Chapter 1
Bayesian Model

Xavi Puig Oriol
Departament d'Estadística i I.O.

Chapter 1. Bayesian Model

1.1 The Statistical Model

A Statistical Model is a list of Probability models indexed by a parameter θ (which can be scalar, vector, matrix..) that we know belongs to a parameter space Ω,

$$M = \{p(\tilde{y} \mid \theta),\ \theta \in \Omega\}$$

Statistical Model

Probability Model

Parameter

Parameter Space

The list of probability distributions of the Statistical Model shares the same Sample Space.

1.1 The Statistical Model

Example 1

Parameter Space

$$M = \{p(\tilde{y} \mid \theta) = Binomial(n = 10,\ \theta), \theta \in (0,1)\}$$

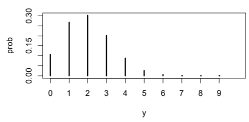
A probability model that belongs to this statistical model is for example:

$$Binomial(n = 10, \theta = 0.2)$$

Using this probability model we can calculate probabilities on the sample space

Sample Space

$$p(\tilde{y} = y \mid \theta = 0.2) = \binom{10}{y} 0.2^{y} (1 - 0.2)^{10-y} \qquad \text{for all} \quad y \in \{0,1,2,\ldots,10\}$$



Note: From a probability model we can calculate probabilities, but we cannot do it from a statistical model.

1.1 The Statistical Model

Example 2

$$M = \{p(\tilde{y} \mid \theta) = Binomial(n = 10, \theta), \theta \in \{0.3, 0.7\}\} = \{Binomial(n = 10, \theta = 0.3), Binomial(n = 10, \theta = 0.7)\}$$

Example 3

$$M = \{p(\tilde{y} \mid \theta = (\mu, \sigma)) = Normal(\mu, \sigma), \mu \in (-\infty, \infty), \sigma \in (0, \infty)\}$$

Example 4

$$M = \{p(\tilde{y} \mid \theta = (\beta_0, \beta_1, \sigma), x) = Normal(\beta_0 + \beta_1 x, \sigma), (\beta_0, \beta_1) \in \Re^2, \sigma \in (0, \infty)\}$$

1.1 The Statistical Model

Given a Statistical Model and a Data Set, the "game" of the statistics is to guess which model

$$p(\tilde{y} \mid \theta^*) \in M$$

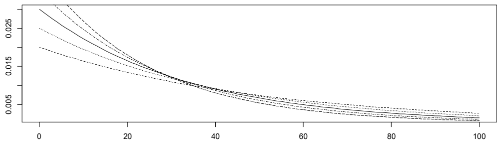
generated the data, where $\theta^* \in \Omega$ is the true value of the parameter.

1.1 The Statistical Model

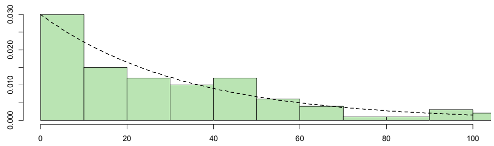
Example

Assuming that the time until death after suffering a very severe heart attack can be characterized by an exponential model,

$$M = \{p(\tilde{y} \mid \theta) = \exp(\theta) = \theta e^{-\theta \tilde{y}}, \theta \in (0, \infty)\}$$



After getting the data, we try to guess θ^*, which is the true population shape



Chapter 1. Bayesian Model

1.2 The four problems of Statistics

1.2 The four problems of Statistics

1.2.1 Data Collection

Design how to collect the Data.

This leads you to choose the statistical model from which you will observe the data. The statistical model will be determined by the nature of the data (discrete or continuous) and by the type of sampling.

$$M = \{p(\tilde{y} \mid \theta),\, \theta \in \Omega\}$$

1.2 The four problems of Statistics

1.2.2 Model Validation

After choosing the model $M = \{p(\tilde{y} \mid \theta), \theta \in \Omega\}$

and a collecting the data $\tilde{y} = y$

you have to decide if the model is valid or not.

That is, you have to decide if the probability model that generated the data belongs to M and therefore the model is correct,

$$p(\tilde{y} \mid \theta^*) \in M$$

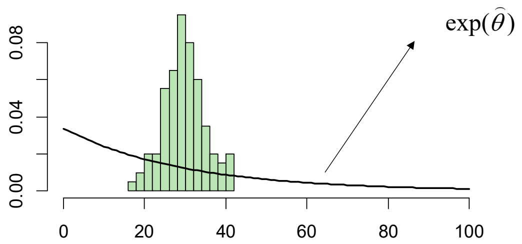
1.2 The four problems of Statistics

1.2.2 Model Validation

Example

Model Chosen $M = \{p(\tilde{y} \mid \theta) = \exp(\theta), \theta \in \Re_+\}$

Data Colected $y_1 = 20.2, y_2 = 35.7, \ldots, y_{100} = 27.4$



It is unlikely that Data has been generated by an Exponential Model. Therefore, the Model is not Valid.

1.2 The four problems of Statistics

1.2.3 Statistical Inference

After collecting the data and checking the model, Statistical Inference tries to guess the true parameter value, θ^*, that is, it tries to guess the model that generated the Data,

$$p(\tilde{y} \mid \theta^*) \in M$$

Statistical Inference mainly consists of:

a) Point Estimation

b) Interval Estimation

c) Hypothesis Testing

d) Prediction

1.2 The four problems of Statistics

1.2.4 Results Presentation

Finally, the results must be interpreted and displayed in an understandable way according to the audience and/or client.

Chapter 1. Bayesian Model

1.3 Bayesian Model

A Bayesian statistician starts with a Statisical Model,

$$M = \{p(\tilde{y} \mid \theta), \theta \in \Omega\}$$

but on top of this he considers θ as a random variable and is willing to choose a probability distribution over Ω, $\pi(\theta)$, before looking at the data, that captures prior knowlege about θ.

We call $\pi(\theta)$ the prior distribution or simply the prior.

1.3 Bayesian Model

Example 1

We want to know the % of smokers older tan 18 years in Barcelona. For this purpose we take a random sample of 200 citizens of Barcelona and count the observed number of smoking citizens.

