Introduction à la modélisation statistique bayésienne

Un cours en R, Stan, et brms

Ladislas Nalborczyk (LPC, LNC, CNRS, Aix-Marseille Univ)

Planning

Cours n°01 : Introduction à l'inférence bayésienne

Cours n°02 : Modèle Beta-Binomial

Cours n°03 : Introduction à brms, modèle de régression linéaire

Cours n°04 : Modèle de régression linéaire (suite)

Cours n°05 : Markov Chain Monte Carlo

Cours n°06 : Modèle linéaire généralisé

Cours n°07 : Comparaison de modèles

Cours n°08 : Modèles multi-niveaux

Cours n°09 : Modèles multi-niveaux généralisés

Cours n°10 : Data Hackathon

Rappels

Principes de l'analyse bayésienne :

- On dispose d'un ensemble de données à analyser
- On suppose un modèle génératif défini par un ensemble de paramètres
- On dispose d'une connaissance a priori quant à la valeur de ces paramètres

> The first idea is that Bayesian inference is **reallocation of credibility across possibilities**. The second foundational idea is that the possibilities, over which we allocate credibility, are **parameter values** in meaningful mathematical models (Kruschke, 2015).

Rappels

Inférence bayésienne : On infère (ou plutôt on déduit) la probabilité que le paramètre ait telle ou telle valeur sachant les données (et le prior) via le théorème de Bayes.

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

Objectif de l'analyse de données bayésienne : Faire évoluer une connaissance a priori sur les paramètres $p(\theta)$ en une connaissance a posteriori $p(\theta \mid y)$, intégrant l'information contenue dans les nouvelles données via $p(y \mid \theta)$.

Rappels

Les étapes de l'analyse de données bayésienne :

1. Définir le modèle - Identifier les paramètres du modèle génératif, définir une distribution a priori pour ces paramètres.

2. Mettre à jour le modèle - Calculer la distribution a posteriori des paramètres (ou une bonne approximation).

3. Interpréter la distribution postérieure - Comparaison de modèles, estimation des paramètres, vérification des prédictions du modèle.

Objectif du cours : Illustrer les différentes étapes de cette démarche à l'aide d'un modèle simple (un seul paramètre), le modèle Beta-Binomial.

Le modèle Beta-Binomial

Pourquoi ce modèle ?

- Le modèle Beta-Binomial couvre de nombreux problèmes de la vie courante :
  - Réussite / échec à un test
  - Présence / absence d'effets secondaires lors du test d'un médicament
- C'est un modèle simple
  - Un seul paramètre
  - Solution analytique

Loi de Bernoulli

S'applique à toutes les situations où le processus de génération des données ne peut résulter qu'en deux issues mutuellement exclusives (e.g., un lancer de pièce). À chaque essai, si on admet que $\Pr(\text{face}) = \theta$, alors $\Pr(\text{pile}) = 1 - \theta$.

Depuis Bernoulli, on sait calculer la probabilité du résultat d'un lancer de pièce, du moment que l'on connait le biais de la pièce θ. Admettons que $Y = 0$ lorsqu'on obtient pile, et que $Y = 1$ lorsqu'on obtient face. Alors Y est distribuée selon une loi de Bernoulli :

$$p(y \mid \theta) = \Pr(Y = y \mid \theta) = \theta^{y}(1 - \theta)^{(1-y)}$$

En remplaçant y par 0 ou 1, on retombe bien sur nos observations précédentes :

$$\Pr(Y = 1 \mid \theta) = \theta^{1}(1 - \theta)^{(1-1)} = \theta \times 1 = \theta$$

$$\Pr(Y = 0 \mid \theta) = \theta^{0}(1 - \theta)^{(1-0)} = 1 \times (1 - \theta) = 1 - \theta$$

Schéma de Bernoulli

Si l'on dispose d'une suite de lancers $\{Y_i\}$ indépendants et identiquement distribués (i.e., chaque lancer a une distribution de Bernoulli de probabilité θ), l'ensemble de ces lancers peut être décrit par une **distribution binomiale**.

Par exemple, imaginons que l'on dispose de la séquence de cinq lancers suivants : Pile, Pile, Pile, Face, Face. On peut recoder cette séquence en $\{0, 0, 0, 1, 1\}$.

Rappel : La probabilité de chaque 1 est θ est la probabilité de chaque 0 est $1 - \theta$.

Quelle est la probabilité d'obtenir 2 faces sur 5 lancers ?

Schéma de Bernoulli

Sachant que les essais sont indépendants les uns des autres, la probabilité d'obtenir cette séquence est de $(1-\theta) \times (1-\theta) \times (1-\theta) \times \theta \times \theta$, c'est à dire : $\theta^2(1-\theta)^3$.

On peut généraliser ce résultat pour une séquence de n lancers et y "succès" :

$$\theta^y(1-\theta)^{n-y}$$

On a jusqu'ici considéré seulement une seule séquence résultant en 2 succès pour 5 lancers, mais il existe de nombreuses séquences pouvant résulter en 2 succès pour 5 lancers (e.g., $\{0,0,1,0,1\}$, $\{0,1,1,0,0\}$)...

Coefficient binomial

Le **coefficient binomial** nous permet de calculer le nombre de combinaisons possibles résultant en y succès pour n lancers de la manière suivante (lu “y parmi n” ou “nombre de combinaisons de y parmi n”)[1] :

$$\binom{n}{y} = C_y^n = \frac{n!}{y!(n-y)!}$$

Par exemple pour $y = 1$ et $n = 3$, on sait qu’il existe 3 combinaisons possibles : $\{0, 0, 1\}, \{0, 1, 0\}, \{1, 0, 0\}$. On peut vérifier ça par le calcul, en appliquant la formule ci-dessus.

$$\binom{3}{1} = C_1^3 = \frac{3!}{1!(3-1)!} = \frac{3 \times 2 \times 1}{1 \times 2 \times 1} = \frac{6}{2} = 3$$

```r
# computing the total number of possible configurations in R
choose(n = 3, k = 1)
```

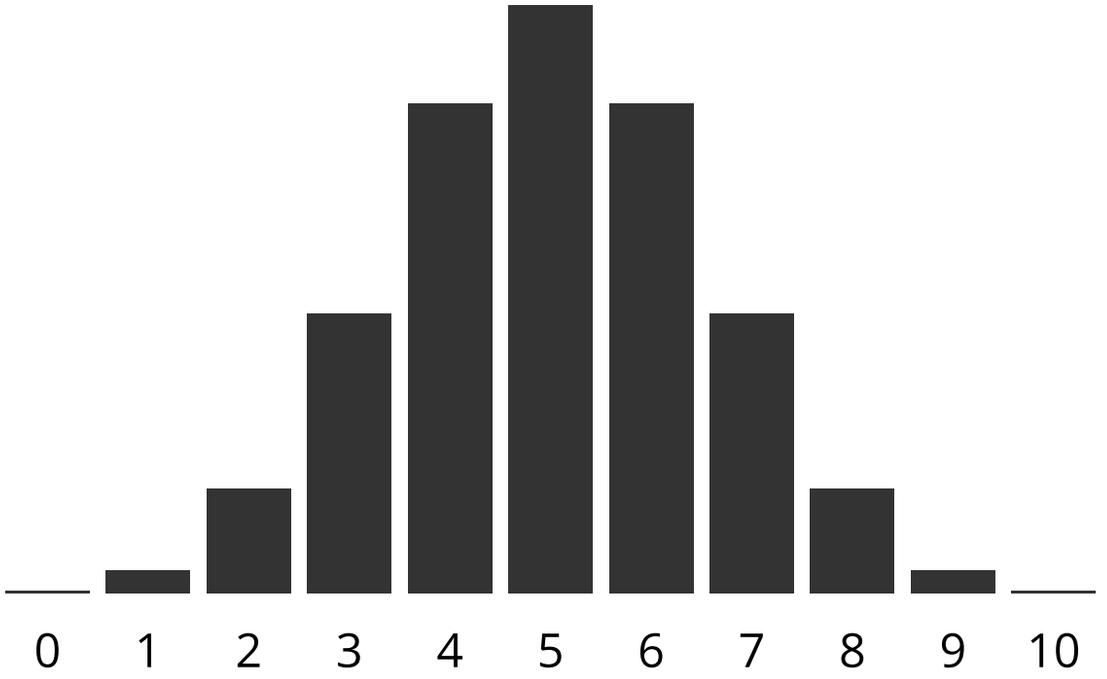
```
[1] 3
```

1. La fonction factorielle notée $n!$ associe à tout entier naturel n l’entier :

Loi binomiale

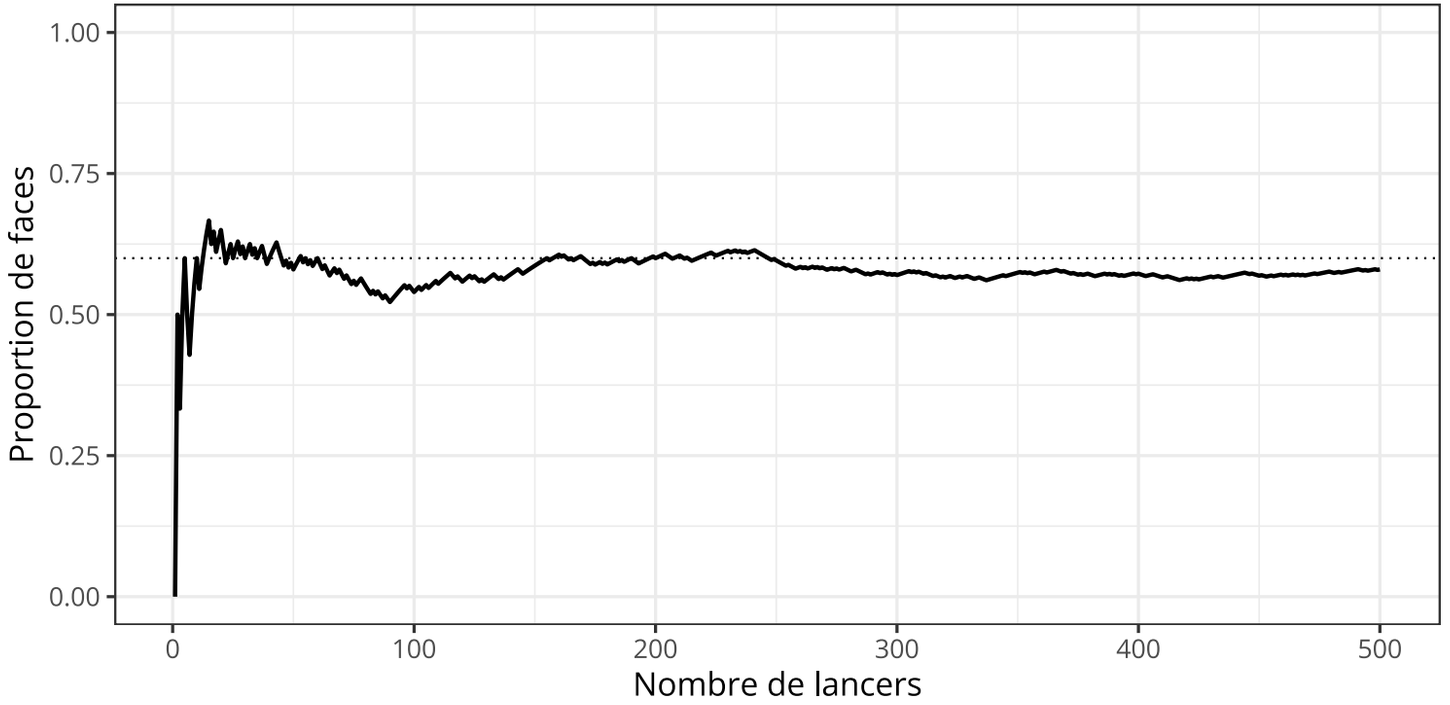
$$p(y \mid \theta) = \Pr(Y = y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$

La loi binomiale nous permet de calculer la probabilité d'obtenir y succès sur n essais, pour un θ donné. Exemple de la distribution binomiale pour une pièce non biaisée ($\theta = 0.5$), indiquant la probabilité d'obtenir n faces sur 10 lancers (en R: `dbinom(x = 0:10, size = 10, prob = 0.5)`).



Générer des données à partir d'une distribution binomiale

```r
library(tidyverse)
set.seed(666) # for reproducibility

sample(x = c(0, 1), size = 500, prob = c(0.4, 0.6), replace = TRUE) %>% # theta = 0.6
        data.frame() %>%
        mutate(x = seq_along(.), y = cummean(.) ) %>%
        ggplot(aes(x = x, y = y) ) +
        geom_line(lwd = 1) +
        geom_hline(yintercept = 0.6, lty = 3) +
        labs(x = "Nombre de lancers", y = "Proportion de faces") +
        ylim(0, 1)
```



Définition du modèle (likelihood)

Fonction de vraisemblance (likelihood)

- Nous considérons y comme étant le nombre de succès
- Nous considérons le nombre d'observations n comme étant une **constante**
- Nous considérons θ comme étant le **paramètre** de notre modèle (i.e., la probabilité de succès)

La fonction de vraisemblance s'écrit de la manière suivante :

$$\mathcal{L}(\theta \mid y, n) = p(y \mid \theta, n) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \propto \theta^y (1-\theta)^{n-y}$$

Vraisemblance versus probabilité

On lance à nouveau une pièce de biais θ (où θ représente la probabilité d'obtenir Face). On lance cette pièce deux fois et on obtient une Face et un Pile.

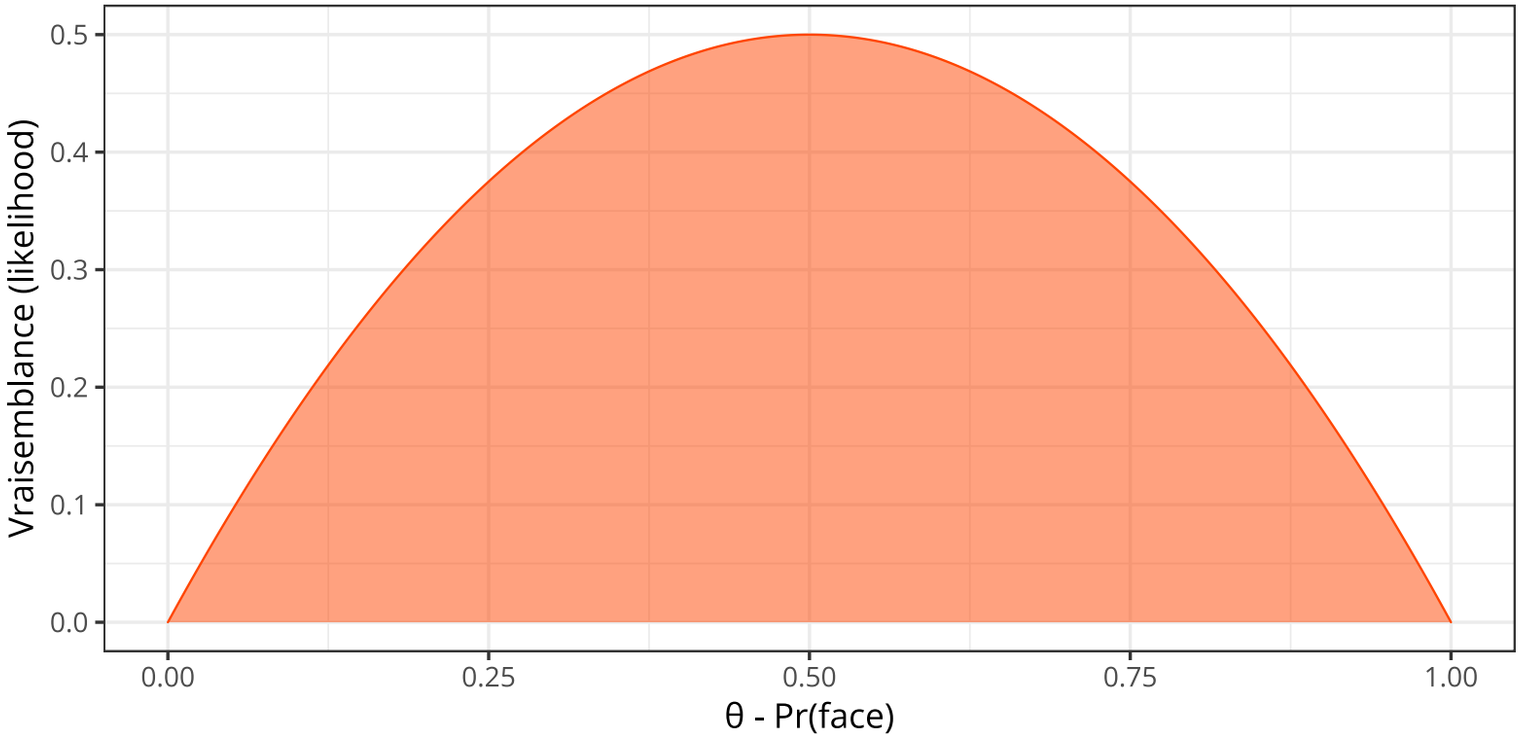
On peut calculer la probabilité d'observer une Face sur deux lancers de pièce **en fonction de** différentes valeurs de θ de la manière suivante :

$$\begin{aligned}\Pr(F, P \mid \theta) + \Pr(P, F \mid \theta) &= 2 \times \Pr(P \mid \theta) \times \Pr(F \mid \theta) \\ &= \theta(1-\theta) + \theta(1-\theta) \\ &= 2\theta(1-\theta)\end{aligned}$$

Cette probabilité est définie pour un jeu de données fixe et une valeur de θ variable. On peut représenter cette fonction visuellement.

Vraisemblance versus probabilité

```r
# Représentation graphique de la fonction de vraisemblance de theta pour y = 1 et n = 2

y <- 1 # nombre de faces
n <- 2 # nombre d'essais

data.frame(theta = seq(from = 0, to = 1, length.out = 1e3) ) %>%
  mutate(likelihood = dbinom(x = y, size = n, prob = theta) ) %>%
  ggplot(aes(x = theta, y = likelihood) ) +
  geom_area(color = "orangered", fill = "orangered", alpha = 0.5) +
  labs(x = expression(paste(theta, " - Pr(face)") ), y = "Vraisemblance (likelihood)")
```



Vraisemblance versus probabilité

Si on calcule l'aire sous la courbe de cette fonction, on obtient :

$$\int_0^1 2\theta(1-\theta)\mathrm{d}\theta = \frac{1}{3}$$

```r
f <- function(theta) {2 * theta * (1 - theta) }
integrate(f = f, lower = 0, upper = 1)
```

```
0.3333333 with absolute error < 3.7e-15
```

Quand on varie θ, la fonction de vraisemblance **n'est pas** une distribution de probabilité valide (i.e., son intégrale n'est pas égale à 1). On utilise le terme de **vraisemblance**, pour distinguer ce type de fonction des fonctions de densité de probabilité. On utilise la notation suivante pour mettre l'accent sur le fait que la fonction de vraisemblance est une fonction de θ, et que les données sont fixes :
$\mathcal{L}(\theta \mid \text{data}) = p(\text{data} \mid \theta)$.

