One-parameter models

Normal model, Poisson model, and prior choice

Teacher: Matilde Trevisani

DEAMS

A.A. 2024/2025
(aggiornato: 2025-04-08)

Agenda (about 3 lectures)

One-parameter models

- General approach to Bayesian data modelling
- A first example
- Note on accumulation of evidence
- Binomial model
- Note on impact of more evidence
- Summarizing posterior distributions
- Conjugacy
- Interplay between priors and data
- Normal model
- Poisson model
- Other models
- Prior specification

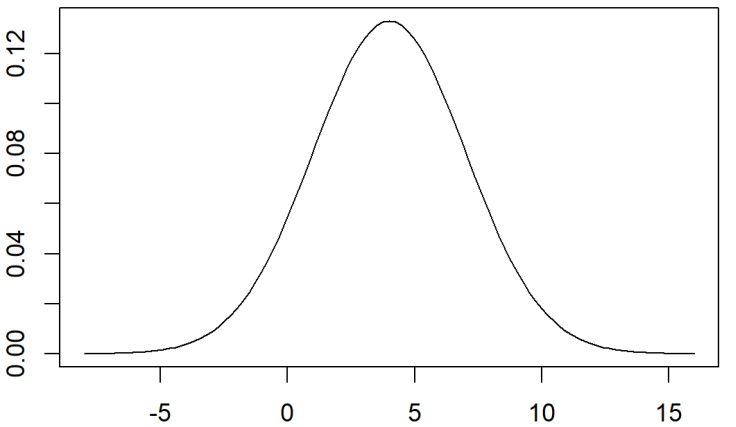
Normal Data

Normal / Gaussian

- Observations y with real values
- Mean θ and variance σ^2 (or standard deviation σ)

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

$$y \sim \mathrm{N}(\theta, \sigma^2)$$



Reasons to use the normal distribution

- Widely used in statistical modeling
- Fundamental because many advanced models and analysis paradigms are based on or involve normal distributions
- Normal distribution often justified by the central limit theorem
- Most often used for computational convenience or tradition

Central limit theorem

De Moivre, Laplace, Gauss, Chebysev, Liapounov, Markov, et al.

Under certain conditions the sum (and the mean) of random variables approaches the Gaussian distribution as $n \to \infty$

Problems

- does not hold for all distributions, e.g., Cauchy
- may require large n, e.g. with Binomial when θ approaches 0 or 1
- does not hold if one of the variables has much larger scale

Model for a normal distribution with unknown mean and known variance

Likelihood

Assume that observations $y = (y_1, \ldots, y_n)$ are normal (with a known variance σ^2) and exchangeable, i.e.,

$$y_1, \ldots, y_n | \theta \sim N(y_i | \theta, \sigma^2) \text{ iid}$$

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

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

Sufficient statistics

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

Data (y) influence through the sample mean $\bar{y}$

- The sample mean is a sufficient statistic for θ (μ)
- σ^2/n is the variance of the sample mean

The natural conjugate prior of an exponential distribution

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

$$p(y|\theta) \propto e^{-n\frac{\theta^2}{2\sigma^2}} e^{\frac{\theta}{\sigma^2} n\bar{y}} = g(\theta)^n e^{\phi(\theta)^T t(y)}$$

and the conjugate prior is

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

that is, the conjugate family is the Gaussian family:

$$\theta \sim N(\mu_0, \sigma_0^2)$$

Normal distribution with unknown mean θ
Conjugate prior for θ

Let assume σ^2 known, the Gaussian family is the conjugate family for the unknown mean of a Gaussian distribution

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

$$\text{Prior} \quad p(\theta) \propto \exp\left(-\frac{1}{2\sigma_0^2}(\theta-\mu_0)^2\right)$$

$$\exp(a)\exp(b) = \exp(a+b)$$

$$\text{Posterior} \quad p(\theta|y) \propto \exp\left(-\frac{1}{2}\left[\frac{(y-\theta)^2}{\sigma^2} + \frac{(\theta-\mu_0)^2}{\sigma_0^2}\right]\right)$$

$$\propto \exp\left(-\frac{1}{2\sigma_n^2}(\theta-\mu_n)^2\right)$$

Calcula

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

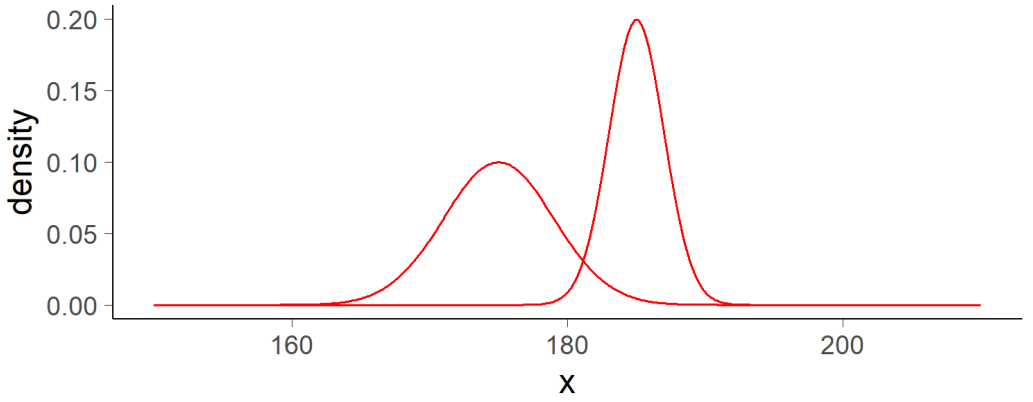
Posterior distribution (cont.)

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

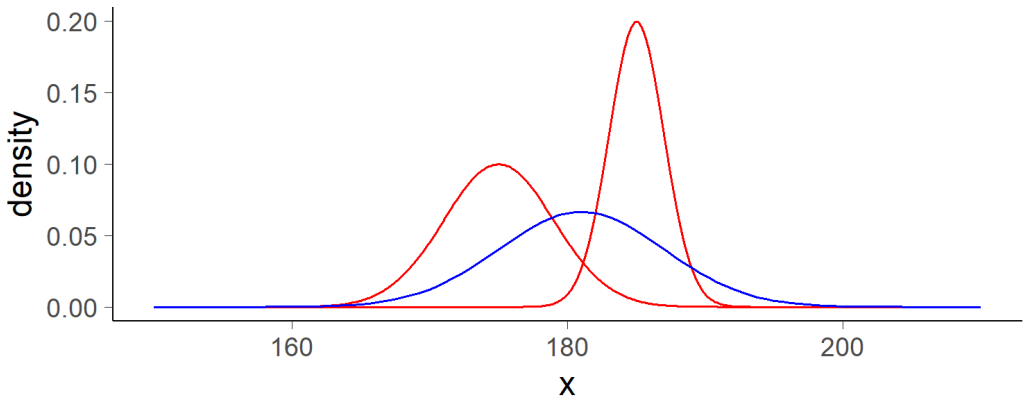
$$\theta|y \sim \mathrm{N}(\mu_n, \sigma_n^2), \quad \text{where} \quad \mu_n = \frac{\frac{1}{\sigma_0^2}\mu_0 + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}} \quad \text{and} \quad \frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}$$

- posterior precision = prior precision + data precision
  (precision=1/variance, data precision is precision of the ML estimator $\bar{y}$)
- posterior mean is a weighted mean with precisions as weights
- $\mu_n = \mu_0 + (\bar{y}-\mu_0)\frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}$ or $\mu_n = \bar{y} - (\bar{y}-\mu_0)\frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}$
- $\mu_n = \mu_0$ if $\bar{y} = \mu_0$ or $\sigma_0^2 = 0$; in limit: $\mu_n \underset{\sigma_0\to 0}{\longrightarrow} \mu_0$

