Frequentist and Bayesian approaches to uncertainty (lecture material)

Chapter · August 2020

CITATIONS
0

READS
25

1 author:

Marc Fischer

22 PUBLICATIONS **79** CITATIONS

SEE PROFILE

Saint-Étienne School of Mines

Frequentist and Bayesian approaches to uncertainty

Marc Fischer (PhD)

Contents

Chapter 1

Introduction

Probabilities permeate every aspect of our modern society and daily life. Ideally, we would prefer to be absolutely sure that certain propositions are true or that some events will happen. Alas, absolute certainty only exists in the realm of pure mathematics. A logic limited to binary variables is utterly unable to describe and to deal with the phenomena of the real world. Instead, the following type of questions promptly comes to mind:

- What is the probability that 1E+06 photon emissions will take place in the next 50 seconds?
- How probable is it that John is the murderer?
- How likely is it that my husband is cheating on me?
- How plausible is it that we live in a computer simulation run by some freak aliens?

In each case, there is really no way to reach absolute certainty. However, it would be equally mistaken to think that "anything goes", that all options have the same degree of plausibility. Whilst stochastic reasoning appeared probably very early after the emergence of humankind on our planet, it was a specific correspondence between the French mathematicians Blaise Pascal and Pierre de Fermat about games of chance which opened up the door to the development of modern probability theory back in 1654. Surprisingly enough, no consensus about the nature of probability has ever been reached since that time. Broadly speaking, statistics and data science can currently be divided into two main camps. On the one hand, **frequentists** define probabilities as relative frequencies that would be observed if the conditions of an experiment could be repeated an endless number of times. On the other hand, Bayesians see probabilities as *degrees of belief* a perfectly rational agent should have with respect to some facts. One fundamental consequence of these diverging views is that while frequentists will consider a physical

parameter (such as a heat diffusion coefficient) as an *unknown constant*, Bayesians will see it as a *random variable* in its own right with its own probability distribution. This results in major differences in the way real-world problems are treated by proponents of these two approaches. The purpose of this first lecture is to provide the reader with an introduction to the underlying philosophical as well as practical aspects of frequentism and Bayesianism. In Chapter 2, the basics of frequentist reasoning are briefly recapped. In Chapter 3, the Bayesian approach is presented. In Chapter 4, we'll be dealing with arguments for and against Bayesianism.

Chapter 2

Frequentism in a nutshell

Since frequentist methods were presented and explained in detail during your first year lecture on statistics and data science, we shall only briefly summarise the main aspects of this approach here. All along, we will be considering an experiment which gives rise to a random variable X. We also assume that X is the result of a process characterised by a set of unknown physical parameters $\theta \in \Theta$. We consider that the conditions of the experiment are repeated n times, thereby giving birth to the random variable vector $(X_1, X_2, ..., X_n)$. The frequentist probability that $X = x$ is the limit of the frequency of that event when an infinite number of trials is considered:

$$p(X = x) = \lim_{n \to \infty} f_n(X_i = x) \tag{2.1}$$

This definition relies on the empirical observation that the frequency of an event (such as the tossing of a fair coin in Fig. 2.2) tends to converge towards a stable value, which is identified as the frequentist probability of that event. Based on this notion, the frequentist approach boils down to the following question: *If hypothesis H is true, what is the probability that we'd get the data D we actually see?* Intuitively, we would like the likelihood $p(D \mid H)$ to be high enough, i.e. above a given threshold. Finding such a systematic threshold does not prove possible, however, as that quantity can take on very small values **even if** H is true. Let us consider a concrete example to illustrate this. We toss a coin 100,000 times. Our goal is to verify whether or not it's perfectly symmetrical (H: $\theta = 0.5$). It lands 50,005 times on heads and 49,995 times on tails. We know that $p(D|H) = \binom{n}{k}\theta^{50005}(1-\theta)^{49995}$, which leads to $p(D|H) = 0.002522$. The value is very small because there is a very large number of equally probable outcomes. Consequently, we modify the question and ask instead: *If hypothesis H is true, what is the probability that we'd get a result* ***which is at least as extreme*** *as the one we actually*

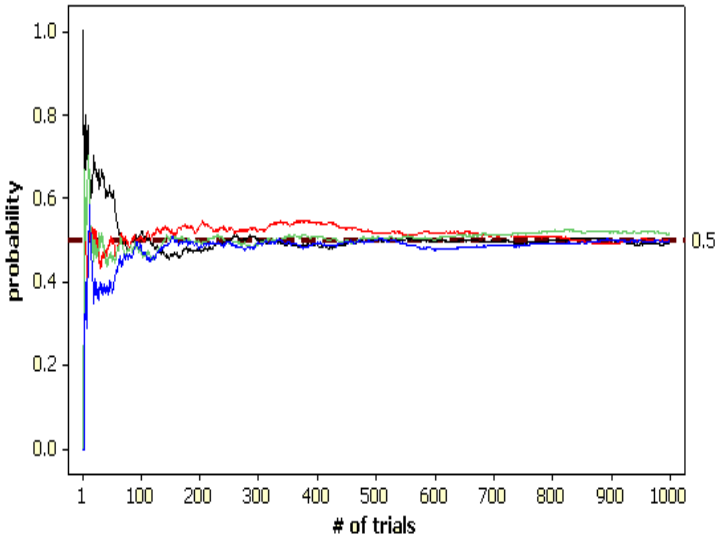


Figure 2.1: Convergence towards the probability

see? This leads to the definition of the p-value [1].

$$p - value = p(N_{heads} - 50000 \geq 5 \ \mid \theta = 0.5) \tag{2.2}$$

We find a p-value equal to 0.4861, which means that we could very plausibly get results that are at least as extreme as the one we saw if the coin were perfectly symmetrical. We cannot, however, draw the **positive conclusion** that the coin is symmetrical from the p-value alone because it doesn't consider the other competing hypotheses. Likewise, a small p-value doesn't necessarily mean that the hypothesis is probably false.

To compare different hypotheses with one another, a widespread technique used by frequentists is the Maximum Likelihood Estimation (MLE). In the case we're studying, this would mean finding $\hat{\theta} = \arg\max_{\theta \in \Theta} p(D \mid \theta) = \arg\max_{\theta \in \Theta} \binom{n}{k} \theta^{50005}(1-\theta)^{49995}$, which leads to the empirical mean: $\hat{\theta} = 50005/100000$. If we are comparing a finite number of models $M_1, M_2, ...M_n$, a straightforward application of the MLE would simply lead us to the model which is the most compatible with the data at hand, that is to say the model associated with the highest $p(D \mid M)$. Most of the time, it is not enough to determine a single optimal value for a parameter k of a model because of the uncertainties of the data D at our disposal. Beyond that, we also need to compute a *confidence interval*

[1] since $n_{heads} > 50000$, we choose to perform a one-sided test contrasting H_0: $\theta = 0.5$ and H_1: $\theta > 0.5$ while ignoring H_2: $\theta < 0.5$

[u(D),v(D)] such that $p(u(D) \leq k \leq v(D)) = \gamma \in [0; 1]$. It is a common but mistaken interpretation of confidence intervals to say that the probability that k belongs to the interval $[u(D), v(D)]$ is equal to γ. For a frequentist, the only random variables are the lower estimator $u(D)$ and the upper estimator $v(D)$. k is not a random variable but an unknown constant and as such its value cannot have any probability. Instead, a confidence interval should be seen as a *random interval* which contains the true parameter value k γ-100% of the time, as illustrated in Fig. 2.2.

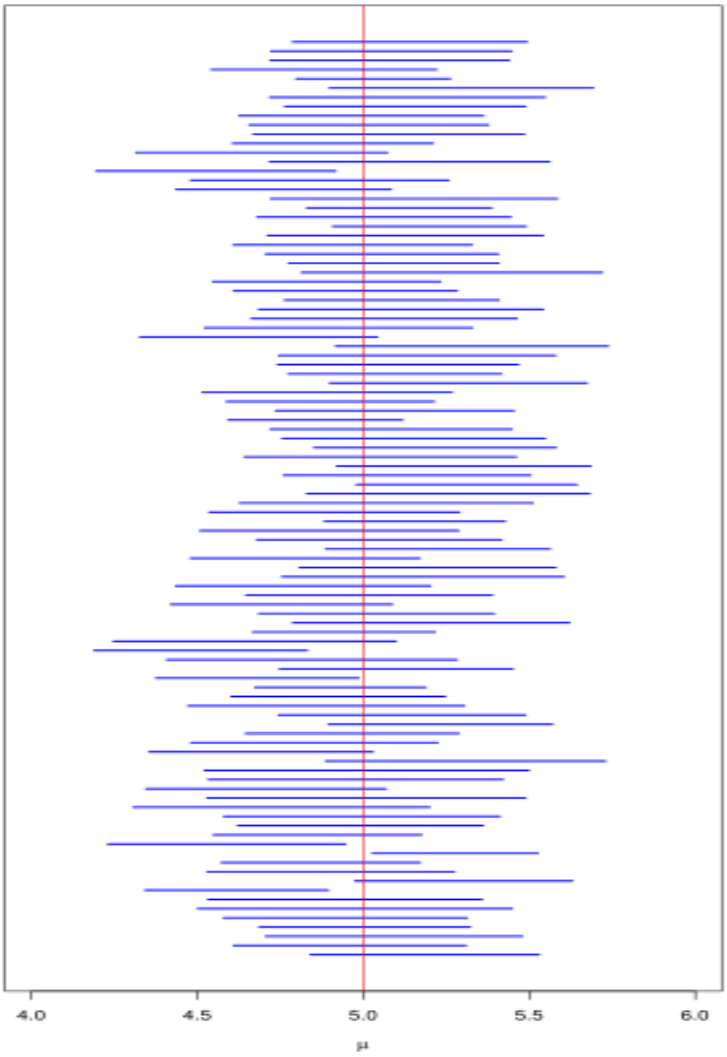


Figure 2.2: Confidence intervals (frequentism)

Chapter 3

Bayesianism in a nutshell

3.1 Discrete parameter space

We shall first consider the case where both the outcome and the parameter space are discrete. We just saw that frequentists are interested in the probability of obtaining the results we see given that some hypothesis is true. Bayesians go beyond that step and want to determine the probability that the hypothesis $\mathrm{H_i}$ is true given the data, that is to say $p(H_i|D)$. That quantity is obtained by updating a prior (initial) belief with the data we now have at our disposal. This is achieved by using Bayes' theorem, which is given by Eq. 3.1 for a finite discrete parameter space (which comes down to considering n hypotheses $\mathrm{H_1}, ..., \mathrm{H_n}$).