The Statistical Model of this experiment is:

$$\tilde{y} \mid \theta \sim Binomial(n = 200, \theta), \quad \theta \in (0,1)$$

Where:

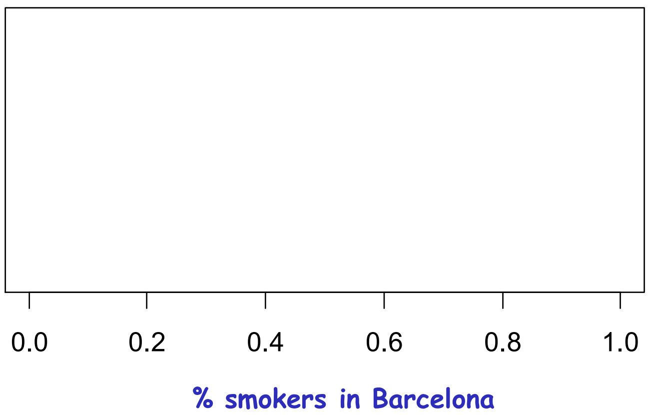
θ is the smoking rate of Barcelona

$\theta*100$ is the percentage of smokers in Barcelona

The prior distribution should be defined using the knowledge of the experts

1.3 Bayesian Model

Example 1



1.3 Bayesian Model

A Bayesian Model is a list of probability distributions sorted from more likely to less likely according to $\pi(\theta)$,

$$M_B = \{p(\tilde{y} \mid \theta), \theta \in \Omega; \pi(\theta)\}$$

Hence, if $\pi(\theta_1) > \pi(\theta_2)$

it means that we believe that the true model is more likely

to be $p(\tilde{y} \mid \theta_1)$

instead of $p(\tilde{y} \mid \theta_2)$

1.3 Bayesian Model

Example 2

We are interested in:

a) The probability of getting head after tossing a coin

b) The probability of a pin landing with the point up after tossing it

c) The probability of something we know nothing about

?

1.3 Bayesian Model

Example 2

We are interested in knowing the probability of:

a) 

b) 

c) **?**

In all the cases the same experiment is performed. The experiment is repeated 10 times and the number of "successes" is counted.

Therefore, in all three cases the Statistical Model is the same:

$$\tilde{y} \mid \theta \sim Binomial(n=10, \theta), \quad \theta \in (0,1)$$

Chapter 1. Bayesian Model

1.4 The Likelihood function

All the information that the data, y, has about the parameter, θ, is in the likelihood function (or likelihood).

The likelihood function is a function proportional to the probability distribución of $\tilde{y}$, $p(\tilde{y} \mid \theta)$, evaluated at the data, $p(y \mid \theta)$, and expressed as a function of θ,

proportional

$$L_y(\theta) \propto p(y \mid \theta)$$

Data observed

1.4 The Likelihood function

The likelihood function sorts the parameter space from less likely to more likely.

So, if $L_y(\theta_1) > L_y(\theta_2)$

it means that according to the data information the true value of the parameter is more likely to be θ_1 than θ_2 .

1.4 The Likelihood function

Example 1

We want to know the probability of getting a head after tossing a coin.

The experiment consists of tossing the coin 10 times and counting the observed heads.

The Statistical Model for this experiment is:

$$\tilde{y} \mid \theta \sim Binomial(n = 10, \theta), \quad \theta \in (0,1)$$

Where:

θ is the probability of getting head.

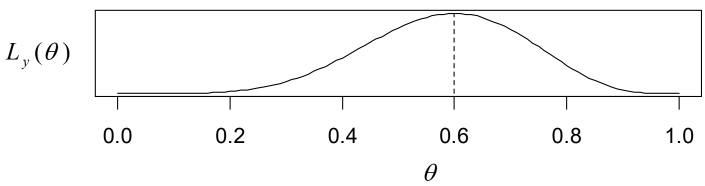
After performing the experiment, we observed 6 heads, y=6.

1.4 The Likelihood function

Example 1 (continued)

The likelihood function is:

$$L_y(\theta) \propto p(y \mid \theta) = \binom{10}{6} \theta^6 (1-\theta)^4 \propto \theta^6 (1-\theta)^4$$



The maximum likelihood value for θ is 0.6, so for example, θ^* is more likely to be 0.6 than 0.4.

1.4 The Likelihood function

Example 2

We want to estimate the mean of the height of Catalan adults, μ. Assuming that their height, y, is $y|\mu \sim Normal(\mu, \sigma=8)$, where σ=8 means that 99,7% of the population height falls into a range of 48 cm.

Supposing the Statistical model is:

$$\tilde{y} \mid \mu \sim Normal(\mu, \sigma = 8), \quad \mu \in \Re$$

A sample of size $n = 10$ is taken, and the height values observed in centimeters are:

168, 175, 155, 183, 158, 170, 172, 167, 184 and 171

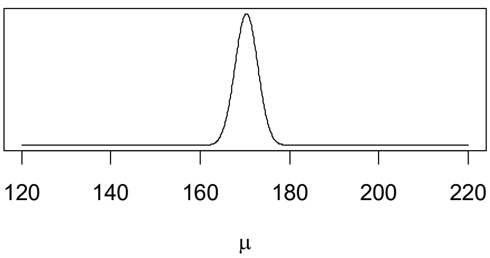
1.4 The Likelihood function

Example 2 (continued)

$\sigma = 8$
n = 10
y = (168, 175, 155, 183, 158, 170, 172, 167, 184, 171)

$$L_y(\mu) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \propto e^{-\frac{1}{2\cdot 8^2}\sum_{i=1}^{10}(y_i-\mu)^2}$$

So, the likelihood function graph is:



1.4 The Likelihood function

The standardized likelihood function is defined as:

$$L_y^{std}(\theta) = \frac{L_y(\theta)}{\int L_y(\theta)\partial\theta}$$

The integral of the standardized likelihood function is 1,

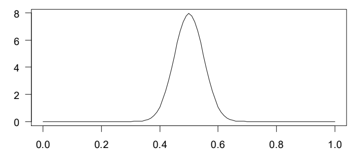
$$\int L_y^{std}(\theta)\partial\theta = \int \frac{L_y(\theta)}{\int L_y(\theta)\partial\theta}\partial\theta = \frac{\int L_y(\theta)\partial\theta}{\int L_y(\theta)\partial\theta} = 1$$

