

Notes de cours - Introduction à la modélisation statistique bayésienne

Un cours en R avec brms

Ladislas Nalborczyk

Dernière mise à jour : 14-10-2022

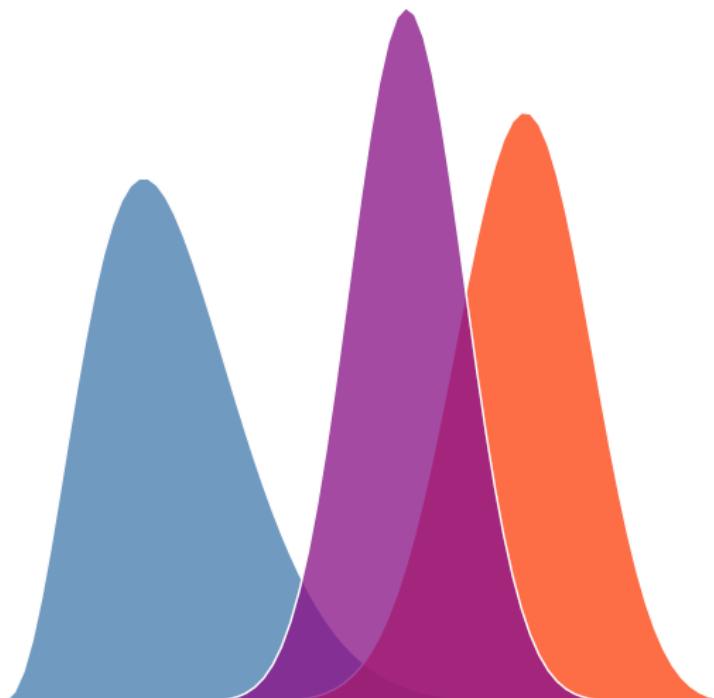


Table des matières

Table des matières	i
Préface	1
1 Introduction à l'inférence bayésienne	3
1.1 Qu'est-ce qu'une probabilité?	4
1.1.1 Axiomes des probabilités	4
1.1.2 Interprétations probabilistes	4
1.2 Logique et raisonnement scientifique	9
1.2.1 Introduction à la logique	9
1.2.2 Quelques syllogismes connus	10
1.2.3 Qu'est-ce qu'une théorie scientifique?	13
1.2.4 Test d'hypothèse nulle et raisonnement scientifique	19
1.2.5 L'approche par comparaison de modèles	21
1.3 Problème du sac de billes (McElreath, 2016b)	24
1.3.1 Énumérer les possibilités	24
1.3.2 Accumulation d'évidence	26
1.3.3 Incorporer un prior	27
1.3.4 Des énumérations aux probabilités	29
1.4 Rappels de théorie des probabilités	30
1.4.1 Probabilité conjointe	31
1.4.2 Probabilité marginale	32
1.4.3 Probabilité conditionnelle	32
1.4.4 Dérivation du théorème de Bayes	33
1.4.5 Loi de probabilité, cas discret	34
1.4.6 Loi de probabilité, cas continu	34
1.4.7 Aparté, qu'est-ce qu'une intégrale?	35
1.4.8 Notations, terminologie	36
1.5 Quelques exemples d'application	38
1.5.1 Diagnostique médical (Gigerenzer, 2002)	38
1.5.2 Problème de Monty Hall	41
2 Modèle beta-binomial	43
2.1 Coefficient binomial	43
2.2 Le modèle Beta-Binomial	43

TABLE DES MATIÈRES

2.2.1	Loi de Bernoulli	44
2.2.2	Processus de Bernoulli	44
2.2.3	Processus de Bernoulli	44
2.2.4	Coefficient binomial	45
2.2.5	Loi binomiale	45
2.2.6	Générer des données à partir d'une distribution binomiale	45
2.2.7	Définition du modèle (likelihood)	47
2.2.8	Vraisemblance versus probabilité	47
2.2.9	Définition du prior	48
2.2.10	La distribution Beta	49
2.2.11	Interprétation des paramètres du prior Beta	49
2.2.12	Prior conjugué	50
2.2.13	Dérivation analytique de la distribution a posteriori	51
2.2.14	Un exemple pour digérer	51
2.2.15	Influence du prior sur la distribution postérieure	52
2.2.16	Ce qu'il faut retenir	52
2.2.17	La vraisemblance marginale (the devil is in the denominator)	53
2.2.18	La distribution postérieure, solution analytique	55
2.2.19	La distribution postérieure, grid method	56
2.2.20	Échantillonner la distribution postérieure	58
2.2.21	La distribution postérieure, résumé	59
2.2.22	Utiliser les échantillons pour résumer la distribution postérieure	60
2.2.23	Highest density interval (HDI)	61
2.2.24	Region of practical equivalence (ROPE)	61
2.2.25	Model checking	62
2.2.26	Posterior predictive checking	64
2.3	Conclusions	65
3	Modèle de régression linéaire	67
3.1	Langage de la modélisation	67
3.2	Un premier modèle	67
3.3	Loi normale	69
3.3.1	D'où vient la loi normale?	70
3.4	Modèle gaussien	72
3.5	Visualiser le prior	74
3.6	Échantillonner à partir du prior	75
3.7	Fonction de vraisemblance	76
3.8	Distribution postérieure	78
3.8.1	Distribution postérieure - grid approximation	78
3.8.2	Distribution postérieure - distributions marginales	79
3.9	Introduction à brms	80
3.9.1	Rappels de syntaxe	81
3.9.2	Quelques fonctions utiles	82

3.9.3	Un premier exemple	83
3.9.4	En utilisant notre prior	83
3.9.5	En utilisant un prior plus informatif	85
3.9.6	Précision du prior (heuristique)	86
3.9.7	Récupérer et visualiser les échantillons de la distribution postérieure . .	86
3.9.8	Récupérer les échantillons de la distribution postérieure	86
3.9.9	Visualiser la distribution postérieure	87
3.9.10	Visualiser la distribution postérieure	88
3.9.11	Ajouter un prédicteur	89
3.10	Régression linéaire à un prédicteur continu	89
3.10.1	Différentes notations équivalentes	90
3.10.2	Représenter les prédictions du modèle	92
3.10.3	Représenter l'incertitude sur μ via fitted()	92
3.10.4	Intervalles de prédiction (incorporer σ)	94
3.10.5	Deux types d'incertitude	95
3.11	Régression polynomiale	95
3.11.1	Modèle de régression polynomiale	97
3.11.2	Représenter les prédictions du modèle	98
3.12	Modèle de régression, taille d'effet	99
3.13	Conclusions	100
4	Modèle de régression linéaire, suite	103
4.1	Régression multiple	103
4.1.1	Associations fortuites	103
4.1.2	Régression multiple	108
4.1.3	Toujours plus de prédicteurs	111
4.1.4	Prédicteurs catégoriels	119
4.1.5	Prédicteurs catégoriels, nombre de catégories > 3	122
4.1.6	Interaction	126
A	Glossaire	131
B	Notations	133
Bibliographie		135

Préface

Ce document regroupe les notes de l'édition 2022 de la formation doctorale "Introduction à la modélisation statistique bayésienne", co-organisée par le collège des écoles doctorales de l'Université Grenoble Alpes et la Maison de la Modélisation et de la Simulation, Nanoscience, et Environnement (MaiMoSiNE).

Ces notes de cours sont organisées en suivant la structure du cours, avec un chapitre par cours, résumant les notions essentielles et donnant quelques éléments de contexte et/ou techniques supplémentaires, pour les plus curieux.

Vous pouvez télécharger la version PDF de ce document en cliquant sur l'icône PDF dans la barre d'outils figurant tout en haut de la version en ligne de ce document. Les slides, données, et codes utilisés sont également disponibles sur le répertoire Github du cours : <https://github.com/lmalborczyk/IMSB2022>.

Le contenu de ce livret est largement inspiré du livre *Statistical rethinking* (McElreath, 2016b, 2020) et des cours partagés librement en ligne par l'auteur. Par ailleurs, le format (HTML) du livret est en partie repris du livre *Introduction to Econometrics with R* (Hanck et al., 2018), disponible [en ligne](#).

Enfin, ce document est diffusé sous une licence *Creative Commons Attribution - Pas d'utilisation commerciale - Partage dans les mêmes conditions* (<https://creativecommons.org/licenses/by-nc-sa/3.0/fr/>). Cela signifie donc que vous êtes libre de recopier / modifier / redistribuer le contenu, à condition que vous citiez la source et que vos modifications soient elles-mêmes distribuées sous la même licence (autorisant ainsi d'autres à pouvoir réutiliser à leur tour vos ajouts).

Introduction à l'inférence bayésienne

"The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning"

Nate Silver

Notre environnement est rempli d'incertitude. Quel temps fera-t-il demain? Qui sera le prochain président de la République? Vais-je apprendre quelque chose en lisant ce livret? Tout individu est capable de se faire une idée intuitive de la réponse à ces questions (avec plus ou moins de succès) sans avoir jamais lu aucun livre traitant formellement de théorie des probabilités. Cependant, un examen plus détaillé des processus menant à ces réponses révèle une complexité et une diversité insoupçonnées. Qu'est-ce qu'une probabilité? À quoi le concept de probabilité réfère-t-il, concrètement, dans le monde? Et surtout, à quoi est-ce que tout cela pourrait bien nous servir dans l'analyse de données expérimentales?

Dans ce premier chapitre, nous allons approfondir cette réflexion sur le concept de probabilité en se basant sur plusieurs définitions ayant été proposées au fil du temps. Puis, nous discuterons plus particulièrement de l'inférence bayésienne, une approche de l'inférence statistique qui utilise les probabilités comme langage pour décrire l'incertitude (et où le concept de *probabilité* est à comprendre dans son acception *épistémique*). Nous proposerons également un bref rappel de théorie des probabilités avant de dériver le théorème de Bayes à partir des règles élémentaires du calcul probabiliste. Le théorème de Bayes est le "moteur" de l'inférence statistique bayésienne, permettant de mettre à jour un état de connaissance a priori (i.e., avant d'observer certaines données) en un état de connaissance a posteriori (i.e., après avoir observé ces données). Nous illustrerons ce mécanisme par plusieurs exemples concrets.

