Bayesian Methods in Economics and Finance (August 26-30, 2019) Bertinoro

prof. Gaetano Carmeci[1]

[1]DEAMS, University of Trieste, Italy

Fundamentals of Bayesian Econometrics
Bertinoro, August 26th 2019

Outline of the talk

1. Preliminaries on probability calculus
   - Bayes' theorem
   - Bayes' theorem: continuous variables
   - Classical and Bayesian inference
   - An important detail, which probability?
   - de Finetti's representation theorem
2. Models
   - Beta-Binomial
   - Normal-normal
   - Normal-normal, known variance
   - Interval estimate
   - Normal-normal, known mean, unknown variance
   - Normal-normal, two unknowns
3. Prediction
4. Multivariate normal
5. Multivariate t

Outline of the talk

Bayesian Econometrics I

- Bayesian econometrics is based on a few simple rules of probability.
- All the things an econometrician is interested in, such as estimate the parameters of a model, compare different models or obtain a prediction from a model, involve the same rules of probability.
- Bayesian methods are thus universal and can be used any time a researcher is interested in using data to learn about a phenomenon.
- In particular, at the heart of Bayesian econometrics lies the *Bayes' theorem*

Probability: Bayes' theorem I

- Bayes' theorem is a rule to compute **conditional probabilities**.
- In other words, it links probability measures on different spaces of events: given two events E and H, the **probability of H conditional on E** is the probability given to H *knowing that, or better, assuming that E is true* (i.e. E is the new universe (Ω)).
- More precisely,
  - I've given a probability measure on E, H and the algebra over E and H,
  - I am told that E has occurred,
  - how do I change (if I change) my opinion on H:

$$P(H) \rightarrow P(H|E) = ?$$

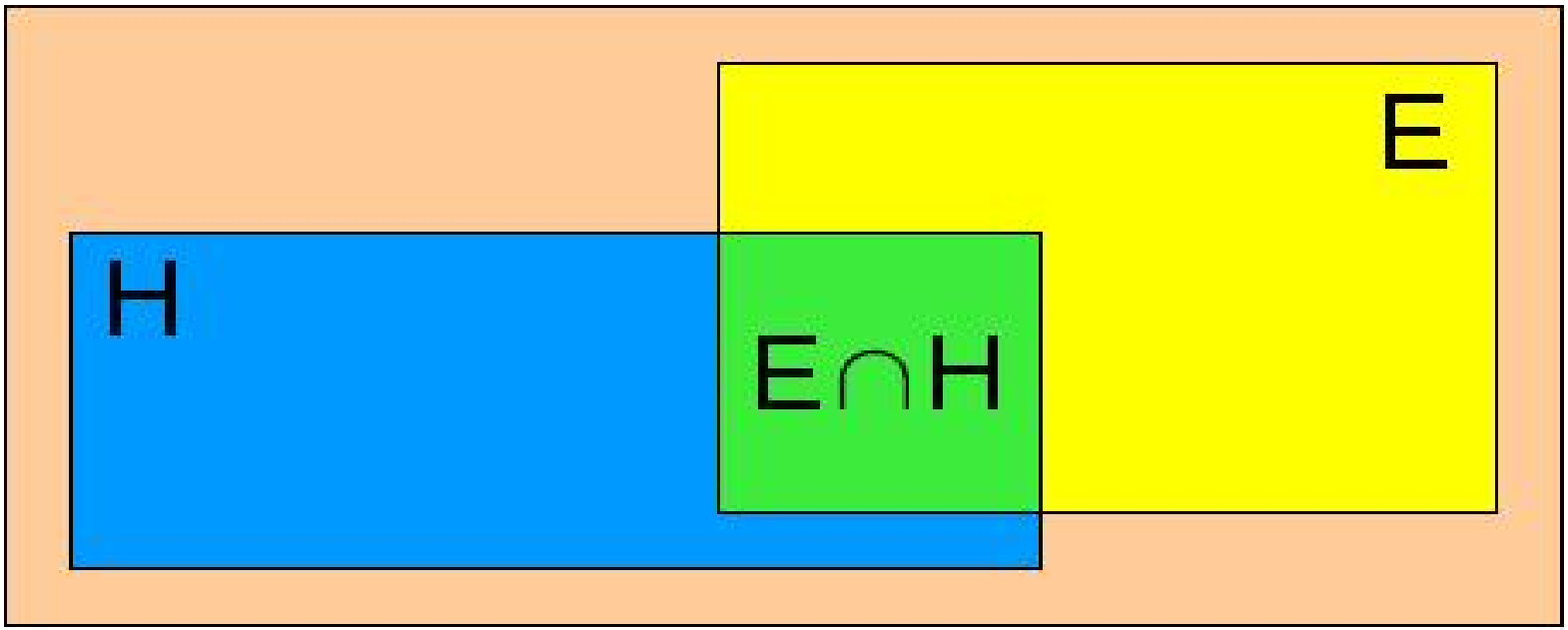
Probability: Bayes' theorem II

Theorem

Bayes' theorem (for events):Let E and H be two events, assume $P(E) \neq 0$, then

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P(H)P(E|H)}{P(E)}$$

Probability: Bayes' theorem III



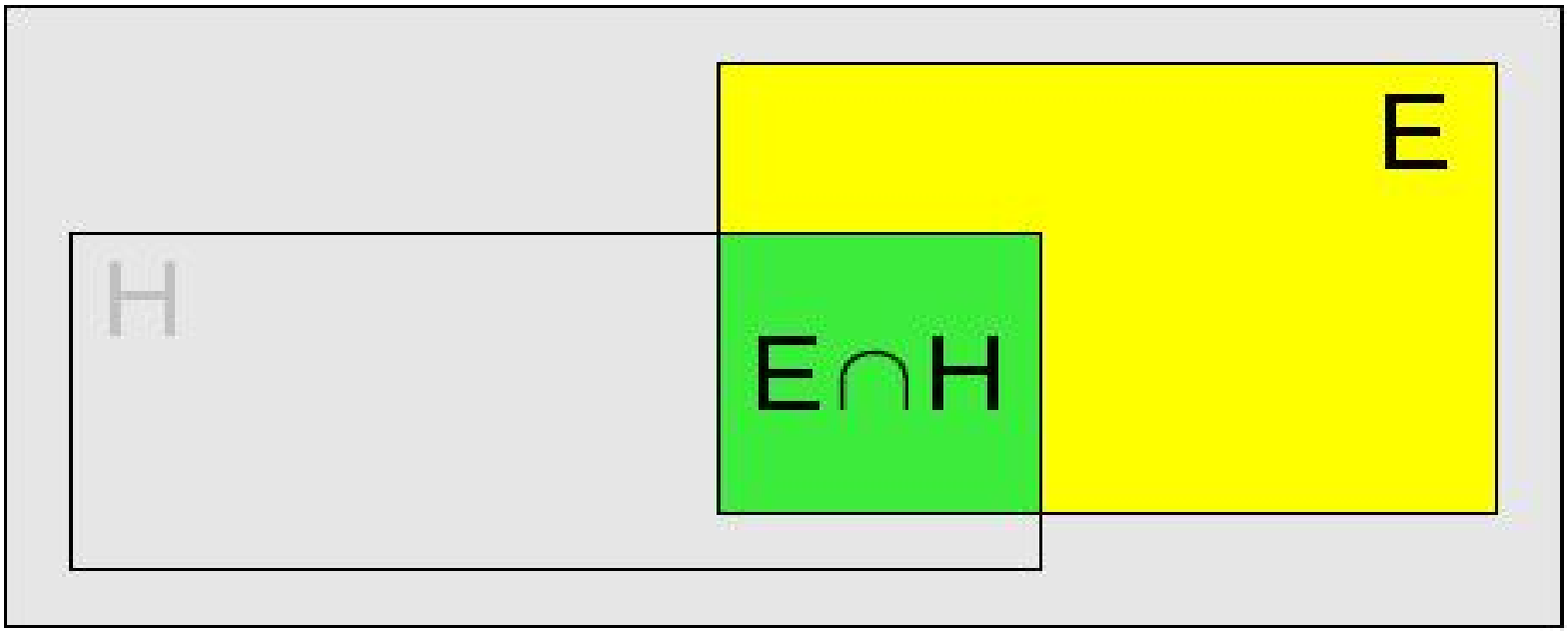
Probability: Bayes' theorem IV

Theorem

Bayes' theorem (for events):Let E and H be two events, assume $P(E) \neq 0$, then

$$P(H|E) = \frac{P(H \cap E)}{P(E)} = \frac{P(H)P(E|H)}{P(E)}$$

Probability: Bayes' theorem V



Bayes' theorem: continuous variables I

We have considered Bayes' theorem for events, we now extend it to continuous random variables.

Theorem

Bayes' theorem: *If*
(i) $\pi(\theta)$ *probability density function*
(ii) $f(y|\theta)$ *probability density function of* y *given* θ
then

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_\Theta f(y|\theta)\pi(\theta)d\theta} \propto \pi(\theta)f(y|\theta)$$

Note that $\int_\Theta \pi(\theta|y)d\theta = 1$.

Bayesian paradigm I

Consider a (parametric) model, i.e. a family of probability distributions indexed by a parameter $\theta \in \Theta$ such that within this family it is assumed to lie the distribution of y:

$$f(y|\theta), \quad \theta \in \Theta.$$

This way to proceed it is not different from the classical paradigm, but here distributions are defined conditional on the value of the parameter. Notice that a parameter is not a r.v. in the classical setting, whereas here all unknown quantities are treated as r.v. The likelihood

$$L(\theta; y) \propto f(y|\theta),$$

is defined as in the classical paradigm.