Vraisemblance versus probabilité

Vraisemblance versus probabilité
pour deux lancers de pièce

	Nombre de Faces (y)			
theta	**0**	**1**	**2**	**Total**
0	1.00	0.00	0.00	1
0.2	0.64	0.32	0.04	1
0.4	0.36	0.48	0.16	1
0.6	0.16	0.48	0.36	1
0.8	0.04	0.32	0.64	1
1	0.00	0.00	1.00	1
Total	2.20	1.60	2.20	

Notons que la vraisemblance de θ pour une donnée particulière est égale à la probabilité de cette donnée pour cette valeur de θ. Cependant, la *distribution* de ces vraisemblances (en colonne) n'est pas une distribution de probabilité. Dans l'analyse bayésienne, **les données sont considérées comme fixes** et la valeur de θ est considérée comme une **variable aléatoire**.

Définition du modèle (prior)

Comment définir un prior dans le cas du lancer de pièce ?

Aspect sémantique → *le prior doit pouvoir rendre compte* :

- D'une absence d'information
- D'une connaissance d'observations antérieures concernant la pièce étudiée
- D'un niveau d'incertitude concernant ces observations antérieures

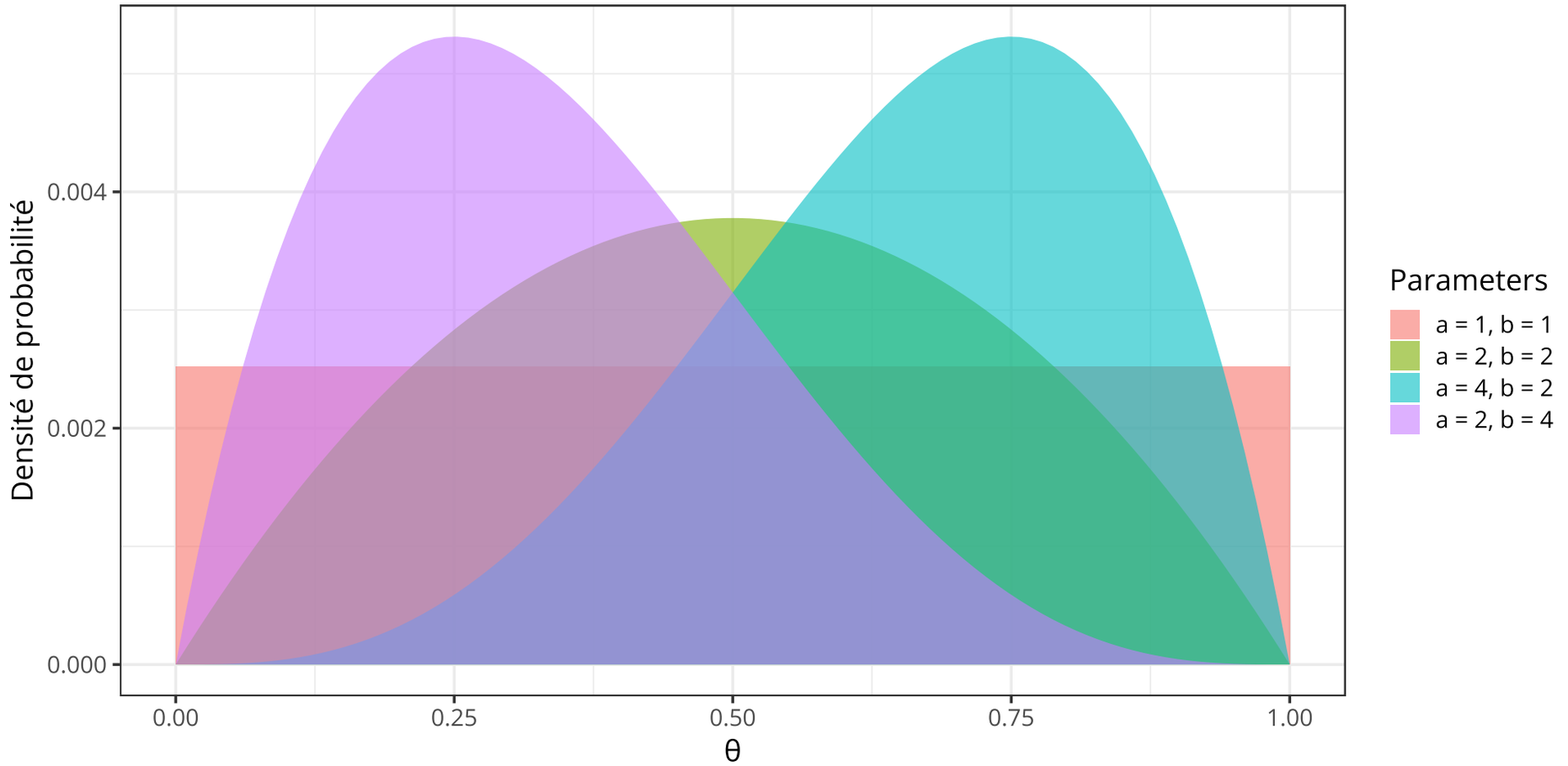
Aspect mathématique → *pour une solution entièrement analytique* :

- Les distributions a priori et a posteriori doivent avoir la même forme
- La vraisemblance marginale doit pouvoir se calculer analytiquement

La distribution Beta

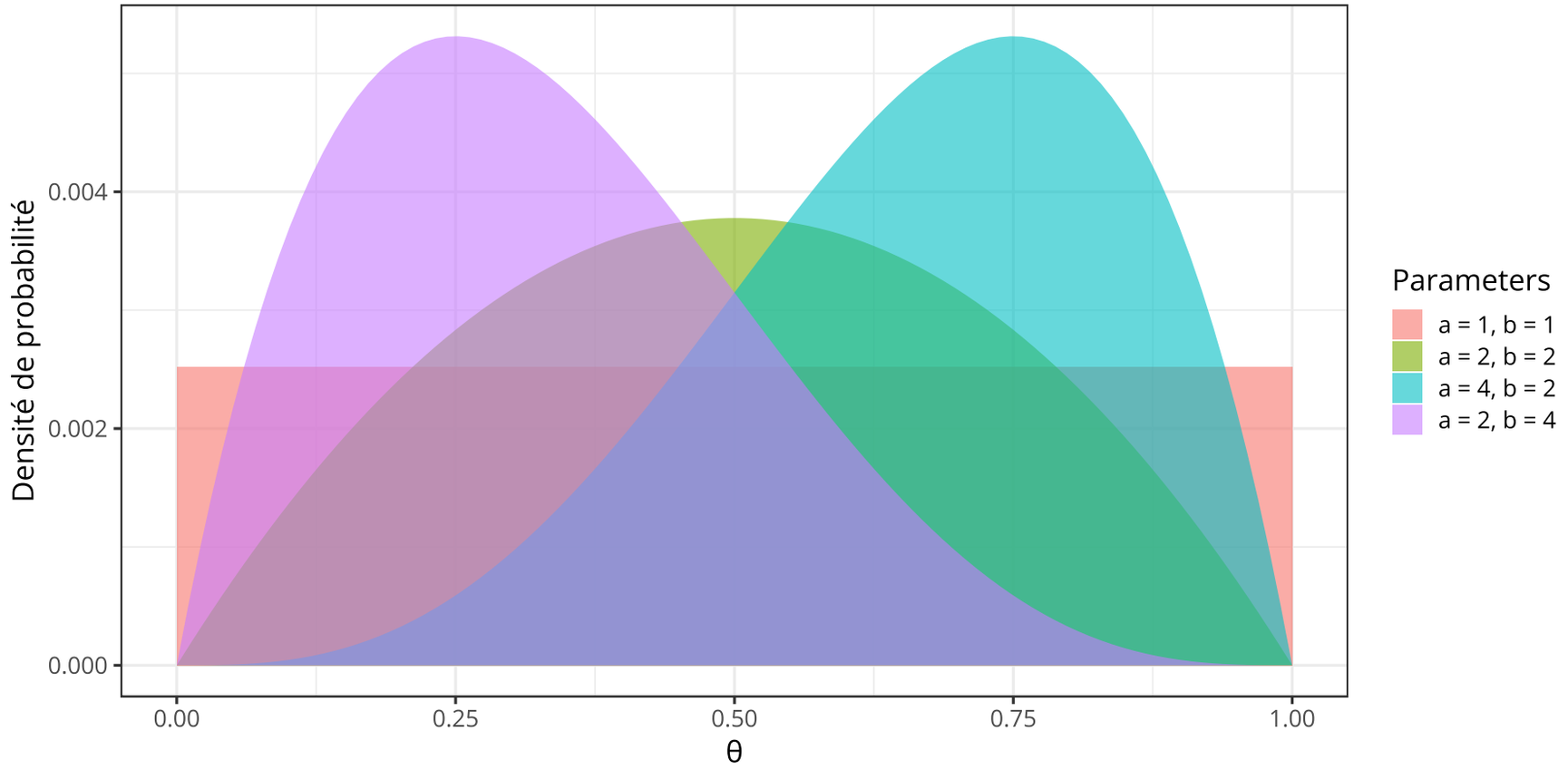
$$
\begin{aligned}
p(\theta \mid a, b) &= \text{Beta}(\theta \mid a, b) \\
&= \theta^{a-1}(1-\theta)^{b-1}/B(a,b) \\
&\propto \theta^{a-1}(1-\theta)^{b-1}
\end{aligned}
$$

où a et b sont deux paramètres tels que $a \geq 0, b \geq 0$, et $B(a,b)$ est une constante de normalisation.



Interprétation des paramètres du prior Beta

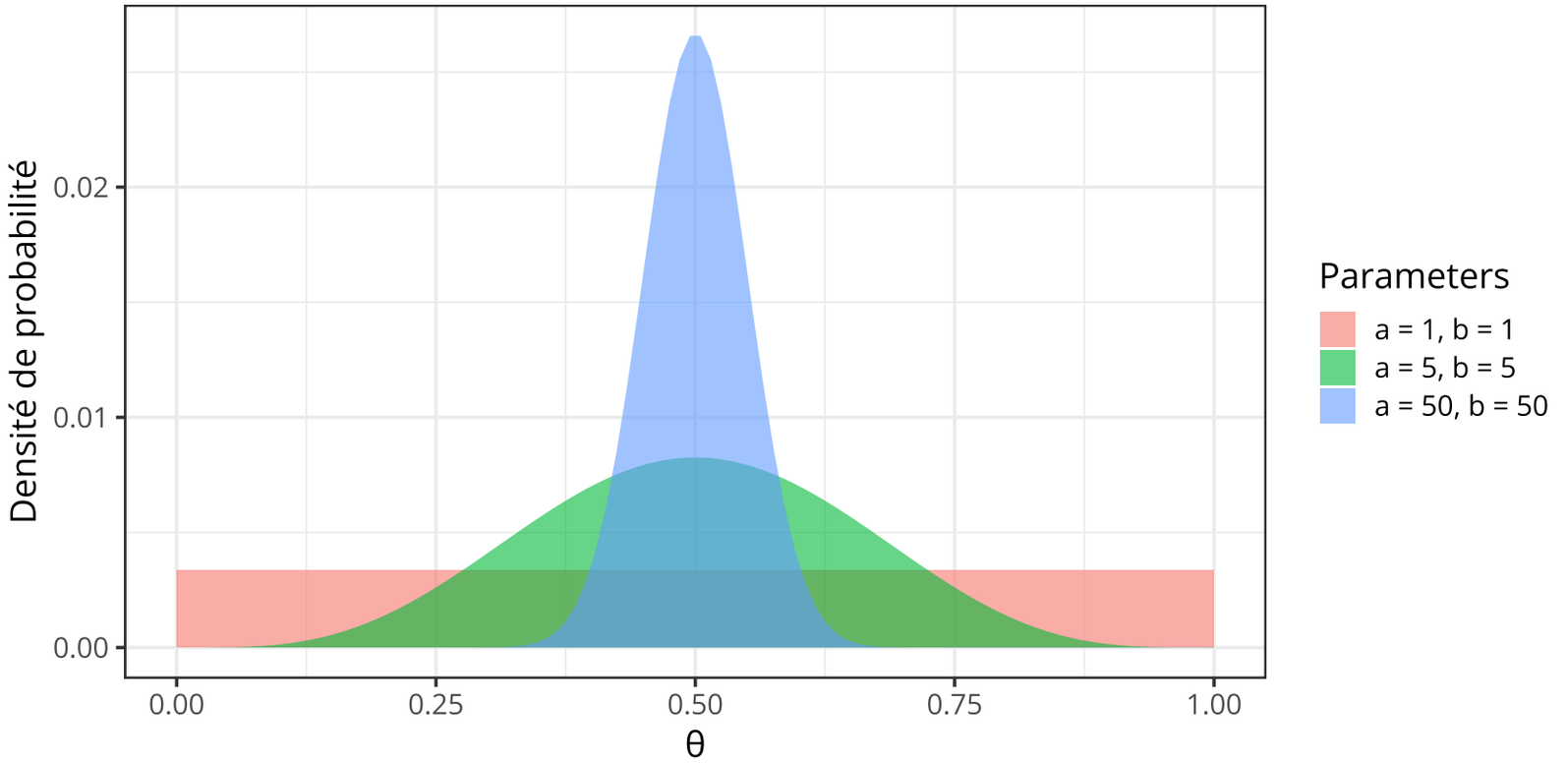
- On peut exprimer l'absence de connaissance a priori par $a = b = 1$ (distribution orange).
- On peut exprimer un prior en faveur d'une absence de biais par $a = b \geq 2$ (distribution verte).
- On peut exprimer un biais en faveur de *Face* par $a > b$ (distribution bleue).
- On peut exprimer un biais en faveur de *Pile* par $a < b$ (distribution violette).



Interprétation des paramètres du prior Beta

Le niveau de certitude augmente avec la somme $\kappa = a + b$.

- Aucune idée sur la provenance de la pièce : $a = b = 1$ -> **prior plat**.
- En attendant le début de l'expérience, on a lancé la pièce 10 fois et observé 5 "Face" : $a = b = 5$ -> **prior peu informatif**.
- La pièce provient de la banque de France : $a = b = 50$ -> **prior fort**.

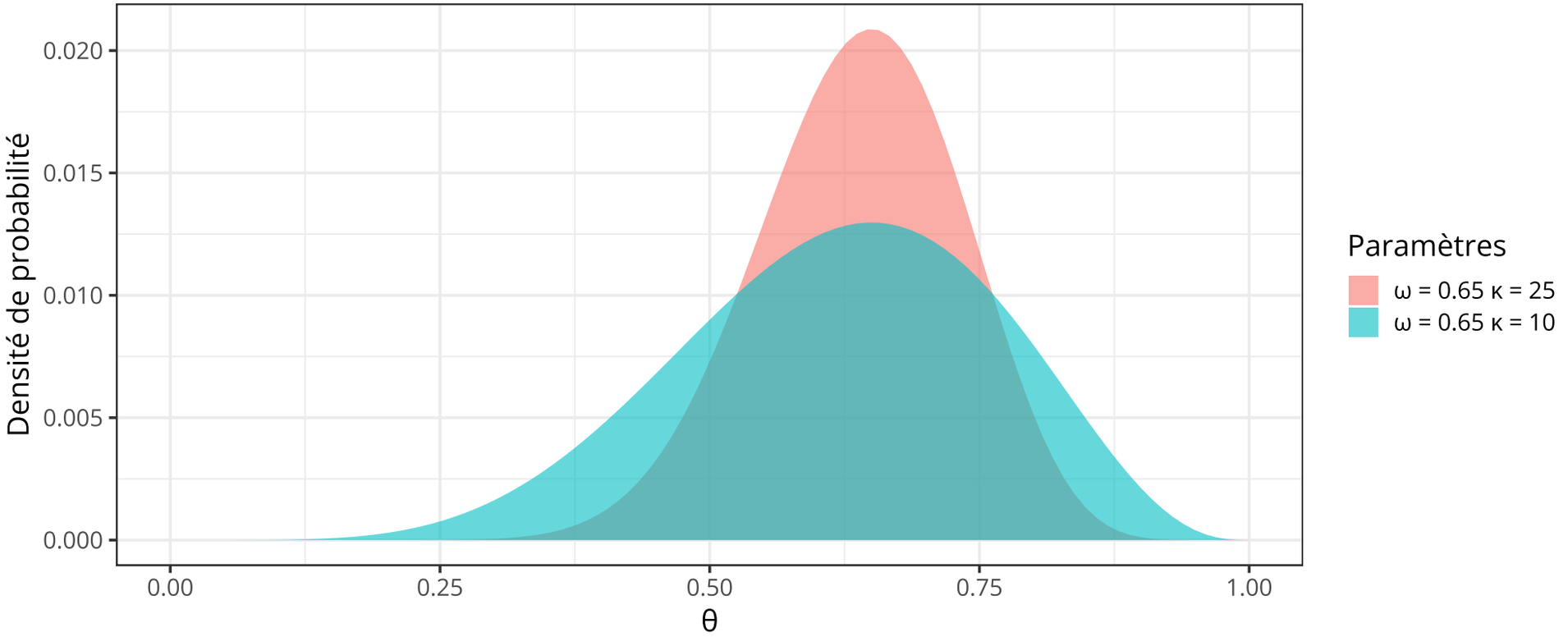


Interprétation des paramètres du prior Beta

Supposons que l'on dispose d'une estimation de la valeur la plus probable ω du paramètre θ. On peut reparamétriser la distribution Beta en fonction du mode ω et du niveau de certitude κ :

$$a = \omega(\kappa - 2) + 1$$
$$b = (1 - \omega)(\kappa - 2) + 1 \qquad \text{pour } \kappa > 2$$

Si $\omega = 0.65$ et $\kappa = 25$ alors $p(\theta) = \text{Beta}(\theta \mid 15.95, 9.05)$.
Si $\omega = 0.65$ et $\kappa = 10$ alors $p(\theta) = \text{Beta}(\theta \mid 6.2, 3.8)$.



