Frequentist and Bayesian approaches to uncertainty

Marc Fischer

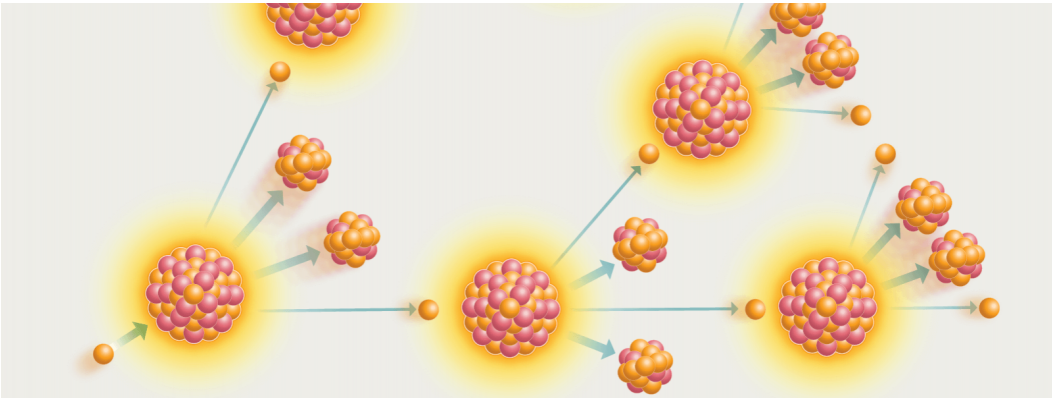
marc.fischer@emse.fr

Mines Saint-Etienne

1. Introduction

Most problems of the modern world and of our individual lives involve probabilities.

"What is the probability that this heavy atom will lose neutrons and cause an uncontrollable chain reaction?"



“How likely is it that a burglar will break into my house while I’m dreaming?”

“How plausible is it that Hubert is the murderer?”

- We would all prefer to be absolutely certain that some propositions are true or that some events will take place.

- Alas, absolute certainty can only be reached in the realm of pure mathematics.

- Real-world problems always possess an inherent uncertainty.

- Probability theory has been originally developed to model games of chance.

- Surprisingly enough, no consensus has ever been reached about the very **nature** of probability.

- Currently, there are two main camps.

1) Frequentists believe that probabilities are the limits of frequencies.

2) Bayesians believe that probabilities are rational degrees of belief.

- The Bayesian approach itself can be divided into

  - precise Bayesianism: probabilities are single values: $p(rain) = 0.35$.

- imprecise Bayesianism: interval probabilities: $p(rain) = [0.30;0.40]$.

Goals of this lecture

- Understand the philosophical differences between frequentism, precise Bayesianism and imprecise Bayesianism.
- Know how to apply these frameworks to concrete problems.
- Understand the advantages and limitations of each approach.
- Get used to working in English :-)

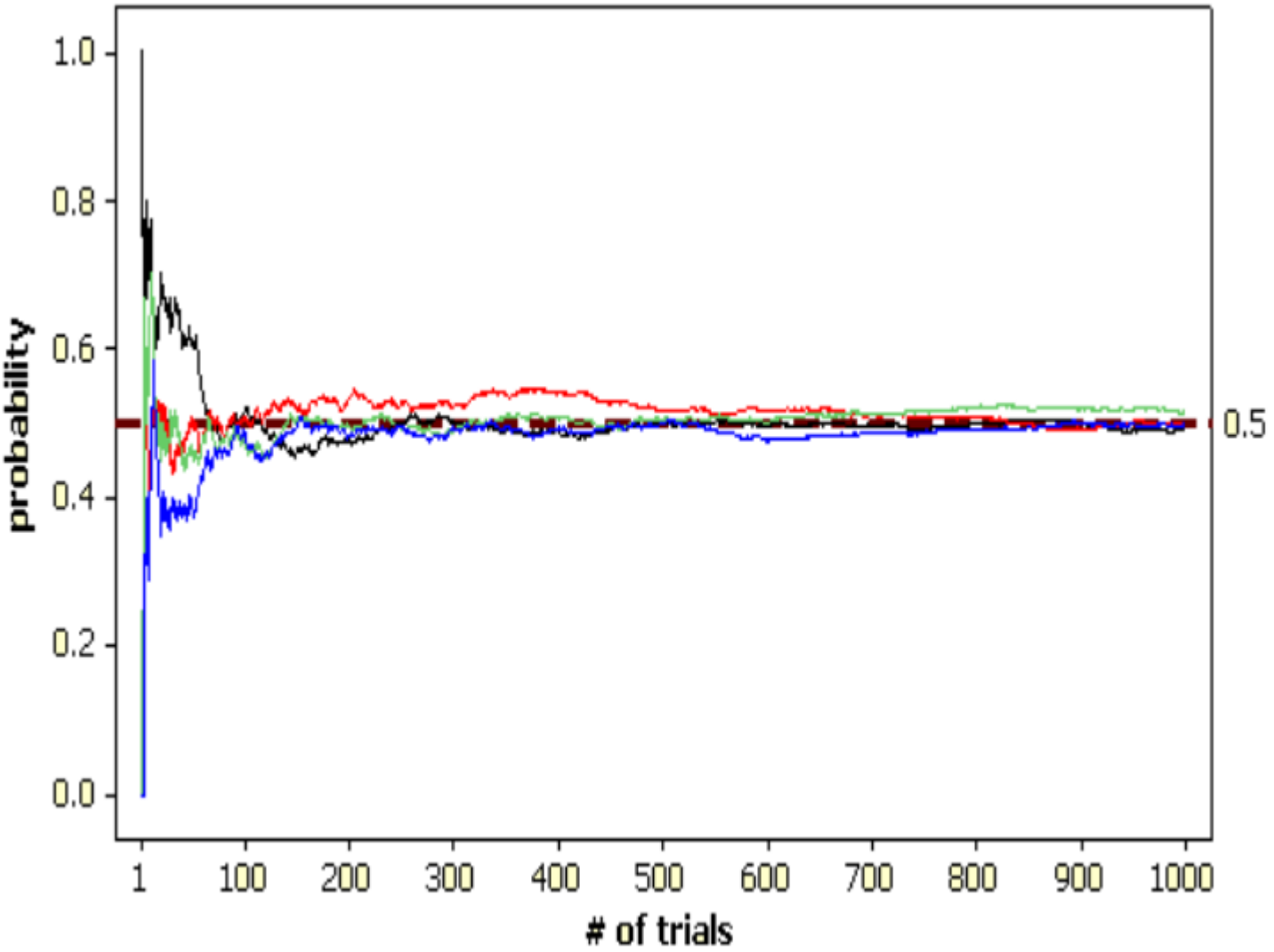
2. Evaluation

- There will be two practical works (TP) in R and Python.
- On the 29/09, you'll give a presentation on an article that interests you.
- The **final mark** will be a combination of the practical works and of the presentation.
- You're encouraged to read the **lecture material ("Frequentism_Bayesianism.pdf")** before the practical works and the presentation.
- The **lecture material** is more detailed than this PowerPoint presentation.

3) Frequentism in a nutshell

- Normally, you're already familiar with frequentism.
- Your first year lecture about statistics and data science was entirely based on a frequentist framework.
- Frequentists define a probability as the limit of the frequency of an event if it could be repeated an infinite number of times.

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


1.0
0.8
0.6
0.4
0.2
0.0
probability
0.5
1
100
200
300
400
500
600
700
800
900
1000
# of trials

- While facing uncertainties, frequentists ask themselves the following question: "*If hypothesis H is true, what is the probability that we'd get the data D we actually see?*"

- Example: coin tossing.
- We're interested in the probability θ that the coin lands on heads.
- We know it was tossed n = 1E+05 times and that it landed k = 49995 times on heads.
- Hypothesis H: the coin is symmetrical $\rightarrow$ $\theta = 0.50$.

- Ideally, we'd like to base all our conclusions solely on the likelihood *p(D|H)*.
- However, it is not possible to define a **threshold** as this value can be arbitrarily small even if hypothesis H is true.

- In the case of the coin, $p(D|H) = \binom{n}{k}\theta^{50005}(1-\theta)^{49995}$, $p(D|H) = 0.002522$.

- As a consequence, we modify the question: "*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 see?*"

- This leads to the definition of the **p-value.**

Distance between prediction (H_0) and reality

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

- In the case of the coin, *p-value* = 0.4861.
- This means we could very plausibly observe our data if the coin were symmetrical.
- We **cannot,** however, draw the positive conclusion that the coin is probably symmetrical.
- Likewise, a small *p-value* doesn't necessarily mean that the hypothesis H is false!
- We therefore need a way to compare different hypotheses with one another.

- Frequentists use the Maximum Likelihood Estimation (MLE) method to that end.
- In the discrete case, if we compare *n* hypotheses H_1, H_2,…, H_n, the maximum likelihood estimator is the hypothesis H_{max} that maximises the probability of observing the data: $H_{max} = \underset{H_i}{\operatorname{argmax}}\, p\left(D \middle| H_i\right)$.

- If the parameter space is continuous, there is an infinite number of possible values and hence also possible hypotheses.

- We must still maximise the likelihood which is then the probability (or probability density) of observing the data given a specific value of the parameter.

- In the case of the coin, we have

$$\hat{\theta} = \arg\max_{\theta \in \Theta} p(D \mid \theta) = \arg\max_{\theta \in \Theta} \binom{n}{k} \theta^{50005} (1-\theta)^{49995}$$

- This naturally leads to the empirical mean $\hat{\theta} = 50005/100000$.

- Usually, a single value doesn't suffice to estimate a parameter.

- For that reason, frequentists also use a confidence interval $[u(D), v(D)]$.
  $p(u(D) \leq k \leq v(D)) = \gamma \in [0; 1]$

- Wrong interpretation: "*the probability that k belongs to the interval* $[u(D),v(D)]$ *is equal to* γ .

- In frequentism, *k* is an unknown parameter but **not a random variable**.

- Consequently, it doesn't have any probability distribution.

- It is *u(D)* and *v(D)* that are the random variables which form a random interval.

- Right interpretation: "If we repeat the experiment/sampling an endless number of times, *k* will belong to $[u(D),v(D)]$ in $(100\gamma)\%$ of all cases.

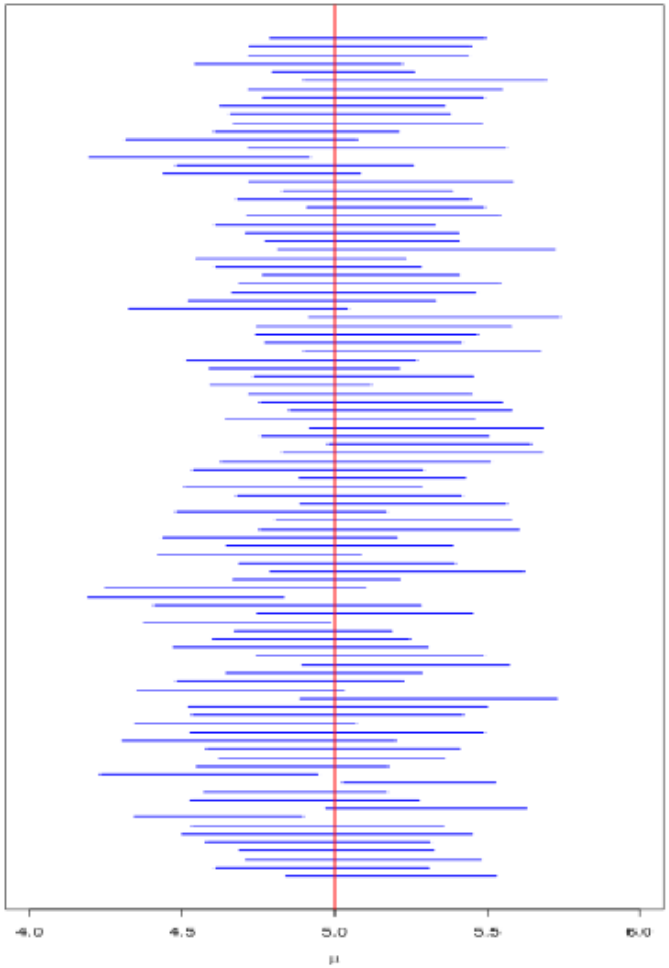


Figure 1: Random confidence intervals

4. Bayesianism in a nutshell.

4.1. Discrete parameter space

- Frequentists ask themselves: "*What is the probability of observing the data D given that hypothesis H is true?*"

- Bayesians go one step further and ask themselves: "What is the probability of hypothesis H given the data D?"