Bayesian paradigm II

A prior distribution is set on the parameter θ,

$$\pi(\theta)$$

which is independent of observations (it is called prior since it comes before seeing the data: this is the new thing!).

Prior information and likelihood are combined using Bayes' theorem to obtain the posterior distribution

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_\Theta f(y|\theta)\pi(\theta)d\theta} \propto \pi(\theta)f(y|\theta) \propto \pi(\theta)L(\theta; y)$$

which sums up all the information we have on the parameter θ.

Inference

Within the classical inferential paradigm we distinguish the following procedures

- point estimate;
- interval estimate;
- hypotheses testing;

These distinctions are less substantial in Bayesian inference, where

THE result of inference IS
the POSTERIOR DISTRIBUTION

The two approaches, a brief comparison I

A statistical problem is faced when, given observations, we want to assess what random mechanism generated them In other words,

- there are two or more probability distributions which may have generated the observations;
- analyzing the data we want to infer the actual distribution (or some property of it) which generated the data.

The two approaches, a brief comparison II

In **CLASSICAL INFERENCE**

- the conclusion is not directly derived from the rules of probability calculus (in fact, these rules are used, but the conclusion is not a direct consequence of them)
- the likelihood and the probability distribution of the sample are used;
- the parameter is treated as a constant.

In **BAYESIAN INFERENCE**

- the reasoning and the conclusion is an immediate consequence of probability calculus rules (specifically of Bayes' theorem);
- the likelihood and the prior distribution are used;
- the parameter is a random variable.

Bayesian statistics and subjective probability I

- It should be clear by now that Bayesian statistics is an application of probability calculus.
- In the following we will adopt the most general definition of probability, i.e. the definition of **subjective probability**.

The subjective approach to probability is based on the idea that probability is not an objective property of a phenomenon but rather it has to do with the personal opinion of an individual

Def. of Subjective probability: for an individual, the probability of an event corresponds to his degree of belief on the occurrence of it.

- Since probability is defined as a subjective degree of belief, it depends upon the information available to the individual, and following this line of reasoning it is clear that by random we mean not known for lack of information.

- Therefore the subjective definition of probability seems to be well suited to the Bayesian paradigm, in which:

  - the parameter to be estimated is a well specified quantity but not known for lack of information and so it is treated as a r.v.

  - a probability distribution is (subjectively) specified for the parameter to be estimated, this is called *the prior distribution*

  - having observed the experimental results, the probability distribution on the parameter is updated using Bayes' theorem to combine experimental results (*likelihood*) and prior distribution to obtain the *posterior* distribution.

Exchangeability (B. de Finetti's representation theorem)

Key hypothesis in many statistical analyses: the observed quantities $y_1, \ldots, y_n$, are exchangeable, i.e. their joint probability distribution is invariant to any permutation of indices

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

The exchangeability hypothesis is equivalent to affirm that conditionally on a parameter vector θ, the random variables $y_1, \ldots, y_n$ are independently and identically distributed (de Finetti's theorem)

Notice that unconditionally the random variables are stochastically dependent!

Outline of the talk

Bayesian inference in brief

1. Specify a model, that is
   - the **conditional distribution** $p(y|\theta)$ associating to each state of nature a probability law of the sample
   - the set of states of nature: **parameter space**, Θ

   In the end we have
   $$\{p(y|\theta), \theta \in \Theta\}$$
2. specify a distribution on the parametric space based on pre experimental opinions: the **prior distribution**, $\pi(\theta)$ (a minimum requirement is that $\text{supp}(\pi) = \Theta$);
3. combine the prior distribution and the likelihood according to **Bayes' theorem** to obtain the **posterior distribution**
   $$\pi(\theta|y) \propto p(y|\theta)\pi(\theta)$$
4. derive conclusions from the posterior distribution (point or interval estimates, hypotheses testing, prediction, ...)

Scheme of inference

determine the states of nature, that is the **parameter space**	Θ
specify a distribution on the parametric space based on pre experimental opinions: the **prior distribution**	$\pi(\theta)$ $(\text{supp}(\pi) = \Theta)$
specify the **likelihood function** associating to each state of nature a probability law of the sample	$p(y\|\theta)$ $(\text{supp}(p) = \mathcal{Y})$
combine the prior distribution and the likelihood according to **Bayes' theorem**,	$\pi(\theta\|y) \propto p(y\|\theta)\pi(\theta)$
the **posterior** distribution is then obtained	$\pi(\theta\|y)$ $(\text{supp}(\pi(\cdot\|y) = \Theta)$
from the posterior distribution we can derive point estimates, interval estimates or perform hypotheses testing.	$E(\theta\|y)$

Some models

In the following we'll discuss some simple examples of Bayesian models.

- They will serve the purpose of exemplifying some aspects of Bayesian inference;
- They are particularly useful since inference can be done without recurse to simulation algorithms (which instead is the rule in Bayesian econometrics).

Model for dichotomic data I

- Dichotomic data arise, for example, when you want to estimate the proportion of a population which possesses a given characteristic using a random sample of individuals,
- That is, n individuals are drawn at random from the population (assume the drawings are performed with replacement if the population is finite) and the possession or not of the given characteristic is recorded (this is called a success or a failure, respectively).
- Examples include the vote in a referendum, the fact of having a given gene or to be unemployed and so on.
- Bayesian inference for dichotomic data was considered by Laplace, who wanted to estimate the proportion of female births in a human population.

Model for dichotomic data

Sample: n IID replications of a dichotomic phenomenon

$$Y_1, \ldots, Y_n$$

$$Y_i|\theta \sim i.i.d.\ \text{Bernoulli}(\theta) \text{ where } pr(Y_i = 1|\theta) = \theta$$

Parameter space: $\Theta = [0, 1]$

Prior distribution: Any distribution with $[0, 1]$ support; for the moment we choose, for mathematical convenience,

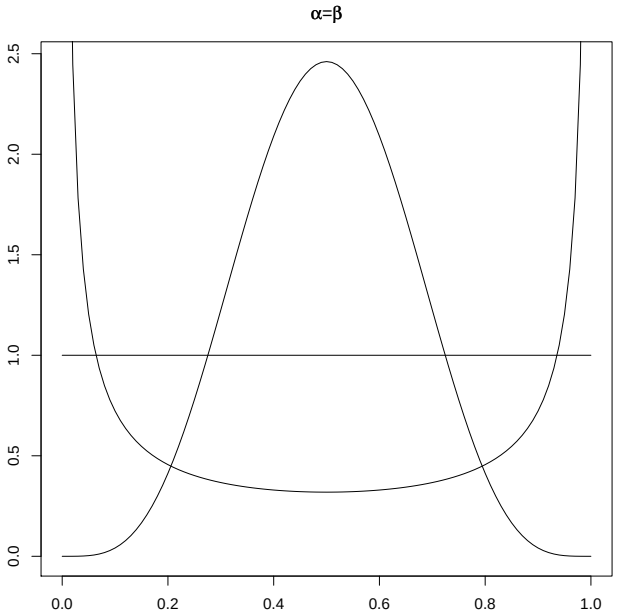
$$\theta \sim \text{Beta}(\alpha, \beta)$$

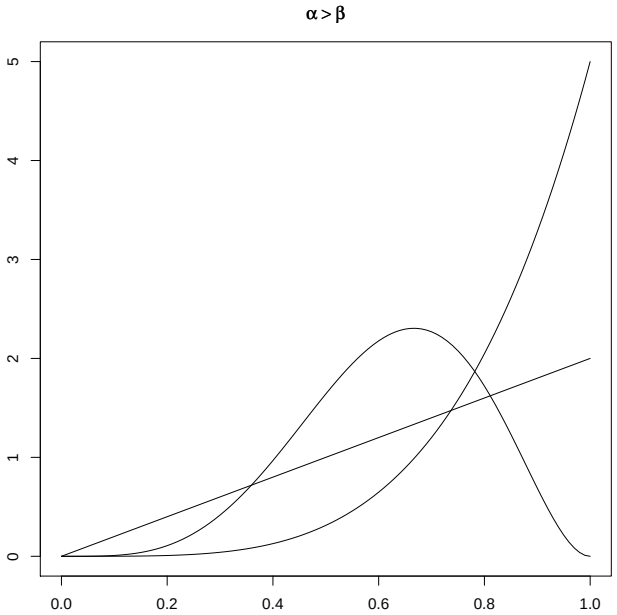
The Beta distribution

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

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

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

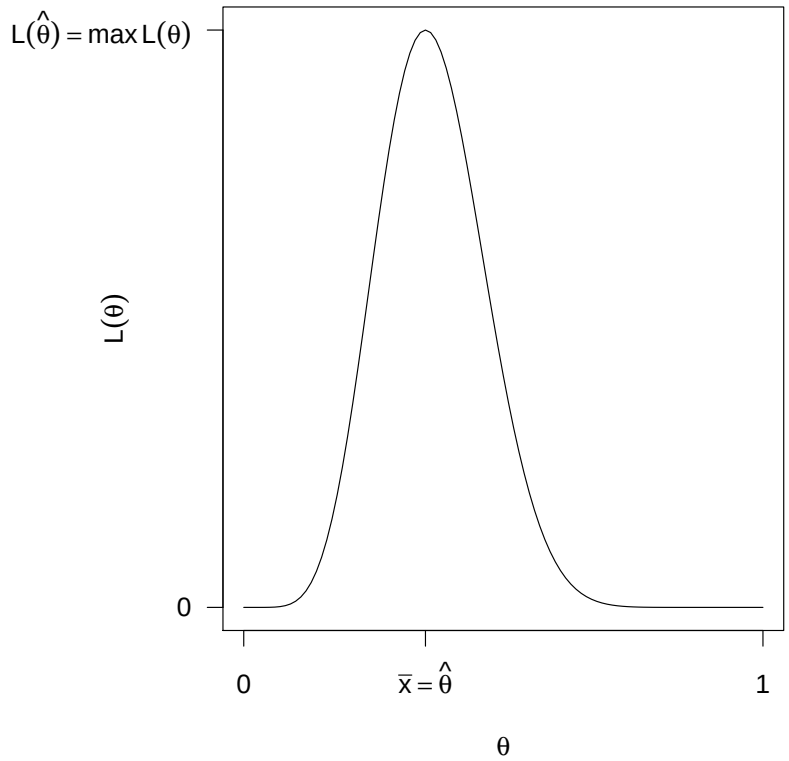




Model for dichotomic data: likelihood

The likelihood is given by

$$L(\theta) = p_\theta(y) = \theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}$$



