also called a credibility interval.)

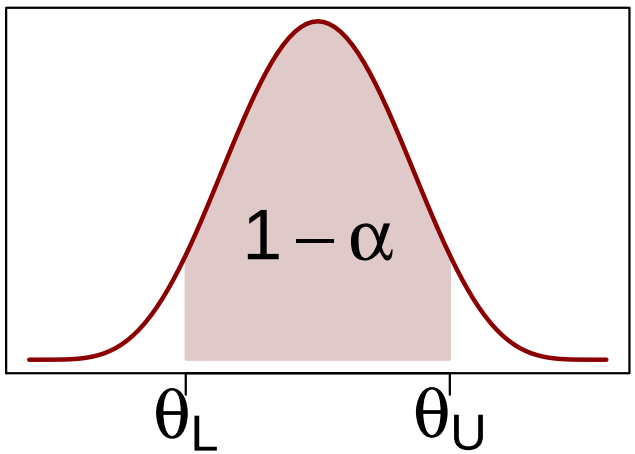
It is somehow the analogue of a confidence interval in classical statistics, but notice the different interpretation, here we say that the unknown parameter lies in the interval with the given probability (rather than saying that the interval is random ...).

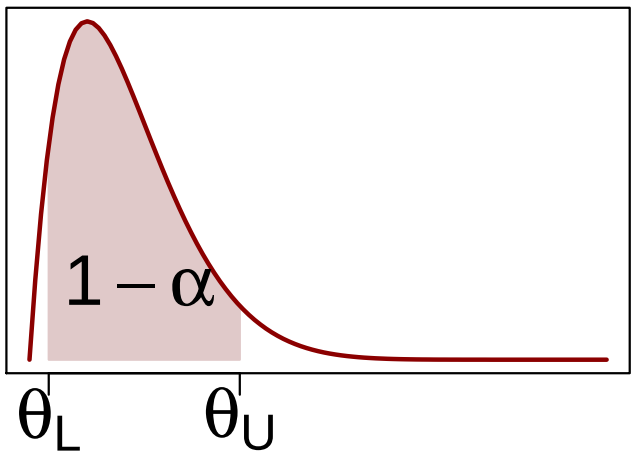
Central posterior intervals

The easiest way to obtain a posterior interval of probability $1 - \alpha$ is to set

- θ_L $\alpha/2$ quantile of $\pi(\theta|y)$
- θ_U $1 - \alpha/2$ quantile of $\pi(\theta|y)$

Below, two examples of central posterior intervals based on quantiles.





High posterior density regions

A different summary of posterior uncertainty is the highest posterior density region: the set of values that contains probability $1 - \alpha$ but also have posterior density higher than values outside.

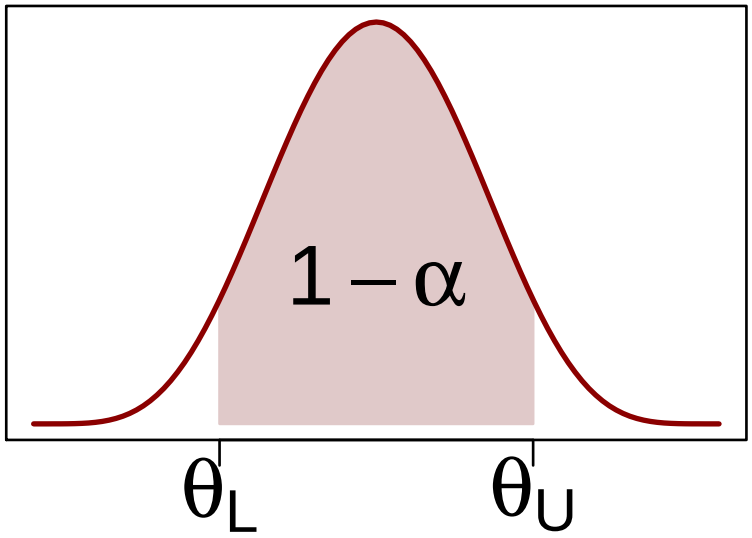
In formulas

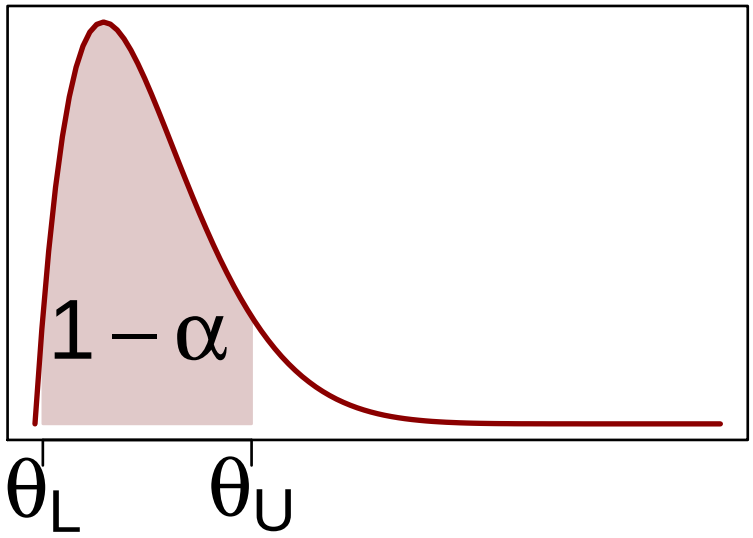
$$\{\theta | \pi(\theta|y) > c_\alpha\}$$

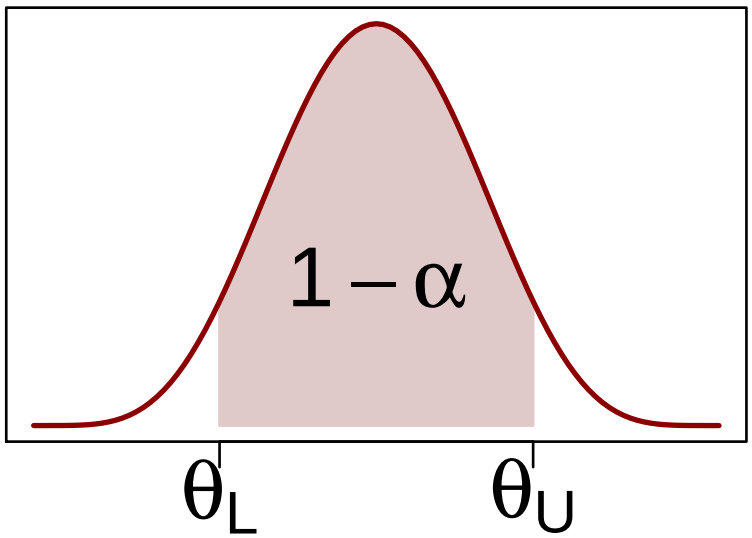
where c_α is such that

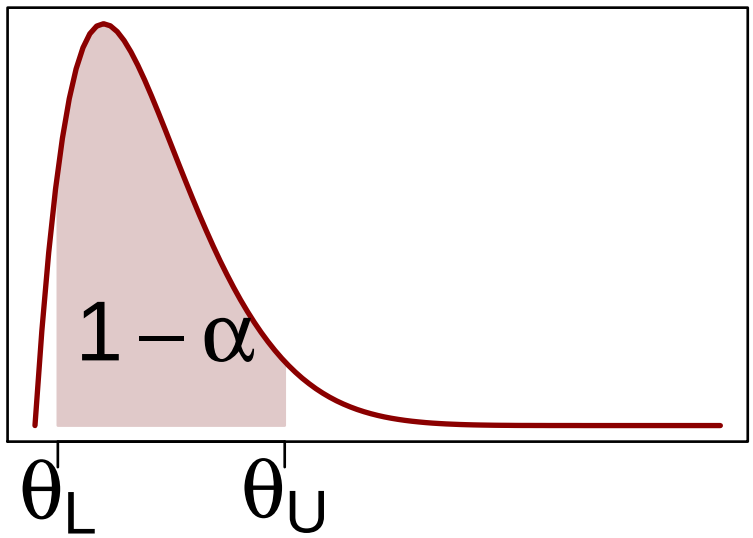
$$\int_{\theta|\pi(\theta|y)>c_\alpha} \pi(\theta|y) = 1 - \alpha$$

HPD regions v. Central posterior intervals









Prior

We considered a uniform prior on θ, this has been the choice of both Bayes and Laplace, who (loosely speaking) justified it

- Bayes based on the fact that it implies a uniform predictive prior on y
- Laplace based on the so called 'principle of insufficient reason' because he had no information about θ

Afterwards different approaches to the prior specification have been considered, in what follows we discuss different choices and look at their consequences, keeping in mind the following

- a prior need only to reasonably summarize the knowledge we have on θ
- if this information is scarce, the effect of the prior should vanish as enough data are collected

Conjugate prior

A convenient type of prior is the kind that leads to a posterior in the same family, this property is called conjugacy.

This is not available for any likelihood (more on that later), for the Binomial model it is represented by the Beta distribution:

If $\theta \sim \text{Beta}(\alpha, \beta)$ then $\theta|y \sim \text{Beta}(\alpha + y, \beta + n)$

as is easily checked:

$$\begin{aligned}\pi(\theta|y) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n+\beta-1}\end{aligned}$$

Posterior mean

Let us synthesize the posterior distribution using the expectation

$$
\begin{aligned}
E(\theta|y) &= \int \theta\pi(\theta|y)d(\theta) = \frac{\alpha + y}{\alpha + \beta + n} \\
&= \frac{\alpha + \beta}{\alpha + \beta + n}\frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n}\frac{y}{n} \\
&= \frac{\alpha + \beta}{\alpha + \beta + n} + \frac{n}{\alpha + \beta + n}
\end{aligned}
$$

The posterior mean is a weighted average of the prior expectation and the ML estimate, where

- ML estimate prevails if n is large;
- ML estimate prevails if α and β are small (the variance of the prior distribution is large). It is worth noting that $\alpha + \beta$ can be interpreted as the equivalent number of observation of the prior distribution.

Posterior variance

The posterior variance is

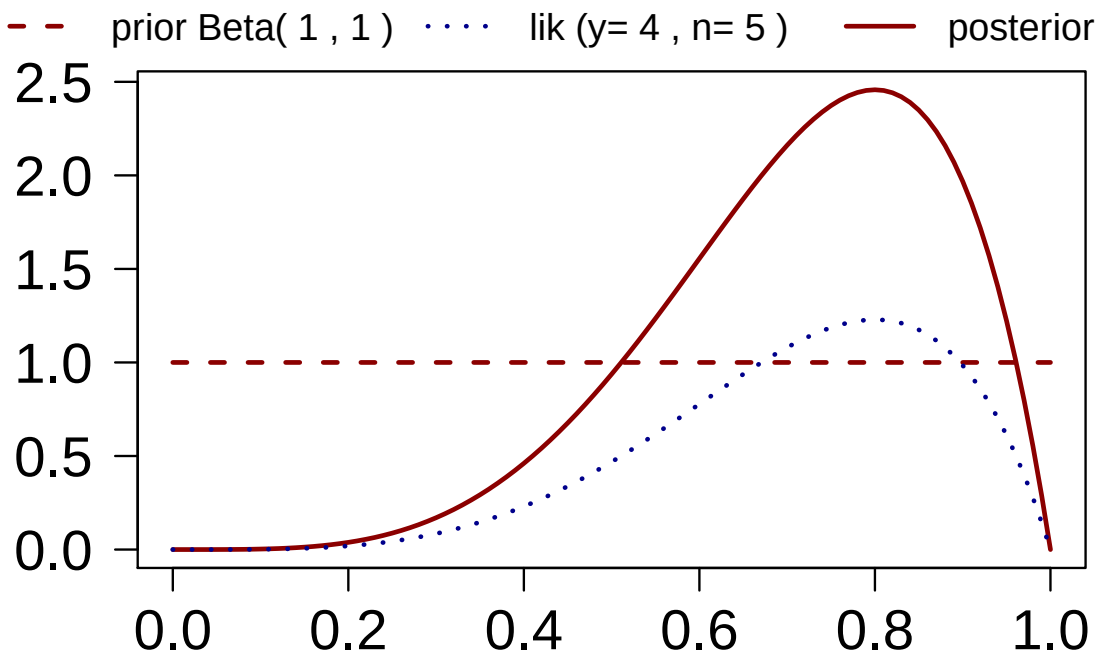
$$V(\theta|y) = \frac{(\alpha + y)(\beta + n - y)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)} = \frac{E(\theta|y)(1 - E(\theta|y))}{\alpha + \beta + n + 1}$$

As y and n gets big

- $E(\theta|y) \approx y/n$
- $V(\theta|y) \approx \frac{1}{n}\frac{y}{n}\left(1 - \frac{y}{n}\right)$

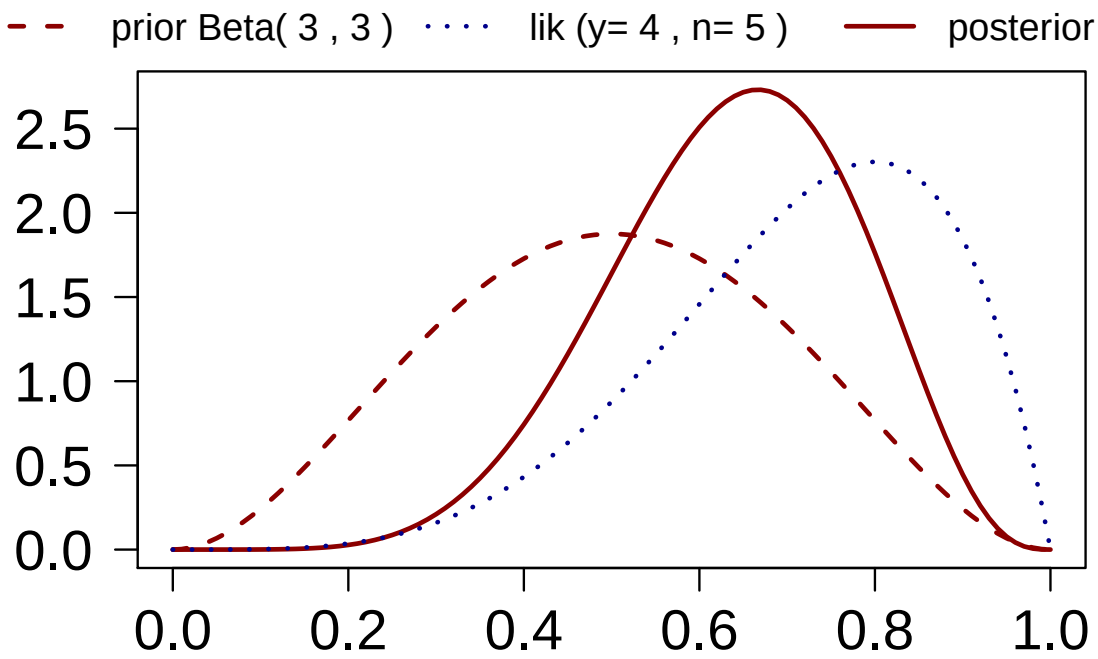
Different priors

With a uniform prior we get this



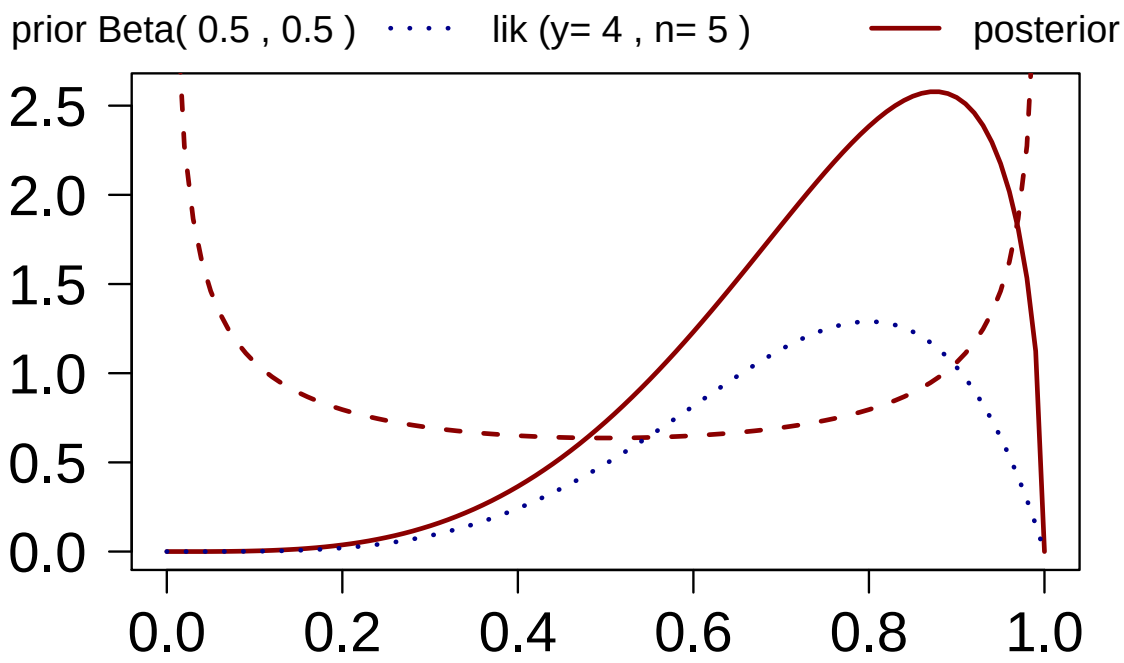
Different priors

We may have a different opinion: we may think that θ is more likely not to be extreme