Prior conjugué

Formellement, si $\mathcal{F}$ est une classe de distributions d'échantillonnage $p(y|\theta)$, et $\mathcal{P}$ est une classe de distributions a priori pour θ, alors $\mathcal{P}$ est **conjuguée** à $\mathcal{F}$ si et seulement si :

$$p(\theta \mid y) \in \mathcal{P} \text{ for all } p(\cdot \mid \theta) \in \mathcal{F} \text{ and } p(\cdot) \in \mathcal{P}$$

(p.35, Gelman et al., 2013). En d'autres termes, un prior est appelé **conjugué** si, lorsqu'il est converti en une distribution a posteriori en étant multiplié par la fonction de vraisemblance, il conserve la même forme. Dans notre cas, le prior Beta est un prior conjugué pour la vraisemblance binomiale, car le posterior est également une distribution Beta.

> “
>
> Le résultat du produit d'un prior Beta et d'une fonction de vraisemblance Binomiale est proportionnel à une distribution Beta. On dit que la distribution Beta est **un prior conjugué** de la fonction de vraisemblance Binomiale.

Dérivation analytique de la distribution a posteriori

Soit un prior défini par : $p(\theta \mid a, b) = \text{Beta}(a, b) = \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \propto \theta^{a-1}(1-\theta)^{b-1}$

Soit une fonction de vraisemblance associée à y “Face” pour n lancers :

$$p(y \mid n, \theta) = \text{Bin}(y \mid n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y} \propto \theta^y (1-\theta)^{n-y}$$

Alors (en omettant les constantes de normalisation) :

$$\begin{aligned}
p(\theta \mid y, n) &\propto p(y \mid n, \theta)\, p(\theta) && \text{Théorème de Bayes} \\
&\propto \text{Bin}(y \mid n, \theta)\, \text{Beta}(\theta \mid a, b) \\
&\propto \theta^y (1-\theta)^{n-y}\, \theta^{a-1}(1-\theta)^{b-1} && \text{Application des formules précédentes} \\
&\propto \theta^{y+a-1}(1-\theta)^{n-y+b-1} && \text{En regroupant les puissances des termes identiques}
\end{aligned}$$

Ici, on a ignoré les constantes qui ne dépendent pas de θ (i.e., le nombre de combinaisons dans la fonction de vraisemblance binomiale et la fonction Beta $B(a, b)$ dans le prior Beta).[1] En les prenant en compte, on obtient en effet une distribution a posteriori Beta de la forme suivante :

$$p(\theta \mid y, n) = \text{Beta}(y + a, n - y + b)$$

Un exemple pour digérer

On observe $y = 7$ réponses correctes sur $n = 10$ questions. On choisit un prior $\text{Beta}(1, 1)$, c'est à dire un prior uniforme sur $[0, 1]$. Ce prior équivaut à une connaissance a priori de 0 succès et 0 échecs (i.e., prior plat).

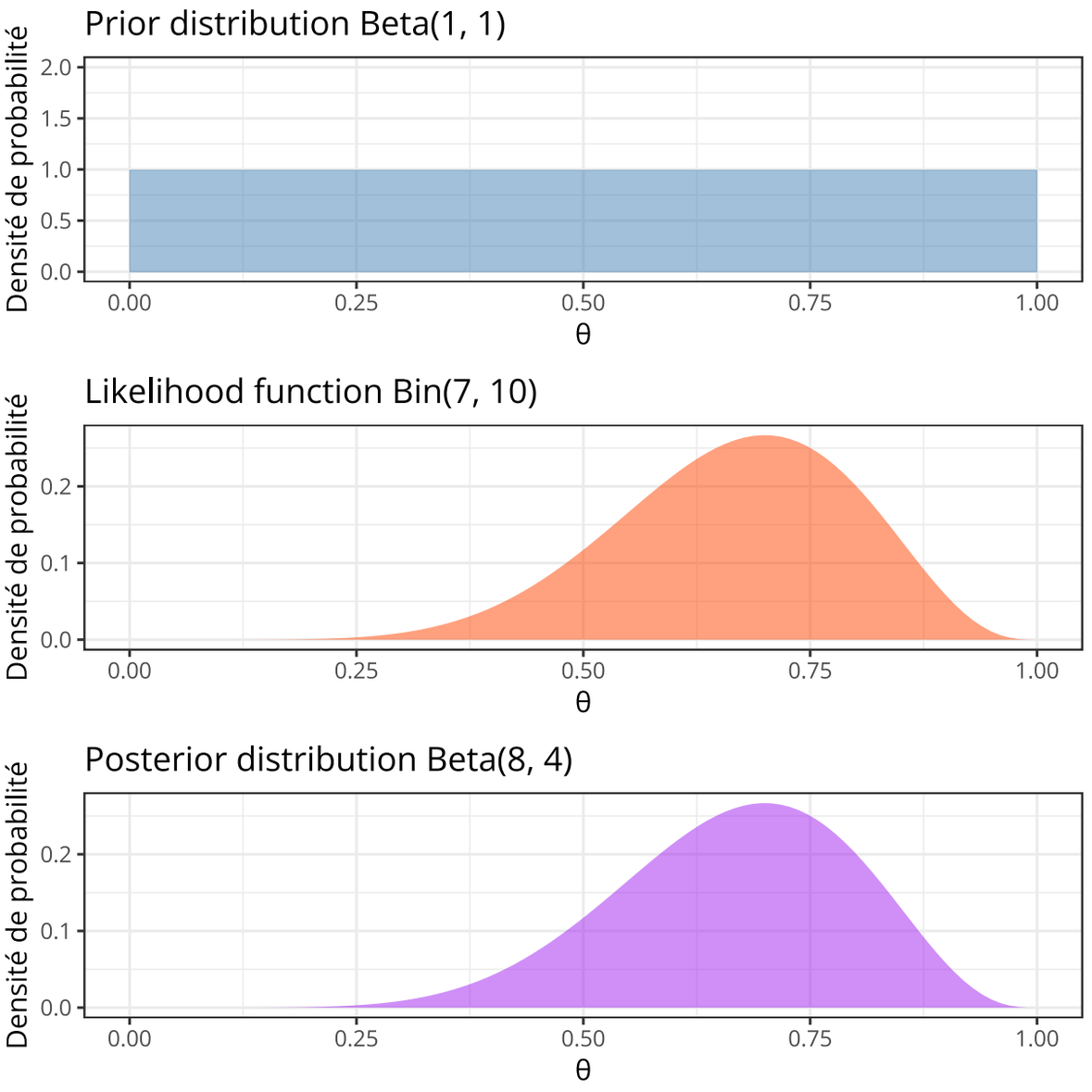
La distribution postérieure est donnée par :

$$\begin{aligned} p(\theta \mid y, n) &\propto p(y \mid n, \theta)\, p(\theta) \\ &\propto \text{Bin}(7 \mid 10, \theta)\, \text{Beta}(\theta \mid 1, 1) \\ &= \text{Beta}(y + a, n - y + b) \\ &= \text{Beta}(8, 4) \end{aligned}$$

La moyenne de la distribution postérieure est donnée par :

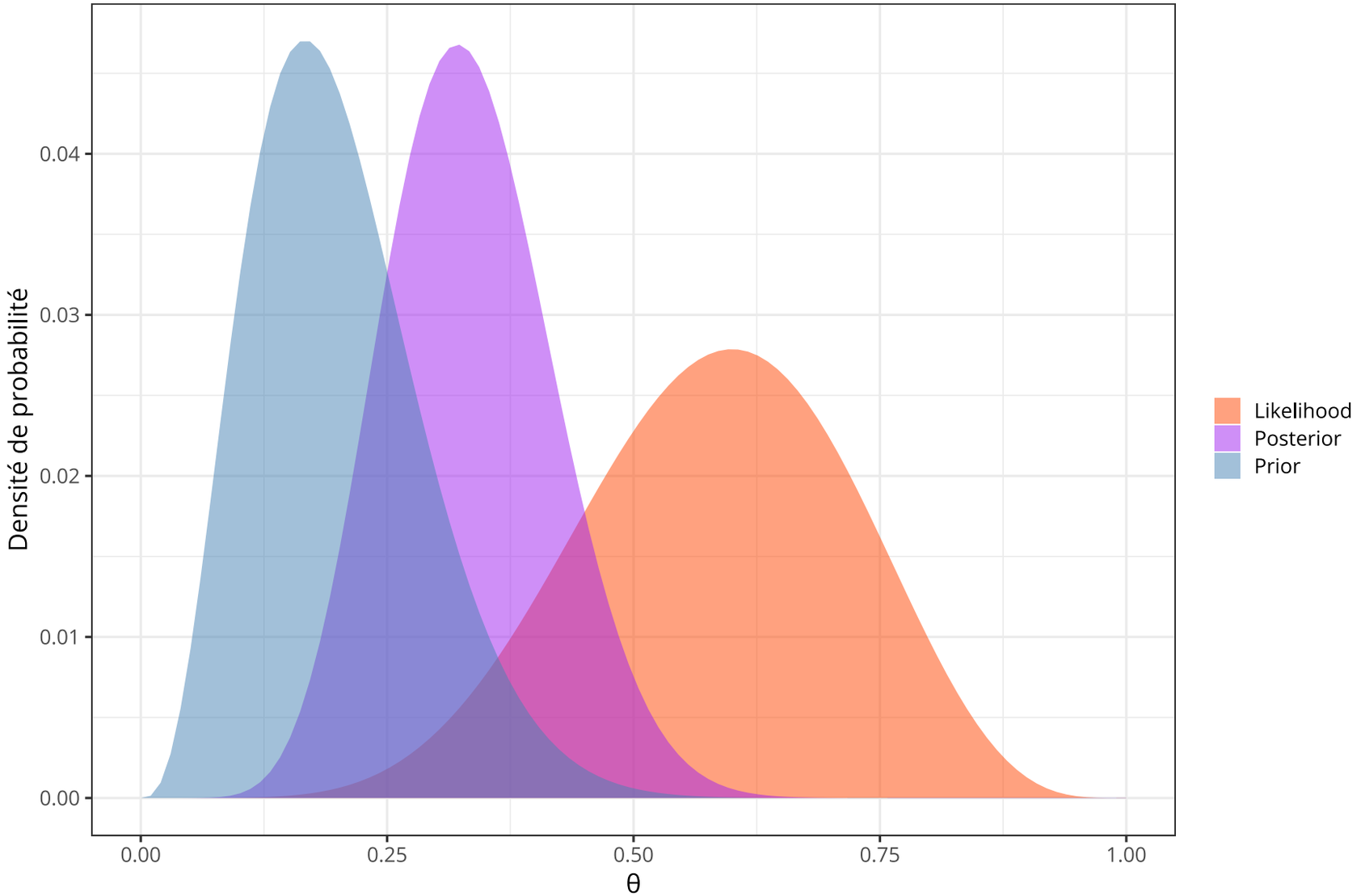
$$\underbrace{\frac{y + a}{n + a + b}}_{posterior} = \underbrace{\frac{y}{n}}_{data} \underbrace{\frac{n}{n + a + b}}_{weight} + \underbrace{\frac{a}{a + b}}_{prior} \underbrace{\frac{a + b}{n + a + b}}_{weight}$$

Un exemple pour digérer



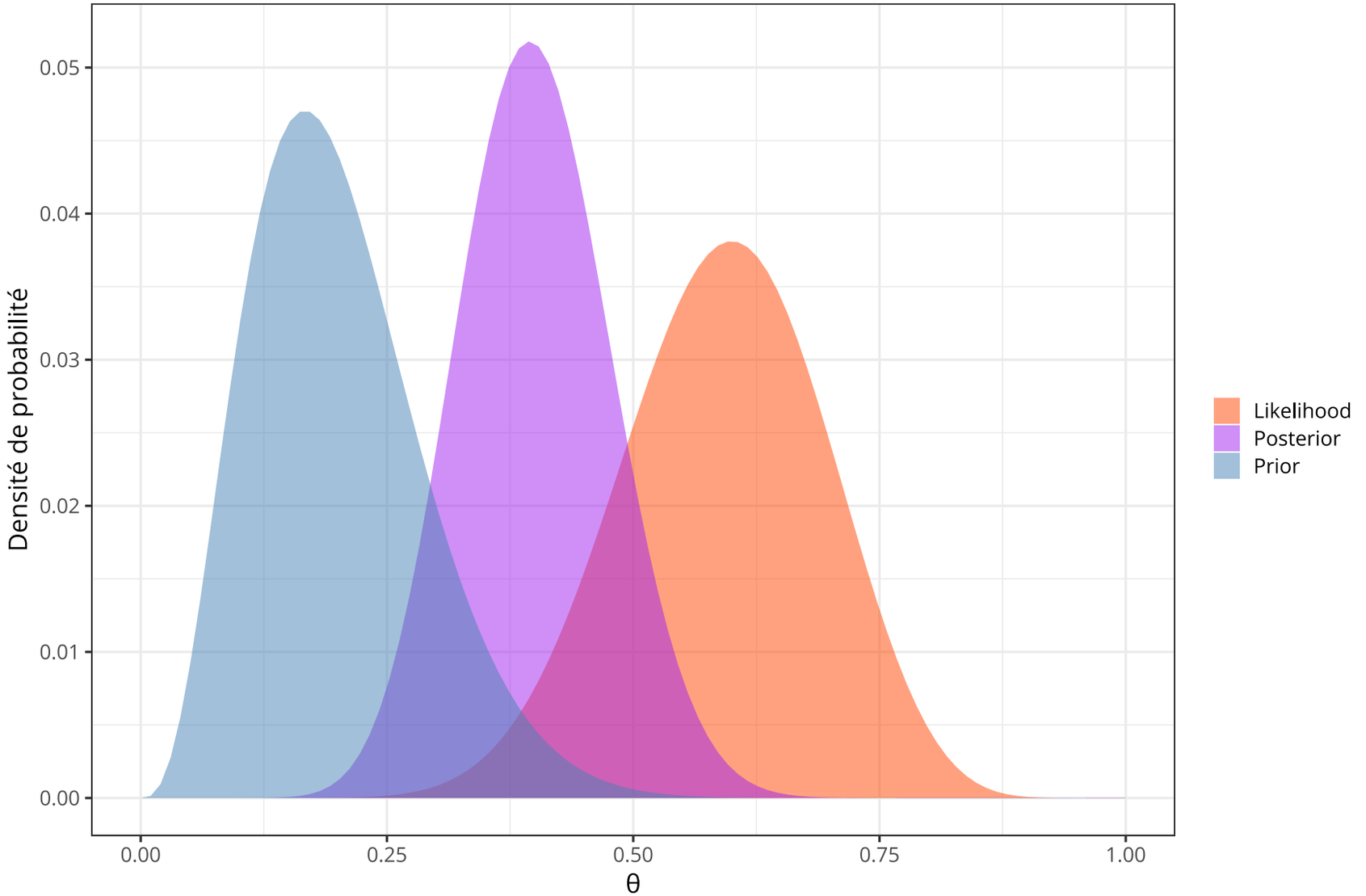
Influence du prior sur la distribution postérieure

Cas $n < a + b, (n = 10, a = 4, b = 16)$.



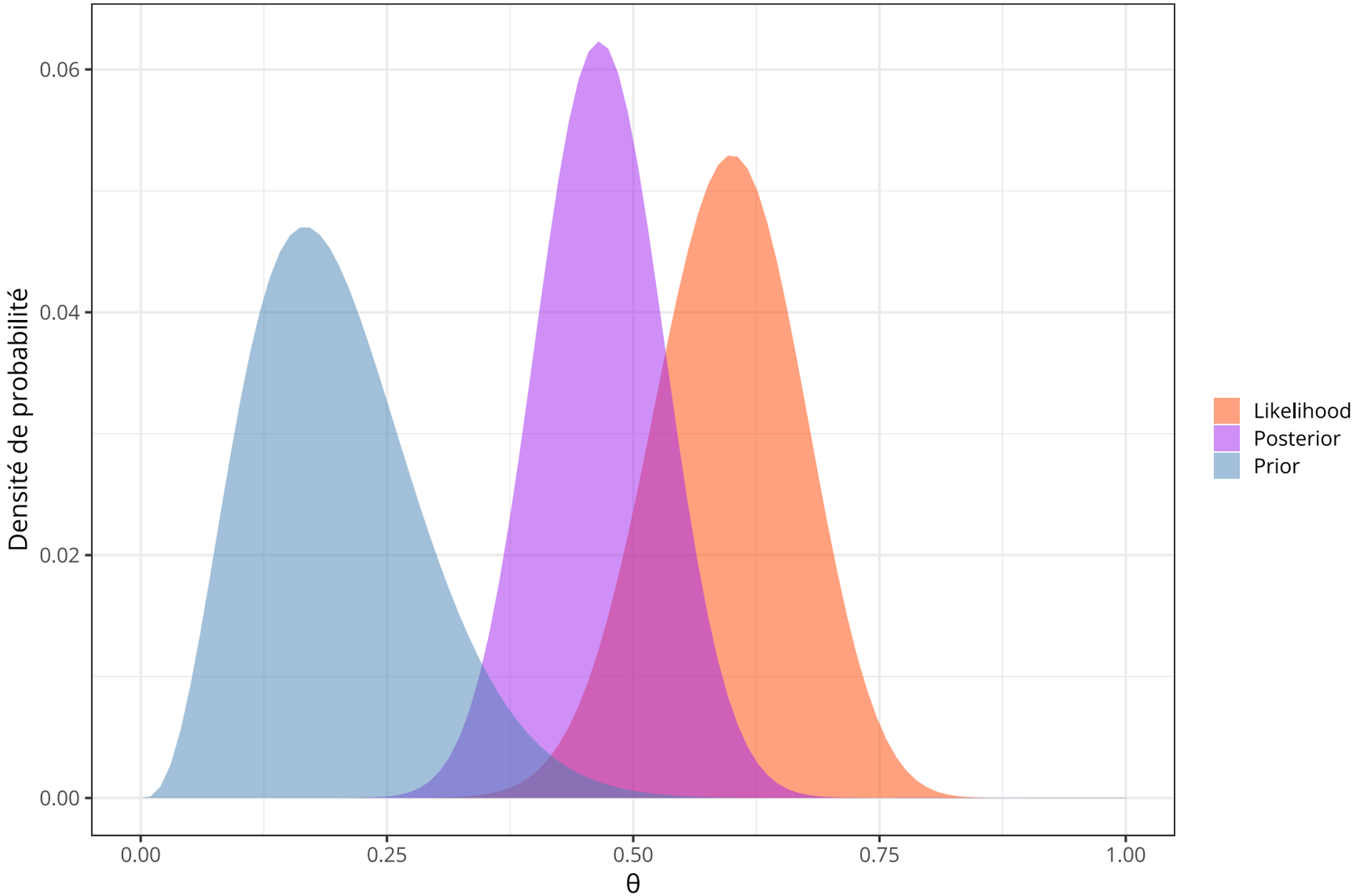
Influence du prior sur la distribution postérieure

Cas $n = a + b, (n = 20, a = 4, b = 16)$.



