Bayesian Inference

(An essential introduction - part 2)

L. Egidi

Fall 2021

University of Trieste

Bayesian model

- A **prior distribution** is defined on the parameter θ

$$\pi(\theta).$$

- We assume an i.i.d. sample $y = (y_1, y_2, \ldots, y_n)$ is obtained from a distribution belonging to a family indexed by θ, and

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

  is then proportional to the likelihood function.

- Bayes theorem allows us to combine prior information and likelihood to give the **posterior distribution**

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

The posterior distribution then sums up all the information (and the uncertainty) about the parameter θ. Inference on the parameter amounts at studying and appropriately summarizing the posterior.

Again on the proportion of seeds

Consider again the example of the estimation the proportion p of High quality seeds in a box. Recall that for inference on p we assumed that:

- $\pi(p)$ is $Beta(\alpha, \beta)$, the prior
- $f(y_i|p)$ are $Be(p)$ and $f(y|p) \propto Bi(n, p)$ is the likelihood

(where $y = \sum_i^n y_i$ and y_i are i.i.d).

- $\pi(p|y)$, the posterior, is $Beta(y + \alpha, \beta + n - y)$.

In our example, $n = 30, y = 23$, the prior is $Beta(7, 4)$.

Then the posterior is a $Beta(30, 11)$

As stated the posterior distribution summarizes what we know about the parameter combining prior knowledge and experimental data.

So inference on p derives from the analysis of this distribution. And we will use it to illustrate the procedures for point and interval estimation

Point estimation of the the proportion of seeds

If we want to select a single value as a point estimate of p, let say $\hat{p}$, we are back to a classical problem: how to select a single number to summarize a distribution $\pi(p)$.

Classical solutions are:

- the expected value $E(p|y)$ of the posterior distribution of the rv p, $E(p|y) = \int_0^1 p\pi(p|y)dp$;
- the median Me of the posterior distribution, $Me : \int_0^{Me} \pi(p|y)dp = 0.5$;
- the mode Mo of the posterior distribution, *i.e.*, the value of p for which $\pi(p|y)$ is maximum.

One can choose one of these as point estimate and, provided that the posterior is unimodal, they provide an appropriate synthesis.

Obviously the three values are equivalent if the posterior is symmetric and unimodal.

Bayes risk

More formally, Bayes estimators can be defined as the quantity that minimizes the posterior expected value of a loss function $L(\theta, \widehat{\theta})$

The quantity $E_{\pi|y}(L(\theta, \widehat{\theta}))$ is called **Bayes risk**.

1. When $L(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$,

i.e. a quadratic loss is used, the quantity that minimizes the Bayes risk is the posterior mean.

2. Posterior median can be also justified as the quantity that minimizes a "linear" loss function, with $a > 0$, defined as $L(\theta, \widehat{\theta}) = a|\theta - \widehat{\theta}|$.

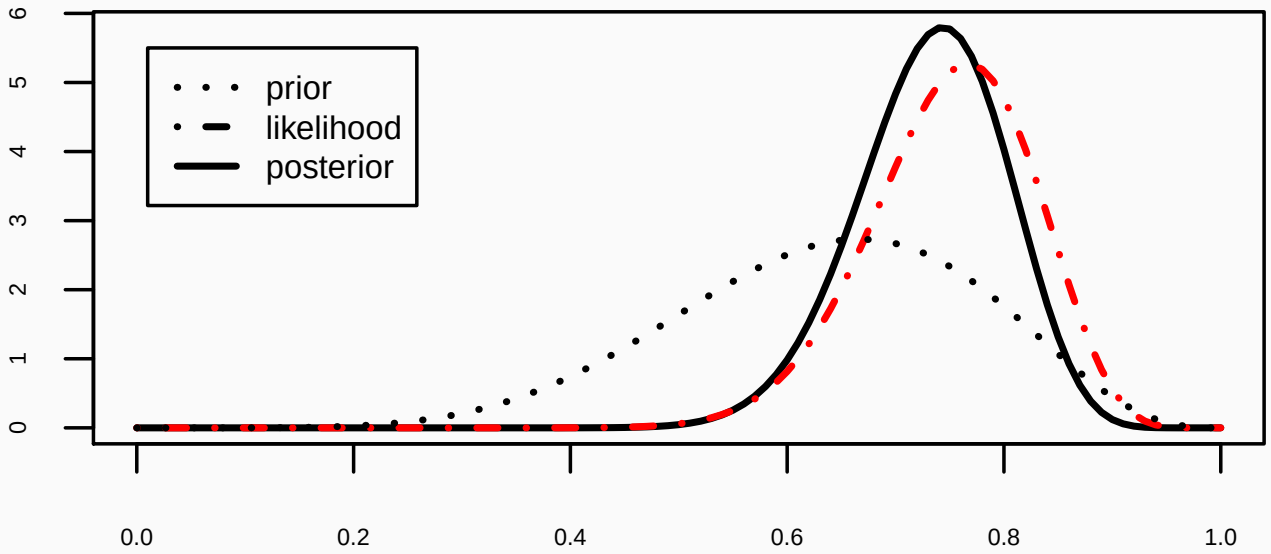
3. Posterior mode (MAP: Maximum A-posteriori Probability) can be justified as the minimizer of the trickier loss function of the form