1.1 Qu'est-ce qu'une probabilité?

1.1.1 Axiomes des probabilités

Pour quantifier notre incertitude vis à vis de la survenue de certains *événements* (e.g., obtenir un chiffre pair lors d'un lancer de dé), on assigne des *probabilités* à ces événements. On définit une probabilité comme une valeur numérique assignée à un *événement* A , compris comme une possibilité appartenant à l'univers Ω (l'ensemble de tous les événements possibles). Les probabilités telles que définies et utilisées dans ce cours se conforment aux axiomes suivants ([Kolmogorov, 1933](#))¹ :

- **Axiome n°1** : $\Pr(A) \geq 0$. La probabilité d'un événement A ne peut pas être négative.
- **Axiome n°2** : $\Pr(\Omega) = 1$. La somme des probabilités de tous les événements possibles est égale à 1 (et donc chaque probabilité individuelle ne peut dépasser 1).
- **Axiome n°3** : $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2)$. La probabilité d'obtenir *soit* l'événement A_1 , *soit* l'événement A_2 (sachant que ces événements sont *incompatibles*) est égale à la somme de la probabilité de chacun de ces deux événements.

Le dernier axiome est également connu comme la **règle de la somme** et n'est valide dans cette forme que pour des événements deux à deux *incompatibles* ou *mutuellement exclusifs*.² Il se généralise à des événements non mutuellement exclusifs de la manière suivante : $\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2)$. Autrement dit et pour résumer, une probabilité est une valeur numérique positive, bornée entre 0 et 1, et qui respecte la règle de la somme.

1.1.2 Interprétations probabilistes

Bien que nous ayons donné ci-dessus une définition des probabilités, cela ne nous donne aucune indication sur la manière d'interpréter ce genre de valeur. Qu'est-ce qu'une probabilité? Quelle genre de chose est une probabilité? À quoi dans le monde fait référence une probabilité? Cette question fait encore aujourd'hui couler beaucoup d'encre en philosophie de la connaissance. Nous discutons ci-dessous brièvement de quelques interprétations possibles du concept de probabilité.

1.1.2.1 Interprétation classique (ou théorique)

La première interprétation proposée est généralement la première définition que nous rencontrons dans notre cursus scolaire. Il s'agit également d'une interprétation précédent l'observation de données. En effet, dans ce cadre, une probabilité peut être calculée avant même d'avoir observé quelque donnée que ce soit, par simple connaissance du système étudié. Plus précisément, on définit la probabilité comme le rapport entre le nombre de cas favorables sur

¹On notera au passage que les axiomes de Kolmogorov représentent un exemple parmi d'autres d'ensemble de règles permettant de définir ce qu'est une probabilité, mais que ce n'est pas le seul (bien que ce soit le plus communément utilisé). Par exemple, les [axiomes de Cox](#) offrent un cadre alternatif.

²Deux événements A_1 et A_2 sont dits *incompatibles* si l'on ne peut avoir les deux en même temps, c'est à dire si $\Pr(A_1 \cap A_2) = 0$.

le nombre de cas possibles. Par exemple, si on s'intéresse à la probabilité de l'événement “Obtenir un chiffre pair” lors du lancer d'un dé (non pipé), alors cette probabilité peut se calculer de la manière suivante :

$$\Pr(\text{pair}) = \frac{\text{nombre de cas favorables}}{\text{nombre de cas possibles}} = \frac{3}{6} = \frac{1}{2}.$$

Cette définition fonctionne bien pour les situations dans lesquelles il n'y a qu'un nombre **fini** de résultats possibles **équiprobables** (i.e., de même probabilité). Cependant, si on applique cette définition telle quelle à des situations plus complexes, on se rend compte que sa portée est limitée. Par exemple, si on applique cette définition à la question suivante : “Quelle est la probabilité qu'il pleuve demain?”, on se retrouve avec le calcul suivant.

$$\Pr(\text{pluie}) = \frac{\text{pluie}}{\{\text{pluie, non-pluie}\}} = \frac{1}{2}$$

Et on se rend compte assez facilement que cette définition ne s'applique pas à des situations de prédiction météorologique, où il peut exister un grand nombre d'évènements possibles n'ayant pas nécessairement la même probabilité.

1.1.2.2 Interprétation fréquentiste (ou empirique)

L'interprétation fréquentiste du concept de probabilité propose que la probabilité est ce vers quoi tend le rapport présenté dans la section précédente lorsque le nombre d'essais tend vers l'infini (i.e., lorsque le nombre d'essais devient très important). Autrement dit, la probabilité est définie de la manière suivante :

$$\Pr(x) = \lim_{n_t \rightarrow \infty} \frac{n_x}{n_t}$$

où n_x est le nombre d'occurrences de l'événement x et n_t le nombre total d'essais. L'interprétation **fréquentiste** postule que, à long-terme (i.e., quand le nombre d'essais s'approche de l'infini), la fréquence relative va converger *exactement* vers ce qu'on appelle “probabilité”.

```
library(tidyverse)

sample(c(0, 1), 500, replace = TRUE) %>%
  data.frame %>%
  mutate(x = seq_along(.), y = cumsum(.) / seq_along(.) ) %>%
  ggplot(aes(x = x, y = y), log = "y") +
  geom_line(lwd = 1) +
  geom_hline(yintercept = 0.5, lty = 3) +
  xlab("Nombre de lancers") +
  ylab("Proportion de faces") +
  ylim(0, 1) +
  theme_bw(base_size = 12)
```

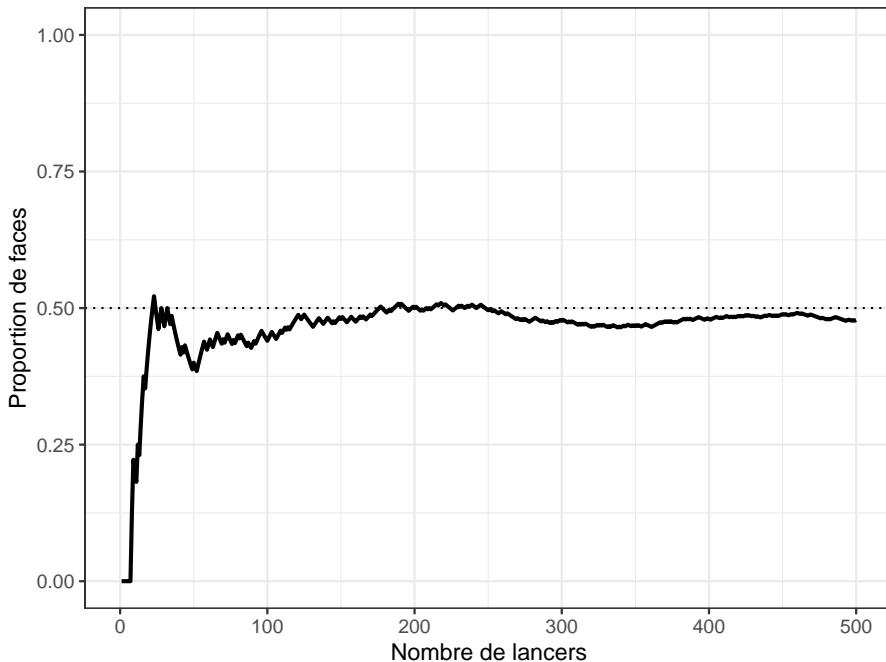


Figure 1.1 : Illustration de l’interprétation fréquentiste du concept de probabilité. Lorsque le nombre d’essais augmente (en abscisse), la fréquence relative (en ordonnée) converge vers la probabilité d’obtenir Face.

La Figure 1.1 illustre l'**interprétation fréquentiste** du concept de probabilité en montrant que la fréquence relative converge vers une fréquence donnée (la “probabilité”) quand le nombre d’essais augmente. Une conséquence importante de cette définition est que le concept de probabilité s’applique uniquement aux **collectifs** (i.e., aux séquences), et non aux événements singuliers. La probabilité est définie comme la limite d’une fréquence relative, et ne permet pas de parler de la probabilité d’un événement unique.

L’interprétation fréquentiste rencontre d’autres problèmes, comme celui de la classe de référence. Considérons par exemple la question suivante : “Quelle est la probabilité que je vive jusqu’à 80 ans ?” Pour répondre à cette question, nous avons besoin de définir la classe de référence à partir de laquelle l’individu examiné (“je”) provient. Autrement dit, je peux estimer la probabilité qu’un individu ayant mes caractéristiques vive jusqu’à 80 ans, mais il faut d’abord déterminer quelles sont les caractéristiques importantes au regard de la question posée. Selon mon sexe biologique, ma nationalité, ou mon statut socio-économique, la réponse à cette question peut varier drastiquement. L’interprétation fréquentiste ne propose pas de procédure stricte permettant de définir la classe de référence pertinente.

Par ailleurs, cette définition ne s’applique pas directement aux événements qui ne peuvent pas se répéter. Par exemple, quelle est la probabilité que j’apprenne quelque chose pendant cette formation ? La réponse à cette question ne peut s’établir qu’en considérant des facteurs extérieurs à la question (e.g., le niveau de connaissance a priori), et ne peut pas reposer sur une évaluation sur le long-terme de la fréquence d’occurrence de l’événement “apprendre quelque chose” (cela n’aurait pas de sens de refaire la formation 1000 fois pour calculer le nombre de fois où on aura appris quelque chose).

Une autre limite importante de l'interprétation fréquentiste est la question de la résolution (ou précision) du calcul de cette probabilité. À partir de combien de lancers (d'une pièce par exemple) a-t-on une bonne approximation de la probabilité? On sait qu'une classe finie d'événements de taille n ne peut produire que des fréquences relatives de précision $1/n$. Jusque quand devrions-nous donc continuer "d'échantillonner le long-terme" avant d'avoir une estimation précise de la probabilité?