Model for dichotomic data: the posterior distribution

The posterior distribution is easily obtained

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

$$\Downarrow$$

$$\pi(\theta|y) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum_i y_i)\Gamma(\beta+n-\sum_i y_i)}\theta^{\alpha+\sum_i y_i-1}(1-\theta)^{\beta+n-\sum_i y_i-1}$$

This is a Beta distribution with parameters $(\alpha + \sum_i y_i)$ and $(\beta + n - \sum_i y_i)$.

Point estimate

- In many cases we may be interested in summarizing the posterior distribution, either
  - because we want to highlight some features of it or
  - because the posterior distribution as a whole may be too difficult to analyse (this may occur even for a simple posterior like this one, depending on who do we need to communicate with).
- a very brute summary is a point estimate which may be thought as a guess on the parameter;
- the posterior expectation, $E(\theta|y)$, is the typical choice;
- the posterior median or posterior mode are valid alternatives;
- their interpretation is clear.

Posterior expectation I

Let us synthesize the posterior distribution using the expectation

$$E(\theta|y) = \int \theta \pi(\theta|y) d(\theta)$$

Being $\pi(\theta|y)$ a Beta distribution, the posterior mean results to be

$$\begin{aligned} &= \frac{\alpha + \sum_i y_i}{\alpha + \beta + n} \\ &= \frac{\alpha + \beta}{(\alpha + \beta + n)} \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \frac{\sum_i y_i}{n} \\ &= \frac{\alpha + \beta}{\alpha + \beta + n} E(\theta) + \frac{n}{\alpha + \beta + n} \hat{\theta} \end{aligned}$$

Posterior expectation II

$$E(\theta|y) = \frac{\alpha+\beta}{\alpha+\beta+n} \underbrace{E(\theta)}_{\text{prior mean}} + \frac{n}{\alpha+\beta+n} \underbrace{\hat{\theta}}_{ML estimate}$$

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 observations of the prior distribution.

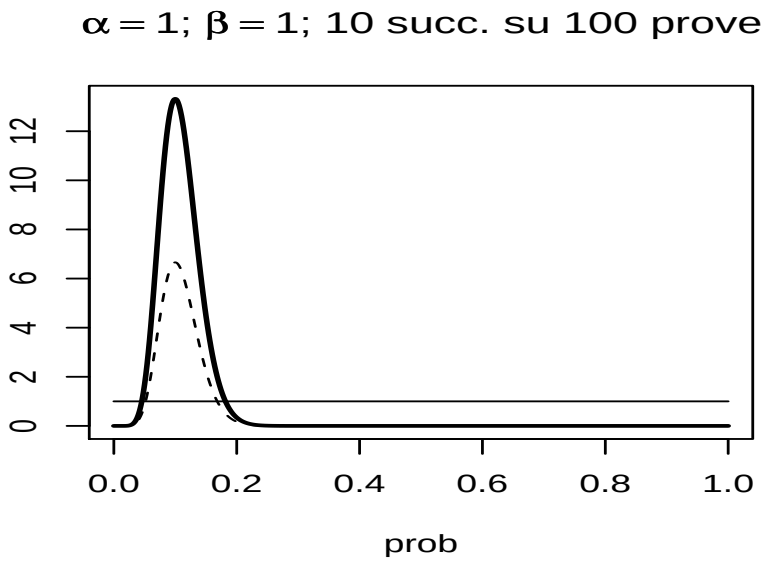
Looking at the whole distribution...

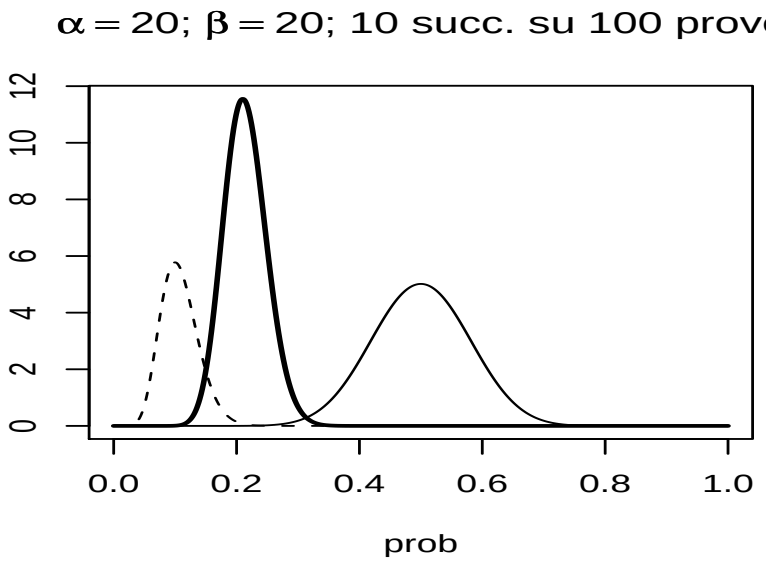
Posterior distribution

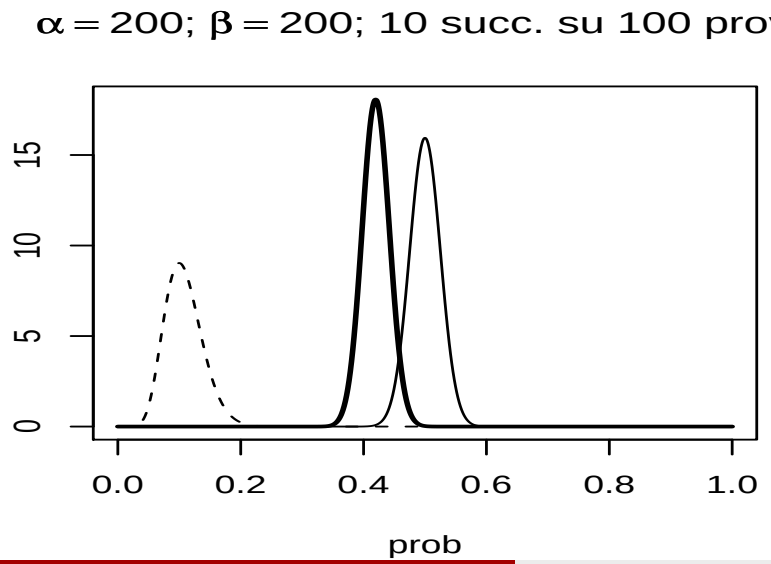
Let us consider the posterior distribution as a whole.The posterior is, in a sense, a compromise between the prior and the likelihood, where

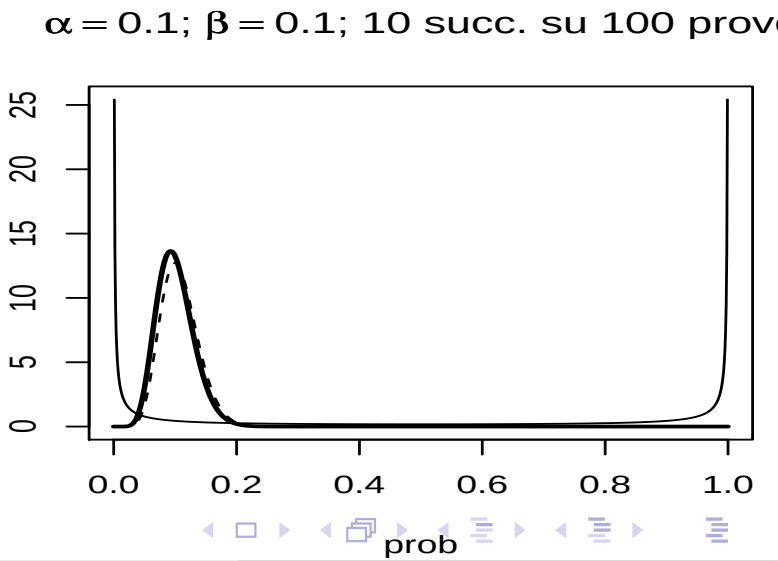
- the likelihood prevails if
  - n is large;
  - α and β are small (the prior is diffuse)