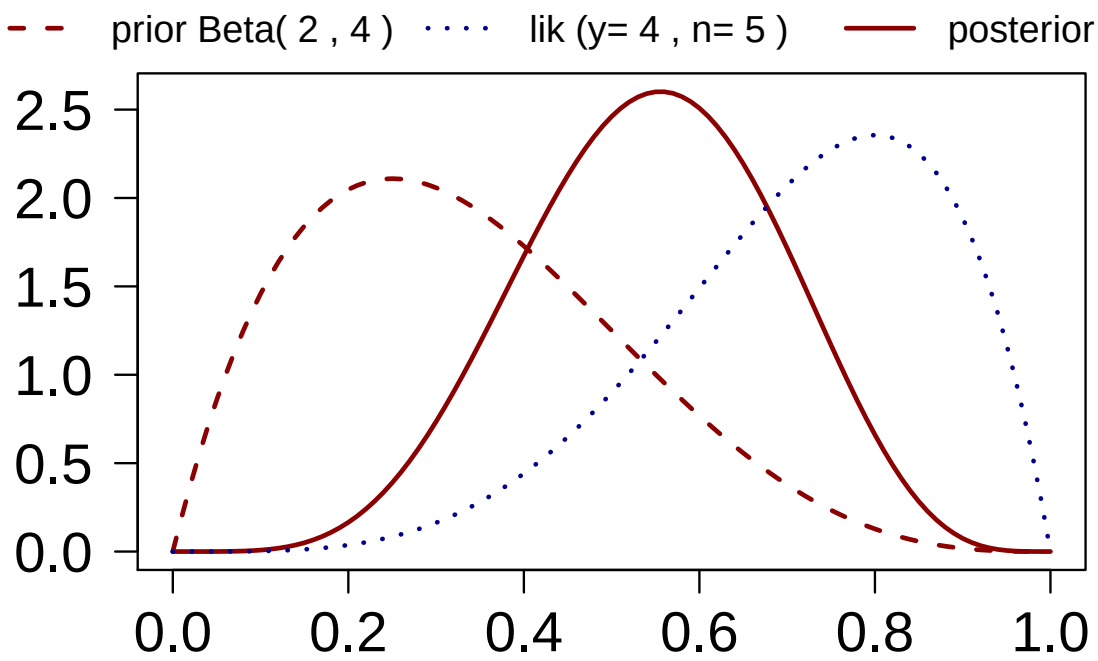
Different priors

... or more likely to be extreme



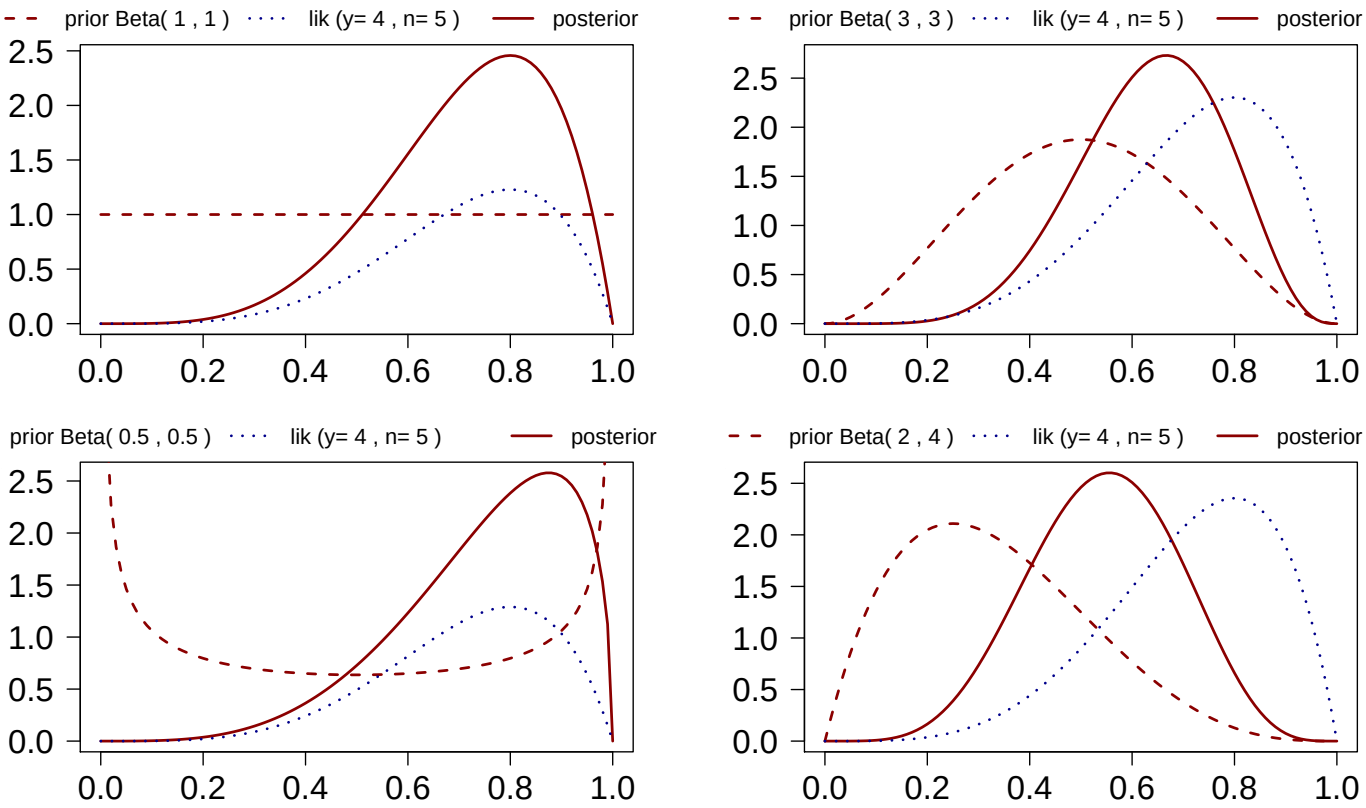
Different priors

Furthermore, we may prefer value below 0.5



Different priors

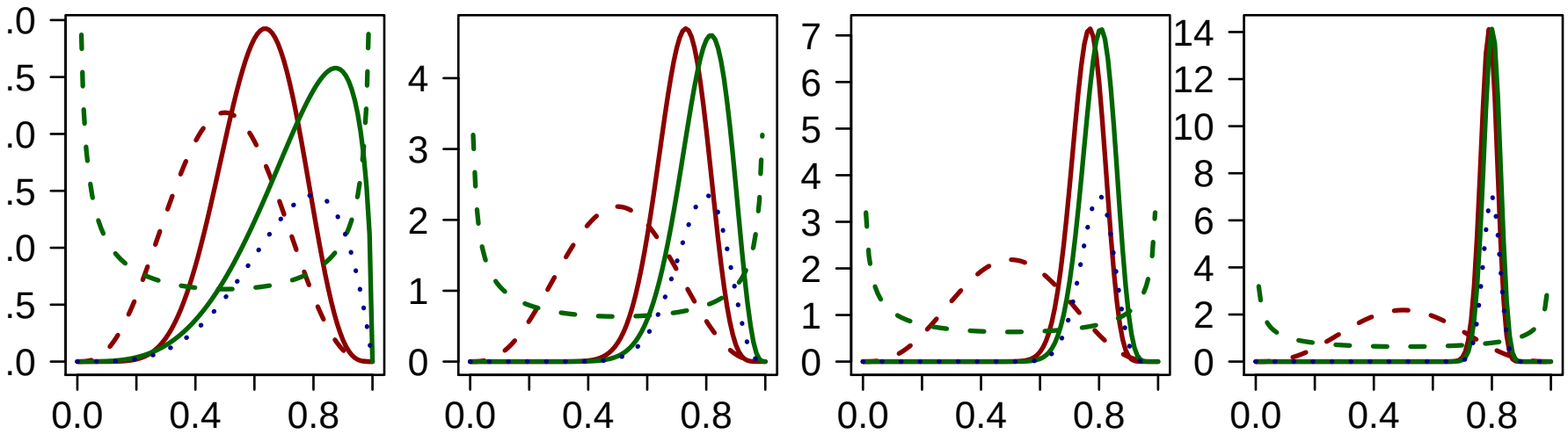
Different priors $\Rightarrow$ different posteriors



Prior effect as n increases

The effect of the prior, however, tend to disappear as enough sample information is entered.

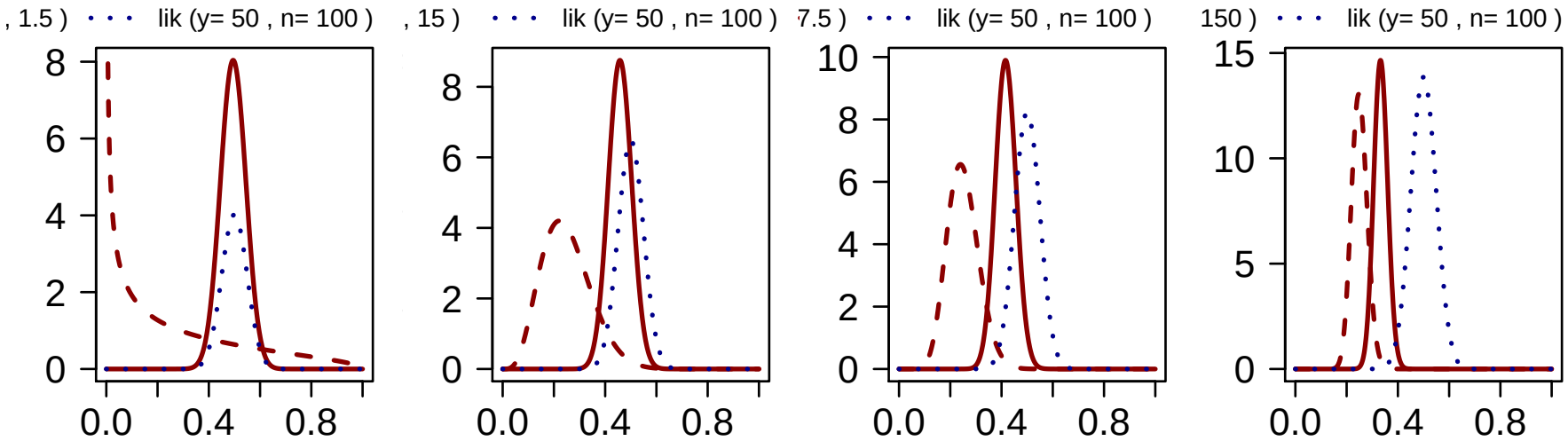
In the following we observe the effect on the posterior of two distinct priors samples of $n = 5, 20, 50, 200$, always with $y/n = 0.8$ and dif



Prior effect as $\alpha + \beta$ increases

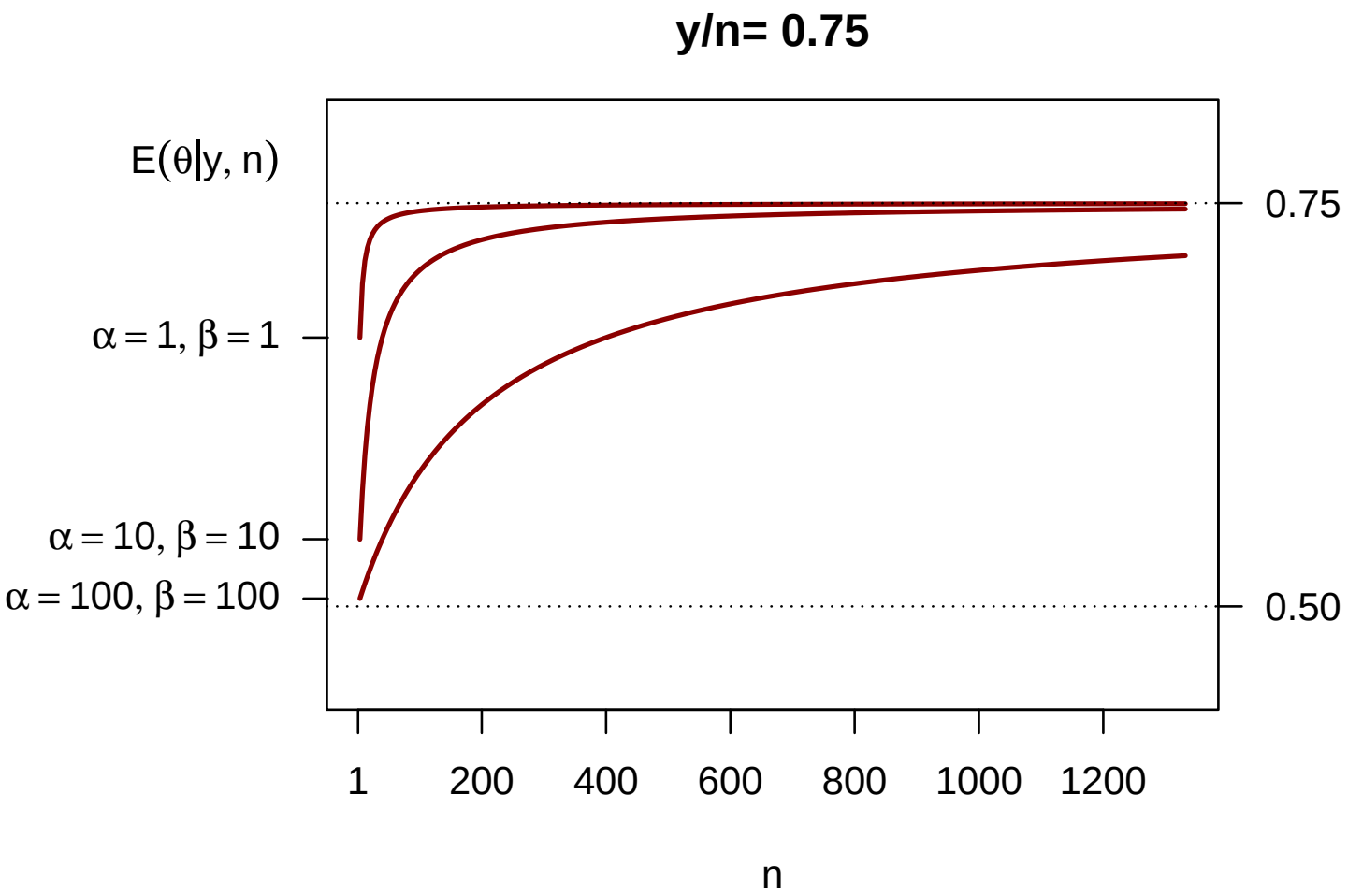
We can see things from another point of view and consider different priors with the same sample.