1.4 The Likelihood function

Example 1 (continued)

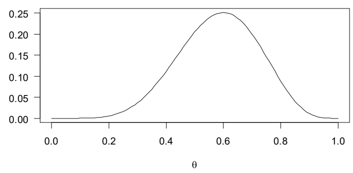
Given the Bayesian Model:

$$\tilde{y} \mid \theta \sim Binomial(n = 10, \theta)$$
$$\theta \sim Beta(50, 50)$$



Observed y=6

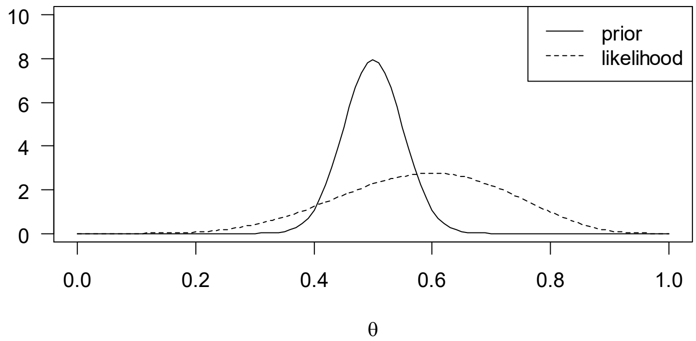
$$L_y(\theta) \propto \binom{10}{6}\theta^6(1-\theta)^4$$



1.4 The Likelihood function

Example 1 (continued) 

It is very illustrative to show the prior distribution and the standardized likelihood function together :



1.4 The Likelihood function

Summarizing…

Statistical Model,

$$M = \{p(\tilde{y} \mid \theta), \theta \in \Omega\}$$

is a (not sorted) list of probability models (one for each posible value of θ).

- The prior, $\pi(\theta)$, sorts the probability models of M using prior information.
- The likelihood, $L_y(\theta)$, sorts the probability models of M using the Data information.

Chapter 1. Bayesian Model

1.5 The Posterior Distribution

- The prior distribution, $\pi(\theta)$, shows the information we have about the parameter **before** looking at the data.

- The posterior distribution, $\pi(\theta|y)$, shows the information we have about the parameter **after** looking at the data.

1.5 The Posterior Distribution

Bayesian Model

$$\tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta)$$
$$\theta \sim \pi(\theta) \longrightarrow \text{Prior Information}$$

⬇

Data Information

$L_y(\theta)$ ➡ **Update**

⬇

Bayesian Model (updated)

$$\tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta)$$
$$\theta \mid y \sim \pi(\theta \mid y) \longrightarrow \text{Posterior Information}$$

1.5 The Posterior Distribution

The posterior distribution is calculated using Bayes' Theorem as:

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

where $p(y)$, is the marginal distribution of $\tilde{y}$ evaluated at the data, which we will also call the prior predictive distribution evaluated at the data,

$$p(\tilde{y} = y) = p(y) = \int p(y,\theta)d\theta = \int p(y \mid \theta)\pi(\theta)d\theta$$

Observation: $p(y)$ is a constant, also called the normalization constant, which makes that the posterior distribution integrates 1.

1.5 The Posterior Distribution

Hence,

$$\pi(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{\underbrace{p(y)}_{\int p(y \mid \theta)\pi(\theta)d\theta}} \propto p(y \mid \theta)\pi(\theta) = L_y(\theta)\pi(\theta)$$

This means that the posterior distribution is a copromise between the prior distribution (information before the data) and the likelihood function (data information),

$$\boxed{\pi(\theta \mid y) \propto L_y(\theta)\pi(\theta)}$$

Note: The most complicated calculation is the denominator. But, in some cases, by calculating only the numerator, we can identify the family of the posterior distribution.

1.5 The Posterior Distribution

Example 1

Given the Bayesian Model:

$$\tilde{y} \mid \theta \sim Binomial(n,\theta) \longrightarrow p(\tilde{y} \mid \theta) = \binom{n}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{n-\tilde{y}}$$

$$\theta \sim Beta(a,b) \longrightarrow \pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

And the observed data : n and y

The posterior distribution is calculated as follows:

$$\pi(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{p(y)} \propto p(y \mid \theta)\pi(\theta)$$

1.5 The Posterior Distribution

Example 1 calculating the denominator

$$\pi(\theta \mid y) = \frac{p(y \mid \theta)\pi(\theta)}{p(y)} = \frac{\binom{n}{y}\theta^{y}(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}}{\int_0^1 \binom{n}{y}\theta^{y}(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\partial\theta} =$$

$$= \frac{\binom{n}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{(a+y)-1}(1-\theta)^{(b+n-y)-1}}{\binom{n}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^1 \theta^{y}(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1}\partial\theta} = \frac{\theta^{(a+y)-1}(1-\theta)^{(b+n-y)-1}}{\int_0^1 \theta^{(a+y)-1}(1-\theta)^{(b+n-y)-1}\partial\theta} =$$

$$= \frac{\theta^{(a+y)-1}(1-\theta)^{(b+n-y)-1}}{\frac{\Gamma(a+y)\Gamma(b+n-y)}{\Gamma(a+y+b+n-y)}\int_0^1 \frac{\Gamma(a+y+b+n-y)}{\Gamma(a+y)\Gamma(b+n-y)}\theta^{(a+y)-1}(1-\theta)^{(b+n-y)-1}\partial\theta} =$$

$$= \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)}\theta^{(a+y)-1}(1-\theta)^{(b+n-y)-1} = Beta(a+y, b+n-y)$$

1.5 The Posterior Distribution

Example 1 (continued)

$$\pi(\theta \mid y) \propto p(y \mid \theta)\pi(\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}$$

$$\propto \theta^{y}(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1} = \theta^{(a+y)-1}(1-\theta)^{(b+n-y)-1}$$

$$\Downarrow$$

$$\pi(\theta \mid y) = Beta(a+y, b+n-y)$$

1.5 The Posterior Distribution

Example 2

We are interested in:

a) The probability of getting head after throwing a coin

b) The probability of a pin landing with the point up after tossing it

c) The probability of something we know nothing about

?