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{p(D)} = \frac{p(D|H_i)p(H_i)}{\sum_{j=1}^{n} p(D|H_j)p(H_j)} \tag{3.1}$$

$\mathrm{p(H_i)}$ is the prior probability, that is to say our degree of belief in the hypothesis H_i **before** considering the data D. As we saw in Chapter 2, $p(D|H_i)$ is the likelihood, i.e. the probability of observing the data D if $\mathrm{H_i}$ were true. $p(H_i|D)$ is the posterior probability, that is the probability of $\mathrm{H_i}$ given the data D. To illustrate all of this, let us consider the following situation which belongs to the daily life of many physicians. A man suffering from very general symptoms goes to his doctor to be checked for a specific disease. We know that in the general population, only 0.1 % of all men have this sickness. The physician decides that the patient must be tested medically. We know that if a man is really sick, the test will be positive in 95 % of all cases, but that if he isn't sick, the test will be negative in 98 % of all cases. We want to know the probability that the patient suffers from the disease given the fact that the test is positive. Would you **intuitively** say it is high or low?

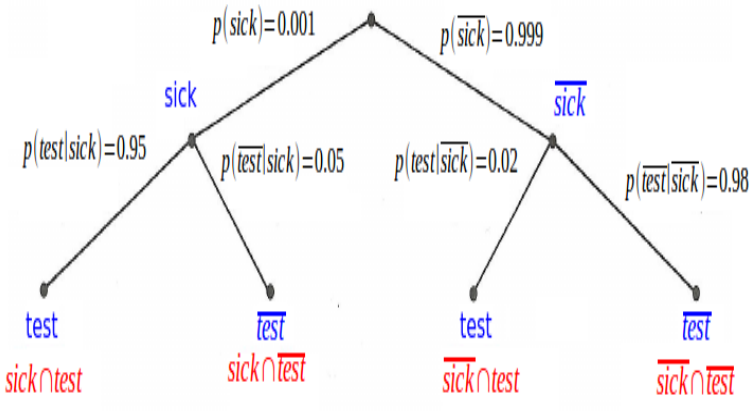


Figure 3.1: Tree diagram

To compute that quantity rigorously, we must use Bayes' theorem as given by Eq. 3.2.

$$p(Sick|Test) = \frac{p(Test|Sick)p(Sick)}{p(Test|Sick)p(Sick) + p(Test|\overline{Sick})p(\overline{Sick})} \tag{3.2}$$

Thanks to the data at our disposal, we know that $p(Test|Sick) = 95$ %, $p(Test|\overline{Sick}) = 2$ %, $p(Sick) = 0.1$ %, and $p(\overline{Sick}) = 99.9$ %. We can represent the situation on a tree diagram (see Figure 3.1). Based on all of this, we find that $p(Sick|Test)$= 4.539 %. If this comes as a surprise to you, you're not alone! Most people would intuitively think that $p(Sick|Test)$ is equal or close to the likelihood $p(Test|Sick)$. An important lesson of this exercise is that we should always beware of our human intuitions while trying to reason probabilistically. The cause of this astonishing result is that the malady is so rare (it affects only 0.1 % of the population) that it is more frequent to find a false positive than a truly sick person.

3.2 Continuous parameter space

3.2.1 Prior updating

The Bayesian methodology can also be very well applied to situations where the parameter space $\theta \in \Theta$ is continuous. The outcome variable X might be either discrete or continuous. Under these conditions, Bayes' theorem consists of Formula 3.3.

$$f(\theta|x) = \frac{f(\theta)L(x|\theta)}{f(x)} = \frac{f(\theta)L(x|\theta)}{\int_{\theta' \in \Theta} f(\theta')L(x|\theta')} \tag{3.3}$$

$f(\theta)$ is the prior probability density, $L(x|\theta)$ is the likelihood function (which is a probability if X is discrete and a probability density if X is continuous)

and $f(\theta|x)$ is the posterior probability density. It is worth noting that for certain families of probability distributions (such as the Poisson distributions and the normal distributions), there are nice analytical formula allowing us to update the parameters of the prior distribution with the likelihood in order to give rise to the posterior distribution. To exemplify this, let us consider a situation where the outcome variable $X = (X_1, X_2, ..., X_n)$ follows a normal distribution: $X_i \sim N(\mu, \sigma^2)$, whereby the standard deviation σ is known and the expected value μ must be estimated. In other words,

$$L(x|\mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right).$$

We must now choose a prior probability distribution for the parameter μ. This might be the strangest aspect of the Bayesian methodology to newcomers: unknown constant parameters must be treated like random variables! To get analytical updating formula, we must choose a prior which is **conjugate** to the likelihood. A prior distribution is conjugate to a likelihood function if and only if the posterior distribution arising out of their combination belongs to the same family as the prior. In that case, this entails choosing a prior normal distribution for μ so that

$$f(\mu) = f_{m,s}(\mu) = \frac{1}{s\sqrt{2\pi}}exp(-\frac{1}{2s^2}(\mu - m)^2).$$

m and s are two constant parameters we're able to choose according to our **prior knowledge**. In that case, it is possible to prove with Formula 3.3 that the posterior distribution is given by

$$f(\mu|x) = N(\mu, m', s') = \frac{1}{s'\sqrt{2\pi}}exp(-\frac{1}{2s'^2}(\mu - m')^2)$$

with

$$m' = \frac{m\sigma^2 + ns^2\overline{x}}{\sigma^2 + ns^2} \qquad s' = \frac{s^2\sigma^2}{\sigma^2 + ns^2}.$$

In Figure 3.2, we can see how the prior distribution $f(\mu)$ (our prior beliefs) is updated with the likelihood $L(x|\mu)$ (the evidence) to give rise to the posterior distribution $f(\mu|x)$ (our posterior beliefs). In the most general case where the prior and posterior probability distributions are not conjugate, we generally use techniques such as Markov chain Monte Carlo methods (MCMC) to generate a series of random variables whose probability density converges towards the unknown posterior probability density [1].

3.2.2 Punctual estimators and credible intervals

In Bayesian reasoning, the posterior probability distribution $f(\theta|x)$ contains all information about the parameter estimation problem. However, it is

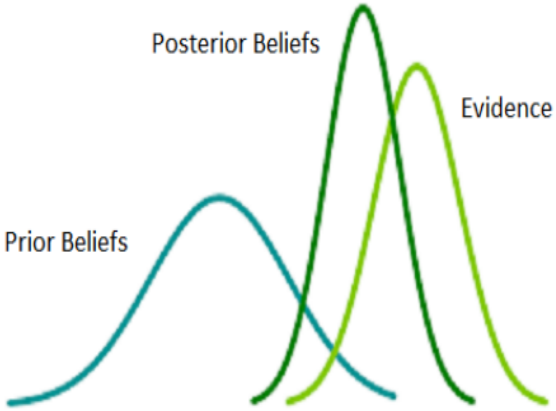


Figure 3.2: Bayesian updating of the prior with the evidence (likelihood)

often not practical to consider it directly, as it can be very hard to visualise it for problems involving ten parameters or more. Instead, there are several Bayesian parameter estimators that allow one to quickly understand the nature of one's results.

Punctual estimators

The **posterior mean** is defined as $\overline{\theta}(x) = \int_{\theta \in \Theta} \theta f(\theta|x) d\theta$ for one parameter and

$$\overline{\theta}(x) = E(T|X = x) = \begin{pmatrix} E(T_1|X = x) \\ ..... \\ E(T_d|X = x) \end{pmatrix}$$

for d parameters. T_i is the random variable θ_i is a realisation of. The **posterior median** is the quantity $\hat{\theta}(x)$ such that

$$p(T \leq \hat{\theta}(x)|X = x) = p(T \geq \hat{\theta}(x)|X = x) = \frac{1}{2}$$

The Bayesian equivalent of the frequentist maximum likelihood estimator is called the **posterior mode**. It is defined as $\theta_{max} = \arg\max_{\theta \in \Theta} f_{T|X=x}(\theta)$.

Credible interval

In Chapter 2, we saw that frequentists use confidence intervals which are *random intervals* that contain the true unknown parameter value k of a model γ-100% of the time. Bayesians regard k as a random variable, and as such they can define a **credible interval** $[t_l, t_u]$ for the most probable values of k according to Eq. 3.4

$$p(k \in [t_l, t_u] = \int_{k=t_l}^{k=t_u} f(k|D) dk = \gamma \tag{3.4}$$

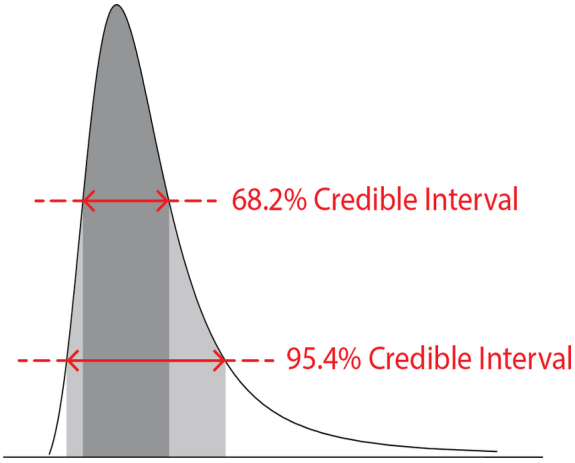


Figure 3.3: Credible interval (Bayesianism)

An example of credible interval for $\gamma = 0.682$ and $\gamma = 0.954$ can be seen in Fig. 3.3.