Effect of the prior



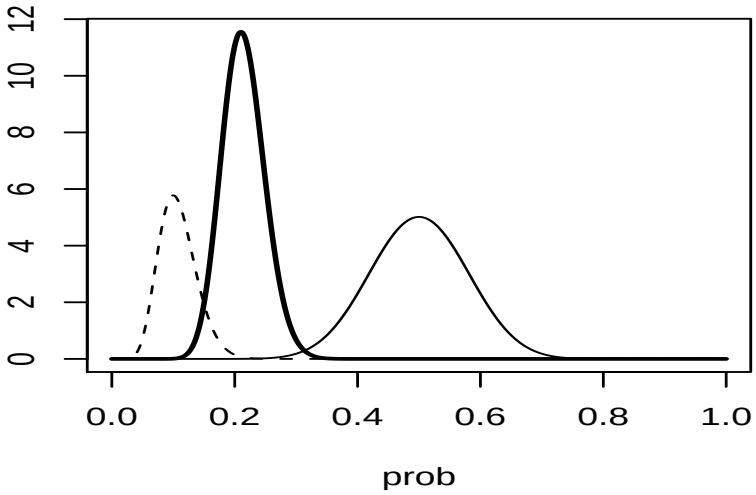




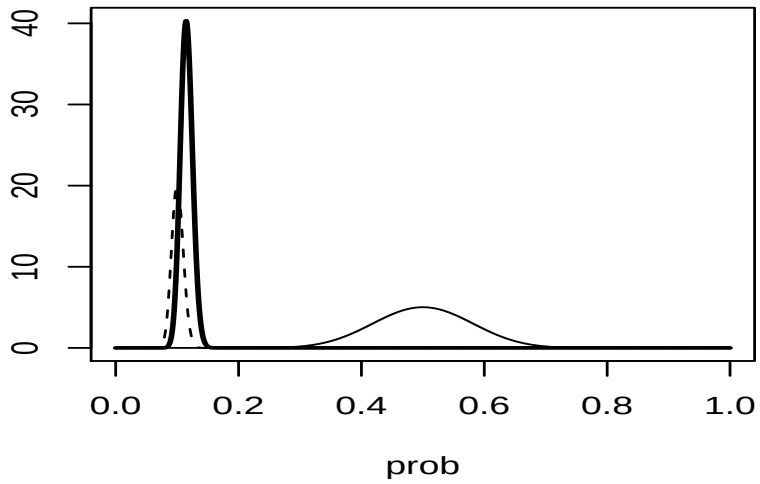


Effect of the likelihood

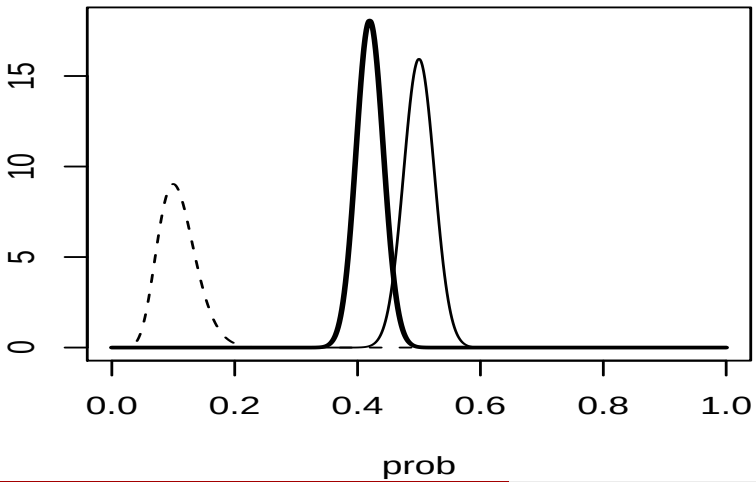
$\alpha = 20$; $\beta = 20$; 10 succ. su 100 prove



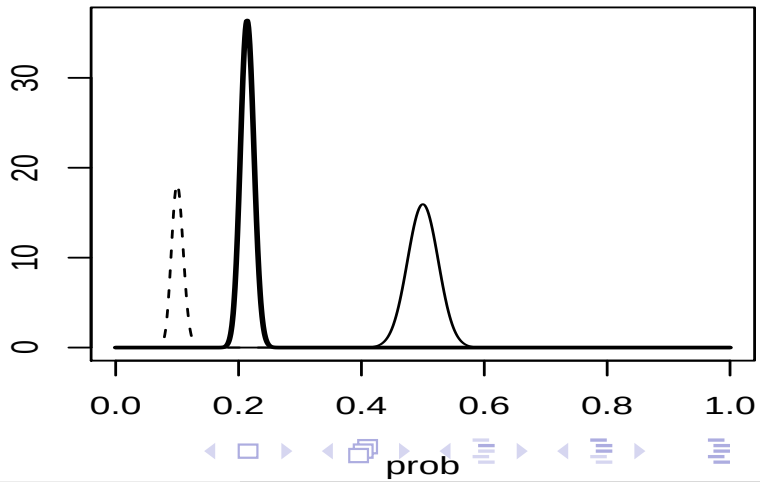
$\alpha = 20$; $\beta = 20$; 100 succ. su 1000 prove



$\alpha = 200$; $\beta = 200$; 10 succ. su 100 prove



$\alpha = 200$; $\beta = 200$; 100 succ. su 1000 prove



Note: prior and posterior distribution have the same functional form

The posterior like the prior is a Beta distribution

Likelihood	Prior	Posterior
$L(\theta; y)$	$\pi(\theta)$	$\pi(\theta \mid y)$
Binomial	Beta(α, β)	Beta$(\alpha + \sum_i y_i, \beta + n - \sum_i y_i)$

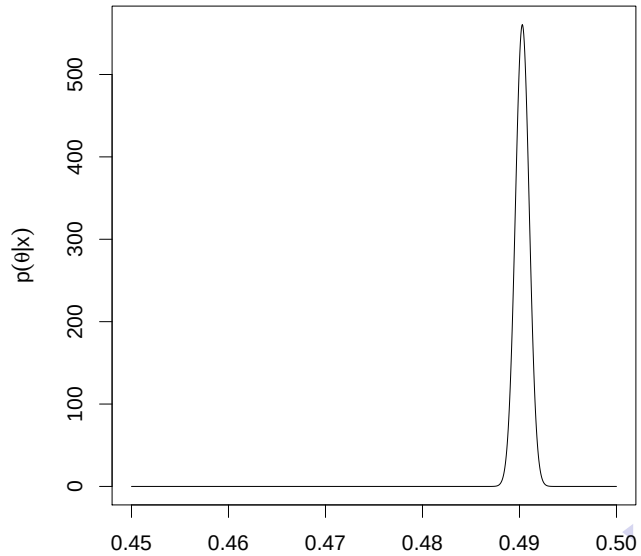
This property (discussed in the following) is called conjugacy.

Laplace example

As noted previously this model was considered by Laplace for estimating the proportion θ of female birth in a population.He considered data on births in Paris from 1745 to 1770: in this period 241,945 females and 251,527 males were born, so

$$\bar{y} = \frac{241,945}{241,945 + 251,527} = 0.4902912, \quad n = 493,472.$$

Assuming as a prior $\theta \sim \text{Beta}(1,1)$ (uniform distribution on $[0,1]$) the posterior is a $\text{Beta}(241,946; 251,528)$



Model for gaussian data

Assume that observations come from a gaussian distribution (variance may be known or unknown)

- $Y_1, \ldots, Y_n \sim iid\mathcal{N}(\mu, \sigma^2)$ conditionally to parameter(s) value(s)
- we distinguish three cases
  - μ parameter, σ^2 known;
  - μ known, σ^2 parameter;
  - μ, σ^2 parameters;
- the likelihood L $(= L(\mu)$ or $L(\mu, \sigma^2))$ is

$$L \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]$$

Gaussian model; σ^2 known I

Likelihood:

$$\begin{aligned} L(\mu) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y_i-\mu)^2}{2\sigma^2}\right] \\ &\propto \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right] \end{aligned}$$

Assume a gaussian prior on μ,

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

The posterior distribution is then

$$\pi(\mu|y) \propto L(\mu)\pi(\mu)$$

Gaussian model; σ^2 known II

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

Gaussian model; σ^2 known III

$$\begin{aligned} \pi(\mu|y) &\propto L(\mu)\pi(\mu) \\ &\propto \exp\left(-\frac{1}{2(\sigma^*)^2}(\mu-\mu^*)^2\right) \qquad [\mathcal{N}(\mu^*,(\sigma^*)^2)] \end{aligned}$$

That is, we obtain a gaussian posterior distribution with parameters μ^* and $(\sigma^*)^2$ which are functions of the prior distribution parameters and the data:

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

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

Gaussian model; σ^2 known IV

The posterior mean is a weighted average of the prior mean and the ML estimate, where the weights are the inverse of their respective variances (i.e. precisions)

$$\mu^* = \mu^*_{n,\sigma_0} = \frac{\frac{n}{\sigma^2}\bar{y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\frac{1}{V(\bar{y})}\bar{y} + \frac{1}{V(\mu)}\mu_0}{\frac{1}{V(\bar{y})} + \frac{1}{V(\mu)}}$$

- $\mu^*_{n,\sigma_0} \underset{n\to\infty}{\longrightarrow} \bar{y}$, in large samples the ML estimate dominates the prior
- $\mu^*_{n,\sigma_0} \underset{\sigma_0\to 0}{\longrightarrow} \mu_0$, the weight of the prior mean on the posterior is greater for sharp priors.