In all the cases the same experiment is performed. The same experiment is repeated 10 times and the number of "successes" is counted.

1.5 The Posterior Distribution

Ejemplo 2 (continuación)

The Bayesian Model for each case is:

a) Coin

$$\tilde{y} \mid \theta \sim Binomial(n=10, \theta), \quad \theta \in (0,1)$$
$$\theta \sim Beta(50, 50)$$

b) pin

$$\tilde{y} \mid \theta \sim Binomial(n=10, \theta), \quad \theta \in (0,1)$$
$$\theta \sim Beta(2, 2)$$

c) **?**

$$\tilde{y} \mid \theta \sim Binomial(n=10, \theta), \quad \theta \in (0,1)$$
$$\theta \sim Beta(1, 1)$$

Imagine that seven successes have been observed in all three cases:

$$n = 10 \qquad y = 7$$

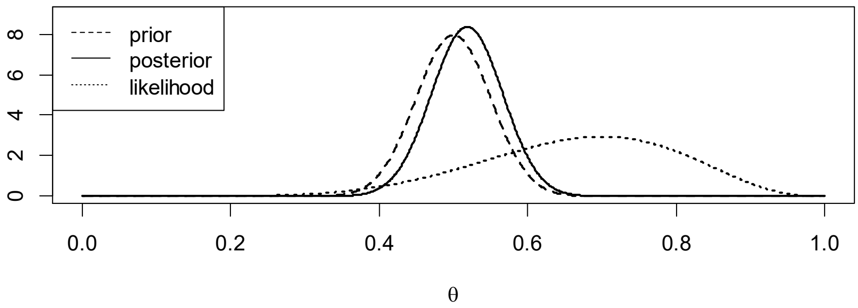
1.5 The Posterior Distribution

Example 2 (continued)

a)  ($n = 10,\ y = 7$)

$$\theta \sim Beta(50, 50) \qquad E[\theta] = 0.5 \qquad V[\theta] = 0.0025$$

$$\theta \mid y \sim Beta(57, 53) \qquad E[\theta \mid y] = 0.52 \qquad V[\theta \mid y] = 0.0022$$



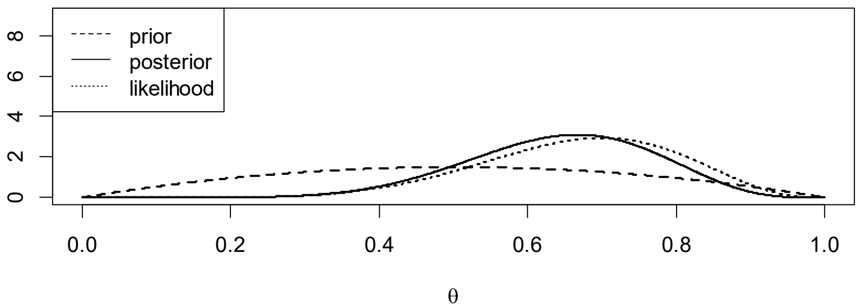
1.5 The Posterior Distribution

Example 2 (continued)

b)  ($n = 10,\ y = 7$)

$$\theta \sim Beta(2, 2) \qquad E[\theta] = 0.5 \qquad V[\theta] = 0.05$$

$$\theta \mid y \sim Beta(9, 5) \qquad E[\theta \mid y] = 0.64 \qquad V[\theta \mid y] = 0.015$$



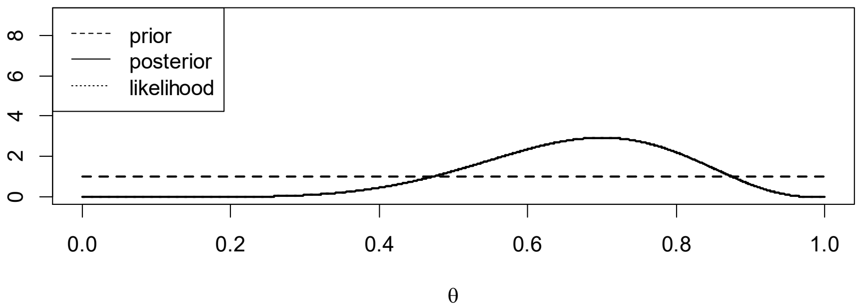
1.5 The Posterior Distribution

Example 2 (continued)

c) **?** ($n = 10$, $y = 7$)

$$\theta \sim Beta(1,1) \qquad E[\theta] = 0.5 \qquad V[\theta] = 0.083$$

$$\theta \mid y \sim Beta(8,4) \qquad E[\theta \mid y] = 0.667 \qquad V[\theta \mid y] = 0.017$$



1.5 The Posterior Distribution

Note: if $\pi(\theta)$ is constant, the posterior Distribution becomes the standardized likelihood function,

$$\pi(\theta \mid y) = \frac{L_y(\theta)\pi(\theta)}{\int L_y(\theta)\pi(\theta)\partial\theta} = \frac{L_y(\theta)}{\int L_y(\theta)\partial\theta} = L_y^{std}(\theta)$$

1.5 The Posterior Distribution

Summarizing……

Bayesian Model

$$\tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta)$$
$$\theta \sim \pi(\theta)$$

↓

Data → **Bayes' Th**

↓

Bayesian Model (updated)

$$\tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta)$$
$$\theta \mid y \sim \pi(\theta \mid y) \longrightarrow \pi(\theta \mid y) = \frac{L_y(\theta)\pi(\theta)}{\int L_y(\theta)\pi(\theta)d\theta}$$

The updated model allows you to calculate point estimations, interval estimations, predictions…

1.5 The Posterior Distribution

Example 2 (Coin)

Bayesian Model

$$\tilde{y} \mid \theta \sim Binomial(n = 10, \theta)$$
$$\theta \sim Beta(50, 50)$$

↓

Data y=7 → **Bayes' Th**

↓

Bayesian Model (updated)

$$\tilde{y} \mid \theta \sim Binomial(\tilde{n}, \theta)$$
$$\theta \mid y \sim Beta(57, 53)$$