We observe a sample with $n = 100$ and $y = 50$, the prior mean is 0.25, $\alpha + \beta$ is 2, 20, 50, 200



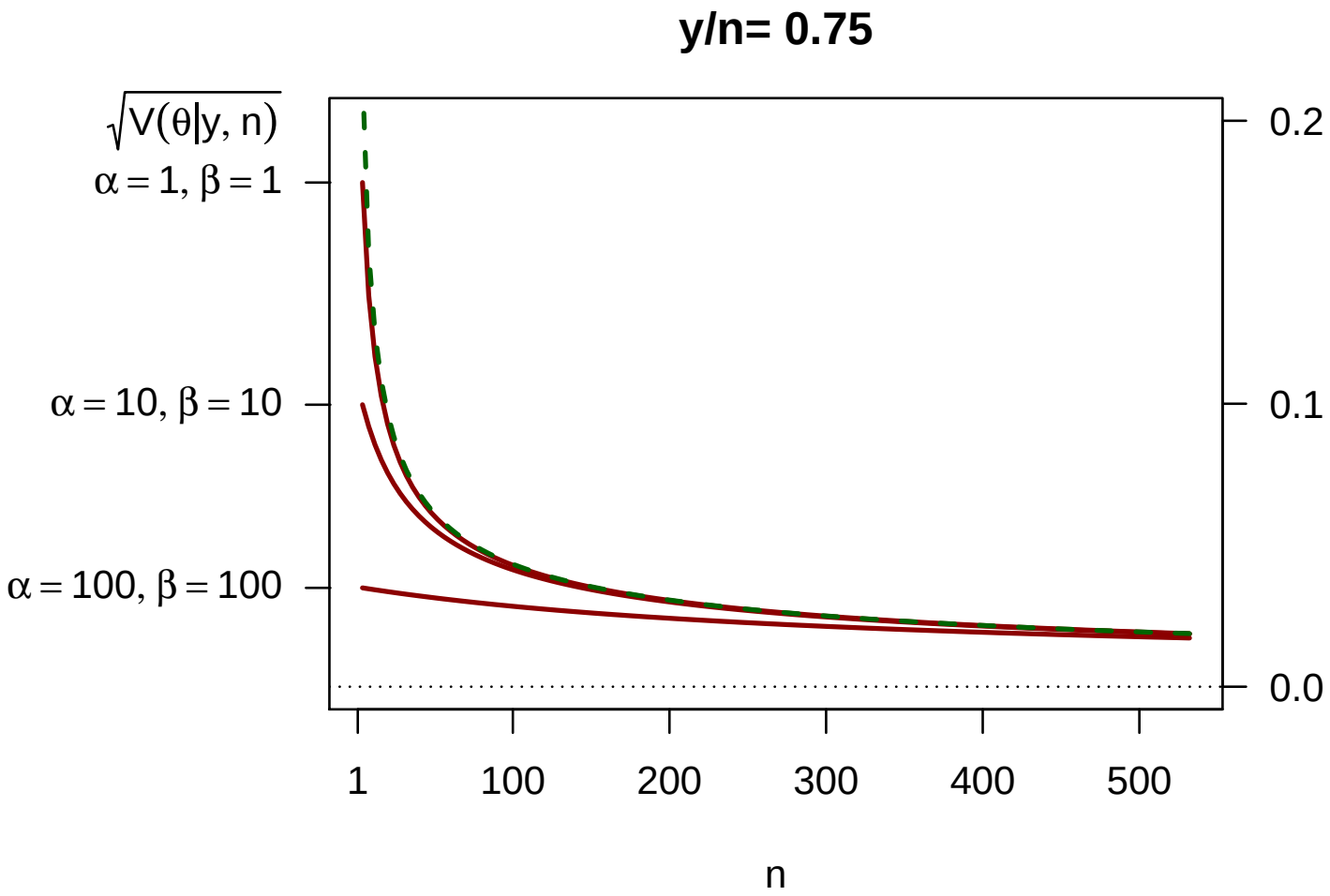
Posterior mean

Posterior mean as n increases for different priors



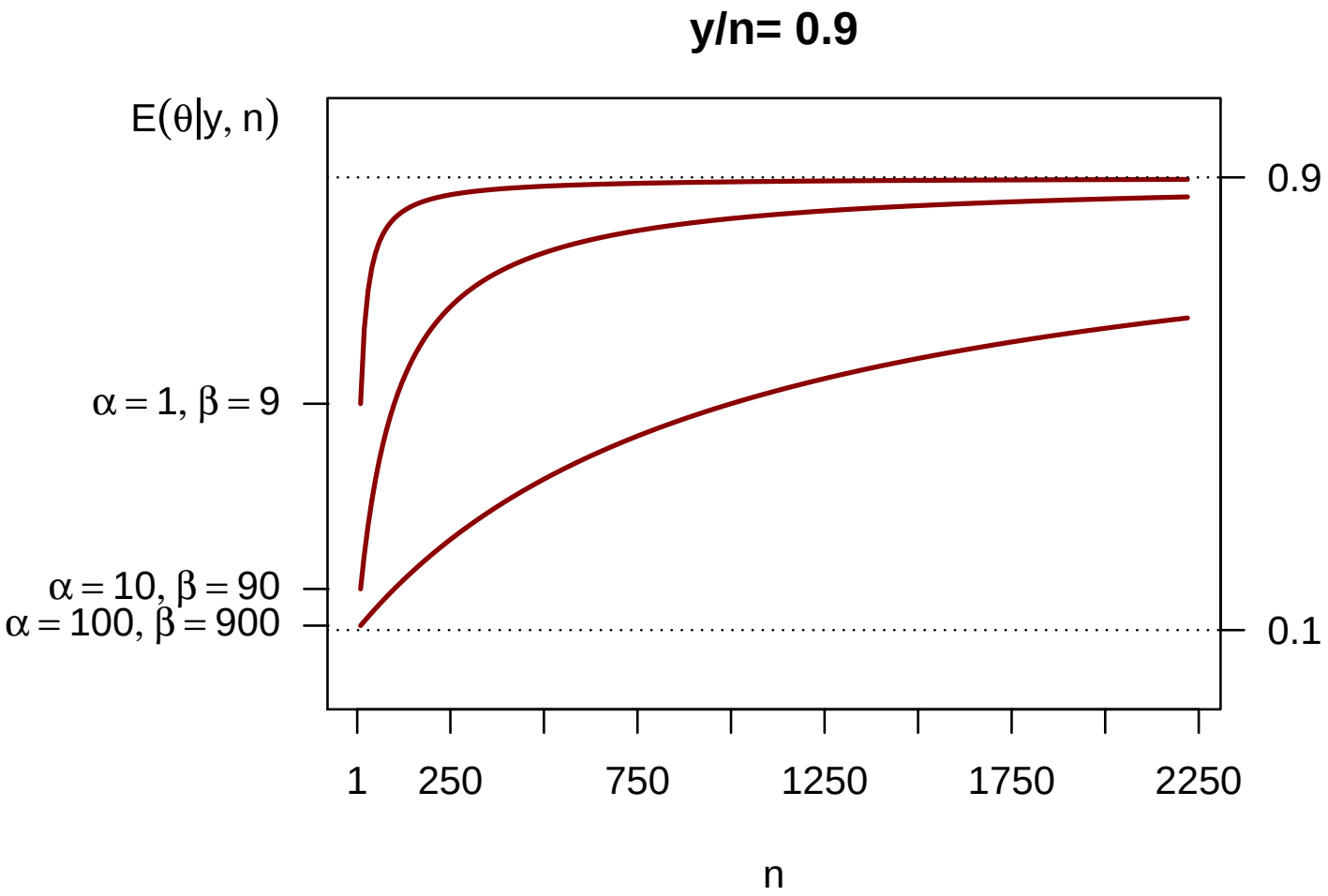
Posterior variance

Posterior mean as n increases for different priors



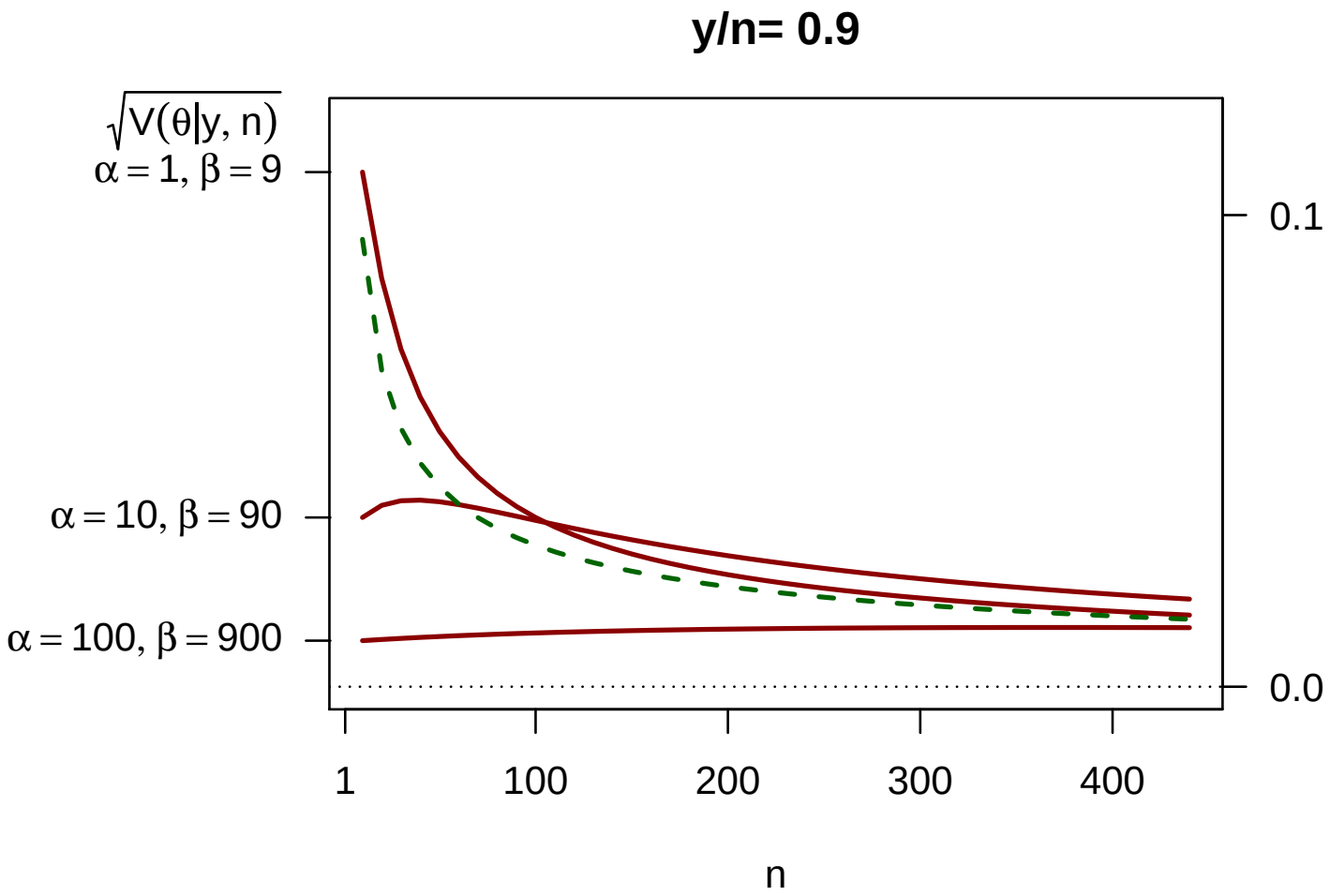
Posterior mean: conflicting priors

Posterior mean as n increases for different priors



Posterior variance: conflicting priors

Posterior mean as n increases for different priors



Indice

Exponential family: definition

Recall that a family of distributions $\mathcal{F} = \{p(y|\theta) : \theta \in \Theta \subset \mathbb{R}^d\}$ is an exponential family if its elements can be written as

$$p(y|\theta) = f(y)g(\theta)e^{\phi(\theta)^T u(y)}$$

where

- $f : \mathbb{R} \to \mathbb{R}$
- $g : \mathbb{R}^d \to \mathbb{R}$
- $\phi : \mathbb{R}^d \to \mathbb{R}^d$
- $u : \mathbb{R}^d \to \mathbb{R}^d$

are known functions.

$\phi(\theta)$ is called the natural parameter of $\mathcal{F}$.

Exponential family: likelihood and sufficient statistic

If a vector of observations $y = (y_1, \ldots, y_n)$ is observed and y_i are IID following a distribution from $\mathcal{F}$

$$p(y|\theta) = \left(\prod_{i=1}^{n} f(y_i) \right) g(\theta)^n \exp \left\{ \phi(\theta)^T \sum_{i=1}^{n} u(y_i) \right\}$$

hence

$$p(y|\theta) \propto g(\theta)^n \exp \left\{ \phi(\theta)^T t(y) \right\}$$

where

$$t(y) = \sum_{i=1}^{n} u(y_i)$$

is a sufficient statistic.

Conjugate distribution for an exponential family

If the prior is of the form

$$\pi(\theta) \propto g(\theta)^{\eta} e^{\phi(\theta)^T \nu}$$

then the posterior is

$$\begin{aligned}\pi(\theta|y) &\propto g(\theta)^{\eta} e^{\phi(\theta)^T \nu} g(\theta)^n \exp\left\{\phi(\theta)^T t(y)\right\} \\ &\propto g(\theta)^{\eta+n} e^{\phi(\theta)^T (\nu + t(y))}\end{aligned}$$

which has the same form as the prior.

It can be shown that only exponential families of distributions have natural conjugate priors.