- Probabilities are no longer seen as frequency limits but as (rational) degrees of belief.

p(rain) = 1: I know it will rain tomorrow.
p(rain) = 0: I know it won't rain tomorrow.
p(rain) = 0.7: I'm not sure but I think it's more likely it'll rain.

As we might have guessed, Bayesianism is based on Bayes' theorem.

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

$p(H_k)$: prior probability of hypothesis H_k.
$p(D|H_k)$: likelihood, probability of the data D given H_k.
$p(D) = \sum_{j=1}^{n} p(D|H_j)p(H_j)$: total probability of observing the data.
$p(H_i|D)$: posterior probability of H_i given the data.

- The posterior probability $p(H_i|D)$ is one of the most important quantities in Bayesianism.

- Let us illustrate all of this through a concrete example.

- A man goes to his physician because he fears to have some specific cancer.

His doctor carries out a test to detect the disease.

- We know that only 0.1% of the population suffers from that cancer.
- If a person is sick, the test is positive in 95% of all cases.
- If a person is not sick, the test is negative in 98% of all cases.
- The physician finally tells the patient that the test is positive.
- Would you be **terrified** in such a situation?

- Thanks to the data, we know that $p(Sick) = 0.1\ \%$,
  $p(\overline{Sick}) = 99.9\ \%$, *p(Test|Sick)* = 95%, and $p(Test|\overline{Sick}) = 2\ \%$.

- According to Bayes' theorem, we find that

$$p(Sick|Test) = \frac{p(Test|Sick)p(Sick)}{p(Test|Sick)p(Sick) + p(Test|\overline{Sick})p(\overline{Sick})} = 4.539\ \%$$

- A study has shown that most people intuitively think that the posterior probability is very high.

$\rightarrow$ Bayes' theorem corrects our fallible human intuitions.

4.2. Continuous parameter space

- Bayesianism can also be successfully applied to problems involving a continuous parameter space.

- $\theta \in \Theta$ is a continuous parameter whereas the observed data x can be either discrete or continuous.

- Bayes' theorem $f(\theta|x) = \dfrac{f(\theta)L(x|\theta)}{f(x)} = \dfrac{f(\theta)L(x|\theta)}{\int_{\theta' \in \Theta} f(\theta')L(x|\theta')d\theta'}$.

- $f(\theta)$: prior probability density of the parameter.

- $f(x) = \int_{\theta' \in \Theta} f(\theta') L(x|\theta') d\theta'$: total probability (density) of the data.

- $L(x|\theta)$: likelihood. Probability density of the data *x* given a fixed parameter value θ .

- $f(\theta|x)$ is the posterior probability density.

Example

- Let's consider *n* independent identically distributed (**iid**) variables $X = (X_1, X_2, ..., X_n)$ with $X_i \sim N(\mu, \sigma^2)$.

- μ is unknown but σ is known.
- The likelihood is given by $L(x|\mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$.

- Now, we must choose a prior probability distribution for μ.
- This is the strangest aspect of Bayesianism for novices!
- We can choose another normal distribution as prior probability density.

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

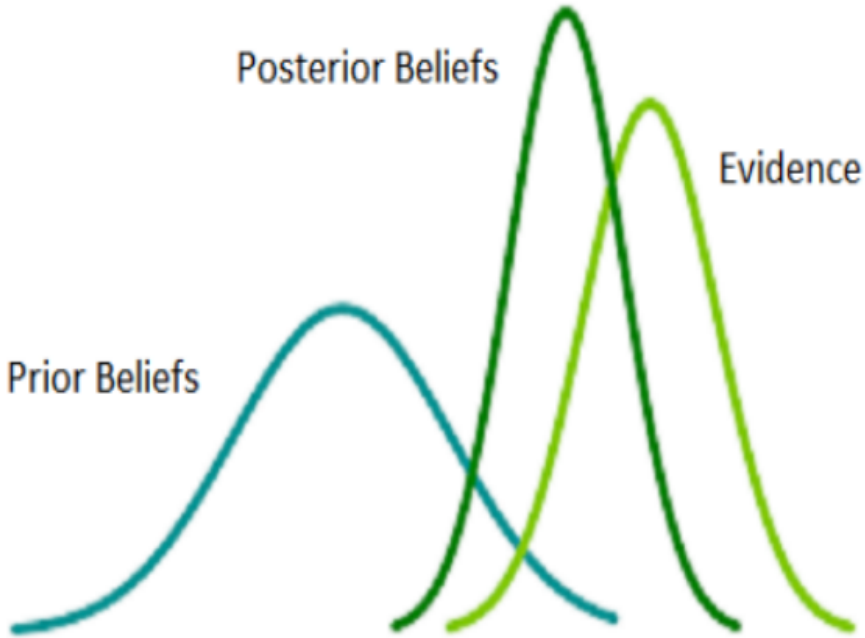
- *m* and *s* are two constants we choose according to our **prior knowledge**.
- Through Bayes' theorem, we can prove that

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

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

- A prior density and a likelihood are said to be <u>conjugate</u> to each other if the posterior density belongs to the same family.

- That's clearly the case here with the normal distribution.



- The posterior distribution is located between the prior and the evidence (likelihood).

- It is also much **narrower** than the prior, which means we have **reduced the uncertainty** of the parameter values thanks to the likelihood.
- Bayesians use several punctual estimators for the parameter value.

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

- 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 several parameters.

- T_i is the random variable corresponding to the realisation θ_i .
- $\overline{\theta}(x)$ is thus the vector of the means.

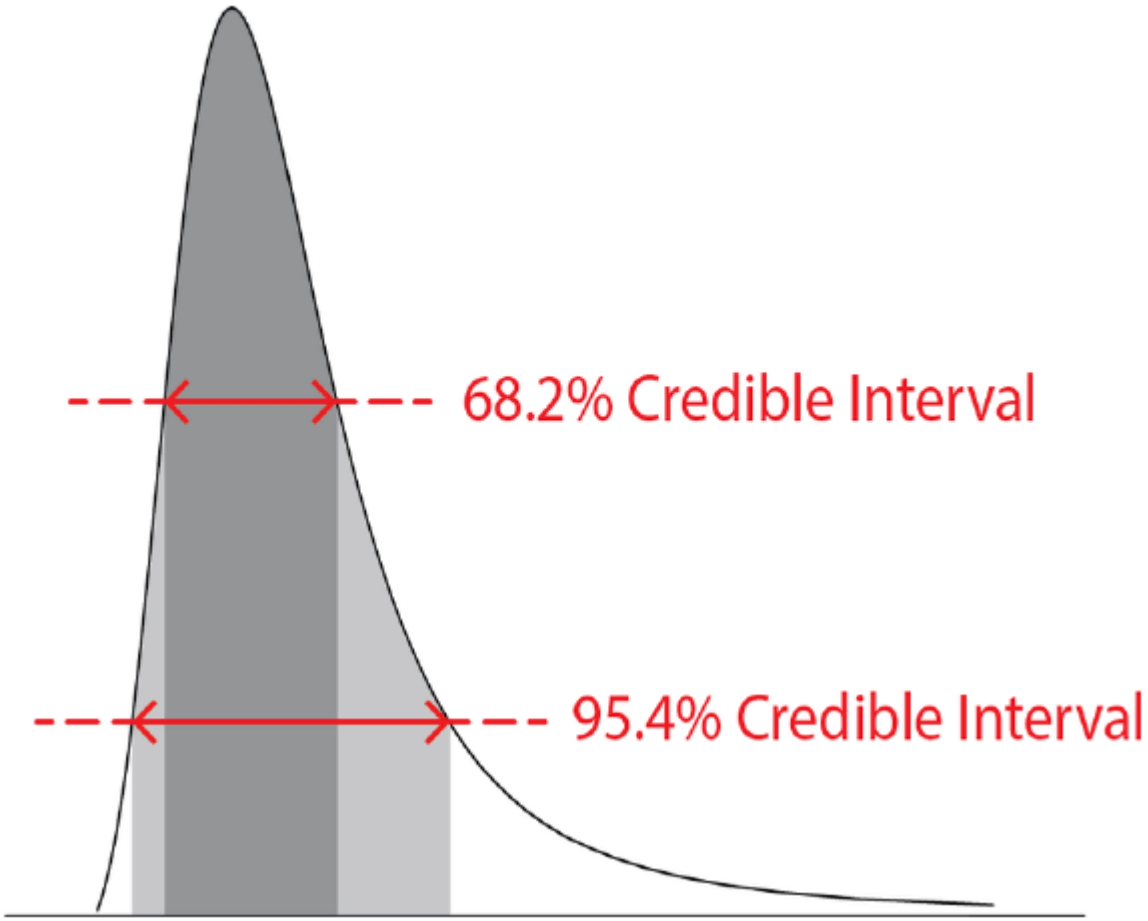
- We can also define the posterior median $\hat{\theta}(x)$ as

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

- The Bayesian equivalent of the confidence interval is the **credible interval**.

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

- This time, it really means the probability that *k* belongs to $[t_l, t_u]$ because the parameter *k* is a random variable in a Bayesian framework.


68.2% Credible Interval
95.4% Credible Interval

- In the previous section, we saw the definition of the *p-value* that many frequentists use to evaluate a model.

- In a Bayesian framework, the **Bayes' factor** plays a similar role.

- Unlike the *p-value,* its goal is to directly compare two models M1 and M2.

- Bayes' theorem implies that

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

- The Bayes' factor B is defined as $\frac{p(M_2|D)}{p(M_1|D)} = B \frac{p(M_2)}{p(M_1)}$

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

- $p(D|M_i)$ is the probability (or probability density) of observing the data if model M_i is true.

- $f(k_i|M_i)$ is the prior probability distribution of k_i given that M_i is the true model.

- $L(D|k_i, M_i)$ is the likelihood, that is the probability density of observing the data D given that model M_i is true and that it takes on the value k_i .

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

- The Bayes' factor should ideally measure how strongly the data D support one model against the other.

- B = 10 ideally means that the data strongly speaks in favour of M_2.

- B = 0.10 ideally means that the data strongly supports M_1.

- One limitation of the Bayes' factor is that we sometimes have no idea about realistic prior distributions for $f(k_1|M_1)$ and $f(k_2|M_2)$.

4.3 Precise Bayesian probabilities as ideal betting behaviour.

- We saw that Bayesians consider probabilities to be rational degrees of belief.
- What on earth are those rational degrees of belief?
- One widespread answer is that they're **betting quotients** of rational agents.

- This is known as the **betting interpretation of probabilities**.

- Let's consider a bet about an event A (such as your favourite horse winning the race or yours becoming the best student of the whole school!)

- If you buy the bet, you pay an amount of money called the $stake \in [0;1]$.
  If A occurs, you receive 1 € and nothing otherwise.
  $\rightarrow$ $profit_b = X(A) - stake$.

- If you sell the bet, you receive $stake \in [0;1]$ €.
  If A occurs, you give 1 € and nothing otherwise.
  $profit_v = stake - X(A)$.

Fundamental axiom of decision theory

- "*A player will always accept a bet if its expected value is equal to or greater than 0*".

- We have $profit_b = X(A) - stake$ and $profit_v = stake - X(A)$.

- Since *X(A)* is a **Bernoulli variable**, we have $\begin{array}{l} E(profit_b) = p(A) - stake \\ E(profit_v) = stake - P(A) \end{array}$

- A player will always accept to buy a bet for $stake = p(A) - \epsilon$, $\epsilon \geq 0$.

- He or she will also always accept to sell the bet for $stake = p(A) + \epsilon$.