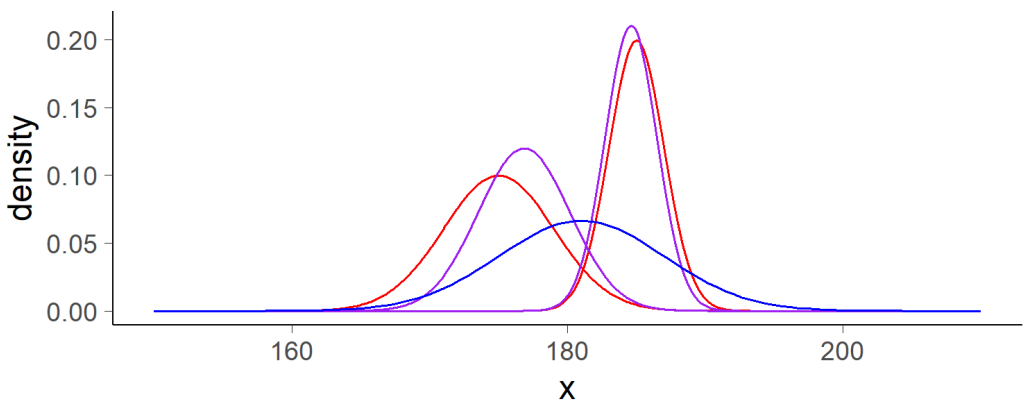
  $\mu_n = \bar{y}$ if $\bar{y} = \mu_0$ or $\sigma^2 = 0$; in limit $\mu_n \underset{n\to\infty}{\longrightarrow} \bar{y}$
- if $\sigma_0^2 = \sigma^2$, the prior corresponds to one virtual observation with value μ_0
- If $\sigma_0 \to \infty$ for a fixed n, or if $n \to \infty$ for fixed σ_0,
  $p(\theta|y) \approx \mathrm{N}(\theta|\bar{y}, \sigma^2/n)$ i.e. posterior approximates the likelihood
- Lastly, $\sigma_n \underset{n\to\infty}{\longrightarrow} 0$ as well as $\sigma_n \underset{\sigma_0\to 0}{\longrightarrow} 0$ (but what the posterior is more concentrated around in the 2 cases?)



(Example with n=1) Suppose you want to guess how tall two people (of *Xland*) are and you are less certain about one than the other $\mu_1 = 175, \mu_2 = 185, \sigma_1 = 4, \sigma_2 = 2$



Suppose we base the prior on the parameters of the male population in *Xland*. $\mu_{pop} = 181, \sigma_{pop} = 6$



The posterior combines prior and observation, if the observation is more precise the prior has less influence, viceversa...

Parameterization in terms of precision

- ***Precision*** $\tau = 1/\sigma^2$ as the inverse of the variance
- $x \mid \mu, \sigma^2 \sim N(\mu, \sigma^2) \rightarrow x \mid \mu, \tau \sim N(\mu, \tau)$
- Conditional probability of the data

$$x_i \mid \mu, \tau \sim N(\mu, \tau)$$

Precision of the data

- Prior distribution

$$\mu \mid \mu_\mu, \tau_\mu \sim N(\mu_\mu, \tau_\mu)$$

Precision of the prior

- Posterior distribution

$$\mu \mid \mathbf{x}, \tau \sim N(\mu_{\mu|\mathbf{x}}, \tau_{\mu|\mathbf{x}})$$