Indice

Likelihood

Assume that observations come from a Gaussian distribution with a known variance (σ^2), so

$$y_1, \ldots, y_n \sim IID\left(\mathcal{N}\left(\theta, \sigma^2\right)\right) \text{ conditional on } \theta$$

the likelihood is given by

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

It is well known that

$$p(y|\theta) \propto \exp\left\{ -\frac{n}{2\sigma^2}(\bar{y} - \theta)^2 \right\}$$

where

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ is the sample mean}$$

Conjugate prior

The Gaussian distribution in exponential form is, for a single observation

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

the likelihood is then, letting $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

$$p(y|\theta) \propto e^{-n\frac{\theta^2}{2\sigma^2}} e^{\frac{\theta}{\sigma^2} n\bar{y}}$$

and the conjugate prior is

$$\pi(\theta) \propto g(\theta)^{\eta} e^{\phi^T \nu} = e^{-\eta\frac{\theta^2}{2\sigma^2}} e^{\frac{\theta}{\sigma^2}\nu} = \exp\left\{ -\frac{\eta}{2\sigma^2} \left(\theta^2 - 2\frac{\nu}{\eta}\theta \right) \right\}$$

that is, the conjugate family is the Gaussian family:

$$\theta \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$

Posterior distribution

$$\pi(\theta|y) \propto p(y|\theta)\pi(\theta) \propto \exp\left\{-\frac{n}{2\sigma^2}(\bar{y}-\theta)^2\right\} \exp\left\{-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right\}$$

$$\propto \exp\left\{-\frac{n}{2\sigma^2}\theta^2 - \frac{1}{2\sigma_0^2}\theta^2 + \frac{\theta\bar{y}n}{\sigma^2} + \frac{\theta\mu_0}{\sigma_0^2}\right\}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\theta^2 + \theta\left(\frac{n}{\sigma^2}\bar{y}+\frac{1}{\sigma_0^2}\mu_0\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}}\left(\theta^2 - 2\theta\frac{\frac{n}{\sigma^2}\bar{y}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)\right\}$$

$$\propto \exp\left\{-\frac{1}{2\left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1}}\left(\theta - \frac{\frac{n}{\sigma^2}\bar{y}+\frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}\right)^2\right\}$$

Posterior distribution (cont.)

$$\pi(\theta|y) \propto p(y|\theta)\pi(\theta)$$
$$\propto \exp\left\{-\frac{1}{2(\sigma_n)^2}(\theta-\theta_n)^2\right\} \qquad \left[\mathcal{N}\left(\theta_n,(\sigma_n)^2\right)\right]$$

that is, we obtain a gaussian posterior distribution with parameters θ_n and σ_n which are a function of prior distribution's parameters and of the data:

$$\theta_n = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}} = \frac{\mu_0\sigma^2 + \bar{y}n\sigma_0^2}{\sigma^2+n\sigma_0^2}$$

$$(\sigma_n)^2 = \left(\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}\right)^{-1} = \frac{\sigma^2\sigma_0^2}{\sigma^2+n\sigma_0^2}$$

Posterior distribution (cont.)

The posterior mean is a weighted average of the prior mean and of the ML estimate, where the weights are the reciprocal of the respective variances

$$\theta_n = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\frac{1}{V(\bar{y})}\bar{y} + \frac{1}{V(\theta)}\mu_0}{\frac{1}{V(\bar{y})} + \frac{1}{V(\theta)}}$$

- $\theta_n \underset{n\to\infty}{\longrightarrow} \bar{y}$ as n grows, the ML estimates weighs more
- $\theta_n \underset{\sigma_0\to 0}{\longrightarrow} \mu_0$ the more concentrated is the prior distribution, the more the prior mean weighs.

Posterior distribution (cont.)

It is interesting to write the posterior mean as

$$\theta_n = \mu_0 + (\bar{y} - \mu_0)\frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}$$

the posterior mean is the prior mean plus an adjustment toward the sample mean.

$$\theta_n = \bar{y} - (\bar{y} - \mu_0)\frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$$

the posterior mean is the sample mean shrunken toward the prior mean.

Posterior distribution (cont.)

The reciprocal of the posterior variance is the sum of the reciprocals of the prior variance and the variance of ML estimator

$$(\sigma_n)^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} = \left(\frac{1}{V(\bar{y})} + \frac{1}{V(\theta)}\right)^{-1}$$

- $\sigma_n \underset{n\to\infty}{\longrightarrow} 0$ as n grows the variance of the posterior diminish (it is more concentrated, where?).
- $\sigma_n \underset{\sigma_0\to 0}{\longrightarrow} 0$ also if the variance of the prior is reduced the posterior is more concentrated, where?

Model for gaussian data: predictive distribution

Consider a new observation $\tilde{y}$, then

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$
$$\propto \int \exp\left\{-\frac{1}{2\sigma^2}(\tilde{y}-\theta)^2\right\} \exp\left\{-\frac{1}{2\sigma_n^2}(\theta-\mu_n)^2\right\} d\theta$$

or, in other words,

- $\tilde{y}|\theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$
- $\theta \sim \mathcal{N}\left(\mu_n, \sigma_n^2\right)$

Then (see here)

$$\tilde{y}|y \sim \mathcal{N}\left(\mu_n, \sigma_n^2 + \sigma^2\right)$$

The uncertainty in the predictive distribution is the uncertainty from the model (σ^2) plus the uncertainty "on" the model (σ_n^2).

Model for gaussian data: predictive distribution mean and variance

Note that the mean of the predictive distribution for $\tilde{y}$ can be derived considering that

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

While the variance is derived from

$$\begin{aligned} V(\tilde{y}|y) &= E(V(\tilde{y}|\theta, y)|y) + V(E(\tilde{y}|\theta, y)|y) \\ &= E(\sigma^2|y) + V(\theta|y) \\ &= \sigma^2 + \sigma_n^2 \end{aligned}$$

which makes the interpretation of the decomposition more transparent.

Note: updates

- Let, a priori, $\theta \sim \mathcal{N}(\mu_0, \sigma_0^2)$.
- given an observation $y_1 \sim \mathcal{N}\left(\theta, \sigma^2\right)$ the posterior is

$$(\theta|y_1) \sim \mathcal{N}\left(\mu_1 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{y_1}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}, \sigma_1^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}\right)$$

- A second observation $y_2 \sim \mathcal{N}\left(\theta, \sigma^2\right)$ becomes available, we can update the posterior using $(\theta|y_1)$ as the prior distribution

$$\pi(\theta|y_1, y_2) \propto \pi(\theta|y_1) f(y_2|\theta)$$

so

$$(\theta|y_1, y_2) \sim \mathcal{N}\left(\mu_2 = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{y_2}{\sigma^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma^2}}, \sigma_2^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma^2}}\right)$$

Note: updates (cont.)

we note that

$$\sigma_2^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma^2}} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} + \frac{1}{\sigma^2}} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{2}{\sigma^2}}$$

$$\mu_2 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{y_1 + y_2}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{2}{\sigma^2}}$$

- That is, the same results is obtained either updating the information in two steps as above and starting from the prior $\mathcal{N}\left(\mu_0, \sigma_0^2\right)$ and updating it using the likelihood of the pairs (y_1, y_2).

Note: sequential updates

Let $p(y|\theta)$ be the model and $\pi(\theta)$ the prior, the posterior is then

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

If a further observation y^*, independent of y and distributed according to $p(y|\theta)$, becomes available, the posterior

$$\pi(\theta|y, x) \propto \pi(\theta)p(y^*|\theta, y)$$

is obtained, being y^* and y independent we can write

$$\begin{aligned}\pi(\theta|y, x) &\propto \pi(\theta)p(y^*|\theta)p(y|\theta) \\ &\propto \pi(\theta|y)p(y^*|\theta)\end{aligned}$$

which is also obtained combining the prior distribution $\pi(\theta|y)$ and the likelihood for y^*.

Note 2: sufficient statistics

Note that, given the prior distribution $\theta \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ the same posterior distribution is obtained with the following observations

- n observations $y_1, \ldots, y_n$ IID from $\mathcal{N}\left(\theta, \sigma^2\right)$;
- 1 observation $\bar{y}$ from a $\mathcal{N}\left(\theta, \sigma^2/n\right)$.

where the second is nothing but the sufficient statistics for the sample $y_1, \ldots, y_n$.

This is intuitive, the posterior depends on the sample only through the likelihood, and the likelihood of the sufficient statistics is equal to the one of the entire sample.

Note 2: sufficient statistics

We can substitute the sample y with any sufficient statistics $t(y)$, we will obtain the same posterior distribution.
If $t(y)$ is sufficient, then the factorization theorem tells us that

$$p(y|\theta) = h(y)g(t(y);\theta)$$

hence

$$\begin{aligned}\pi(\theta|y) &\propto \pi(\theta)p(y|\theta)\\ &\propto \pi(\theta)h(y)g(t(y);\theta)\\ &\propto \pi(\theta)g(t(y);\theta)\\ &\propto \pi(\theta|t(y))\end{aligned}$$

Indice

Likelihood

Although not a realistic situation, this is relevant both as

- an example of inference for a scale parameter
- a building block for the model on Gaussian data with both the mean and the variance unknown

Let then θ be known and

$$y_1, \ldots, y_n \sim IID\left(\mathcal{N}\left(\theta, \sigma^2\right)\right) \text{ conditional on } \sigma^2$$

so the likelihood is

$$\begin{aligned} p(y|\sigma^2) &\propto \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(y_i - \theta)^2 \right\} \right) \\ &\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{n}{2\sigma^2} \sum_{i=1}^{n} (y_i - \theta)^2 \right\} \end{aligned}$$

Prior

Assume an inverse gamma prior on σ^2,

$$\sigma^2 \sim \text{invGamma}(\gamma, \delta)$$

$$\pi(\sigma^2) \propto (\sigma^2)^{-(\gamma+1)} e^{-\delta/\sigma^2}$$

which is the same as saying that

$$\frac{1}{\sigma^2} \sim \text{Gamma}(\gamma, \delta)$$

Posterior

The posterior distribution is then

$$\pi(\sigma^2|y) \propto p(y|\sigma^2)\pi(\sigma^2)$$

$$\pi(\sigma^2|y) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\theta)^2\right\}(\sigma^2)^{-\gamma-1}e^{-\delta/\sigma^2}$$

$$\propto (\sigma^2)^{-n/2-\gamma-1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2+\delta\right]\right\}$$

that is,

$$\sigma^2|y \sim \text{inv-Gamma}\left(\gamma+n/2, \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2+\delta\right)$$

Reparametrization 1

It is convenient to reparametrize the model with the precision $\tau = 1/\sigma^2$, so the prior assumption is

$$\tau \sim \text{Gamma}(\gamma, \delta),$$

the likelihood is

$$p(y|\tau) \propto (\tau)^{n/2} \exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}(y_i - \theta)^2\right\}$$

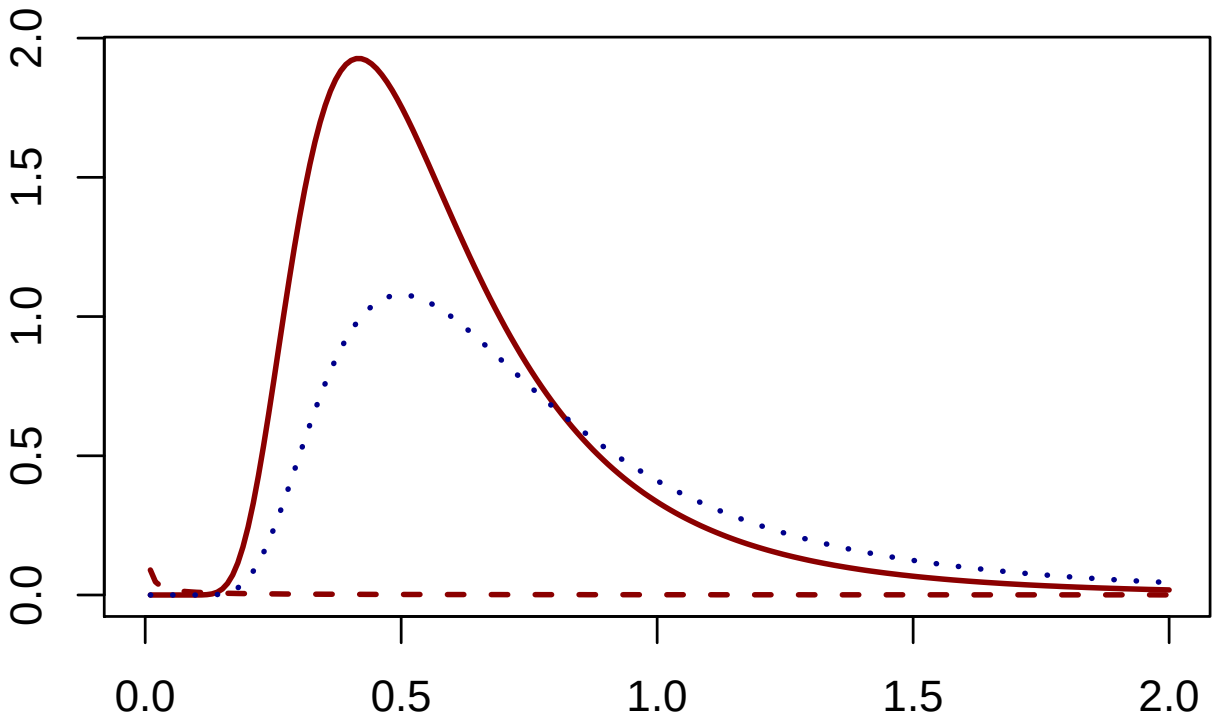
and the posterior is $\text{Gamma}(n/2 + \gamma, \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2 + \delta)$

$$\pi(\tau|y) \propto \tau^{n/2} \exp\left\{-\frac{1}{2}\tau\sum_{i=1}^{n}(y_i - \theta)^2\right\} \tau^{\gamma-1} e^{-\delta\tau}$$

$$\propto \tau^{n/2+\gamma-1} \exp\left\{-\tau\left[\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta)^2 + \delta\right]\right\}$$

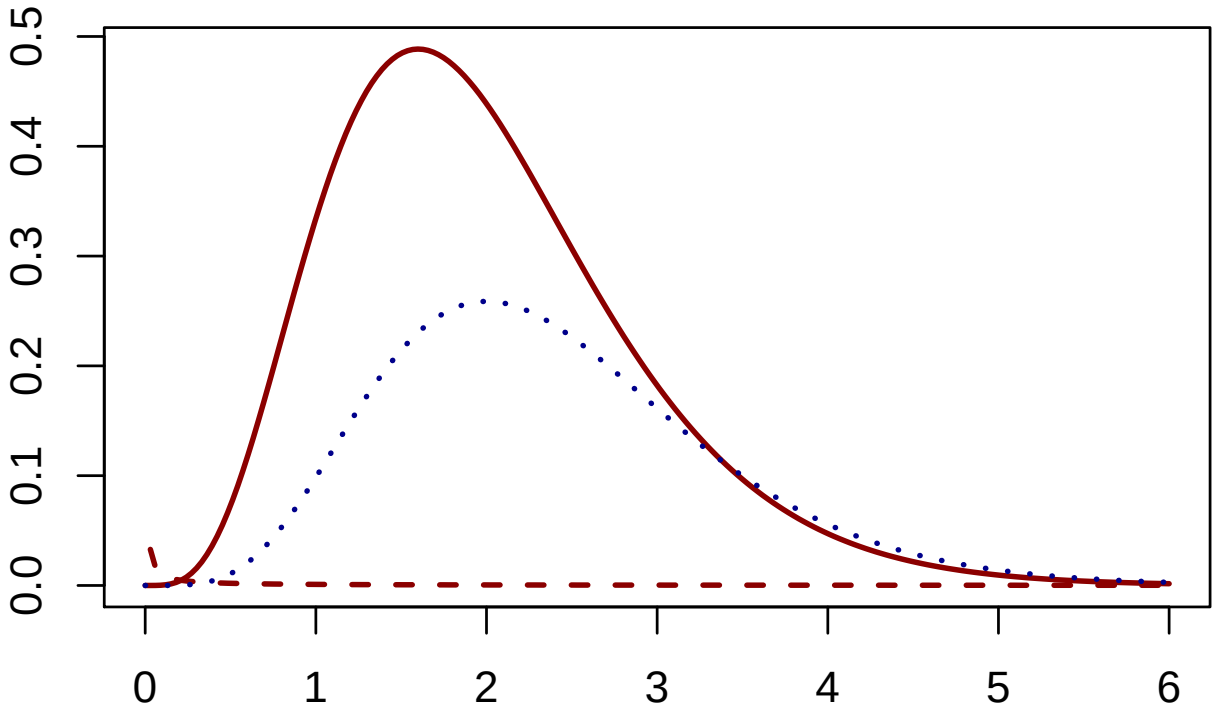
Inference for σ^2

Prior is an inverse gamma with parameters $\gamma = \delta = 10^{-3}$, sample variance is 0.5, $n = 10$



