Bayesian Methods
for Biomedical Research

—

Lecture notes

Boris Hejblum

https://bayesee.borishejblum.science

ISPED summer school, Université de Bordeaux

June 2024

© Boris Hejblum 2024

This document takes inspiration from the two following books: *The Bayesian Choice* by CP Robert and *Le raisonnement bayésien* by E Parent & J Bernier, along with lecture notes from D Commenges.

Contents

Vocabulary

A few words frequently used in the Bayesian paradigm:

paradigm

Refers to a coherent system of representation of the world, a way of seeing things.

a priori

In Bayesian statistics, the Latin expression *a priori* is widely used. It means *previously* in English, or more precisely *based on data prior to the experiment*. Etymologically this expession comes from "*a priori ratione*" which means in Latin *by a preceding reason*, and opposes *a prosteriori*.

a posteriori

The Latin expression *a posteriori* is also widely used in the Bayesian framework. It means *after the fact* in English, or more precisely *by relying on experience, on facts*. Etymologically this expession comes from "*a posteriori ratione*" which means in Latin *by a reason that comes after*, and opposes *a priori*.

elicitation

Action formalizing an expert's knowledge to enable it to be shared, e.g. to incorporate it into a model.

Course objectives

I. <u>Familiarize oneself with the Bayesian framework:</u>

1. understand and assess a Bayesian modeling strategy, and discuss its underlying assumptions

2. rigorously describe expert knowledge by a quantitative prior distribution

II. <u>Be able to study and perform Bayesian analyses in biomedical applications:</u>

1. understand and discuss assumptions and methodological choices in biomedical literature using Bayesian methods, including "under the hood" estimation machinery

2. understand, discuss and reproduce a Bayesian (re-)estimation of a proportion or a relative risk using ®

3. understand and perform a Bayesian linear regression using ® applied to meta-analysis

4. put into perspective the results from a Bayesian analysis described in a scientific article

Disclaimer

This course will **NOT** teach you how to:

1 design an adaptive clinical trial – for this see Berry (2006). Bayesian clinical trials. *Nature Reviews: Drug Discovery* 5:27–36. `https://doi.org/10.1038/nrd1927`

2 carefully perform a meta-analysis – for this, see Harrer, Cuijpers, Furukawa, Ebert (2019). *Doing Meta-Analysis in R: A Hands-on Guide.* `https://bookdown.org/MathiasHarrer/Doing_Meta_Analysis_in_R/`. DOI: 10.5281/zenodo.2551803.

These lecture notes include mathematical and technical details for the sake of completeness that are not necessary for this course.

Nonetheless, these notes are by no means exhaustive, and the curious reader will be referred to more comprehensive textbooks such as *The Bayesian Choice* by C. Robert.

Some motivational examples from biomedical research

Apologies, this section is a lot about COVID-19.

Diagnostic tests

Let us imagine that someone gets a positive result on a COVID-19 test. But knowing the test is imperfect, they might wonder: *What is the probability that they are actually infected by the SARS-CoV-2 ?* This question was recently tackled in a news article in the Guardian[1]. Good *et al.* focus on the opposite question in a short research report in the *Journal of General Internal Medicine* of what to make of a negative test result[2]. In both instances, Bayesian thinking is paramount to account for the epidemiological context and make inform decision based on the test result. This kind of question bears important consequences in the context of the current pandemic where surgery protocols include pre-operative SARS-CoV-2 negative testing[3].

An additional example related to this topic is the work of Gelman & Carpenter[4] who can incorporate uncertainty about sensitivity and specificity of a given tests for estimating prevalence of COVID-19 in California.

Clinical trial design

Houston *et al.*[5] presented the methodology of the ATTACC trial, an adaptive Bayesian randomized controlled trial, in *Clinical Trials*: "*Using a Bayesian framework, the trial will declare results as soon as pre-specified* **posterior probabilities** *for superiority, futility, or harm are reached. The trial uses response-adaptive randomization to* **maximize the probability that patients will receive the more beneficial treatment approach, as treatment effect information accumulates within the trial**."

Study analyses

The REMAP-CAP trial

An important analysis of drug repositioning against COVID-19 was recently in the *New England Journal of Medicine*[6]. Its goal was to evaluate whether IL-6 receptor antagonists *tocilizumab* and *sarilumab* were able to improve survival for critically ill patients with Covid-19 receiving organ support in ICU based on data from the REMAP-CAP clinical trial (`ClinicalTrials.gov` NCT02735707).The authors used the Bayesian framework to analyse their data and use specific Bayesian concepts when communicating their results and conclusions:

"*The median adjusted cumulative odds ratios were 1.64 (95%* **credible interval**, *1.25 to 2.14) for tocilizumab and 1.76 (95%* **credible interval**, *1.17 to 2.91) for sarilumab as compared with control, yielding* **posterior probabilities of superiority to control of more than 99.9%**

1. https://www.theguardian.com/world/2021/apr/18/obscure-maths-bayes-theorem-reliability-covid-lateral-flow-tests-probability

2. Good *et al.* Interpreting COVID-19 Test Results: a Bayesian Approach. *Journal of General Internal Medicine* 35:2490-2491, 2020. DOI: 10.1007/s11606-020-05918-8

3. see for instance Yang & Nguyen, Re-visiting preoperative SARS-CoV-2 testing using a Bayesian approach, *Can J Anesth*, 67:1690–1691, 2020. DOI: 10.1007/s12630-020-01767-5

4. Gelman & Carpenter, Bayesian analysis of tests with unknown specificity and sensitivity, *JRSS C*, 69(5):1269-1283, 2020. DOI: 10.1111/rssc.12435

5. Anti-Thrombotic Therapy to Ameliorate Complications of COVID-19 (ATTACC): Study design and methodology for an international, adaptive Bayesian randomized controlled trial. *Clinical Trials*, 17(5):491-500, 2020. DOI: 10.1177/1740774520943846.

6. REMAP-CAP Investigators, Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19. *New England Journal of Medicine*, 384(16):1491-1502, 2021. DOI: 10.1056/NEJMoa2100433

and of 99.5%, respectively. *An analysis of 90-day survival showed improved survival in the pooled interleukin-6 receptor antagonist groups, yielding a hazard ratio for the comparison with the control group of 1.61 (95% **credible interval**, 1.25 to 2.08) and a **posterior probability of superiority** of more than 99.9%.*"

The BNT162b2 (*Pfizer-BioNTech*) vaccine against COVID-19

The results from the evaluation and of safety of the BNT162b2 vaccine from *Pfizer* & *BioNTech* have recently been published in the *New England Journal of Medicine*[7]. They also rely on Bayesian indicators to present their results, as shown in Figure 1.

Table 2. Vaccine Efficacy against Covid-19 at Least 7 days after the Second Dose.*						
Efficacy End Point	**BNT162b2**		**Placebo**		**Vaccine Efficacy, % (95% Credible Interval)‡**	**Posterior Probability (Vaccine Efficacy >30%)§**
	No. of Cases	Surveillance Time (n)†	No. of Cases	Surveillance Time (n)†		
	(N=18,198)		**(N=18,325)**			
Covid-19 occurrence at least 7 days after the second dose in participants without evidence of infection	8	2.214 (17,411)	162	2.222 (17,511)	95.0 (90.3–97.6)	>0.9999
	(N=19,965)		**(N=20,172)**			
Covid-19 occurrence at least 7 days after the second dose in participants with and those without evidence of infection	9	2.332 (18,559)	169	2.345 (18,708)	94.6 (89.9–97.3)	>0.9999

* The total population without baseline infection was 36,523; total population including those with and those without prior evidence of infection was 40,137.
† The surveillance time is the total time in 1000 person-years for the given end point across all participants within each group at risk for the end point. The time period for Covid-19 case accrual is from 7 days after the second dose to the end of the surveillance period.
‡ The credible interval for vaccine efficacy was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.
§ Posterior probability was calculated with the use of a beta-binomial model with prior beta (0.700102, 1) adjusted for the surveillance time.

Figure 1 – Table 2 from Polack *et al.*, *NEJM*, 2020

7. Polack *et al.*, Safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine, *New England Journal of Medicine*, 383(27):2603-2615, 2020. DOI: 10.1056/nejmoa2034577

Chapter 1

Introduction to Bayesian statistics

Statistics is a mathematical science, whose objective is to describe what has happened and to assess what may happen in the future. It relies on the observation of natural phenomena in order to propose an interpretation, often through probabilistic models.

1.1 Reminder on frequentist statistics

Frequentist statistics refers to the theory of statistics developed largely by Neyman & Pearson, and based on a deterministic view of the parameters of probabilistic models, which are the very objects that statistical inference seeks to estimate. Maximum likelihood estimation is one of the fundamental tools of frequentist statistics, as is the statistical test theory with its associated confidence interval concept.

1.2 The Bayesian paradigm

1.2.1 Bayes' theorem

The word "Bayesian" comes from the name of Reverend Thomas Bayes. In 1763, the latter publishes an article [1] in which he exposes the following theorem:

$$\mathbb{P}(A|E) = \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E|A)\mathbb{P}(A) + \mathbb{P}(E|\bar{A})\mathbb{P}(\bar{A})} = \frac{\mathbb{P}(E|A)\mathbb{P}(A)}{\mathbb{P}(E)}$$

Posterity refers to this theorem as *Bayes's theorem*, even though the latter actually presents a continuous version in his work:

$$g(x|y) = \frac{f(y|x)g(x)}{\int f(y|x)g(x)\,\mathrm{d}x}$$

where X and Y are two random variables whose realisations are denoted x and y respectively, $f(y|x)$ represents the conditional distribution of Y knowing the realization of X, and $g(x)$ is the marginal distribution of X. The French mathematician Laplace also found these results independently. Laplace and Bayes both further described the uncertainty about the parameters θ of a parametric model $f(y|\theta)$ through a probability distribution π. Bayes' theorem is then written:

$$p(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)\,\mathrm{d}\theta}$$

---

1. T. Bayes, 1763. An essay towards solving a problem in the doctrine of chances, *The Philosophical Transactions of the Royal Society*, **53**: 370-418. (posthumous)

The fundamental difference between the frequentist approach and the Bayesian approach is thus that the latter does not consider the parameters as fixed (i.e. for which there is one true value), but rather as random variables. It is a profound philosophical difference, even if there are many bridges between the two approaches.

This way of considering parameters as random variables induces a marginal probability distribution $\pi(\theta)$ on the parameters. This distribution is called the *prior* or the distribution *a priori*. Its specification is both an asset of Bayesian analysis – since it allows the hypotheses on the subject under study to be formalized and taken into account in the modeling – but also a weakness – since it necessarily introduces subjectivity into the analysis. These two sides of the same coin will be put forward in turn by the Bayesians and their detractors.

1.2.2 Bayesian vs. Frequentists: an outdated debate

The ideas of the Reverend Bayes, found independently and then further explored by Laplace, had a profound influence on the development of statistics during the second half of the 18th century and the 19th century. But with the advent of modern statistics at the turn of the 20th century with Galton and Pearson, then with Fisher and Neymann in particular, frequentist theory became dominant. It was only towards the end of the 20th century that Bayesian statistics came back on the scene, notably thanks to the rise of the computer and the development of efficient numerical methods which made it possible to overcome certain limitations previously present in Bayesian analysis.