Precision of the posterior

Accumulation of precision in the posterior

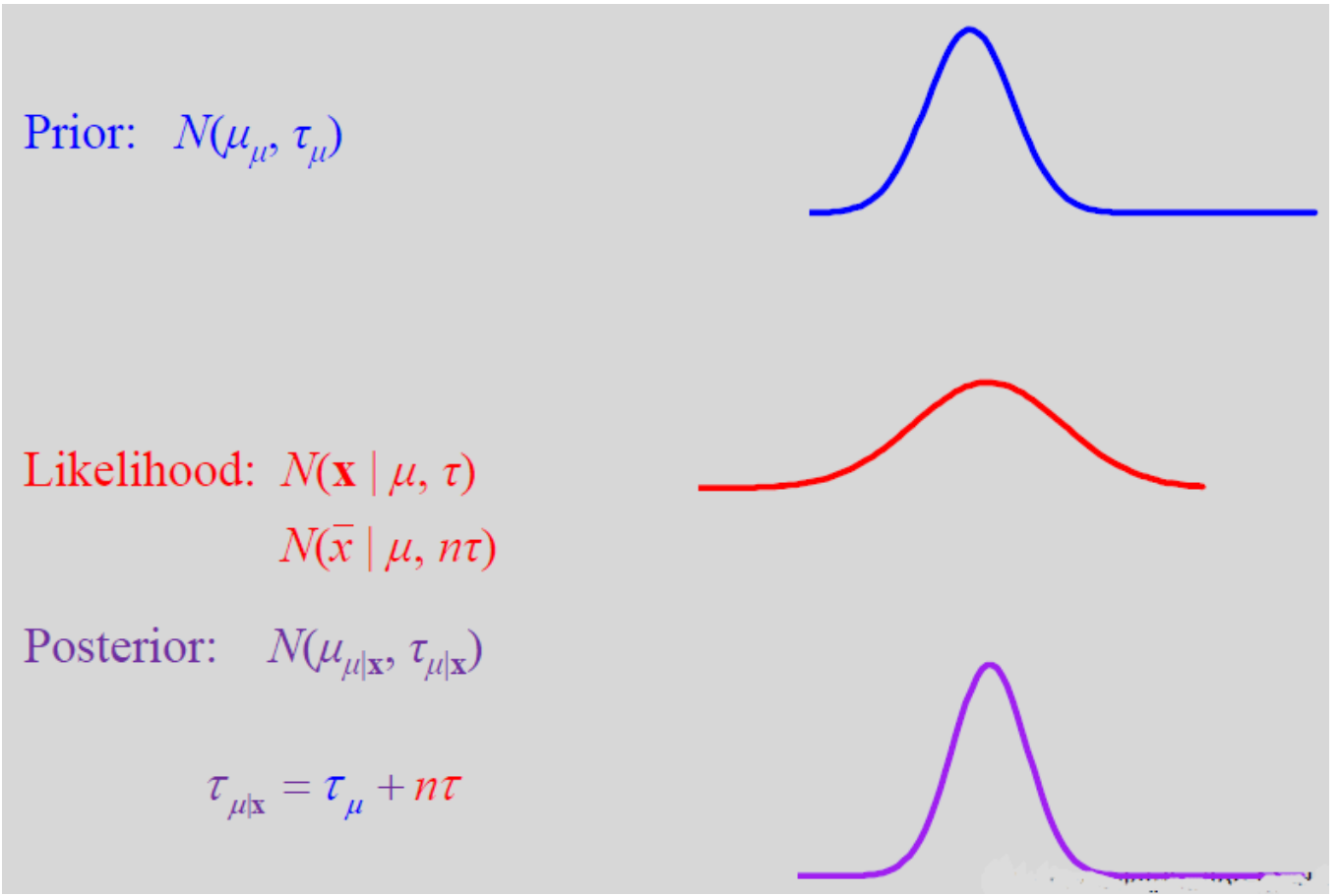
Prior: $N(\mu_\mu, \tau_\mu)$

Likelihood: $N(\mathbf{x} \mid \mu, \tau)$

$N(\bar{x} \mid \mu, n\tau)$

Posterior: $N(\mu_{\mu|\mathbf{x}}, \tau_{\mu|\mathbf{x}})$

$$\tau_{\mu|\mathbf{x}} = \tau_\mu + n\tau$$



Posterior mean

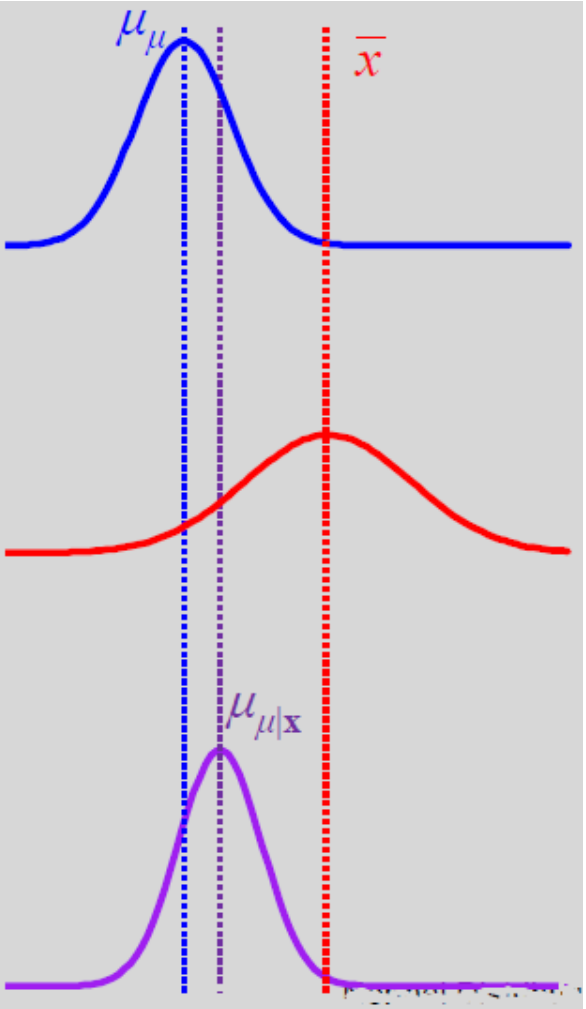
Prior: $N(\mu_\mu, \tau_\mu)$

Likelihood: $N(\mathbf{x} \mid \mu, \tau)$

$N(\bar{x} \mid \mu, n\tau)$

Posterior: $N(\mu_{\mu|\mathbf{x}}, \tau_{\mu|\mathbf{x}})$

$$\mu_{\mu|\mathbf{x}} = \frac{\tau_\mu}{\tau_\mu + n\tau}\mu_\mu + \frac{n\tau}{\tau_\mu + n\tau}\bar{x}$$



Shrinkage toward prior

- The posterior is a *synthesis* of the prior and the likelihood
- It can be seen as an **updating of the prior** in light of the data (as reflected in the likelihood)
- It can be seen as an **augmentation of the information in the data** (as reflected in the likelihood) thanks to the information in the prior
  - In this way, the posterior does not look exactly like the likelihood, it is "shrunken" towards the prior
  - The posterior shows a **shrinkage** towards the prior
    - The extent to which the posterior shrinks towards the prior depends on the relative amount of information in the prior and the data

Scarse information in the data (small n)

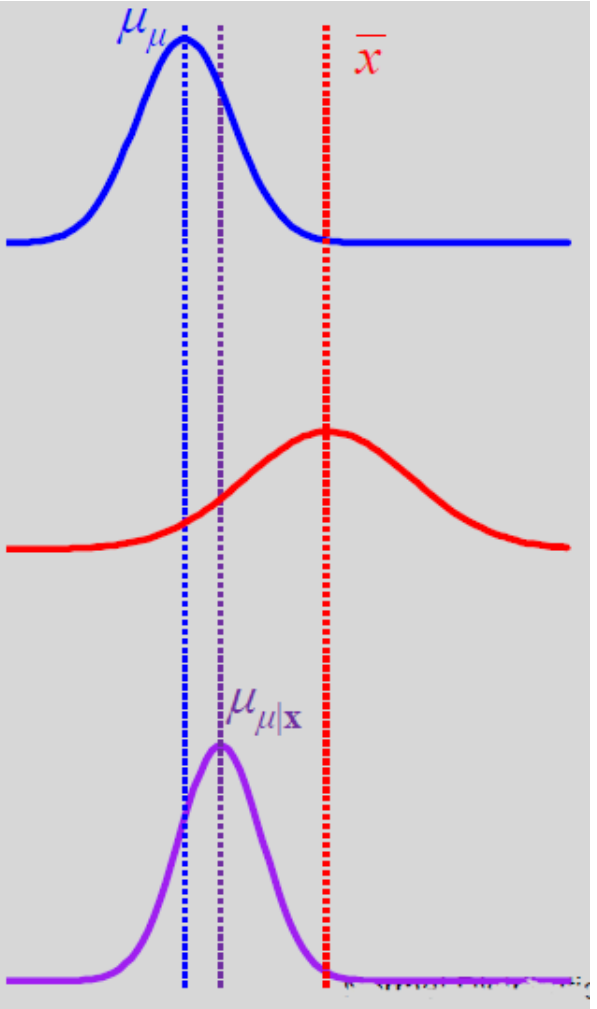
Prior: $N(\mu_\mu, \tau_\mu)$

Likelihood: $N(\mathbf{x} \mid \mu, \tau)$

$N(\bar{x} \mid \mu, n\tau)$

Posterior: $N(\mu_{\mu|\mathbf{x}}, \tau_{\mu|\mathbf{x}})$

$$\mu_{\mu|\mathbf{x}} = \frac{\tau_\mu}{\tau_\mu + n\tau}\mu_\mu + \frac{n\tau}{\tau_\mu + n\tau}\bar{x}$$



Lots of information in the data (large n)

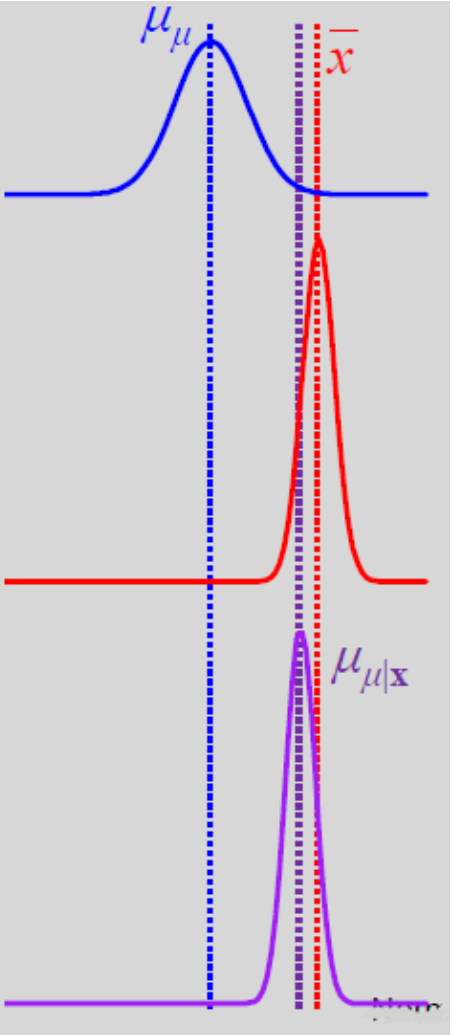
Prior: $N(\mu_\mu, \tau_\mu)$

Likelihood: $N(\mathbf{x} \mid \mu, \tau)$

$N(\bar{x} \mid \mu, n\tau)$

Posterior: $N(\mu_{\mu|\mathbf{x}}, \tau_{\mu|\mathbf{x}})$

$$\mu_{\mu|\mathbf{x}} = \frac{\tau_\mu}{\tau_\mu + n\tau}\mu_\mu + \frac{n\tau}{\tau_\mu + n\tau}\bar{x}$$



Asymptotics & Connections to Frequentist Inference

Asymptotic posterior

$$\mu_{\mu|\mathbf{x}} = \left(\frac{\mu_\mu}{\sigma_\mu^2} + \frac{n\bar{x}}{\sigma^2}\right) \Big/ \left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}\right) \qquad \sigma^2_{\mu|\mathbf{x}} = 1 \Big/ \left(\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma^2}\right)$$

- Asymptotically, as $n \to \infty$, $\sigma^2_{\mu|\mathbf{x}} \to \frac{\sigma^2}{n}$, $\mu_{\mu|\mathbf{x}} \to \bar{x}$
- Asymptotically, as $n \to \infty$, $p(\mu \mid \mathbf{x}) \to N\left(\bar{x}, \frac{\sigma^2}{n}\right)$
- Likewise if $\sigma_\mu^2 \to \infty$

($\sigma_0^2 \to \infty$) A uniform "distribution" for the mean

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

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

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

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

which is the same as we would obtain by assuming

$$p(\theta) \propto k$$

improper priors

Connections With Frequentist Approach

	Bayes, as $n \to \infty$ or $\sigma_\mu^2 \to \infty$	Frequentist, $n \to \infty$
Distribution	$p(\mu \mid \mathbf{x}) \to N\left(\bar{x}, \frac{\sigma^2}{n}\right)$	$p(\bar{X}) \to N\left(\mu, \frac{\sigma^2}{n}\right)$
Point estimate	$\mu_{\mu\mid\mathbf{x}} \to \bar{x}$	$\bar{x}$
Variability	$\sigma_{\mu\mid\mathbf{x}} \to \frac{\sigma}{\sqrt{n}}$	$\sigma_{\bar{x}} \to \frac{\sigma}{\sqrt{n}}$
Interval	$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

As $n \to \infty$ or $\sigma_\mu^2 \to \infty$, will be similar ***numerically***, but very different ***conceptually***

Critiques And Justification

- Asymptotically, the prior becomes irrelevant
  - *Principle of stable estimation*
  - Critique of "the role of the prior" loses force
- "We'll grant you the Bayesian approach asymptotically, our complaint is in finite samples"
- But frequentist approach only justified asymptotically!
  - Parameter estimation, standard errors, data-model fit
- Bayesian approach does not need to invoke asymptotic arguments
  - And if you grant them to the frequentist, shouldn't you grant them to the Bayesian?