3.2.3 Bayes' factor

In Chapter 2, we saw that the p-value is one of the central quantities allowing one to compare two models M_1 and M_2 in a frequentist framework. In a Bayesian framework, the Bayes factor B plays a similar role. Through a straightforward application of Bayes' theorem, it can be shown that the ratio of the posterior probabilities of M_1 and M_2 can be written as in Eq. 3.5.

$$\frac{p(M_2|D)}{p(M_1|D)} = \frac{\int_{k_2 \in K_2} L(D|k_2, M_2) f(k_2|M_2) dk_2}{\int_{k_1 \in K_1} L(D|k_1, M_1) f(k_1|M_1) dk_1} \frac{p(M_2)}{p(M_1)} \tag{3.5}$$

$$\frac{p(M_2|D)}{p(M_1|D)} = B \frac{p(M_2)}{p(M_1)} \tag{3.6}$$

The Bayes factor is defined by Eq. 3.7.

$$B = \frac{P(D|M_2)}{P(D|M_1)} = \frac{\int_{k_2 \in K_2} L(D|k_2, M_2) f(k_2|M_2) dk_2}{\int_{k_1 \in K_1} L(D|k_1, M_1) f(k_1|M_1) dk_1} \tag{3.7}$$

$p(M_1)$ and $p(M_2)$ are the prior probabilities of model M_1 and model M_2, respectively. $f(k_1|M_1)$ is the prior probability distribution of k_1 provided that M_1 is the true model. Likewise, $f(k_2|M_2)$ is the prior probability distribution of k_2 provided that M_2 is the true model. $L(D|k_1, M_1)$ is the likelihood of the data D given that model M_1 is true and its parameter is equal to k_1. Similarly, $L(D|k_2, M_2)$ is the likelihood of the data D given that model M_2 is true and its parameter is equal to k_2. The Bayes factor B is commonly interpreted as the strength of the support of data D for model M_2. For instance, $B = 10$ would mean that the data strongly favour M_2 whereas $B =$

0.12 would mean that the data strongly disfavour M_2 and support M_1. This interpretation is only valid, however, if the prior probability distributions $f(k_1|M_1)$ and $f(k_2|M_2)$ are reasonably well known. As we shall see in 4.2, there are many situations where this isn't the case.

3.3 Precise and imprecise Bayesianism

3.3.1 Bayesian probabilities as ideal betting behaviour

Up until now, we've defined a Bayesian probability as the *degree of belief* in an event or a proposition that someone ought to have to be *rational*. But that doesn't tell us what those rational degrees of belief actually are. One very popular way to approach the problem is to consider a series of bets. This is known as the **betting interpretation of probability**. Let's consider an event A (such as your favourite horse winning the race, a thunderstorm darkening the skies tomorrow afternoon and striking your rival dead or your becoming the best student of the whole school). If you're the one buying the bet (for a given $stake \in [0;1]$), you receive 1 € if it comes true and nothing otherwise. If you are the person selling the bet, you give away 1 € if it comes true and nothing otherwise. For the buyer, the profit is $profit_b = X(A) - stake$, whereas for the vendor it is $profit_v = stake - X(A)$. $X(A)$ is a Bernoulli random variable equal to 1 if A occurs and 0 otherwise. A fundamental axiom of decision theory is that a buyer or a vendor would accept a bet if and only if the **expected value** of the profit is greater than or equal to 0. Since X(A) follows a Bernoulli distribution, we have $E(profit_b) = p(A) - stake$ and $E(profit_v) = stake - P(A)$. This implies that the buyer will always accept to buy the bet for $stake = p(A) - \epsilon$ with $\epsilon \geq 0$ and that the vendor will always accept to sell the bet for $stake = p(A) + \epsilon$ with $\epsilon \geq 0$. An individual will always be ready to *both* buy and sell the bet for $stake = p(A)$, which is called a **fair price**. In the general case, the subjective (Bayesian) probability $p(A)$ of an event A for a given person is the value of the quotient $q \in [0;1]$ such that he or she would always be ready to buy the bet for qS and receive S if A occurs. S is an arbitrary real number that can be **both** positive (i.e. the person is a buyer) and negative (i.e. the person is a vendor). A set of probability values is said to be *coherent* if and only if it cannot result in a series of bets that would be a sure loss for the player that has them. An example of an incoherent set of probabilities would be to believe that $p(A) = 0.7$ and $p(\overline{A}) = 0.6$ at the same time. A person holding such beliefs might be willing to bet $0.7 * 100 = 70$ € on A and $0.6 * 10 = 60$ € on $\overline{A}$. The total profit of those two bets is the random variable

$$profit = 100X(A) - 70 + 100X(\overline{A}) - 60.$$

Since we always have $X(A) + X(\overline{A}) = 1$, we also always have

$$profit = 100 - 70 - 60 = -30.$$

This is an example of a *sure loss*, i.e. a loss the betting agent is bound to always suffer from. The incoherent beliefs leading right to it are called *irrational*. Now, it can be mathematically demonstrated that any person whose degrees of belief violate the Kolmogorov axioms can be the victim of a "dutch book", i.e. a series of bets leading to a sure loss [2]. Given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, the Kolmogorov axioms are given by Formula 3.8.

$$\begin{aligned} \forall A \in \sigma, \quad p(A) &\geq 0 \\ P(\Omega) &= 1 \\ p(A \cup B) = p(A) + p(B) \quad &\text{if } A \cap B = \emptyset \end{aligned} \tag{3.8}$$

From a classical Bayesian standpoint, they are a requirement of rationality.

The Dutch Book Argument for Bayesianism isn't universally accepted, though. A good presentation of the philosophical discussion can be found in [2].

3.3.2 Imprecise probability

Principles

Until now, we've been assuming all along that the probability of an event or of a proposition can be described by a *single* number. There are, however, situations where our knowledge isn't good enough to allow us to be so precise. In the face of ignorance, uncertainty, ambiguity, and limited knowledge, it can be much wiser to consider and use **imprecise probabilities** rather than the precise probabilities we've been dealing with up to now. Imprecise probabilities (IP) consist of probability *intervals* which replace the precise values we are familiar with [3]. For instance, in many cases we might not be able to know that the value of *p(Explosion in chemical plant)* is precisely equal to 0.454. Instead, our knowledge might only justify our belief it belongs to the interval [0.2;0.7]. If we consider the betting interpretation of probabilities presented in 3.3.1, for a price of 1 €, it means that we would always buy the bet on A = "*The chemical plant will explode*" for an amount equal or inferior to the **lower prevision** $\underline{P}(A) = 0.20$ € and always sell the bet for an amount equal to or higher than the **upper prevision** $\overline{P}(A) = 0.70$ €. We would be undecided for any amount of money included in the interval]0.2;0.7[€ which corresponds to our uncertainty / lack of knowledge. We shall see in 4.2 that while precise probabilities are a good description of random uncertainty (we know the risks and their probabilities), they cannot satisfyingly describe epistemic uncertainty / ignorance (we don't know the

probability of the risks). Instead of the Kolmogorov axioms, imprecise probabilities have a set of properties shown in Formula 3.9 (given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$).

$$\begin{aligned} \forall A \in \sigma, \quad 0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1 \\ \text{If } \underline{P}(A) = \overline{P}(A) = P(A) : \text{precise probability} \\ \text{If } \underline{P}(A) = 0 \text{ and } \overline{P}(A) = 1 : \text{ complete ignorance about A} \\ \text{For any disjoint events A and B:} \\ \underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B) \text{ and } \overline{P}(A \cup B) \geq \overline{P}(A) + \overline{P}(B) \\ \underline{P}(\overline{A}) = 1 - \overline{P}(A) \end{aligned} \tag{3.9}$$

$\underline{P}(A)$ can be interpreted as reflecting the evidence in favour of the event A whereas $\underline{P}(\overline{A}) = 1 - \overline{P}(A)$ reflects the evidence against A and thus for $\overline{A}$. $\Delta(A) = \overline{P}(A) - \underline{P}(A)$ reflects our ignorance, i.e. our lack of information about A. In the discrete case, we consider a **set** of prior probabilities

$$S = \{p(H_k) \in [\underline{P}(H_k); \overline{P}(H_k)], \quad k \in [|1; n|] \mid \sum_{i=1}^{n} p(H_i) = 1\}.$$

It is updated by applying Bayes' theorem to **all** $[p(H_1), p(H_2), ..., p(H_n)]$ belonging to it:

$$p(H_i|D) = \frac{p(D|H_i)p(H_i)}{p(D)} = \frac{p(D|H_i)p(H_i)}{\sum_{j=1}^{n} p(D|H_j)p(H_j)}$$

We obtain a set of posterior probabilities defined by $[\underline{P}(H_i|D); \overline{P}(H_i|D)], \ i \in [|1; n|]$ with

$$\underline{P}(H_i|D) = \min_{\{p(H_1),...,p(H_n)\} \in S} \frac{p(D|H_i)p(H_i)}{\sum_{j=1}^{n} p(D|H_j)p(H_j)}$$

$$\overline{P}(H_i|D) = \max_{\{p(H_1),...,p(H_n)\} \in S} \frac{p(D|H_i)p(H_i)}{\sum_{j=1}^{n} p(D|H_j)p(H_j)}$$

Example

Let us reconsider the example of the medical test given in 3.1. Like before, a man suffering from very general symptoms goes to his doctor to be checked for a specific disease. The physician wants to test the patient. We know that if a man is really sick, the test will be positive in 95 % of all cases, but that if he isn't sick, the test will be negative in 98 % of all cases. The difference is now that the proportion of men suffering from the sickness is more uncertain. Due to smaller samples, we only know that between 0.05 % and 0.20 % of the male population is affected by the disease. The formula to

find the posterior probability that the man suffers from the malady is still given by

$$p(Sick|Test) = \frac{p(Test|Sick)p(Sick)}{p(Test|Sick)p(Sick) + p(Test|\overline{Sick})p(\overline{Sick})}$$

This time, however, the prior probabilities are imprecise so that $p(Sick) \in [\underline{P}(Sick); \overline{P}(Sick)] = [5E-04; 2E-03]$ and $p(\overline{Sick}) \in [\underline{P}(\overline{Sick}); \overline{P}(\overline{Sick})] = [0.9980; 0.9995]$. We can also write the posterior probability that the man is sick after a positive test as

$$p(Sick|Test) = \frac{p(Test|Sick)p(Sick)}{p(Test|Sick)p(Sick) + p(Test|\overline{Sick})(1 - p(Sick))}$$

The lower bound of $p(Sick|Test)$ is given by

$$\underline{P}(Sick|Test) = min_{p(Sick)\in[5E-04;2E-03]} \frac{p(Test|Sick)p(Sick)}{p(Test|Sick)p(Sick) + p(Test|\overline{Sick})(1 - p(Sick))}$$

whereas the upper bound of $p(Sick|Test)$ is

$$\overline{P}(Sick|Test) = max_{p(Sick)\in[5E-04;2E-03]} \frac{p(Test|Sick)p(Sick)}{p(Test|Sick)p(Sick) + p(Test|\overline{Sick})(1 - p(Sick))}.$$

We find that $\underline{P}(Sick|Test) = 0.02321$ and $\overline{P}(Sick|Test) = 0.0869$.