- He'll be ready to **both buy and sell** the bet for $stake = p(A)$.
  This leads to the definition of the Bayesian probability *p(A)*.

p(A) *is the quotient* ***q*** *such that a player would always be ready to pay* ***qS*** *to receive* ***S*** *if* ***A*** *occurs.* $S \in \mathbb{R}$ *can be both positive (buyer) and negative (vendor).*

- An important notion in Bayesian philosophy is that of irrational beliefs.

- A combination of beliefs is irrational if it can lead to a **sure loss of** money.

- Suppose that someone believes that $p(A) = 0.\bar{7}$ and $p(\overline{A}) = 0.6$ at the same time.

- That person would be ready to bet $0.7 * 100 = 70$ € on A to receive 100 € and $0.6 * 10 = 60$ € on $\overline{\text{A}}$ to receive 100 € at the same time.
- The total profit of the combination of the two bets is the random variable $profit = 100X(A) - 70 + 100X(\overline{A}) - 60$.

- Since $X(A) + X(\overline{A}) = 1$, we always have $profit = 100 - 70 - 60 = -30$.

- Such a bet would thus <u>always</u> result in a financial loss for the player.
  → **sure loss**.

- It can be shown that an ensemble of beliefs must respect the **Kolmogorov axioms** to avoid a sure loss.

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

4.4 Imprecise Bayesian probabilities

- Until now, we've assumed all along that probabilities are **single-valued**.
- However, this can be unrealistic in many situations.
- Suppose we're interested in the probability that a chemical plant explodes.

- If we've very good information, we might know that *p(explosion)* = 0.454.
- In many cases, we cannot be so precise.
- Instead, we might only know that $p(explosion) \in [0.2; 0.7]$.
- Using the betting interpretation of probability, it means we'd always be willing to buy the bet for $stake \leq 0.2$ € and sell the bet for $stake \geq 0.7$ € so that we'd get or give 1 € if the event occurs, respectively.
- We're undecided for $stake \in]0.2; 0.7[$. This corresponds to our uncertainty.

- Imprecise Bayesian probabilities follow their own axioms.

$$\forall A \in \sigma, \quad 0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$$

$$\text{If } \underline{P}(A) = \overline{P}(A) = P(A) : \text{precise probability}$$

$$\text{If } \underline{P}(A) = 0 \text{ and } \overline{P}(A) = 1 : \text{ complete ignorance about A}$$

$$\text{For any disjoint events A and B:}$$

$$\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B) \text{ and } \overline{P}(A \cup B) \geq \overline{P}(A) + \overline{P}(B)$$

$$\underline{P}(\overline{A}) = 1 - \overline{P}(A)$$

Rule for updating imprecise probabilities

- Imprecise probabilities are an **ensemble of precise probabilities**.

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

- We update it by applying Bayes' theorem to **all** members of the ensemble.

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

- We obtain an ensemble of posterior probabilities.

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

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

Example

- Let's reconsider the case of the medical test.

- We still have $p(Test|Sick) = 95\ \%$ and $p(Test|\overline{Sick}) = 2\ \%$.

- This time, however, we only know that $p(\textit{cancer}) \in [0.05\,;0.20]$. (due to the bad quality of the samples).

- Applying the updating rule, we find that

$$\underline{P}(Sick|Test) = 0.02321 \text{ and } \overline{P}(Sick|Test) = 0.0869$$

Complete ignorance

- Knowing absolutely nothing about an event A means that $p(A) \in [0;1]$.
- However, we cannot directly update this interval using the formula we saw.
- If 0 belongs to the ensemble of prior probabilities, 0 will always belong to the ensemble of posterior probabilities (exercise).
- If 1 belongs to the prior probabilities, 1 will always belong to the posterior probabilities (exercise).

- Updating [0;1] thus always leads to [0;1] as if we had learned absolutely nothing through the data.
  There are two possible strategies to avoid this problem.

1) We only keep [0;1] to suspend belief **while we have no data**.
When the first data arrive, we use them to replace [0;1] rather than update it.

Example: we know nothing about a coin.
So long as we haven't tossed it, we keep the belief that $p(heads) \in [0;1]$.

If after having tossed the coin 10 times, it landed 4 times on heads, we can consider that $p(heads) \in [0.40;0.60]$.

2) We use a **near-ignorance prior** $[\epsilon;1-\epsilon]$, $\epsilon \geq 0$ from the very beginning.

Example: [0.001;0.999].

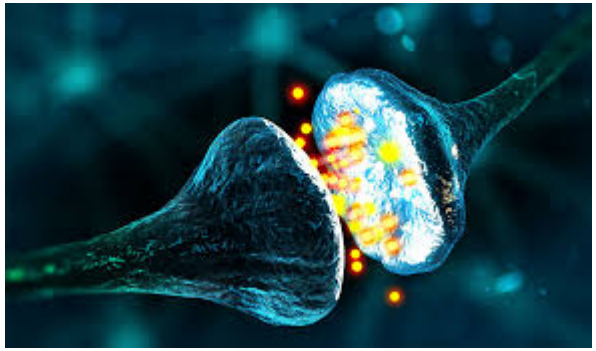
5) Why should we choose one approach rather than another?

5.1) Use of prior information

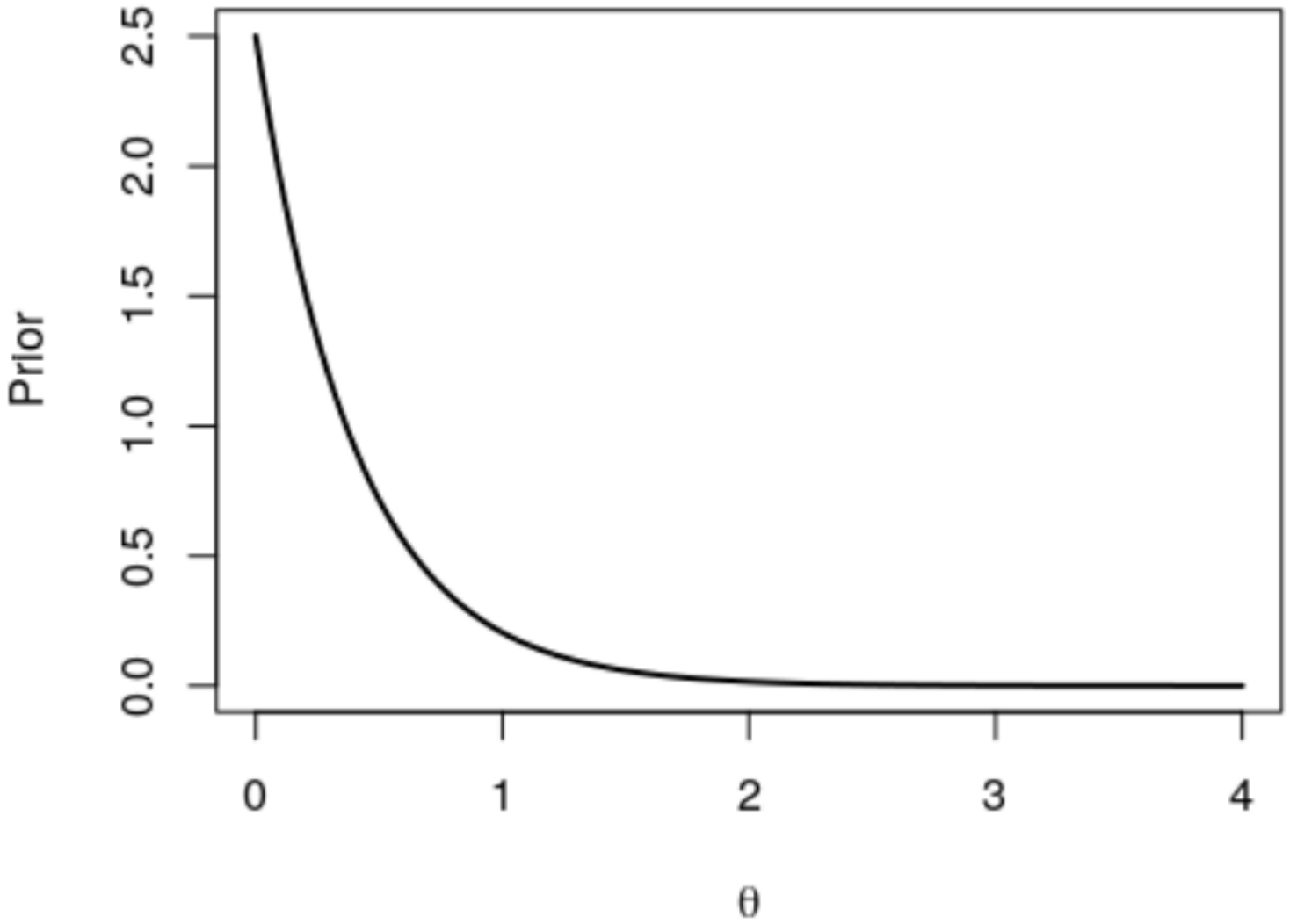
- Frequentists only use the data D they determined to derive conclusions about a parameter value.

- It often happens, however, that we also have prior information at our disposal.
- Bayes' theorem allows us to **combine that prior information with the new data**.

- Example: firing rate of a neuron.
- We want to estimate the firing rate θ (Hz) of a neuron.



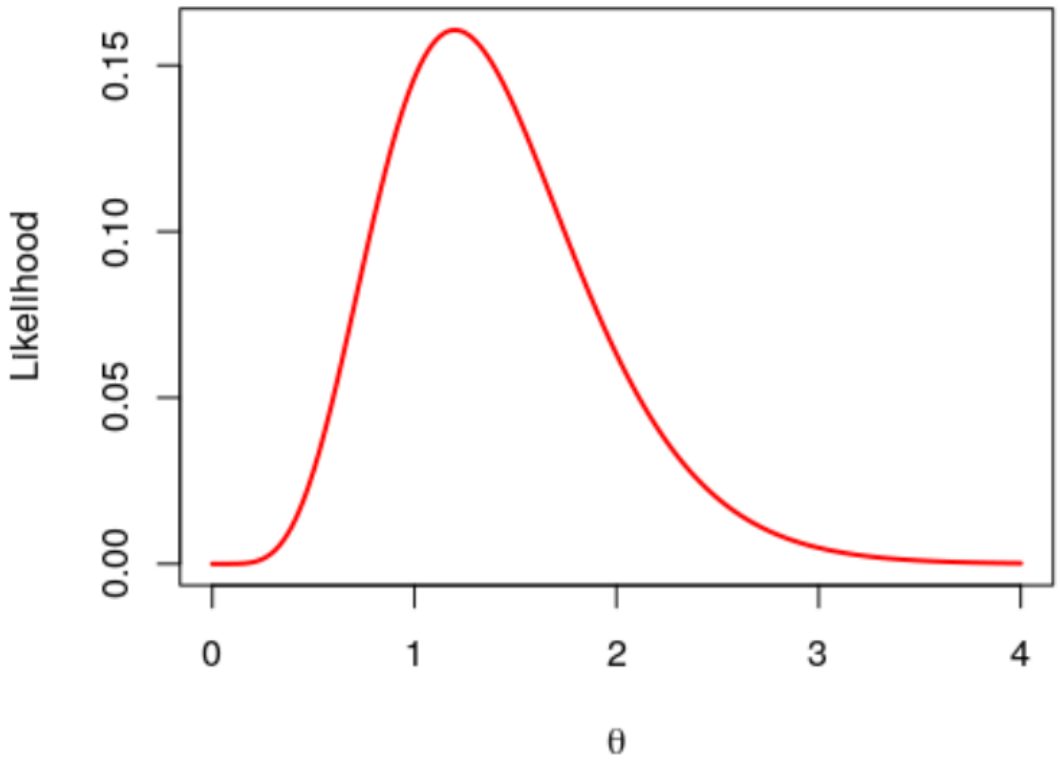
- The true value is $\theta = 0.76$ Hz but we don't know it.
- However, we know through the study of similar neurons that θ follows an exponential distribution whose inverse mean is given by $\tau = \frac{1}{0.4} = 2.5$ s
- This leads to the definition of the prior density.


Prior
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3
4
θ

- We measure the number y of activation events during the period t.
- The likelihood follows a Poisson distribution

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

- We measure $y = 6$ for $t = 5$ s.



- The Maximum Likelihood Estimator leads to $\hat{\theta}(y) = \frac{y}{t}$.