The principle of stable estimation, or precise measurement, specifies that when a likelihood function is sharply peaked in an interval over which a prior density is relatively flat, the posterior density does not differ much from the normed likelihood function.

Summary of Normal Distribution Model With Unknown Mean

- Exchangeability yields factorization
- Sufficiency of sample mean
- Normal prior as conjugate
- Precision as the inverse of the variance
  - This is what some software (BUGS, JAGS) uses
- Posterior as the synthesis of prior and data
  - Shrinkage to the prior
- Asymptotics and connections with frequentist approaches

Normal model - Predictive distribution

Often the predictive distribution is more interesting than the posterior distribution. The posterior distribution describes the uncertainty in the parameters, while the predictive distribution also describes the uncertainty about the future event

In the case of a Gaussian distribution with known variance σ^2 the model is

$$y \sim N(\theta, \sigma^2),$$

where σ^2 describes the aleatory uncertainty.
With a uniform prior the posterior is

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

where σ^2/n describes the epistemic uncertainty related to θ, while the posterior predictive distribution of a new $\tilde{y}$ is

$$p(\tilde{y}|y) \sim N(\tilde{y}|\bar{y}, \sigma^2 + \sigma^2/n),$$

where the uncertainty is the sum of the aleatory uncertainty (σ^2) and the epistemic uncertainty (σ^2/n).

Normal model - predictive distribution - conjugate prior for the mean

With a normal prior, the predictive distribution is

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

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

Then,

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

(go to theorem)

We can derive also the moments of 1st and 2nd order as follows

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

Again, the predictive variance = variance of observational model σ^2 + posterior variance σ_n^2 (variance on the model).

Model for a normal distribution with known mean and unknown variance

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

$$y_1, \ldots, y_n | \sigma^2 \sim N(y_i | \theta, \sigma^2) \text{ iid}$$

so the likelihood is

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

where the sufficient statistic is $v = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta)^2$, which is the $\hat{\sigma}^2_{MLE}$.

Conjugate prior

The corresponding conjugate prior is the **inverse-gamma** (IG)

$$\sigma^2 \sim \text{Inv-gamma}(\alpha, \beta)$$

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

which is the same as saying that

$$\frac{1}{\sigma^2} \sim \text{Gamma}(\alpha, \beta)$$

Note that

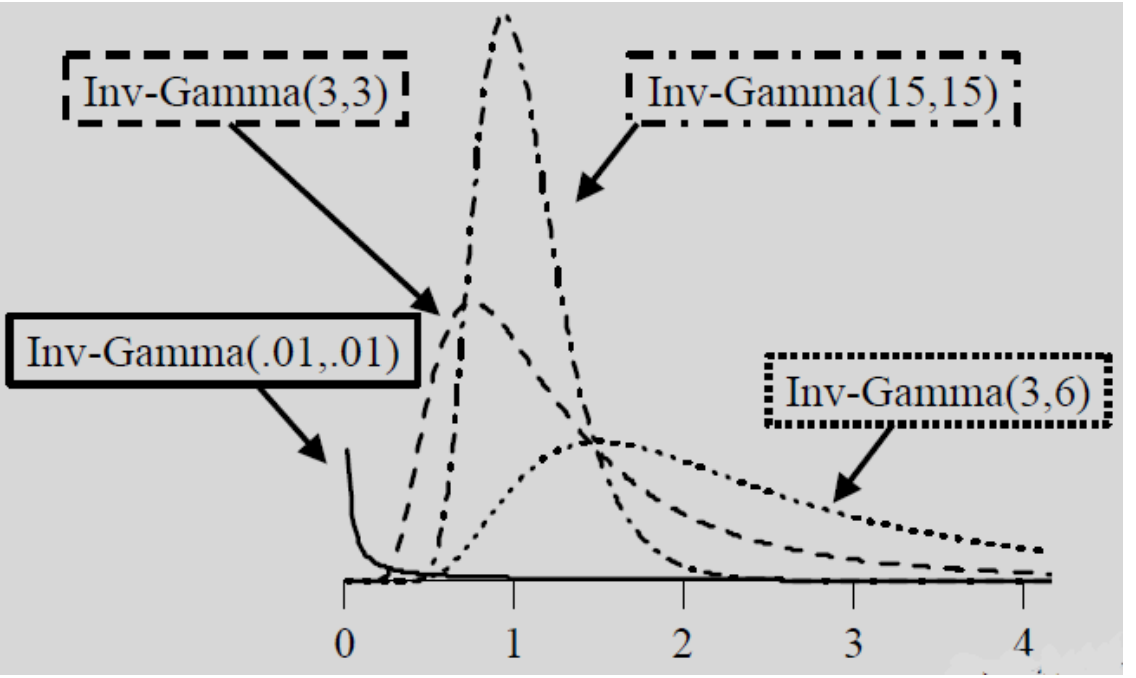
$E(\sigma^2) = \frac{\beta}{\alpha-1}$ for $\alpha > 1$,

$V(\sigma^2) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$,

$\text{Mode}(\sigma^2) = \frac{\beta}{\alpha+1}$

Inverse gamma prior for the variance

$p(\sigma^2) = \text{Inv} - \text{gamma}(\alpha, \beta)$, with $\sigma^2 \geq 0, \alpha > 0, \beta > 0$



Integral is finite if $\alpha > 0$, density is finite if $\alpha \geq 1$

Non-informative if $\alpha, \beta \to 0$

Posterior

The posterior distribution is then

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

$$p(\sigma^2|y) \propto (\sigma^2)^{-n/2} \exp\left\{-\frac{n}{2\sigma^2}v\right\}(\sigma^2)^{-\alpha-1}e^{-\beta/\sigma^2}$$

$$\propto (\sigma^2)^{-n/2-\alpha-1} \exp\left\{-\frac{1}{\sigma^2}\left[\frac{n}{2}v+\beta\right]\right\}$$

that is,

$$\sigma^2|y \sim \text{Inv-gamma}\left(\frac{n}{2}+\alpha, \frac{n}{2}v+\beta\right)$$