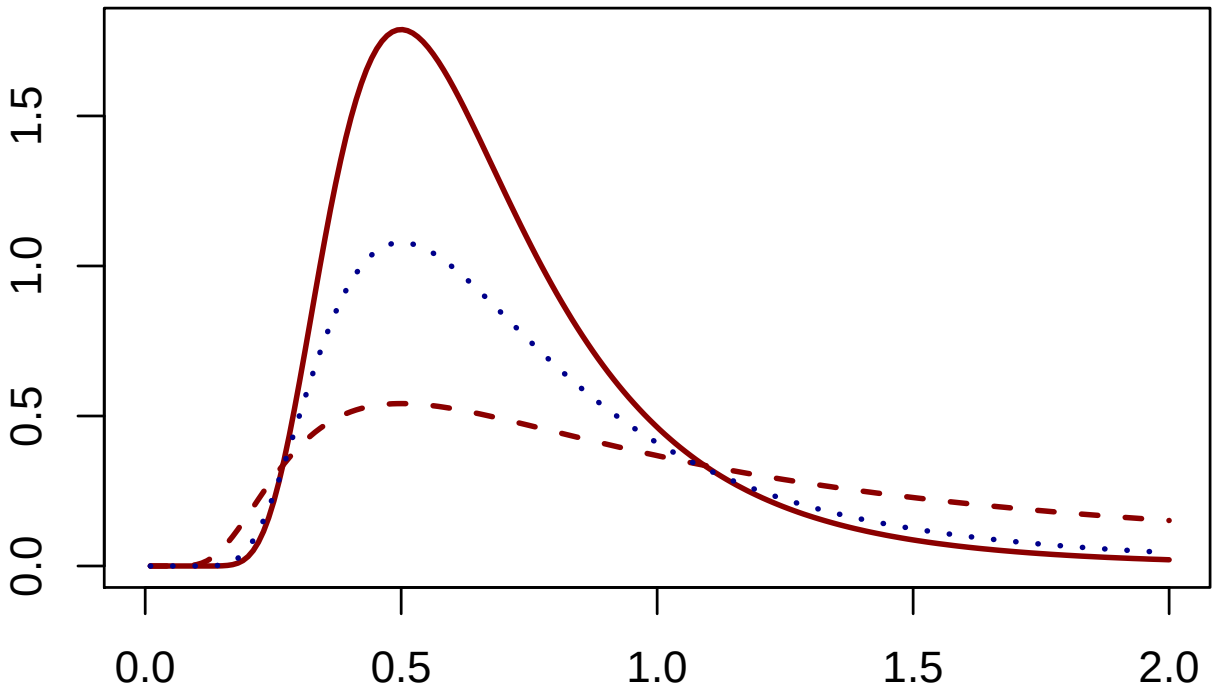
Inference for $\tau = 1/\sigma^2$

Prior is a gamma with parameters $\gamma = \delta = 10^{-3}$, sample variance is 0.5, $n = 10$



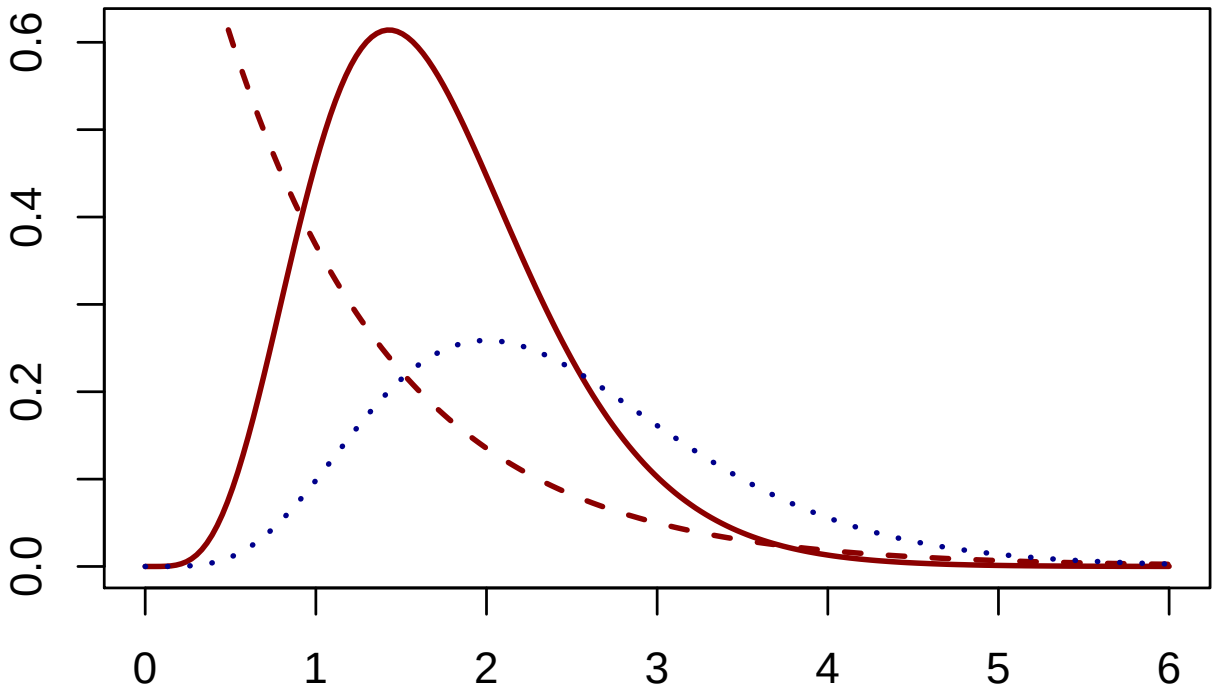
Inference for σ^2

Prior is an inverse gamma with parameters $\gamma = \delta = 1$, sample variance is 0.5, $n = 10$



Inference for $\tau = 1/\sigma^2$

Prior is a gamma with parameters $\gamma = \delta = 1$, sample variance is 0.5, $n = 10$



Reparametrization 2

Another convenient parametrization is to write

$$\sigma^2 =_d \frac{\sigma_0^2 \nu_0}{X}, \quad X \sim \chi^2_{\nu_0}$$

following Gelman we call this inv-$\chi^2(\nu_0, \sigma_0^2)$.
(This corresponds to $\nu_0 = 2\gamma$ and $\sigma_0^2 = \delta/\gamma$.)

The posterior is then

$$\sigma^2|y \sim \text{inv-}\chi^2\left(\nu_0 + n, \frac{\nu_0\sigma_0^2 + n\hat{\sigma}^2_{MLE}}{\nu_0 + n}\right)$$

the scale parameter being a weighted average of the prior variance σ_0^2 and the MLE with weight given by ν_0 and n.

Indice

Count data: Poisson

Assume that $y_i|\theta \sim \text{Poisson}(\theta)$, that is

$$p(y_i|\theta) = \frac{\theta^{y_i} e^{-\theta}}{y_i!}$$

then if $y = (y_1, \ldots, y_n)$ are observed and these are iid conditionally on θ, then, if we let

$$t(y) = \sum_{i=1}^{n} y_i$$

the likelihood is

$$\begin{aligned} p(y|\theta) &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &\propto \theta^{t(y)} e^{-n\theta} \\ &= (e^{-\theta})^n e^{t(y)\log\theta} \end{aligned}$$

Posterior for Poisson data

The likelihood

$$p(y|\theta) \propto (e^{-\theta})^n e^{t(y)\log\theta}$$

belongs to an exponential family with natural parameter $\phi(\theta) = \log\theta$, the conjugate prior is a Gamma distribution, let

$$\pi(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta}$$

then the posterior is

$$\theta|y \sim \text{Gamma}(\alpha + n\bar{y}, \beta + n)$$

Distribution of y

Note that

$$p(y)\pi(\theta|y) = p(y|\theta)\pi(\theta) \Rightarrow p(y) = \frac{p(y|\theta)\pi(\theta)}{\pi(\theta|y)}$$

which in the Poisson case means, for a single observation y

$$\begin{aligned} p(y) &= \frac{\text{Poisson}(\theta)\text{Gamma}(\alpha,\beta)}{\text{Gamma}(\alpha+y, 1+\beta)} \\ &= \frac{\Gamma(\alpha+y)\beta^{\alpha}}{\Gamma(\alpha)y!(1+\beta)^{\alpha+y}} \\ &= \binom{\alpha+y-1}{y}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y} \end{aligned}$$

which is a Negative binomial

Extension of the Poisson model: exposure

Assume data y_i are such that

$$y_i|\theta \sim \text{Poisson}(x_i\theta)$$

where $x_i > 0$, which is observed, is the exposure of unit i.
The likelihood is

$$p(y|\theta) \propto \theta^{\sum_i y_i} e^{-\theta \sum_i x_i}$$

if the prior is again a Gamma(α, β) then the posterior is

$$\theta|y \sim \text{Gamma}(\alpha + \sum_i y_i, \beta + \sum_i x_i)$$

Poisson example: inference for an incidence rate

In a US city with a population of 200 000 we observed 3 deaths of asthma in a year, that is an observed mortality rate for asthma of 1.5 per 100 000.

We want to combine this datum with prior information on asthma mortality and we use the Bayesian paradigm.

We assume a Poisson model (typical in the epidemiology context), the likelihood is

$$y|\theta \sim \text{Poisson}(2\theta)$$

where θ is the asthma mortality rate in cases per 100 000.

Poisson example: prior elicitation

We want to choose a prior within the conjugate family, so

$$\pi(\theta) = \text{Gamma}(\alpha, \beta)$$

In order to assess appropriate hyperparameters α and β we use the fact that according to epidemiological literature

(1) rates above 1.5 per 100 000 are rare

(2) typical mortality rate is around 0.6 per 100 000

Fact (2) suggests

$$E(\theta) = \frac{\alpha}{\beta} = 0.6$$

Fact (1) suggests that we should keep $P(\theta < 1.5)$, for example

$$\alpha = 3; \quad \beta = 5$$

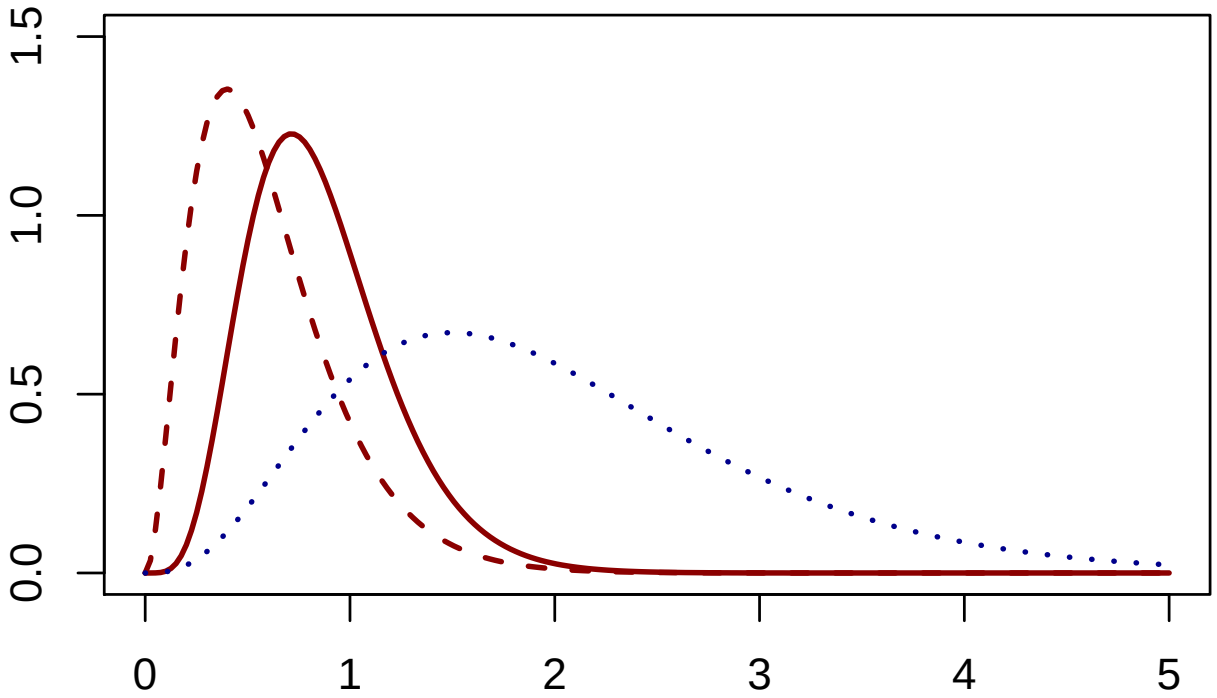
leads to $P(\theta < 1.44) = 0.975$.

Poisson example: prior elicitation

Starting from a prior $\pi(\theta) = \text{Gamma}(3, 5)$ the observation $y = 3$ with exposure $x = 2$ leads to the posterior

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

whose mean is 0.86.

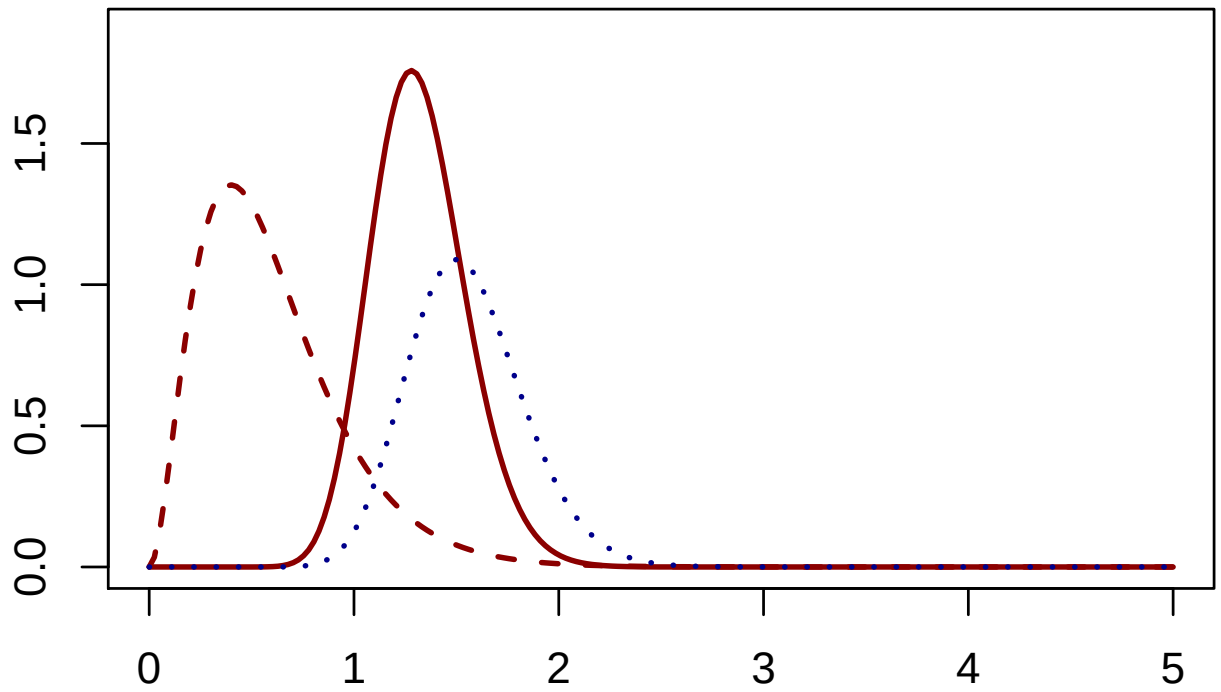


Poisson example: additional observations

Suppose now that the rate 1.5 per 100 000 is observed for 10 years (same city), so $y = 30$ and $x = 20$ and the posterior is

$$\theta|y \sim \text{Gamma}(33, 25)$$

whose mean is 1.32.