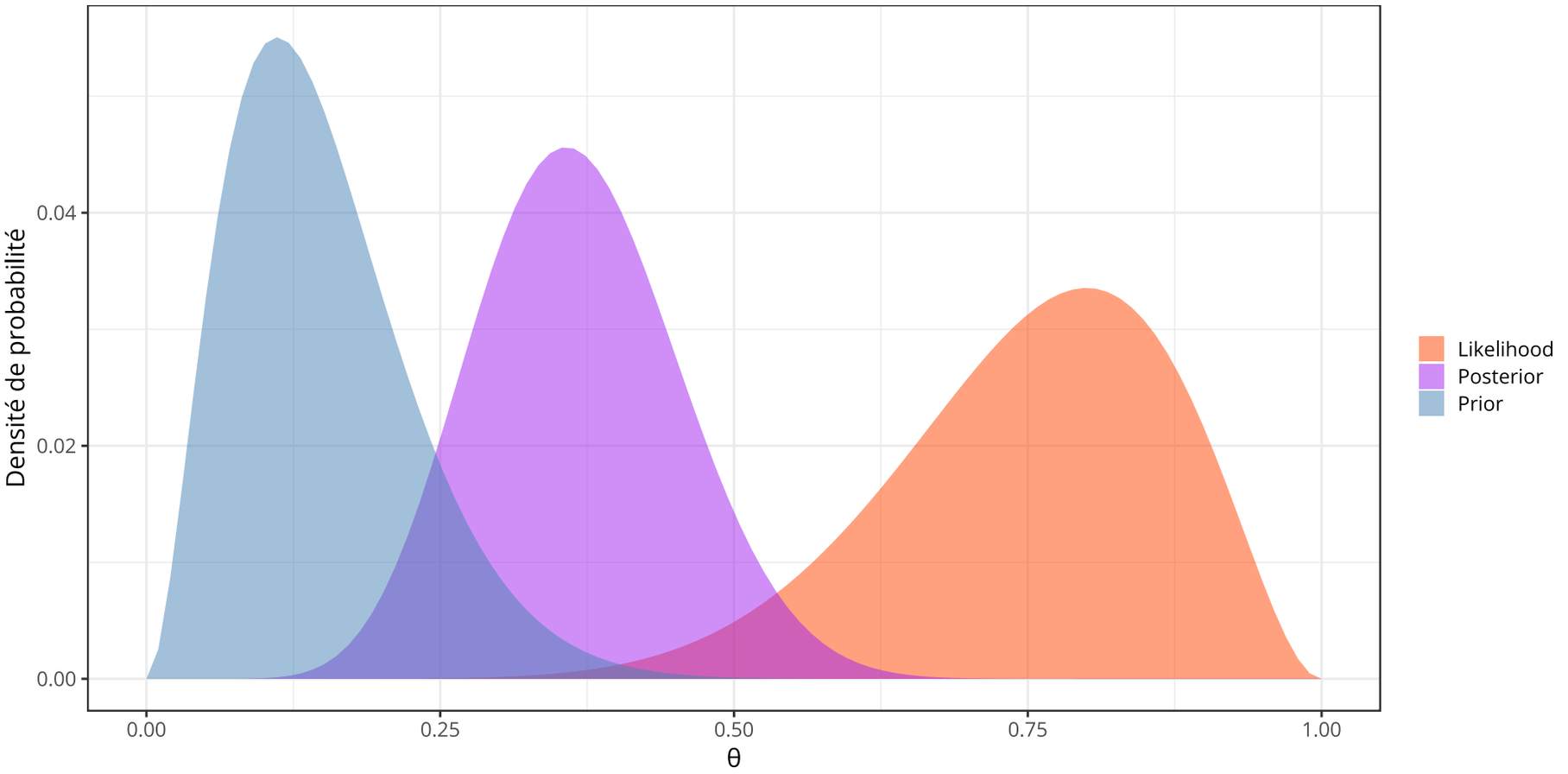
Influence du prior sur la distribution postérieure

Cas $n > a + b, (n = 40, a = 4, b = 16)$.



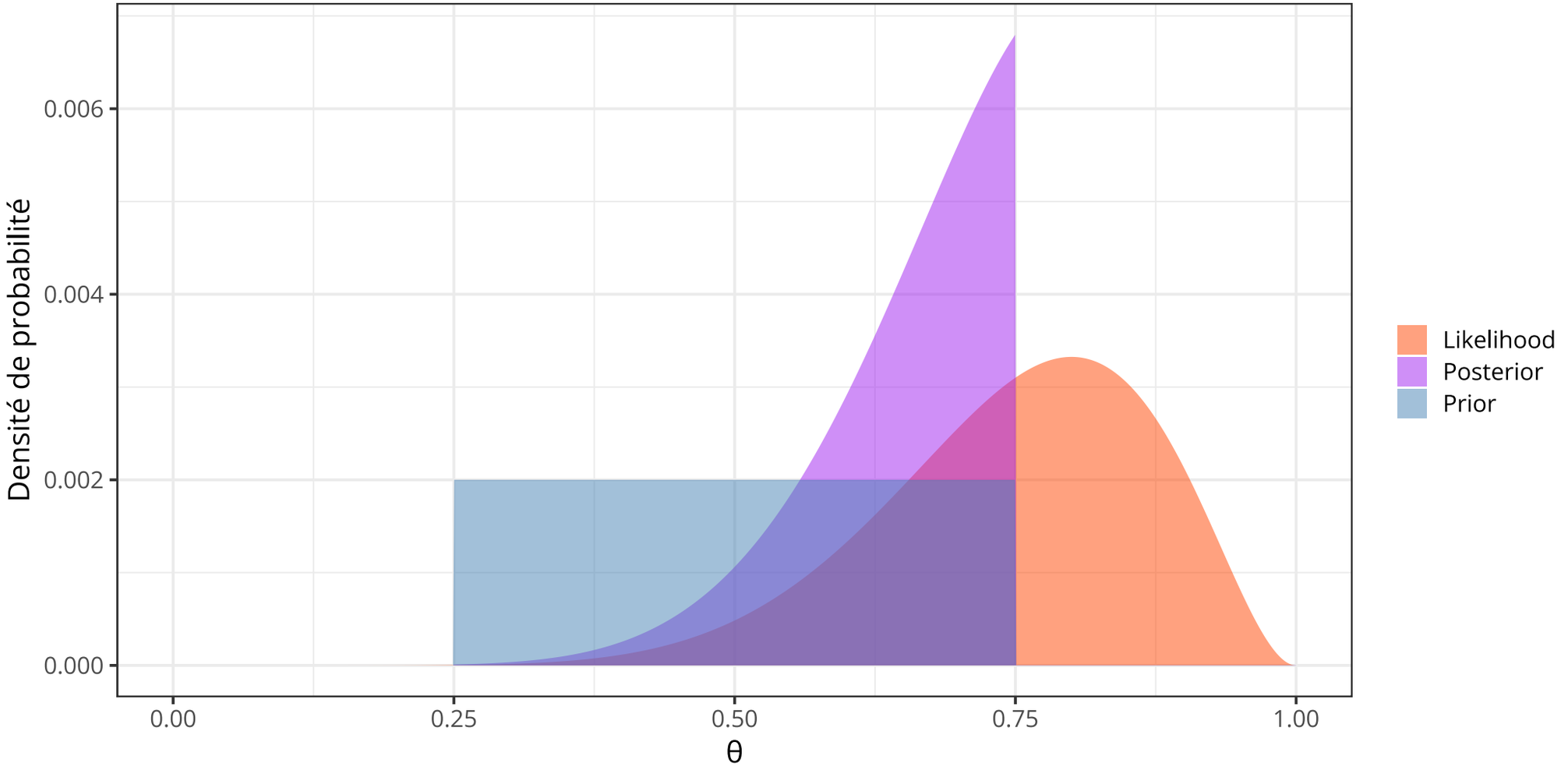
Ce qu'il faut retenir

> The posterior distribution is always a compromise between the prior distribution and the likelihood function (Kruschke, 2015).



Ce qu'il faut retenir

Plus on a de données, moins le prior a d'influence dans l'estimation de la distribution a posteriori (et réciproquement). **Attention :** Lorsque le prior accorde une probabilité de 0 à certaines valeurs de θ, le modèle est incapable d'apprendre (ces valeurs sont alors considérées comme "impossibles")...



La vraisemblance marginale

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Marginal likelihood}} \propto \text{Likelihood} \times \text{Prior}$$

$$p(\theta \mid \text{data}) = \frac{p(\text{data} \mid \theta) \times p(\theta)}{p(\text{data})} \propto p(\text{data} \mid \theta) \times p(\theta)$$

Si on zoom sur la vraisemblance marginale (aussi connue comme **evidence**)...

$$p(\text{data}) = \int p(\text{data}, \theta)\, \mathrm{d}\theta \qquad \text{Marginalisation sur le paramètre } \theta$$

$$p(\text{data}) = \int p(\text{data} \mid \theta) \times p(\theta)\, \mathrm{d}\theta \qquad \text{Application de la règle du produit}$$

La vraisemblance marginale

Petit problème : $p(\text{data})$ s'obtient en calculant la somme (pour des variables discrètes) ou l'intégrale (pour des variables continues) de la densité conjointe $p(\text{data}, \theta)$ sur toutes les valeurs possibles de θ. Cela se complique lorsque le modèle comprend plusieurs paramètres continus...

Par exemple pour deux paramètres discrets :

$$p(\text{data}) = \sum_{\theta_1} \sum_{\theta_2} p(\text{data}, \theta_1, \theta_2)$$

Et pour un modèle avec deux paramètres continus :

$$p(\text{data}) = \int_{\theta_1} \int_{\theta_2} p(\text{data}, \theta_1, \theta_2) \mathrm{d}\theta_1 \mathrm{d}\theta_2$$

La vraisemblance marginale

Trois méthodes pour résoudre (contourner) ce problème :

- Solution analytique ⟶ Utilisation d'un prior conjugué (e.g., le modèle Beta-Binomial).
- Solution discrétisée ⟶ Calcul de la solution sur un ensemble fini de points (grid method).
- Solution approchée ⟶ On échantillonne "intelligemment" l'espace conjoint des paramètres (e.g., méthodes MCMC, Cours n°05).

Distributions discrètes

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters[note 1]	Posterior predictive[note 2]
Bernoulli	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i, \beta + n - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$p(\tilde{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
Binomial	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	$\mathrm{BetaBin}(\tilde{x} \mid \alpha', \beta')$ (beta-binomial)
Negative binomial with known failure number, r	p (probability)	Beta	α, β	$\alpha + \sum_{i=1}^{n} x_i, \beta + rn$	$\alpha - 1$ total successes, $\beta - 1$ failures[note 1] (i.e., $\frac{\beta - 1}{r}$ experiments, assuming r stays fixed)	
Poisson	λ (rate)	Gamma	k, θ	$k + \sum_{i=1}^{n} x_i, \frac{\theta}{n\theta + 1}$	k total occurrences in $\frac{1}{\theta}$ intervals	$\mathrm{NB}(\tilde{x} \mid k', \theta')$ (negative binomial)
			α, β[note 3]	$\alpha + \sum_{i=1}^{n} x_i, \beta + n$	α total occurrences in β intervals	$\mathrm{NB}\left(\tilde{x} \mid \alpha', \frac{1}{1 + \beta'}\right)$ (negative binomial)
Categorical	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + (c_1, \ldots, c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i[note 1]	$p(\tilde{x} = i) = \frac{{\alpha_i}'}{\sum_i {\alpha_i}'} = \frac{\alpha_i + c_i}{\sum_i \alpha_i + n}$
Multinomial	$\boldsymbol{p}$ (probability vector), k (number of categories; i.e., size of $\boldsymbol{p}$)	Dirichlet	$\boldsymbol{\alpha}$	$\boldsymbol{\alpha} + \sum_{i=1}^{n} \mathbf{x}_i$	$\alpha_i - 1$ occurrences of category i[note 1]	$\mathrm{DirMult}(\tilde{\mathbf{x}} \mid \boldsymbol{\alpha}')$ (Dirichlet-multinomial)
Hypergeometric with known total population size, N	M (number of target members)	Beta-binomial[4]	$n = N, \alpha, \beta$	$\alpha + \sum_{i=1}^{n} x_i, \beta + \sum_{i=1}^{n} N_i - \sum_{i=1}^{n} x_i$	$\alpha - 1$ successes, $\beta - 1$ failures[note 1]	
Geometric	p_0 (probability)	Beta	α, β	$\alpha + n, \beta + \sum_{i=1}^{n} x_i$	$\alpha - 1$ experiments, $\beta - 1$ total failures[note 1]	

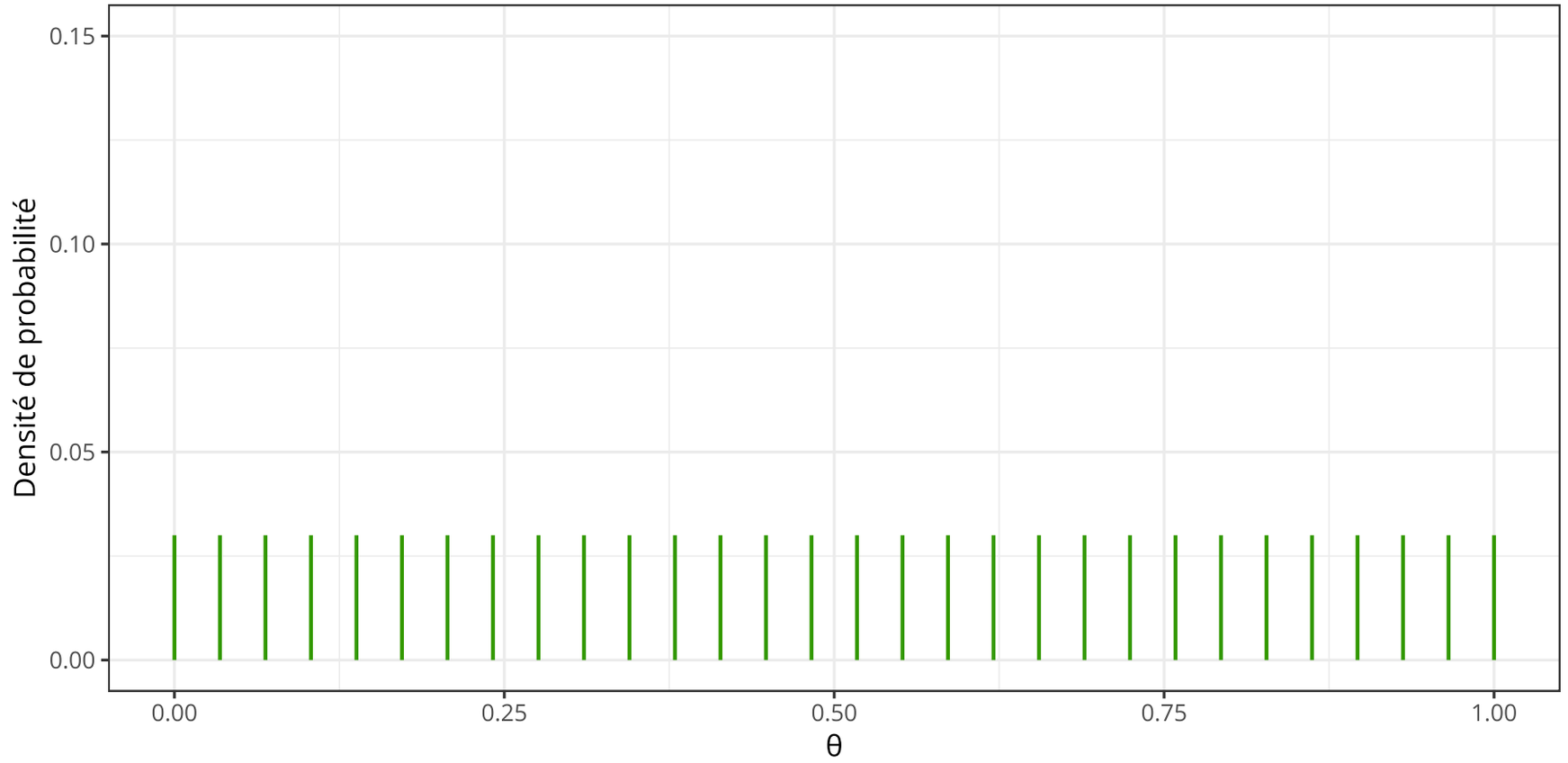
Distributions continues

Likelihood	Model parameters	Conjugate prior distribution	Prior hyperparameters	Posterior hyperparameters	Interpretation of hyperparameters	Posterior predictive[note 4]
Normal with known variance σ^2	μ (mean)	Normal	$\mu_0,\ \sigma_0^2$	$\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right)\Big/\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right),$ $\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\tilde{x}\vert\mu_0', \sigma_0^{2\prime}+\sigma^2)$[5]
Normal with known precision τ	μ (mean)	Normal	$\mu_0,\ \tau_0$	$\left(\tau_0\mu_0+\tau\sum_{i=1}^n x_i\right)\Big/(\tau_0+n\tau),\ \tau_0+n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}\left(\tilde{x}\vert\mu_0', \frac{1}{\tau_0'}+\frac{1}{\tau}\right)$[5]
Normal with known mean μ	σ^2 (variance)	Inverse gamma	$\alpha,\ \beta$ [note 5]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	variance was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\vert\mu, \sigma^2=\beta'/\alpha')$[5]
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	$\nu,\ \sigma_0^2$	$\nu+n,\ \frac{\nu\sigma_0^2+\sum_{i=1}^n(x_i-\mu)^2}{\nu+n}$	variance was estimated from ν observations with sample variance σ_0^2	$t_{\nu'}(\tilde{x}\vert\mu, \sigma_0^{2\prime})$[5]
Normal with known mean μ	τ (precision)	Gamma	$\alpha,\ \beta$[note 3]	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}^n(x_i-\mu)^2}{2}$	precision was estimated from 2α observations with sample variance β/α (i.e. with sum of squared deviations 2β, where deviations are from known mean μ)	$t_{2\alpha'}(\tilde{x}\vert\mu, \sigma^2=\beta'/\alpha')$[5]
Normal[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0,\ \nu,\ \alpha,\ \beta$	$\frac{\nu\mu_0+n\bar{x}}{\nu+n},\ \nu+n,\ \alpha+\frac{n}{2},$ $\beta+\frac{1}{2}\sum_{i=1}^n(x_i-\bar{x})^2+\frac{n\nu}{\nu+n}\frac{(\bar{x}-\mu_0)^2}{2}$ • $\bar{x}$ is the sample mean	mean was estimated from ν observations with sample mean μ_0; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β	$t_{2\alpha'}\left(\tilde{x}\vert\mu', \frac{\beta'(\nu'+1)}{\nu'\alpha'}\right)$[5]

Problème : Cette solution est très contraignante. Idéalement, le modèle (likelihood + prior) devrait être défini à partir de l'interprétation que l'on peut faire des paramètres de ces distributions, et non pour faciliter les calculs...