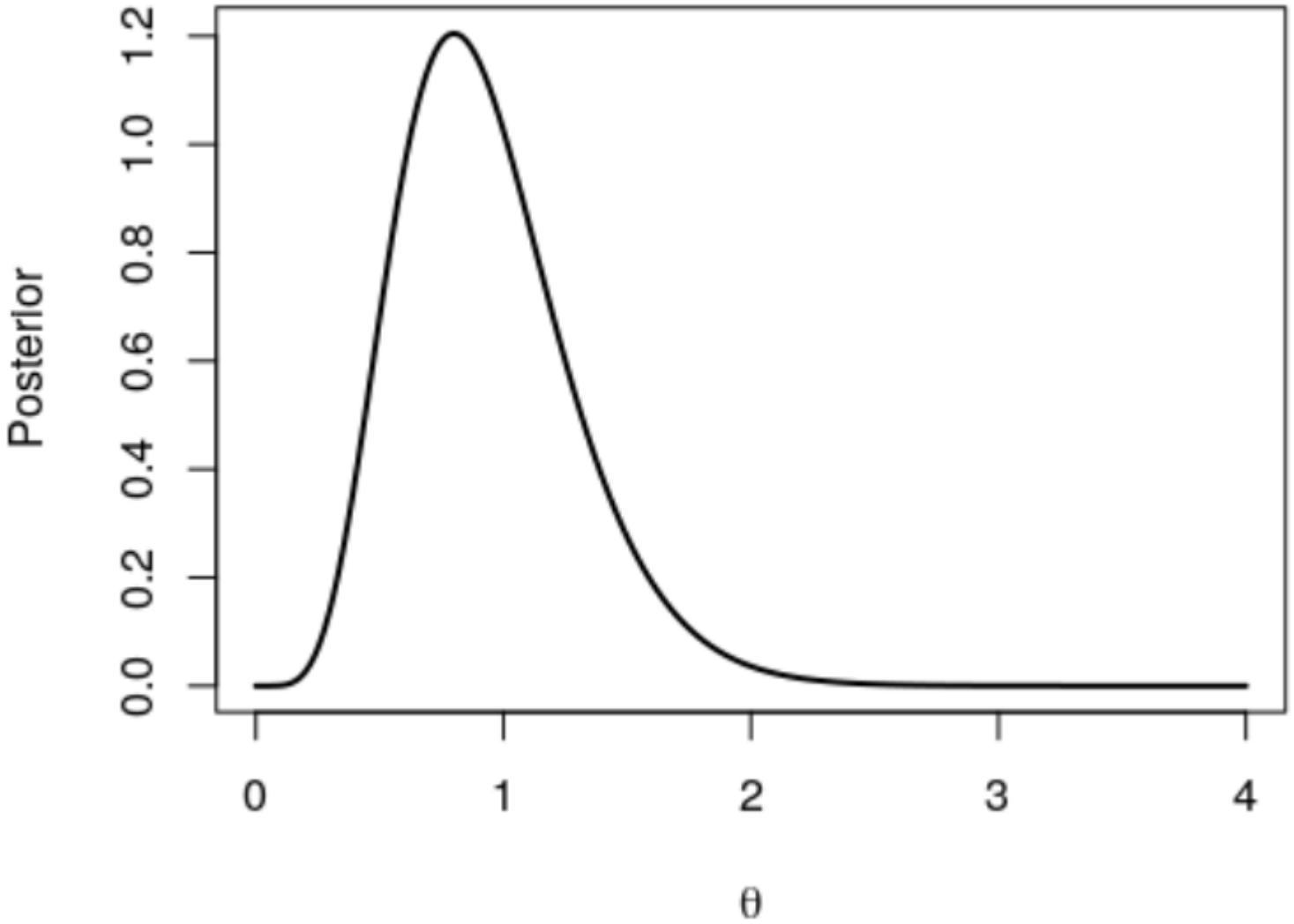
- Using the frequentist method, the estimated value would thus be $\hat{\theta}(y) = 1.2$ Hz.

- The Bayesian approach consists of combining the prior and the likelihood.

- $$f(\theta|x) = \frac{f(\theta)L(x|\theta)}{f(x)} = \frac{f(\theta)L(x|\theta)}{\int_{\theta' \in \Theta} f(\theta')L(x|\theta')d\theta'}$$ (Bayes' theorem)

- We can prove analytically that the posterior distribution is given by

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


Posterior
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0
1
2
3
4
θ

- Maximum likelihood estimate $\hat{\theta}(y) = \frac{y}{t} = 1.2 \text{ Hz}$ (frequentism).

- Maximum posterior estimate $\hat{\theta_p} = \frac{y}{t+\tau} = 0.828 \text{ Hz}$ (Bayesianism).

- True value $\theta = 0.76 \text{ Hz}$.

- We see that the incorporation of prior information leads to a more realistic estimate for θ .

5.2) The strong likelihood principle

- Suppose we're interested in the probability θ that a coin is symmetrical.
- We want to compare two hypotheses: H_0: $\theta = 0.5$ and H_1: $\theta > 0.5$.
- We toss the coin $n = 12$ times. The coin lands $k = 9$ times on heads.
- We want to know if the data allow us to reject H_0, by using the **p-value**.
- There are two ways to approach that problem.

1) Binomial distribution

- X is the number of times the coin landed on heads during the 12 trials.

$\rightarrow X \sim B(12, \theta)$ and $L_1(\theta|X = 9) = \binom{12}{9}\theta^9(1-\theta)^3$.

- The p-value is defined as

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

2) Negative binomial distribution

- We can also view the experiment in the following manner:
  "*We toss the coin **until** we get 3 tails*".

- Let Y be the number of heads before the end of the experiment.

- The likelihood is given by $$L_2(\theta|Y=9) = \binom{11}{9}\theta^9(1-\theta)^3 \quad .$$

- The p-value is

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

- We thus both have $p - value_1 = 0.075$ and $p - value_2 = 0.0325$.

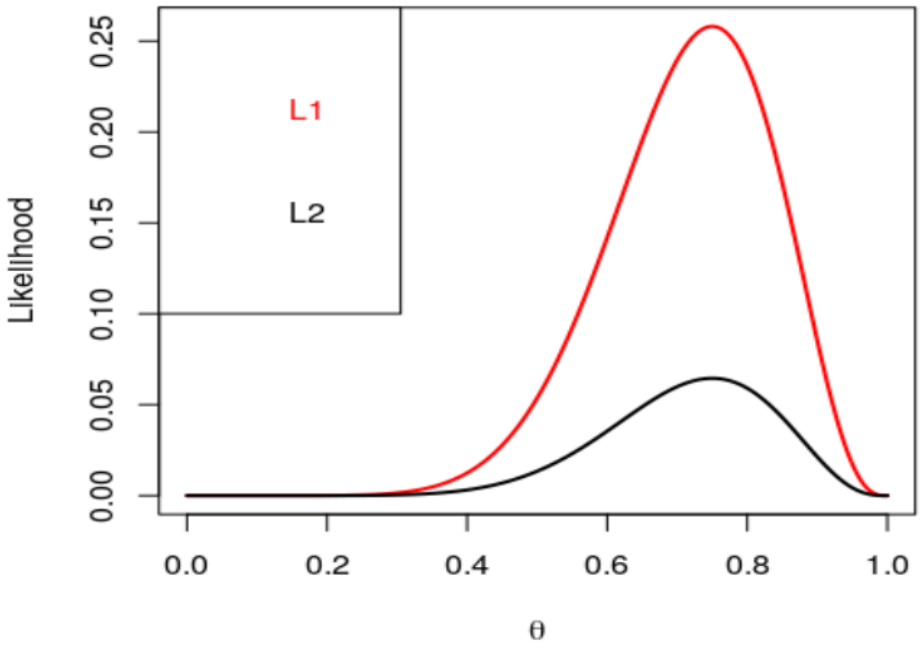
- We're now facing a **paradox**. Both descriptions of the problem are equally valid.

- Strong likelihood principle: all statistical information derived from an experiment is entirely contained within the likelihood function.

- If a likelihood function is proportional to another, we should draw exactly the same conclusions from both situations.
- This is clearly the case here:

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

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



- However, using the *p-value* at a level of confidence of 95% leads us to conflicting conclusions.

- $p - value_1 = 0.075$ leads us to keep $H_0 (\ \theta\ = 0.5)$.

- $p - value_2 = 0.0325$ leads us to accept H_1 (θ > 0.5).

- Let us now use the Bayesian methodology to deal with this problem.

- We choose to describe our prior knowledge through a **beta prior**:

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

- We choose a = 3 and b = 7 because we know that most coins are biased towards θ = 0.3.

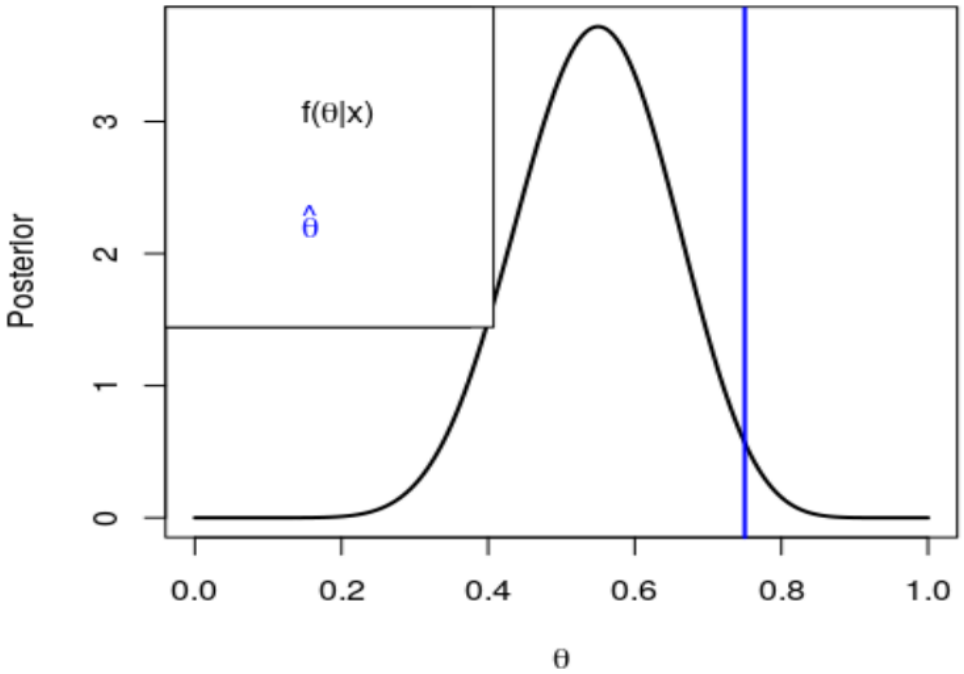
- If we use the same prior, if two likelihoods are proportional to each other, the posterior will be the same.

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

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

- Frequentism violates the strong likelihood principle.

- Bayesianism does not violate the strong likelihood principle.



Remark: the maximum posterior estimate is far from the Maximum Likelihood Estimate because of the incorporation of our prior knowledge.

5.3) The principle of indifference

- There are many situations where we have no information that would help us choose a prior distribution.

- In such a case, many Bayesians use the **Principle Of Indifference** (POI).

- If we know absolutely nothing about the hypotheses H_1, H_2,…, H_n, we should assign them the same prior probability.

- Consequently, $p(H_1)=p(H_2)=\ldots p(H_n)=\frac{1}{n}$.

- If a parameter is continuous, the principle of indifference leads to a uniform prior which is also called *flat prior*.

- If we only know that a parameter *k* belongs to $[0.1;1000]$, this leads to the following prior:

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

5.4) Confusing ignorance with knowledge

- The problem of the principle of indifference is that it leads one to mistake ignorance for knowledge.
- Let us consider the following situation.
- We're going to toss **three coins** and want to choose three prior probabilities $p_0(heads)$.

- Our state of knowledge is different in each case.

- **Coin A**: we know through precise physical measurements that it's symmetrical.

- **Coin B**: we tossed it one million times and saw that the relative frequency oscillates closely around 50%.