1.1.2.3 Interprétation propensionniste

Selon l'interprétation propensionniste du concept de probabilité, les propriétés fréquentistes (i.e., à long terme) des objets (e.g., une pièce) seraient provoquées par des *propriétés physiques intrinsèques* aux objets. Par exemple, une pièce biaisée va engendrer une fréquence relative (et donc une probabilité, selon l'interprétation fréquentiste) biaisée en raison de ses propriétés physiques. Pour les propensionnistes, les probabilités représentent ces caractéristiques intrinsèques, ces **propensions** à générer certaines fréquences relatives, et non les fréquences relatives en elles-mêmes.

Une conséquence intéressante de cette définition (et un progrès par rapport à l'interprétation fréquentiste) est que ces propriétés sont les propriétés d'événements individuels... et non de séquences! L'interprétation propensionniste nous permet donc de parler de la probabilité d'événements uniques.

1.1.2.4 Interprétation logique

L'interprétation logique du concept de probabilité, comme l'interprétation classique, postule que les probabilités peuvent être déterminées a priori en examinant les caractéristiques du système étudié (e.g., les caractéristiques de l'objet, l'ensemble des événements possibles, etc). Cependant, contrairement à l'interprétation classique, l'interprétation logique permet de rendre compte des événements non équiprobables. Cette interprétation se propose de réaliser cet objectif en généralisant la logique binaire (vrai / faux) au monde probabiliste, en étudiant le degré de support logique fourni par un ensemble de preuves pour une hypothèse donnée. Considérons par exemple l'argument logique ci-dessous.

Prémisse n°1 : Considérons une salle dans laquelle sont présents 10 étudiants

Prémisse n°2 : Neuf étudiants portent un t-shirt vert

Prémisse n°3 : Un étudiant porte un t-shirt rouge

Prémisse n°4 : Une personne est tirée au sort...

Conclusion n°1 : L'étudiant tiré au sort porte un t-shirt

Cette conclusion est *vraie* et l'argument qui y est attachée est *valide*. Pour rappel, on dit qu'un argument est *valide* lorsqu'il n'existe aucune situation logiquement possible dans laquelle tous les prémisses de l'argument soient vrais et sa conclusion fausse (Talbot, 2015). Autrement

dit, si on considère les prémisses 1 à 4 comme vrais, alors il est (logiquement) impossible pour cette conclusion d'être fausse.

Conclusion n°2 : L'étudiant tiré au sort porte un t-shirt rouge

Cette conclusion est *fausse* et l'argument qui y est attachée est *invalid*e.

Conclusion n°3 : L'étudiant tiré au sort porte un t-shirt vert

Cette conclusion est également *fausse* et l'argument qui y est attaché est *invalid*e. Cependant, on pourrait dire intuitivement que cette conclusion est "un peu moins fausse" que la conclusion précédente (je m'excuse en avance pour les logiciens qui me lisent), dans le sens où elle est "plus fortement" impliquée par les prémisses 1 à 4 que la conclusion n°2.

Bien que les règles de la logique formelle n'autorisent pas des conclusions à être "plus ou moins vraies", l'interprétation logique du concept de probabilité cherche justement à étendre les règles de la logique aux événements continus, et se propose d'utiliser le langage des probabilités dans ce but. Autrement dit, la probabilité représente donc le *degré de support logique* qu'une conclusion peut avoir, relativement à un ensemble de prémisses (Carnap, 1950; Keynes, 1921).

Une conséquence intéressante de cette interprétation est que toute probabilité est **conditionnelle**, que ce soit à de l'information a priori ou, par exemple, à un ensemble de prémisses.

1.1.2.5 Interprétation bayésienne

Selon l'interprétation bayésienne (subjective), la probabilité est **une mesure du degré de croyance** (ou *crédence*) ou d'*incertitude*. Un événement *certain* aura donc une probabilité de 1 et un événement *impossible* aura une probabilité de 0.

So to assign equal probabilities to two events is not in any way an assertion that they must occur equally often in any "random experiment"; as Jeffrey emphasized, it is only a formal way of saying "I don't know" (Jaynes, 1986).

Pour parler de probabilités, dans ce cadre, nous n'avons donc plus besoin de nous référer à la limite d'occurrence d'un événement (à sa fréquence). La probabilité est un concept abstrait faisant référence à un état de connaissance et / ou permettant de quantifier l'incertitude liée à cet état de connaissance.

1.1.2.6 Interprétations probabilistes - résumé

Pour résumer, les différentes interprétations discutées ci-dessus peuvent être classées dans deux grandes catégories :

Interprétation épistémique : toute probabilité est conditionnelle à de l'information disponible (e.g., prémisses ou données). La probabilité est utilisée comme moyen de quantifier l'incertitude.

Interprétation logique (e.g., Keynes, Carnap), interprétation bayésienne (e.g., Jeffreys, de Finetti, Savage).

Interprétation physique : les probabilités dépendent d'un état du monde, de caractéristiques physiques, elles sont indépendantes de l'information disponible (ou de l'incertitude).

Interprétation classique (e.g., Laplace, Bernouilli, Leibniz), interprétation fréquentiste (e.g., Venn, Reichenbach, von Mises).

Les plus curieux d'entre vous seront ravis de trouver plus d'informations dans cet excellent article de la *Stanford Encyclopedia of Philosophy* ([Hájek, 2019](#)) sur les différentes interprétations du concept de probabilité.

1.2 Logique et raisonnement scientifique

1.2.1 Introduction à la logique

Le but de cette section est de proposer une courte (et très incomplète) introduction à la logique, afin de pouvoir analyser dans une section ultérieure l'argument central de l'inférence fréquentiste et d'illustrer les similarités entre la logique et le fonctionnement de l'inférence bayésienne. Commençons tout d'abord commencer par définir les termes utilisés (cf. [Talbot, 2015](#)).

Définition 1.1. Un *argument* est un ensemble de propositions dans lequel une proposition est affirmée sur la base d'autres propositions.

Un argument (du moins tel que défini ici) est donc composé de différentes propositions. Parmi ces propositions, certaines vont être utilisées pour *affirmer* une autre proposition.

Définition 1.2. La *conclusion* est l'affirmation faite sur la base d'autres propositions.

Définition 1.3. Les *prémisses* d'un argument sont les raisons offertes qui permettent d'affirmer la conclusion.

Pour résumer, un *argument* est un ensemble de propositions, parmi lesquelles des *prémisses* sont utilisées pour affirmer une *conclusion*. Les propositions qui composent un *argument* (i.e., les prémisses et la conclusion) peuvent être *vraies* ou *fausses* mais l'argument ne peut pas être *vrai* ou *faux*. Un argument est seulement *valide* ou *invalidé*.

Définition 1.4. Un argument est dit *valide* si et seulement si il n'existe aucune situation logiquement possible dans laquelle tous les prémisses de l'argument soient vrais et sa conclusion fausse.

Définition 1.5. Un argument est dit *invalid*e si et seulement si il existe aucune situation logiquement possible dans laquelle tous les prémisses de l'argument soient vraies et sa conclusion fausse.

Pour résumer, un argument est un ensemble de *propositions*, dont certaines d'entre elles (les prémisses) sont utilisées pour affirmer (ou justifier) une autre (la conclusion). Ces propositions peuvent être vraies ou fausses mais un argument peut seulement être *valide* ou *invalid*e. Un argument est dit *valide* lorsqu'il est logiquement *impossible* pour la conclusion d'être fausse (sachant que les prémisses sont *vraies*).

1.2.2 Quelques syllogismes connus



Figure 1.2 : Un pingouin s'essayant à la logique. Source : <https://www.pinterest.com/pin/465418942711158498/>.

Un syllogisme est un raisonnement logique qui met en relation *au moins* trois propositions : au moins deux prémisses et une conclusion. Afin d'illustrer les définitions proposées ci-dessus, nous allons maintenant examiner quelques exemples. Savez-vous reconnaître les arguments valides et invalides ?

Argument n°1

- Prémissie n°1 : Si un suspect ment, il transpire.
- Prémissie n°2 : (On observe que) Ce suspect transpire.
- Conclusion : Par conséquent, ce suspect ment.

Cet argument est *invalid*e car (on applique la définition 1.4) il existe des situations dans lesquelles à la fois les prémisses 1 et 2 sont vraies, et pourtant la conclusion est fausse. Par exemple, il se peut que le suspect transpire pour d'autres raisons que le mensonge (e.g., la température de la salle d'interrogatoire).

Argument n°2

- Prémissse n°1 : Si un suspect transpire, il ment.
- Prémissse n°2 : (On observe que) Ce suspect ne transpire pas.
- Conclusion : Par conséquent, ce suspect ne ment pas.

Cet argument est également *invalid*e car il existe des situations dans lesquelles à la fois les prémisses 1 et 2 sont vraies, et pourtant la conclusion est fausse. Par exemple, il se peut que le suspect fasse partie des gens qui ne transpirent pas lorsqu'ils mentent.

Argument n°3

- Prémissse n°1 : Tous les menteurs transpirent.
- Prémissse n°2 : (On observe que) Ce suspect ne transpire pas.
- Conclusion : Par conséquent, ce suspect n'est pas un menteur.

Cet argument est *valide* car il n'existe aucune situation dans laquelle à la fois les prémisses 1 et 2 sont vraies et la conclusion serait fausse. Autrement dit, si les prémisses 1 et 2 sont vraies, il est *logiquement impossible* pour la conclusion d'être fausse. Nous allons maintenant examiner quelques raisonnements valides et invalides connus afin de nous aider à les repérer plus facilement.