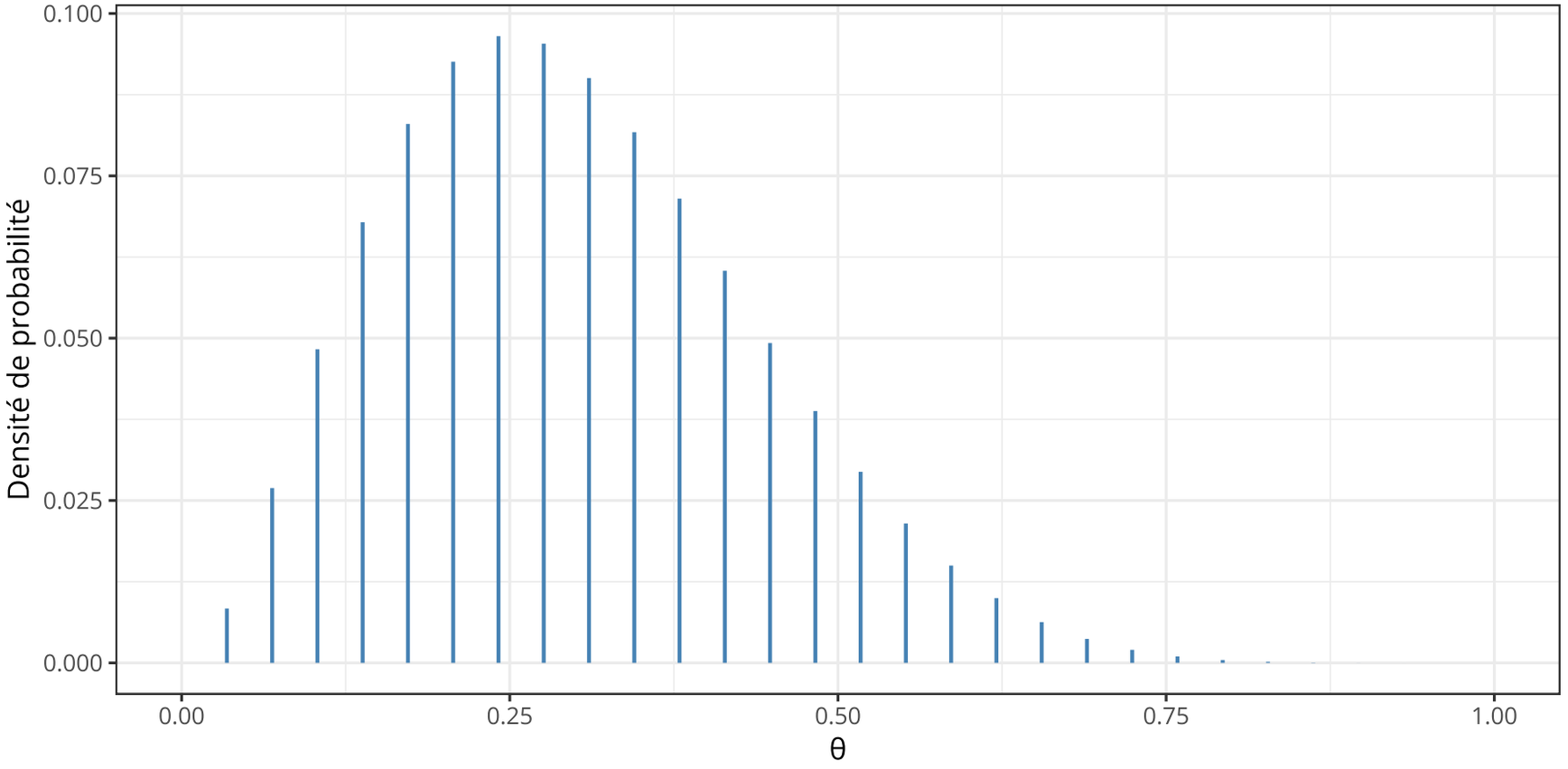
La distribution postérieure, grid method

- **Définir la grille**
- Calculer la valeur du prior pour chaque valeur de la grille
- Calculer la valeur de la vraisemblance pour chaque valeur de la grille
- Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation



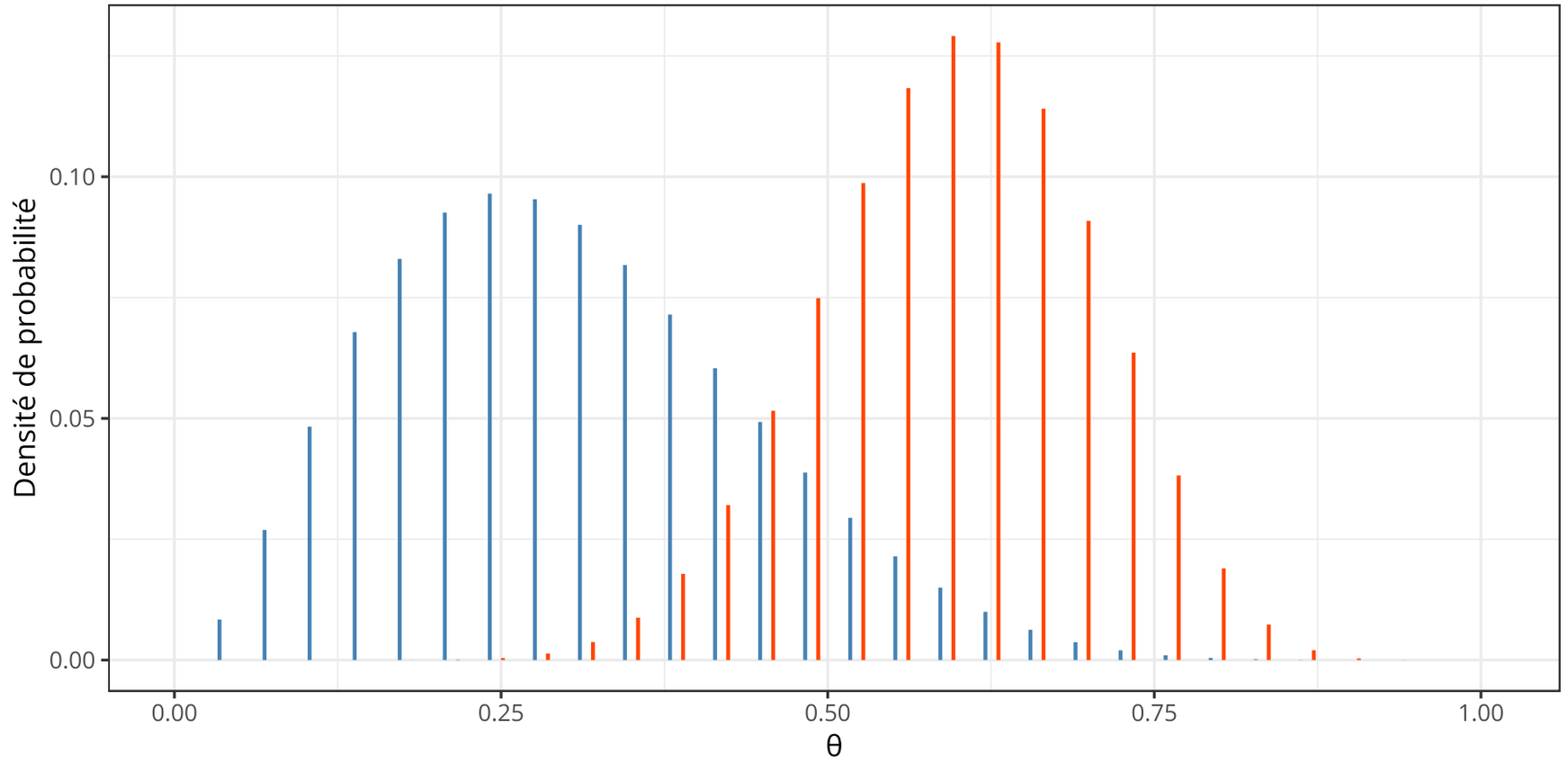
La distribution postérieure, grid method

- Définir la grille
- **Calculer la valeur du prior pour chaque valeur de la grille**
- Calculer la valeur de la vraisemblance pour chaque valeur de la grille
- Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation



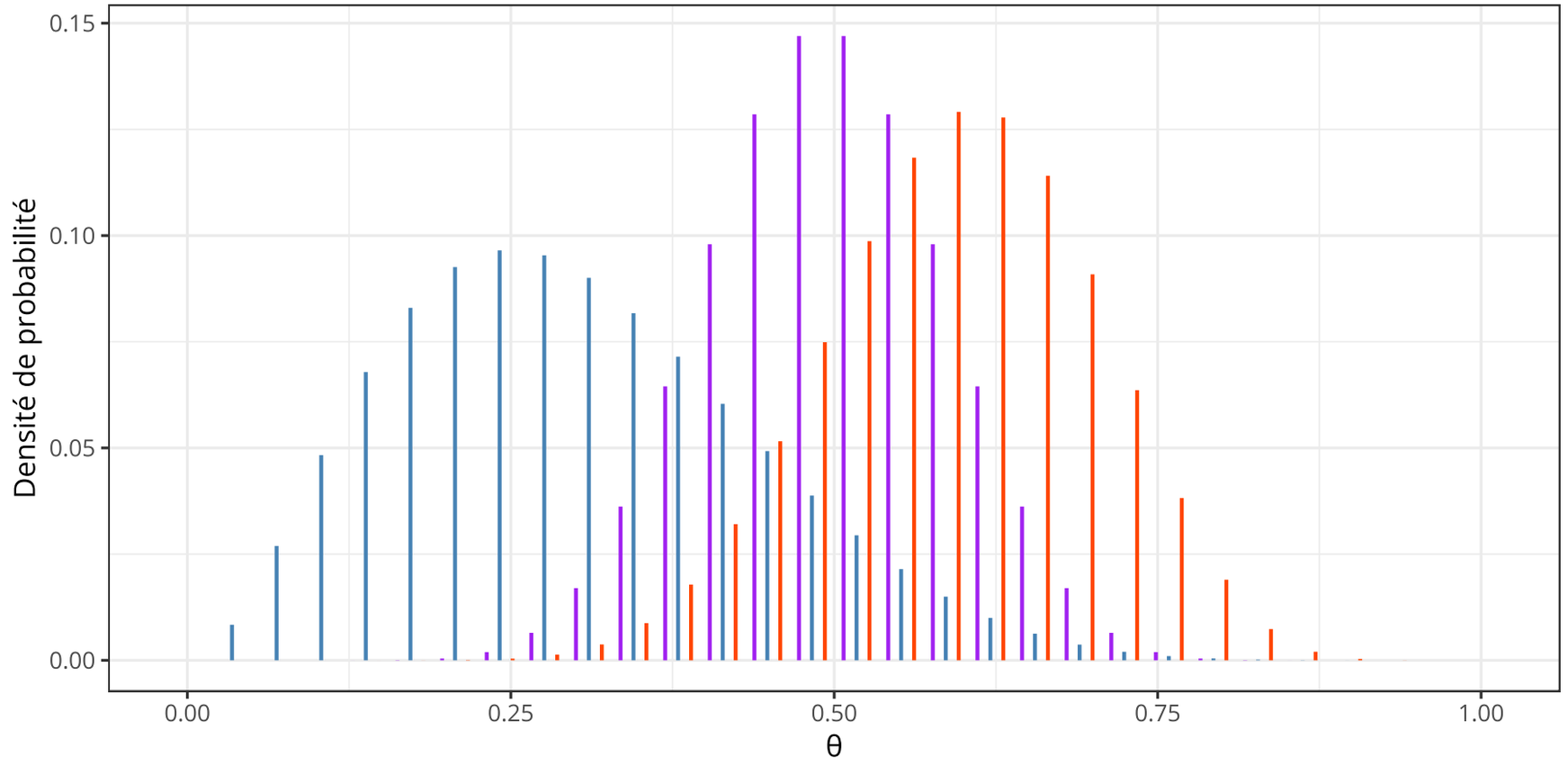
La distribution postérieure, grid method

- Définir la grille
- Calculer la valeur du prior pour chaque valeur de la grille
- **Calculer la valeur de la vraisemblance pour chaque valeur de la grille**
- Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation



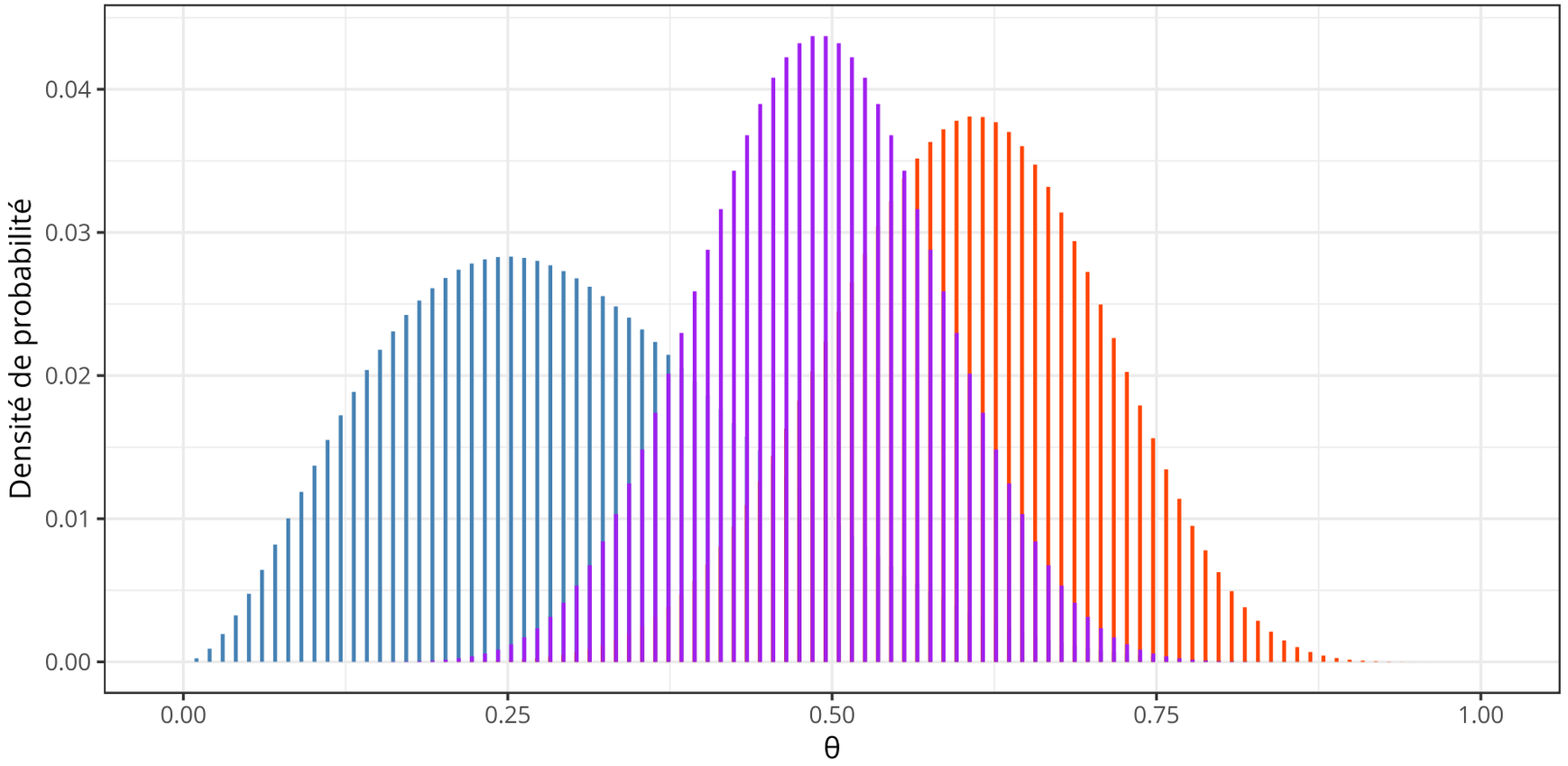
La distribution postérieure, grid method

- Définir la grille
- Calculer la valeur du prior pour chaque valeur de la grille
- Calculer la valeur de la vraisemblance pour chaque valeur de la grille
- **Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation**



La distribution postérieure, grid method

- Définir la grille
- Calculer la valeur du prior pour chaque valeur de la grille
- Calculer la valeur de la vraisemblance pour chaque valeur de la grille
- **Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation**



La distribution postérieure, grid method

Problème du nombre de paramètres... En affinant la grille on augmente le temps de calcul :

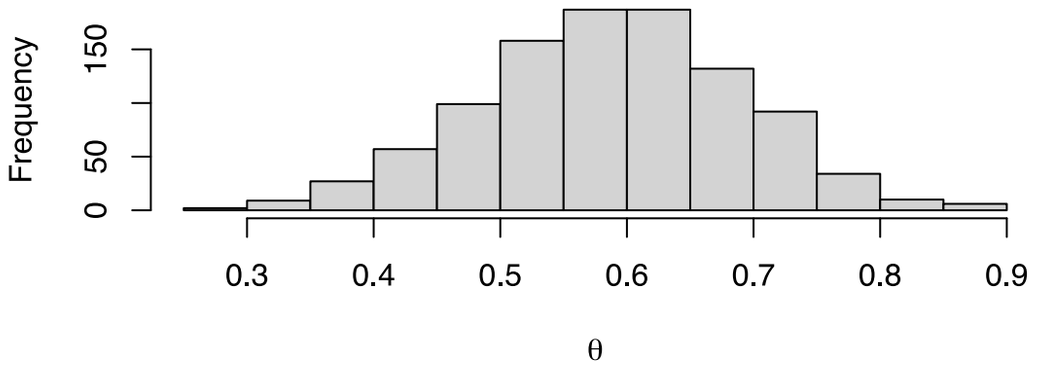
- 3 paramètres avec une grille de 10^3 noeuds = une grille de 10^9 points de calcul
- 10 paramètres avec une grille de 10^3 noeuds = une grille de 10^{30} points de calcul

Le “superordinateur” chinois Tianhe-2 réalise $33,8 \times 10^{15}$ opérations par seconde. Si on considère qu’il réalise 3 opérations par noeud de la grille, il lui faudrait 10^{14} secondes pour parcourir la grille une fois (pour comparaison, l’âge de l’univers est approximativement de $(4,354 \pm 0,012) \times 10^{17}$ secondes)...