Under the influence of Fisher in particular, who firmly rejected Bayesian reasoning, the 20th century saw the statistical community split in two between supporters of the Bayesian approach and supporters of the frequentist approach (considering the parameters as fixed), with sometimes virulent debates opposing the two communities.

Today, these quarrels are outdated, thanks in part to the practical successes of both approaches on modern and complex problems. In addition, a number of methods, such as empirical Bayes methods, lie at the boundary between the two approaches and bridge the gap between them. Today's (bio)statistician must therefore be pragmatic and versatile, integrating Bayesian analysis into his/her toolbox to solve the problems he/she faces.

"*Être ou ne pas être bayésien, là n'est plus la question: il s'agit d'utiliser à bon escient les outils adaptés quand cela est necessaire* – To be, or not to be, Bayesian, that is no longer the question: it is a matter of wisely using the right tools when necessary." Gilbert Saporta

Chapter 2

Bayesian modeling

2.1 Refresher on frequentist modeling

Let us consider a series of *iid* (independent and identically distributed) random variables $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, of which we observe a sample $\boldsymbol{y} = (y_1, \ldots, y_n)$. A frequentist model for their probability distribution is the following probability density family: $f(y|\theta)$, $\theta \in \Theta$. This model assumes there is a "true" distribution of Y characterized by the "true" value of the parameter θ^* which is written $f(y|\theta^*)$. We then want an estimator $\widehat{\theta}$, often one that has good asymptotic properties (generally unbiased for θ^* and with as little variance as possible).

2.2 Historical motivating example

Laplace looked into the probability of birth of girls (rather than boys). To do so, he used the births observed in Paris between 1745 and 1770, during which 241,945 girls and 251,527 boys were born. The question is then: "When a child is born, is it equally likely to be a girl or a boy?"

2.3 Construction of a Bayesian model

The first step in building a model is always to identify the question you want to answer. Once this step is completed, it is a matter of determining what kind of observations are available and will be able to inform our response to the question of interest.

2.3.1 The sampling model

Let us denote the observations avalable $\boldsymbol{y}$. Like a frequentist model, a parametric Bayesian model consists of first proposing a probabilistic model underlying the generation of these observations: $Y_i \stackrel{iid}{\sim} f(y|\theta)$. The latter is called the "sampling model"

In the historical example, Laplace proposed a sampling model based on Bernoulli's law. Let be Y_i the random variable whose value is 1 if the new born i is a girl, and 0 if it is a boy: $Y_i \sim \text{Bernoulli}(\theta)$, where $\theta \in [0, 1]$.

2.3.2 *Prior* distribution

In Bayesian modeling, compared to frequentist modeling, we add a probability distribution (defined on the parmaeters space Θ), called *prior* distribution:

$$\theta \sim \pi(\theta)$$
$$Y_i|\theta \overset{iid}{\sim} f(y|\theta)$$

θ will thus be treated like a random variable, but which is never observed !

In the historical application, Laplace first considered a uniform prior on the probability θ that a newborn would be a girl rather than a boy: $\theta \sim \mathcal{U}_{[0,1]}$

2.3.3 *Posterior* distribution

The purpose of such a Bayesian modeling is to infer the *posterior* distribution of the parameters, i.e. the law of θ conditionally on the observations: $p(\theta|Y)$, which is called distribution *a posteriori* or *posterior* distribution. It is calculated from the sampling model $f(y|\theta)$ – from which we obtain the likelihood $f(\boldsymbol{y}|\theta)$ for all observations – and the *prior* $\pi(\theta)$ thanks to Bayes' theorem:

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

where $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\theta)\pi(\theta)\,\mathrm{d}\theta$ is the marginal law of $\boldsymbol{Y}$.

Example with a uniform *prior*

In the historical example, the likelihood is thus:

$$f(\boldsymbol{y}|\theta) = \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{(1-y_i)} = \theta^{S}(1-\theta)^{n-S}$$

where $S = \sum_{i=1}^{n} y_i$. We then get the following *posterior*:

$$p(\theta|\boldsymbol{y}) = \frac{\theta^{S}(1-\theta)^{n-S}}{f(\boldsymbol{y})}$$

It can be shown that $f(\boldsymbol{y}) = \int_0^1 \theta^S(1-\theta)^{n-S}\,\mathrm{d}\theta = \frac{1}{\binom{n}{S}(n+1)}$ thanks to a series of integration by parts, where $\binom{n}{S} = \frac{n!}{S!(n-S)!}$. The distribution *a posteriori* is finally the following: $p(\theta|\boldsymbol{y}) = \binom{n}{S}(n+1)\theta^S(1-\theta)^{n-S}$. To answer the question of interest, we can then calculate:

$$P(\theta \geq 0.5|\boldsymbol{y}) = \int_{0.5}^{1} p(\theta|\boldsymbol{y}) = \binom{n}{S}(n+1)\int_{0.5}^{1} \theta^S(1-\theta)^{n-S}\,\mathrm{d}\theta$$

Unfortunately, this integral (said to be "incomplete") has no analytical solution. An approximation by a normal distribution however allowed Laplace to conclude that the probability of birth of a girl is lower than that of a boy[1], since he obtained: $P(\theta \geq 0.5|\boldsymbol{y}) \approx 1.15\ 10^{-42}$

---

1. This conclusion has since been confirmed and seems to be valid for the human species in general.

Example of the Beta distribution conjugacy

Let us now use a different *prior* distribution, for instance the Beta(α, β) distribution whose density is written: $f(\theta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ (for $\alpha > 0$ and $\beta > 0$).
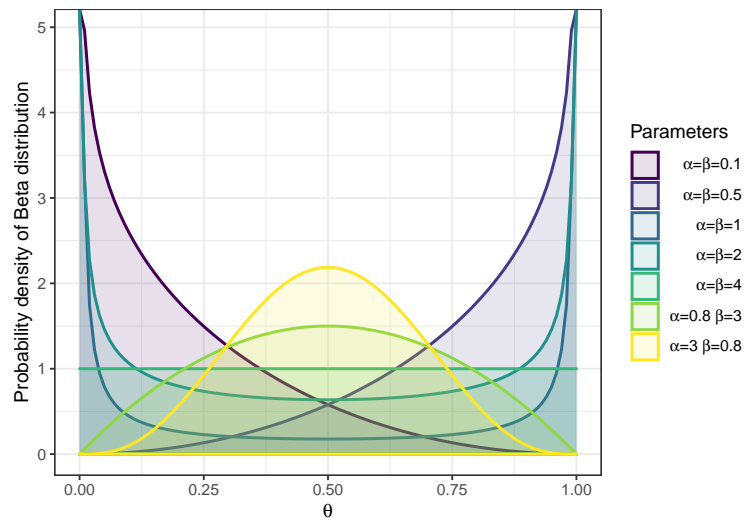
Examples of various parametrizations for the Beta distribution

We notice that the uniform distribution is a special case of the Beta distribution when α and β are both equal to 1. If one recalculate the *posterior* distribution with the *prior* $\pi = \text{Beta}(\alpha, \beta)$, one easily gets:

$$p(\theta|\mathbf{y}) \propto \theta^{\alpha+S-1}(1-\theta)^{\beta+(n-S)-1}$$

where the notation "$\propto$" means "proportional to" (i.e. up to a multiplicative constant). We recognize, up to a normalization constant, the form of a Beta distribution, whose parameter pair would be $(\alpha + S, \beta + (n - S))$. Thus, we deduce from this that $\theta|\mathbf{y} \sim \text{Beta}(\alpha + S, \beta + (n - S))$. This is called a conjugated distribution because the *posterior* and the *prior* belong to the same parametric family.

We can now evaluate the impact of this Beta *prior* on our result based on the choice of hyperparameters α and β. We notice that the *prior* doesn't seem to affect our result here.

Interpretation of the *prior*	Parameters of the Beta distribution	$P(\theta \geq 0.5|\boldsymbol{y})$
#boys > #girls	$\alpha = 0.1, \beta = 3$	$1.08 \ 10^{-42}$
#boys < #girls	$\alpha = 3, \quad \beta = 0.1$	$1.19 \ 10^{-42}$
#boys = #girls	$\alpha = 4, \quad \beta = 4$	$1.15 \ 10^{-42}$
#boys $\neq$ #girls	$\alpha = 0.1, \beta = 0.1$	$1.15 \ 10^{-42}$
non-informative	$\alpha = 1, \quad \beta = 1$	$1.15 \ 10^{-42}$

Table 2.1 – For 493,472 newborns including 241,945 girls

That is because we have a lot of observations at hand. The impact of the *prior* on the *posterior* then becomes very small compared to the amount of information provided by the observations. If we imagine that we had observed only 20 births, including 9 girls, then we notice a much greater influence of the *prior*.

Interpretation of the *prior*	Parameters of the Beta distribution	$P(\theta \geq 0.5 \| \boldsymbol{y})$
#boys > #girls	$\alpha = 0.1, \beta = 3$	0.39
#boys < #girls	$\alpha = 3,\quad \beta = 0.1$	0.52
#boys = #girls	$\alpha = 4,\quad \beta = 4$	0.46
#boys ≠ #girls	$\alpha = 0.1, \beta = 0.1$	0.45
non-informative	$\alpha = 1,\quad \beta = 1$	0.45

Table 2.2 – For 20 newborns including 9 girls
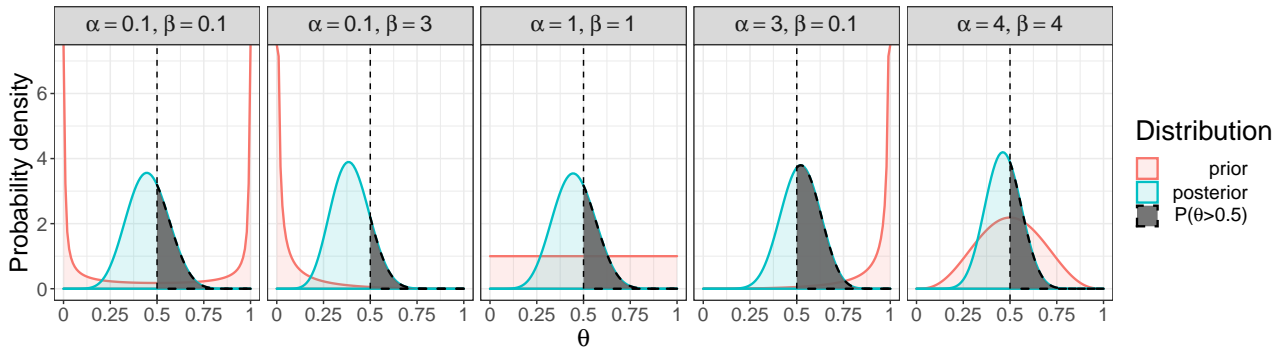
Impact of different Beta priors for 20 observed births

2.3.4 The thorny question of the *prior* choice

An essential feature of the Bayesian approach is thus to have a distribution on the parameters. In Bayesian inference, we start from a distribution *a priori*, and the information contained in the observations is used to obtain the distribution *a posteriori*. The *a priori* distribution brings flexibility compared to a frequentist model, by allowing external knowledge to be incorporated into the model. For example, this may solve identifiability problems sometimes encountered by a purely frequentist approach when the information provided by the observations is not sufficient to estimate all the parameters of interest.