1.2.2.1 Arguments invalides

Le premier raisonnement fallacieux que nous allons étudier est connu comme le sophisme de l'**affirmation du conséquent**. Ce raisonnement vise à inférer la réalisation d'un *antécédent* sur la base de la réalisation du *conséquent*. Considérons l'exemple suivant :

- Prémissse n°1 : S'il a plu, alors le sol est mouillé (A implique B).
- Prémissse n°2 : Le sol est mouillé (B).
- Conclusion : Donc il a plu (A).

Dans cet argument, la prémissse n°1 nous dit que l'antécédent (A) implique le conséquent (B). La prémissse n°2 *affirme* le conséquent B. La conclusion consiste à affirmer l'antécédent A sur la base de ces deux prémisses. Or cet argument est invalide, car (en l'occurrence) le sol pourrait être mouillé pour d'autres raisons que la pluie. Il s'agit du même genre de raisonnement que l'argument n°1 discuté dans la section précédente et il peut s'écrire dans une forme générale de la manière suivante :

$$\frac{A \Rightarrow B, B}{A}$$

où $A \Rightarrow B$ se lit "A implique B" et se comprend comme dans la phrase "Si A, alors B".

Un deuxième argument fallacieux relativement répandu est connu comme le sophisme de la **négation de l'antécédent** et consiste à affirmer une négation du conséquent (i.e., non B) sur la base d'une négation de l'antécédent (i.e., non A). Considérons l'exemple suivant :

- Prémissse n°1 : S'il a plu, alors le sol est mouillé (A implique B).
- Prémissse n°2 : Il n'a pas plu (non A).
- Conclusion : Donc le sol n'est pas mouillé (non B).

Dans cet exemple comme dans le précédent, la prémissse n°1 nous dit que l'antécédent (A) implique le conséquent (B). La prémissse n°2 affirme une négation de l'antécédent (i.e., non A ou $\neg A$). La conclusion consiste à affirmer une négation du conséquent (i.e., $\neg B$) sur la base de ces deux prémisses. Cet argument est également invalide car (en l'occurrence) le sol pourrait être mouillé pour d'autres raisons que la pluie. Autrement dit :

$$\frac{A \Rightarrow B, \neg A}{\neg B}$$

où $\neg A$ représente la négation de A (i.e., non A).

1.2.2.2 Arguments valides

Attardons-nous maintenant sur deux des raisonnements valides les plus connus. Le premier est connu comme le **modus ponens** et consiste à déduire un conséquent sur la base d'une implication (e.g., A implique B) et de l'affirmation d'un antécédent. Considérons l'exemple suivant :

- Prémissse n°1 : Si on est lundi, alors John ira au travail (A implique B).
- Prémissse n°2 : On est lundi (A).
- Conclusion : Donc John ira au travail (B).

Comme dans les deux exemples précédents, la prémissse n°1 nous dit que l'antécédent (A) implique le conséquent (B). La prémissse n°2 *affirme* l'antécédent A. La conclusion consiste à affirmer le conséquent B sur la base de ces deux prémisses. Cet argument est *valide* (cf. définition 1.4) car il n'existe aucune situation *logiquement possible* dans laquelle les deux prémisses seraient vraies et la conclusion fausse. Autrement dit, cet argument est valide car il est *impossible* pour la conclusion d'être fausse, sachant que les prémisses sont vraies. Le modus ponens peut s'écrire de la manière suivante :

$$\frac{A \Rightarrow B, A}{B}.$$

Le deuxième argument valide que nous allons discuter est connu comme le **modus tollens**, dont l'importance s'avère capitale dans le raisonnement scientifique, ou du moins dans sa version idéalisée (cf. section suivante). Cet argument consiste à déduire la négation de l'antécédent sur la base d'une implication et de la négation du conséquent. Considérons l'exemple suivant :

- Prémissse n°1 : Si mon chien détecte un intrus, alors il aboie (A implique B).

- Prémissse n°2 : Mon chien n'a pas aboyé (non B).
- Conclusion : Donc il n'a pas détecté d'intrus (non A).

Comme dans les exemples précédents, la prémissse n°1 nous dit que l'antécédent (A) implique le conséquent (B). La prémissse n°2 affirme la negation du conséquent (i.e., non B). La conclusion consiste à affirmer la negation de l'antécédent (i.e., non A) sur la base de ces deux prémisses. Cet argument est *valide* (cf. définition 1.4) car il n'existe aucune situation *logiquement possible* dans laquelle les deux prémisses seraient vraies et la conclusion fausse. Autrement dit, cet argument est valide car il est *impossible* pour la conclusion d'être fausse, sachant que les prémisses sont vraies. Dans notre exemple, si le chien n'a pas aboyé, c'est *nécessairement* qu'il n'a pas détecté d'intrus. Cependant, cela ne veut pas dire qu'aucun intrus a visité notre maison, seulement que le chien n'a pas détecté d'intrus. Le modus tollens peut s'écrire de la manière suivante :

$$\frac{A \Rightarrow B, \neg B}{\neg A}$$

Ayant défini ce qu'est un argument, ce qui le compose et ce qui fait sa validité, nous disposons maintenant des outils nécessaires afin d'étudier la "logique" du raisonnement scientifique, et d'essayer de voir la place occupée par l'analyse de données dans ce raisonnement.

1.2.3 Qu'est-ce qu'une théorie scientifique ?

Qu'est-ce qu'une théorie scientifique ? D'un point de vue très général, une théorie scientifique peut être définie comme un ensemble de propositions logiques qui postulent des relations causales entre des phénomènes observables. Dans un premier temps, ces propositions sont formulées en termes abstraits et généraux (e.g., "tout objet répond à la force de gravité de manière similaire"), mais mènent ensuite à des propositions concrètes et testable empiriquement (e.g., "la vitesse de chute de deux objets devrait être la même, toute chose étant égale par ailleurs"). Il existe cependant de nombreux "types" de théories scientifiques. Par exemple, P. E. Meehl (1986) liste trois types de théories :

- *Functional-dynamic theories* : les théories qui relient "les états aux états" ou "les événements aux événements". Par exemple, ce type de théorie décrit comment un changement sur une variable affecte une ou plusieurs autres variable(s).
- *Structural-compositional theories* : les théories qui expliquent "de quoi est composé quelque chose", de quels genres d'objets un plus gros objet est fait (i.e., sa structure), ou comment ces différentes parties sont assemblées.
- *Evolutionary theories* : les théories qui s'intéressent à l'histoire et au développement des choses (e.g., la théorie de l'évolution, la chute de Rome, etc).

Malgré cette diversité, et sans consensus clair sur ce qui fait une bonne ou une mauvaise théorie, la philosophie des sciences nous offre cependant des outils *conceptuels* utiles pour évaluer

les théories, identifier ce qui les rend plus ou moins “fortes”, et évaluer ce qui fait un *test sévère* d'une théorie. Mais comment-on pouant nous évaluer les théories et comment créer des test sévères et pertinents ?

1.2.3.1 On ne peut pas les confirmer

Un premier “problème” avec les théories scientifiques est que nous ne pouvons pas les “confirmer”. En effet, selon Campbell (1990), le raisonnement scientifique (naif) aurait la forme logique suivante :

- Prémissse n°1 : Si la théorie de Newton (A) est “vraie”, alors on devrait observer observer que les marées ont la période B, la trajectoire de Mars la forme C, la trajectoire d'une boule de canon la forme D, etc.
- Prémissse n °2 : Nos observations confirment B, C, et D.
- Conclusion : Donc la théorie de Newton est “vraie”.

Or cet argument est invalide. Comme nous l'avons vu dans la section précédente, il s'agit du raisonnement fallacieux d'affirmation du conséquent. Une manière de s'en rendre compte est de représenter visuellement la forme de cet argument (cf. Figure 1.3).

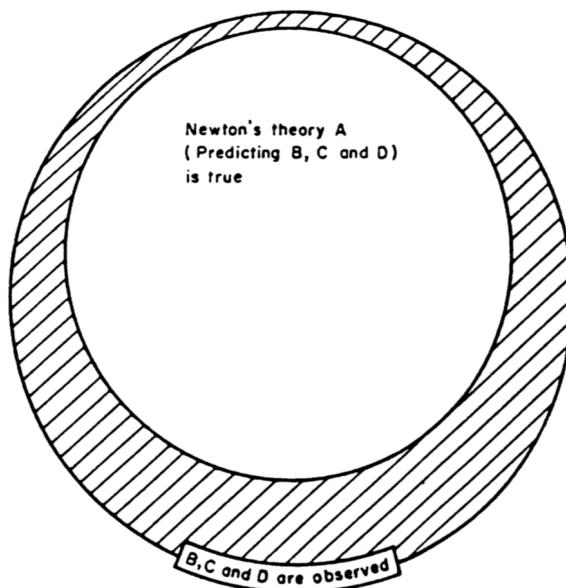


Figure 1.3 : Relation entre observations et théorie de la gravitation de Newton, selon Campbell (1990).

L'invalidité de cet argument provient de l'existence de la zone hachurée de la Figure 1.3, qui contient les autres explications possibles pour les observations que nous avons réalisées. En d'autres termes, observer B, C, et D ne nous permet pas de conclure que la théorie de Newton (A) est vraie, car ces observations pourraient avoir été générées par d'autres phénomènes que ceux postulés par la théorie de Newton. Cependant, observer B, C, D peut tout de même être informatif, selon certaines conditions, au regard de la théorie de Newton. Par exemple, si nous

n'avions pas observé B, C, et D, alors nous aurions pu conclure que A était fausse. Donc observer B, C, et D fait que A reste “plausiblement vraie”. Une autre manière de le dire est que A a “survécu” au test des observations B, C, et D. Nous verrons un peu plus loin comment la capacité des théories à survivre à un test empirique peut être utilisé comme métrique d'évaluation des théories.