Échantillonner la distribution postérieure

Pour échantillonner (de manière intelligente) une distribution postérieure, on peut utiliser différentes implémentations des méthodes MCMC (e.g., Metropolis-Hastings, Hamiltonian Monte Carlo) que l'on discutera au Cours n°05. En attendant, on va travailler avec des échantillons de la distribution postérieure i) pour s'habituer en préparation aux méthodes MCMC et ii) car c'est plus simple de calculer une moyenne ou un intervalle de crédibilité sur des échantillons plutôt qu'en calculant des intégrales.

```r
p_grid <- seq(from = 0, to = 1, length.out = 1000) # creates a grid
prior <- rep(1, 1000) # uniform prior
likelihood <- dbinom(x = 12, size = 20, prob = p_grid) # computes likelihood
posterior <- (likelihood * prior) / sum(likelihood * prior) # computes posterior
samples <- sample(x = p_grid, size = 1e3, prob = posterior, replace = TRUE) # sampling
hist(samples, main = "", xlab = expression(theta) ) # histogram
```



La distribution postérieure, résumé

Cas analytique :

```r
a <- b <- 1 # paramètres du prior Beta
n <- 9 # nombre d'observations
y <- 6 # nombre de succès
p_grid <- seq(from = 0, to = 1, length.out = 1000)
posterior <- dbeta(p_grid, y + a, n - y + b) # plot(posterior)
```

Grid method :

```r
p_grid <- seq(from = 0, to = 1, length.out = 1000)
prior <- rep(1, 1000) # uniform prior
likelihood <- dbinom(x = y, size = n, prob = p_grid)
posterior <- (likelihood * prior) / sum(likelihood * prior) # plot(posterior)
```

Échantillonner la distribution postérieure pour la décrire :

```r
samples <- sample(x = p_grid, size = 1e4, prob = posterior, replace = TRUE) # hist(samples)
```


La distribution postérieure, résumé

Méthode analytique

- La distribution postérieure est décrite explicitement
- Le modèle est fortement contraint

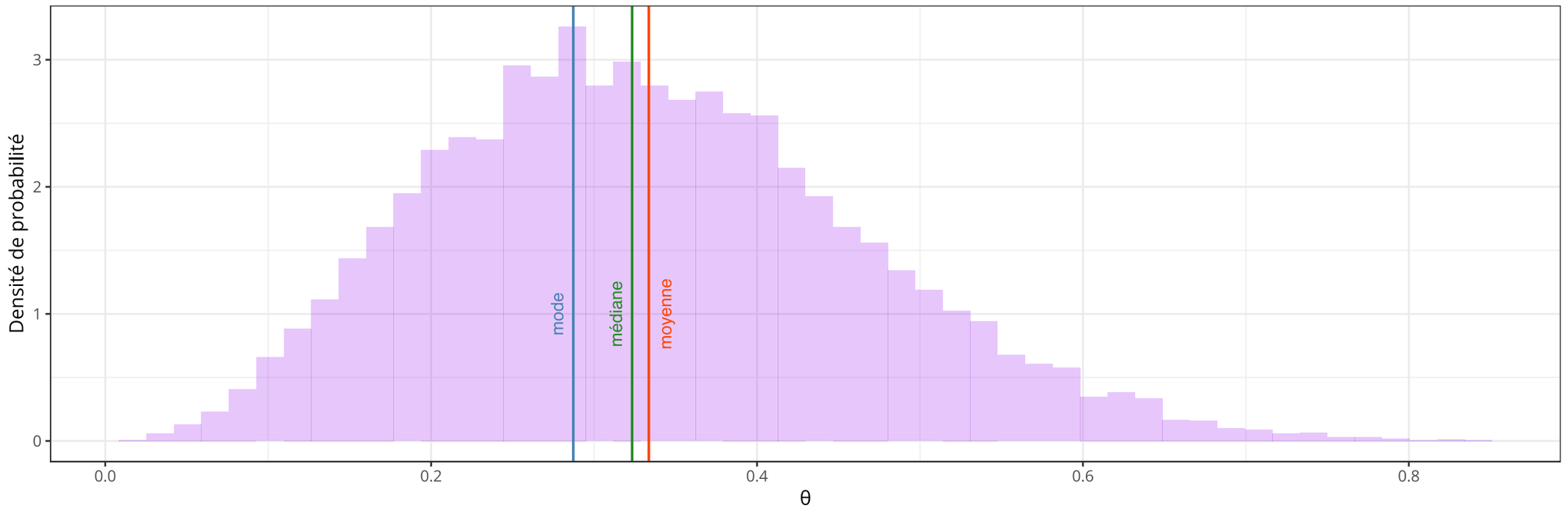
Méthode sur grille

- La distribution postérieure n'est donnée que pour un ensemble fini de valeurs
- Plus la grille est fine, meilleure est l'estimation de la distribution postérieure
- Compromis "Précision - Temps de calcul"

Utiliser les échantillons pour résumer la distribution postérieure

Estimation de la tendance centrale : À partir d'un ensemble d'échantillons d'une distribution postérieure, on peut calculer la moyenne, le mode, et la médiane. Par exemple pour un prior uniforme, 10 lancers, et 3 Faces.

```
1  mode_posterior <- find_mode(samples) # en bleu
2  mean_posterior <- mean(samples) # en orange
3  median_posterior <- median(samples) # en vert
```



Utiliser les échantillons pour résumer la distribution postérieure

Quelle est la probabilité que le biais de la pièce θ soit supérieur à 0.5 ?

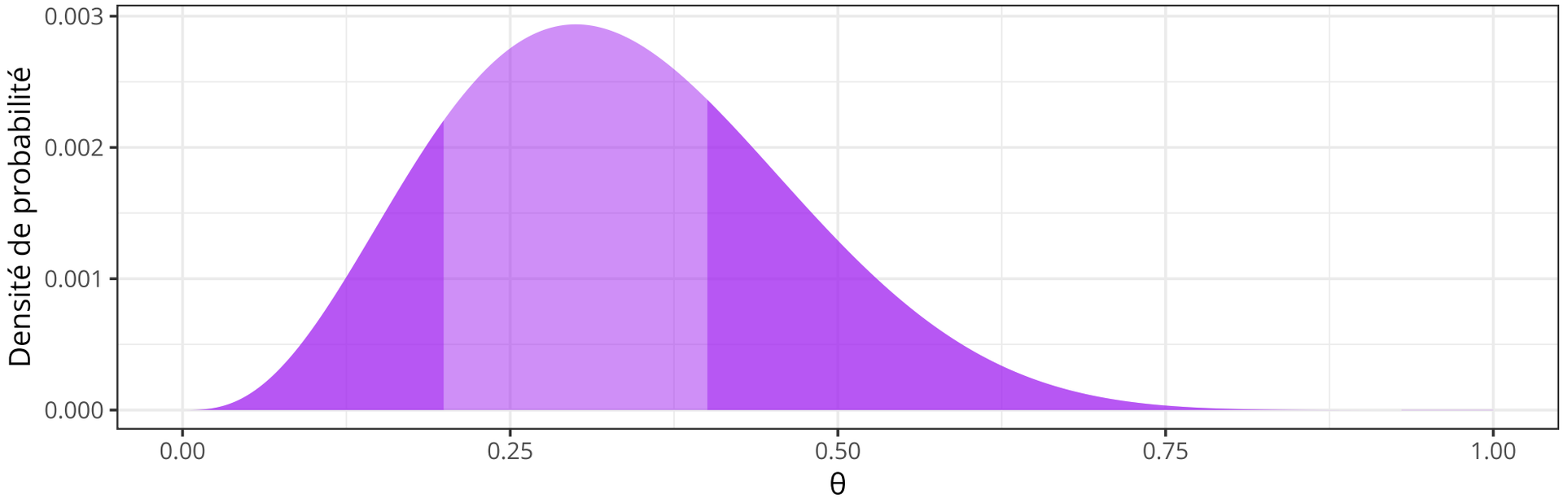
```r
sum(samples > 0.5) / length(samples) # équivalent à mean(samples > 0.5)
```

```
[1] 0.112
```

Quelle est la probabilité que le biais de la pièce θ soit compris entre 0.2 et 0.4 ?

```r
sum(samples > 0.2 & samples < 0.4) / length(samples)
```

```
[1] 0.5482
```



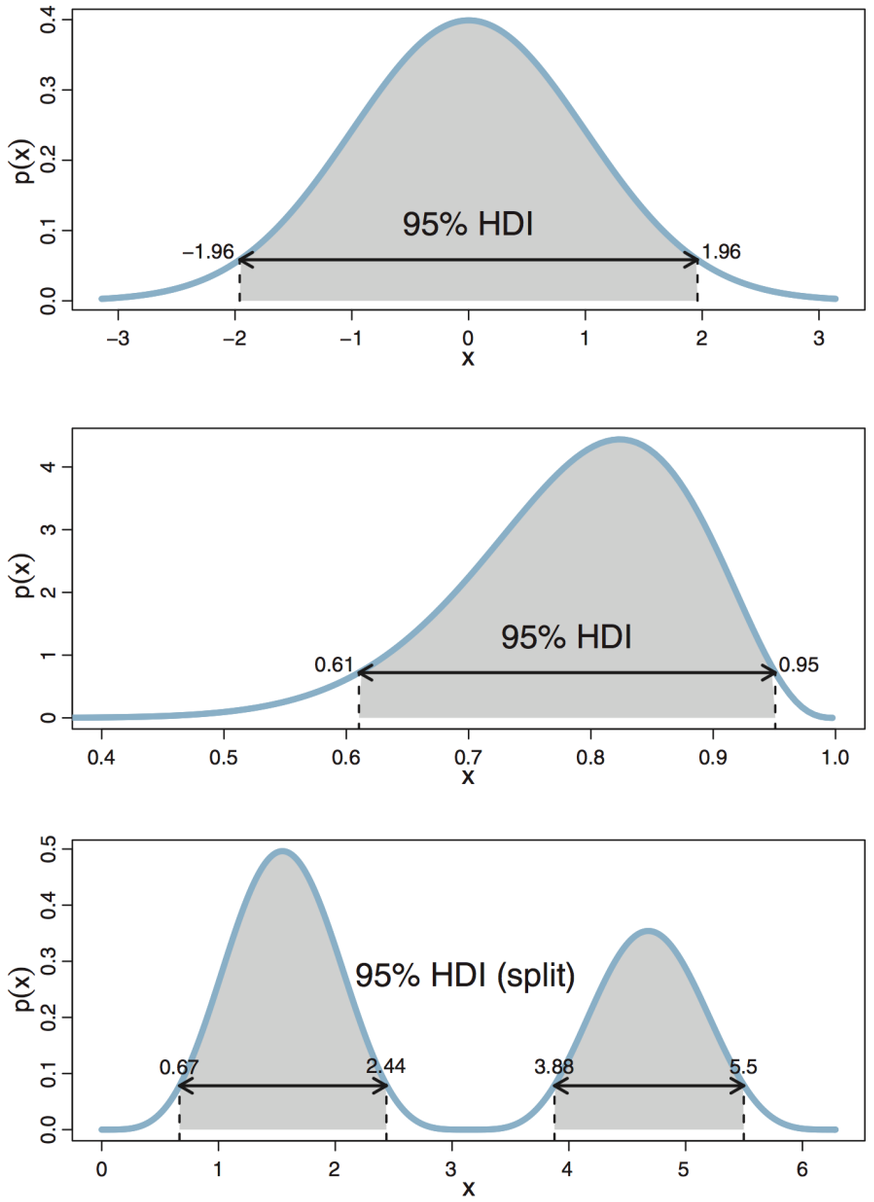
Highest density interval (HDI)

Highest density interval (HDI) :

- Le HDI indique les valeurs du paramètre qui sont les plus probables (sachant les données et le prior)
- Plus le HDI est étroit et plus le degré de certitude est élevé
- La largeur du HDI diminue avec l'augmentation du nombre de mesures

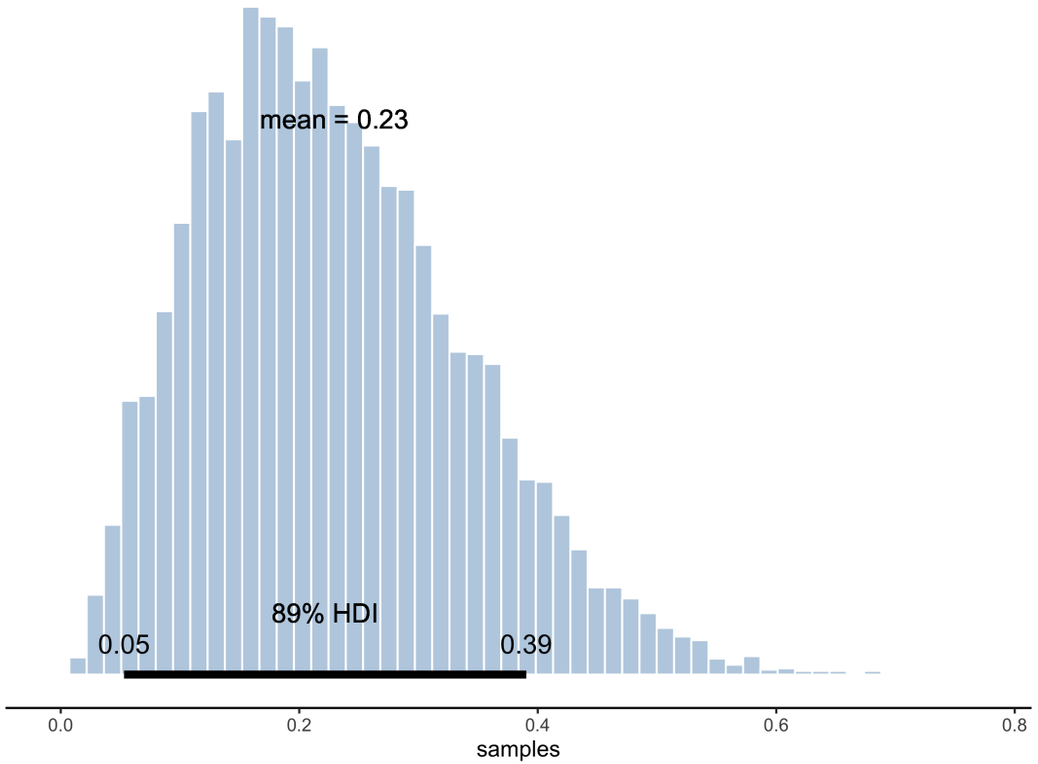
> Définition : les valeurs du paramètre θ contenues dans un HDI à 89% sont telles que $p(\theta) > W$ où W satisfait la condition suivante :
>
> $$\int_{\theta\,:\,p(\theta)>W} p(\theta)\,\mathrm{d}\theta = 0.89.$$

Highest density interval (HDI)



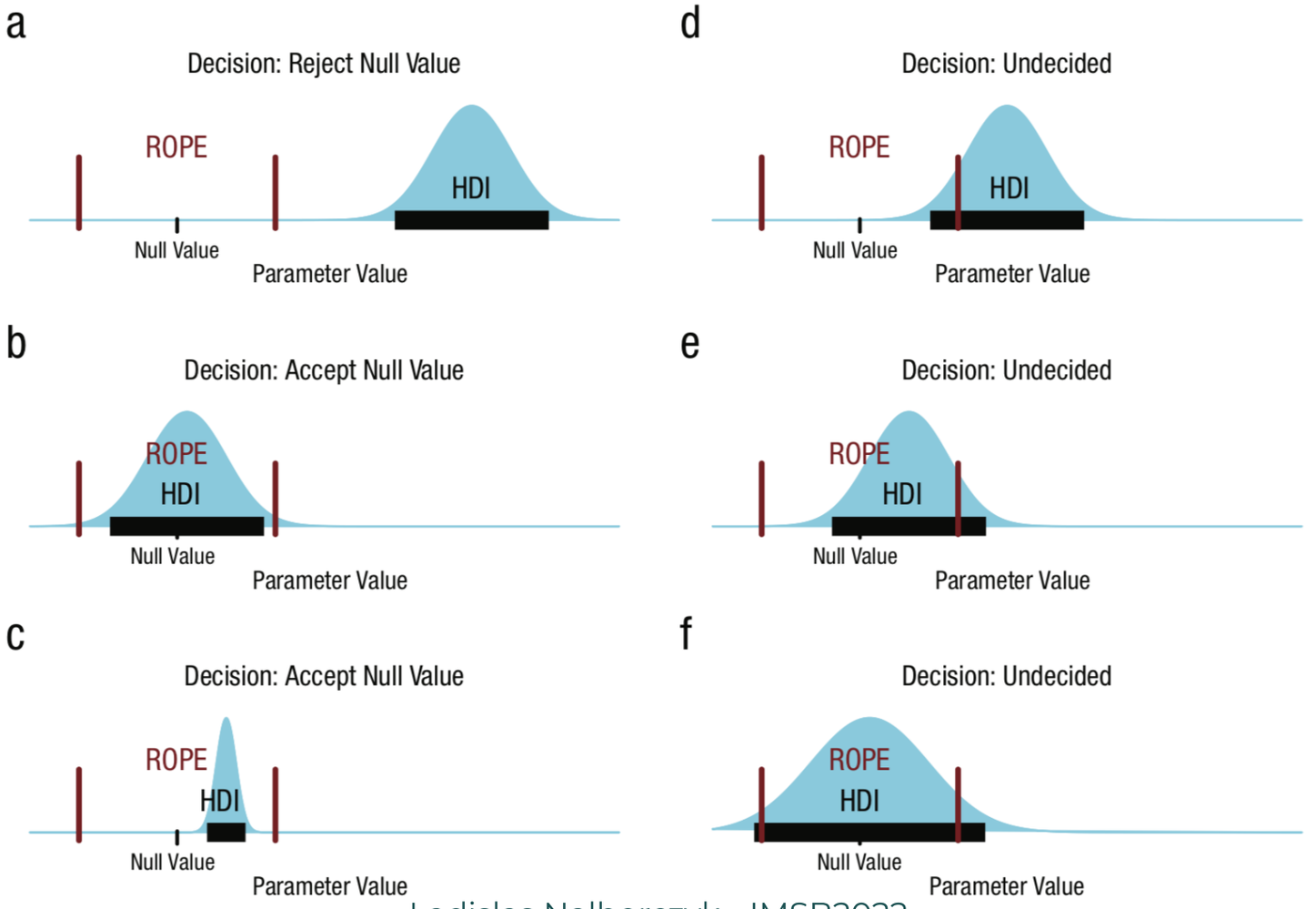
Highest density interval (HDI)

```r
library(imsb)

set.seed(666)
p_grid <- seq(from = 0, to = 1, length.out = 1e3)
pTheta <- dbeta(p_grid, 3, 10)
massVec <- pTheta / sum(pTheta)
samples <- sample(x = p_grid, size = 1e4, replace = TRUE, prob = pTheta)

posterior_plot(samples = samples, credmass = 0.89)
```