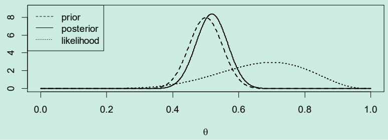
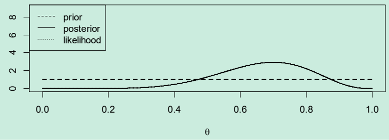
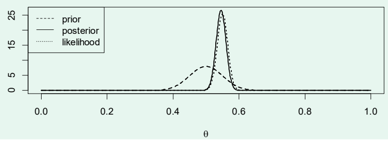
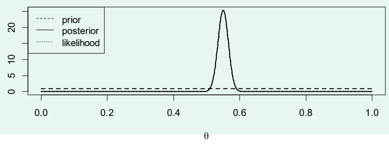
The updated model allows you to calculate point estimations, interval estimations, predictions…

1.5 The Posterior Distribution

Example 2 (coin)

For each of the following scenarios (data and prior) :

- Think about what the prior distribution, the likelihood function and the posterior distribution look like:

Data	Beta(50, 50)	Uniform(0, 1)
n=10 y=7		
n=1000 y=550		

1.5 The Posterior Distribution

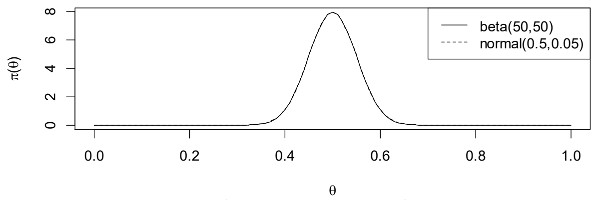
Ejemplo 2 (Coin)

Model A

$$\tilde{y} \mid \theta \sim Binomial(n = 10, \theta)$$
$$\theta \sim Beta(50, 50)$$

Model B

$$\tilde{y} \mid \theta \sim Binomial(n = 10, \theta)$$
$$\theta \sim Normal(0.5, 0.05)$$



Data observed: $n = 10$ and $y = 7$

What will the posterior distribution look like in each model, given that both prior distributions look identical?

Chapter 1. Bayesian Model

1.6 The prior and posterior predictive distribution

Example 1

We want to learn about the height of adult Catalans. Assume that their height, y, is $y|\mu \sim Normal(\mu, \sigma=8)$, where σ=8 means that 99,7% of the population height falls in a range of 48 cm.

$\tilde{y}$:= the height in centimeters of an adult

Assuming the Statistical Model:

$$\tilde{y} \mid \mu \sim Normal(\mu, \sigma = 8), \quad \mu \in \Re$$

Defining the Bayesian Model is equivalent to specifying the prior distribution for μ.

$$M_B = \{\tilde{y} \mid \mu \sim Normal(\mu, \sigma = 8), \quad \mu \in \Re; \pi(\mu)\}$$

1.6 The prior and posterior predictive distribution

Example 1 (continued)

We want to learn about the height of adult Catalans. Assume that their height, y, is $y|\mu \sim Normal(\mu, \sigma=8)$, where σ=8 means that 99,7% of the population height falls in a range of 48 cm.

$$\tilde{y} \mid \mu \sim Normal(\mu, \sigma = 8), \quad \mu \in \Re$$
$$\mu \sim ?$$

1.6 The prior and posterior predictive distribution

Example 1 (continued)

We want to learn about the height of adult Catalans. Assume that their height, y, is $y|\mu \sim Normal(\mu, \sigma=8)$, where σ=8 means that 99,7% of the population height falls in a range of 48 cm.

$$\tilde{y} \mid \mu \sim Normal(\mu, \sigma = 8), \quad \mu \in \Re$$
$$\mu \sim Normal(m = ?, s = ?)$$

Defining the Bayesian Model means choosing the Statistical Model and the prior distribution by specifying the values of its parameters.

Google: "Normal Distribution calculator"
(https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html)

1.6 The prior and posterior predictive distribution

Before observing the data, we start from the Bayesian Model:

$$M_B = \{p(\tilde{y} \mid \theta), \theta \in \Omega; \pi(\theta)\}$$

Where $\pi(\theta)$, represents the degreee of knowledge or uncertainty about θ (about Ω).

The marginal distribution of $\tilde{y}$, called the prior predictive distribtuion, represents our knowledge and uncertainty about the sample space before observing the data.

1.6 The prior and posterior predictive distribution

We can calculate the prior predictive distribution as follows:

$$p(\tilde{y}) = \int p(\tilde{y}, \theta) d\theta = \int p(\tilde{y} \mid \theta) \pi(\theta) d\theta$$

Observations:

- $p(\tilde{y}) = E_{\pi(\theta)}[p(\tilde{y} \mid \theta)]$

  is a weighed average of all possible porbability models, where the weights are determined by $\pi(\theta)$.

- $p(\tilde{y})$ allows to make predictions before observing the data.

1.6 The prior and posterior predictive distribution

Example 1 (continued)

$$\tilde{y} \mid \mu \sim Normal(\mu, \sigma = 8), \quad \mu \in \Re$$
$$\mu \sim Normal(m = ?, s = ?)$$

$\mu = E[\tilde{y}]$ is the expected value of y, that is, the population mean

$\pi(\mu)$ is the prior information about μ, that is, about the population mean (not about y)

$p(\tilde{y})$ is the prior information about y, that is, about the observable space (the sample space).

If we know something about the parameters, it means that we know something about what the data will look like and vice versa. The prior predictive distribution translates the information about the parameter space into the sample space.