1.2.3.2 On ne peut pas les réfuter (au sens strict)

Armés de nos connaissances en logique, nous avons donc établi qu'une théorie scientifique ne peut pas être *confirmée*. Peut-être pourrions nous alors les *réfuter*? A-t-on des moyens de montrer qu'une théorie est fausse? Qu'est-ce que cela veut dire pour une théorie d'être “fausse”? Selon la pensée influente de Popper, une théorie est falsifiable (ou réfutable) si et seulement si il existe au moins un falsificateur potentiel (i.e., au moins une proposition possible qui soit en contradiction logique avec elle). En d'autres termes, une théorie peut être considérée comme réfutable s'il peut être démontré qu'elle est fausse.

Notons au passage que la falsifiabilité de Popper concerne le problème de la *démarcation* (c'est-à-dire ce qu'est la science et ce qui est la pseudoscience) et définit les pseudosciences comme étant composées de théories non falsifiables (c'est-à-dire des théories qui ne permettent pas d'être réfutées). Mais lorsqu'il s'agit de décrire *comment* la science fonctionne (visée descriptive) ou comment la science *devrait* fonctionner (visée prescriptive), le standard falsificationniste ne fonctionne pas vraiment. En fait, il est quasiment unanimement impossible d'appliquer le falsificationnisme déductif dans des contextes scientifiques concrets ([McElreath, 2016b](#)). Dans les sections suivantes, nous discutons de quatre problèmes qui nous empêchent de réfuter des théories, à savoir : i) la distinction entre modèle théorique et modèle statistique, ii) le problème de la mesure, iii) la nature probabiliste des hypothèses scientifiques, et enfin iv) le problème de Duhem-Quine.

1.2.3.3 Modèles théoriques et modèles statistiques

Un modèle statistique est un appareil utilisé pour relier, pour faire le lien entre un modèle théorique et certaines données. Il peut être défini comme une instanciation d'une théorie en un ensemble d'énoncés ou de propositions probabilistes ([Rouder et al., 2016](#)). En général, il n'existe pas de relation univoque entre modèles théoriques et modèles statistiques. Autrement dit, un modèle théorique donné peut être représenté (i.e., implémenté) par différents modèles statistiques et réciproquement, différents modèles statistiques peuvent être construits à partir du même modèle théorique. Par conséquent, la confirmation ou réfutation d'un modèle statistique ne permet pas l'induction strict au modèle théorique.

Par exemple, une pratique statistique courante en sciences expérimentales est le test d'hypothèse nulle fréquentiste (*Null Hypothesis Significance Testing* ou NHST), qui consiste à tester une hypothèse nulle (souvent l'hypothèse d'absence d'effet) afin de confirmer (ou plutôt, corroborer) une hypothèse théorique alternative d'intérêt. Or, le fait de rejeter l'hypothèse d'absence d'effet ne fournit qu'une très faible *corroboration* de l'hypothèse d'intérêt, comme de nombreuses théories peuvent potentiellement prédire un effet est non-nul. L'hypothèse

d'absence d'effet (i.e., l'hypothèse que l'effet est précisément égal à 0) est beaucoup plus contraignante (restrictive) que l'hypothèse alternative selon laquelle l'effet n'est pas 0.

1.2.3.4 Le problème de la mesure

La logique de la réfutation est assez simple et repose sur la puissance du *modus tollens*. Appliquée au raisonnement scientifique, cet argument peut être présenté de la manière suivante :

- Prémissse n°1 : Si ma théorie T est correcte, alors je devrais observer certaines données D .
- Prémissse n°2 : J'observe d'autres données que celles prédictes par ma théorie $\neg D$.
- Conclusion : Donc, ma théorie est fausse $\neg T$.

Cet argument est parfaitement valide pour les propositions logiques, qui peuvent être soit vraies, soit fausses. Cependant, le premier problème qui apparaît lorsqu'on applique ce raisonnement à des cas concrets de raisonnement scientifique est le problème de l'erreur d'observation (ou erreur de mesure). Toute observation est sujette à de l'erreur, surtout lorsqu'on étudie des phénomènes nouveaux (McElreath, 2016b).

Considérons un instant un exemple qui nous vient de Physique, lorsqu'a été rapportée l'observation de **neutrinos** plus rapides que la vitesse de la lumière (McElreath, 2016b, 2020). Selon Einstein, aucun objet ne peut voyager plus vite que la lumière. Par conséquent, l'observation de certaines particules (en l'occurrence des neutrinos) qui voyageraient à une vitesse supérieure à celle de la lumière pourraient être considérée comme une réfutation flagrante de la théorie de la relativité restreinte.

En 2011, une large équipe de physiciens de renommée internationale ont pourtant annoncé la détection de neutrinos voyageant plus rapidement que la vitesse de la lumière. De manière intéressante, la première réaction de la communauté scientifique ne fut pas d'annoncer que la théorie d'Einstein était réfutée. Bien au contraire, la plus grande partie de la communauté s'est demandée "D'où est-ce que vient l'erreur dans les mesures réalisées par cette équipe?" (McElreath, 2016b, 2020). L'équipe ayant réalisé ces mesures a de ses voeux appelé à des réplications indépendantes de leurs résultats. Deux ans plus tard et après plusieurs ré-analyses et réplications contradictoires, la communauté était unanime que les résultats qui semblaient contredire la théorie d'Einstein étaient en fait dû à une erreur de mesure (l'équipe ayant réalisé les premières mesures a réalisé plus tard que l'erreur provenait **d'un câble mal branché**).

Cette anecdote de l'histoire des sciences nous apprend au moins deux choses. Premièrement, il est intéressant d'analyser la réaction de la communauté de l'annonce de ces résultats. La théorie de la relativité restreinte ayant accumulé de nombreux succès prédictifs au cours du dernier siècle, la survenue d'une observation si dramatiquement incompatible avec la théorie était perçue comme hautement improbable aux yeux des experts. Cela nous renseigne sur la manière dont les théories scientifiques gagnent en "crédence" aux yeux d'une communauté d'expert, et également comme cette crédence ou l'historique d'une théorie affecte ou influence la manière d'interpréter les observations empiriques. Deuxièmement, cela souligne

le fait qu'une observation ou un ensemble d'observation peut difficilement compter comme une réfutation stricte d'une théorie, car il existe (presque) toujours une probabilité de se tromper ou une erreur irréductible dans la précision de la mesure. De manière générale, le problème avec l'erreur de mesure est de savoir si la réfutation d'une théorie T par un ensemble d'observations D est véritable ou simplement superficielle. Sachant que toute mesure est sujette à de l'erreur, toute conclusion scientifique qui repose sur une ou des mesure(s) ne peut qu'apporter une réfutation partielle (exprimée en termes probabilistes) d'une théorie, et non une réfutation stricte (comme en logique formelle).

1.2.3.5 Hypothèses probabilistes

Un autre problème émerge lorsqu'on essaye d'appliquer le modus tollens aux hypothèses scientifiques. Ce problème (désigné comme "illusion permanente" par Gigerenzer, 1993) est que la plupart des hypothèses scientifiques ne sont pas vraiment de la forme "tous les cygnes sont blancs" mais sont plutôt de la forme suivante :

- Mon hypothèse est que 90% des cygnes sont blancs.
- Si mon hypothèse est correcte, alors on ne devrait *probablement pas* observer des cygnes noirs.

Sachant cette hypothèse, que peut-on conclure si on observe un cygne noir ? Et bien pas grand chose. Un autre exemple classique en sciences expérimentales est celui de la logique du test d'hypothèse nulle (Cohen, 1994) :

- Prémissse n°1 : Si l'hypothèse nulle est vraie, alors ces données sont peu probables.
- Prémissse n °2 : On observe ces données.
- Conclusion : Donc l'hypothèse nulle est improbable.

Cependant, à cause du prémissse probabiliste (prémissse n°1), cet argument est invalide et sa conclusion est fausse. Pour s'en rendre compte, considérons un autre exemple (Cohen, 1994; Pollard & Richardson, 1987) :

- Prémissse n°1 : Si un individu est Américain, il est peu probable qu'il soit membre du Congrès.
- Prémissse n°2 : Cet individu n'est pas membre du Congrès.
- Conclusion : Cet individu n'est probablement pas Américain.

Cette conclusion est saugrenue est l'argument est invalide, car il oublie de considérer l'alternative, qui est que si cet individu n'était pas Américain, la probabilité qu'il soit membre du Congrès serait de 0. Cet argument est identique au précédent :

- Prémissse n°1 : Si l'hypothèse nulle est vraie, alors ces données sont peu probables.
- Prémissse n °2 : On observe ces données.
- Conclusion : Donc l'hypothèse nulle est improbable.

Et cet argument est invalide pour les mêmes raisons que le précédent, à savoir i) que le prémissse n°1 est probabiliste (et non discret) et ii) qu'il ne considère pas l'hypothèse alternative. Ainsi, même sans erreur de mesure, on se rend compte que le problème d'hypothèse probabiliste nous empêche de réfuter ce genre d'hypothèse via le modus tollens.

1.2.3.6 Forme logique du test expérimental d'une théorie

Un dernier (mais non des moindres) problème est connu comme la “thèse de Duhem-Quine” ou le “problème de l’indétermination”. En pratique, lorsqu’une théorie T est testée, il est nécessaire de faire appel à des hypothèses sous-jacentes (i.e., non explicites) ou à d’autres théories. Ces théories “auxiliaires” nous aident à “connecter” la théorie d’intérêt T avec le “monde réel”, afin de faire des prédictions concrètes (e.g., “les cygnes blancs et noirs passent la même proportion de leur temps à se ballader, donc la probabilité de les observer dans la nature devrait être égale”). Ces théories auxiliaires sont souvent des théories à propos des outils que nous utilisons (e.g., “le BDI est ou un outil valide pour mesurer le niveau de symptômes dépressifs chez des patients souffrant de dépression chronique”).