This is a great advantage of the Bayesian approach. But on the other hand, the choice of this distribution on the parameters introduces an intrinsic subjectivity into the analysis, which can be critisized. For example, a statistician working for a pharmaceutical company could choose a *prior* distribution giving a high probability that a drug is effective, which will necessarily influence the result. The choice (or elicitation) of this *prior* distribution is therefore sensitive.

First of all, let us make two theoretical remarks:

1 the support of the *posterior* must be included in the support of the *prior*. In other words, if $\pi(\theta) = 0$, then $p(\theta|\boldsymbol{y}) = 0$.

2 in general we assume the independence of the different parameters *a priori* (when there is more than one parameter – which is almost always the case in applications), which allows to elicit the *priors* parameter by parameter.

Prior Elicitation

There exist strategies to communicate with non-statistical experts to transform their *prior* **knowledge** into *prior* **distribution**.

The simplest method is to ask the experts to give weights (or probabilities) to ranges of values: this is the "histogram method". However, when the parameter can take values in an

unbounded scale, this method might give a zero *prior* for parameter values that are nevertheless possible. . .

Another approach is to give ourselves a parametric family of distributions $p(\theta|\eta)$ and to choose η so that the *prior* distribution is in agreement with what the experts think for specific characteristics of the problem (for example, the mean and variance, or simple quantiles such as quartiles, could coincide with their views). This solves the support problem raised by the histogram method. However, the choice of the parametric family can be important. For example, a normal $\mathcal{N}(0, 2.19)$ distribution has the same quartiles as a Cauchy $\mathcal{C}(0, 1)$ distribution (namely -1, 0, 1). But these two *priors* can give quite different distributions *a posteriori*. One strategy for determining quartiles is to ask the following questions for instance:

— for the median: *Can you determine a value such that theta is as likely to be above or to be below ?*
— then for the first quartile: *Suppose you are told that θ is below [a given median value], can you then determine a new value such that θ is as likely to be above or to be below?*
— similarly the third quartile is determined. . .

Software exists to help elicit *prior* distribution by experts: see for example the academic tool SHELF [2].

One can also elicit *priors* from the literature. The idea is to define the moments of the *prior* such that they give a reasonable probability to the parameter values that have been identified in the literature. If we propose a normal distribution $\mathcal{N}(\mu, \sigma^2)$, we can for example choose μ and σ so that the smallest value given in the literature is equal to $\mu - 1.96\sigma$ and the largest to $\mu + 1.96\sigma$ (a trained eye will have recognised the 2.5% and 97.5% quantiles of the normal distribution). A more elaborate approach is to maximize the likelihood of literature values. . .

The quest for non-informative *priors*

For some parameters (or even all parameters) it's common that one have no prior knowledge whatsoever. One can then try to define a "non-informative" *prior* distribution. For example if the parameter is the probability that a coin will fall on heads or tails, a non-informative distribution could (at first glance) be the uniform distribution on $[0, 1]$ (Bayes' historical choice in 1763). However, two major difficulties emerge:

1 **Improper distributions**

   The first challenge is that this can lead to consider improper distributions. An improper distribution is characterized by a density which does not sum to one. For example, for a mean parameter of a normal distribution, it may seem natural to define a constant *prior* with density $\pi(\theta) = c$ (i.e. all possible values on $]-\infty, +\infty[$ have the same probability). Of course $\int_{-\infty}^{\infty} c \, d\theta = \infty$, and such a choice does not define a probability distribution ! It is however **acceptable because the *posterior* is** (most of the time) **proper**. Indeed:

$$p(\theta|y) = \frac{f(y|\theta)c}{\int f(y|\theta)c \, d\theta}$$

   If $\int f(y|\theta)c \, d\theta = K$ (as it is often the case), then $p(\theta|y) = \frac{f(y|\theta)}{K}$ is a proper density (i.e. which sums to 1).

---

2. SHELF Software at `http://www.tonyohagan.co.uk/shelf/` and the user-friendly Ⓡ package https://CRAN.R-project.org/package=SHELF

2 Non-invariant distributions

The second challenge comes from the non-invariance of the uniform distribution for non-linear transformations of parameters. Indeed if we make a transformation of the parameters $\gamma = g(\theta)$, the density of γ is written: $\pi(\gamma) = |J| \pi(\theta)$, where $|J|$ is the Jacobian of the transformation (i.e. the determinant of the Jacobian matrix $J = \frac{\partial g^{-1}(\gamma)}{\partial \gamma}$). For example if we take a uniform density equal to 1 for θ on $(0, +\infty)$ and we do the transformation $\gamma = \log(\theta)$, we have $g^{-1}(\gamma) = e^{\gamma}$ and $|J| = e^{\gamma}$. So we have $\pi(\gamma) = e^{\gamma}$, which is not the characterization of a uniform distribution. Hence the following paradox: if the uniform distribution for θ reflects a total absence of *a priori* knowledge on θ, we should also have a total absence of *a priori* knowledge on γ, which should translate into a uniform distribution on γ. But that cannot be true. Thus the uniform distribution cannot generally be the distribution representing an absence of *prior* knowledge. This is a central argument which led Fisher, in 1922, to propose the maximum likelihood estimator, possessing an invariance property for non-linear transformations of parameters.

NB: This does not mean that one cannot take a uniform distribution as their *prior*, but one must keep in mind that the uniform distribution only applies to a specific parameterization. . .

To tackle these challenges, various solutions have been proposed. They have shown that there is no such thing as a completely non-informative *prior* distribution, but some can be considered as **weakly informative**.

Jeffreys' *priors*

Perhaps the most successful approach to weakly informative *priors* is that of Jeffreys. The latter proposed a procedure to find a *prior* distribution with an invariance property with respect to parameterization. In the univariate case, Jeffreys' *prior* is defined by:

$$\pi(\theta) \propto \sqrt{I(\theta)}$$

where I is Fisher's information matrix (as a reminder, $I(\theta) = -\mathbb{E}_{Y|\theta}\left[\frac{\partial^2 \log(f(y|\theta))}{\partial \theta^2}\right]$). Jeffreys' *prior* is therefore invariant for bijective transformations of the parameters. That is, if we consider another parameterization $\gamma = g(\theta)$ (for which there is reciprocal bijection g^{-1}), we always get:

$$\pi(\gamma) \propto \sqrt{I(\gamma)}$$

while $\pi(\gamma)$ still corresponds to the same *prior* on θ.

In the multidimensional case (the most common) Jeffreys' *prior* is defined as:

$$\pi(\theta) \propto \sqrt{|I(\theta)|}$$

where $|I(\theta)|$ is the determinant of Fisher's information matrix $I(\theta)$. However this method is rarely used in practice because on the one hand calculations can be hard, and on the other hand it can give somewhat curious results. Indeed, in the case of a normal likelihood where we have 2 parameters θ and σ for example, Jeffreys' multidimensional *prior* is $1/\sigma^2$, which is different from $\pi(\sigma) = 1/\sigma$ obtained in the unidimensional case. . . In practice the tendency is generally to apply Jeffreys' *prior* separately for each parameter, and to define the joint distribution *a priori* by multiplying the *priors* for each parameter (thus making an independence hypothesis *a priori*). For the normal example with two parameters, we get $\pi(\theta, \sigma) = 1/\sigma$. But we notice it's not really Jeffreys' two-dimensional prior anymore. . .

Diffuse *priors*

In practice, a very common alternative for giving a weakly informative *prior* is the use of parametric distribution (such as the normal distribution) with very large variance parameters (which approaches the uniform law while avoiding the problem of improper distributions).

2.4   Going further

2.4.1   Hyper-*priors* & hierarchical models

In classical Bayesian modeling, we consider two hierarchical levels: first $\pi(\theta)$, then $f(\boldsymbol{y}|\theta)$. It is possible to add a level by also putting a *prior* onto the η parameter of $\pi(\theta)$, called a hyper-parameter: $\pi(\theta|\eta)$. Applying the Bayesian approach, we can give this hyper-parameter a *prior* distribution, then called hyper-*prior* and denoted $h(\eta)$. The distribution is:

$$p(\theta|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\theta)\pi(\theta)}{f(\boldsymbol{y})} = \frac{\int f(\boldsymbol{y}|\theta)\pi(\theta|\eta)h(\eta)d\eta}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\theta)\int \pi(\theta|\eta)h(\eta)d\eta}{f(\boldsymbol{y})}$$

We notice that this hierarchical modeling with three hierarchical levels is equivalent to a Bayesian modeling with two levels and a *prior* distribution which becomes: $\pi(\theta) = \int \pi(\theta|\eta)h(\eta)d\eta$. Nevertheless, this hierarchical construction can facilitate the modeling stage as well as the elicitation of the *prior*. It is even possible to build models with more than three levels, considering that the distribution of η depends itself on "hyper-hyper-parameters", and so on... A typical use case for hierarchical Bayesian modeling is the inclusion of random effects in the linear model. Latent class models are another example. We remark here that the boundary between frequentist and Bayesian modeling is becoming thinner, and that it is mainly a matter of interpretation of the model parameters (and therefore of the results).

If we go back once more to the historical example of birth sex with a Beta *prior*, one can propose two Gamma hyper-*priors* for α and β:

$$\alpha \sim \text{Gamma}(4, 0.5)$$
$$\beta \sim \text{Gamma}(4, 0.5)$$
$$\theta|\alpha, \beta \sim \text{Beta}(\alpha, \beta)$$
$$Y_i|\theta \overset{iid}{\sim} \text{Bernoulli}(\theta)$$

2.4.2   Empirical Bayes

The empirical Bayes strategy consists of eliciting the *prior* according to its empirical marginal distribution, and therefore to estimate the *prior* from the data. This means giving ourselves hyper-parameters and trying to estimate them through frequentist methods (for example by maximum likelihood) by $\hat{\eta}$, before plugging this estimate into the *prior* distribution and thus obtaining the *posterior* distribution $p(\theta|\boldsymbol{y}, \hat{\eta})$. This empirical Bayes approach that combines Bayesian and frequentist may seem to go against the idea of an *a priori*, since the data are already used to define the *prior*. Nevertheless, one can see the empirical Bayes strategy as an approximation of the completely Bayesian approach. Compared to a weakly informative *prior* it gives a more concentrated *posterior* (decreased variance), at the cost of introducing a bias in the estimate (we use the data twice !). This approach illustrates once again the trade-off between bias and variance that is typical in any estimation procedure.