$$L(\theta, \widehat{\theta}) = \begin{cases} 0 & \text{if } |\widehat{\theta} - \theta| < c, \\ 1 & \text{otherwise,} \end{cases}$$

as c goes to 0.

Estimating the quantity of seeds by the posterior mean

```
curve(dbeta(x,7+23,4+7),xlab="p", ylab="density",lty=1,lwd=2,
      cex.axis=.5, cex.lab=.6, ann=F)
curve(dbeta(x,23+1,7+1),add=TRUE,lty=4,lwd=2, col=2)
curve(dbeta(x,7,4),add=TRUE,lty=3,lwd=2)
legend(.01,5.5,c("prior","likelihood","posterior"), lty=c(3,4,1), lwd=c
```



Using Posterior mean as an estimator of p

Recall that if a rv $p \sim Beta(\alpha, \beta)$ then the prior mean is $E_\pi(p) = \frac{\alpha}{\alpha+\beta}$, being $\pi(\theta|y)$ a Beta distribution, the posterior mean is

$$
\begin{aligned}
&= \frac{\alpha + y}{\alpha + \beta + n} \\
&= \frac{\alpha + \beta}{\alpha + \beta + n} \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \frac{y}{n} \\
&= \frac{\alpha + \beta}{\alpha + \beta + n} E_\pi(p) + \frac{n}{\alpha + \beta + n} \hat{p}_{ML}
\end{aligned}
$$

$$
E(p|y) = \frac{\alpha + \beta}{\alpha + \beta + n} \underbrace{E_\pi(P)}_{\text{prior expectation}} + \frac{n}{\alpha + \beta + n} \underbrace{\hat{p}}_{MLE}
$$

A closer look to posterior distribution

We have seen that 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.

The posterior distribution as a whole is a compromise between the prior and the likelihood, and the likelihood prevails if

- n is large;
- α and β are both close to 1 (the prior is diffuse)

To appreciate the quality of the posterior mean as an estimate we can look at the posterior variance (or at standard deviation)

$V(\theta) = \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)}$ that for large n is $\approx \frac{1}{n}\frac{y}{n}(1-\frac{y}{n})$

Estimating the score for Tripadvisor ratings

Tripdavisor uses a formula for calculating and comparing the ratings of restaurants by its users. It derives from using a weighted mean that relies upon the Bayesian idea

The following formula was the base to calculate $W = \frac{Rv+Cm}{v+m}$

where:

- W = weighted rating
- R = average rating (stars) for the restaurant (1 to 5) - *the likelihood*
- v = number of votes/ratings for the restaurant = (votes)
- m = weight given to the prior estimate (in this case, the number of votes for a stable average rating)
- C = the mean vote across the whole pool - *the prior*

W is just the weighted arithmetic mean of R and C with weights v and m.

As the number of ratings surpasses m, the confidence of the average rating surpasses the confidence of the prior knowledge, and the weighted bayesian rating W approaches a straight average R.