- **Coin C**: We know **absolutely nothing** about that coin.

- Applying the principle of indifference leads us to the following conclusion:

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 characterised it physically in great detail.

- This is truly **absurd**: “If we know absolutely nothing about the coin, we cannot know it is perfectly symmetrical”.

- If we use imprecise probabilities, we can make a distinction between the three coins.

- Coin A and B: $p(heads) = 0.5$ (Knowledge)

- Coin C: $p(heads) \in [0;1]$ (Complete ignorance).

5.5) Could second order probabilities avoid this problem?

- Many classical Bayesians would try to avoid the problem through the use of second order probabilities.

- Instead of only considering the binary event $E \in \{Heads, Tails\}$ as a random variable, we also consider the Bernoulli parameter $\theta = p(heads) \in [0;1]$ as a random variable.

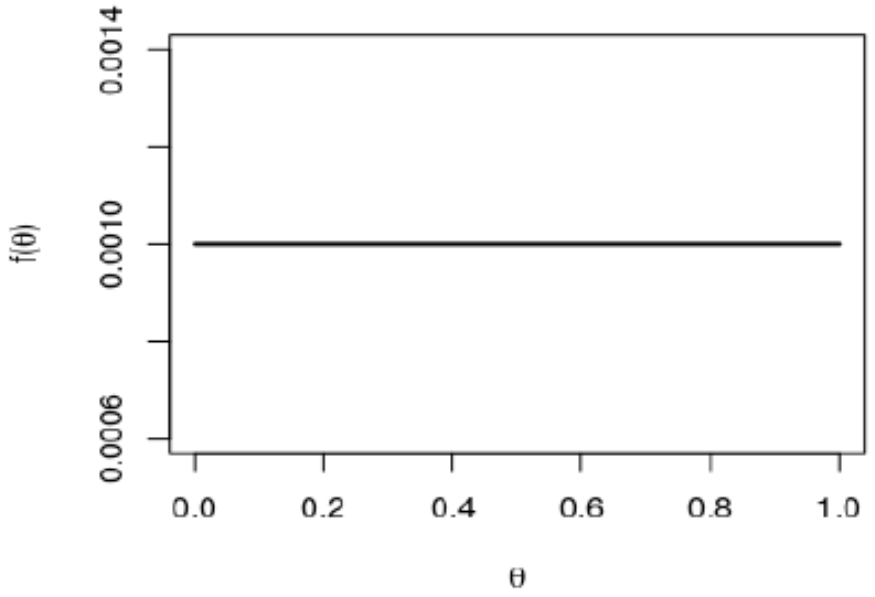
- Applying the principle of indifference to θ leads to a **uniform (flat) prior** equal to 1 on [0;1] and 0 otherwise.

- Let us now suppose that we tossed the coin $n = 2k$ times and that it landed on heads k times.
- We know that the likelihood is given by

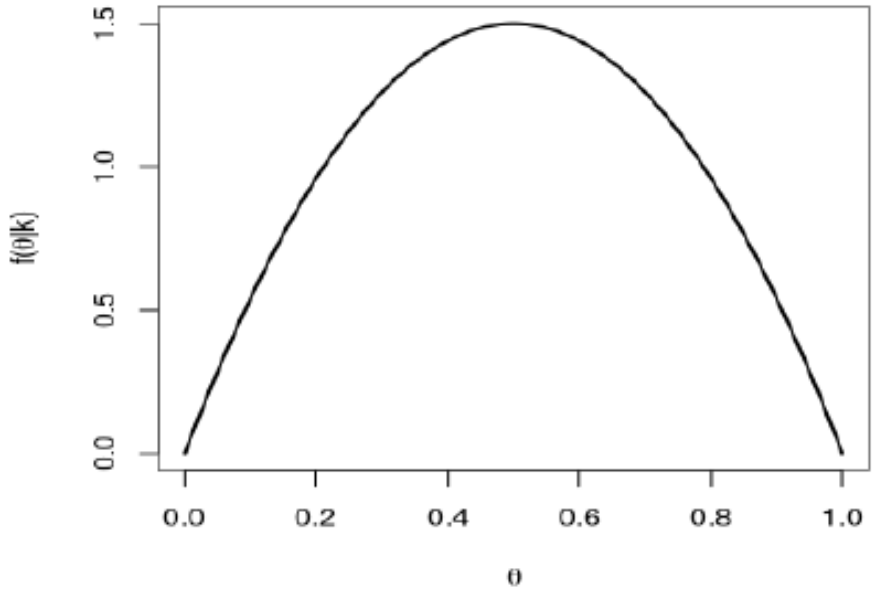
$$L_1(X = k|\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}.$$

- If we use a uniform prior, the posterior is given by $Be_{k+1,n-k+1}(\theta)$ so that $f(\theta|X = k) = \dfrac{\theta^k(1-\theta)^{n-k}}{B_{k+1,n-k+1}}$ with $B_{k+1,n-k+1} = \displaystyle\int_{x=0}^{1} \theta^k(1-\theta)^{n-k}d\theta$.

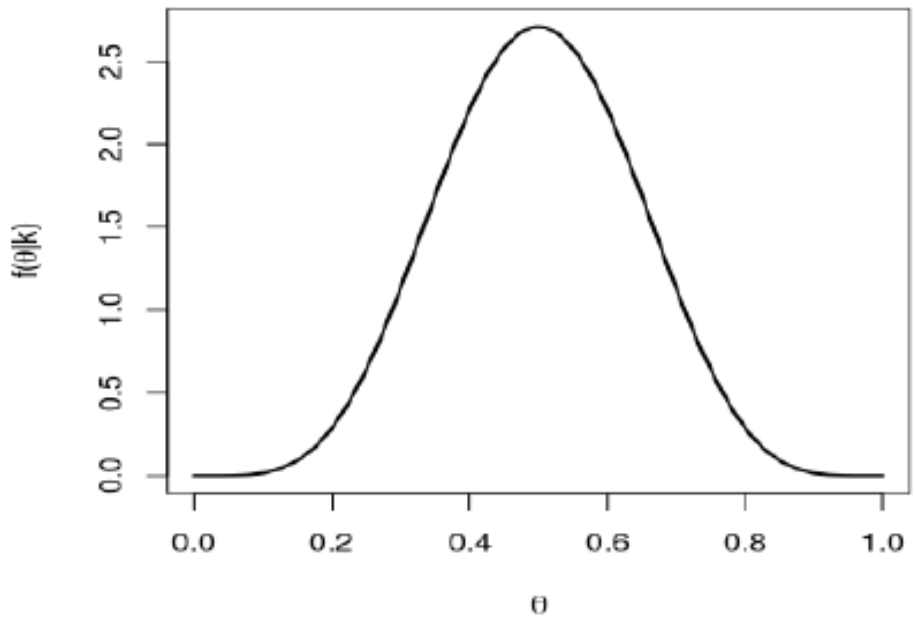
- Let us now plot the posterior density for different values of n.



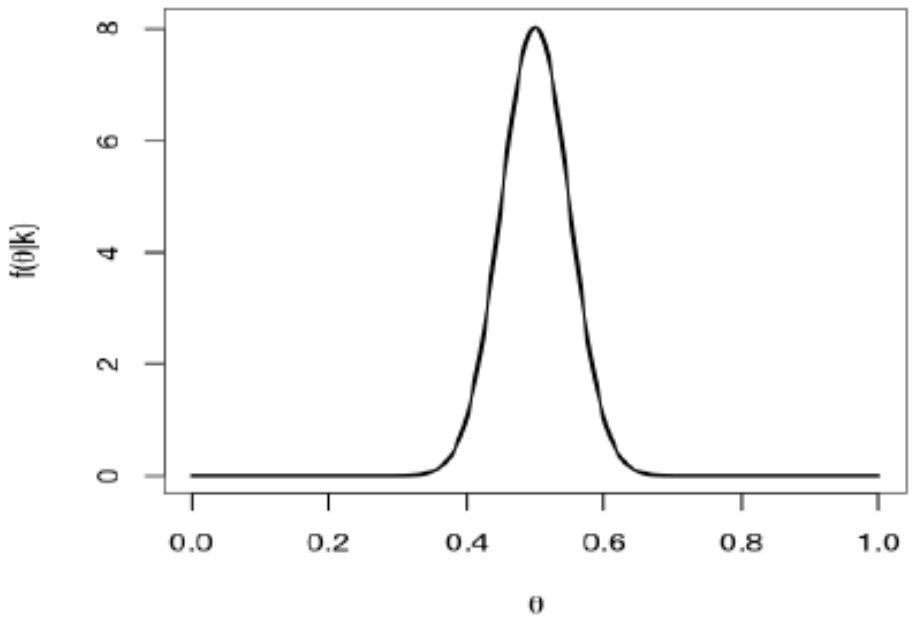
(a) Prior



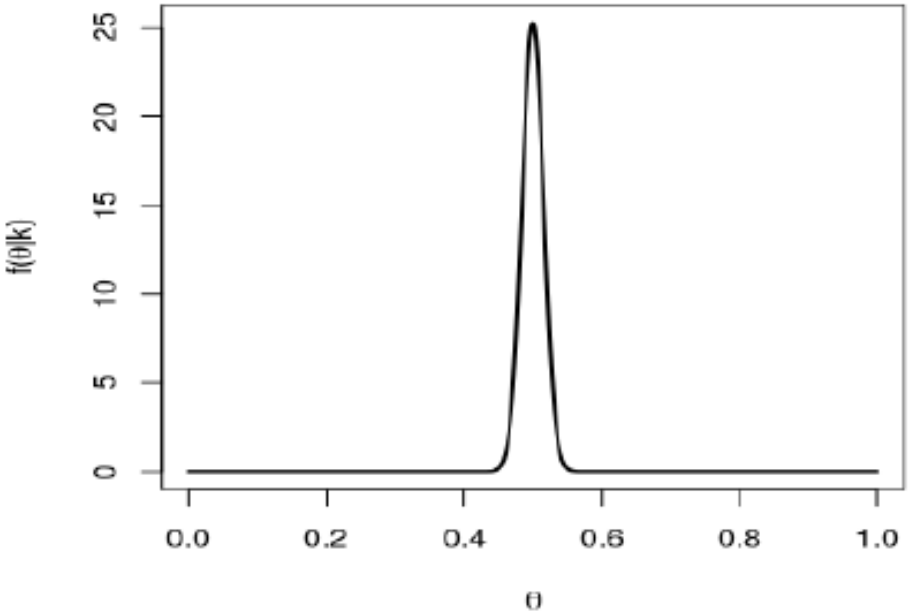
(b) Posterior: $n = 2$



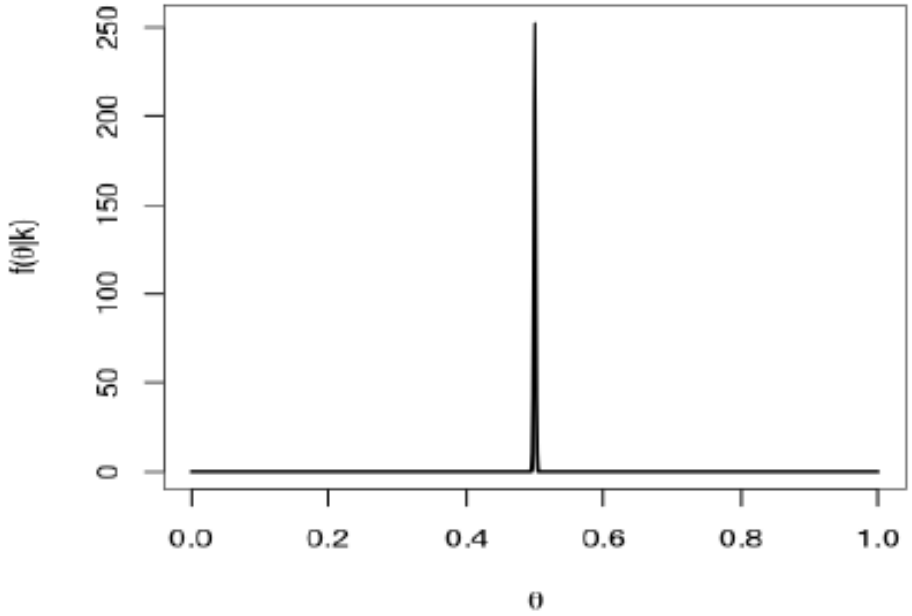
(c) Posterior: $n = 10$



(d) Posterior: $n = 100$



(e) Posterior: $n = 1000$



(f) Posterior: $n = 100000$

We see that while *p(heads)* = 0.5 in all cases, the probability density of θ is very different.