2.4.3 Sequential Bayes

Note that Bayes' theorem can be used sequentially. Omitting the denominator (which does not depend on θ) one can write: $p(\theta|\boldsymbol{y}) \propto f(\boldsymbol{y}|\theta)\pi(\theta)$. If $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$, one gets: $p(\theta|\boldsymbol{y}) \propto f(\boldsymbol{y}_2|\theta)f(\boldsymbol{y}_1|\theta)\pi(\theta) \propto f(\boldsymbol{y}_2|\theta)p(\theta|\boldsymbol{y}_1)$. The *posterior* knowing $\boldsymbol{y}_1$ becomes the *prior* for the new observation $\boldsymbol{y}_2$. So we can update the information on θ as the observations arrive.

2.5 Bayesian Inference

Once Bayesian modeling is complete, the *posterior* distribution (obtained by choosing the *prior* distribution, the sampling model and the observed data) is available. This distribution contains all the information on θ conditionally on the model and the data. One can nevertheless be interested in summaries of this distribution, for example in a central parameter of this distribution such as its expectation, its mode or its median... Those are akin to point estimators obtained by frequentist analysis.

2.5.1 Decision theory

Statistical decision theory is generally used when estimating an unknown parameter θ. The decision then deals with the choice of a point estimator $\widehat{\theta}$. In order to determine the optimal $\widehat{\theta}$, a **cost function** is defined (with a value in $[0, +\infty[$) representing the penalty associated with the choice of a particular $\widehat{\theta}$ (that is, the associated decision). In order to determine the optimal $\widehat{\theta}$ (i.e. the optimal decision) one will want to minimize the chosen cost function. Note that a large number of different cost functions are possible, and that each of them results in a different optimal point estimator, and therefore a specific optimal decision. This adds an additional layer of subjectivity to an analysis, and it is good practice to perform a sensitivity analysis to quantify the impact of the cost function choice on the result of an analysis.

2.5.2 Point estimate

We now present several point estimators widely used in Bayesain inference.

Posterior mean

The *posterior* mean is defined as:

$$\mu_P = \mathbb{E}(\theta|\boldsymbol{y}) = \mathbb{E}_{\theta|\boldsymbol{y}}(\theta)$$

It is the estimator with the smallest *posterior* variance (in the Bayesian meaning: $\mathbb{E}_{\theta|\boldsymbol{y}}(\theta - \hat{\theta})^2$). It is therefore the optimal point estimator in the sense of the quadratic error (quadratic loss function). Note that the calculation of this expectation is not always easy, because it assumes the calculation of an integral...

Maximum *A Posteriori*

The maximum has been used a lot, especially since it is easier (or at least less difficult) to compute. Indeed, it does not require any integral calculation, but just a simple maximization of $f(\boldsymbol{y}|\theta)\pi(\theta)$ (because the denominator $f(\boldsymbol{y})$ does not depend on θ). This mode estimator is called the **maximum *a posteriori*** (often denoted **MAP**).

BF	Strength of evidence
< 1	Negative (supports H_0)
1 to $10^{1/2}$	Barely worth mentioning
$10^{3/2}$ to 10	Substantial
10 to $10^{3/2}$	Strong
103/2 to 100	Very strong
> 100	Decisive

Table 2.3 – Jeffreys' scale for interpreting Bayes factors

Posterior median

The median is also a possible summary of the *posterior* distribution. As its name suggests, this is the median of $p(\theta|\boldsymbol{y})$. This is the optimal point estimator in the sense of the absolute error (linear loss function) $|\theta - \hat{\theta}|$.

2.5.3 Credibility interval

Finally we can define a set of values with a high *posterior* probability of occurrence. Such a set is called a **credibility set**. If the *posterior* distribution is unimodal, such a set is an interval. For example, a **95% credibility interval** is an interval $[t_{inf}; t_{sup}]$ such that $\int_{t_{inf}}^{t_{sup}} p(\theta|\boldsymbol{y}) \, \mathrm{d}\theta = 0.95$. We're usually interested in the shortest possible 95% credibility interval (also called Highest Density Interval).

Let us recall here the interpretation of a frequentist confidence interval at a 95% level, which is interpreted as follows, with respect to all the intervals of this level that could have been observed:

> *95% of the intervals computed on all possible samples (all those that can be observed) contain the true value θ*

☙ one cannot interpret a realization of a confidence interval in probabilistic terms ! It is a common mistake... The credibility interval is interpreted much more naturally, as an interval that has a 95% chance of containing θ (for a 95% level, obviously)

2.5.4 Bayes Factor

The **Bayes Factor** is the marginal likelihood ratio between two hypotheses (e.g. H_1 and H_0):

$$BF_{10} = \frac{f(\boldsymbol{y}|H_1)}{f(\boldsymbol{y}|H_0)}$$

It is interpreted in terms of favored support for either hypothesis from the observed data $\boldsymbol{y}$. It can be used in a Bayesian analysis to perform model selection, and notably to quantify the benefit of incorporating one additional parameter in the model. Jeffreys proposed a scale for interpreting the values of Bayes factors that is shown in Table 2.3. The **posterior odds** between those two hypotheses can then be computed as:

$$\frac{p(H_1|\boldsymbol{y})}{p(H_0|\boldsymbol{y})} = BF_{10} \times \frac{p(H_1)}{p(H_0)}$$

If the prior probability is the same for both hypotheses $p(H_0) = p(H_1)$, then the posterior odds are equal to the Bayes Factor.

2.5.5 Asymptotic – and frequentist– properties of the *posterior* distribution

Doob's convergence theorem

A very interesting result is the asymptotic behavior of the *posterior* distribution under certain hypotheses (*iid* observations, densities three times differentiable, existence of moments of order 2). There is a first result, Doob's convergence theorem, which ensures that the distribution concentrates around the true value of the parameter when $n \to \infty$. We can note it (convergence in distribution):

$$p(\theta|\boldsymbol{y}_n) \xrightarrow{\mathcal{L}} \delta_{\theta^*}$$
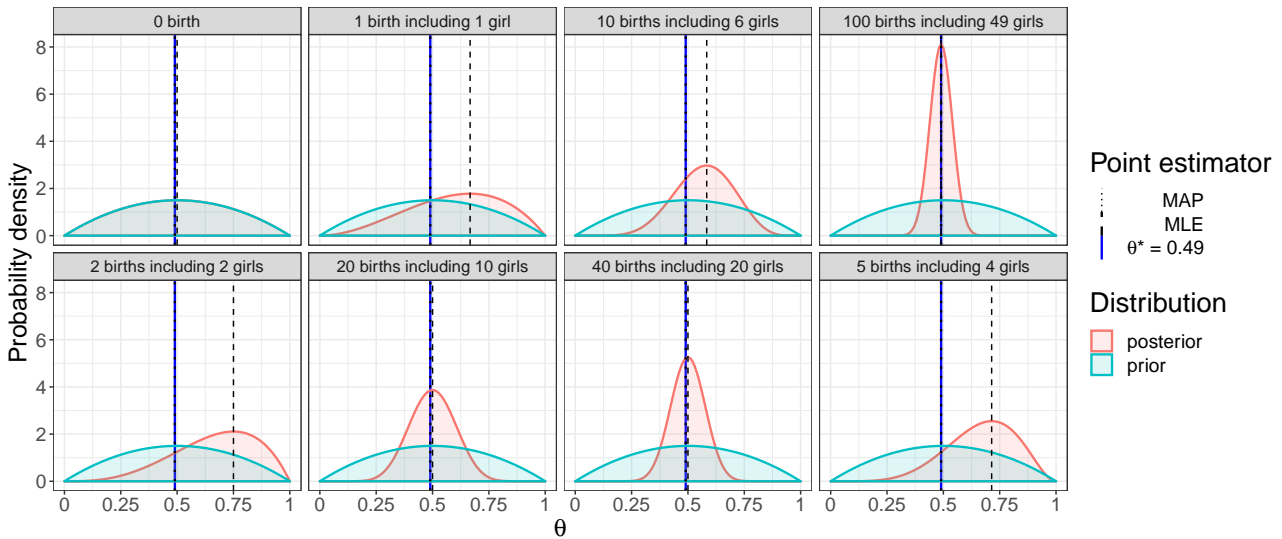
Historical example: posterior concentration around θ* with n

Bernstein-von Mises Theorem

A richer result characterizes the asymptotic distribution of θ: the **Bernstein-von Mises theorem** (also called **Bayesian central-limit theorem**). For a large n the *posterior* distribution $p(\theta|\boldsymbol{y})$ can be approximated by a normal distribution centered at the mode $\hat{\theta}$ and for variance the inverse of the Hessian (i.e. the second derivative) of $p(\theta|\boldsymbol{y})$ with respect to θ taken at the mode θ. One can then write the following approximation:

$$p(\theta|\boldsymbol{y}) \approx \mathcal{N}(\hat{\theta}, I(\hat{\theta})^{-1})$$

This results is important for two reasons:

— it can be used to explain why Bayesian methods and frequentist procedures based on maximum likelihood give, for large enough n, very close results. Thus, in dimension 1, the asymptotic credibility interval is $[\hat{\theta}\sqrt{I(\hat{\theta})^{-1}}]$, and compared to the frequentist confidence interval constructed from the estimator's asymptotic distribution: $[\hat{\theta}_{MLE} \pm 1.96\sqrt{I(\hat{\theta}_{MLE})^{-1}}]$ (where $I(\hat{\theta}_{MLE})$ is here the observed Fisher information matrix, and corresponds to the previous definition for uniform *priors*). We note that they are both identical (for a uniform *prior*). For these *priors*, we note that we also have $\hat{\theta} = \hat{\theta}_{MLE}$ (and even if we don't take uniform *priors*, the estimators and intervals are very close, since the weight of the *prior* becomes negligible when $n \to \infty$). The theoretical interpretation of these intervals obviously remains different.

— it means that we can approximate the *posterior* distribution by a normal distribution, whose mean and variance we can calculate simply using the MAP, and thus facilitate numerical calculations of Bayesian inference.

2.6 Conclusion and perspective on Bayesian modeling

2.6.1 Essential concepts

1 **Bayesian modeling**:

$$\theta \sim \pi(\theta) \quad \text{the } prior$$
$$Y_i|\theta \overset{iid}{\sim} f(y|\theta) \quad \text{sampling model}$$

2 **Bayes' formula**:

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

where $p(\theta|\boldsymbol{y})$ is the *posterior* distribution, $f(\boldsymbol{y}|\theta)$ is the likelihood (inherited from the sampling model), $\pi(\theta)$ is the *prior* distribution on the parameters θ and $f(\boldsymbol{y}) = \int f(\boldsymbol{y}|\theta)\pi(\theta)$ is the marginal distribution of the data, i.e. the normalizing constant (with respect to θ).