Updating of imprecise probabilities in the case of complete ignorance

If any prior probability $p(H_k)$ is equal to 0, Bayes' theorem implies that the posterior probability $p(H_k|D)$ will also always be equal to 0, regardless of the nature and strength of the data D (this can be easily verified by writing down the formula). Likewise, if any prior probability $p(H_k)$ is equal to 1, Bayes' theorem entails that the posterior probability $p(H_k|D)$ will also always be equal to 1, regardless of the data D. This has an important consequence. We have just seen that the interval $[\underline{P}(H_k); \overline{P}(H_k)] = [0; 1]$ is the appropriate description of a state of complete ignorance, and in 4.2, we'll get familiar with strong arguments supporting this view. Nevertheless, the imprecise Bayesian updating formula we just saw will always give us [0;1] as posterior probability interval if [0;1] is our prior probability intervall. In other words, we are unable to learn anything from any new data. There are two main strategies an imprecise Bayesian can resort to to overcome this problem.

- We can use a near-ignorant imprecise prior $[\epsilon; 1 - \epsilon]$ where $\epsilon \in]0; 1[$ is very small. Example: [0.0001;0.9999].

- We keep the whole interval [0;1] to *suspend belief* so long as we don't have any data at our disposal. When the first data show up, we **replace** [0;1] by a narrower interval. For instance, if an unknown coin tossed 10 times lands 4 times on heads, our new prior probability interval might become [0.30;0.50].

Chapter 4

Why should we choose one approach rather than the other?

4.1 Prior knowledge and the problem of the p-values

4.1.1 Use of prior information

When we encounter a new problem, it often occurs that we already have relevant prior knowledge thanks to our experience with similar cases. While a frequentist would only use the data/observations *at hand* to compare models or estimate a parameter, it is possible in a Bayesian framework to systematically **combine** these new data with our prior knowledge via Bayes' theorem. To illustrate this, let us consider the estimation of the action potential rate (firing rate) θ ($.s^{-1}$) of a neuron. The true value is $\theta = 0.76$ Hz but we don't know that. However, we know through our careful study of the firing rates of **similar neurons** that they follow an exponential distribution whose inverse mean is known: $\tau = \frac{1}{0.4} = 2.5$ s. This allows us to define the prior exponential distribution $f_\tau(\theta) = \tau e^{-\tau\theta}$ which is shown in Figure 4.1 For a period $t = 5$ s, we counted the number of action potentials $y = 6$. We know that y is the realisation of a random variable Y that is Poisson-distributed so that

$$f_{Y|T=\theta}(y) = \frac{(\theta t)^y}{y!} e^{-\theta t}, y \in \mathbb{N}_0 \ .$$

The likelihood is shown in Fig. 4.2. The frequentist maximum likelihood estimator is given by

$$\hat{\theta}(y) = \frac{y}{t}.$$

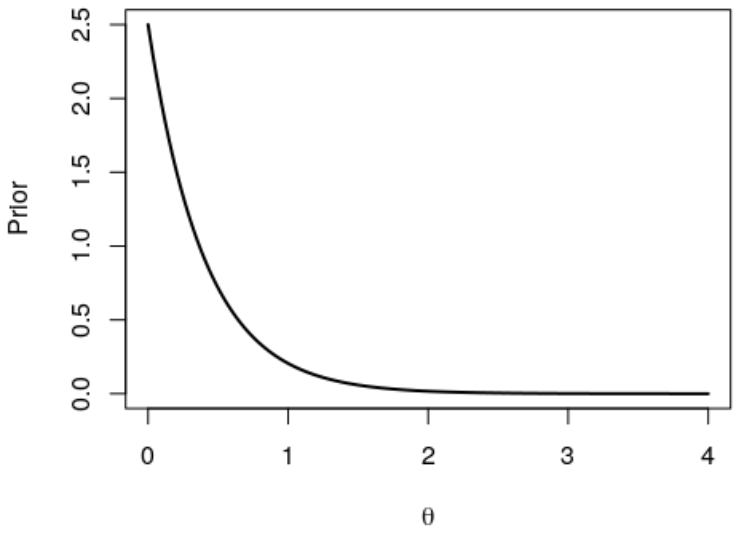


Figure 4.1: Prior distribution of the firing rate

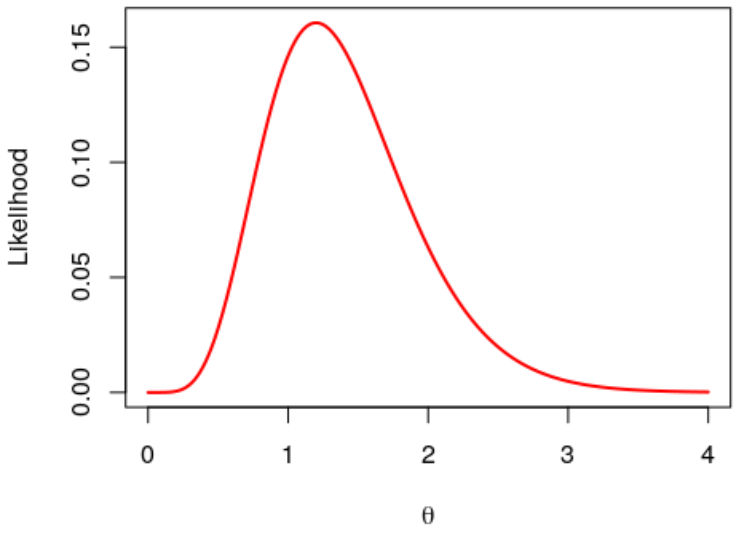


Figure 4.2: Likelihood of the firing rate

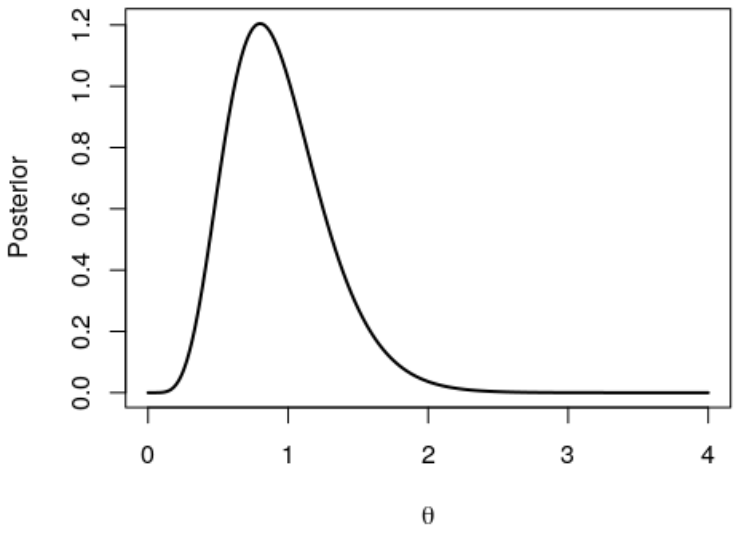


Figure 4.3: Posterior distribution of the firing rate

It is equal to $\hat{\theta}(y) = 1.2$ Hz. Now, if we're good Bayesians, we want to also use the prior distribution based on our *experience with similar neurons* to tackle this parameter estimation problem. By applying Bayes' theorem to the situation

$$f(\theta|y) = \frac{f_\tau(\theta)L(y|\theta)}{f(y)} = \frac{f_\tau(\theta)L(y|\theta)}{\int_{\theta' \in \Theta} f_\tau(\theta')L(y|\theta')} \tag{4.1}$$

we can prove mathematically that the posterior is given by

$$f(\theta|y) = \frac{\theta^y (t+\tau)^{y+1}}{y!} e^{-\theta(t+\tau)} \tag{4.2}$$

The posterior distribution is represented in Fig. 4.3. The **posterior mode** (which is the Bayesian equivalent of the frequentist Maximum Likelihood Estimator) is given by

$$\hat{\theta_p} = \frac{y}{t+\tau}.$$

By applying this formula, we find that $\hat{\theta_p} = 0.828$ Hz. It is closer to the true value ($\theta = 0.76$ Hz) than the frequentist prediction $\hat{\theta}(y) = 1.2$ Hz. Our use of a prior probability distribution *based on our previous experience* allowed us to get a more realistic estimation of the firing rate of the neuron we are studying.

4.1.2 The strong likelihood principle

Suppose that we toss an unknown coin 12 times. We want to determine the probability θ that it lands on heads, which is also a measure of its evenness/symmetry. We got 9 "heads" and 3 "tails". We want to compare $H_0 : \theta = \frac{1}{2}$ with $H_1 : \theta > \frac{1}{2}$. There are **two ways** to describe the situation and deal with the problem.

- Binomial distribution: let X be the number of times the coin landed on heads during the 12 trials. We know that $X \sim B(12, \theta)$ so that the likelihood is given by

$$L_1(\theta|X = 9) = \binom{12}{9}\theta^9(1-\theta)^3.$$

  The p-value is the probability that if the coin were perfectly symmetrical (i.e. $\theta = 0.5$), we would obtain a result that is at least as extreme as the one we got. Therefore,

$$p - value_1 = p_{\theta=0.5}(X \geq 9) = \sum_{j=9}^{12}\binom{12}{j}\theta^j(1-\theta)^{12-j} = 0.075.$$

- Negative binomial distribution. We can also conceive of the experiment as the tossing of a coin until we get 3 "tails". Thus, if Y is the number of "heads", we need 9 of them before the last toss which must be a "tail". As a consequence, the likelihood function is

$$L_2(\theta|Y = 9) = \binom{11}{9}\theta^9(1-\theta)^3$$

$$p - value_2 = p_{\theta=0.5}(Y \geq 9) = \sum_{j=9}^{\infty}\binom{2+j}{j}\theta^j(1-\theta)^3 = 0.0325$$

We are confronted with a paradox. On a fundamental level, both description of the situation are equally correct. What is more, the likelihood function of the second model $L_1(\theta|X = 9)$ is proportional to the likelihood function of the first model $L_2(\theta|Y = 9)$. The **strong likelihood principle** stipulates that all information related to a set of new data D is contained within the likelihood function. As a consequence, if a likelihood function is proportional to another one, the conclusions we can draw from each one should be exactly the same. And yet, in the first case, $p-value_1 = 0.075$ so that we'd keep the null hypothesis $H_0 : \theta = \frac{1}{2}$ at a confidence level of 95 %, while in the second case, $p - value_1 = 0.0325$, which would lead us to reject H_0 in favour of H_1 at a confidence level of 95 %. Consequently, it can be seen that frequentist methods based on p-values can lead to *inconsistent answers*. This stems