Bayesian interval estimation

The posterior distribution can be summarized by posterior expectation and variance:

- these roughly correspond to 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 big difference in interpretation where credibility intervals are much more natural and close to common sense.
- Most non statisticians actually *interpret confidence intervals as if they were credibility intervals*.

Classical confidence interval vs credibility interval

Classical interval estimate (confidence interval)

An interval is associated to the sample y such that with a confidence level $1-\alpha$, contains the true value of the parameter.

Interpretation: if N 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.

An interval is associated to the sample y such that it **contains the true value of the parameter with probability** $1-\alpha$.

Bayesian interval estimate (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,

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

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, generally, the smallest one.

HPD (High Posterior Density) region

A better (smaller) interval is defined as

High posterior density (HPD)

The high posterior density credibility region is a set $C \subset \Theta$ such that

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

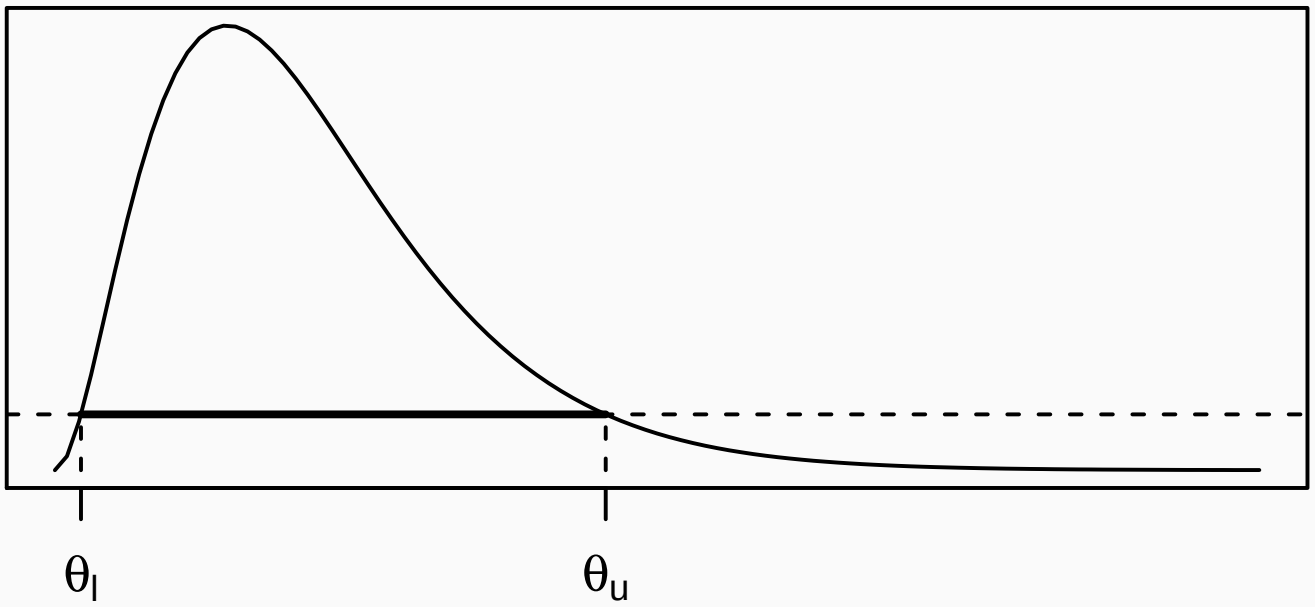
and

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

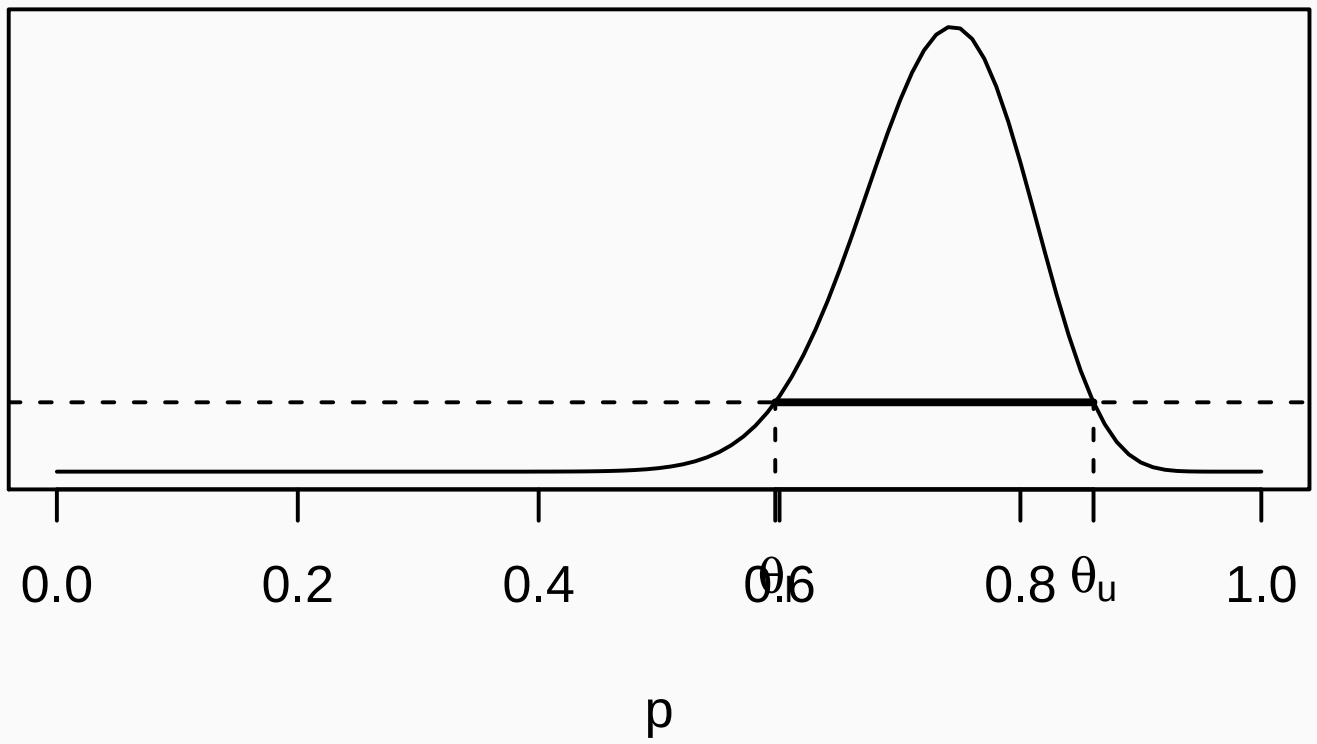
if $\theta_1 \in C$ and $\theta_2 \notin C$.

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

HPD region

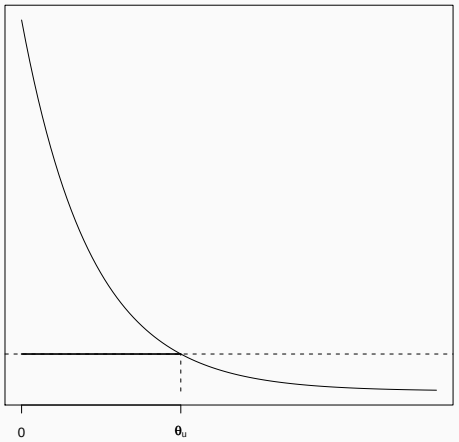


.95 HPD region for p in the seeds example

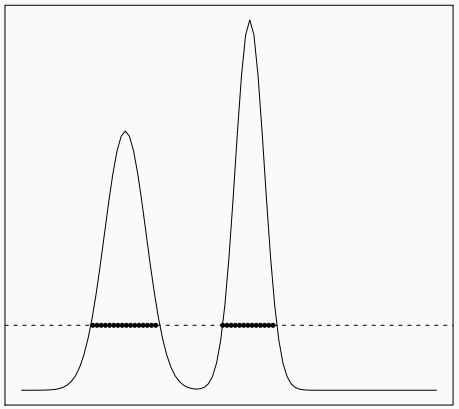


Special cases

monotone posterior



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 $I = \int_C \pi(\theta|y)d\theta$