3 **The *posterior* distribution is given by**:

$$p(\theta|\boldsymbol{y}) \propto f(\boldsymbol{y}|\theta)\pi(\theta)$$

4 **Weekly informative *priors***

5 ***Posterior* mean, MAP, and credibility intervals**

2.6.2 Usefulness of the Bayesian approach as an analytical tool

The Bayesian framework is a statistical tool for data analysis, on the same footing as other methodologies such as random forests, dimension reduction methods, latent class models, etc. It is particularly useful when few observations only are available and frequentist methods do not yield any or satisfactory results (e.g., logistic regression with very little or no event, i.e. a lot or even only 0 in the case of extremely rare events) and/or when there is important knowledge *a priori* that can be integrated into a model with few observations (for example the model used by *FiveThirtyEight* to predict the results of the 2008 American elections in each American state, in some of which only few polls were conducted, or in genomic studies where the number of observations available for each gene is generally relatively small while many genes are observed). In a few instances, the Bayesian framework allows the definition of models with features that currently cannot be replicated in the frequentist framework, e.g. in Bayesian nonparametrics. Like any statistical method, Bayesian analysis has advantages and disadvantages that will be more or less important depending on the application considered.

Chapter 3

Numerical computation for Bayesian analysis

3.1 Estimating the posterior distribution is often costly

3.1.1 Multidimensional parameters

In real-world data analysis, there are often several parameters. The vector of parameters is thus $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$. Bayes' formula yielding the *posterior* distribution from both the prior and the likelihood is still valid: it gives the joint posterior ditribuion of all the parameters. All the information is encoded in this joint distribution. Unfortunately, its numerical calculation is not always easy – especially for complex models (and in some models, even the likelihood is difficult to compute). In addition, to obtain the joint posterior, it is also necessary to compute the normalizing constant $f(\boldsymbol{y}) = \int_{\Theta}^{d} f(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$. An analytical solution is only available in very special cases (especially when using conjugate distributions); and in the majority of practical cases, this normalizing integral must in fact be calculated numerically. If θ is d-dimension, this means calculating an integral of complexity d, which is difficult when d is large (serious numerical problems appear as soon as $d > 4$).

An even more challenging issue arises when drawing conclusions from this joint distribution. In general we are interested in the possible values for each parameter. This means that we need the marginal, unidimensional distribution of each parameter. For a given parameter, to obtain it, it is then necessary to integrate out the $d - 1$ other parameters from the joint distribution (and this d times, for each of the d parameters). The problem is all the more difficult because it is necessary to calculate these integrals for each possible value of the parameter, in order to reconstitute the full posterior probability density. In complex problems, a sufficiently precise calculation of these integrals seems unfeasible, and algorithms based on sampling simulations are generally used, in particular the so-called **Markov chain Monte Carlo** (MCMC) algorithms.

3.1.2 Computational Bayesian statistics

Finding the posterior appears simple in theory thanks to Bayes' theorem. But in practice the calculation of the normalizing integral to the denominator is often extremely difficult. Finding an analytical expression is only possible in a few very special cases, and numerical evaluation can be just as difficult, especially when the dimension of the parameter space increases.

Computational Bayesian statistics is looking for solutions to be able to estimate the distribution *a posteriori*, even when only the numerator in Bayes' theorem is known (non-normalized posterior). The main methods used are based on sampling algorithms to generate a sample distributed according to the posterior distribution. Among these algorithms, two main categories

can be distinguished: (i) first, direct sampling methods, where a sample is generated from a simple (e. g. uniform) distribution and then transformed so that the result is distributed according to the posterior; (ii) Monte Carlo Markov chain methods (MCMC), where a Markov chain is constructed on the space of parameters whose invariant probability distribution matches the posterior distribution.

3.1.3   Monte Carlo method

Monte Carlo (1955) was the encrypted name of a project by John von Neumann and Stanislas Ulam at the *Los Alamos Scientific Laboratory* to use random numbers to estimate quantities that are difficult (or impossible) to calculate analytically.

Using the law of large numbers, the aim is to construct a Monte Carlo sample to calculate various functions using the probability distribution followed by this very sample. Indeed $\mathbb{E}[f(X)] = \int_x f(x)p(x)dx$. However, thanks to the law of large numbers, we have: $\mathbb{E}[f(X)] = \frac{1}{N}f(x)$ provided that the x forms a *iid* sample according to p, the probability distribution of X. We can thus calculate a certain number of integrals, provided we are able to sample according to p_X.

Example: Estimating $\pi = 3.14\ldots$ with random numbers
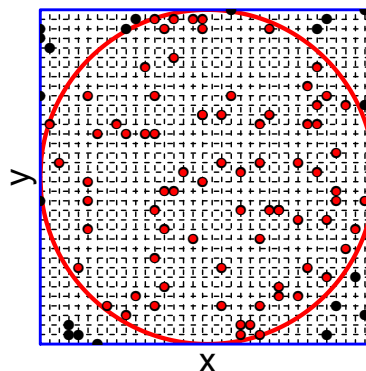
Casino roulette (in Monte Carlo ?)   A target with a 36×36 superimposed grid

1. The probabilty to be inside the disk is the ratio between the disk and the square surfaces:
$$p_C = \frac{\pi R^2}{(2R)^2} = \frac{\pi}{4}$$

2. Let us sample n points $((x_{11}, x_{21}), \ldots, (x_{1n}, x_{2n})) = (P_1, \ldots, P_n)$ in the 36×36 coordinate system defined by the grid, thanks to the roulette which will générate the coordinate one by one.

3. Put those samples onto the grid and count how many lands inside the disk.

4. Compute the ratio (i.e. the estimated probability of being inside the disk) :

$$\widehat{p}_C = \frac{\sum P_i \in circle}{n}$$

If $n = 1000$ and we find 786 points are in the circle, then we have $\widehat{\pi} = 4 \times \frac{786}{1000} = 3.144$. We could further improve our estimate by increasing the resolution of our grid, and also by

increasing our number n of sampled points. Indeed, we have $\lim_{n \to +\infty} \widehat{p}_C = p_C$ according to the law of large numbers.

We thus built a Monte Carlo sample, from which we can calculate many functions, including π which corresponds to 4 times the probability of being in the circle !

Similarly, direct or MCMC sampling methods seek to construct a Monte Carlo sample following the posterior distribution, in order to calculate a number of functions (posterior averages, credibility intervals, *etc*) from it.

3.2   Direct sampling methods

3.2.1   Generation of random numbers according to usual probability distributions

There are several ways to generate so-called random numbers according to known distributions. The vast majority of computer programs does not generate completely random numbers. Rather, we are talking about **pseudo-random numbers**, which seem random but are actually generated according to a deterministic process (which depends in particular on a "seed").

The Uniform distribution

To generate a pseudo-random sample according to the uniform distribution on $[0, 1]$, we can give the example of the linear congruential algorithm (Lehmer, 1948):

> 1 Generate a sequence of integers y_n such as:
> $$y_{n+1} = (ay_n + b) \text{ mod. } m$$
> 2 $x_n = \dfrac{y_n}{m-1}$
>
> Choose a, b and m so that y_n has a very long period and that $(x_1, \ldots, x_n)$ can be considered as *iid*

where y_0 is the so-called "seed". We notice that we necessarily have $0 \leq y_n \leq m-1$. In practice we take m very large (for example 2^{19937}, the default in R which uses the Mersenne-Twister variation of this algorithm). In this course, we will not focus on the generation of pseudo-random according to the uniform distribution on $[0, 1]$, this is a tool that we will consider reliable and that is used by the different algorithms detailed later on.

Other distributions

To sample according to the binomial distribution $Bin(n, p)$, we can use the **relationships between the different usual distributions**, starting from $U_i \sim \mathcal{U}_{[0,1]}$:

$$Y_i = \mathbb{1}_{U_i \leq p} \sim \text{Bernoulli}(p)$$

$$X = \sum_{i=1}^{n} Y_i \sim Bin(n, p)$$

To sample according to the Normal law $N(0.1)$, we can use the Box-Müller:
If U_1 and U_2 are 2 uniform variables $]0; 1]$ independent, then

$$Y_1 = \sqrt{-2 \log U_1} \cos(2\pi U_2),$$

$$Y_2 = \sqrt{-2 \log U_1} \sin(2\pi U_2)$$

are independent and each follow the Normal law $N(0.1)$.

3.2.2 Sampling according to a distribution defined analytically

Inverse transform sampling

Definition: **Generalized inverse**

For a function F defined on $\mathbb{R}$, its generalized inverse is defined as

$$F^{-1}(u) = \inf\{x; F(x) > u\}$$

Property: Let F be the cumulative distribution function (cdf) corresponding to a given probability distribution, and let U be a random variable following a uniform distribution on $[0, 1]$. Then $F^{-1}(U)$ defines a random variable whose cumulative distribution function is F.

We deduce from the above property that if we know the distribution function of the law according to which we want to simulate, and if we are able to reverse it, then we can generate a sample according to this law from a uniform sample on $[0, 1]$.

Example: We want to sample according to the exponential distribution of parameter λ.

We know the density function of the exponential probability distribution, which is $f(x) = \lambda exp(-\lambda x)$, as well as the corresponding cumulative distribution function (its integral), which is $F(x) = 1 - exp(-\lambda x)$.

Let's pose $F(x) = u$. One notices then that $x = -\dfrac{1}{\lambda} \log(1 - u) = F^{-1}(u)$.

If $u \sim \mathcal{U}_{[0,1]}$, then $x \sim Exp(\lambda)$.

Acceptance-rejection method

The acceptance-rejection method consists in using an instrumental distribution g, which we know how to sample from, in order to sample according to the target distribution f. The general principle is to choose g close to f, to propose samples according to g, and to accept some and reject others in order to obtain a sample according to the f distribution in the end.
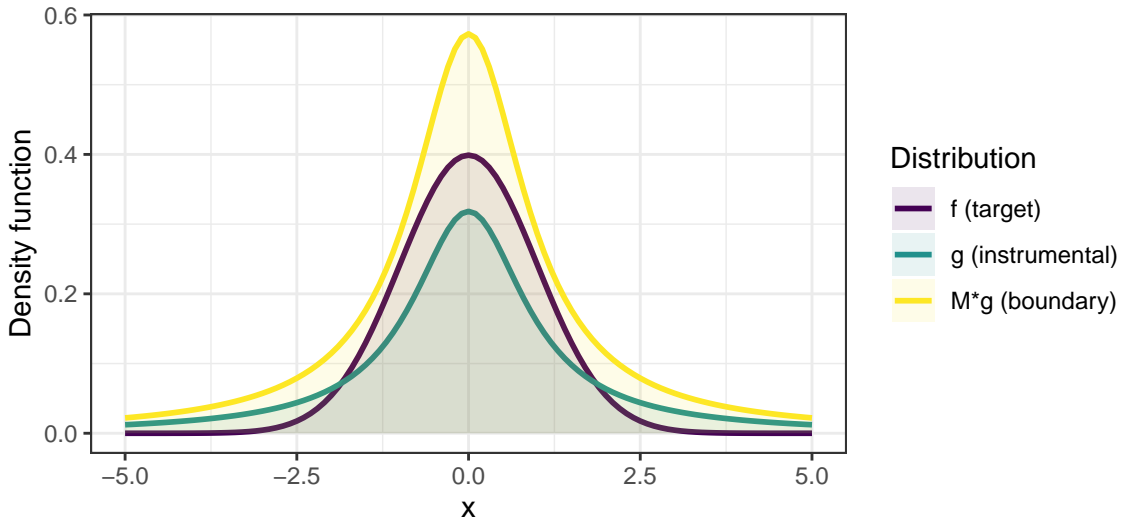

Let f be the density function from a probability distribution of interest.