Region of practical equivalence (ROPE)

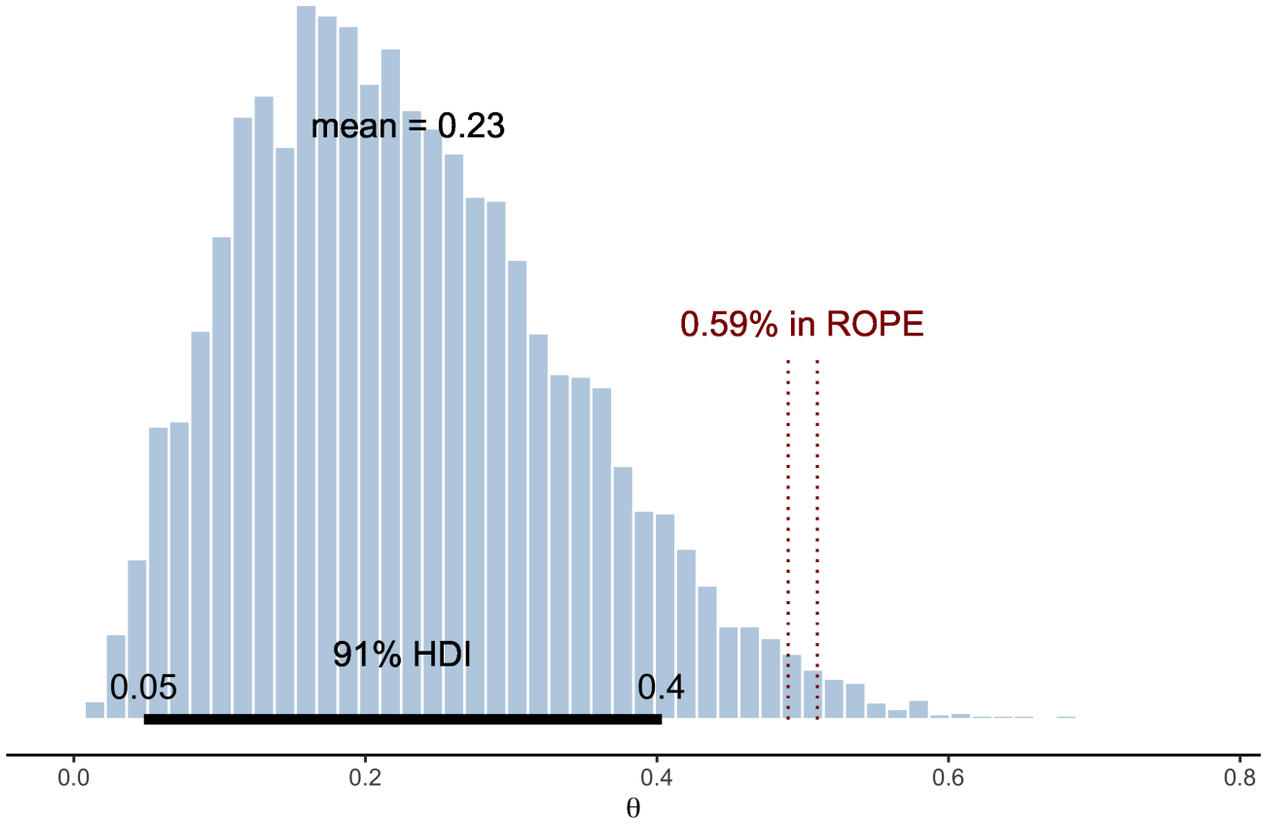
Cette procédure permet d'accepter ou de rejeter une valeur nulle (null value). La région d'équivalence pratique ou **region of practical equivalence** (ROPE) définit un intervalle de valeurs qu'on considère comment étant "équivalentes" à la valeur nulle. La figure ci-dessous résume les décisions possibles issues de cette procédure (Kruschke, 2018).



Region of practical equivalence (ROPE)

La valeur du paramètre (e.g., $\theta = 0.5$) est rejetée si le HDI est entièrement hors de la ROPE. La valeur du paramètre (e.g., $\theta = 0.5$) est acceptée si le HDI est entièrement dans la ROPE. Si le HDI et la ROPE se chevauchent, on ne peut pas conclure...

```
1  posterior_plot(samples = samples, rope = c(0.49, 0.51) ) +
2      labs(x = expression(theta) )
```



Comparaison de modèles

On lance une pièce 200 fois et on obtient 115 “Faces”. Est-ce que la pièce est biaisée ? Nous construisons deux modèles et essayons de savoir lequel rend le mieux compte des données.

$$\begin{cases} \mathcal{M}_0 : Y \sim \text{Binomial}(n, \theta = 0.5) & \text{La pièce n'est pas biaisée} \\ \mathcal{M}_1 : Y \sim \text{Binomial}(n, \theta \neq 0.5) & \text{La pièce est biaisée} \end{cases}$$

Le facteur de Bayes (Bayes factor) fait le rapport des vraisemblances (marginales) des deux modèles.

$$\frac{p(\mathcal{M}_0 \mid \text{data})}{p(\mathcal{M}_1 \mid \text{data})} = \frac{p(\text{data} \mid \mathcal{M}_0)}{p(\text{data} \mid \mathcal{M}_1)} \times \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}$$

Comparaison de modèles

Le facteur de Bayes (Bayes factor) fait le rapport des vraisemblances (marginales) des deux modèles.

$$\frac{p(\mathcal{M}_0 \mid \text{data})}{p(\mathcal{M}_1 \mid \text{data})} = \frac{p(\text{data} \mid \mathcal{M}_0)}{p(\text{data} \mid \mathcal{M}_1)} \times \frac{p(\mathcal{M}_0)}{p(\mathcal{M}_1)}$$

Soit dans notre exemple :

$$\text{BF}_{01} = \frac{p(\text{data} \mid \mathcal{M}_0)}{p(\text{data} \mid \mathcal{M}_1)} = \frac{0.005955}{0.004975} \approx 1.1971.$$

Le rapport de probabilités a augmenté de 20% en faveur de $\mathcal{M}_0$ après avoir pris connaissance des données. Le facteur de Bayes peut également s'interpréter de la manière suivante : Les données sont environ 1.2 fois plus probables sous le modèle $\mathcal{M}_0$ que sous le modèle $\mathcal{M}_1$.

Model checking

Les deux rôles de la fonction de vraisemblance :

- C'est une fonction de θ pour le calcul de la distribution postérieure : $\mathcal{L}(\theta \mid y, n)$
- Lorsque θ est connu / fixé, c'est une distribution de probabilité : $p(y \mid \theta, n) \propto \theta^y (1-\theta)^{(n-y)}$

On peut utiliser cette distribution de probabilité pour générer des données... !

Par exemple : Générer 10.000 valeurs à partir d'une loi binomiale basée sur 10 lancers et une probabilité de Face de 0.6 :

```r
samples <- rbinom(n = 1e4, size = 10, prob = 0.6)
```

Model checking

Dans un modèle bayésien, il existe **deux sources d'incertitude** lorsqu'on génère des prédictions :

- Incertitude liée au processus d'échantillonnage
  -> On tire une donnée issue d'une distribution Binomiale
- Incertitude sur la valeur de θ elle-même
  -> L'incertitude quant à la valeur de θ est représentée par une distribution de probabilité (postérieure)

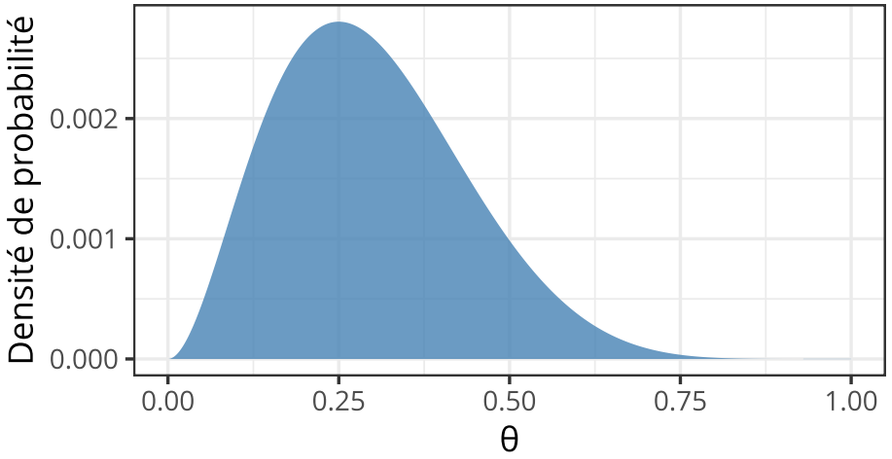
Par exemple : Générer 10000 valeurs à partir d'une loi binomiale basé sur 10 lancers et une probabilité de Face décrite par la distribution postérieure de θ :

```r
posterior <- rbeta(n = 1e4, shape1 = 16, shape2 = 10)
samples <- rbinom(n = 1e4, size = 10, prob = posterior)
```


Prior and posterior predictive checking

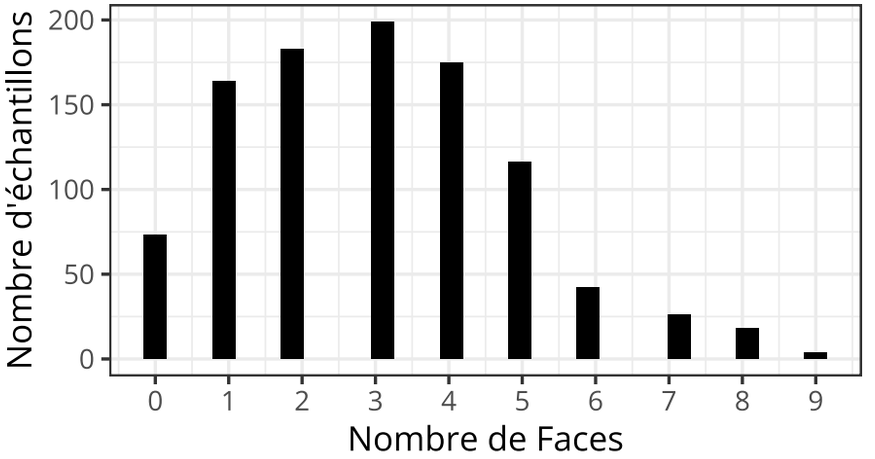
Prior distribution

rbeta(n = 1e4, shape1 = 3, shape2 = 7)



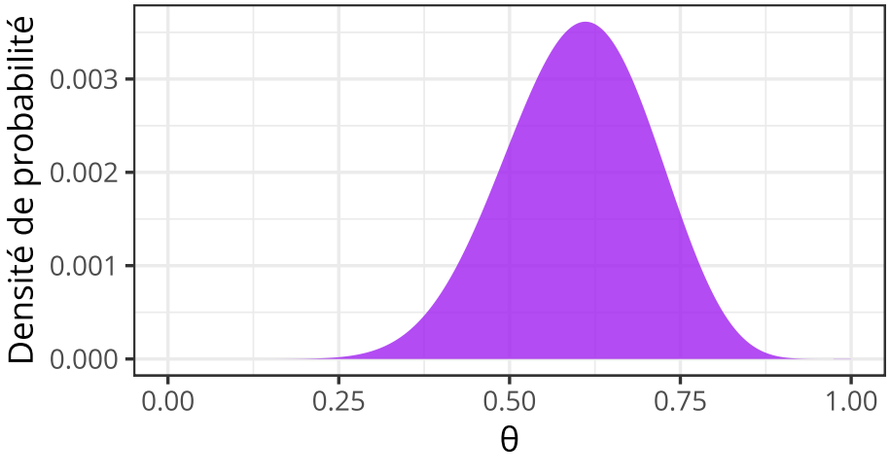
Prior predictive distribution

rbinom(n = 1e4, size = 10, prob = prior)



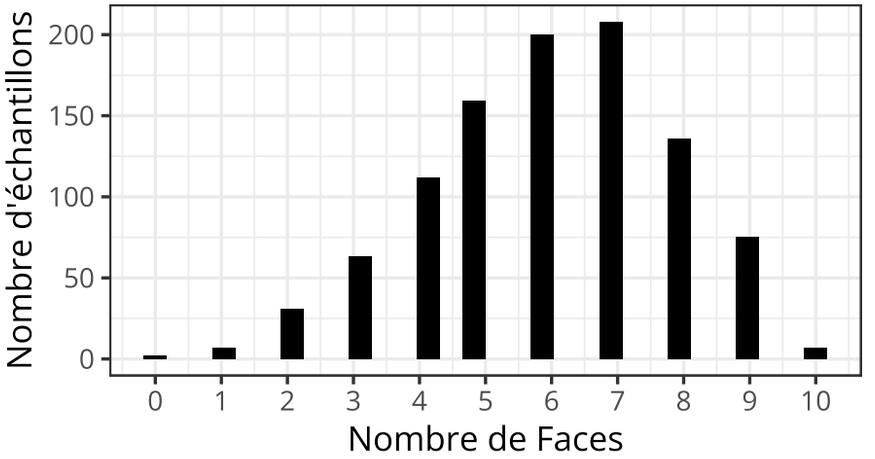
Posterior distribution

rbeta(n = 1e4, shape1 = 12, shape2 = 8)

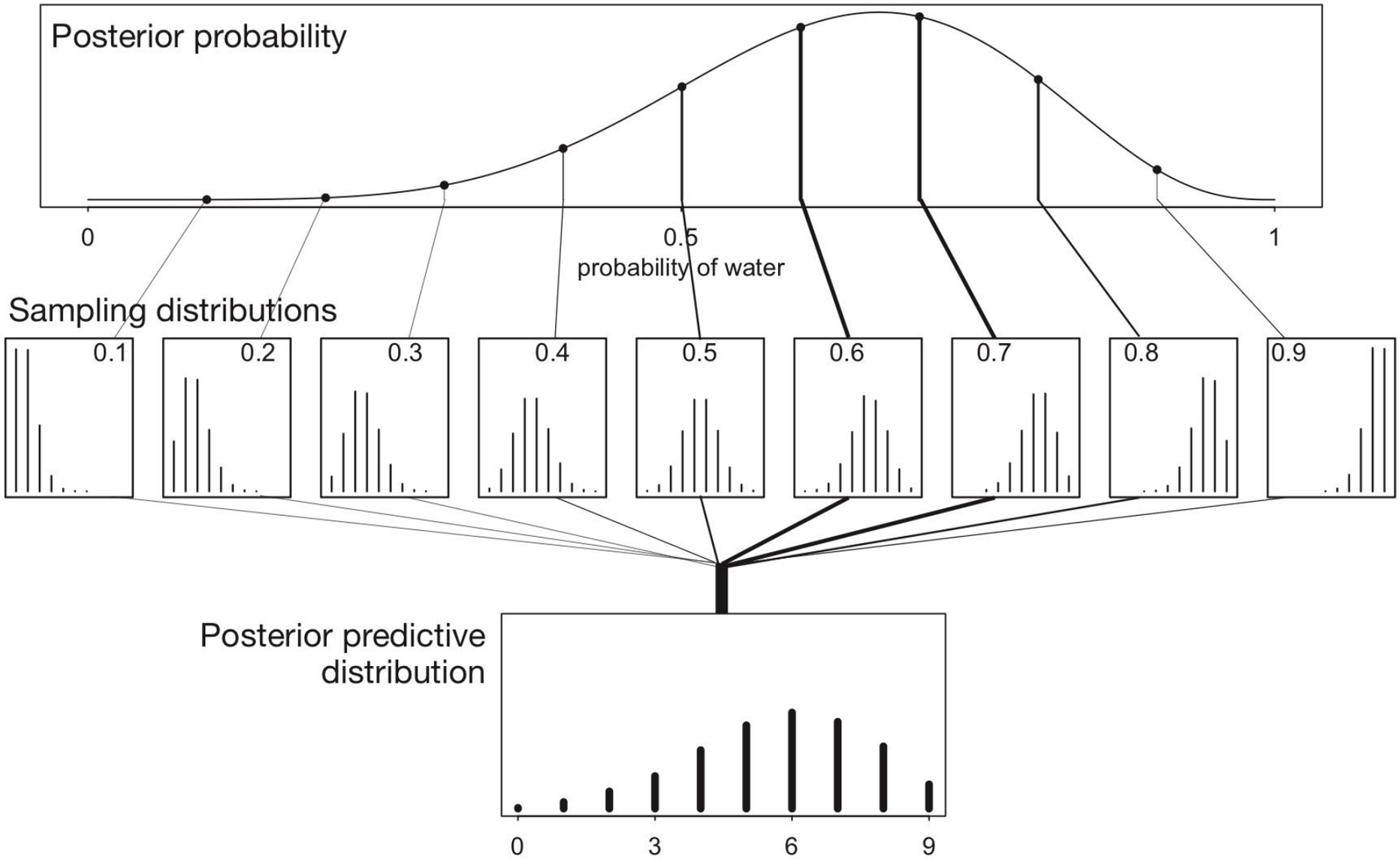


Posterior predictive distribution

rbinom(n = 1e4, size = 10, prob = posterior)



Posterior predictive checking



Travaux pratiques

Un analyste qui travaille dans une fabrique de célèbres petits pains suédois a lu un livre qui soulevait une épineuse question... Pourquoi la tartine tombe toujours du côté du beurre ? À défaut de proposer une réponse plausible, il se propose de vérifier cette assertion. La première expérience qu'il réalise consiste à faire tomber une tartine beurrée de la hauteur d'une table. Les résultats obtenus sont disponibles dans le jeu de données `tartine1` du paquet `imsb`.