The posterior mean is $E(\sigma^2|y) = \frac{2\beta+nv}{2\alpha+n-2}$

The posterior mode is $\text{Mode}(\sigma^2|y) = \frac{2\beta+nv}{2\alpha+n+2}$

Reparametrization 1

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

$$\tau \sim \text{Gamma}(\alpha, \beta),$$

the likelihood is

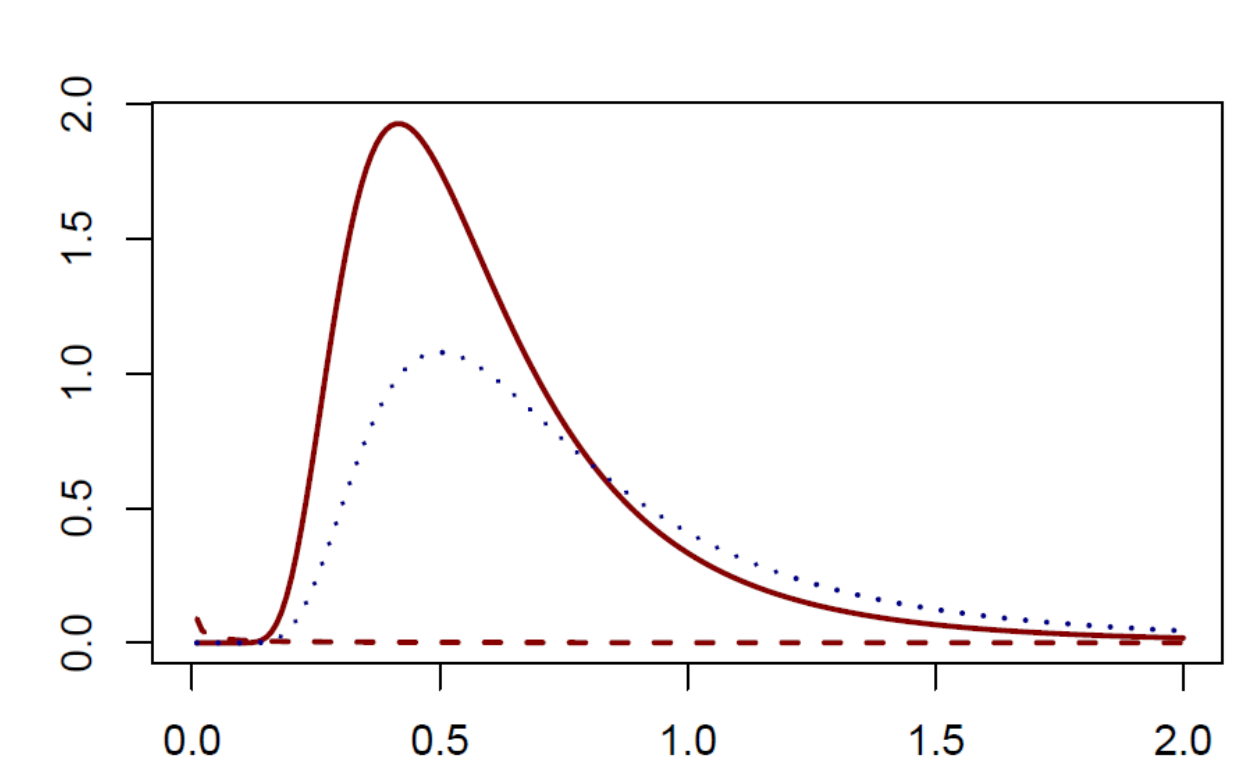
$$p(y|\tau) \propto (\tau)^{n/2} \exp\Big\{-\frac{nv}{2}\tau\Big\}$$

and the posterior is $\text{Gamma}(n/2 + \alpha, nv/2 + \beta)$:

$$\begin{aligned} p(\tau|y) &\propto \tau^{n/2} \exp\Big\{-\frac{nv}{2}\tau\Big\}\tau^{\alpha-1}e^{-\beta\tau} \\ &\propto \tau^{n/2+\alpha-1} \exp\Big\{-\tau\left[\frac{nv}{2} + \beta\right]\Big\} \end{aligned}$$