Let g be the density function from a probability distribution (from which one knows how to sample) such that, for all x:

$$f(x) \leq Mg(x)$$

While $i \leq n$:

    1 Sample $x_i \sim g$ and $u_i \sim \mathcal{U}_{[0,1]}$

    2 If $u_i \leq \dfrac{f(x_i)}{Mg(x_i)}$, **accept** the draw:
       $y_i := x_i$
       else **reject** it and return to 1.

$(y_1, \ldots, y_n) \overset{iid}{\sim} f$

Example of a proposal and a target distribution for the accept–reject algorithm

The smaller the M, the lower the rejection rate and the more efficient the algorithm is (in the sense that it requires less iteration to obtain a sample size of n). It is therefore advisable to choose g as close as possible to f, especially when the dimension increases (the impact of M being all the more important then). Nevertheless, the proposal law will necessarily have heavier queues than the target law, in all dimensions of the parameter space. Because of the scourge of size, when the number of parameters increases, the acceptance rate decreases very quickly.

Exercise 1: Construct a pseudo sample of size n according to the following discrete law(multinomial distribution with m elements $\{x_1, \dots, x_m\}$):

$$p_1 \delta_{x_1}(x) + p_2 \delta_{x_2}(x) + \dots p_m \delta_{x_m}(x) \qquad \text{with } \sum_{i=1}^{m} p_i = 1$$

where $\delta_y(x) = 1$ if $x = y$, and 0 otherwise.

Exercise 2: Using the inversion method, generate a sample size according to a Cauchy's law (whose density is $f(x) = \dfrac{1}{\pi(1 + x^2)}$), knowing that $\arctan'(x) = \dfrac{1}{(1 + x^2)}$.

Exercise 3: Write an acceptance-rejection algorithm to simulate the realization of a sample size n of a normal law $N(0.1)$ using a Cauchy law as a proposal. Find the optimal M value.

3.3 MCMC algorithms

The principle of MCMC algorithms is to build a Markov chain visiting the parameter space, whose invariant probability law is the posterior distribution.

3.3.1 Markov chains: a primer

A (discrete-time) Markov chain is a discrete-time stochastic process. It can be defined as a sequence of random variables X_0, X_1, X_2, X_3, ... (all defined on the same space) with the **Markov property** ("memoryless"):

$$p(X_i = x | X_0 = x_0, X_1 = x_1, \dots, X_{i-1} = x_{i-1}) = p(X_i = x | X_{i-1} = x_{i-1})$$

The set of possible values for X_i is called **state space** and is noted E.

Remark: Continuous-time Markov chains can also be defined, then requiring discrete state space, but such mathematical objects are beyond the scope of this course.

A Markov chain is defined by 2 parameters:
1 its initial distribution $p(X_0)$
2 its transition probabilities $T(x, A) = p(X_i \in A | X_{i-1} = x)$

Remark In the following, we will only consider Markov chains **homogeneous**, i. e. who checks:

$$p(X_{i+1} = x | X_i = y) = p(X_i = x | X_{i-1} = y)$$

Property: A Markov chain is said to be **irreducible** if all sets of non-zero probability can be reached from any starting point (i.e. any state is accessible from any other).

Property: A Markov chain is said to be **recurrent** if the trajectories (X_i) pass an infinite number of times in any set of non-zero probability of the state space.

Property: A Markov chain is said to be **aperiodic** if nothing induces periodic behavior of the trajectories.

Definition: A probability distribution $\tilde{p}$ is called **invariant law** (or **stationary law**) for a Markov string if it verifies the following property: if X_i follows $\tilde{p}$, then X_{i+1} (and the following items) are necessarily distributed according to $\tilde{p}$.

*Remark*: A Markov chain can admit several stationary laws.

Ergodic theorem: A positive irreducible and recurrent Markov chain (i.e. the average return time is finite) admits a single invariant probability distribution $\tilde{p}$ and converges almost certainly towards it (if it is also aperiodic, then it converges in law towards $\tilde{p}$).

Example: Doudou the hamster

Let us assume that Doudou's state (a hamster) every minute follows a Markov chain with three possible states: sleep (S), eat (E), or work out (W). Thus, its state in one minute depends only on its current state, and not what it was doing before. Suppose that the transition probability matrix is then the following:

$$P = \begin{pmatrix} X_i/X_{i+1} & S & E & W \\ S & 0.9 & 0.05 & 0.05 \\ E & 0.7 & 0 & 0.3 \\ W & 0.8 & 0 & 0.2 \end{pmatrix}$$

1) Is the Markov chain irreducible ? recurrent ? aperiodic ?
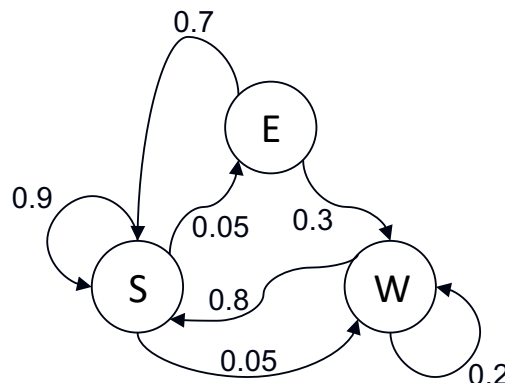
2) Suppose that Doudou is now asleep. What will it be doing in 2 minutes ? in 10 minutes ?

$$x_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}^T \qquad x_2 = x_0 P^2 = \begin{pmatrix} 0.885 \\ 0.045 \\ 0.070 \end{pmatrix}^T \qquad x_{10} = x_2 P^8 = x_0 P^{10} = \begin{pmatrix} 0.884 \\ 0.044 \\ 0.072 \end{pmatrix}^T$$

3) Suppose now that Doudou is working out. What is he going to be doing in 10 minutes ?

$$x_0 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}^T \qquad x_{10} = x_0 P^{10} = \begin{pmatrix} 0.884 \\ 0.044 \\ 0.072 \end{pmatrix}^T$$

Here, since the chain is aperiodic, recurrent and irreducible, therefore there is a stationary distribution: $\tilde{p} = \tilde{p}P$.

3.3.2 MCMC sampling

MCMC algorithms: general principle

The general principle of MCMC algorithms is as follows: to produce an acceptable approximation of an integral – or other functional – of a distribution of interest (such as the posterior), one only needs to sample a Markov chain whose limit distribution is this very distribution of interest (i.e. the posterior), and then to apply the Monte Carlo method to it.

A **twofold convergence** is thus required:

1 the Markov chain must first converge to its stationary distribution: $\forall X_0, X_n \xrightarrow[n \to +\infty]{\mathcal{L}} \tilde{p}$

2 once this stationary distribution is reached, the Monte Carlo convergence must also happen:
$$\frac{1}{N} \sum_{i=1}^{N} f(X_{n+i}) \xrightarrow[N \to +\infty]{} \mathbb{E}[f(X)]$$

$$\overbrace{X_0 \to X_1 \to X_2 \to \cdots \to X_n}^{\text{Markov chain convergence}} \to \overbrace{X_{n+1} \to X_{n+2} \to \cdots \to X_{n+N}}^{\text{Monte Carlo sample}}$$

MCMC algorithms uses an acceptance-rejection framework:

---

1 Initialise $x^{(0)}$

2 For $t = 1 \ldots n + N$:
    1. Propose a new candidate $y^{(t)} \sim q(y^{(t)}|x^{(t-1)})$
    2. Accept $y^{(t)}$ with probability $\alpha(x^{(t-1)}, y^{(t)})$:
        $x^{(t)} := y^{(t)}$
        if $t > n$, "save" $x^{(t)}$ (as part of the final Monte Carlo sample)

where q is the instrumental distribution for proposing new samples and α is the acceptance probability.

General organisation of MCMC algorithms

For the instrument proposal distribution q, there is not universal optimal choice, but an infinity of possible distributions (some better than others). In order to ensure convergence towards the targeted distribution $\tilde{p}$: (i) the support of q must cover all of the support values of $\tilde{p}$, (ii) q must not generate periodic values. Ideally, q is chosen so that its calculation is simple and fast (for example, you can choose q to be symmetric).

Metropolis-Hastings algorithm

The Metropolis-Hastings is the workhorse of MCMC algorithms: it is a very simple and general algorithm for sampling according to uni- or multi-dimensional distributions.

---

1 Initialise $x^{(0)}$

2 For $t = 1 \ldots n + N$:
  1. Sample $y^{(t)} \sim q(y^{(t)}|x^{(t-1)})$
  2. Compute the acceptance probability
  $$\alpha^{(t)} = \min\left\{1, \frac{\tilde{p}(y)}{q(y^{(t)}|x^{(t-1)})}\bigg/ \frac{\tilde{p}(x^{(t-1)})}{q(x^{(t-1)}|y^{(t)})}\right\}$$
  3. Acceptance-rejection step:
     Sample a value $u^{(t)} \sim \mathcal{U}_{[0,1]}$
     $$x^{(t)} = \begin{cases} y^{(t)} \text{ if } u^{(t)} \leq \alpha^{(t)} \\ x^{(t-1)} \text{ else} \end{cases}$$

One can reformulate the acceptance probability $\alpha^{(t)}$ as: $\alpha^{(t)} = \min\left\{1, \frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})}\frac{q(x^{(t-1)}|y^{(t)})^{(t)}}{q(y^{(t)}|x^{(t-1)})}\right\}$. It can thus be computed knowing $\tilde{p}$ only up to a constant (since it simplifies in the above ratio). Note that this is particularly useful when the target $\tilde{p}$ is actually the posterior distribution of some Bayesian model.

There are particular cases where the computation of $\alpha^{(t)}$ can be simplified, such as:

- **Independant Metropolis-Hastings**: $q(y^{(t)}|x^{(t-1)}) = q(y^{(t)})$
- **Random walk Metropolis-Hastings**: $q(y^{(t)}|x^{(t-1)}) = g(y^{(t)} - x^{(t-1)})$
  If g is symetric ($g(-x) = g(x)$), the computation of the acceptance probability $\alpha^{(t)}$ then simplifies: $\frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})}\frac{q(y^{(t)}|x^{(t-1)})}{q(x^{(t-1)}|y^{(t)})} = \frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})}\frac{g(y^{(t)} - x^{(t-1)})}{g(x^{(t-1)} - y^{(t)})} = \frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})}$

The Metropolis-Hastings algorithm is a very simple and general algorithm for one-dimensional or multi-dimensional sampling. The choice of the instrumental distribution is crucial – but difficult – and has a considerable impact on the algorithm's performance (e.g. a high rejection rate often implies very long computation times). Moreover, it is an algorithm that becomes ineffective when the dimension of the problem becomes too large. The simulated annealing algorithm and the Gibbs sampler are algorithms that partially overcome some of these limits.

Simulated annealing

In order to go beyond some of the limitations of the Metropolis-Hastings algorithm, the acceptance probability computation can be changed during the algorithm progression. The idea is to first have a high acceptance probability $\alpha^{(t)}$, in order to explore the whole state space (i.e.