1.6 The prior and posterior predictive distribution

Example 1 (continued)

$$p(\tilde{y}) = \int p(\tilde{y}, \mu) = \int p(\tilde{y} \mid \mu)\, \pi(\mu) \partial\mu =$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\tilde{y}-\mu)^2} \frac{1}{\sqrt{2\pi s^2}} e^{-\frac{1}{2s^2}(\mu - m)^2} \partial\mu =$$

$$= \cdots = Normal\left(m, \sqrt{\sigma^2 + s^2}\right)$$

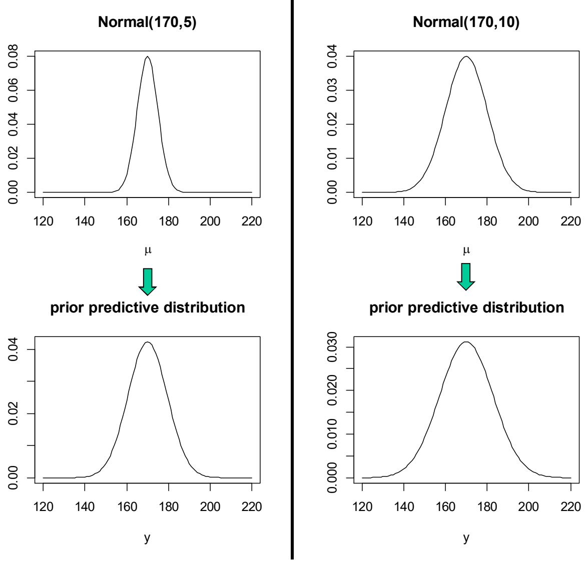
$p(\tilde{y})$ is the prior predictive distribution
is the probability distribution on the sample space

1.6 The prior and posterior predictive distribution

Example 1 (continued)

$\pi(\mu)$

$p(\tilde{y})$



1.6 The prior and posterior predictive distribution

Defining the Bayesian Model means choosing the Statistical Model and the prior distribution.

The prior's parameters can be chosen in two ways

1. Using/drawing the prior distribution in case the information we have is about the parameter space
2. Using/drawing the prior predictive distribution in case the information we have is about the sample space

1.6 The prior and posterior predictive distribution

Once the data has been observed, we update the model changing $\pi(\theta)$ by $\pi(\theta|y)$. Now,

$\pi(\theta \mid y)$ shows everything we know about the parameter

$p(\tilde{y} \mid y)$ shows everything we know about the data behavior

The posterior predictive distribution

1.6 The prior and posterior predictive distribution

The posterior predictive distribtuion is calculated as follows:

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

Observation:

$p(\tilde{y} \mid y)$ will be used to make predictions

1.6 The prior and posterior predictive distribution

Example 2

Given the Bayesian Model:

$$\tilde{y} \mid \theta \sim Binomial(n,\theta) \longrightarrow p(\tilde{y} \mid \theta) = \binom{n}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{n-\tilde{y}}$$

$$\theta \sim Beta(a,b) \longrightarrow \pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}$$

We calculate the prior predictive distribution as follows:

$$p(\tilde{y}) = \int p(\tilde{y},\theta)d\theta = \int p(\tilde{y} \mid \theta)\pi(\theta)d\theta$$

1.6 The prior and posterior predictive distribution

Example 2 (continued)

$\tilde{y}$ future observation

$\tilde{n}$ Number of future experiments

$$p(\tilde{y}) = \int p(\tilde{y},\theta)d\theta = \int p(\tilde{y} \mid \theta)\pi(\theta)d\theta = \int_0^1 \binom{\tilde{n}}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{\tilde{n}-\tilde{y}} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} \partial\theta =$$

$$= \binom{\tilde{n}}{\tilde{y}} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \theta^{(a+\tilde{y})-1}(1-\theta)^{(b+\tilde{n}-\tilde{y})-1}\partial\theta =$$

$$\int_0^1 Beta(a+\tilde{y}, b+\tilde{n}-\tilde{y})\partial\theta = 1$$

$$= \binom{\tilde{n}}{\tilde{y}} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+\tilde{y})\Gamma(b+\tilde{n}-\tilde{y})}{\Gamma(a+b+\tilde{n})} \underbrace{\int_0^1 \frac{\Gamma(a+b+\tilde{n})}{\Gamma(a+\tilde{y})\Gamma(b+\tilde{n}-\tilde{y})} \theta^{(a+\tilde{y})-1}(1-\theta)^{(b+\tilde{n}-\tilde{y})-1}\partial\theta}_{} =$$

$$= \binom{\tilde{n}}{\tilde{y}} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+\tilde{y})\Gamma(b+\tilde{n}-\tilde{y})}{\Gamma(a+b+\tilde{n})} \Longrightarrow BetaBinomial(a,b,\tilde{n})$$

See the formulary

1.6 The prior and posterior predictive distribution

Example 2 (continued)

$$\begin{array}{l} \tilde{y} \mid \theta \sim Binomial(n,\theta) \\ \theta \sim Beta(a,b) \end{array} \longrightarrow \tilde{y} \sim BetaBinomial(a,b,\tilde{n})$$

after observing the data: n , y

$$\pi(\theta \mid y) = Beta(a+y, b+n-y)$$

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta)\pi(\theta \mid y)d\theta = \ldots =$$

$$= \binom{\tilde{n}}{\tilde{y}} \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)} \frac{\Gamma(a+y+\tilde{y})\Gamma(b+n-y+\tilde{n}-\tilde{y})}{\Gamma(a+b+n+\tilde{n})} =$$

$$= BetaBinomial(a+y, b+n-y, \tilde{n})$$

1.6 The prior and posterior predictive distribution

Example 3

We are interested in:

a) The probability of getting head after throwing a coin

b) The probability of a pin landing with the point up after throwing it

c) The probability of something we know nothing about

?

In all the cases the same experiment is performed. The same experiment is repeated 10 times and the number of "successes" is counted.

1.6 The prior and posterior predictive distribution

Example 3 (continued)

The Bayesian Model for each case is:

a) Coin

$$\tilde{y} \mid \theta \sim Binomial(n = 10, \theta), \quad \theta \in (0,1)$$
$$\theta \sim Beta(50, 50)$$

b) pin

$$\tilde{y} \mid \theta \sim Binomial(n = 10, \theta), \quad \theta \in (0,1)$$
$$\theta \sim Beta(2, 2)$$

c) ?

$$\tilde{y} \mid \theta \sim Binomial(n = 10, \theta), \quad \theta \in (0,1)$$
$$\theta \sim Beta(1, 1)$$