- Complete ignorance corresponds to a flat (uniform) prior.
- The higher n is, the more precise the prior becomes.
- For $n \to +\infty$, it tends towards a singularity at $\theta = 0.5$.
- A classical (precise) Bayesian would argue that we can distinguish knowledge from ignorance in this way.
- However, this only pushes back the problem.
- Let's consider two other 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 $p(H_1) = \int_{\theta=0}^{0.01} 1 d\theta = 0.01$ and $p(H_2) = \int_{\theta=0.40}^{0.60} 1 d\theta = 0.20$.

- Applying the principle of indifference leads thus 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".*

- It's the same absurd creation of knowledge out of ignorance.

5.6) Arbitrariness of the uniform prior

- Let's consider now a (very simple) problem of chemical kinetics.
- We're interested in the reaction R + R $\rightarrow$ P_2 with $\frac{d[R]}{dt} = -2k[R]^2$.
- [R] $(\mathrm{mol/m^3})$ is the concentration of the reactant.
- k $(\mathrm{mol.cm^{-3}.s^{-1}})$ is the rate coefficient we want to estimate.
- We only know a **lower and an upper bound** for $k \in [5\mathrm{E}{+}04; 5\mathrm{E}{+}09]$.

- Applying the principle of indifference (POI) to k leads to

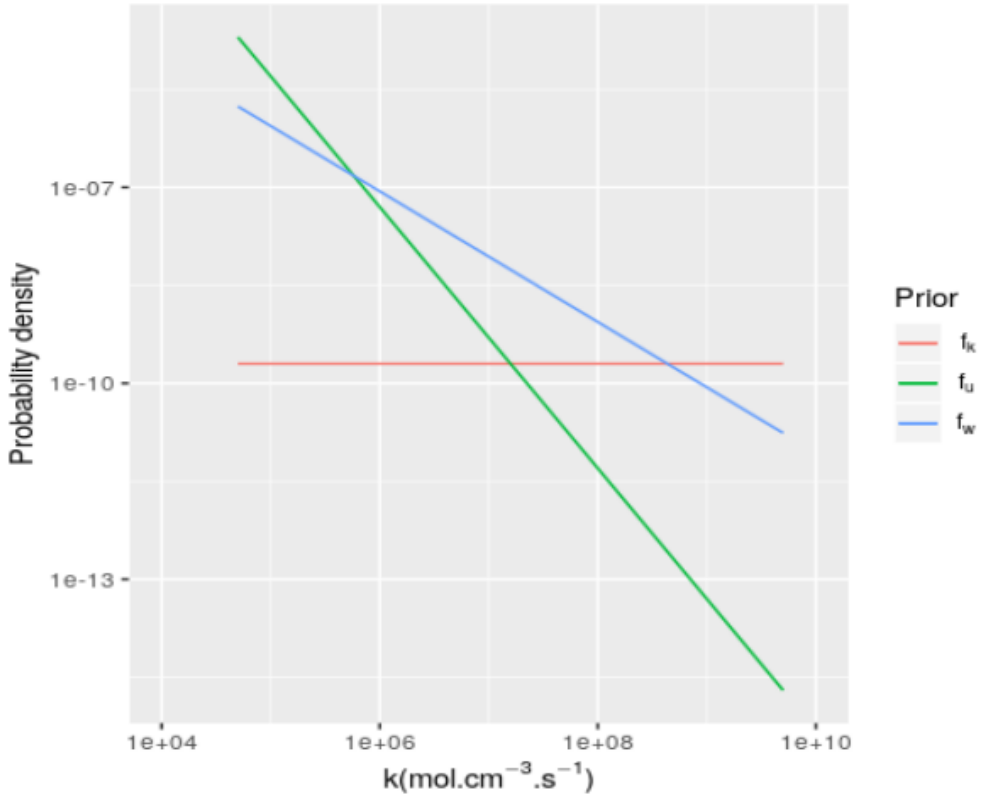
$$f_k(k) = \frac{1}{5\text{E}{+}09 - 5\text{E}{+}04} \approx 2\text{E}{-}10 \cdot$$

- If we are ignorant about $k \in [5\text{E}{+}04; 5\text{E}{+}09]$, we're also ignorant about $u = \frac{1}{k} \in [2\text{E-}10\,;2\text{E-}05]$ and $w = \log_{10}(k) \in [\log_{10}(5\text{E+}04)\,;\log_{10}(5\text{E+}04)]$.

- It would be equally justifiable to apply the POI to u and w.

- However, if we do that, the results we get aren't consistent.

$$f_k(k) \approx 2\text{E}{-10} \qquad f_u(k) = 5\text{E}{+04}k^{-2} \qquad f_w(k) = \frac{0.20}{ln(10)k}$$



- Let's suppose we're interested in the probability that $k \in [5\text{E+}04\,;5\text{E+}06]$.
- First prior: $p(k \in [5\text{E+}04; 5\text{E+}06]) = 2\text{E}{-}10(5\text{E+}06 - 5\text{E+}04) = 0.00099$
- Second prior: $p(k \in [5\text{E+}04; 5\text{E+}06]) = 5\text{E+}04(\frac{1}{5\text{E+}04} - \frac{1}{5\text{E+}06}) = 0.99$.
- Third prior: $p(k \in [5\text{E+}04; 5\text{E+}06]) = 0.20(log_{10}(5\text{E+}06) - log_{10}(5\text{E+}04)) = 0.4$.
- The value depends thus strongly on the choice of the prior.
- During the **first practical work** (TP1), we'll see how the choice of the prior can impact the evaluation and comparison of models based on experimental data.

5.7) Jeffreys' prior

- The Jeffreys' prior is independent on parametrisation.
- Let M be a model predicting a random vector $X \in \mathbb{R}^n$ as a function of θ.
- $x = (x_1, ..., x_n)$ is a realisation of $X = (X_1, ...X_n)$.
- $L(x|\theta)$ is the likelihood of the model. $\theta^* = \theta^*(x)$: MLE estimate.
- The observed Fisher information is defined by $I_{\theta^*}(x) = -\frac{d^2}{d\theta^2} ln(L(x|\theta))(\theta^*)$.
- It measures the information contained in the observed data x.

- The Fisher information is the **expected value** of the observed Fisher information.

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

- It is the information we'd expect from the random vector $X = (X_1, ...X_n)$.

- The Jeffreys' prior is the probability density proportional to $\sqrt{det(I(\theta))}$ so that $p(\theta) \propto \sqrt{det(I(\theta))}$.

- Since the Fisher information is independent on the parametrisation, the Jeffreys prior is also independent on it.

- If we considered $\theta_2 = \frac{1}{\theta}$ or $\theta_3 = \ln(\theta)$, we'd get the same distribution.
- The Jeffreys' prior avoids the problem of arbitrariness.
- Unfortunately, it isn't a legitimate representation of ignorance.
- If we consider the problem of coin tossing (with a binomial likelihood), we find that $p(\theta) = beta(0.5, 0.5)$.
- This is an extremely biased prior that isn't neutral at all.

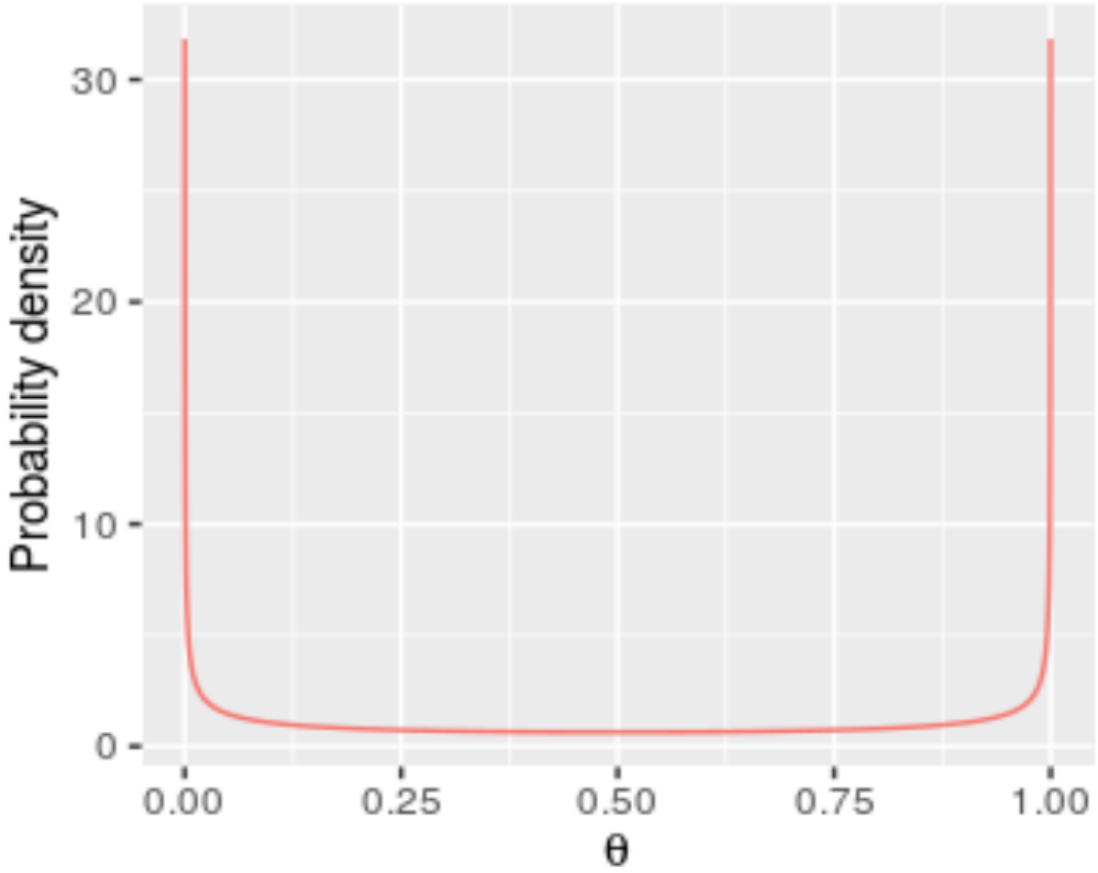


Figure 2: Jeffreys' prior for coin tossing

5.8) Washing out the priors

- We've seen so far that a single prior density cannot represent ignorance.
- The good news is that the effect of the prior becomes smaller and smaller as we add more data.
- This effect is called "***washing out the priors***".
- Illustration: let's suppose we're in a casino.
- There are three types of coin tossing machines:

Type 1: if we randomly choose a machine, the probability that the coin lands on heads is **uniform**.

This means that $\theta \approx 0.32$, $\theta \approx 0.15$ and $\theta \approx 0.67$ etc. happen with the same frequency.