"travel far"), and then to decrease it when the algorithm converges so that the new accepted values are concentrated around the optimal mode. This is done by introducing a "temperature" into the Metropolis-Hastings algorithm, which varies at each iteration and is noted $T(t)$:

---

1 Initialise $x^{(0)}$

2 For $t = 1 \ldots n + N$:

  a. Sample $y^{(t)} \sim q(y^{(t)}|x^{(t-1)})$

  b. Compute the acceptance probability
$$\alpha^{(t)} = \min \left\{ 1, \left(\frac{\tilde{p}(y^{(t)})}{\tilde{p}(x^{(t-1)})} \frac{q(x^{(t-1)}|y^{(t)})}{q(y^{(t)}|x^{(t-1)})} \right)^{\frac{1}{T(t)}} \right\}$$

  c. Acceptance-rejection step: sample a value $u^{(t)} \sim \mathcal{U}_{[0,1]}$
$$x^{(t)} := \begin{cases} y^{(t)} \text{ if } u^{(t)} \leq \alpha^{(t)} \\ x^{(t-1)} \text{ else} \end{cases}$$

For example, one can use $T(t) = T_0 \left(\frac{T_f}{T_0} \right)^{\frac{t}{n}}$ with T_0 the initial temperature, n the number of iterations above which one thinks to reach convergence, and T_f the temperature after n iterations. This algorithm is particularly useful when local optimums are present. An example of possible temperature values could be $T_0 = 10$ and $T_f = 0.1$ (but appropriate values depend on the actual scale of the problem).

Gibbs sampler

When the dimension (of x) increases, it becomes very difficult to propose probable values in algorithms using the acceptance-rejection strategy. The idea behind the Gibbs sampler is to generate x coordinate by coordinate, while conditioning on the last values obtained. Therefore, x must admit a decomposition such that $x = (x_1, \ldots, x_d)$, and the distributions $p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$ must be known and easy to sample from. Unlike the Metropolis-Hastings algorithm, the Gibbs sampler does not really rely on an acceptance-rejection strategy but accepts all sampled proposals ($\alpha = 1$). The proposal distributions are imposed here: they are the conditional probability distributions of each coordinate. The Gibbs sampler is thus a coordinate-wise update algorithm:

---

1 Initialise $x^{(0)} = (x_1^{(0)}, \ldots, x_d^{(0)})$

2 For $t = 1 \ldots n$:

  1. sample $x_1^{(t)} \sim p(x_1|x_2^{(t-1)}, \ldots, x_d^{(t-1)})$

  2. sample $x_2^{(t)} \sim p(x_2|x_1^{(t)}, x_3^{(t-1)}, \ldots, x_d^{(t-1)})$

  3. ...

  4. sample $x_i^{(t)} \sim p(x_i|x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_{i+1}^{(t-1)}, \ldots, x_d^{(t-1)})$

  5. ...

  6. sample $x_d^{(t)} \sim p(x_d|x_1^{(t)}, \ldots, x_{d-1}^{(t)})$

Remark: if conditional distributions are unknown for some coordinates, they can be sampled by introducing an acceptance-rejection step for this coordinate only. Such algorithms are known as Metropolis-within-Gibbs.

3.4 Use of MCMC algorithms for Bayesian inference in practice

The implementation of Metropolis-Hastings, Gibbs, or Metrolpolis within Gibbs algorithms thus enable one to sample according from the posterior distribution of a Bayesian model. In particular, the prior can be used as the porposal while the target distribution is the posterior: x is then replace by θ, and $\tilde{p}$ by $p(\theta|\boldsymbol{y})$, respectively. A number of software programs such as *JAGS*(`http://mcmc-jags.sourceforge.net/`), STAN (`http://mc-stan.org/`) or *WinBUGS* (`https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/`) offer an implementation of such algorithms for a great number of modeling strategies and can be applied without having to (re)program the MCMC algorithm itself.

The BUGS project (*Bayesian inference Using Gibbs Sampling*: `https://www.mrc-bsu.cam.ac.uk/software/bugs/`) was initiated in 1989 by the BioStatistics Unit of the MRC (*Medical Research Council*) at the University of Cambridge (UK) to provide flexible software for the Bayesian analysis of complex statistical models using MCMC algorithms. Its most famous implementation is WinBUGS, a point-&- click software available under the *Windows* operating system. OpenBUGS is an implementation running under *Windows*, *Mac OS* or *Linux*. JAGS (*Just another Gibbs Sampler*) is another, more recent, implementation that also relies on the BUGS language. A very useful resource is the JAGS user manual (`http://sourceforge.net/projects/mcmc-jags/files/Manuals/3.x/jags_user_manual.pdf`). Finally, we should also note the software STAN, recently developed at Columbia University, which is similar to BUGS only in its interface, relying on innovative MCMC algorithms, such as Hamiltonian Monte Carlo or Variational Bayes.

3.4.1 MCMC convergence

Sampling according to the *posterior* with an MCMC algorithm features 2 steps:

The **burn-in** phase: The **phase de chauffe** (*burn-in*) : this warm-up phase corresponds to the first iterations of the MCMC algorithm, which should not be retained in the Monte Carlo sample analysis. Indeed, these do not come from distribution. This phase therefore corresponds to the time needed for the Markov chain to converge towards its stationary law. Its length varies from model to model. There are no consequences for taking too long a heating phase, apart from its computational burden.

The **sampling phase**: it must be long enough to allow a good estimate of the *posterior* distribution, especially for low probability ranges.

The mathematical properties of Markov chains guarantee the convergence of MCMC algorithms, but do not give an indication of the number of iterations required to achieve this convergence. While there is no way to guarantee this convergence in finite time, there are a number of tools available to diagnose the non-convergence of a Markov chain towards its stationary law. They must therefore be used when interpreting the outputs of an MCMC algorithm to avoid situations where the chain has not converged.

One way to monitor the convergence of an MCMC sampling algorithm is to generate several strings (in parallel and independently) with different initial values. If the algorithm works, then these different (Markov) strings must converge to the same stationary distribution (the *posterior* distribution). After enough iterations, it should be impossible to distinguish between these different channels. For each string, the n first values are considered to belong to the **burn-in** phase of the algorithm, necessary for the Markov string to first converge to its stationary law from the initial values. They are therefore not retained, and we are interested in the following N observations that will constitute our Monte Carlo samples.

Monte Carlo error

The Monte Carlo error characterizes the uncertainty introduced by MCMC sampling. For a given parameter, it quantifies the variability expected in its estimation if we would generate several (Markov) **chains**, i.e. several *posterior* Monte Carlo samples (thanks to an MCMC algorithm, with different initializations and each time the same number N of iterations). The Monte Carlo standard-errors give an idea of this variability. If the standard errors have very different values from one chain to another, then the sampler must be run for longer. The exact length of sampling required to obtain a given standard error will depend on the efficiency and mixing of the sampler. It is important that this Monte Carlo error be small with respect to the estimated variance of the *posterior* distribution.

Gelman-Rubin statistic

One way to evaluate the convergence of an MCMC sampler is to compare the variation between different chains to the variation within the same chain after a number of iterations. If the algorithm has converged, the between-chains variation should be close to zero.

Let $\theta_{[c]} = (\theta_{[c]}^{(1)}, \ldots, \theta_{[c]}^{(N)})$ the N-sample obtained from chain $c = 1, \ldots, C$ of an MCMC algorithm sampling θ. The **Gelman-Rubin statistic** is then:

$$R = \frac{\frac{N-1}{N}W \frac{1}{N}B}{W}$$

with $B = \frac{N}{C-1}\sum_{c=1}^{C}(\bar{\theta}_{[C]} - \bar{\theta}.)^2$ the between-chains variance, $\bar{\theta}_{[c]} = \frac{1}{N}\sum_{t=1}^{N}\theta_{[c]}^{(t)}$, $\bar{\theta}. = \frac{1}{C}\sum_{c=1}^{C}\bar{\theta}_{[C]}$, and $W = \frac{1}{C}\sum_{c=1}^{C}s_{[c]}^2$ the within-chain variance, $s_{[c]}^2 = \frac{1}{N-1}\sum_{t=1}^{N}(\theta_{[c]}^{(t)} - \bar{\theta}_{[C]})^2$. When $N \to +\infty$ while $B \to 0$, R gets close to 1. One will thus want to run an MCMC algorithm for a sufficient number of iterations in order to reach a value of R close enough to 1, for instance between 1 and 1.01 (or 1.05).

The Gelman-Rubin statistic is a ratio (therefore without unit) which makes it a summary that can be interpreted simply and in the same way for any MCMC sampler and any Bayesian model. Another advantage is that it does not require any tuning parameter (unlike Monte Carlo errors). The Gelman-Rubin statistic is therefore a good way to diagnose the convergence of an MCMC algorithm. Nevertheless, its calculation may be unstable and it cannot guarantee convergence on its own. It is a general tool, for the general monitoring of a Markov chain.

Note that other statistics (e.g. Geweke's statistic) are sometimes used instead of, or in addition to, Gelman-Rubin's – which remains the most popular.

Graphical diagnostics

In addition to the Gelman-Rubin statistic, a number of graphical diagnostics can be used to evaluate the non-convergence of an MCMC algorithm:

- the **trace**: refers to the representation of the successive values of the string. When more than one independent chain is generated from different initializations, the traces of the different chains must stabilize and overlap once convergence is achieved.

- **non-parametric density estimators**: according to Bernstein-von Mises' convergence theorem, the *posterior* distribution must be unimodal. For this we can use a non-parametric (kernel) density estimator on the generated Monte Carlo sample to check that the *posterior* distribution is indeed unimodal and sufficiently smooth.

- **runing quantiles**: similarly to the trace, the quantiles of the different chains must stabilize and overlap during the different iterations once convergence is achieved.

- **Gelman-Rubin diagram**: it represents the cumulative Gelman-Rubin statistic over the iterations. Its level must quickly become very close to 1 (ideally < 1.01 or at least < 1.05).
- **autocorrelogram**: when the Markov chain doesn't "mix" very well, it can happen that successive observations are highly correlated from one iteration to the next. This is not a problem in itself, but it greatly reduces the effective sample size for the *posterior* estimations. A common solution is to keep only one iteration out of 2, 5 or 10 (the more correlated the retained samples, the more they spaced-out) using the `thin` parameter (adjusting the spacing between the iterations kept in the MCMC sample).
- **cross-correlation**: One can also look at the correlation between our differences through the different samples. Note that it is common to observe a strong correlation between certain parameters and that this is not necessarily indicative of a problem with the MCMC algorithm (the frequentist approach also estimates correlations, sometimes significant, between the parameters of a model using the Fisher information matrix).

Remark: it is common for diagnoses to be OK for some parameters, but not for others. This is a subjective assessment, and the aim is that the majority of the criteria are met (or more or less met), for a large majority of the parameters.