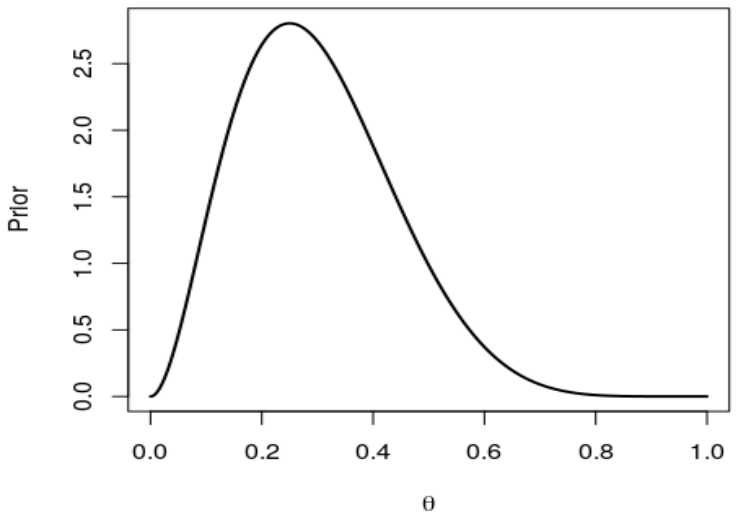


Figure 4.4: Beta prior distribution for $a = 3$ and $b = 7$

from the fact that the p-value doesn't only consider the *actually observed* data but other *unobserved fictional* data as well ($X \in [10; 12]$ for $p - value_1$ and $Y \in [10; \infty]$ for $p - value_2$). The Bayesian approach respects the strong likelihood principle. Let us assume that our prior knowledge about θ can be expressed through a **beta prior distribution** $Be_{a,b}(\theta)$ so that

$$f_{a,b}(\theta) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B_{a,b}}$$

whereby

$$B_{a,b} = \int_{\theta=0}^{1} \theta^{a-1}(1-\theta)^{b-1} d\theta \ .$$

Let us choose $a = 3$ and $b = 7$ because most coins we know of are biased towards $\theta = 0.3$. The prior distribution is represented in Fig. 4.4. The likelihoods $L_1(\theta|X = 9) = \binom{12}{9}\theta^9(1-\theta)^3$ and $L_2(\theta|Y = 9) = \binom{11}{9}\theta^9(1-\theta)^3$ can be seen in Fig. 4.5. According to Bayes' theorem, we have

$$f(\theta|x) = \frac{f(\theta)L(x|\theta)}{f(x)} = \frac{f(\theta)L(x|\theta)}{\int_{\theta' \in \Theta} f(\theta')L(x|\theta')} \ .$$

For the first binomial model, we have thus

$$f(\theta|x) = \frac{\frac{\theta^2(1-\theta)^6}{B_{3,7}}\binom{12}{9}\theta^9(1-\theta)^3}{\int_{\theta' \in [0;1]} \frac{\theta'^2(1-\theta')^6}{B_{3,7}}\binom{12}{9}\theta'^9(1-\theta')^3 d\theta'} \ .$$

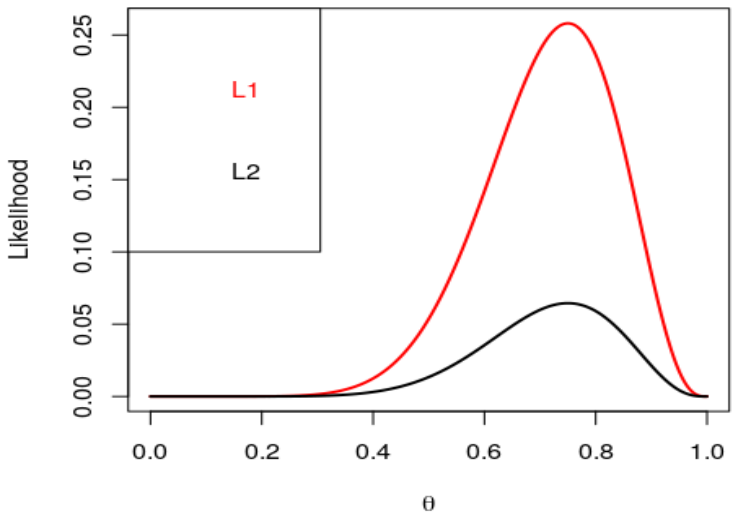


Figure 4.5: Likelihood L_1 (binomial) and L_2 (negative binomial)

$$f(\theta|x) = \frac{\theta^{11}(1-\theta)^9}{\int_{\theta' \in [0;1]} \theta'^{11}(1-\theta')^9 d\theta'}$$

$$f(\theta|x) = Be_{12,10}(\theta)$$

If we now consider the likelihood L_2 of the negative binomial model, $L_2(\theta|Y = 9) = \binom{11}{9}\theta^9(1-\theta)^3$, applying Bayes' theorem also leads us to $f(\theta|x) = Be_{12,10}(\theta)$ since the term $\binom{11}{9}$ disappears. The posterior distribution is shown in Fig. 4.6. It can be seen that the optimal value of $\theta|x$ is far away from the value of the frequentist estimator $\hat{\theta}$. This stems from our choice of a prior biased towards small values of θ, which itself reflects our foreknowledge (also called prior knowledge) about the properties of other coins.

4.2 Classical Bayesianism and the problem of ignorance

4.2.1 The principle of indifference

At this stage, we now know that the main difference between frequentism and Bayesianism is that Bayesians systematically combine a prior probability distribution with the likelihood, thereby giving rise to the posterior distribution that reflects both the prior distribution and the new data rep-

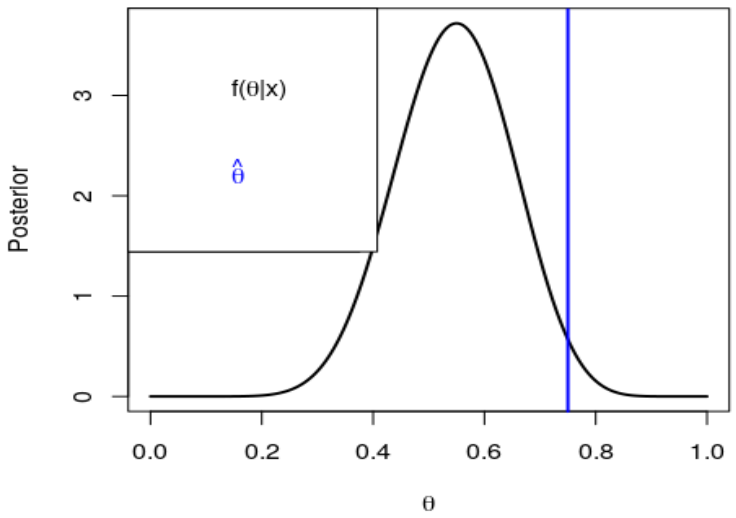


Figure 4.6: Beta posterior distribution ($a = 12$ and $b = 10$)

resented by the likelihood. In situations where we truly have **prior knowledge** at our disposal (such as in 3.1, 4.1.1 and 4.1.2), this is a very good thing, as the results of frequentist estimation based on the likelihood alone can often be far away from the real parameter value, if the sample is too small or if the data aren't sensitive enough to the parameters we're interested in. Nevertheless, there are many situations where the problem we're dealing with is so new and unknown that we don't have any relevant prior knowledge. Classical *precise* Bayesians typically use the **Principal Of Indifference** (POI) to find prior probabilities for situations we are completely ignorant about. The POI stipulates that if we are equally ignorant about the n outcomes $\mathrm{O}_1, ..., \mathrm{O}_\mathrm{n}$ of a situation, we aren't allowed to favour any one of these so that we must attribute the same probability to every possibility: $p(O_i) = \frac{1}{n}, i \in [|1; n|]$. So according to the POI, if you throw an unknown dice, you must assume that the probability you'll get a 6 is equal to 1/6 since there are only six outcomes. In the case of a continuous parameter, the principle of indifference leads to a *uniform prior distribution* (flat prior), since no parameter value can be privileged over another. For example, if the only thing we know about a chemical kinetic parameter k is that it belongs to the interval [0.1;1000], applying the principal of indifference would produce the prior distribution f_0 represented in Fig. 4.7, which is such that

$$f_0 = \frac{1}{1\mathrm{E}+03-1\mathrm{E}-01} \approx 1E-03.$$

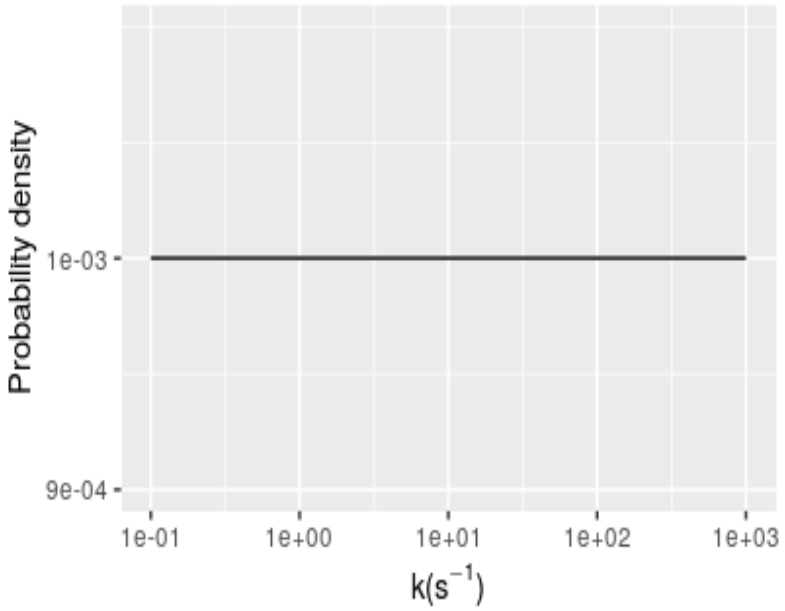


Figure 4.7: Uniform prior for $k \in [0.1; 1000]$

4.2.2 Problem: mistaking ignorance for knowledge

Ever since it was formulated by French mathematician Pierre-Simon Laplace and British economist John Maynard Keynes, the principle of indifference has faced many challenges and objections. The main problem is that it leads one to confuse knowledge and ignorance. To illustrate this, we can consider the simple case of coin tossing (see Fig. 4.8).

Figure 4.8: Coin tossing

The situation can be described as follows:

- Through careful physical measurements, we know that coin A is perfectly symmetrical.

- We tossed coin B one million times and saw that it tends to land on heads half of the time.
- We know absolutely nothing about coin C.

For both coin A and B, our extensive knowledge allows us to justifyingly say that the probability that the coin will land on heads is 0.5. If we were now to apply the principle of indifference to coin C, we would also conclude that the probability of its landing on heads is equal to 0.5. This, however, comes down to saying:

> Since I know absolutely nothing about coin C, I know that the probability of its landing on heads is 0.5 as if I had thoroughly characterised it physically.