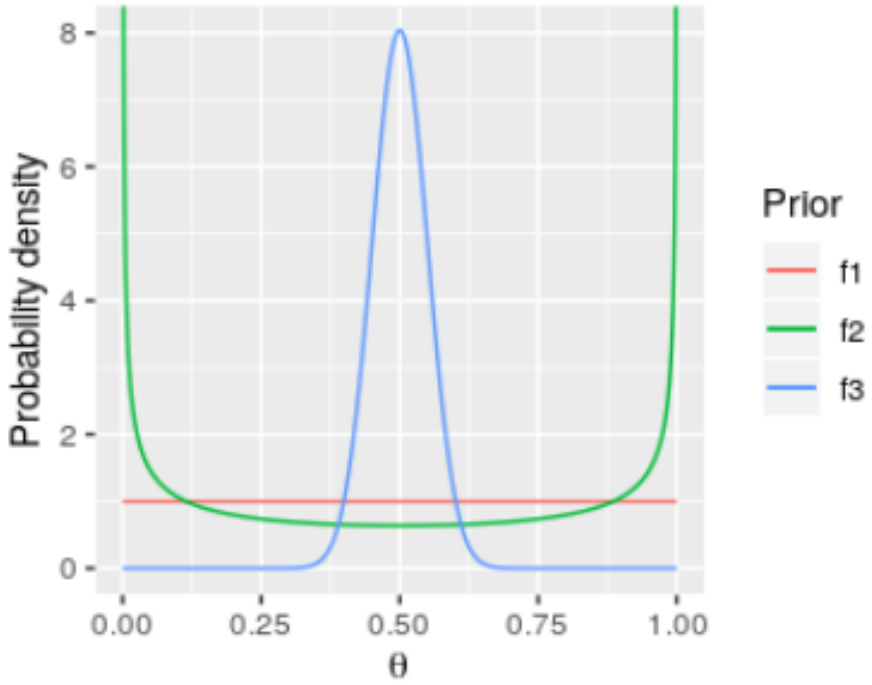
Type 2: θ is **strongly biased** towards either heads ($\theta \approx 1$) or tails ($\theta \approx 0$) .

Type 3: the coins are **relatively fair**: θ isn't too far from 0.50.

We aren't told the category of the machine we're gonna use.

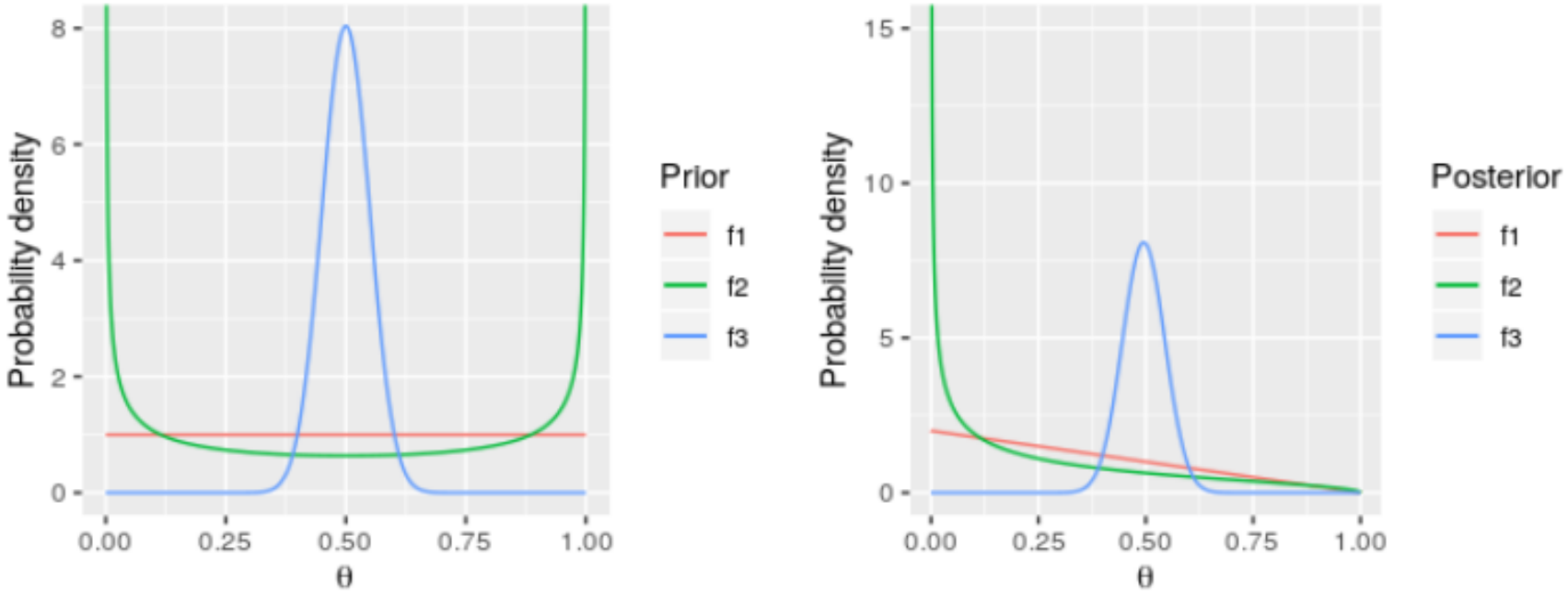
We choose to represent the situation through three priors.

They correspond to the three possible machines.

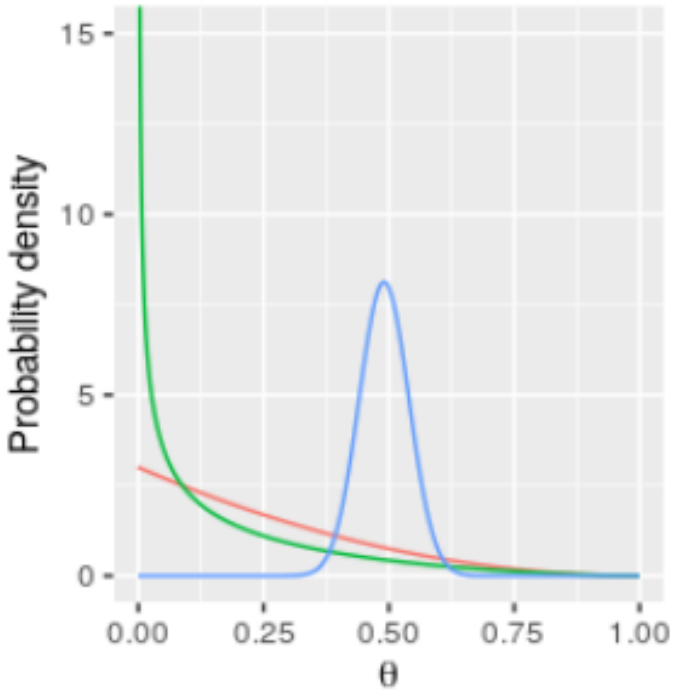


Question: which prior corresponds to which machine?

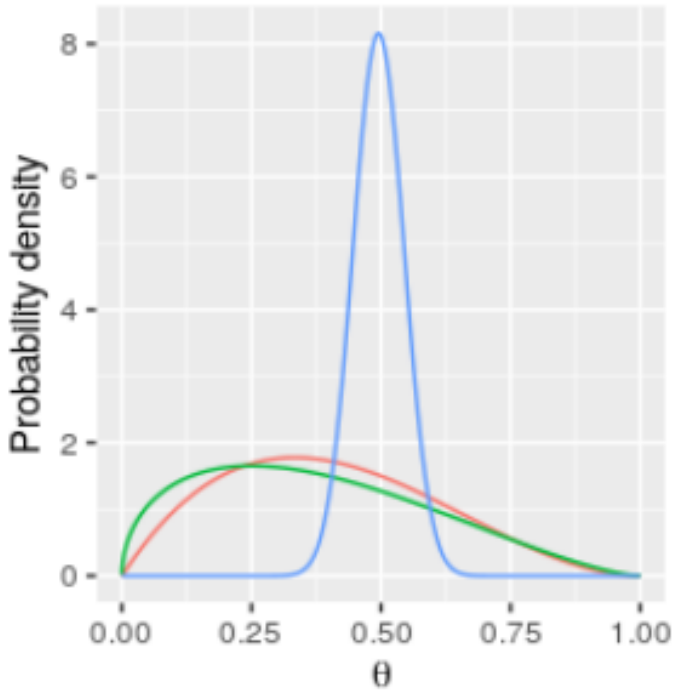
We look then at how fast the three corresponding posteriors converge.