Lorsque nous testons une théorie qui prédit que “Si O_1 ” (une manipulation expérimentale), “Alors O_2 ” (une observation prédictive), ce que nous voulons dire en fait est que l’on devrait observer cette relation *si et seulement si* tous les éléments auxiliaires sont corrects. Ainsi, la structure logique du test empirique d’une théorie T peut être décrit de la manière suivante ([Paul E. Meehl, 1990a, 1978, 1997](#)) :

$$(T \wedge A_t \wedge C_p \wedge A_i \wedge C_n) \rightarrow (O_1 \supset O_2)$$

où “ \wedge ” représente une conjonction (“et”), “ \rightarrow ” représente une déduction logique, et “ \supset ” représente l’implication logique (e.g., “Si O_1 , Alors O_2 ”). A_t est une conjonction de théories auxiliaires, C_p est connu comme le *ceteribus paribus* (i.e., on postule qu’il n’existe pas de facteurs extérieurs non pris en compte et qui pourraient “masquer” l’effet d’intérêt) A_i est une théorie auxiliaire concernant les outils utilisés pour mesurer l’effet d’intérêt, et C_n est un énoncé à propos des conditions particulières de l’expérience réalisée (i.e., on postule qu’il n’existe pas de bruit ou erreur systémique dans le protocole expérimental).

En d’autres termes, une *conjonction* de tous les éléments du côté gauche de la formule ci-dessus (ce qui inclut notre théorie T) implique la partie droite de la formule, c’est à dire “Si O_1 , Alors O_2 ”. Si l’expérience réalisée nous révèle que cette relation ne tient pas, alors on aimerait pouvoir conclure que notre hypothèse T est réfutée (en appliquant le modus tollens).

Or, une négation de la partie droite de cette formule nous permet seulement d'affirmer une négation de l'**intégralité** de la partie gauche. Autrement dit, ne pas observer une prédition

empirique d'une théorie nous permet de réfuter l'ensemble $T \wedge A_t \wedge C_p \wedge A_i \wedge C_n$, ce qui est très différent d'une réfutation de T (Paul E. Meehl, 1990a). En termes plus formels, une négation de la conjonction (de gauche) est logiquement équivalent à déclarer une disjonction des conjoints (i.e., soit l'un ou l'autre des composants de la partie gauche est faux).

Pour résumer, ne pas observer quelque chose qui était prédit par une théorie ne permet pas de montrer que cette théorie est fausse, mais cela permet de montrer que la conjonction de la théorie et des hypothèses auxiliaires est fausse. Une conséquence des quatre problèmes soulevés dans cette section et que la réfutation d'une théorie scientifique n'est jamais logique, mais elle est **consensuelle** (McElreath, 2016b, 2020). Une proposition théorique est considérée comme réfutée lorsqu'une communauté d'experts a accumulé un grand nombre de preuves variées, issus de protocoles et de groupes de recherches variés, au fil des décennies. Ce travail d'accumulation des preuves s'accompagne de discussions critiques indissociables du travail de développement théorique. En somme, la réfutation d'une théorie est un résultat social, issue d'une communauté d'experts, et n'est (presque) jamais le résultat d'une déduction logique formelle.

1.2.4 Test d'hypothèse nulle et raisonnement scientifique

Une croyance répandue en sciences expérimentales est que l'utilisation de la procédure NHST est bien alignée avec la philosophie scientifique de Popper (et implicitement, qu'il s'agit là de quelque chose de souhaitable). Cependant, le parallèle entre la procédure NHST et la philosophie Poppérienne est très approximatif. La logique de la procédure NHST peut être résumée de la manière suivante :

1. On suppose l'hypothèse d'absence d'effet \mathcal{H}_0 .
2. On génère un nombre infini d'échantillons sous cette hypothèse.
3. On compare les données que nous avons observées dans notre expérience à la distribution contrefactuelle des données sous l'hypothèse nulle \mathcal{H}_0 .

Si les données observées semblent suffisamment invraisemblables conditionnellement à \mathcal{H}_0 (où “suffisamment” correspond au niveau α du test), nous pouvons rejeter l'hypothèse nulle en toute sécurité et considérer ce rejet comme une corroboration de l'hypothèse alternative \mathcal{H}_1 (quelle que soit l'hypothèse alternative).

En d'autres termes, la seule hypothèse réellement testée (dans la procédure NHST classique) est l'hypothèse nulle, qui est rarement d'intérêt pour le chercheur en train de la tester. Ainsi, afin de réellement aligner cette procédure avec la méthode falsificationniste, il faudrait tester les prédictions de l'hypothèse théorique \mathcal{T} réellement d'intérêt, et non les prédictions d'une hypothèse épouvantail \mathcal{H}_0 .³ Comme résumé par P. E. Meehl (1986) :

“[...] we have been brainwashed by Fisherian statistics into thinking that refutation of H0 is a powerful way of testing substantive theories”.

³À ce propos, et comme nous le verrons un peu plus tard, il est possible et relativement facile de généraliser la procédure de calcul des p-valeurs à n'importe quel modèle statistique dans le cadre bayésien, voir par exemple cet article de blog : <http://www.barelysignificant.com/post/ppc/>.

Pour résumer, Fidler et al. (2018) décrivent quatre raisons qui questionne le parallèle entre la procédure NHST et la méthode falsificationniste.

1. L'hypothèse nulle \mathcal{H}_0 (ou plutôt la *nil hypothesis*, c'est à dire l'hypothèse que la valeur du paramètre testé est précisément 0) est très probablement fausse car en sciences sociales en particulier “tout tend à être associé avec tout”, un phénomène également connu comme le *crud factor* (Paul E. Meehl, 1990b).
2. La procédure NHST ne teste pas réellement l'hypothèse d'intérêt \mathcal{T} , mais seulement l'hypothèse épouvantail \mathcal{H}_0 .
3. De nombreuses hypothèses théories sous-jacentes ne sont pas assez bien développées ou formalisées et ne permettent pas de formuler des hypothèses statistiques.
4. Même si tous les points précédents ont été résolus, nous aurions encore besoin de soumettre ces hypothèses à des tests sévères, ce qui nécessiterait des études bien alimentées et bien conçues, ce qui n'est pas la norme actuelle ⁴.

Le point n°2 soulevé ci-dessus nous dit que la procédure NHST ne suit pas le falsificationnisme Poppérien car il ne soumet pas la théorie testée à un risque de falsification *sévère*, mais seulement à un danger très faible. En d'autres termes, le test d'hypothèse nulle (tel que pratiqué dans la procédure NHST classique) ne soumet pas la théorie sous-jacente (i.e., celle qu'on vise à évaluer via le test de l'hypothèse statistique dérivée de cette première) à un test fort (Paul E. Meehl, 1990a, 1967, 1997). Autrement dit, étant donné que l'hypothèse nulle est facilement réfutée, l'hypothèse alternative est facilement corroborée statistiquement, et donc peu corroborée au niveau théorique.

Ceci étant dit, cette critique porte sur une manière d'utiliser la procédure NHST (qui est aujourd'hui la plus répandue en Psychologie), mais ce n'est pas la seule manière d'utiliser cette procédure. On pourrait très bien utiliser cette procédure afin d'essayer de réfuter notre hypothèse d'intérêt \mathcal{T} . Paul E. Meehl (1967) distingue entre l'usage *fort* et *faible* du test d'hypothèse nulle en comparant l'usage de la procédure NHST en physique et en psychologie. L'usage *faible* du test d'hypothèse nulle, décrit ci-dessus, correspond à tenter de réfuter une hypothèse nulle qui ne nous intéresse pas afin de corroborer une hypothèse alternative d'intérêt.

L'usage fort du test d'hypothèse nulle nécessite cependant d'avoir une théorie suffisamment développée, à même de prédire une valeur numérique précise pour une observation, ou tout du moins un intervalle réduit de valeurs possibles, ou alors certaines formes de fonctions (e.g., quadratique ou cubique) entre les variables d'intérêt (Paul E. Meehl, 1997). Dans ce genre de situation, le test d'hypothèse nulle pourrait agir comme un test Poppérien *risqué*, au sens où l'hypothèse testée est soumises à un risque élevé de réfutation (Paul E. Meehl, 1997). Notons également que la procédure NHST peut être adaptée afin de réfuter des intervalles de valeurs,

⁴La puissance statistique moyenne des tests d'hypothèse réalisés en Psychologie est estimée à moins de 50% (e.g., Szucs & Ioannidis, 2017).

via les tests d'équivalence (Lakens et al., 2018; Rogers et al., 1993) ou la procédure HDI+ROPE (Kruschke, 2015).

Bien entendu, dans certaines (rares) situations, l'hypothèse d'absence d'effet est théoriquement d'intérêt, et donc viser à réfuter cette hypothèse via un test d'hypothèse nulle représenterait une tentative sérieuse de réfutation. Par exemple il existe certaines théories en Psychologie qui prédisent que certains comportement seraient invariants selon certaines situations (Morey et al., 2018). Ces hypothèses peuvent être testées *sévèrement* par un test d'hypothèse nul, car ce dernier les exposerait à un haut risque de réfutation.

Pour résumer, la vision naïve du falsificationnisme consiste à penser que la science progresse par falsification logique (et que donc la statistique devrait viser la falsification). Cependant, comme discuté dans cette section, cette perspective se retrouve face à plusieurs problèmes difficilement surmontables, que nous résumons ci-dessous/

- Premier problème : Les hypothèses théoriques ne sont pas les modèles (hypothèses statistiques). Un modèle statistique est un appareil utilisé pour relier, pour faire le lien entre un modèle théorique et certaines données. Il peut être défini comme une instanciation d'une théorie en un ensemble d'énoncés ou de propositions probabilistes (Rouder et al., 2016).
- Deuxième problème : En général, il n'existe pas de relation univoque entre modèles théoriques et modèles statistiques. Autrement dit, un modèle théorique donné peut être représenté (i.e., implémenté) par différents modèles statistiques et réciproquement, différents modèles statistiques peuvent être construits à partir du même modèle théorique. Par conséquent, la confirmation ou réfutation d'un modèle statistique ne permet pas l'induction strict au modèle théorique.
- Troisième problème : Les hypothèses scientifiques sont souvent probabilistes, ce qui invalide l'emploi du modus tollens.
- Quatrième problème : Les mesures permettant de tester une théorie sont sujettes à des erreurs, ce qui empêche également la réfutation stricte d'hypothèses (cf. l'anecdote des neutrinos).