Gaussian model; σ^2 known V

Interestingly, the posterior mean can be written as

$$\mu^* = \mu^*_{n,\sigma_0} = \mu_0 + (\bar{y} - \mu_0)\frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2} = \bar{y} - (\bar{y} - \mu_0)\frac{\sigma^2}{\sigma^2 + n\sigma_0^2}$$

i.e., the posterior mean is equal to the prior mean plus an adjustment that is proportional to the difference between the sample mean and the prior mean. Equivalently, the posterior mean is equal to the sample mean minus an adjustment that is proportional to the difference between the sample mean and the prior mean.
The inverse of the posterior variance, i.e. the posterior precision, is the sum of the prior precision and the precision of ML estimator

$$(\sigma^*)^2 = (\sigma^*_{n,\sigma_0})^2 = \left(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\right)^{-1} = \left(\frac{1}{V(\bar{y})} + \frac{1}{V(\mu)}\right)^{-1}$$

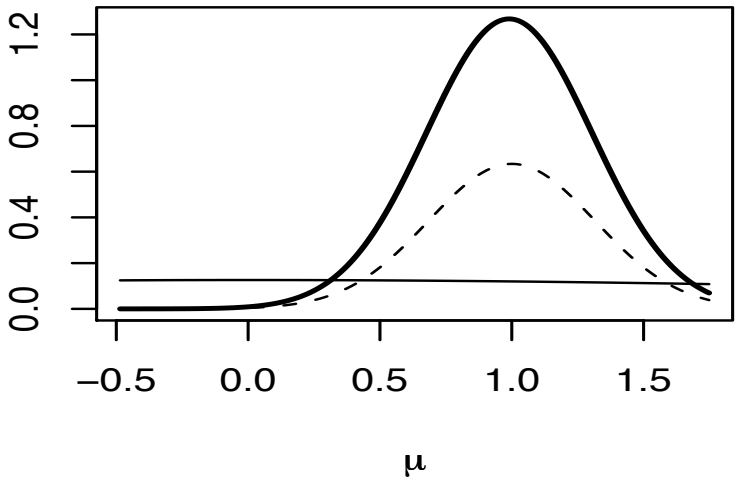
Gaussian model; σ^2 known VI

- $\sigma^*_{n,\sigma_0} \underset{n\to\infty}{\longrightarrow} 0$, in large samples the posterior will be more and moreconcentrated (around the ML estimate).
- $\sigma^*_{n,\sigma_0} \underset{\sigma_0\to 0}{\longrightarrow} 0$, given the sample, the sharper the prior the less dispersed the posterior (around the prior mean!!)

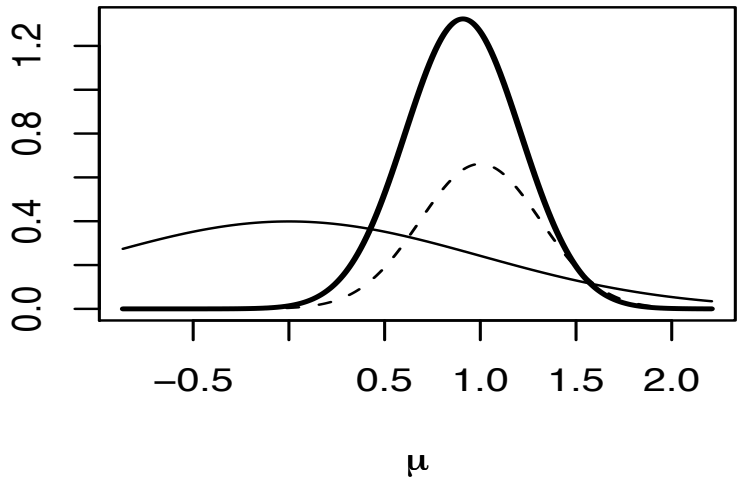
The effects of varying the prior and the number of observations are best seen in the following pictures (notice that there is a mistake in the pictures legend!!!)

Effect of prior

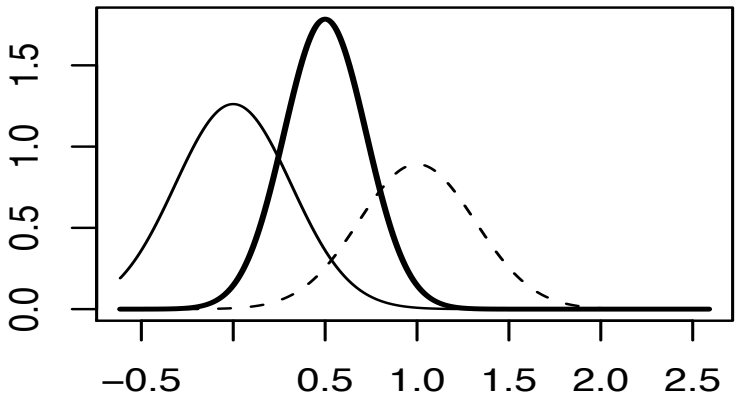
$\mu_0 = 0; \sigma_0^2 = 0.1$; n=10; $\sigma^2 = 1$



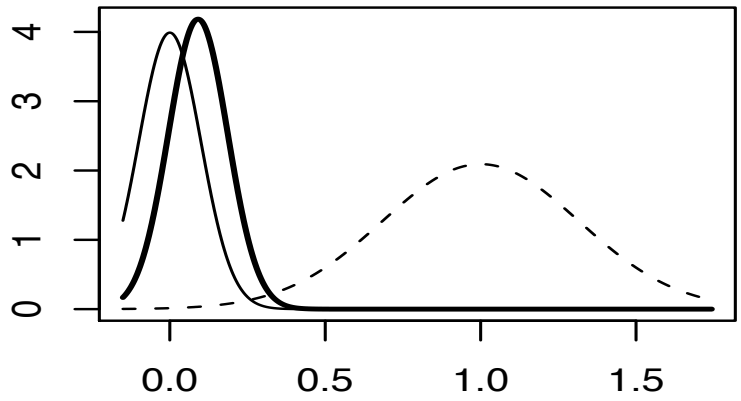
$\mu_0 = 0; \sigma_0^2 = 1$; n=10; $\sigma^2 = 1$



$\mu_0 = 0; \sigma_0^2 = 10$; n=10; $\sigma^2 = 1$

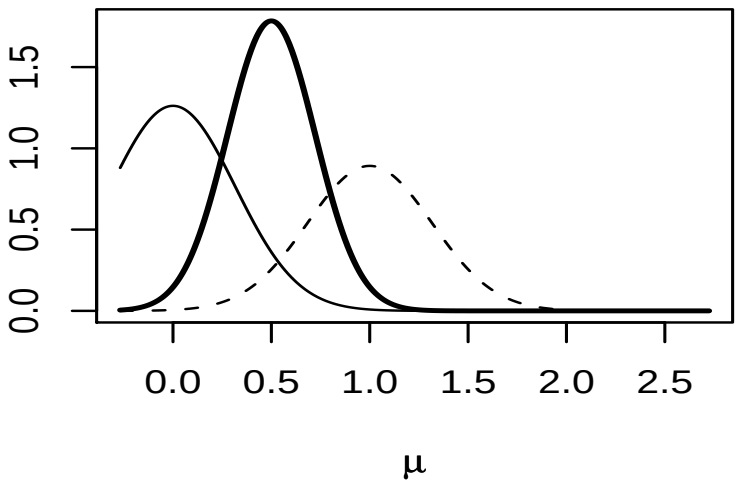


$\mu_0 = 0; \sigma_0^2 = 100$; n=10; $\sigma^2 = 1$

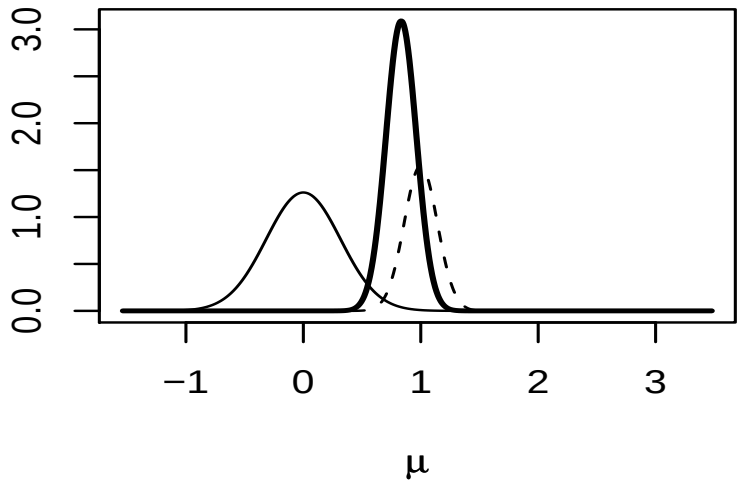


Effect of likelihood

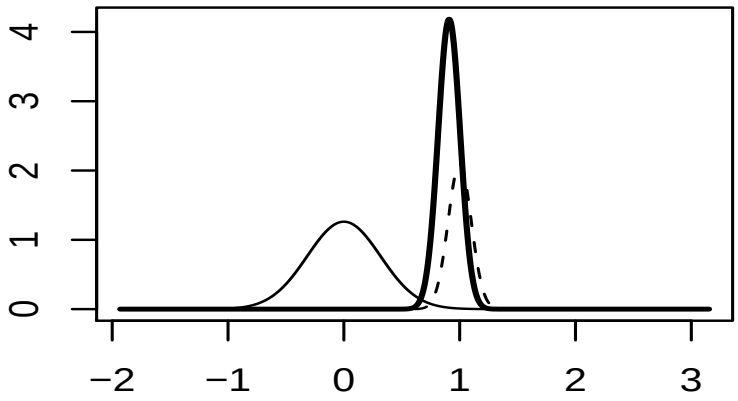
$\mu_0 = 0$; $\sigma_0^2 = 10$; n=10; $\sigma^2 = 1$