That sentence seems clearly **absurd**: if we know absolutely nothing about coin C, then we cannot know it is as likely to land on heads as on tails. According to the late Bayesian philosopher of science Wesley Salmon, the principle of indifference amounts to magical thinking [4]:

> Knowledge of probabilities is concrete knowledge about occurrences; otherwise it is useless for prediction and action. According to the principle of indifference, this kind of knowledge can result immediately from our ignorance of reasons to regard one occurrence as more probable than another. This is epistemological magic. Of course, there are ways of transforming ignorance into knowledge – by further investigation and the accumulation of more information. It is the same with all "magic": to get the rabbit out of the hat you first have to put him in. The principle of indifference tries to perform "real magic".

It is in such situations that **imprecise probabilities** turn out to be a very useful and fruitful framework. While an imprecise Bayesian would agree that p(A) = p(B) = 0.5, he or she would describe complete ignorance in the third case as $p(C) = [0; 1]$. In other words, while we know nothing about the coin, we represent our level of knowledge by the whole interval [0;1].

4.2.3 Can second order probabilities come to our rescue?

A classical Bayesian using only precise probabilities might answer that it isn't the principle of indifference which is mistaken but that our application to the coin problem was all too simplistic. Instead of directly applying it to the binary variable $X \in \{heads, tails\}$, we could apply it to

$$\theta = p(X = "heads"),$$

which is a continuous variable belonging to the interval [0;1]. It is called the second-order probability. This straightforwardly leads to a uniform prior on

[0;1] which is equal to 1 on this interval. Let us now suppose that, *after* having chosen this uniform prior, we toss the coin n times and count $X = k$ heads. As we saw in 4.1.2, the likelihood is given by

$$L_1(X = k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

It can be shown through Bayes' theorem that the posterior probability distribution is given by the beta distribution $Be_{k+1,n-k+1}(\theta)$ so that

$$f(\theta|X = k) = \frac{\theta^k(1-\theta)^{n-k}}{B_{k+1,n-k+1}}$$

whereby

$$B_{k+1,n-k+1} = \int_{x=0}^{1} \theta^k(1-\theta)^{n-k}d\theta \ .$$

Let us now represent the probability distribution for different values of (n,k): (0,0), which corresponds to our complete ignorance in the absence of any data, (1,2), (5,10), (50,100), (500,1000), and (50000,100000). The results are shown in Figure 4.9.

In all cases, we have $p(heads) = p(tails) = 0.5$. However, there are strong differences between the probability distributions of $\theta = p(heads)$. While it is completely flat in the absence of data, the posterior probability distribution becomes sharper and sharper as we add more information. A classical Bayesian would argue that the difference between ignorance and knowledge lies in the flatness of $f(\theta)$. Total ignorance corresponds to a flat (uniform) probability distribution whereas knowledge corresponds to a sharp probability distribution (complete knowledge corresponds to a distribution which is singular at $\theta = 0.5$). However, such an answer merely pushes back the problem. To see why, let us consider a uniform prior and two hypotheses:

- H_1: the coin is strongly biased towards heads, i.e. $\theta < 0.01$.
- H_2: the coin is more or less fair, i.e. $0.40 \leq \theta \leq 0.60$

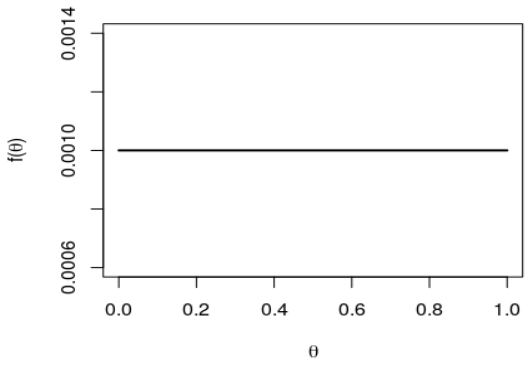
We have $p(H_1) = \int_{\theta=0}^{0.01} 1d\theta = 0.01$ and $p(H_2) = \int_{\theta=0.40}^{0.60} 1d\theta = 0.20$. This leads to

$$\frac{p(H_2)}{p(H_1)} = 20.$$

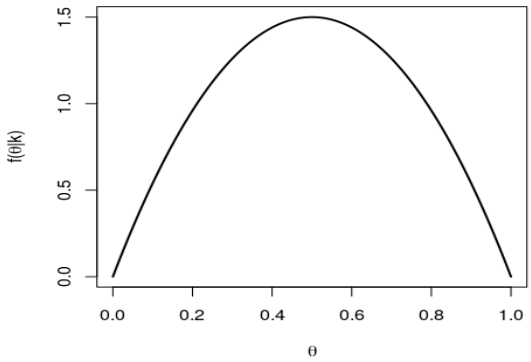
Applying the principle of indifference would lead straight to the following conclusion:

> Since we know absolutely nothing about the coin, it is 20 times more likely it is relatively fair than that it is strongly biased towards heads.

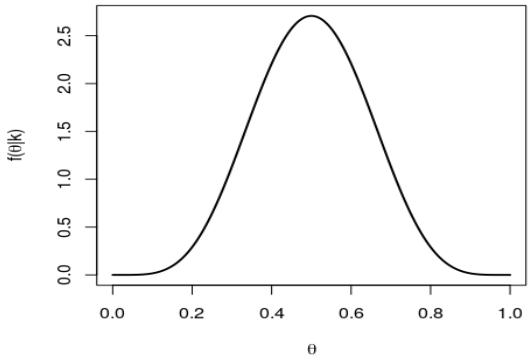
We are thus still confronted with the absurd creation of knowledge out of sheer ignorance.



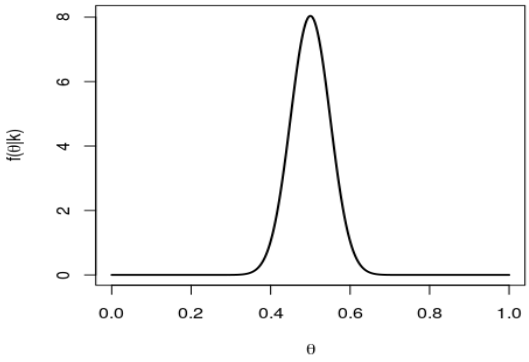
(a) Prior



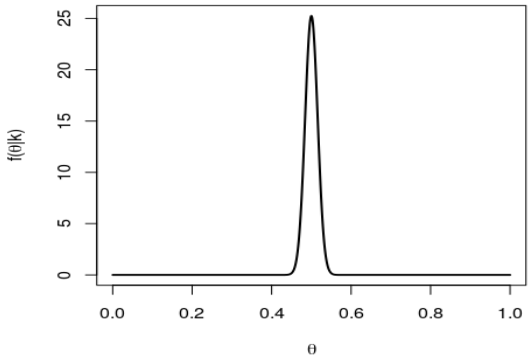
(b) Posterior: $n = 2$



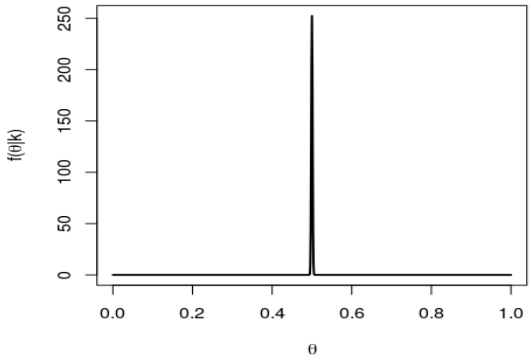
(c) Posterior: $n = 10$



(d) Posterior: $n = 100$



(e) Posterior: $n = 1000$



(f) Posterior: $n = 100000$

Figure 4.9: Prior and evolution of the posterior

4.2.4 Arbitrariness of the prior

Another concern related to the application of the principle of indifference is the fact that we get contradictory results depending on the parameter we apply it to. Let us consider a problem of chemical kinetics to illustrate this. We are interested in the elementary reaction $\mathrm{R}+\mathrm{R} \rightarrow \mathrm{P}_2$. We measured the evolution of the concentration of species R as a function of time. We know that the system is governed by the following differential equation:

$$\frac{d[R]}{dt} = -2k[R]^2$$

where $[R]_0$ (mol/m^3) is the initial concentration and k (mol.cm^{-3}.s^{-1}) is the reaction coefficient we want to estimate. Our prior knowledge is limited to the fact that $k \in [5\text{E}{+}04; 5\text{E}{+}09]$. We could also reformulate the problem as the estimation of other parameters such as $u = 1/k$ so that $u \in [2\text{E}{-}10; 2\text{E}{-}05]$ or $w = log_{10}(k)$ so that $w \in [log_{10}(5\text{E}{+}04); log_{10}(5\text{E}{+}09)]$. It would be *equally justifiable* to apply the principle of indifference to any of these variables. However, if we do so, the results we get aren't consistent with one another. The uniform prior on k is given by

$$f_k(k) = \frac{1}{5\text{E}{+}09 - 5\text{E}{+}04} \approx 2\text{E}{-}10.$$

The uniform prior on u is given by

$$f_u(u) = \frac{1}{2\text{E}{-}05 - 2\text{E}{-}10} \approx 5\text{E}{+}04.$$

The cumulative prior probability distribution is therefore approximately given by

$$F_u(u) = p(U \leq u) = 5\text{E}{+}04(u - 2\text{E}{-}10).$$

We thus have

$$p(U \leq u) = p(K^{-1} \leq k^{-1}) = p(k \leq K) = 1 - F_u(k) = 5\text{E}{+}04(k^{-1} - 2\text{E}{-}10).$$

$$F_u(k) \approx 1 - 5\text{E}{+}04k^{-1}.$$

A simple derivation leads to $f_u(k) = 5\text{E}{+}04k^{-2}$. Likewise, the uniform prior on w is given by

$$f_w(w) = \frac{1}{log_{10}(5\text{E}{+}09) - log_{10}(5\text{E}{+}04)} = 0.20.$$