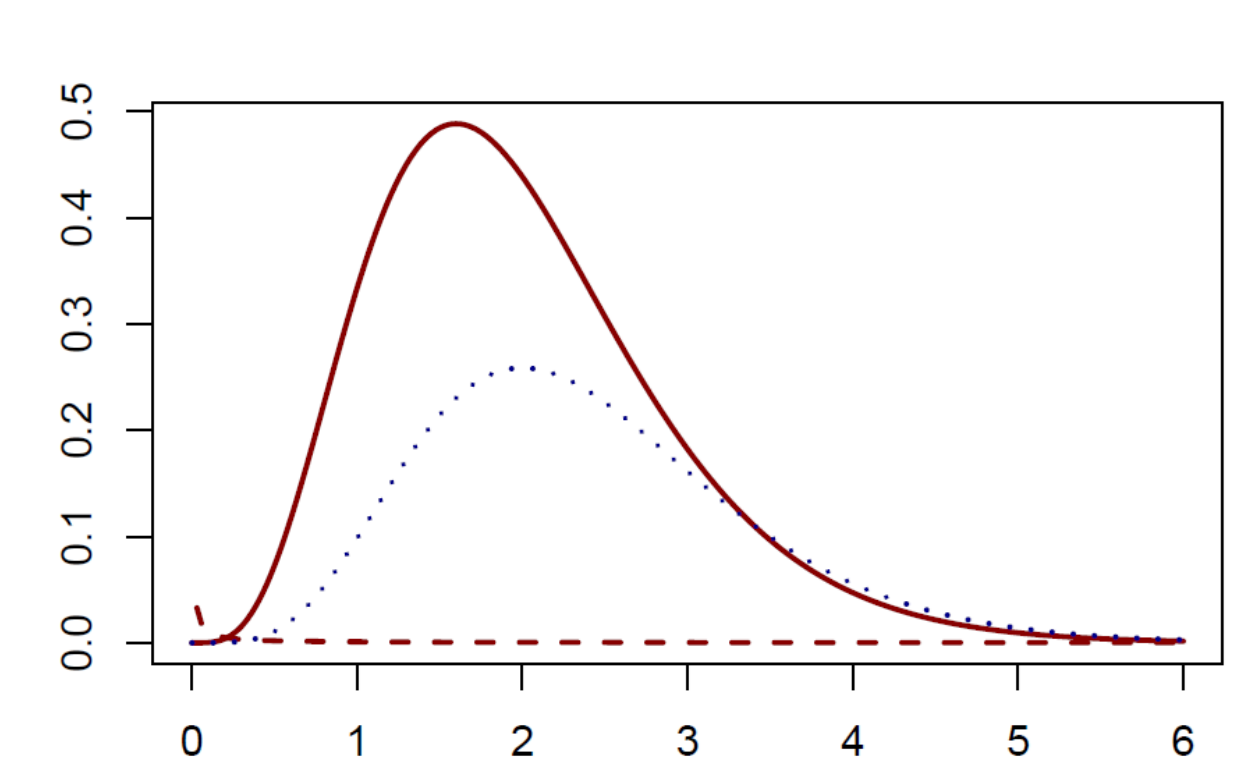
Inference for σ^2

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



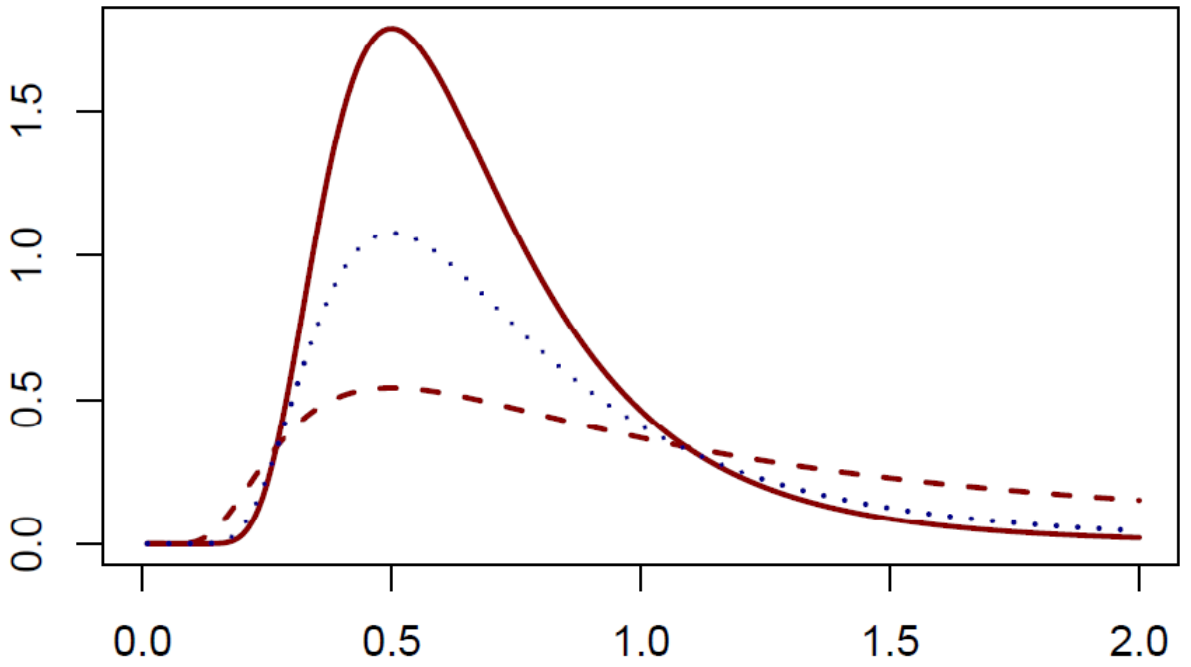
Inference for $\tau = 1/\sigma^2$

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



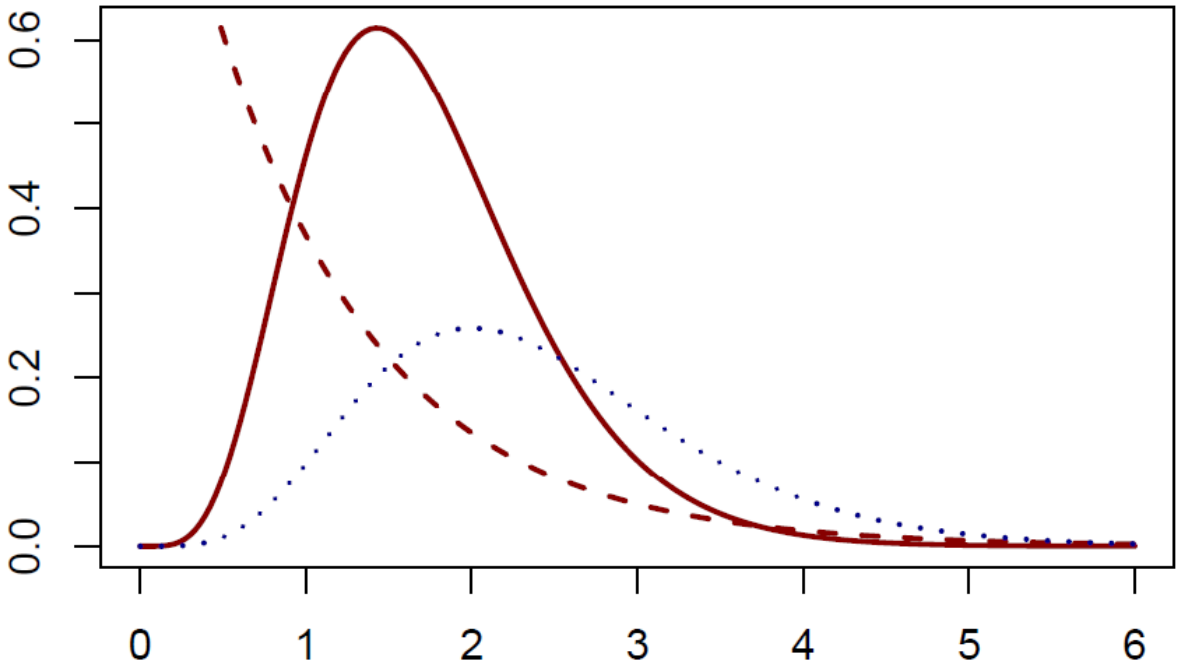
Inference for σ^2

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



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

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



Reparametrization with $\text{Inv-}\chi^2$: prior

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 $\text{Inv-}\chi^2(\nu_0, \sigma_0^2)$ (**inverse-scaled** χ^2).

This corresponds to $\nu_0 = 2\alpha$ and $\sigma_0^2 = \beta/\alpha$, or, in other terms, implies that $\sigma^2 \sim \text{Inv-gamma}(\nu_0/2, (\nu_0/2)\sigma_0^2)$, the density is

$$p(\sigma^2) = \frac{((\nu_0\sigma_0^2)/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} (\sigma^2)^{-\nu_0/2-1} \exp\{-(\nu_0\sigma_0^2)/(2\sigma^2)\}$$

Note that

$$E(\sigma^2) = \frac{\nu_0\sigma_0^2}{\nu_0 - 2}, \ \nu_0 > 2; \quad \text{Mode}(\sigma^2) = \frac{\nu_0\sigma_0^2}{\nu_0 + 2}$$

$$V(\sigma^2) = \frac{(\nu_0\sigma_0^2)^2}{(\nu_0-2)^2(\nu_0-1)}, \ \nu_0 > 2$$

Reparametrization with $\text{Inv-}\chi^2$: posterior

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.

Then

$$E(\sigma^2|y) = \frac{\nu_0\sigma_0^2 + n\hat{\sigma}^2_{MLE}}{\nu_0 + n - 2}$$

$$\text{Mode}(\sigma^2|y) = \frac{\nu_0\sigma_0^2 + n\hat{\sigma}^2_{MLE}}{\nu_0 + n + 2}$$

The priori can then be interpreted as information equivalent to ν_0 observations with variance σ_0^2 (a non-informative prior corresponds to $\nu_0 = 0$).

Summary of priors for the variance

Distribution	Density	Mean	Mode
$z \sim \text{Gamma}(\alpha, \beta)$	$\frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{\alpha-1} e^{-\beta z}$	$\frac{\alpha}{\beta}$	$\frac{\alpha-1}{\beta},\ \alpha \geq 1$
$z \sim \text{inv-Gamma}(\alpha, \beta)$	$\frac{\beta^{\alpha}}{\Gamma(\alpha)} z^{-\alpha-1} e^{-\beta/z}$	$\frac{\beta}{\alpha-1},\ \alpha > 1$	$\frac{\beta}{\alpha+1}$
$z \sim \text{inv-}\chi^2(\nu, \sigma)$ $z \sim \text{inv-Gamma}(\nu/2, \nu\sigma/2)$	$\frac{(\nu\sigma/2)^{(\nu/2)}}{\Gamma(\nu/2)} z^{-\nu/2-1} e^{-\nu\sigma/(2z)}$	$\frac{\nu\sigma}{\nu-2},\ \nu > 2$	$\frac{\nu\sigma}{\nu+2}$

Poisson Model

Count data: Poisson likelihhod

Assume that y_i is a count modelled as a Poisson variate and that $y = (y_1, \ldots, y_n)$ are observed and

$$y_1, \ldots, y_n | \theta \sim \text{Poisson}(y_i|\theta) \text{ iid}$$

then, if we let $t(y) = \sum_{i=1}^n y_i$, the likelihood is

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

where $t(y)$ is the sufficient statistics.

The likelihood

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

belongs to an exponential family with natural parameter $\phi(\theta) = \log \theta$

Conjugate prior to Poisson likelihood

The conjugate prior has to be

$$p(\theta) \propto \theta^{\alpha-1} e^{-\beta\theta} \propto \text{Gamma}(\alpha, \beta)$$

which is equivalent to a total count of $\alpha - 1$ in β prior observations.

The posterior is then

$$p(\theta|y) = \text{Gamma}(\alpha + \sum y_i, \beta + n)$$

The posterior mean being

$$E(\theta|y) = \frac{\alpha + n\bar{y}}{\beta + n}$$

Distribution of y is Negative Binomial

Prior predictive distribution of a Poisson model has a negative binomial distribution.

Note that $p(y) = \frac{p(y|\theta)p(\theta)}{p(\theta|y)}$, then, for a single poisson observation y,

$$p(y) = \frac{\text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha,\beta)}{\text{Gamma}(\theta|\alpha+y,\beta+1)}$$

which reduces to

$$p(y) = \binom{\alpha+y-1}{y}\left(\frac{\beta}{\beta+1}\right)^{\alpha}\left(\frac{1}{\beta+1}\right)^{y}$$

known as negative binomial

$$y \sim \text{Neg-Bin}(\alpha,\beta)$$

The negative binomial distribution describes the number of failures that occur until a predefined number of successes α occurs in a Bernoulli process with parameter $\beta/(\beta+1)$.