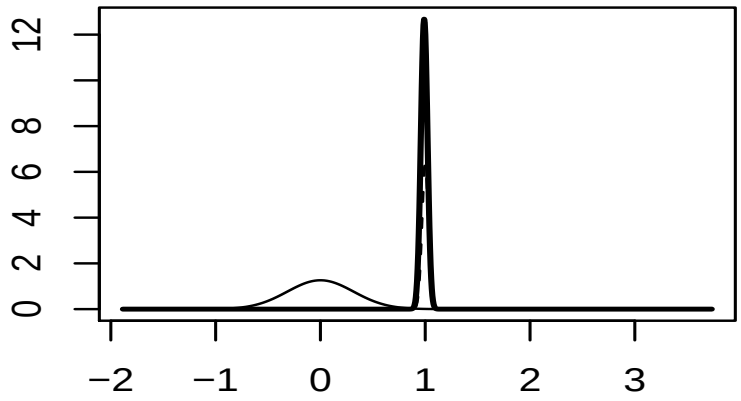
$\mu_0 = 0$; $\sigma_0^2 = 10$; n=50; $\sigma^2 = 1$



$\mu_0 = 0$; $\sigma_0^2 = 10$; n=100; $\sigma^2 = 1$



$\mu_0 = 0$; $\sigma_0^2 = 10$; n=1000; $\sigma^2 = 1$



Note 1: noninformative prior distribution

The Normal distribution is only one possible distribution for μ, any other distribution with support the real line is suitable. An extreme alternative is a constant prior, which is interesting for the results that are obtained.

- Let $\pi(\mu) \propto$ constant (it's an improper prior, because it is not a distribution)
- The posterior is

$$\begin{aligned} \pi(\mu|y) &\propto L(\mu)\pi(\mu) \\ &\propto \exp\left[-\frac{n}{2\sigma^2}(\bar{y}-\mu)^2\right] \end{aligned}$$

that is

- the posterior is $\mu \sim \mathcal{N}(\bar{y}, \frac{\sigma^2}{n})$
- the posterior mean is equal to the ML estimate
- the result of the frequentist approach: $\bar{y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Note 2: updates I

- Let, a priori, $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$
- Given an observation $y_1 \sim \mathcal{N}(\mu, \sigma^2)$ the posterior is

$$(\mu|y_1) \sim \mathcal{N}(\mu_1, \sigma_1^2)$$

where $\mu_1 = \frac{\frac{\mu_0}{\sigma_0^2} + \frac{y_1}{\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}$ and $\sigma_1^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}$

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

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

so

$$(\mu|y_1, y_2) \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

where $\mu_2 = \frac{\frac{\mu_1}{\sigma_1^2} + \frac{y_2}{\sigma^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma^2}}$ and $\sigma_2^2 = \frac{1}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma^2}}$

Note 2: updates II

- Notice 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 posterior is obtained either updating the information in two steps as above starting from the prior $\mathcal{N}(\mu_0, \sigma_0^2)$ or updating it directly using the likelihood of the pair (y_1, y_2).

Note 2: updates III

Sequential updates: Let $f(y|\theta)$ be the model and $\pi(\theta)$ the prior, the posterior is then

$$\pi(\theta|y) \propto \pi(\theta)L(\theta; y)$$

If a further observation x, independent of y and distributed according to $f(x|\theta)$, becomes available, the posterior can be written as

$$\pi(\theta|y, x) \propto \pi(\theta)L(\theta; y, x)$$

Given that x and y conditional on θ are independent we can write

$$\begin{aligned} \pi(\theta|y, x) &\propto \pi(\theta)L(\theta; x)L(\theta; y) \\ &\propto \pi(\theta|y)L(\theta; x) \end{aligned}$$

i.e. the posterior can be also obtained by combining the prior distribution $\pi(\theta|y)$ and the likelihood for x.

Note 3: sufficient statistics I

Note that, given the prior distribution $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ the same posterior distribution is obtained with one of the following information

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

where the latter represents the sufficient statistics (together with n) 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 likelihood of the whole sample.

Note 3: sufficient statistics II

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

$$\begin{aligned} & & L(\theta; y) &\propto L(\theta; t(y)) \\ \textbf{Factorization theorem}: & & f(y|\theta) &= h(y)g(t(y); \theta) \end{aligned}$$

Hence

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

Bayesian and classical interval estimation I

- We have seen that the information in the posterior distribution can be summarized by its posterior expectation and standard deviation;
- These roughly correspond to the point estimate and its standard error in classical inference (although the interpretation is a bit different!).
- Given that θ is a random variable, it is natural to think at an analogue of confidence intervals;
- This analogue is called credibility interval.
- there is a great difference in interpretation, where credibility interval is a more natural object. According to some authors, most of non statisticians actually interpret confidence intervals as if they were credibility intervals.

Bayesian and classical interval estimation II

Classical interval estimate (confidence interval) An interval is associated to the sample y such that with a confidence level $1-\alpha$, it contains the true value of the parameter.
Interpretation: if $N = 100$ samples were observed and for each of them a $1-\alpha$ confidence interval were obtained, on average $100(1-\alpha)$ of them would contain the true value of the parameter.

Bayesian interval estimate (credibility interval) An interval is associated to the sample y such that it contains the true value of the parameter with probability $1-\alpha$. Interpretation: straightforward and intuitive (we don't need to think to a large number of hypothetical samples).

Bayesian and classical interval estimation III

Let us compare the formal definitions

Def: confidence interval

A confidence interval for θ is a pair of statistics $L(Y), U(Y) \in \Theta$ such that

$$P(L(Y) \leq \theta \leq U(Y)) \geq 1 - \alpha \quad \forall \theta$$

where the probability is with respect to the distribution of Y.

Def: credibility interval

A credibility interval for θ is a pair of statistics $L(Y), U(Y) \in \Theta$ such that

$$P(L(Y) \leq \theta \leq U(Y)) \geq 1 - \alpha$$

where the probability is with respect to the distribution of θ.

Bayesian and classical interval estimation IV

Confidence interval:

(assume L^{-1}, U^{-1} exist)

$$\begin{aligned} P(L(Y) \leq \theta \leq U(Y)) &= P(U^{-1}(\theta) \leq Y \leq L^{-1}(\theta)) \\ &= \int_{U^{-1}(\theta)}^{L^{-1}(\theta)} p(y|\theta)dy \end{aligned}$$

Credibility interval:

$$P(L(Y) \leq \theta \leq U(Y)) = \int_{L(Y)}^{U(Y)} \pi(\theta|y)d\theta$$

Credibility intervals

Given a distribution for θ, $\pi(\theta|y)$ there is not a unique interval satisfying the condition

$$P(L(Y) \leq \theta \leq U(Y)) = \int_L^U \pi(\theta|y)d\theta = 1 - \alpha$$

the easiest choice is to set L and U equal to the quantiles $\alpha/2$ and $1 - \alpha/2$ of $\pi(\theta|y)$, that is, such that

$$\int_{-\infty}^{L} \pi(\theta|y)d\theta = \int_U^{+\infty} \pi(\theta|y)d\theta = \alpha/2$$

this interval satisfies the condition but is not in general the smallest one.

Credibility intervals: HPD (Highest Posterior Density) I

A better (meaning smaller) interval is defined as **Highest Posterior Density** (HPD). The highest posterior density credibility region is a set $C \subset \Theta$ such that

$$P(\theta \in C) = 1 - \alpha$$

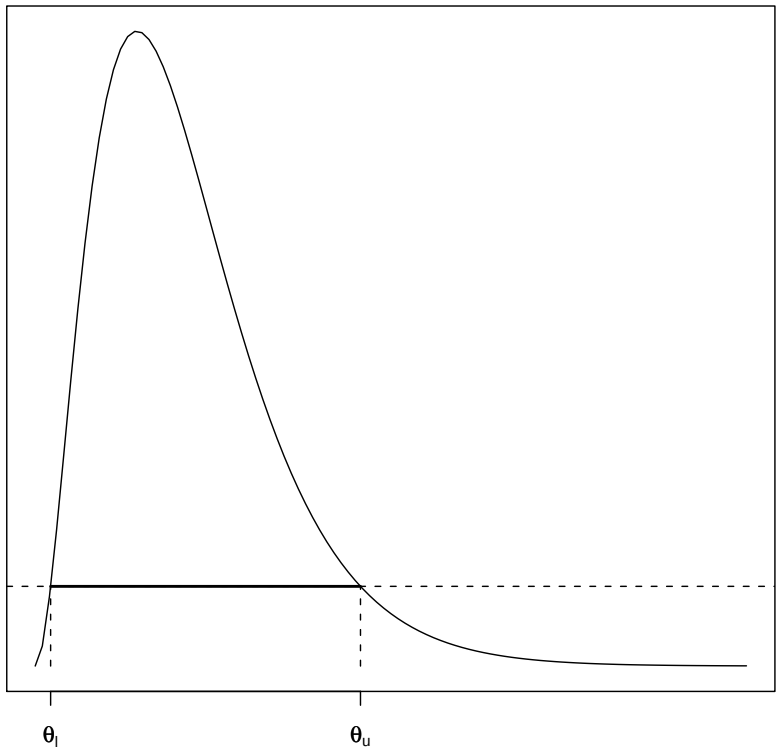
and

$$\pi(\theta_1|y) > \pi(\theta_2|y)$$

if $\theta_1 \in C$ and $\theta_2 \notin C$

Credibility intervals: HPD (Highest Posterior Density) II

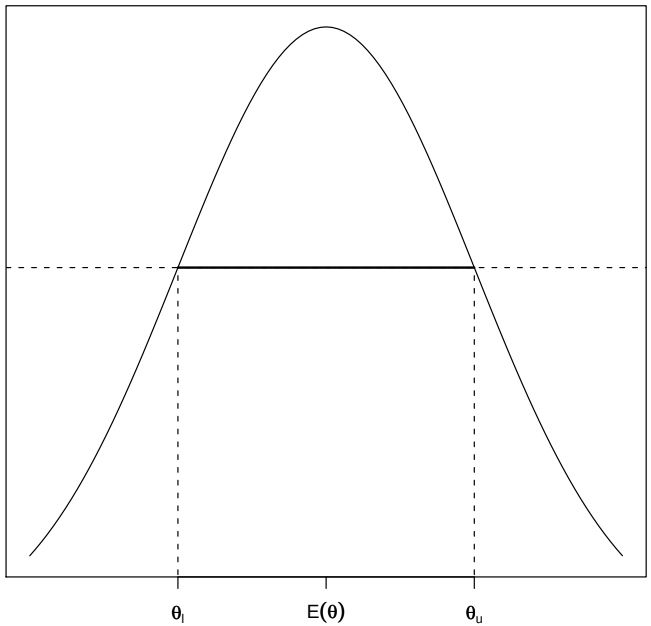
Given $\pi(\theta|y)$ the HPD interval C is obtained including the values of θ of highest density