Indice

Prior distribution

The prior distribution is a novelty in Bayesian statistics with respect to classical statistics.
It is

- an opportunity, since we can formally include information other than observations in inference
- a problem, since we must include in inference informations which do not come from the experiment (observations).

From what we already discussed we know that

- Attitude toward subjective priors are the most various, from essential to unacceptable.
- (Reasonably specified) prior information vanishes as the sample size tends to infinity. This helps but is not a panacea, we have finite samples, so in practice our inference will be affected by the prior.

Sensitivity of results to prior's choice

The following example, due to Berger, shows that the same experimental result may lead to different conclusions depending on the prior distribution.

$$Y_1, \ldots, Y_n \ \ IID(\mathcal{N}(\theta, 1)) \text{ hence } L(\theta) \propto \exp\left\{-\tfrac{1}{2}(\bar{y} - \theta)^2\right\}$$

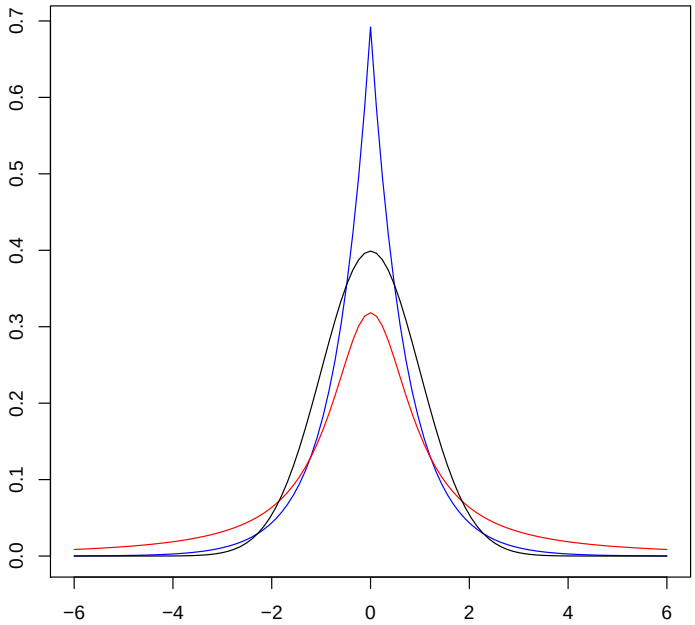
Let us fix the quartiles of the prior distribution:

$$IQ = -1 \ \ ; \quad Me = 0 \ \ ; \quad IIIQ = 1$$

There are infinite probability distribution coherent with the above values, let us consider

- Gaussian: $\theta \sim \mathcal{N}(0, 2.19)$
- Laplace: $\theta \sim La(1.384)$
- Cauchy: $\theta \sim Ca(0, 1)$

Sensitivity of results to prior's choice (cont.)



Gaussian

$$\pi(\theta) = \frac{1}{\sqrt{2\pi}\tau} \exp\left\{-\frac{1}{2\tau^2}\theta^2\right\}$$

Laplace

$$\pi(\theta) = \frac{\lambda}{2} \exp\left\{-\lambda|\theta|\right\}$$

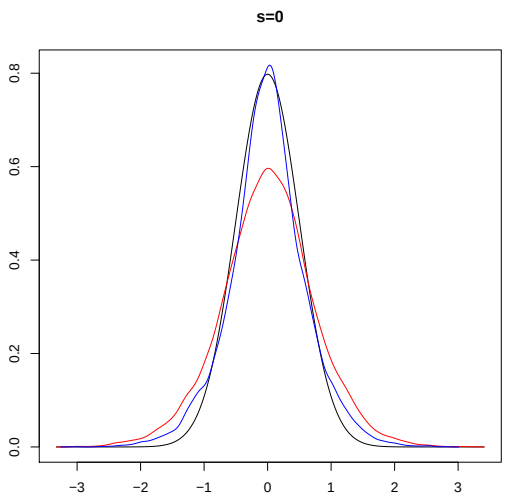
Cauchy

$$\pi(\theta) = \frac{1}{\pi(1+\theta^2)}$$

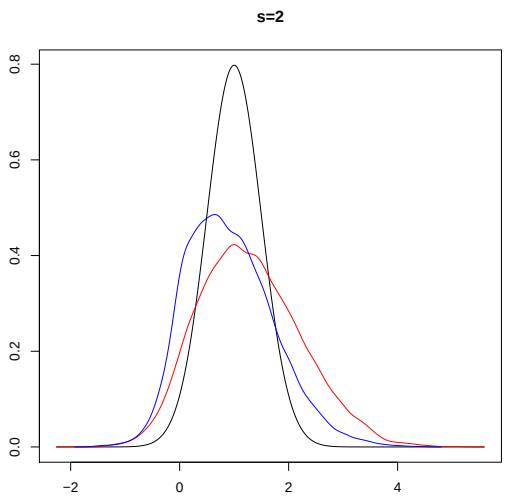
Sensitivity of results to prior's choice (cont.)

Assume $n = 1$, consider three different samples

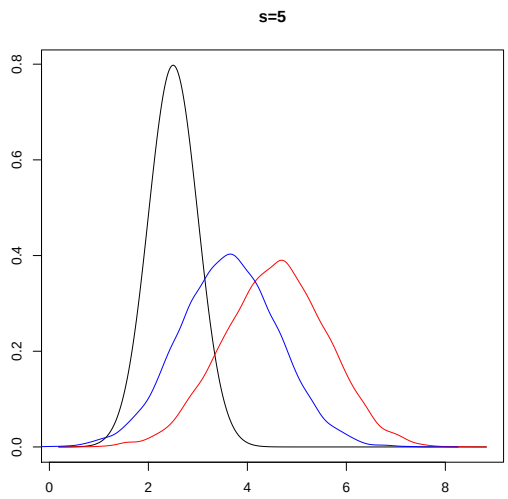
$y = 0$



$y = 2$



$y = 5$



Caution: relatively similar prior could lead to different posterior.

How do we choose a prior?

Any probability distribution (and not only) can be a prior for $\theta \in \Theta$.

A reasonable requirement is that $\text{supp}(\pi(\theta)) = \Theta$ (note that the support of the posterior distribution is, whatever the likelihood, a subset of the support of the prior distribution).

Typical choices are

- conjugate distributions;
- non informative (reference) priors
  - uniform prior
  - Jeffreys prior
  - improper prior
- weakly informative distributions

Conjugate priors: pros and cons

A family of distributions $f(\theta;\nu)$ is a natural conjugate for the likelihood $L(\theta)$ if, assuming $\pi(\theta) = f(\theta;\nu)$ the posterior distribution is in the same family, that is $\pi(\theta|y) = f(\theta;\nu_1)$ for some ν_1.

+ the main advantage is that solutions are available in closed form and are easily obtained;
- restricting the choice to the conjugate family may be too restrictive;
• conjugate families are less relevant today due to the use of MCMC and similar method to explore posterior distribution (closed forms are not needed anymore).

Conjugate priors examples

- Beta + Binomial;
- Gaussian + Gaussian (for the mean, variance known);
- Gamma + Poisson.

Non informative (reference) priors

Abandon the idea that the prior distribution is meant to reflect the opinion of the researcher prior to observing any data.

Rather, we want to model the absence of any opinion (whether this is realistic is disputable).

This is a relevant issue also as a possible answer to the objection which are put forward by those who do not like the results of inference to depend on subjective opinions: the rationale is to **let the data speak for themselves**.

These kind of priors have been called non informative or reference priors are sometimes associated to adjectives such as vague, flat or noninformative.

Problem is, it is not so obvious what "non informative" means.

Non informative priors: uniform

An intuitive solution is to assume

$$\pi(\theta) \propto k$$

so that no values of θ are privileged (principle of insufficient reason).

There are two difficulties

- What if the parameter space is not limited?
  - If the parameter space is not limited a constant has an infinite integral and so is not a probability distribution.
- Is it really non informative?

A uniform "distribution" for the mean

Consider the inference for the mean in a Gaussian sample starting from a prior $\theta \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$ the posterior is

$$\mathcal{N}\left(\frac{\mu_0\sigma^2 + \bar{y}n\sigma_0^2}{\sigma^2 + n\sigma_0^2}, \frac{\sigma^2\sigma_0^2}{\sigma^2 + n\sigma_0^2}\right)$$

if σ_0^2 is big relative to σ^2/n this is approximately

$$\mathcal{N}\left(\bar{y}, \frac{\sigma^2}{n}\right)$$

which is the same as we would obtain by assuming

$$\pi(\theta) \propto k$$

Improper prior

If we apply (the algebra of) Bayes theorem

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

with a function $\pi(\theta)$ which is not a valid probability distribution, then

- $\pi(\theta|y)$ is not necessarily a valid distribution (and if it is not then it is not useful)
- if $\pi(\theta|y)$ is a valid distribution then it is reasonable to interpret it as a posterior distribution

In practice: the uniform prior may work even if the parameter space is limited (on a case by case basis).

"Informativeness' of the uniform distribution

The non informative nature of the uniform distribution in general is disputable.

Let

$$\pi(\theta) \propto k$$

consider the reparametrization $\psi = \psi(\theta)$, then

$$\pi(\psi) = \pi(\theta^{-1}(\psi)) \left| \frac{d\theta}{d\psi} \right|$$

which is not uniform in general.

That is, assuming that uniform means non informative, by specifying a uniform distribution for the parameter θ, we are specifying an informative prior on its transform $\psi = \psi(\theta)$.

Jeffreys' prior

The above issue may be overcome by posing

$$\pi(\theta) = \sqrt{\mathcal{I}(\theta)}$$

where H is Fisher information, that is

$$[\mathcal{I}(\theta)] = -E\left(\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2}\right)$$

with this, for any parametrization $\psi = \psi(\theta)$

$$\pi(\psi) = \sqrt{\mathcal{I}(\psi)} = \sqrt{\mathcal{I}(\theta)}\left|\frac{d\theta}{d\psi}\right|$$

Jeffreys' prior: example

Consider a Binomial experiment, so the log-likelihood is

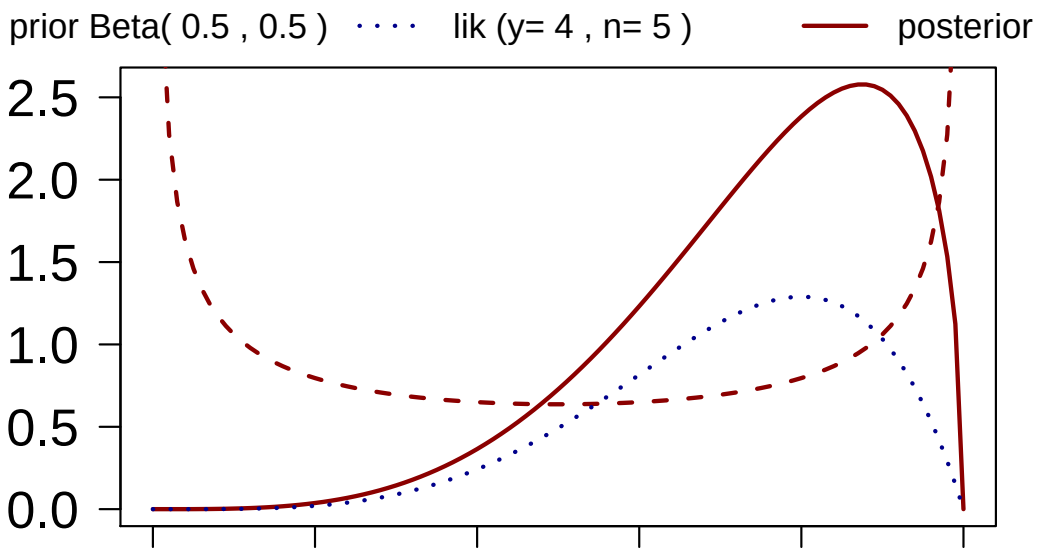
$$\log p(y|\theta) = y \log \theta + (n - y) \log(1 - \theta)$$

then

$$[\mathcal{I}(\theta)] = -E\left(\frac{d^2}{d\theta^2} \log p(y|\theta)\right) = \frac{n}{\theta(1-\theta)}$$

the prior is then a Beta$(1/2, 1/2)$

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



Binomial, reparametrization

Consider the reparametrization

$$\psi = \log\left(\frac{\theta}{1-\theta}\right) \in \mathbb{R}$$

If we assumed a uniform prior on θ then

$$\pi(\psi) = \pi(\theta^{-1}(\psi))\left|\frac{d\theta}{d\psi}\right| = \frac{e^{\psi}}{(1+e^{\psi})^2}$$

If the jJeffrey's prior is chosen then it implies

$$\pi(\psi) = \frac{e^{\psi/2}}{1+e^{\psi}}$$

which is also equal to

$$\sqrt{\mathcal{I}(\psi)}$$

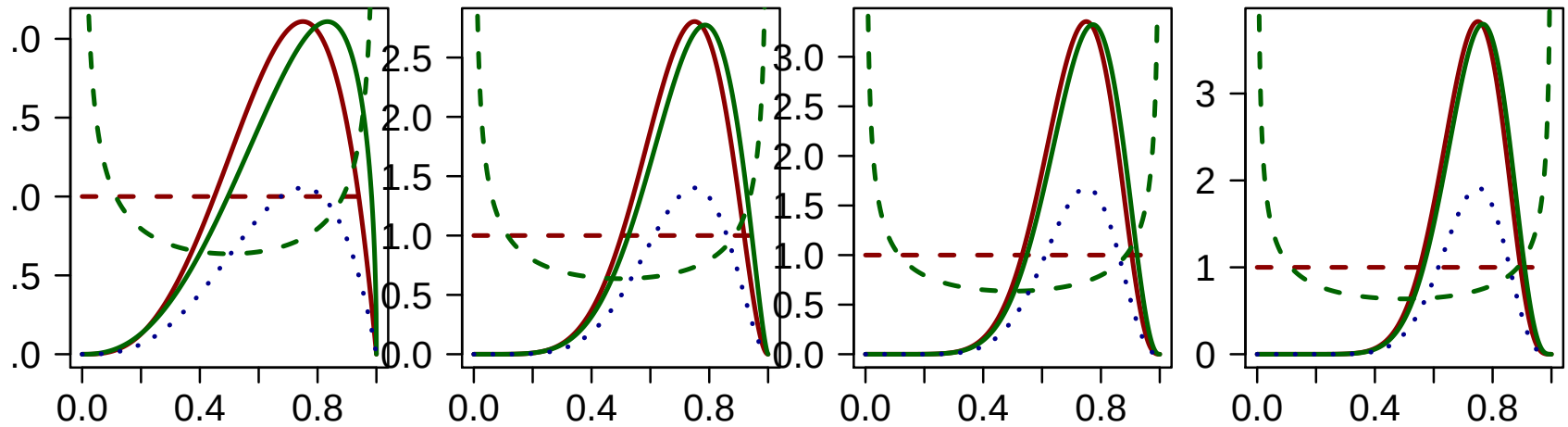
Binomial, reparametrization (cont.)