Negative Binomial is a mixture of Poisson

The above derivation shows how the negative binomial distribution is a mixture of Poisson distributions with θ rates following a Gamma distribution.

$$\text{Neg-Bin}(\alpha, \beta) = \int \text{Poisson}(y|\theta)\text{Gamma}(\theta|\alpha, \beta)d\theta$$

Poisson model with exposure

Assume we know for each observation y_i the value of an explanatory variable x_i (> 0), e.g.

- y_i is the incidence of a desease in unit i
- x_i is the exposure in unit i
- θ is the incidence *rate*

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

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

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

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

Note: model non exchangeable in y_i's but exchangeable in couples $(x, y)_i$.

Example: estimate of incidence rate from count data

- In a given year, in a US city of 200,000 inhabitants, we observed $y = 3$ deaths for asthma
- Let θ be the underlying long-term asthma mortality rate (measured in cases per 100,000 persons per year)
- The exposure is then $x = 2$
- Then, $y|\theta \sim Pois(2\theta)$
- (prior elicitation) If we assume $\theta \sim \text{Gamma}(3, 5)$ (assuming that the city and year are exchangeable with other years and cities from which we get information about θ)
- Then, $\theta|y \sim \text{Gamma}(6, 7)$
- *Do you notice a* shrinkage *toward the prior compared to the "raw" mortality rate?*
- Imagine that we have more data: $y = 30$ in the following 10 years (in the same city, with a constant population of 200,000). How does the posterior change?

Prior elicitation

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

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

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

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

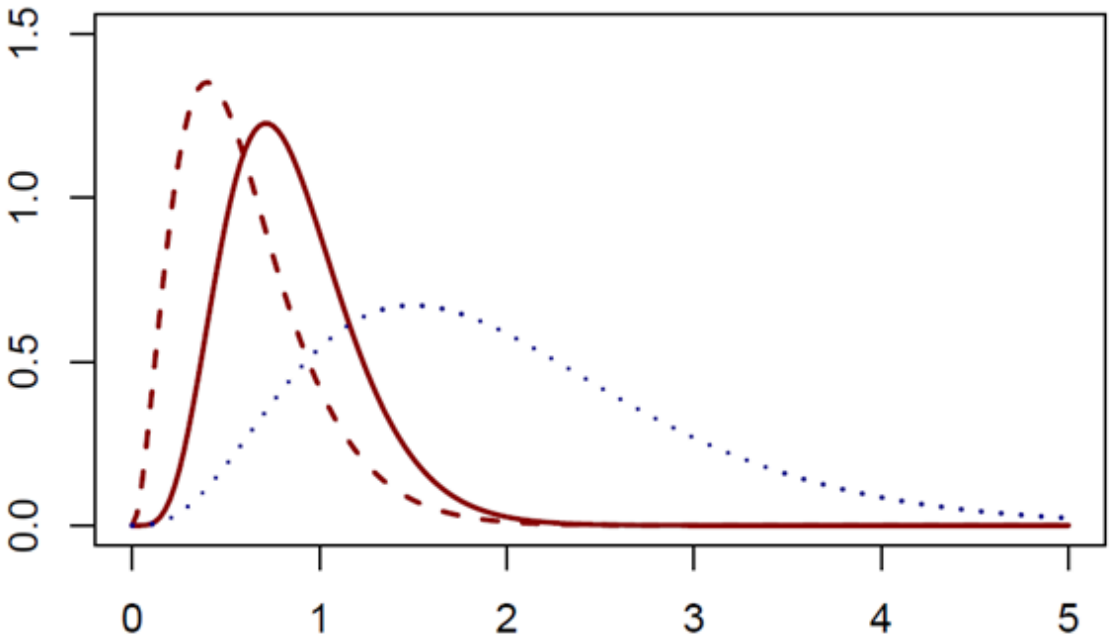
leads to $P(\theta < 1.44) = 0.975$.

Shrinkage to the prior

Starting from a prior $p(\theta) = \text{Gamma}(3, 5)$ 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



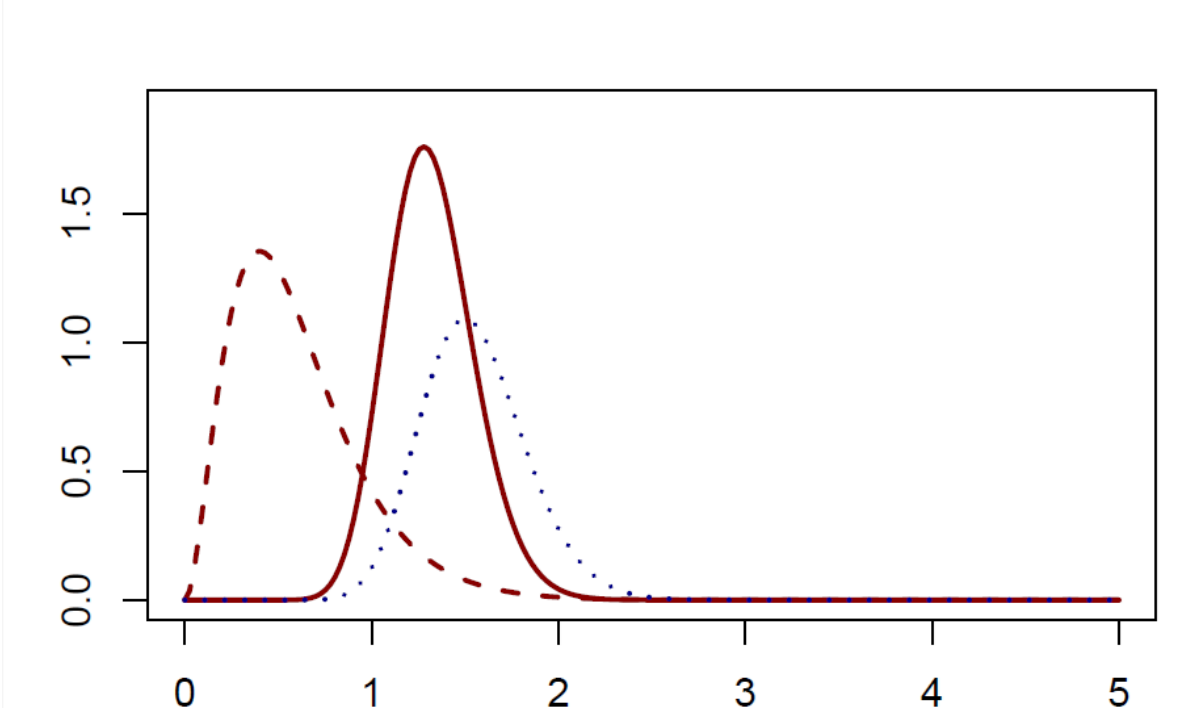
The posterior mean is shrunken towards the prior mean 0.6 away from the observed mortality rate 1.5.

More (exchangeable) observations

With a constant population of 200,000 and assuming that the ten-year results are independent with a constant long-run rate θ, the posterior distribution of θ is then

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

whose mean is 1.32.