Effective sample size

In practice, a sample generated from an MCMC algorithm is *iid* only in very special cases. Indeed, the Markov "memoryless" property generally leads to a correlation between the values generated one after the other (dependent sampling). For a fixed sample size of N, this auto-correlation decreases the amount of information available, and slows down the convergence of the law of large numbers in the Monte Carlo method – compared to a purely independent sample. An indicator to quantify this information is the **effective sample size** which is calculated as follows:
$$ ESS = \frac{N}{1 + 2 \sum_{k=1}^{+\infty} \rho(k)} $$

where $\rho(k)$ stands for auto-correlation with *lag* of rank k.

One solution used in practice to reduce these auto-correlation problems is not to retain all the values successively sampled by an MCMC algorithm, but to space out the iterations retained. For example, only the values sampled every 2, 5, or 10 iterations can be retained, which will decrease the dependency within the generated Monte Carlo sample.

3.4.2 Inference from MCMC sampling

Estimation

Using MCMC algorithms, we're able to obtain a Monte Carlo sample of the *posterior* distribution for a given Bayesian model. Thus one can use the Monte Carlo method to obtain different ***posterior* estimates**: point estimates (*posterior* mean, *posterior* median, ...), credibility intervals (in particular thanks to the ® package `HDInterval` which makes it possible to calculate the narrowest credibility interval for a given level, i.e. the *Highest Density Interval – HDI*)), cross-correlations between parameters, etc.
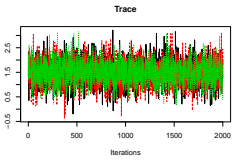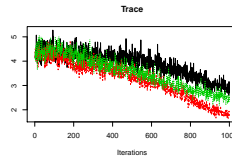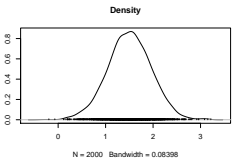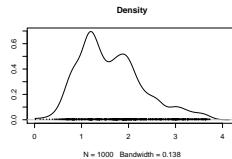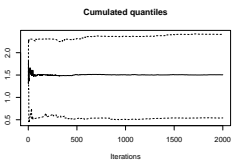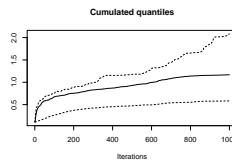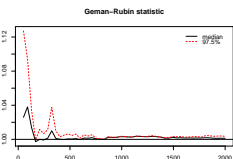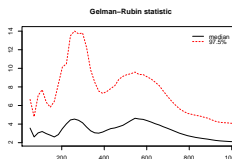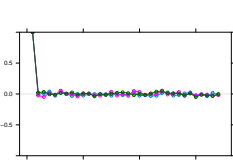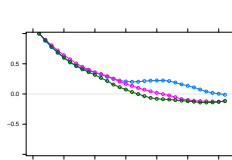
Graph name	😁	☹️	R	Potential solutions
Trace			`coda::traceplot()`	↗ `n.iter` and/or ↗ *burn-in*
Density			`coda::densplot()`	↗ `n.iter` and/or ↗ *burn-in*
Running quantiles			`coda::cumuplot()`	↗ `n.iter` and/or ↗ *burn-in*
Gelman-Rubin			`coda::gelman.plot()`	↗ `n.iter` and/or ↗ *burn-in*
Auto-correlation			`coda::acfplot()`	↗ `thin` and/or ↗ `n.iter` and/or ↗ *burn-in*

Table 3.1 – Reference examples for graphical diagnostics of convergence

Deviance Information Criterion (DIC)

The ***Deviance Information Criterion*** (DIC) relies on the deviance [1], which is defined as: $D(\theta) = -2\log(p(\theta|\boldsymbol{y})) + C$ where C is a constant. DIC is then defined as:

$$DIC = \overline{D(\theta)} + p_D$$

where $p_D = \left(D(\bar{\theta}) - \overline{D(\theta)}\right)$ represents a penalty for the actual number of parameters. In particular, the DIC makes it possible to compare different models on the same data (the lower the DIC, the better the model), and to make modeling choices in the Bayesian context.

3.5 Other methods

3.5.1 Variational Bayes

Variational Bayes inference is an approximation technique of the full Bayesian approach that focuses on the estimation of *posterior* means, and the uncertainty around them. It is based on a parametric approximation of the *posterior* distribution that minimizes Kullback-Leibler's divergence from the true *posterior* distribution. The computation of the variational Bayes solution thus amounts to a classical optimization problem, whose numerical computation is generally very fast ; this can make it an appealing solution in big data problems. Nevertheless, the quality of the variational approximation will depend on the adequacy of the chosen parametric model, for which there are no guarantees. In addition, this approach usually requires a relatively extensive analytical study of the *posterior* distribution.

3.5.2 Approximate Bayes Computation (ABC)

Approximate Bayes Computation(ABC) is another alternative to MCMC methods, which uses the sampling model to avoid having to calculate the likelihood in the numerator of the Bayes formula, by instead sampling observations according to the generative model of the data. One then obtains a sample of the *posterior* distribution by keeping the θ parameter values, generated from the *prior* distribution, having generated the samples sufficiently close to the actually observed data. The difficulty of this approach lies in the formalization of "close enough", which induces an approximation compared to the exact Bayesian approach which would retain all the values of θ (but whose computation cost by this method is then often very high).

---

1. M Plummer, Penalized loss functions for Bayesian model comparison, *Biostatistics*, 2008

Chapter 4

Bayesian analyses in biomedical applications: some real-world use cases

In this chapter, we cover the basics of three different real-world use cases, that illustrate settings where the Bayesian approach can be particularly useful in biomedical science.

4.1 *Post-mortem* analysis of an under-powered randomized trial: a case-study

The randomized clinical trial *EOLIA* (Combes *et al.*, *NEJM*, 2018) evaluated a new treatment for severe acute respiratory distress syndrome (ARDS) by comparing the mortality rate after 60 days among 249 patients randomized between a control group (receiving conventional treatment, i.e. mechanical ventilation) and a treatment group receiving ExtraCorporeal Membrane Oxygenation (ECMO) – the new treatment studied. A frequentist analysis of the data concluded to a Relative Risk of death of 0.76 in the ECMO compared to control (in Intent to Treat), with $CI_{95\%} = [0.55, 1.04]$ and the associated p-value of 0.09.

	Group	
	ECMO	Control
group size n	124	125
number of deaths at 60 days	44	57

Table 4.1 – Observed data in the EOLIA trial

Goligher *et al.* (*JAMA*, 2018) performed a Bayesian re-analysis of these data, further exploring the evidence and how it can be quantified and summarized with a Bayesian approach.

4.2 Bayesian meta-analysis

4.2.1 Introduction to meta-analysis

A meta-analysis is an analysis of analyses, producing a single quantitative summary of studies answering the same research question. This can be particularly appealing, especially in biomedical applications where medical therapies effects are often evaluated in multiple different studies. Ideally, one would pool individual observations from multiple studies (while accounting for potential differences in the pooled experiments), but most of the times only aggregated summary statistics estimates (often denoted as effect sizes) are available, alongside some sort of uncertainty measures (generally the corresponding standard errors).

Study Heterogeneity

One of the main difficulties for performing a meta-analysis is the variations of the observed effects. Those can either come from within-study uncertainty or real heterogeneity in effect size between the different studies. It is often the case that the different studies were conducted on various populations, and therefore there is potential extra-variability between them. In addition, studies are often of various sample sizes, a parameter which will also impact the estimate and its variability.

Meta-analysis random-effects model

The random-effects model is the one of the most common approach to meta analysis. The model can be written as follows:

$$y_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$$
$$\theta_i \sim \mathcal{N}(\mu, \tau^2)$$

It can be seen as a hierarchical generalization of the fixed effect model $y_i \sim \mathcal{N}(\mu, \sigma_i^2)$ which assume the exact same mean for each study, while the random-effects models allow for between study variability through the parameter τ: $y_i \sim \mathcal{N}(\mu, \sigma_i^2 + \tau^2)$.

4.2.2 Bayesian meta-analysis in practice

Meta-analysis: a perfect usecase for Bayesian analysis ?

A Bayesian approach to inference is very attractive in this context, especially when a meta-analysis is based only on few studies. Indeed, the Bayesian approach allows for integrating previous knowledge in the form of informative *priors*, and can mitigate some of the computationnal shortcomings encountered when dealing with few observations.

bayesmeta R package

The recent 🅡 package `bayesmeta` has been implemented to perform such Bayesian meta-analysis. It has a companion Shiny app availbale at: `http://ams.med.uni-goettingen.de:3838/bayesmeta/app/`

4.2.3 Example dataset: Crins *et al.*, 2014

In 2014, Crins *et al.* published a meta-analysis of controlled studies on interleukin-2 receptor antagonists for pediatric liver transplant recipients. They estimated random-effects meta-analysis models to assess (among other things) the incidence of i) acute rejection, ii) steroid-resistant rejection, iii) post-transplant lymphoproliferative disease, and iv) patient death, with or without IL-2RA.

4.2.4 Going further

Scientific literature search

An very important part preceding any meta-analysis is to perform a rigorous and exhaustive search of the scientific literature. This is no easy task and methologies and tools have been developed to that end. An important aspect is that quite often estimates along with their standard errors are not given right away, so one has to transform these oneself.

Evidence synthesis

Meta-analysis is a method part of the broader field of evidence synthesis that aims at synthesizing the scientific evidence on a particular subject. For instance, meta-regression is an extension of meta-analysis to take into account the effect of covariates. Other approaches, such as mechanistic modeling can also be considered for performing evidence synthesis.

Meta-analysis and more generally evidence synthesis methods are still active research domains, and should be used with care and thoughfulness. For example, one of the debated property of the random-effects model is that it will effectively reduce the weight of studies with larger sample sizes (and smaller standard errors around their estimates) compared to the fixed effect model (Serghiou & Goodman, *JAMA*, 2018), which can be argued as either being a bug or a feature depending on the context (depending on how trustworthy are the studies).

4.3   Adaptative phase I/II trials: CRM and Bayesian analysis

4.3.1   Introduction to Continuous Reassessment Methods (CRM) for dose finding

Continuous Reassessment Methods (CRM) can be used in Phase I dose-escalation trials, where the objective is to identify the optimal dose (i.e. with the greater efficacy while maintaining an acceptable toxicity: the higher the dose the greater the efficacy but at the same time the greater the toxicity). The idea is then to select iteratively the dose that the next recruited patient will be given, based past accumulating observations from patients previously included. This trial design was first introduced by O'Quigley *at al.* in 1990, and is being increasingly used (although still in a minority of such trials).

A strong argument in favor of CRM is that they allow to treat each patient ethically by always giving the dose that is best supported by the current evidence, while searching for the optimal dose. The Bayesian approach is particularly well suited for such studies thanks to its ability to easily formalize prior knowledge and its chain rule (i.e. sequential Bayes approach).

4.3.2   Critical reading of Kaguelidou *et al.*, *PLOS ONE*, 2016

Kaguelidou *et al.* conducted a dose-finding study of Omeprazole on gastric pH in neonates with Gastro-Esophageal Acid Reflux (GEAR) in order to determine the minimum effective dose. They used a CRM design to select the drug dose as close as possible to the predefined target level of efficacy (with a credibility interval of 95%).