The cumulative prior probability distribution is

$$F_w(w) = p(W \leq w) = p(ln(K) \leq ln(k)) = p(K \leq k) = 0.20(w - log10(5\text{E}{+}04)).$$

$$F_w(k) \approx 0.2log10(k) - 0.939794$$

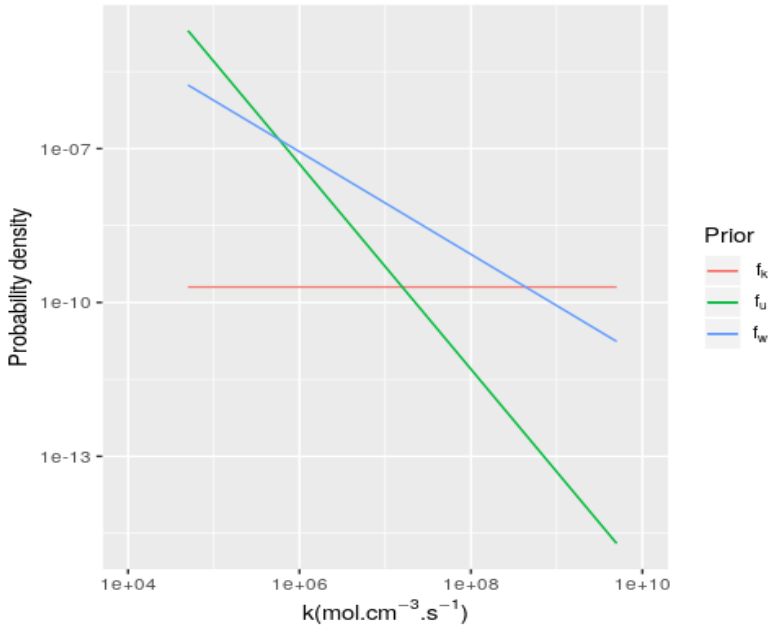


Figure 4.10: Uniform priors for $k \in$ [5E+04;5E+09]

This leads to

$$f_w(k) = \frac{0.20}{ln(10)k}$$

The three priors can be seen in Figure 4.10. Let us now calculate $p(k \in$ [5E+04;5E+06]). If we choose f_k, we find that

$$p(k \in [5E{+}04; 5E{+}06]) = 2E{-}10(5E{+}06 - 5E{+}04) = 0.00099.$$

If we use f_u, we have

$$p(k \in [5E{+}04; 5E{+}06]) = 5E{+}04(\frac{1}{5E{+}04} - \frac{1}{5E{+}06}) = 0.99.$$

And if we choose f_w, we get

$$p(k \in [5E{+}04; 5E{+}06]) = 0.20(log_{10}(5E{+}06) - log_{10}(5E{+}04)) = 0.4$$

We see thus that the choice of the prior has a huge impact on these probabilities. Consequently, it is misguided to call such uniform priors "uninformative priors" or "non-informative priors" since they always express a specific information [5, 6, 7]. None of these values alone can be considered an adequate representation of our ignorance. Instead, complete ignorance should **theoretically** be represented by the union of all possible probability densities which are equal to 0 outside the interval [5E+04;5E+09]. Updating a set of prior probability distributions in a situation of complete ignorance

is far from being a trivial task [8]. As we saw in 3.3.2, it demands either the use of near-ignorance priors or a narrower selection of priors thanks to experimental information.

4.2.5 Jeffreys' Prior

Aware of the problem of the arbitrariness of uniform priors, British mathematician Sir Harold Jeffrey sought to define a prior that would be completely independent upon parametrisation, i.e. upon the way the problem is expressed [9]. Let us consider a model M (accounting for a random vector $X = (X_1, ...X_n) \in \mathbb{R}^n$ made of n independent identically distributed random variables) which depends on a parameter θ. Let's suppose we've measured a realisation of X that we call $x = (x_1, ..., x_n)$. $L(x|\theta)$ designates the likelihood that is only a function of θ since x is known. Let $\theta^* = \theta^*(x)$ be the maximum likelihood estimate of the parameter. For fixed values of the data x, the **observed** Fisher information is defined by Equation 4.3.

$$I_{\theta^*}(x) = -\frac{d^2}{d\theta^2} ln(L(x|\theta))(\theta^*) \tag{4.3}$$

The observed Fisher information measures the quality of the information contained in x to estimate the true value of θ.

The expected Fisher information (or simply Fisher information) is the expected value of $I_{\theta*}(X)$ with respect to $X(\theta)$.

$$I(\theta) = E_{X|\theta}(I_{\theta*}(X)) = \int_{x \in \mathbb{R}^n} I_{\theta*(x)}(x) L(x|\theta) dx \tag{4.4}$$

It is the expected information brought to us by the random variable X, that is to say the information we'd expect before knowing the values of the sample x.

The Jeffreys' prior is a probability distribution defined in such a way that it is proportional to the square root of the determinant of the Fisher information.

$$p(\theta) \propto \sqrt{det(I(\theta))} \tag{4.5}$$

Let us consider a monotonous transformation $\phi = h(\theta)$ which associates one parameter ϕ to every parameter θ. If $p_1(\theta)$ is the Jeffreys' prior of θ and $p_2(\phi)$ is the Jeffreys' prior of ϕ and P_1 and P_2 are the corresponding cumulative probability distributions, we always have $P_1(\theta) = P_2(h(\theta))$ for any θ belonging to the parameter space. This means that Jeffreys' prior is invariant under a change of coordinates for the parameter vector. The paradoxes we saw in 4.2.4 can no longer be observed with this prior. Unfortunately, it isn't a legitimate representation of ignorance. To see why, let us consider again coin tossing. Like in 4.2.3, we'll consider the probability of getting a

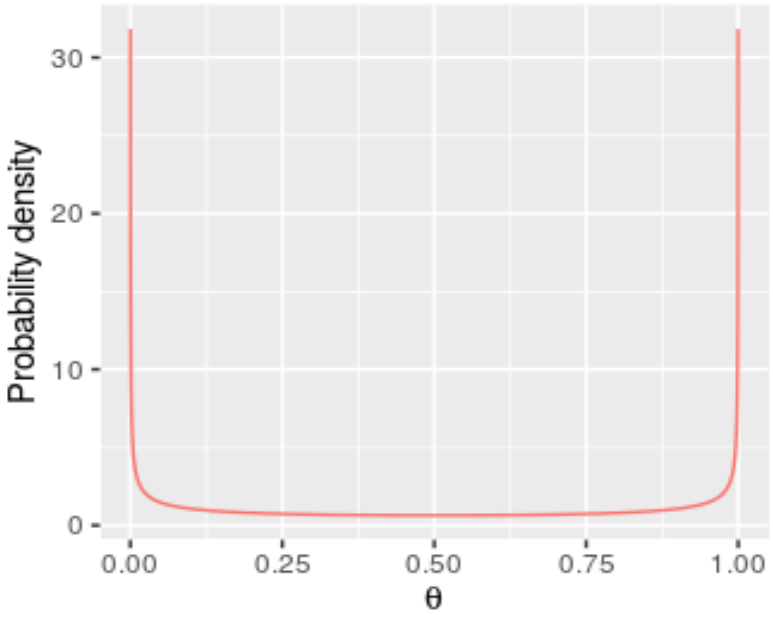


Figure 4.11: Jeffreys' prior for coin tossing

head as the parameter θ. The likelihood is

$$L(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

The Jeffreys' prior associated to that situation is given by the function $Beta(0.5, 0.5)$ which is shown in Figure 4.11. It can be clearly recognised that this distribution is **not neutral** but strongly favours extreme values. It is thus misleading to refer to Jeffreys' prior as a non-informative prior since it expresses a very specific information [5].

4.2.6 Washing out the priors

We have seen two reasons why a precise Bayesian framework is *conceptually inadequate* to represent a state of complete ignorance. However, in cases where we have many *relevant* data at our disposal, the prior will often only have a negligible effect on the posterior probability distribution. In that situation, we say that the data **wash out** the prior or that the posterior is *data dominated*.
To illustrate this, we'll consider that we just sneaked into a casino where there are three types of coin tossing machines.

1. One machine where the probability that the coin lands on heads θ is **uniformly distributed**. This means that for different games, it

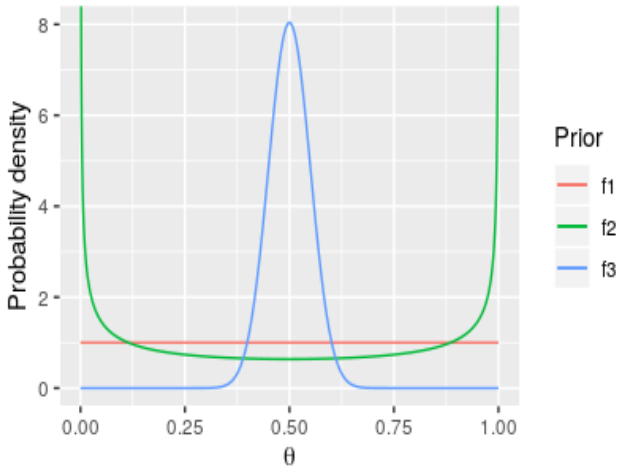


Figure 4.12: Priors for the casino

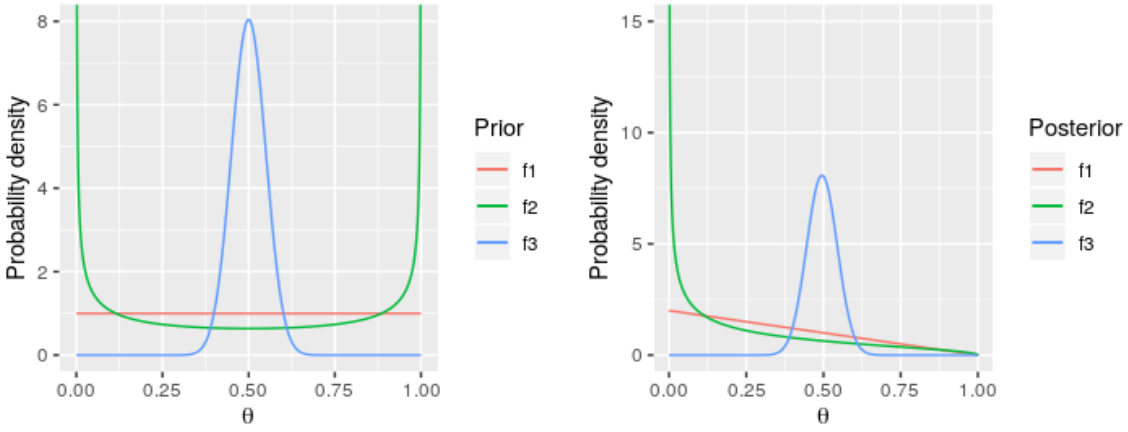
can be close to 0.115, 0.251, 0.712, 0.396, 0.407...or any other number between 0 and 1 with the same probability. [1]

2. One machine is rigged and only tosses coins that are extremely biased towards either heads or tails.

3. One machine relies only on relatively fair coins, so that θ cannot be too far from 0.50.

The owner of the casino greets us enthusiastically and proposes us to get started on a brand-new machine he recently acquired. However, he mischievously forgets to tell us which type of machine it is. In such a situation, it is appropriate to use an *imprecise prior* made of the three priors corresponding to the three machines. They're represented in Figure 4.12. As we get more data, we keep updating the three priors. The results can be seen in Figures 4.13 and Figure 4.14. While the two first priors quickly converge, the third one needs much more time. In general, one of the main reasons of the success of classical (precise) Bayesian methods that only use one prior boils down to the huge amount of data that is being drawn upon so that the effects of the prior becomes completely negligible. Nevertheless, in situations where there are less data, it is always wise to employ several priors. This is known of as **robust Bayesian analysis**, that is to say a Bayesian analysis that is robust to the choice of the prior. It is worth noting that all data are not necessarily *relevant* for the estimation of a given parameter of a model. In chemistry, for example, it can often occur that concentration profiles are highly sensitive to some reaction parameters whereas other parameters have

[1] In reality, this won't be a continuous but a very large discrete ensemble.