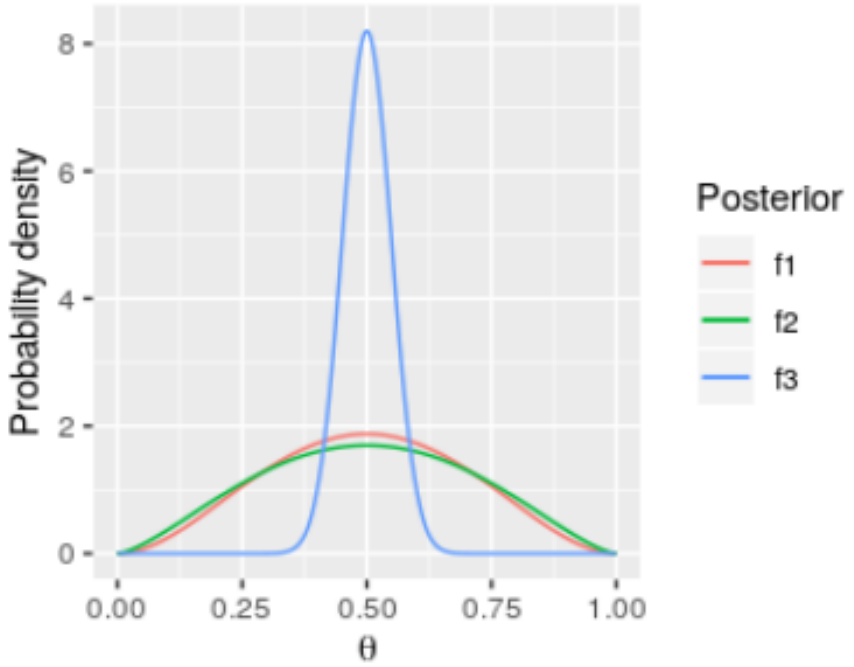
(a) n = 0 (b) n = 1


Probability density
15
10
5
0
0.00
0.25
0.50
0.75
1.00
θ
Posterior
f1
f2
f3

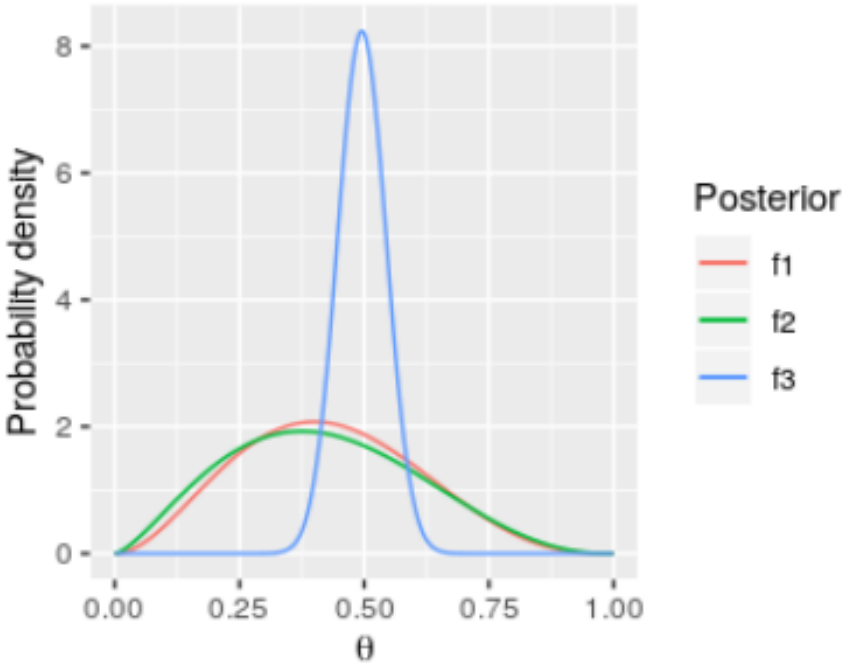
(c) n = 2


Probability density
8
6
4
2
0
0.00
0.25
0.50
0.75
1.00
θ
Posterior
f1
f2
f3

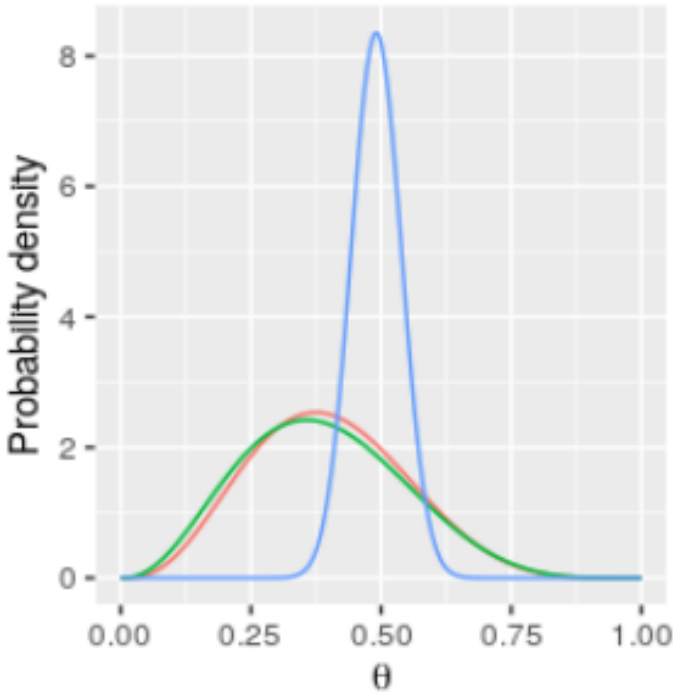
(d) n = 3


Probability density
8
6
4
2
0
0.00
0.25
0.50
0.75
1.00
θ
Posterior
f1
f2
f3

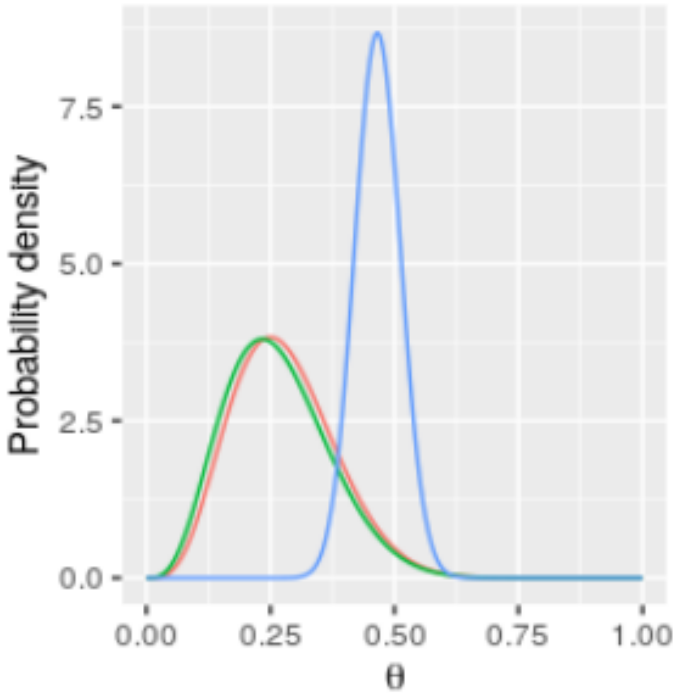
(e) n = 4


Probability density
8
6
4
2
0
0.00
0.25
0.50
0.75
1.00
θ
Posterior
f1
f2
f3

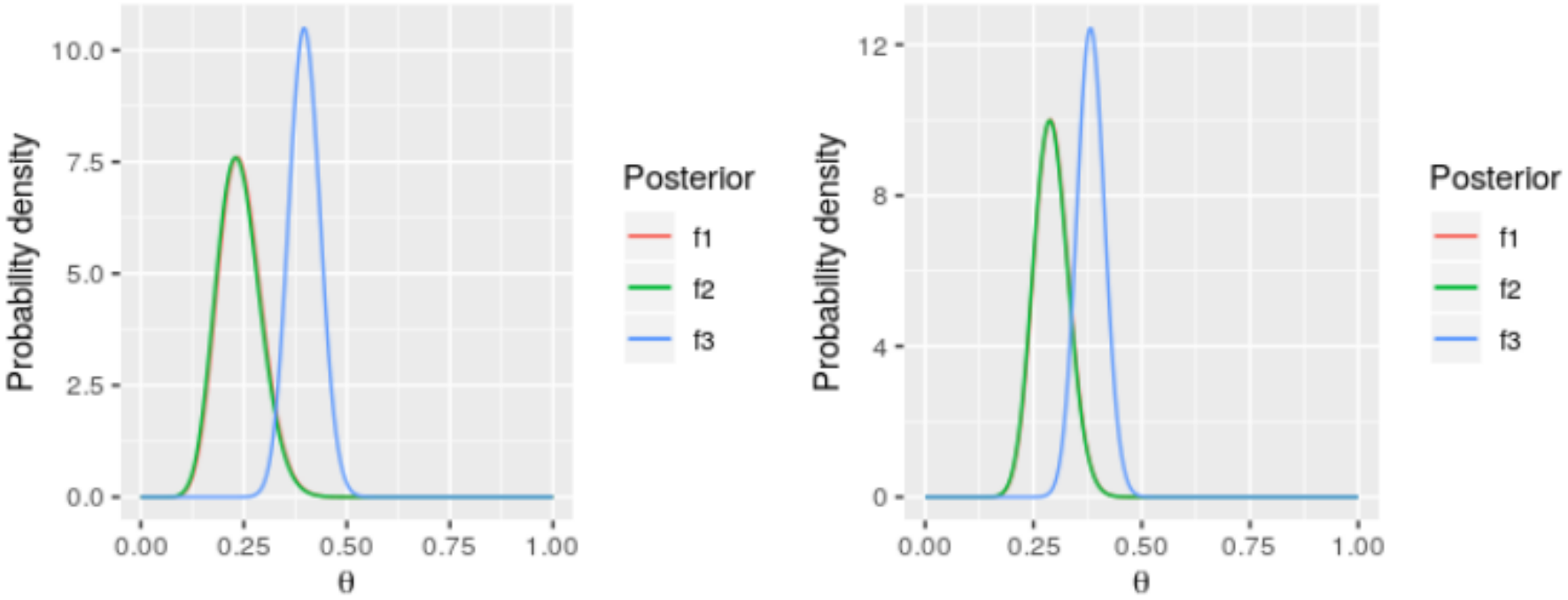
(f) n = 5

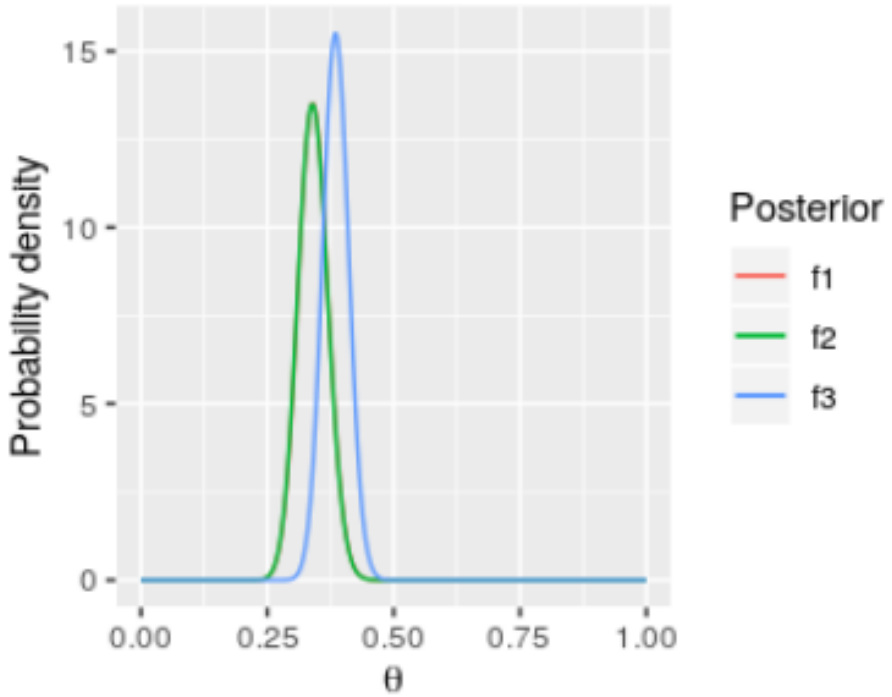

Probability density
8
6
4
2
0
0.00
0.25
0.50
0.75
1.00
θ
Posterior
f1
f2
f3

(a) n = 8

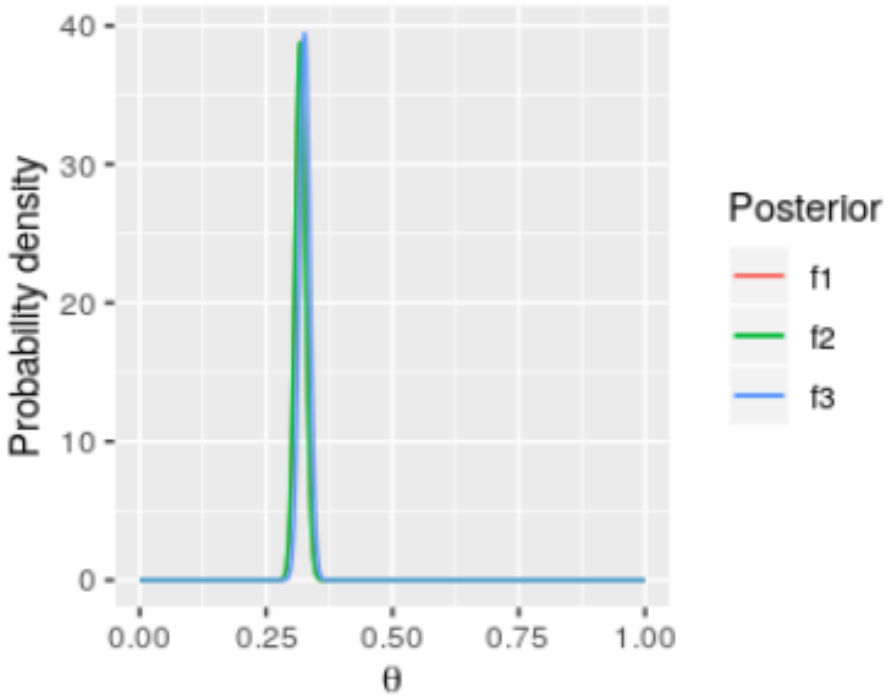

Probability density
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
θ
Posterior
f1
f2
f3

(b) n = 16


Probability density
10.0
7.5
5.0
2.5
0.0
0.00
0.25
0.50
0.75
1.00
θ
Posterior
f1
f2
f3
Probability density
12
8
4
0
0.00
0.25
0.50
0.75
1.00
θ
Posterior
f1
f2
f3



(e) n = 256

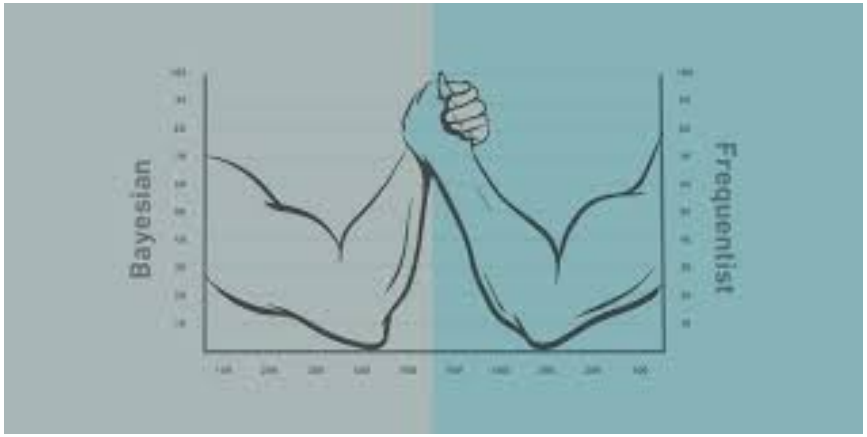


(f) n = 2048

- The convergence towards a unique probability density is guaranteed.
- It can, however, take a lot of time!

6. Conclusion

- Nowadays, probabilistic reasoning plays a vital role in engineering.
- No consensus about the best approach has yet emerged.
- The main division concerns Bayesians versus frequentists.



- The major limitations of frequentism are its inability to include prior knowledge and the inconsistent results it can produce through the *p-values*.

- Precise Bayesianism cannot correctly describe complete ignorance.

- Imprecise Bayesianism is a richer framework. But it demands more time and computational power and a higher level of abstraction.

- Using comparatively both frequentist and (imprecise) Bayesian methods can be useful to gain deeper insights into the nature of a problem.