Récupérer les données

Première tâche : Récupérer les donées.

```r
# importer les données
data <- open_data(tartine1)

# description sommaire des données
str(data)
```

```
'data.frame':   500 obs. of  2 variables:
 $ trial: int  1 2 3 4 5 6 7 8 9 10 ...
 $ side : int  1 1 0 1 0 0 1 1 1 0 ...
```


Questions

- La tartine n'ayant que deux faces, le résultat s'apparente à un tirage sur une loi binomiale de paramètre θ inconnu. Quelle est la distribution postérieure du paramètre θ au vu de ces données, sachant que l'analyste n'avait aucun a priori (vous pouvez également utiliser vos propres connaissances a priori).
- Calculer le HDI à 95% de la distribution postérieure et en donner une représentation graphique (indice : utilisez la fonction `imsb::posterior_plot()`).
- Peut-on rejeter l'hypothèse nulle selon laquelle $\theta = 0.5$? Répondez à cette question en utilisant la procédure HDI+ROPE.
- Importer les données `tartine2` du paquet `imsb`. Mettre à jour le modèle en utilisant le mode de la distribution postérieure calculée précédemment.

Proposition de solution - Question 1

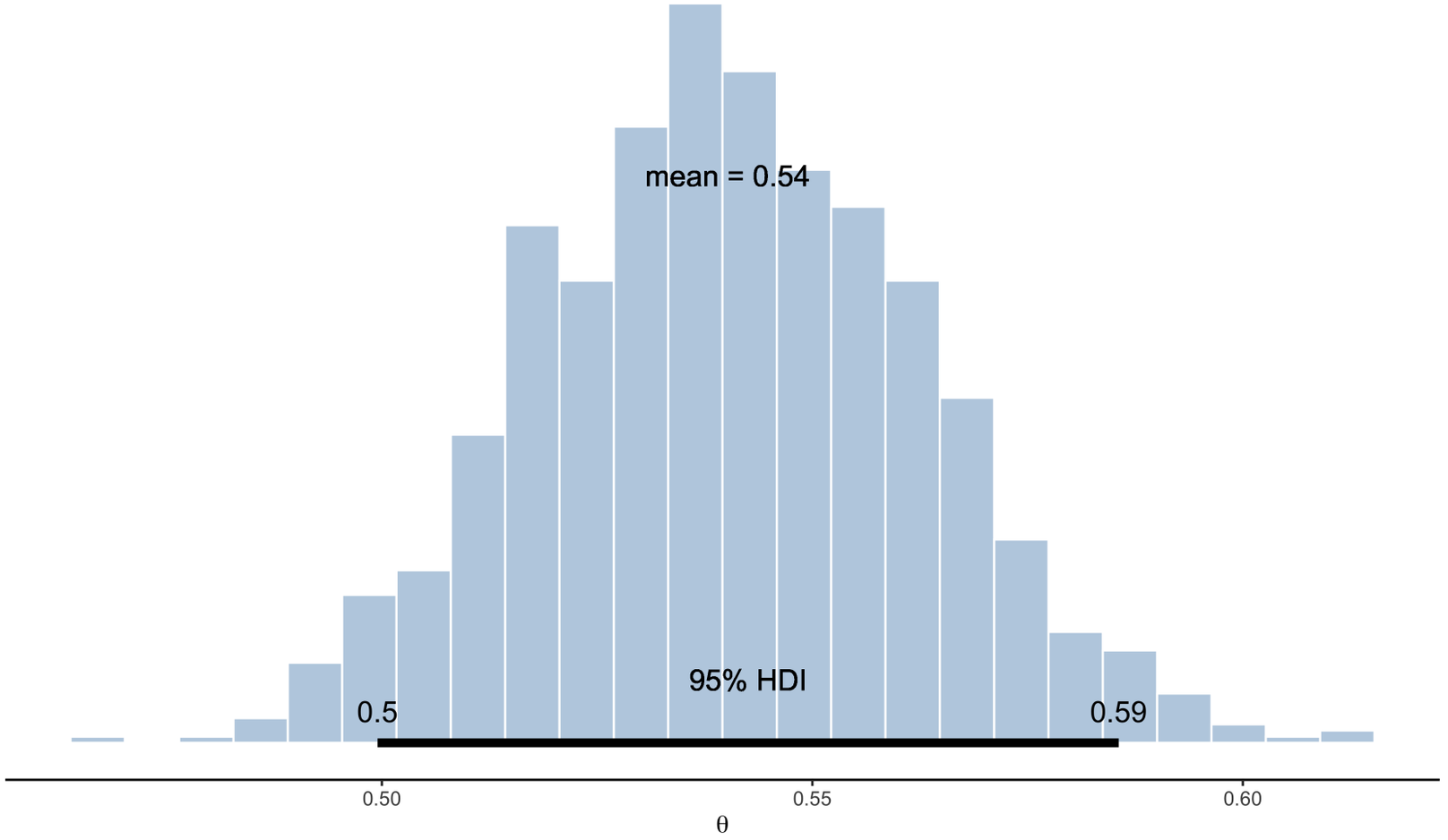
La tartine n'ayant que deux faces, le résultat s'apparente à un tirage sur une loi binomiale de paramètre θ inconnu. Quelle est la distribution postérieure du paramètre θ ?

```r
# nombre d'essais
nbTrial <- length(data$trial)

# nombre de "succès" (i.e., la tartine tombe du côté du beurre)
nbSuccess <- sum(data$side)

# taile de la grille
grid_size <- 1e3

# génère la grille
p_grid <- seq(from = 0, to = 1, length.out = grid_size)

# prior uniforme
prior <- rep(1, grid_size)

# calcul de la vraisemblance
likelihood <- dbinom(x = nbSuccess, size = nbTrial, prob = p_grid)

# calcul du posterior
posterior <- likelihood * prior / sum(likelihood * prior)
```


Proposition de solution - Question 2

Calculer le HDI à 95% de la distribution postérieure et en donner une représentation graphique.

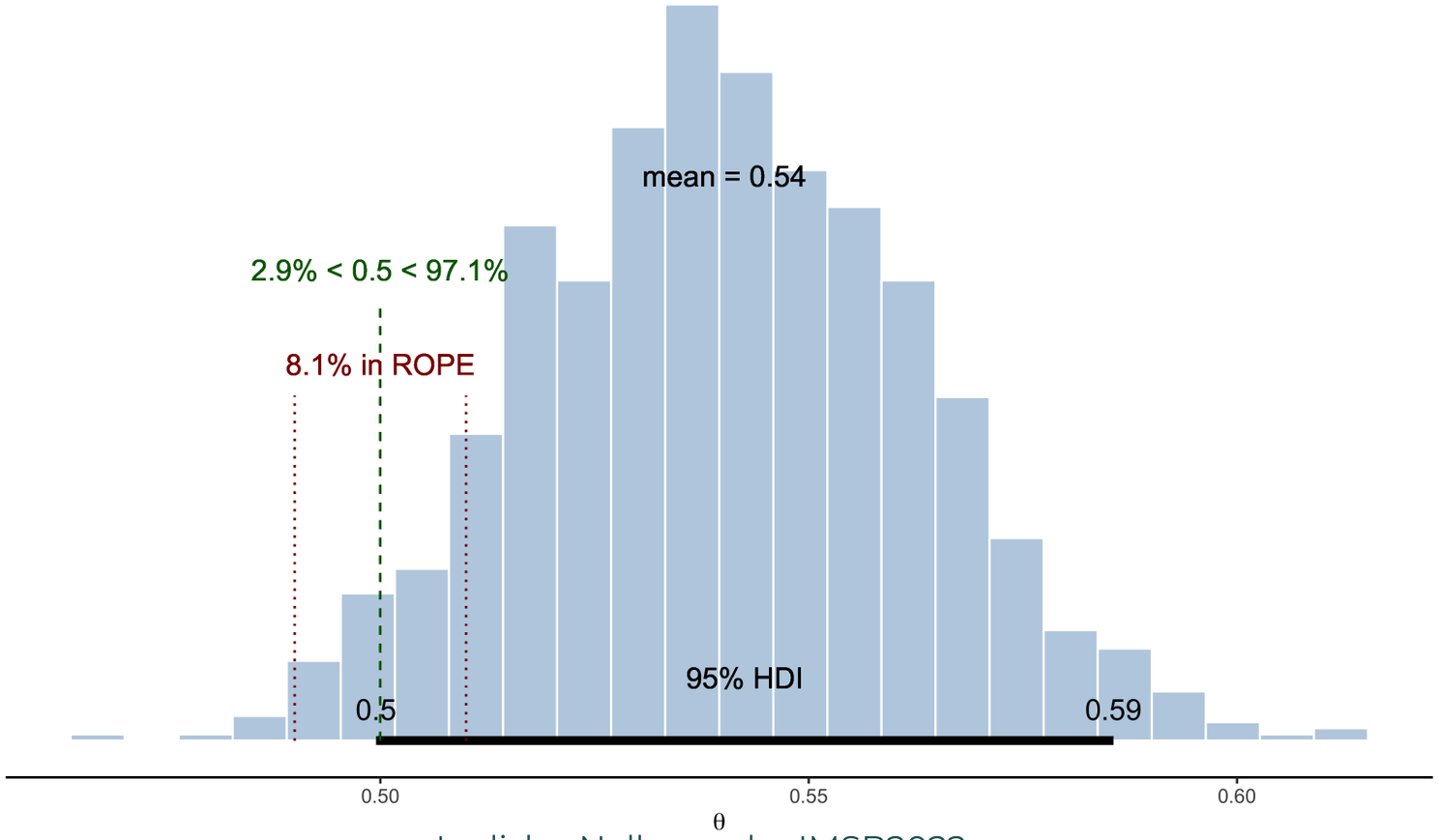
```
samples <- sample(x = p_grid, prob = posterior, size = 1e3, replace = TRUE)
posterior_plot(samples = samples, credmass = 0.95) + labs(x = expression(theta) )
```



Proposition de solution - Question 3

Peut-on rejeter l'hypothèse nulle selon laquelle $\theta = 0.5$? Non, car le HDI recouvre partiellement la ROPE...

```
1  posterior_plot(
2    samples = samples, credmass = 0.95,
3    compval = 0.5, rope = c(0.49, 0.51)
4    ) + labs(x = expression(theta) )
```



Proposition de solution - Question 4

À ce stade, on ne peut pas conclure. L'analyste décide de relancer une série d'observations afin d'affiner ses résultats.

```r
data2 <- open_data(tartine2)
str(data2)
```

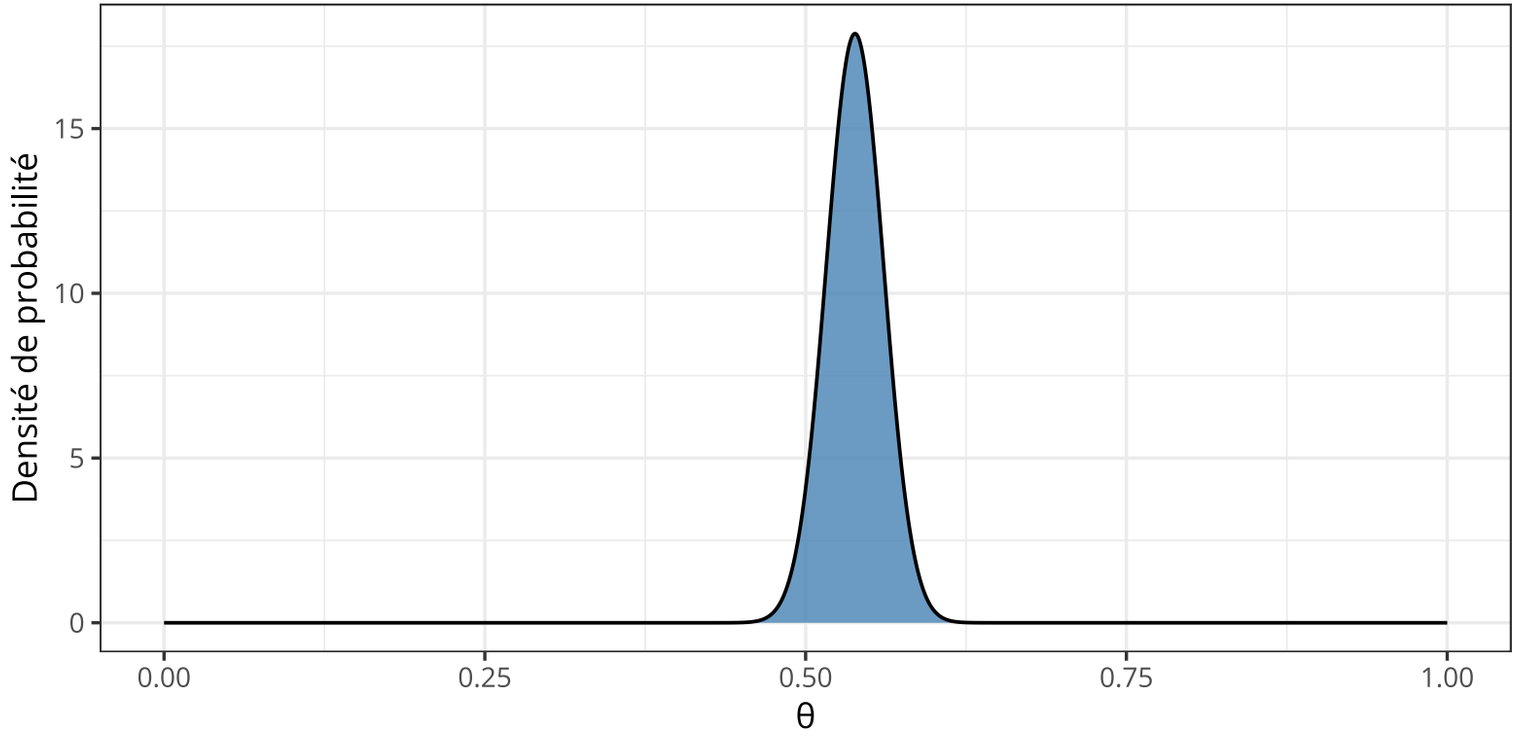
```
'data.frame':   100 obs. of  2 variables:
 $ trial: int  1 2 3 4 5 6 7 8 9 10 ...
 $ side : int  0 0 1 0 0 1 1 1 0 0 ...
```

```r
nbTrial2 <- length(data2$trial) # nombre d'essais
nbSucces2 <- sum(data2$side) # nombre de succès
```


Proposition de solution - Question 4

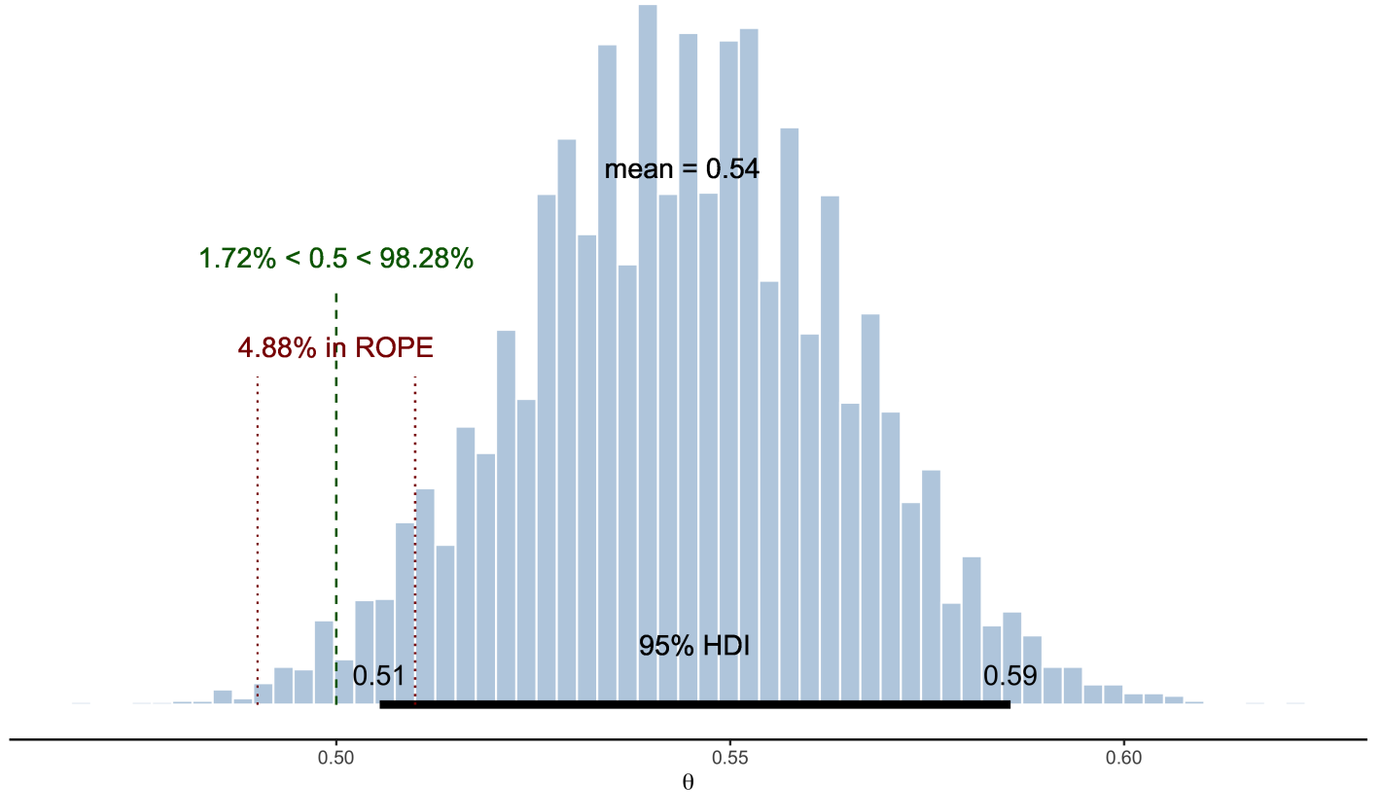
On utilise le posterior précédent comme prior de ce nouveau modèle.

```
mode1 <- find_mode(samples)
prior2 <- dbeta(p_grid, mode1 * (nbTrial - 2) + 1, (1 - mode1) * (nbTrial - 2) + 1)

data.frame(x = p_grid, y = prior2) %>%
  ggplot(aes(x = x, y = y) ) +
  geom_area(alpha = 0.8, fill = "steelblue") +
  geom_line(size = 0.8) +
  labs(x = expression(theta), y = "Densité de probabilité")
```



Proposition de solution - Question 4 (suite)

```r
likelihood2 <- dbinom(x = nbSucces2, size = nbTrial2, prob = p_grid)
posterior2 <- likelihood2 * prior2 / sum(likelihood2 * prior2)
samples2 <- sample(p_grid, prob = posterior2, size = 1e4, replace = TRUE)

posterior_plot(
  samples = samples2, credmass = 0.95,
  compval = 0.5, rope = c(0.49, 0.51)
  ) + labs(x = expression(theta) )
```



Références

Gelman, A., Carlin, J. B., Stern, H., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, third edition*. CRC Press, Taylor & Francis Group.

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kruschke, J. K. (2015). *Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan* (2nd Edition). Academic Press.