- if $I < 1 - \alpha$ $\quad$ $k_m \leftarrow k_i$ (shorter interval) return to 1
- if $I > 1 - \alpha$ $\quad$ $k_M \leftarrow k_i$ (longer interval), return to 1
- if $I = 1 - \alpha$ $\quad$ STOP $\quad$ C is the solution

Model for gaussian data

Assume that observations come from a gaussian distribution (variance known)

- $Y_1, \ldots, Y_n \sim iid \quad N(\mu, \sigma^2)$ conditional to parameter(s) value(s).

μ is the parameter, σ^2 is known;

the likelihood L $(= L(\mu))$ is

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

Gaussian model; σ^2 known

Likelihood:

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

Assume a Gaussian prior on μ,

$$\mu \sim N(\mu_0, \sigma_0^2).$$

The posterior distribution is then

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

Gaussian model; σ^2 known

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

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

Gaussian model; σ^2 known

$$\begin{aligned} \pi(\mu|y) &\propto L(\mu)\pi(\mu) \\ &\propto e^{-\frac{1}{2(\sigma^*)^2}(\mu-\mu^*)^2} \times N(\mu^*, (\sigma^*)^2) \end{aligned}$$

that is, we obtain a Gaussian posterior distribution with parameters μ^* and σ^* which are a function of prior distribution's parameters and of 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

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:

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

- $\mu^*_{n,\sigma_0} \underset{n\to\infty}{\longrightarrow} \bar{y}$ as n grows, the ML estimates weights more;
- $\mu^*_{n,\sigma_0} \underset{\sigma_0\to 0}{\longrightarrow} \mu_0$ the more concentrated is the prior distribution, the more the prior mean weights.

It is interesting to write the posterior mean as

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

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

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

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

Gaussian model; σ^2 known

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

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

- $\sigma^*_{n,\sigma_0} \underset{n\to\infty}{\longrightarrow} 0$ as n grows the variance of the posterior diminish
- $\sigma^*_{n,\sigma_0} \underset{\sigma_0\to 0}{\longrightarrow} 0$ also if the variance of the prior is reduced the posterior is more concentrated

Hypoteses testing

Suppose you want to test the Hypothesis

$H_0 : \theta \in \Theta_0; \qquad H_1 : \theta \in \Theta_1$

Θ_0 and Θ_1 form a partition of the parameter space.

The beliefs about the two hypotheses are summarized by the posterior odds ratio

$$\frac{p_0}{p_1} = \frac{P(\theta \in \Theta_0 | y)}{P(\theta \in \Theta_1 | y)} = \frac{\int_{\Theta_0} \pi(\theta|y) d\theta}{\int_{\Theta_1} \pi(\theta|y) d\theta}$$

A measure of the evidence provided by the data in support of H_0 is the **Bayes factor**

$$BF = \frac{\text{posterior odds}}{\text{prior odds}} = \frac{p_0/p_1}{\pi_0/\pi_1},$$

where π_0 and π_1 are respectively $\int_{\Theta_0} \pi(\theta) d\theta$ and $\int_{\Theta_1} \pi(\theta) d\theta$ the probability of the two hypotheses prior observing the data. Note that you can then evaluate the posterior probability that the null hypothesis is true $p_0 = \frac{\pi_0 BF}{\pi_0 BF + 1 - \pi_0}$

Coniugacy

Note that in the two examples considered prior and posterior distributions have the same functional form.

For the seeds example the posterior distribution is a Beta like the prior as well as for the Normal mean example

Likelihood	Prior	Posterior
$L(\theta; y)$	$\pi(\theta)$	$\pi(\theta \mid y)$
Binomial	$Beta(\alpha, \beta)$	$Beta(\alpha + \sum_i y_i, \beta + n - \sum_i y_i)$
Normal	$N(\mu, \sigma^2)$	$N(\mu^*, (\sigma^*)^2)$

This property relates the family of the prior distribution with the likelihood and is called **conjugacy**. Example of conjugate families are:

- Beta prior and Binomial likelihood
- Normal and Normal
- Gamma prior and Poisson likelihood

The prior distribution

The idea of using prior knowledge in Bayesian statistics is a critical issue and it is a new element in comparison with classical statistics.