In fact

$$\begin{aligned}
\sqrt{\mathcal{I}(\psi)} &= \sqrt{-E\left(\frac{d^2 \log p(y|\psi)}{d\psi^2}\right)} \\
&= \sqrt{-E\left(\frac{d^2}{d\psi^2}\left(y\psi - \log(1+e^{\psi})\right)\right)} \\
&= \sqrt{E\left(\frac{e^{\psi}}{(1+e^{\psi})^2}\right)} \\
&= \sqrt{\frac{e^{\psi}}{(1+e^{\psi})^2}} \\
&= \frac{e^{\psi/2}}{1+e^{\psi}}
\end{aligned}$$

Sensitivity to the prior choice: Jeffrey's v. uniform

Samples imply $\hat{\theta} = 0.75$, $n = 4, 8, 12, 16$.



Sensitivity to the prior choice

Consider a Beta-Binomial model where 4 successes are observed on $n = 10$ trials, so that the ML estimate is 0.4, consider as a prior a $Beta(\alpha, \beta)$ with $\alpha = \beta$ (so $E(\theta) = 0.5$), compare below the effect of different choices on the posterior means and variances

	$\alpha + \beta$	$V(\theta)$	$E(\theta\|y)$	$V(\theta\|y)$
Jeffrey	1	0.1250	0.409	0.0201
Uniform	2	0.0833	0.417	0.0187
	5	0.0417	0.433	0.0153
	10	0.0227	0.450	0.0118
	20	0.0119	0.467	0.0080
	50	0.0049	0.483	0.0041
	100	0.0025	0.491	0.0023

Weakly informative prior

The rationale is that we usually do not really need to start from complete ignorance (which is what reference priors try to describe).

On the contrary there usually is some information

- for the probability of a female birth we are pretty sure it is not 0.1 or 0.9,

The idea is than to use a prior conveying less information than what we actually have

- for the probability of a female birth we may use $\pi(\theta) \sim \mathcal{N}\left(0.5, 0.1^2\right)$, or $\pi(\theta) \sim \text{Beta}(20, 20)$
- for the inference on the mean $\pi(\theta) \sim \mathcal{N}\left(0, A^2\right)$ with A large (where what large means depends on the problem)

Prior distribution, in brief

There are some situations in which it is sensible to put relevant information into the prior distribution (especially with few data).

In general, even if we had information, it may be deemed inconvenient to include it in the model (prior), possible reasons include

- difficulties to elicit the prior
- mathematical simplicity

In this case we have a number of options

- uniform / improper priors
- non informative priors (Jeffrey's priors)
- weakly informative priors (possibly conjugate)

These are all valid options none of which is clearly superior, in fact, if we have enough data to rely exclusively on them, then the choice among relatively flat priors should not matter.

On the contrary it is advisable to avoid automatic use of a particular specification and do some sensitivity analysis.

Relation between prior and posterior

Note that $E(\theta) = E(E(\theta|y))$, and

$$V(\theta) = E(V(\theta|y)) + V(E(\theta|y))$$

that is, the posterior variance is, on average, smaller than the prior variance ($V(\theta) > E(V(\theta|y))$.
In particular it is smaller the greater is the variation of $E(\theta|y)$ across y.

Indice

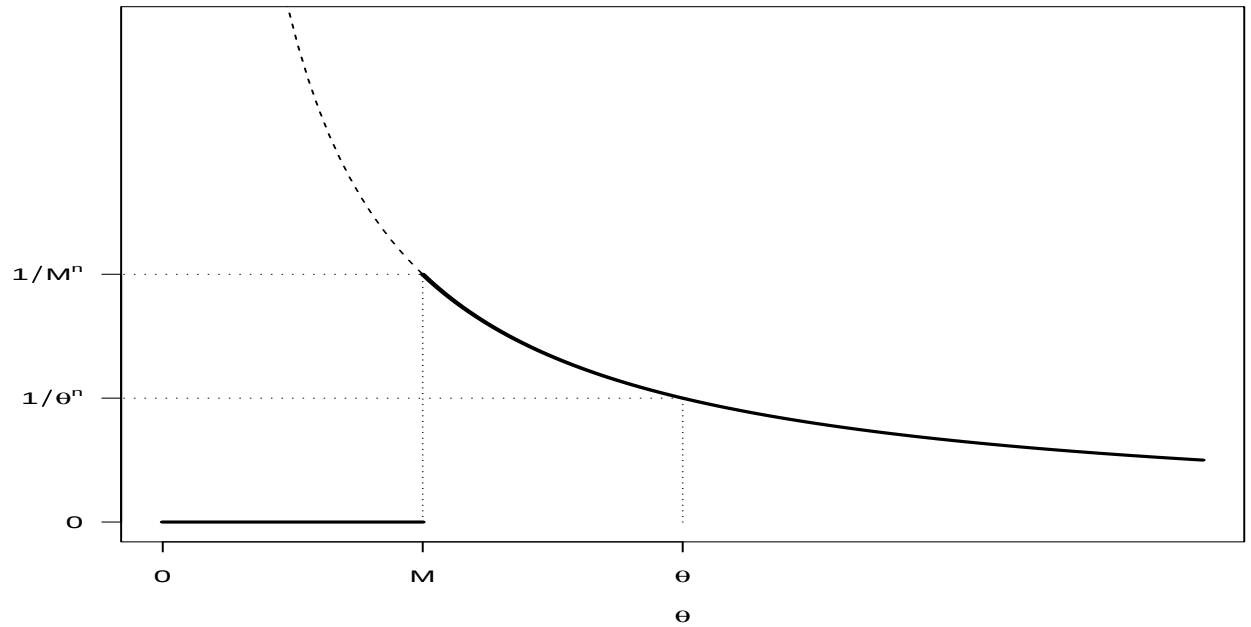
Model for uniform data: example

Let

$$y_1, \ldots, y_n | \theta \sim IID(\text{Unif}(\theta)), \quad \theta \in \Theta = \mathbb{R}^+$$

The likelihood is ($M = y_{(n)}$ in the plot)

$$p(y|\theta) = p_\theta(y) = \prod_{j=1}^{n} \frac{1}{\theta} \mathbb{I}(y_j \leq \theta) = \frac{1}{\theta^n} \mathbb{I}\left(y_{(n)} \leq \theta\right)$$



Model for uniform data: a priori, a posteriori

We consider the improper prior distribution

$$\pi(\theta) \propto \theta^{-\alpha}, \quad \alpha > 0$$

The posterior is then

$$\pi(\theta|y) = kp(y|\theta)\pi(\theta) = k\theta^{-n}\mathbb{I}\left(y_{(n)} \leq \theta\right)\theta^{-\alpha} = k\theta^{-(n+\alpha)}\mathbb{I}\left(y_{(n)} \leq \theta\right)$$

where

$$\frac{1}{k} = \int \theta^{-(n+\alpha)}\mathbb{I}\left(y_{(n)} \leq \theta\right) d\theta = \int_{y_{(n)}}^{+\infty} \theta^{-(n+\alpha)} d\theta = \frac{1}{y_{(n)}^{\alpha+n-1}(\alpha+n-1)}$$

hence

$$\begin{aligned}\pi(\theta|y) &= y_{(n)}^{\alpha+n-1}(\alpha+n-1)\theta^{-(n+\alpha)}\mathbb{I}\left(y_{(n)} \leq \theta\right) \\ &= y_{(n)}^{-1}(\alpha+n-1)\left(\frac{\theta}{y_{(n)}}\right)^{-(n+\alpha)}\mathbb{I}\left(y_{(n)} \leq \theta\right)\end{aligned}$$

Model for uniform data: Bayesian and classical estimates

In studying Bayesian statistics it is interesting to compare the Bayesian and classical estimates.

We compare in what follows the posterior expectation (a standard Bayesian estimator) with three natural estimators derived from the classic paradigm:

- MLE: $\hat{\theta}_{ML}$
- unbiased estimator derived from MLE: $\hat{\theta}_{ND}$
- efficient estimator: $\hat{\theta}_{EFF}$ best among linear functions of MLE

Model for uniform data: MLE

The likelihood is maximized when $\hat{\theta}_n = y_{(n)}$, whose distribution, for $y > 0$, is

$$P(\hat{\theta}_n \leq y) = P(y_{(n)} \leq y) = \bigcap_i P(y_i \leq y) = \prod_i \min\left(\frac{y_i}{\theta}, 1\right)$$

$$= \min\left(\left(\frac{y}{\theta}\right)^n, 1\right)$$

with density

$$f_{\hat{\theta}_n}(y) = \frac{n y^{n-1}}{\theta^n} \mathbb{I}(y < \theta)$$

and so $E(\hat{\theta}|\theta) = \frac{n\theta}{n+1}$

Model for uniform data: alternative estimators

- MLE is biased, $E(\hat{\theta}|\theta) = \frac{n\theta}{n+1}$, an unbiased estimator is obtained as

$$\hat{\theta}_{ND} = \frac{n+1}{n}\hat{\theta}_{ML} = \frac{n+1}{n}y_{(n)}$$

- furthermore, the best estimator (in terms of MSE) among those which can be written in the form $c\hat{\theta}_{ML}$ is

$$c^* = \frac{n+2}{n+1} = \text{argmin}E(c\hat{\theta}_{ML} - \theta)^2$$

and so $\hat{\theta}_{EFF} = \frac{n+2}{n+1}y_{(n)}$

Model for uniform data: Bayesian and classical estimates

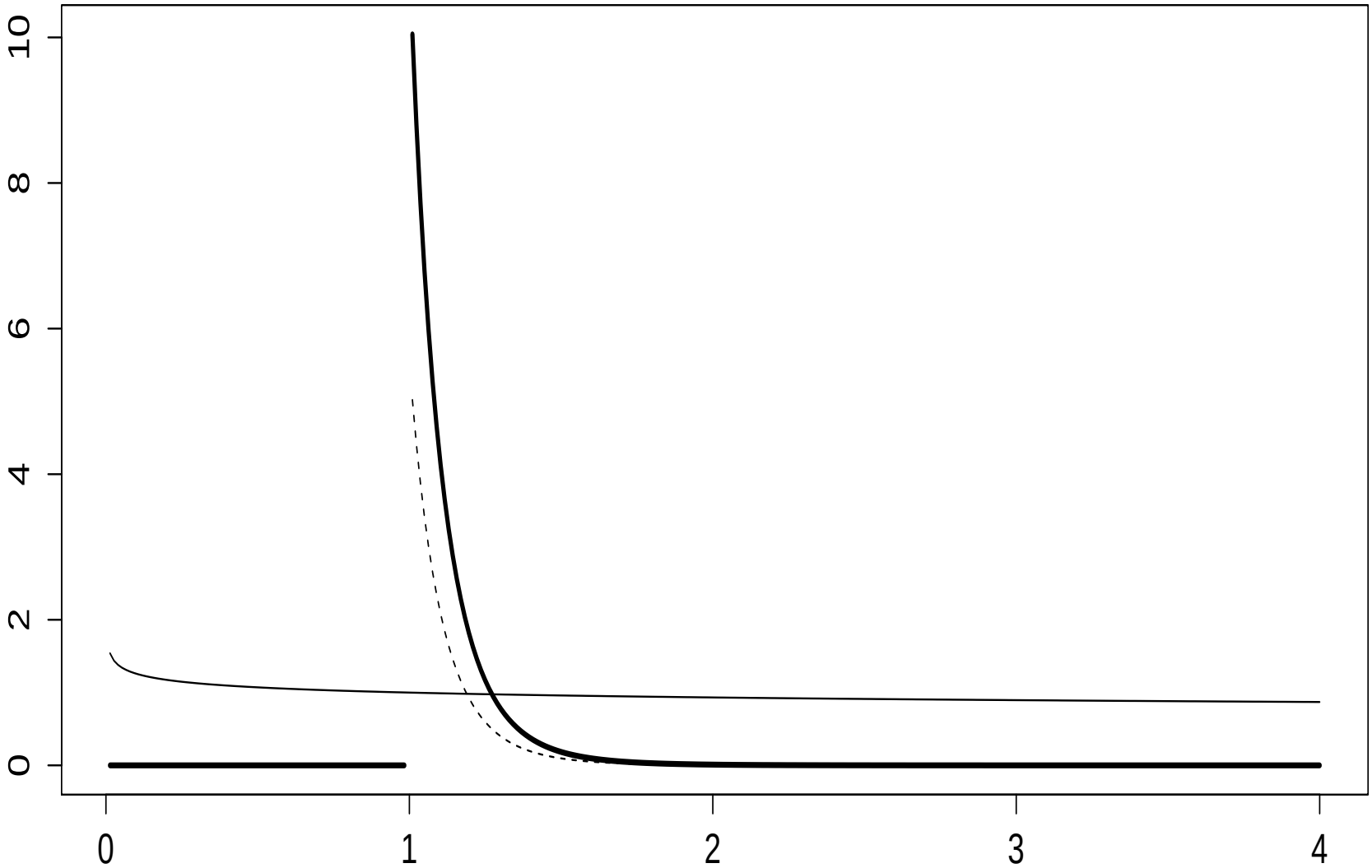
Posterior expectation

$$E(\theta|y) = \int_{y_{(n)}}^{+\infty} y_{(n)}^{\alpha+n-1}(\alpha+n-1)\theta^{-(n+\alpha)}\mathbb{I}\left(y_{(n)} \leq \theta\right) d\theta$$
$$= \frac{\alpha+n-1}{\alpha+n-2} y_{(n)}$$