Seven successes have been observed in all three cases: $n = 10$, $y = 7$

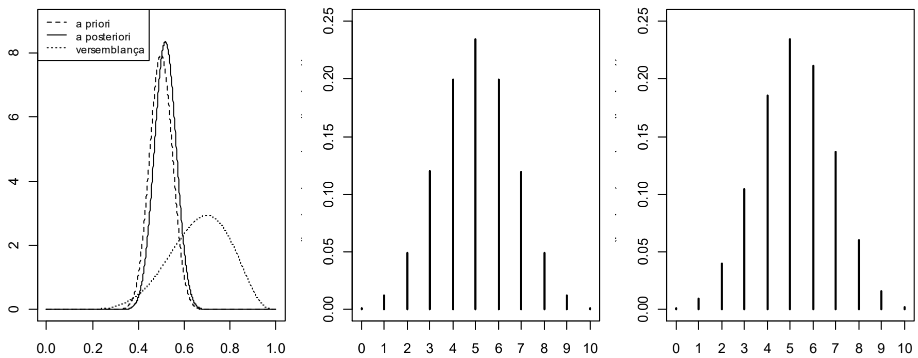
We will make predictions for a future sample of size 10: $\tilde{n} = 10$

1.6 The prior and posterior predictive distribution

Example 3 (continued)

a) Coin ($\tilde{n}$ = 10)

$\pi(\theta), \pi(\theta \mid y), L_y(\theta)$ $\quad$ $p(\tilde{y})$ $\quad$ $p(\tilde{y} \mid y)$

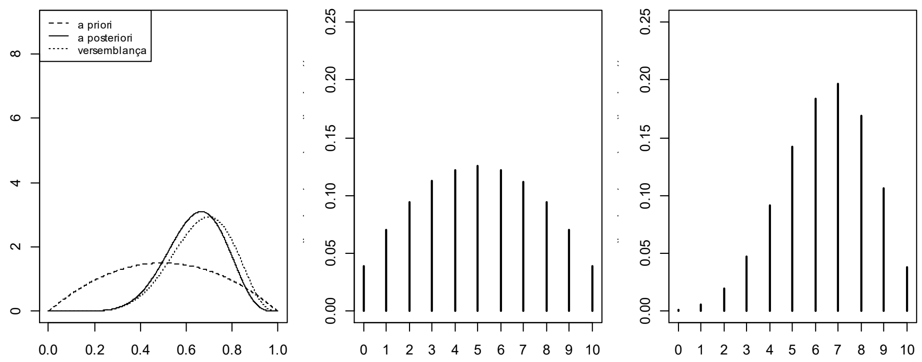


1.6 The prior and posterior predictive distribution

Example 3 (continued)

b) pin ($\tilde{n}$ = 10) 

$\pi(\theta), \pi(\theta \mid y), L_y(\theta)$ $\quad p(\tilde{y})$ $\quad p(\tilde{y} \mid y)$

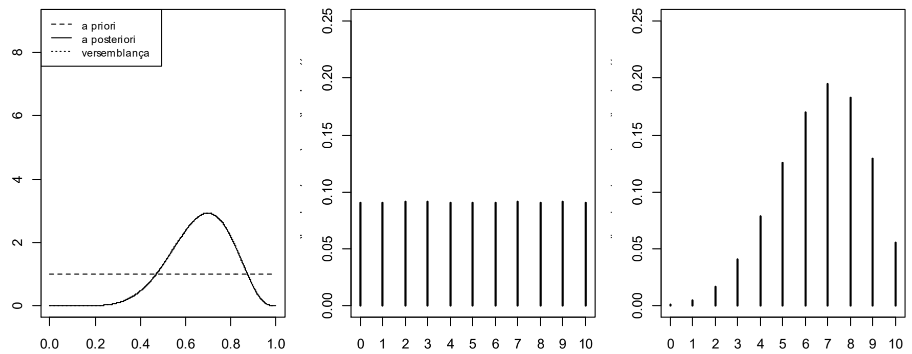


1.6 The prior and posterior predictive distribution

Example 3 (continued)

c) **?** ($\tilde{n}$ = 10)

$\pi(\theta), \pi(\theta \mid y), L_y(\theta)$ $\quad p(\tilde{y})$ $\quad p(\tilde{y} \mid y)$



1.6 The prior and posterior predictive distribution

Predictive distributions can be calculated in two ways:

a) Analytically

b) By simulation

---

a) Analytically: solving the integrals

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

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

1.6 The prior and posterior predictive distribution

b) By simulation

To approximate the prior predictive distribution using simulations, first of all you must set the number of simulations, M (the larger the better), then from j=1 to M:

1. simulate

$\theta^{(j)}$ from $\pi(\theta)$, and

2. simulate

$\tilde{y}^{(j)}$ from $p(\tilde{y} \mid \theta^{(j)})$

These $\tilde{y}^{(1)},\ldots,\tilde{y}^{(M)}$ simulated values are simulations from the prior predictive distribution. Hence, using these simulations we can calculate everything we want/need (moments, probabilities, etc.). And, by graphing it we can reproduce the shape of its probability distribution.

1.6 The prior and posterior predictive distribution

Example 1 (continued)

$$\tilde{y} \mid \mu \sim Normal(\mu, \sigma = 8), \quad \mu \in \Re$$
$$\mu \sim Normal(m = 170, s = 5)$$

Setting $M=100.000$

1. We simulate M values from a $Normal(170, 5)$

   $\mu^{(1)} = 171.2, \mu^{(2)} = 167.8, \ldots, \mu^{(100.000)} = 170.4$

2. For each simulated value of μ , in step 1, we simulate a new value from a

   $Normal(\mu^{(j)}, \sigma = 8)$

   $\tilde{y}^{(1)} = 182.4 , \tilde{y}^{(2)} = 158.8, \ldots, \tilde{y}^{(100.000)} = 170.1$

   $\Downarrow$

   It is a simulated value from a $Normal(171.2, 8)$

1.6 The prior and posterior predictive distribution

To approximate the posterior predictive distribution using simulations, first of all, you must set the number of simulations, *M* (the larger the better), then for *j*=1 to *M*:

1. Simulate

   $\theta^{(j)}$ de $\pi(\theta \mid y)$, i

2. Simulate

   $\tilde{y}^{(j)}$ de $p(\tilde{y} \mid \theta^{(j)})$

These $\tilde{y}^{(1)}, \ldots, \tilde{y}^{(M)}$ simulated values are simulations from the prior predictive distribution. Hence, using these simulations we can calculate everything we want/need (moments, probabilities, etc.). And, by graphing it we can reproduce the shape of its probability distribution.