Enfin, la falsification concerne le problème de la démarcation, pas celui de la méthode. La science est une technologie sociale, la falsification est **consensuelle**, et non pas logique.

1.2.5 L'approche par comparaison de modèles

En connaissance des limitations de l'approche par test d'hypothèse nulle telle que présentée dans la section précédente, nous adoptons dans ce livre une approche dite par *comparaison de modèle*.

En bref, en place d'une approche mécanique du test d'hypothèse nulle, nous proposons une méthode qui met l'accent sur l'estimation de paramètres, la comparaison de modèles statistiques et théoriques (sensés et d'intérêt), et l'extension (amélioration) continue du modèle

(Burnham & Anderson, 2004; e.g., Burnham & Anderson, 2002; Cumming, 2014, 2012; Gelman et al., 2013; Gelman & Hill, 2006; Judd et al., 2009; Kruschke, 2015; Kruschke & Liddell, 2018a, 2018b; McElreath, 2016a). En d'autres notre approche est une approche de **modélisation statistique** plutôt qu'une approche de **test statistique** (e.g., Noël, 2015). Cette approche vise à modéliser le processus sous-jacent ayant généré les données observées (i.e., le *processus de génération des données* ou PGD) plutôt qu'à tester si la valeur de certains paramètres d'un modèle inapproprié est égal à une valeur arbitraire (e.g., $\theta = 0$). Cette approche n'est cependant pas incompatible avec l'approche falsificationniste tel que décrite en philosophie des sciences. En effet, certains statisticiens bayésiens comme Gelman & Shalizi (2013) suggèrent que la réfutation des modèles statistiques joue un rôle important dans le processus de modélisation et d'amélioration des modèles (nous y reviendrons à plusieurs reprises, en particulier lorsque nous discuterons l'utilisation des *prior* et *posterior predictive checks*).

Afin d'illustrer l'approche par comparaison de modèles, considérons l'exemple suivant. On s'intéresse au lien entre deux variables aléatoires continues x et y . On réalise une expérience et on collecte 10 observations nous permettant d'étudier cette relation. L'hypothèse de modélisation la plus classique est de postuler une relation linéaire entre x et y . La droite minimisant la somme des erreurs au carré est représentée par la Figure 1.4.

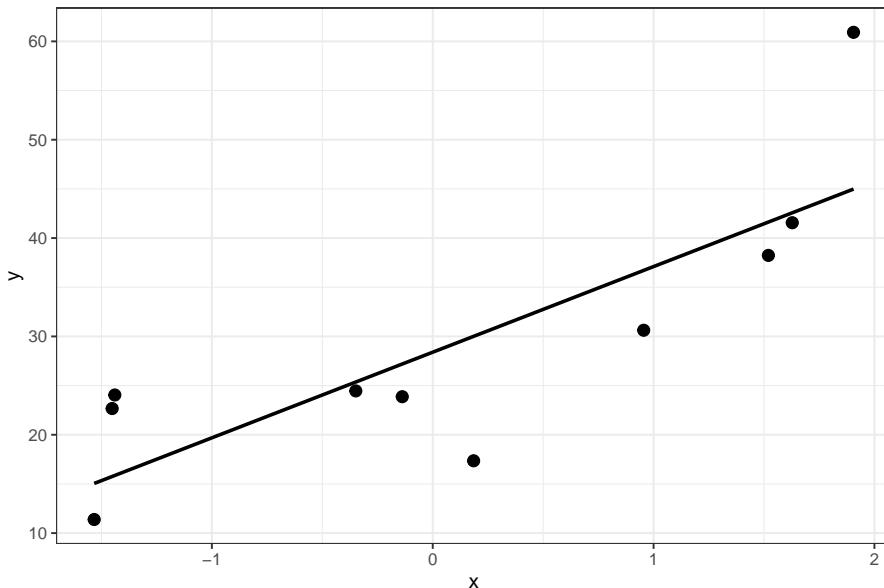


Figure 1.4 : Scatterplot des 10 observations obtenues dans notre expérience et droite des moindres carrés décrivant la relation entre x et y estimée sur la base de ces 10 observations.

Cette description peut-être *améliorée* (où “améliorer” consiste à réduire l'erreur) pour mieux prendre en compte les données qui s'écartent de la prédition linéaire. La figure 1.5 représente la prédition d'un modèle polynomial (quadratique).

Cette “amélioration” du modèle statistique via une augmentation de la complexité de ce dernier peut être poursuivi. On sait qu'un ensemble de N points peut être *exhaustivement* (i.e., sans erreur) décrit par une fonction polynomiale d'ordre $N - 1$ (cf. Figure 1.6). Augmenter la complexité du modèle améliore donc la précision de notre description des données mais réduit également la généralisabilité de ses prédictions (il s'agit du dilemme classique entre biais

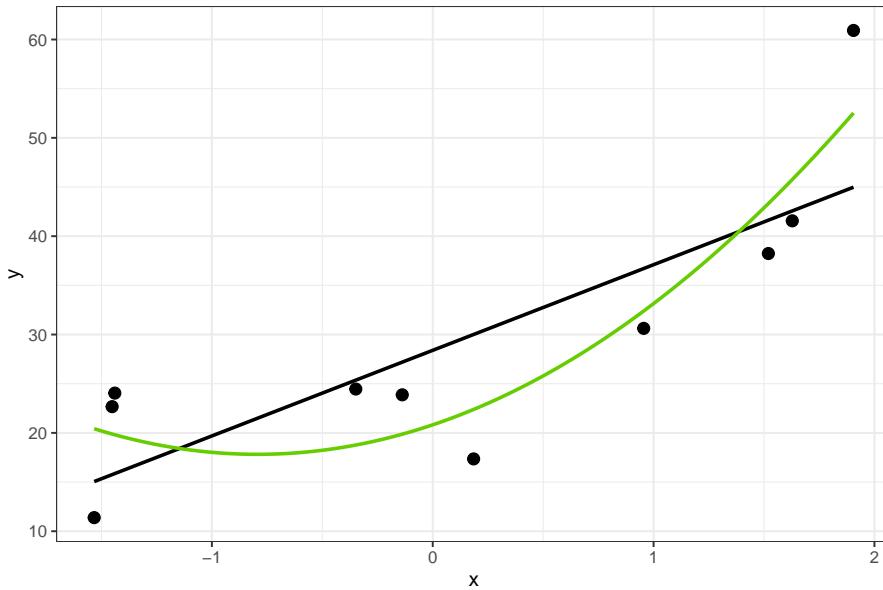


Figure 1.5 : Prédiction quadratique décrivant la relation entre x et y estimée sur la base des 10 observations collectées pour notre expérience.

et variance).

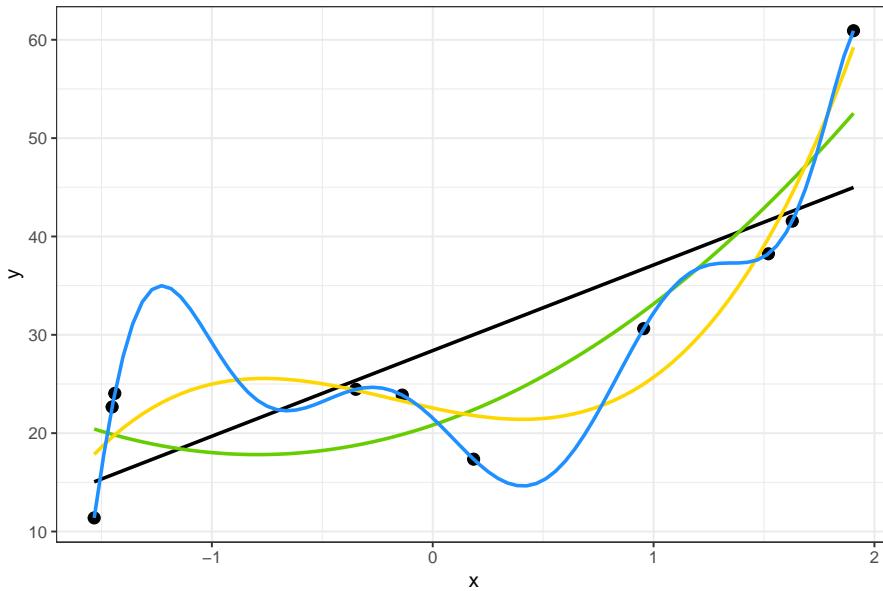


Figure 1.6 : Différentes prédictions de complexité croissante estimées sur la base des 10 observations collectées pour notre expérience.

Nous avons donc besoin d'outils qui prennent en compte le rapport entre la qualité de la description des données et la complexité du modèle, c'est à dire qui évaluent la parcimonie du modèle. Dans cette perspective, nous ferons au Chapitre 6 un détour par la théorie de l'information, qui nous permettra introduire des outils comme l'AIC (et ses différentes extensions).

Pour résumer, notre approche consistera donc à construire des modèles statistiques comme des implémentations mathématiques (probabilistes) de modèles théoriques que nous souhaitons comparer. L'inférence statistique bayésienne consistera à mettre à jour notre état de

connaissance concernant les valeur des paramètres de ces modèles (mais également notre état de connaissance vis à vis de la validité relative de ces modèles) en fonction des données observées. Ces modèles seront comparés en fonction de i) leur puissance prédictive et ii) leur complexité. Au lieu d'essayer de réfuter un modèle épouvantail (i.e., l'hypothèse nulle), on comparera des modèles intrinsèquement intéressants, qu'on essayera de réfuter afin de les améliorer. La section suivante présente un premier exemple permettant de saisir l'idée centrale de l'inférence bayésienne.