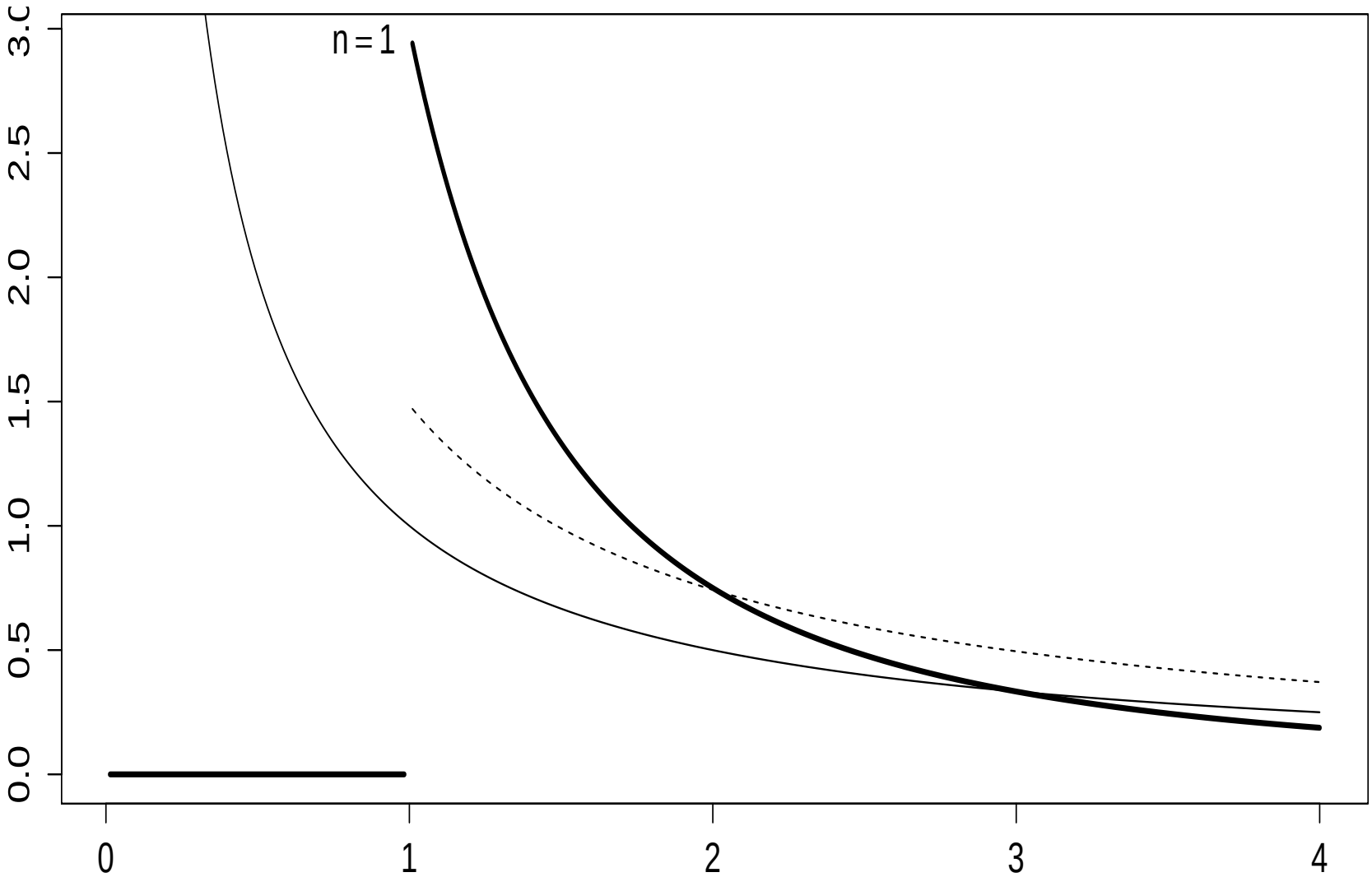
This is

- MLE if $\alpha \to \infty$
- unbiased estimator if $\alpha = 2$
- efficient estimator if $\alpha = 3$

Laplace example: a priori effect



Laplace example: a priori effect



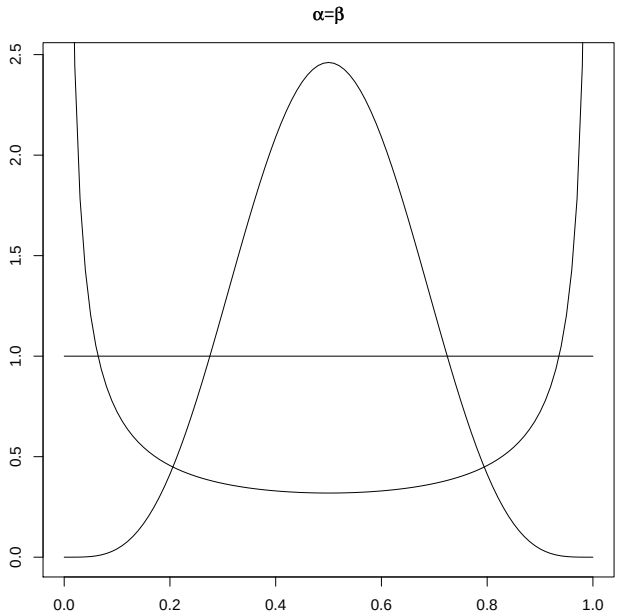
Indice

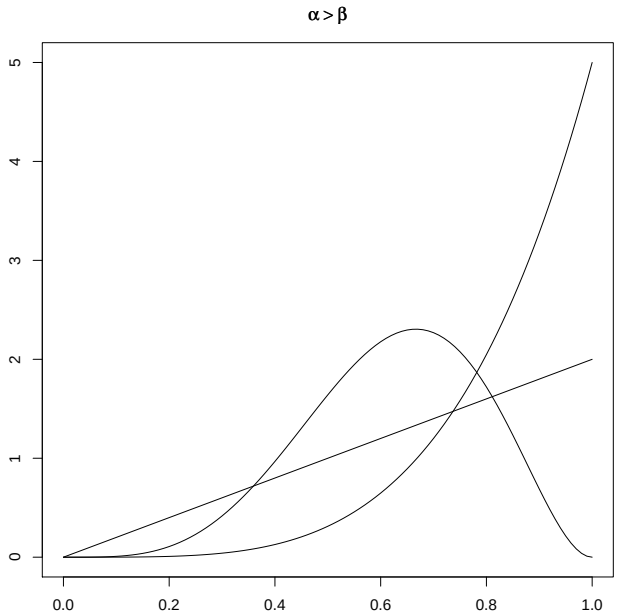
The Beta distribution

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

where $0 < \theta < 1$ e $\alpha, \beta > 0$,

$$E(\theta) = \frac{\alpha}{\alpha+\beta} \quad V(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$





back

Theorem: mixture of normals

Theorem

If $Y|\theta \sim \mathcal{N}\left(\theta, \sigma^2\right)$ and $\theta \sim \mathcal{N}\left(\mu, \tau^2\right)$ then

$$Y \sim \mathcal{N}\left(\mu, \sigma^2 + \tau^2\right).$$

This is easily seen, let

$$Z = Y - \theta|\theta; \text{ then } \quad Z \sim \mathcal{N}\left(0, \sigma^2\right) \quad \forall\theta$$

consider $X = Z + \theta$

- X is a sum of normal r.v. so it is normal,
- $E(X) = E(Z) + E(\theta) = 0 + \mu = \mu$
- $V(X) = V(Z) + V(\theta) + 2\text{Cov}(Z, \theta) = \sigma^2 + \tau^2 + 2\text{Cov}(Z, \theta) = \sigma^2 + \tau^2$
- since $\text{Cov}(Z, \theta) = E(Z\theta) = E(E(Z\theta|\theta)) = E(E((X - \theta)\theta|\theta)) = 0$

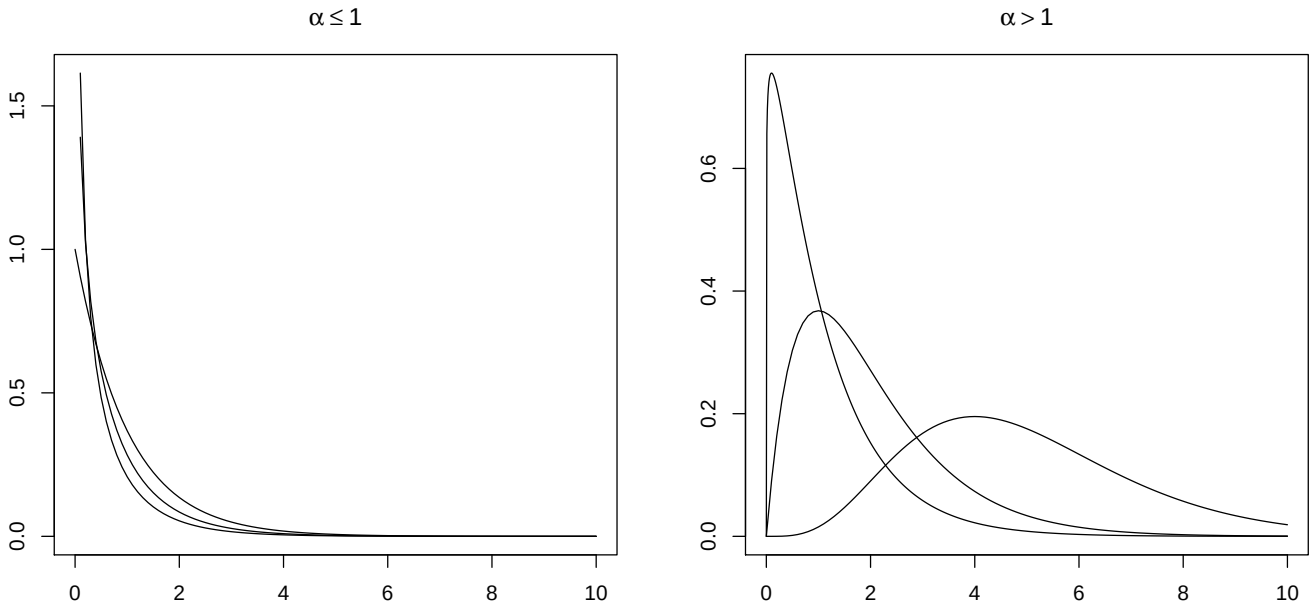
back

Gamma

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$$

where $\theta > 0$, $\alpha, \beta > 0$,

$$E(\theta) = \frac{\alpha}{\beta} \quad V(\theta) = \frac{\alpha}{\beta^2}$$

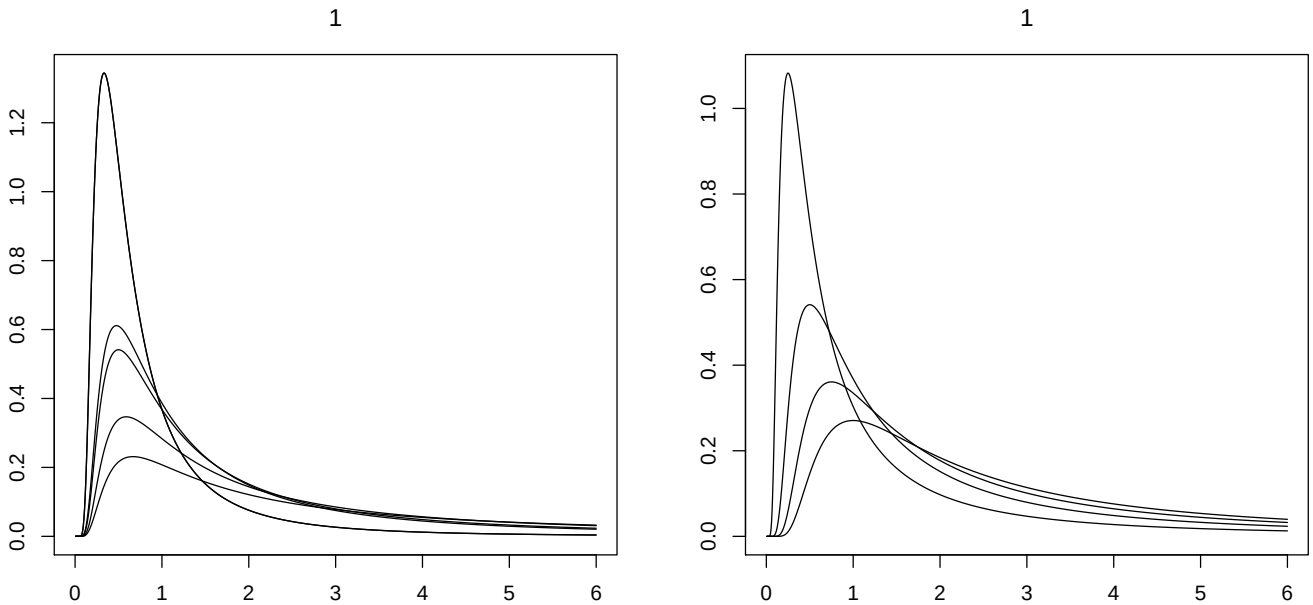


inverse Gamma

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{-(\alpha+1)}e^{-\beta/\theta}$$

where $\theta > 0$, $\alpha, \beta > 0$,

$$E(\theta) = \frac{\beta}{\alpha - 1} \quad V(\theta) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$



Definition of statistic

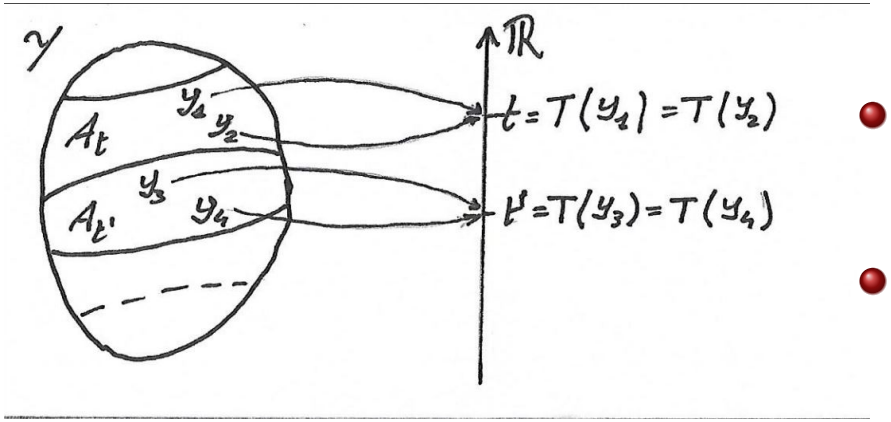
Definition: statistic

A **statistic** is a function of the data (not depending on the parameter)

$$T : \mathcal{Y} \to \mathcal{T} \subset \mathbb{R}^k$$

A statistic T is characterized by a partition of the sample space with elements

$$\mathcal{Y}_t = \{y \in \mathcal{Y} | T(y) = t\}, \quad t \in \mathcal{T}$$



- a bijective transformation of T has the same partition.
- a non bijective transformation of T is associated to a coarser partition.

Sufficient statistic

Generally speaking a statistic summarizes the sample, it is sufficient for a parameter θ if it does not summarize too much, i.e. once the statistic is known, knowing more of the sample would not add any information about θ

Definition: sufficient statistic

Within the model $(\mathcal{Y}, p_\theta, \Theta)$ a statistic $T(\cdot) : \mathcal{Y} \to \mathbb{R}^k$ is sufficient for θ if and only if

$$p(y|T = t) \text{ does not depend on } \theta$$

Characterizations of sufficiency

Theorem

Given a statistical model $(\mathcal{Y}, p_\theta, \Theta)$ and $T : \mathcal{Y} \to \mathbb{R}^k$ a statistic. The folowing statements are equivalent:

(i) $T(y) = T(z) \Rightarrow L(\theta, y) \propto L(\theta, z)$;

(ii) (Neyman factorization theorem)
there exist two function h e g such that
$p_\theta(y) = h(y)g(T(y), \theta)$;

(iii) T is sufficient
$p_\theta(y|T = t) = p(y|T = t) \ \ \forall\theta$;

Minimal suffciency

Definition: minimal sufficient statistic

For a model $(\mathcal{Y}, p_\theta, \Theta)$ a statistic $T(\cdot) : \mathcal{Y} \to \mathbb{R}^k$ is **minimal sufficient** for θ if and only if

$$T(y_1) = T(y_2) \Leftrightarrow L(\theta; y_1) \propto L(\theta; y_2)$$

per ogni $\theta \in \Theta$.

Equivalently

- Every sufficient statistic is a function of T.
- Its partition is the same as the likelihood partition.
- $\frac{L(\theta;y)}{L(\theta;z)} = c(y,z) \Leftrightarrow T(y) = T(z)$

back