Credibility intervals: HPD (Highest Posterior Density) III

unimodal symmetric posterior

$$C = \left[F^{-1}(\alpha/2), F^{-1}(1-\alpha/2)\right]$$



Credibility intervals: HPD (Highest Posterior Density) IV

For example in the normal-normal model, since

$$\pi(\theta|y) = \mathcal{N}(\mu^*, \sigma^{2*})$$

The interval defined by the quantiles is centered on the mean and has extremes

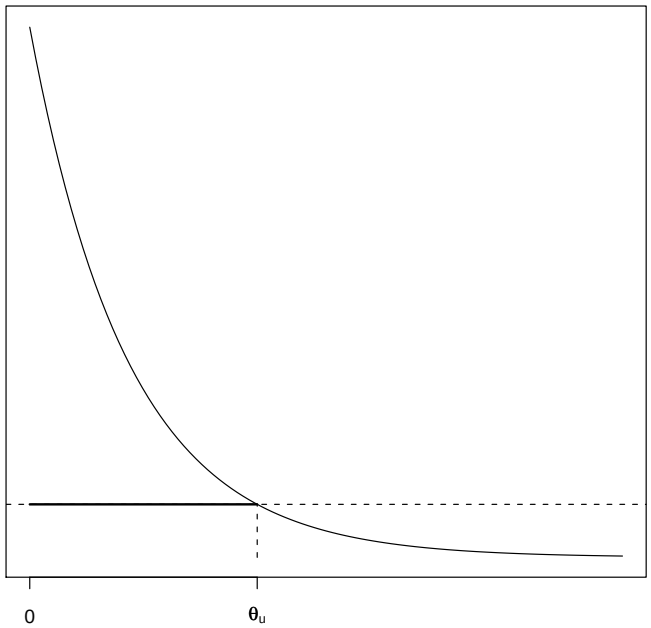
$$C = \left[\mu^* - \sigma^* \Phi^{-1}\left(1 - \frac{\alpha}{2}\right), \mu^* + \sigma^* \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\right]$$

(This is true both for the conjugate and the noninformative prior.) It's straightforward to show (think at the graph of the gaussian density), that this interval is HPD too.

Credibility intervals: HPD (Highest Posterior Density) V

monotone posterior

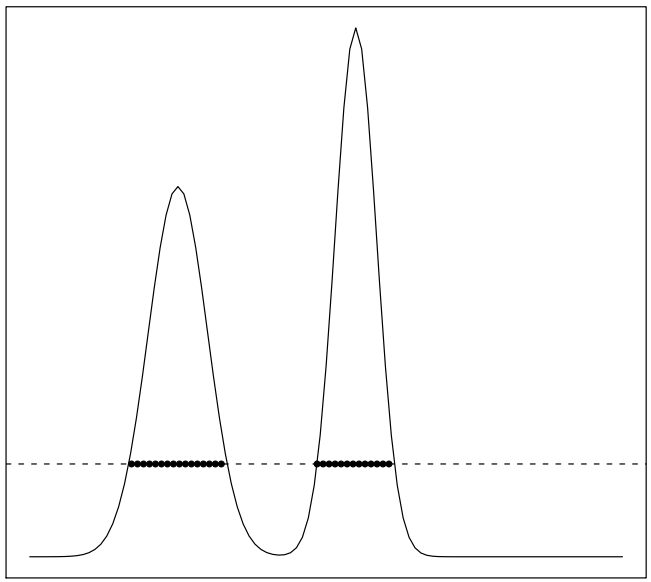
$$C = \left[0, F^{-1}(1-\alpha)\right]$$



Credibility intervals: HPD (Highest Posterior Density) VI

multimodal posterior

the HPD region is not necessarily an interval but can be the union of disjoint intervals



Finding the HPD region

For a unimodal posterior (not necessarily symmetric) we may use an algorithm to find the interval: start from $k_m = 0$, $k_M = \max_\theta \pi(\theta|y)$ then at step i

1. $k_i = (k_m + k_M)/2$
2. determine $C = \{\theta|\pi(\theta|y) > k_i\}$
3. compute $I = \int_C \pi(\theta|y)d\theta$
4. if $I < 1-\alpha$ $k_m \leftarrow k_i$ (shorter interval) return to 1
   if $I > 1-\alpha$ $k_M \leftarrow k_i$ (longer interval), return to 1
   if $I = 1-\alpha$ STOP, C is the solution

Gaussian model; σ^2 unknown I

Likelihood:

$$\begin{aligned} L(\sigma^2) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\ &\propto (\sigma^2)^{-n/2} \exp\left[-\frac{n}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \end{aligned}$$

Gaussian model; σ^2 unknown II

Assume an *inverse gamma* prior on σ^2,

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

denoted as

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

with support the positive real line, where $\gamma > 0$ is called the *shape parameter* and $\delta > 0$ the *scale parameter*. Notice that in this parametrization $\nu = 2\gamma$ corresponds to the *degrees of freedom* of the distribution. The mean of the r.v. is $\frac{\delta}{\gamma-1}$ for $\gamma > 1$ and the variance is $\frac{\delta^2}{(\gamma-1)^2(\gamma-2)}$ for $\gamma > 2$ (See pdf graph)

Gaussian model; σ^2 unknown III

Assume an inverse gamma prior on σ^2,

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

The posterior distribution is then

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

$$\begin{aligned}
\pi(\sigma^2|y) &\propto (\sigma^2)^{-n/2} \exp\left[-\frac{n}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right](\sigma^2)^{-\gamma-1}e^{-\delta/\sigma^2} \\
&\propto (\sigma^2)^{-n/2-\gamma-1} \exp\left[-\frac{1}{\sigma^2}\left[\frac{n}{2}\sum_{i=1}^{n}(y_i-\mu)^2+\delta\right]\right]
\end{aligned}$$

that is, an inverse gamma with parameters $\gamma + n/2$ and $\frac{n}{2}\sum_{i=1}^{n}(y_i-\mu)^2 + \delta$.

Gaussian model; σ^2 unknown IV

It might be convenient to reparameterize the model in terms of the precision $\tau = 1/\sigma^2$, so the likelihood is

$$L(\tau) \quad \propto \quad (\tau)^{n/2} \exp\left[-\frac{n}{2}\tau \sum_{i=1}^{n}(y_i - \mu)^2\right]$$

The conjugate prior assumption is then $\tau \sim \text{Gamma}(\gamma, \delta)$ i.e.

$$\pi(\tau) \propto \tau^{\gamma-1} e^{-\delta\tau}$$

with support the positive real line, where $\gamma > 0$ is called the *shape parameter* and $\delta > 0$ the *inverse scale parameter* or *rate parameter* . Notice that in this parametrization $\nu = 2\gamma$ corresponds to the *degrees of freedom* of the distribution. The mean of the r.v. is $\frac{\gamma}{\delta}$ and the variance is $\frac{\gamma}{\delta^2}$

Gaussian model; σ^2 unknown V

An alternative common parameterization of the Gamma distribution used in MATLAB and R is the following: $\tau \sim \text{Gamma}(k, \theta)$ where

$$\pi(\tau) \propto \tau^{k-1} e^{-\frac{\tau}{\theta}}$$

with support the positive real line, where $k > 0$ is called the *shape parameter* and $\theta > 0$ the *scale parameter*. Notice that in this parameterization $\nu = 2\gamma$ corresponds to the *degrees of freedom* of the distribution. So, $k = \gamma$ and $\theta = \frac{1}{\delta}$. The mean of the r.v. is $k\theta$ and the variance is $k\theta^2$.
(See pdf graph)

Gaussian model; σ^2 unknown VI

The relationship between the Gamma(k, θ) and the IGamma(γ, δ) is the following:

- if $x \sim \text{Gamma}(k, \theta)$, i.e.

$$p(x) \propto x^{k-1} e^{-\frac{x}{\theta}}$$

- then $y = \frac{1}{x} \sim \text{IGamma}(k, \frac{1}{\theta})$, i.e.

$$p(y) \propto y^{-k-1} e^{-\frac{\frac{1}{\theta}}{y}}$$

Gaussian model; σ^2 unknown VII

Notice that the result above implies that using the first parameterization of the gamma, Gamma(γ, δ) i.e.