Formally this knowledge is introduced by specifying a prior distribution that includes information other than what is directly observed in the process of inference.

The most common concerns are about the use of subjective probability in deriving the prior (for instance, by using experts' opinions for eliciting the prior).

For this reasons some relevant topics in Bayesian statistics refer to:

1. the choice of prior distributions that are diffuse (non informative) in order to give more (or exclusively) weight to experimental data or to obtain results that are consistent with results from likelihood based inference
2. the analysis of the sensitivity of the inference to the alternative choice of the prior (Bayesian robustness)

Objections on the use of prior distributions

One (non-Bayesian statistician) could argue that if I specify a subjective prior distribution, since I can chose any distribution, I can also modify the results and obtain whatever conclusion I want. The result could then be manipulated, it is subjective and hence not scientific.

Counter-objections include

- classical procedures are also subjective, for example in the specification of the model;
- the relevance of the prior distribution is limited and tends to vanish if the sample size increases;
- actually, the information conveyed by the data would outweigh the information in the prior for any reasonable specification;
- a possible compromise is to use standard priors which do not involve personal (subjective) opinions.

Non informative priors

The prior distribution is meant to reflect the opinion of the researcher prior to observing any data. What if there is no opinion? (Whether this is realistic is disputable.)

This is a relevant issue and a possible answer to the objection that the results of inference should not depend on subjective opinions.

It has then been proposed to use 'standard' distributions which, in some sense, bring no (or very limited) information on the parameter.

An intuitive solution is to assume $\pi(\theta) \propto k$ so that no values of θ are privileged (principle of insufficient reason).

- Strictly speaking, this is admissible only if the parameter space is limited.
- If the parameter space is not limited a constant has an infinite integral and so is not a probability distribution.
- It is possible however, that a proper posterior distribution is obtained even starting from an improper prior. If this is the case, the inference is valid.

Jeffreys' prior

The non informative nature of the uniform distribution 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 instead an informative prior on its transform $\psi = \psi(\theta)$.

Jeffreys' prior

The above issue may be overcome by posing

$$\pi(\theta) = \sqrt{\det H(\theta)}$$

where H is the information matrix, that is, the matrix with (i, j) element

$$[H(\theta)]_{ij} = -E\left(\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\theta; y)\right)$$

then, for any parametrization $\psi = \psi(\theta)$

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

Consider, for instance, a Binomial experiment, so the log-likelihood is

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

then

$$[H(\theta)] = -E\left(\frac{d^2}{d\theta^2} l(\theta; y)\right) = \frac{n}{\theta(1 - \theta)}$$

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

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

Bayes computation

To answer the basic questions of statistical inference we need to know the posterior distribution of θ:

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

The principal practical challenge is that $\int_\Theta f(y|\theta)\pi(\theta)d\theta$ is usually intractable for many interesting models and also quantity related to $\pi(\theta|y)$ of direct interest for summarizing inference (such as mean, median, percentiles, probabilities) cannot be evaluated.

There are then two main strategies to overcome the problem:

- approximate the integrals
- find a way to get a (simulated) sample from $\pi(\theta|y)$ without requiring evaluation of the integrals.

The latter strategy is based on the fact that simulating from a density is as good as being able to evaluate the density, and sometimes better. This is achieved mainly by **Monte Carlo Markov Chain methods**.

Monte Carlo Markov Chain

Monte Carlo Markov Chain methods simulate values from a Markov chain whose stationary distribution is exactly the posterior distribution of interest.

Once a sample of simulated values is given this can be used to evaluate all the quantities of interest for Bayesian inference.

Two are the main algorithms to obtain this sample of simulated values:

- Metropolis-Hastings algorithm
- Gibbs sampling

For a comprehensive overview about theory and practice of Bayesian statistics and MCMC methods you could attend the course Bayesian Statistics during your second year, second semester!