1.6 The prior and posterior predictive distribution

Summarizing……

Bayesian Model

$$\tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta)$$
$$\theta \sim \pi(\theta)$$
$$\longrightarrow \quad p(\tilde{y})$$

Data → **Bayes' Th.**

Bayesian Model (updated)

$$\tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta)$$
$$\theta \mid y \sim \pi(\theta \mid y)$$
$$\longrightarrow \quad p(\tilde{y} \mid y)$$

The updated Bayesian Model allows to make point estimations, interval estimations, predictions, …

Chapter 1. Bayesian Model

1.1 The Statistical Model
1.2 The four problems of statistics
1.3 The Bayesian Model
1.4 The likelihood function
1.5 The posterior distribution
1.6 The prior and posterior predictive distribution
1.7 Choosing the prior distribution

1.7 Choosing the Prior Distribution

Before observing the data, we start from the Bayesian Model:

$$M_B = \{p(\tilde{y} \mid \theta), \theta \in \Omega; \pi(\theta)\}$$

Where $\pi(\theta)$, represents our knwoledge or uncertainty about θ (about Ω).

The prior information should be based on previous studies and the subjective knowledge of experts.

In the case of not having information or not being able to agree on the knowledge of the experts, we will use non-informative prior distributions.

1.7 Choosing the Prior Distribution

Conjugate prior:

A prior distribution is the conjugate prior of a statistical model if the posterior distribution is from the same family as the prior distribution.

Examples

Model	prior	Posterior
Binomail(n, θ)	*Beta(a,b)*	*Beta(a+y, b+n-y)*
Poisson(θ)	*Gamma(a,b)*	*Gamma(a+Σy , b+n)*

1.7 Choosing the Prior Distribution

1.7.1 Informative prior distributions

$\pi(\theta)$ represents our knowledge (or uncertainty) about θ and we have to choose its parameters' values (let's call them a and b) and we basically do it in two ways:

1. by trial and error, drawing $\pi(\theta)$.
2. solving a system of equations based on moments and/or quartiles.

Example

$$E[\theta] = f_1(a,b) = k_1$$
$$q_{0.75}(\theta) = f_2(a,b) = k_2$$

1.7 Choosing the Prior Distribution

Example 1

Several previous studies conclude that using a new treatment for curing burns the probability to get better in five days is mainly between 0.4 and 0.8. The Hospital decides to carry out its own clinical trial; The next 40 patients with burns will be treated using this new treatment, and the number of cured patients after 5 days will be counted.

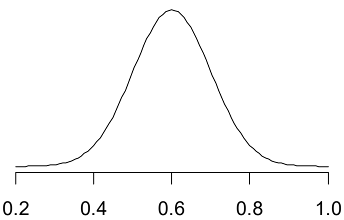
$$\boxed{\begin{array}{c} \tilde{y} \mid \theta \sim Binomial(n = 40, \theta) \\ \theta \sim Beta(a = ?, b = ?) \end{array}}$$

1.7 Choosing the Prior Distribution

Example 1 (continued)

$$\tilde{y}_C \mid \theta_C \sim Binomial(n_C = 40, \theta_C)$$
$$\theta_C \sim Beta(a = ?, b = ?)$$

$\pi(\theta_C)$



$$\underbrace{4\sigma = 0.4}$$
$$\sigma = 0.1$$

$$\theta \sim Beta(a, b)$$
$$E[\theta] = \frac{a}{a+b}$$
$$V[\theta] = \frac{ab}{(a+b)^2(a+b+1)}$$

$$E[\theta] = \frac{a}{a+b} = 0.6$$
$$\sqrt{V[\theta]} = \sqrt{\frac{ab}{(a+b)^2(a+b+1)}} = 0.1$$

$$\Downarrow$$

$$\boxed{\begin{array}{l} a = 13.8 \\ b = 9.2 \end{array}}$$

1.7 Choosing the Prior Distribution

Observations:

1. It is always advisable to plot $\pi(\theta)$.
2. If the information we have is about the sample space then it is also advisable to plot the $p(\tilde{y})$

1.7 Choosing the Prior Distribution

1.7.2 Non-informative priors

It is about choosing $\pi(\theta)$ that least affects the Data information. The most common are:

a) Flat prior

b) Conjugate prior with huge variance

1.7 Choosing the Prior Distribution

a) Flat prior (or Laplace's prior)

All possible values for $\theta \in \Omega$ are equally likely

The posterior distribution becomes the standardized likelihood.

Observation: if Ω is not a closed space and $\pi(\theta)$=const, then its integral is infinite and it is no longer a probability distribution. In this case we say that the prior is improper. However, if the prior is improper it won't be a problem if the posterior is proper (i.e. its integral equals 1).

1.7 Choosing the Prior Distribution

b) Conjugate prior with huge variance

It consists of choosing the parameters' values of the conjugate prior, which makes a huge variance.

1.7 Choosing the Prior Distribution

Example

Given the Bayesian Model: $\tilde{y} \mid \theta \sim Binomial(n, \theta)$

$\theta \sim Beta(a, b)$

and the observed data: $n\ ,\ y$

	Prior	Posterior	Observation
Flat	Beta(1, 1)	Beta(1+y,1+n-y)	Mode(θ|y) is the MLE of θ
Conjugate	Beta(0,001, 0,001)	Beta(y, n-y)	E(θ|y) is the MLE of θ

Chapter 1. Bayesian Model

1.6 The prior and posterior predictive distribution

Summarizing……

Bayesian Model

$$\boxed{\begin{array}{l} \tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta) \\ \theta \sim \pi(\theta) \end{array}} \longrightarrow p(\tilde{y})$$

$\Downarrow$

Data $\Rightarrow$ **Bayes' Th.**

$\Downarrow$

Bayesian Model (updated)

$$\boxed{\begin{array}{l} \tilde{y} \mid \theta \sim p(\tilde{y} \mid \theta) \\ \theta \mid y \sim \pi(\theta \mid y) \end{array}} \longrightarrow p(\tilde{y} \mid y)$$

The updated Bayesian Model allows to make point estimations, interval estimations, predictions, …

Chapter 1
Bayesian Model

Xavi Puig Oriol
Departament d'Estadística i I.O.