Exponential Model

Exponential Model

- y continuous variable, $y \geq 0$, real-valued, e.g. a waiting time
- property: *memoryless*

$$P(y > t + s | y > s,\, \theta) = P(t > t | \theta)\ \forall\, t, s$$

- Sampling model $\mathrm{Exp}(y|\theta)$

$$p(y|\theta) = \theta \exp\left(-\theta y\right) \text{ for } y \geq 0,\ \theta > 0$$

  $E(y) = 1/\theta, \theta = 1/E(y|\theta)$ is called the *rate*
- $\mathrm{Exp}(\theta) = \mathrm{Gamma}(\alpha = 1, \beta = \theta)$
- If n observations, the likelihood is $p(y|\theta) = \theta^n \exp\left(-\theta \sum_i y_i\right)$
- Prior $p(\theta) = \mathrm{Gamma}(\alpha, \beta) \propto \theta^{\alpha-1} e^{-\beta\theta}$
- From the likelihood prior parameters can be interpreted as $\alpha - 1$ exponential observations with total waiting time β
- Posterior $p(\theta|y) = \mathrm{Gamma}(\alpha + 1, \beta + y)$

Prior distribution

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.

$$Y_1, \dots, Y_n \ \text{ iid}(\mathcal{N}(\theta, 1)) \text{ hence } L(\theta) \propto \exp(-\frac{n}{2}(\bar{y} - \theta)^2)$$

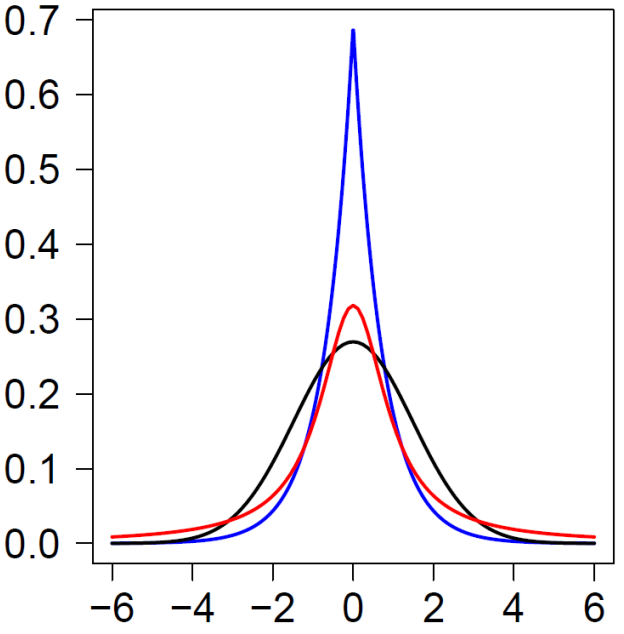
Let us fix the quartiles of the prior distribution:

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

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.)



Laplace
$p(\theta) = \frac{\lambda}{2}\exp(-\lambda|\theta|)$

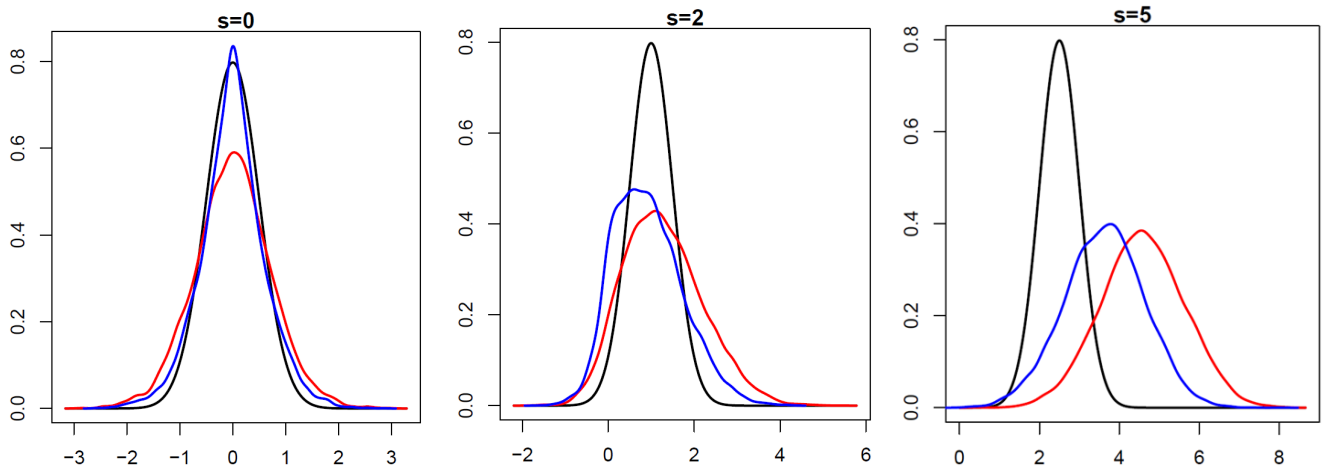
Cauchy
$p(\theta) = \frac{1}{\pi(1+\theta^2)}$

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

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 $\text{supp}(p(\theta)) = \Theta$
(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 $p(\theta) = f(\theta;\nu)$ the posterior distribution is in the same family, that is $p(\theta|y) = f(\theta;\nu_1)$ 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;

$\times$ 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

$$p(\theta) \propto k$$

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?

Improper priors

If we apply (the algebra of) Bayes theorem

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

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

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

  See example of uniform prior for the mean of a normal

In practice: the uniform prior may work even if the parameter space is not 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

$$p(\theta) \propto k$$

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

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

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

$$p(\theta) \propto \sqrt{\mathcal{I}(\theta)}$$

where $\mathcal{I}$ is the 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 $\psi = \psi(\theta)$

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

Jeffreys' prior: example

Consider a Binomial experiment, so the log-likelihood is

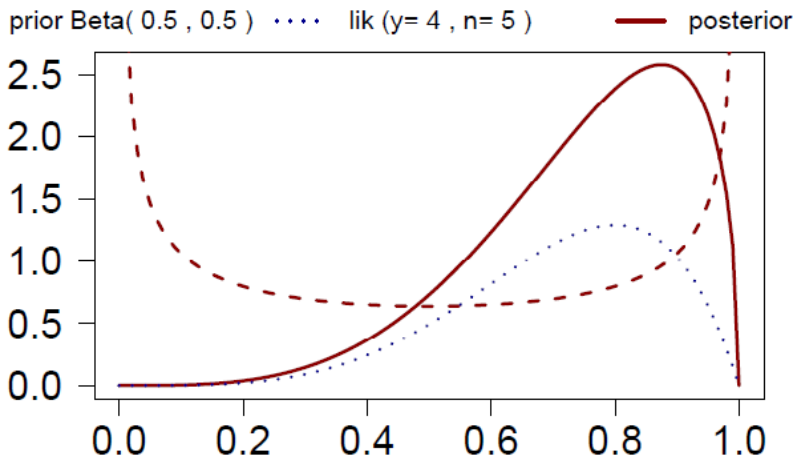
$$\log p(y|\theta) = y \log \theta + (n-y) \log(1-\theta)$$

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 $\text{Beta}(1/2, 1/2)$

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



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

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

If the Jeffrey's prior is chosen then it implies

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

which is also equal to

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

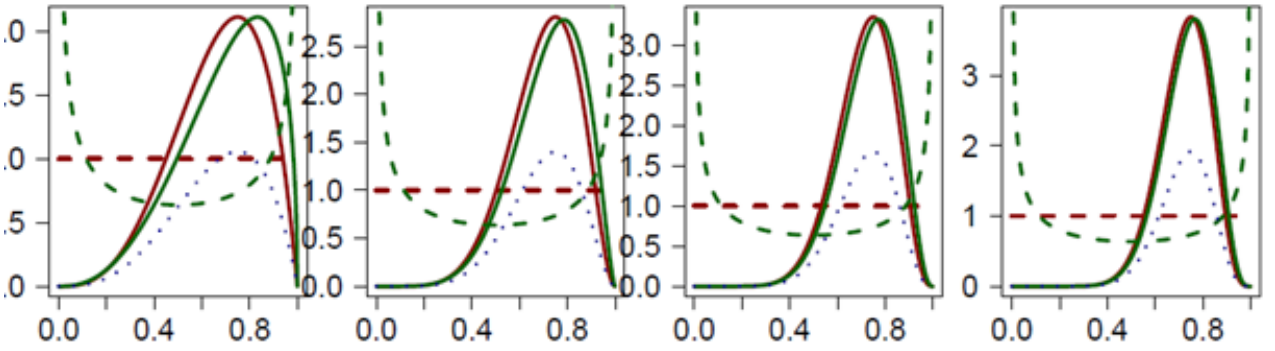
Binomial, reparametrization (cont.)

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 $p(\theta) \sim N(0.5, 0.1^2)$, or $p(\theta) \sim \text{Beta}(20, 20)$
- for the inference on the mean $p(\theta) \sim N(0, A^2)$ 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 wholly 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.

Appendix

Theorem: Mixture of Normals

Theorem

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

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

This is easily seen, let

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

then $Y = Z + \theta$

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

Back to normal predictive