$$p(x) \propto x^{\gamma-1} e^{-\delta x}$$

- $y = \frac{1}{x} \sim$ IGamma(γ, δ), i.e.

$$p(y) \propto y^{-\gamma-1} e^{-\frac{\delta}{y}}$$

Gaussian model; σ^2 unknown VIII

So, using the first parameterization $\tau \sim \text{Gamma}(\gamma, \delta)$, the posterior results to be $\text{Gamma}(n/2 + \gamma, \frac{n}{2}\sum_{i=1}^{n}(y_i - \mu)^2 + \delta)$

$$\begin{aligned}
\pi(\tau|y) &\propto \tau^{n/2} \exp\left[-\frac{n}{2}\tau \sum_{i=1}^{n}(y_i - \mu)^2\right] \tau^{\gamma-1} e^{-\delta\tau} \\
&\propto \tau^{n/2+\gamma-1} \exp\left[-\tau\left(\frac{n}{2}\sum_{i=1}^{n}(y_i - \mu)^2 + \delta\right)\right]
\end{aligned}$$

Two-parameters models

Suppose that a model depends upon two parameters θ_1, θ_2

$$p(y|\theta_1, \theta_2)$$

then the prior is a bivariate distribution $\pi(\theta_1, \theta_2)$ and the posterior is a bivariate distribution as well

$$\pi(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)\pi(\theta_1, \theta_2)$$

Often one of the parameters, say θ_2 is a nuisance parameter, in which case we may be interested in the marginal posterior for θ_1, which is obtained as

$$\pi(\theta_1|y) = \int \pi(\theta_1, \theta_2|y)d\theta_2 \propto \int p(y|\theta_1, \theta_2)\pi(\theta_1, \theta_2)d\theta_2$$

Gaussian model: μ, σ^2 unknown, likelihood I

The likelihood is

$$
\begin{aligned}
L(\mu, \sigma^2) &\propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \\
&\propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_j (y_j - \mu)^2\right] \\
&\propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_j (y_j - \bar{y} + \bar{y} - \mu)^2\right] \\
&\propto (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{y} - \mu)^2\right)\right]
\end{aligned}
$$

where the last row is a consequence of

$$
\sum_j (y_j - \bar{y} + \bar{y} - \mu)^2 = \underbrace{\sum_j (y_j - \bar{y})^2}_{=(n-1)s^2} + n(\bar{y} - \mu)^2 + \underbrace{2(\bar{y} - \mu)\sum_j (y_j - \bar{y})}_{=0}
$$

Gaussian model: μ, σ^2 unknown, likelihood II

It is convenient to reparametrize the model writing $\tau = 1/\sigma^2$, so the likelihood is

$$L(\mu, \tau) \quad \propto \quad \tau^{n/2} \exp\left[-\frac{\tau}{2}\left((n-1)s^2 + n(\bar{y}-\mu)^2\right)\right]$$

the parameter τ is the precision.

Gaussian model: σ^2 unknown, prior

Since there are two parameters, the prior distribution must clearly be bivariate we can use the normal-gamma distribution, that is

- $\mu|\tau \sim \mathcal{N}(\alpha, \frac{1}{\xi\tau})$
- $\tau \sim \text{Gamma}(\gamma, \delta)$

so

$$\begin{aligned}
\pi(\mu, \tau) &= \pi(\mu|\tau)\pi(\tau) \\
&\propto (\xi\tau)^{1/2} \exp\left[-\frac{\xi\tau}{2}(\mu-\alpha)^2\right] \tau^{\gamma-1} \exp\left[-\delta\tau\right] \\
&\propto \xi^{1/2}\tau^{\gamma-1/2} \exp\left[-\tau\left(\frac{\xi}{2}(\mu-\alpha)^2 + \delta\right)\right]
\end{aligned}$$

Gaussian: μ, σ^2 unknown, a posteriori I

$$
\begin{aligned}
\pi(\mu, \tau | y) &\propto L(\mu, \tau)\pi(\mu, \tau) \\
&\propto \tau^{n/2} \exp\left[-\frac{\tau}{2}\left((n-1)s^2 + n(\bar{y}-\mu)^2\right)\right] \times \\
&\qquad \times \xi^{1/2}\tau^{\gamma-1/2} \exp\left[-\tau\left(\frac{\xi}{2}(\mu-\alpha)^2 + \delta\right)\right] \\
&\propto \tau^{\frac{n}{2}+\gamma-\frac{1}{2}} \exp\left[-\frac{\tau}{2}(n(\bar{y}-\mu)^2 + \xi(\mu-\alpha)^2)\right] \times \\
&\qquad \times \exp\left[-\frac{(n-1)\tau}{2}s^2 - \delta\tau\right] \\
&\propto \tau^{\frac{n}{2}+\gamma-\frac{1}{2}} \exp\left[-\frac{\tau}{2}(n+\xi)\left(\mu - \frac{n\bar{y}+\xi\alpha}{n+\xi}\right)^2\right] \times \\
&\qquad \times \exp\left[-\tau\left(\frac{(n-1)}{2}s^2 + \delta + \frac{n\xi}{2(n+\xi)}\right)\right]
\end{aligned}
$$

Gaussian: μ, σ^2 unknown, a posteriori II

$$\begin{aligned}
\pi(\mu, \tau|y) &\propto \tau^{\frac{n}{2}+\gamma-\frac{1}{2}} \exp\left[-\frac{\tau}{2}(n+\xi)\left(\mu - \frac{n\bar{y}+\xi\alpha}{n+\xi}\right)^2\right] \times \\
&\quad \times \exp\left[-\tau\left(\frac{(n-1)}{2}s^2 + \delta + \frac{n\xi}{2(n+\xi)}\right)\right] \\
&\propto \tau^{\gamma^*-\frac{1}{2}} \exp\left[-\frac{\tau}{2}\xi^*(\mu-\alpha^*)^2\right] \exp\left[-\tau\delta^*\right]
\end{aligned}$$

the posterior is then a normal-gamma with parameters $\alpha^* = \frac{n\bar{y}+\xi\alpha}{n+\xi}$; $\xi^* = n + \xi$
$\delta^* = \frac{(n-1)}{2}s^2 + \delta + \frac{n\xi}{2(n+\xi)}$; $\gamma^* = \gamma + \frac{n}{2}$

Gaussian: μ, σ^2 unknown, a posteriori III

One can draw conclusions directly from the bivariate posterior distribution (for instance, a posterior credibility region may be obtained for the pair); however, if we are only interested in the mean μ the following results are relevant:

- conditionally to a value for the precision τ,

$$\pi(\mu|\tau, y) = \mathcal{N}\left(\alpha^*, \frac{1}{\xi^*\tau}\right)$$

- Marginally $\pi(\mu|y)$ is a Student's t distribution with degrees of freedom $2\gamma^* = 2\gamma + n$ and mean $\alpha^* = \frac{n\bar{y}+\xi\alpha}{n+\xi}$

Outline of the talk

Prediction I

Given

- a model for y, $p(y|\theta)$;
- a posterior for θ, $\pi(\theta|y)$.

Suppose you want to predict a future value of y (or any transformation $g(y)$), let's call it y^{new}, consider

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

For exchangeable observations (i.e. independent conditional on θ), this becomes

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

Prediction of y^{new} is based on the so called *predictive density*,

$$p(y^{new}|y) = \int p(y^{new}, \theta|y)d\theta$$

Prediction II

As a simple example consider a gaussian model with known variance

- model y, $p(y|\theta) = \mathcal{N}(\theta, \sigma^2)$, σ^2 known;
- posterior for θ, $\pi(\theta|y) = \mathcal{N}(\mu^*, (\sigma^2)^*)$.

then, using the theorem above

$$y^{\text{new}}|y \sim \mathcal{N}(\mu^*, \sigma^2 + (\sigma^2)^*)$$

which has an easy interpretation: predicting y^{new}, the uncertainty due to the parameter, $(\sigma^2)^*$ adds to the uncertainty due to the model, σ^2.
It is interesting to rewrite the above formula as

$$p(y^{\text{new}}|y) = \int p(y^{\text{new}}|\theta, y)\pi(\theta|y)d\theta$$

the distribution of y^{new} is then the mixture of the conditional distributions of y^{new} given θ, using as mixing distribution the posterior of θ.

Outline of the talk

Multivariate normal distribution

Let $y \in \mathbb{R}^d$ be a random vector and assume

$$y|\mu, \Sigma \sim \mathcal{N}(\mu, \Sigma),$$

then its pdf is given by

$$p(y|\mu, \Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]$$

where μ is the (vector) mean and Σ the covariance matrix (whereas Σ^{-1} is called the precision matrix).

Outline of the talk

Multivariate Student's t distribution

Let $y \in \mathbb{R}^d$ be a random vector and assume

$$y|\mu, \Sigma, \nu \sim t(\mu, \Sigma, \nu),$$

then its pdf is given by

$$p(y|\mu, \Sigma, \nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \pi^{-d/2} \nu^{\frac{\nu}{2}} |\Sigma|^{-1/2} \left[\nu + (y-\mu)^T \Sigma^{-1} (y-\mu)\right]^{-\frac{\nu+d}{2}}$$

where $\mu \in \mathbb{R}^d$ is the mean, for $\nu > 1$, Σ is a PDS dxd matrix called scale matrix, and $\nu > 0$ represents its degrees of freedom. Notice that $Var(y) = \frac{\nu}{\nu-2}\Sigma$, for $\nu > 2$.