(a) n = 0

(b) n = 1

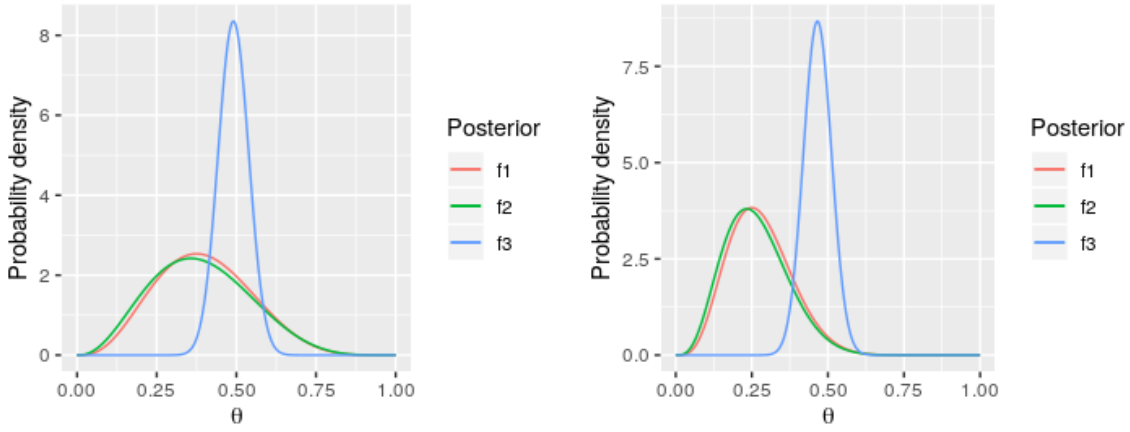
(c) n = 2

(d) n = 3

(e) n = 4

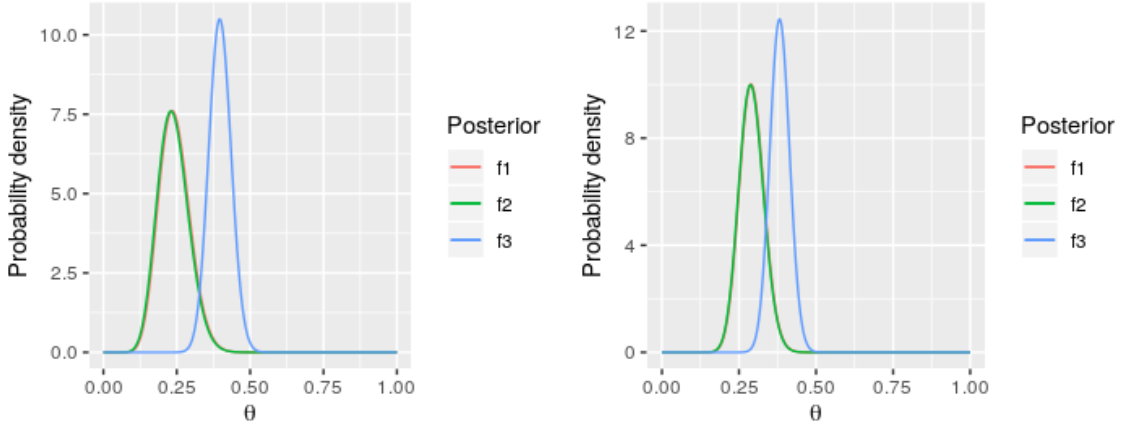
(f) n = 5

Figure 4.13: Evolution of the posteriors



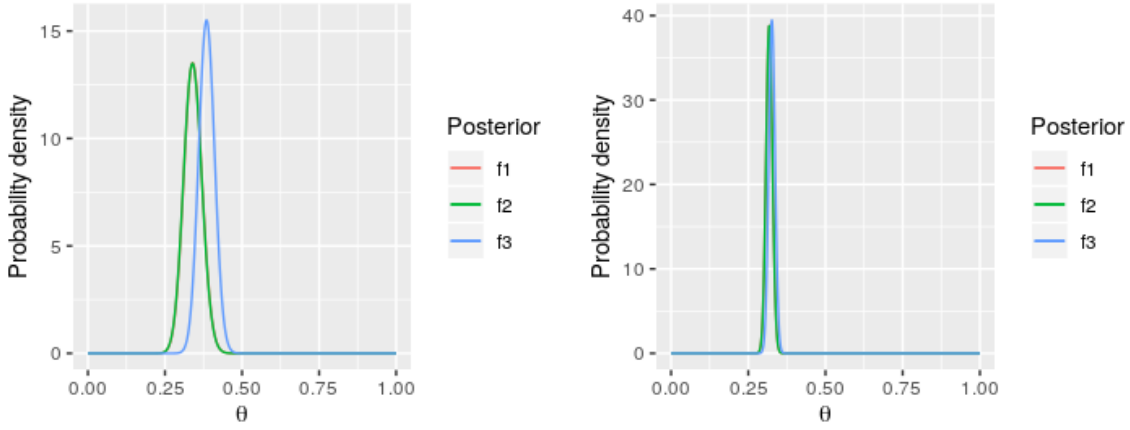
(a) n = 8

(b) n = 16



(c) n = 64

(d) n = 128



(e) n = 256

(f) n = 2048

Figure 4.14: Evolution of the posteriors

only a small influence on them. In that case, even a very large amount of such data will not be sufficient to wash out the priors.

Chapter 5

Conclusion

Nowadays, probabilistic reasoning has become an essential part of every field of engineering. The queuing theory, i.e. the probabilistic study of waiting lines, is being applied to logistic problems in companies as diverse as Amazon, Google, PSA, or the SCNF, to name but a few. Statistical analyses are routinely being employed in production chains to check the quality of products and optimise the production times. Probabilistic models are being constantly developed and refined to predict the preferences and behaviours of clients and optimise marketing strategies. Probabilities are fundamental to the work of any modern environmental health safety engineer since it is entirely based on the notion of risks that no deterministic approach can tackle. Probabilistic and statistical techniques related to big data can be applied to an enormous range of topics such as fraud detection, medical diagnostic, smart farming, the fight against terrorism, the prediction of epidemics, and strategies for reducing burnout rates in enterprises. One key message of this lecture is that **there is no consensus** about which probabilistic methods should be used and there remain many disagreements among theoreticians and practitioners alike. We have seen that the main division concerns frequentists (who only see probabilities as the tendency of a situation to produce limit frequencies if the conditions could be repeated an endless number of times) and Bayesians (who view probabilities as degrees of belief someone must have in order to be rational). Frequentism is often referred to as *classical statistics*, even though it is a later invention than Bayesianism. It is still the dominant approach in many settings, whereby many of its practitioners aren't even aware of the underlying philosophy. It is, for example, often used to analyse defective products in small and medium-sized enterprises. And in many situations, it delivers satisfying results that are sufficient to bring about the desired outcomes. Most progress of medicine and the design of new medication were based on a frequentist framework. However, it also has some shortcomings that can become quite severe in many situations. As we saw in 4.1.1 , it systematically ignores

useful *prior information*. This can be very problematic if the data sample is small or not sensitive enough to some of the parameters we want to estimate. In that case, the Bayesian philosophy allows one to fruitfully combine one's prior knowledge with the new data. In 4.1.2, we were confronted with another problem. The p-value, which is utilised by frequentists to compare two hypotheses, can be inconsistent and depend on the way we describe the problem. Bayesianism avoids this paradox.

However, there are situations where **precise** Bayesianism (that describes the probability of an event as a single value, e.g. $p(rain) = 0.33$) can lead to misleading conclusions. As we realised in 4.2, precise Bayesianism cannot express the difference between knowledge (we studied a coin physically and know it is symmetrical) and ignorance (we know absolutely nothing about the coin). While a probability of 0.5 would be the correct description of the first situation, the second situation can only be described through an imprecise probability interval of [0;1]. Another issue with precise Bayesianism when applied to continuous parameters is that it cannot determine a unique prior probability distribution **truly** representing complete ignorance so that several prior distributions should be used to express our lack of knowledge. In 4.2.6, we saw that with enough data, the influence of the prior distribution will become smaller and smaller, so that in the end we'll get the same results as if we had been performing a precise Bayesian analysis all along. This, however, only works if the data are sensitive to the parameters we want to estimate. The most important message of this lecture is that you should never use probabilistic techniques uncritically like a black box but always seek to understand what you're doing from a conceptual standpoint.

Bibliography

[1] C. Andrieu, N. De Freitas, A. Doucet, M.I. Jordan, Machine learning **50**(1-2), 5 (2003)

[2] S. Vineberg. Dutch book arguments. The Stanford Encyclopedia of Philosophy (Spring 2016 Edition), Edward N. Zalta (ed.) (1999). Available online under https://plato.stanford.edu/archives/spr2016/entries/dutch-book/;

[3] P. Walley, International Journal of Approximate Reasoning **24**, 125 (2000)

[4] W. Salmon, *The Foundations of Scientific Inference* (University of Pittsburgh, 1967)

[5] R.E. Kass, L. Wasserman, Journal of the American Statistical Association **91**(435), 1343 (1996)

[6] J. Norton, Philosophy of Science **75**, 45 (2008)

[7] J. Norton, Philosophy of Statistic **7** (2011)

[8] J.O. Berger, Journal of statistical planning and inference **25**(3), 303 (1990)

[9] H. Jeffreys, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **186**(1007), 453 (1946)