1.3 Problème du sac de billes (McElreath, 2016b)

Imaginons que nous disposions d'un sac contenant 4 billes. Ces billes peuvent être soit blanches, soit bleues. Nous savons qu'il y a précisément 4 billes, mais nous ne connaissons pas le nombre de billes de chaque couleur. Nous savons cependant qu'il existe cinq possibilités (que nous considérons comme nos *hypothèses*) :

Hypothèse n°1 : ○○○○

Hypothèse n°2 : ●○○○

Hypothèse n°3 : ●●○○

Hypothèse n°4 : ●●●○

Hypothèse n°5 : ●●●●

Le but est de déterminer quelle combinaison est la plus probable, **sachant certaines observations**. Imaginons que l'on tire trois billes à la suite, avec remise, et que l'on obtienne la séquence suivante : ●○●.

Cette séquence représente nos données. À partir de ces données, quelle **inférence** peut-on faire sur le contenu du sac? En d'autres termes, que peut-on dire de la probabilité de chaque hypothèse?

1.3.1 Énumérer les possibilités

Une stratégie consiste à compter le nombre de possibilités menant aux données obtenues à chaque tirage. Par exemple, si nous considérons l'hypothèse n°2 (i.e., on se place dans un cadre dans lequel cette hypothèse est "vraie"), on peut représenter l'arbre des issues possibles. La Figure 1.7 représente ces différentes possibilités. Au premier tirage, selon l'hypothèse n°2, nous avions une chance sur 4 d'obtenir une bille bleue.

La Figure 1.8 représente l'ensemble des résultats possibles aux tirages 1 et 2 selon l'hypothèse n°2. On réalise qu'au deuxième comme au premier tirage, on avait une chance sur quatre d'obtenir une bille bleue. Par conséquent on avait $1 \times (4 \times 1) = 4$ chances sur $4^2 = 16$ d'obtenir deux billes bleues. De la même manière, on peut calculer qu'on avait 1×3 chances sur $4^2 = 16$ d'obtenir une bille bleue et une bille blanche. Autrement dit, 3 chemins mènent à la suite "bille bleue puis bille blanche".



Figure 1.7 : Représentation de l'ensemble des issues possibles au premier tirage selon l'hypothèse n°2.

La Figure 1.9 représente l'ensemble des résultats possibles aux tirages 1, 2, et 3 selon l'hypothèse n°2. On réalise qu'à chaque tirage on avait une chance sur quatre d'obtenir une bille bleue. Par conséquent on avait $1 \times (4 \times 1) \times (4 \times 1) = 16$ chances sur $4^3 = 64$ d'obtenir trois billes bleues. De la même manière, on peut calculer qu'on avait $1 \times 3 \times 1$ chances sur $4^3 = 64$ d'obtenir une bille bleue et une bille blanche. Autrement dit, 3 chemins mènent à la suite “bille bleue puis bille blanche puis bille bleue”.

La Figure 1.10 représente le nombre de “chemins” qui mènent au résultat obtenu et confirme que sous l'hypothèse n°2, 3 chemins sur $4^3 = 64$ conduisent au résultat obtenu. Qu'en est-il des autres hypothèses ?

La Figure 1.11 représente le nombre de chemins menant aux données observées pour les hypothèse n°2, n°3, et n°4.

On peut ensuite comparer les hypothèses par leur *propension* à mener aux données observées. Plus précisément, on peut comparer les hypothèses entre elles en comparant le nombre de chemins menant aux données pour chaque hypothèse (cf. Tableau 1.1).

On pourra conclure, au vu des données, que l'hypothèse n°4 est la plus *plausible* car c'est l'hypothèse qui **maximise le nombre de manières possibles d'obtenir les données obtenues**.

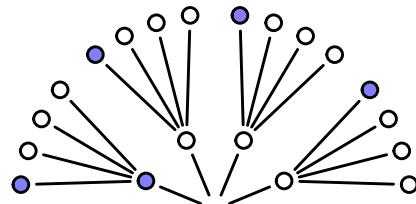


Figure 1.8 : Représentation de l'ensemble des issues possibles aux premier et deuxième tirages selon l'hypothèse n°2.

Tableau 1.1 : Comparer des hypothèses en comparant le nombre de manières qu'elles ont de produire les données observées.

Hypothèse	Façons d'obtenir les données
○ ○ ○○	$0 \times 4 \times 0 = 0$
● ○ ○○	$1 \times 3 \times 1 = 3$
●● ○○	$2 \times 2 \times 2 = 8$
●●●○	$3 \times 1 \times 3 = 9$
●●●●	$4 \times 0 \times 4 = 0$

1.3.2 Accumulation d'évidence

Jusque là, nous avons considéré que toutes les hypothèses étaient équiprobables a priori (suivant le [principe d'indifférence](#)). Cependant, on pourrait avoir de l'information a priori, provenant de nos connaissances (e.g., concernant les particularités des sacs de billes) ou de données antérieures. Imaginons que nous tirions une nouvelle bille du sac. Comment pouvons-nous incorporer cette nouvelle donnée ?

Il suffit d'appliquer la même stratégie que précédemment, et de mettre à jour le dernier compte en le multipliant par ces nouvelles données (cf. Tableau 1.2). Cette procédure illustre

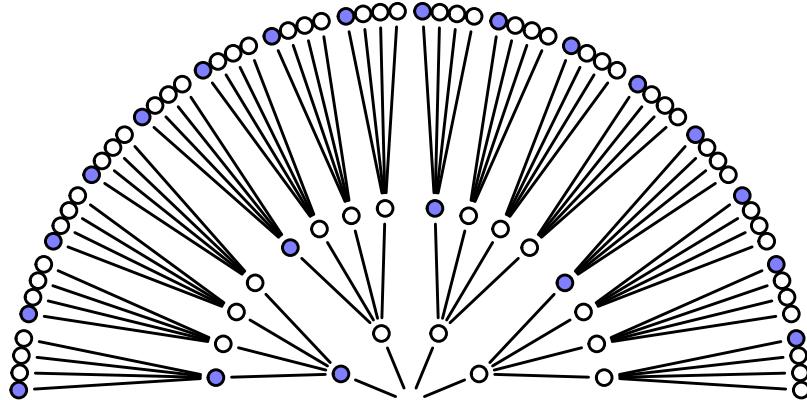


Figure 1.9 : Représentation de l'ensemble des issues possibles selon l'hypothèse n°2 sur l'ensemble des tirages.

un mécanisme central de l'inférence bayésienne qui concerne l'accumulation d'information. Dans le cadre bayésien, cette accumulation se déroule naturellement, où le résultat de nouvelles analyses peut être naturellement incorporé au résultat d'analyses précédentes, pour mener à un état de connaissance mis à jour.

Tableau 1.2 : Illustration de l'accumulation d'information dans le cadre bayésien.

Hypothèse	Façons de produire ●	Compte précédent	Nouveau compte
○ ○ ○○	0	0	$0 \times 0 = 0$
● ○ ○○	1	3	$3 \times 1 = 3$
●● ○ ○	2	8	$8 \times 2 = 16$
●●● ○	3	9	$9 \times 3 = 27$
●●●●	4	0	$0 \times 4 = 0$

1.3.3 Incorporer un prior

Supposons maintenant qu'un employé de l'usine de fabrication des billes nous dise que les billes bleues sont rares. Plus précisément, cet employé nous dit que pour chaque sac contenant 3 billes bleues, ils fabriquent deux sacs en contenant seulement deux, et trois sacs en

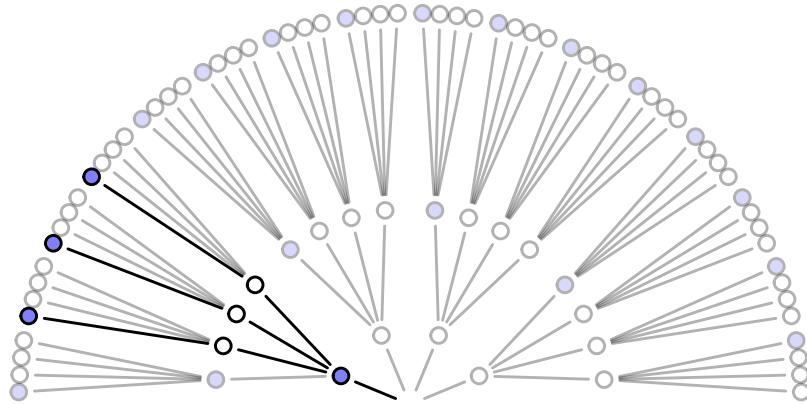


Figure 1.10 : Représentation de l'ensemble des issues possibles selon l'hypothèse n°2 sur l'ensemble des tirages. Les chemins menant aux données observées sont mis en avant.

contenant seulement une. Il nous apprend également que tous les sacs contiennent au moins une bille bleue et une bille blanche. On peut traduire ces informations en langage mathématique, en attribuant un poids de 0 pour les hypothèses n°1 et n°5, et en rendant l'hypothèse n°2 trois fois plus probable que l'hypothèse n°4, et l'hypothèse n°3 deux fois plus probable que l'hypothèse n°4 (cf. Tableau 1.3).

Tableau 1.3 : Illustration de l'intégration d'information a priori dans le cadre bayésien.

Hypothèse	Compte précédent	Prior usine	Nouveau compte
○ ○ ○○	0	0	$0 \times 0 = 0$
● ○ ○○	3	3	$3 \times 3 = 9$
●● ○ ○	16	2	$16 \times 2 = 32$
●●● ○	27	1	$27 \times 1 = 27$
●●●●	0	0	$0 \times 4 = 0$

Cette procédure illustre encore une fois l'intégration d'information antérieure (on dira également *a priori*), que ce soit empirique ou théorique, dans l'analyse de nouvelles données. Ce mécanisme est résumé par le célèbre dicton bayésien : *Yesterday's posterior is today's prior* (Lindley, 2001).

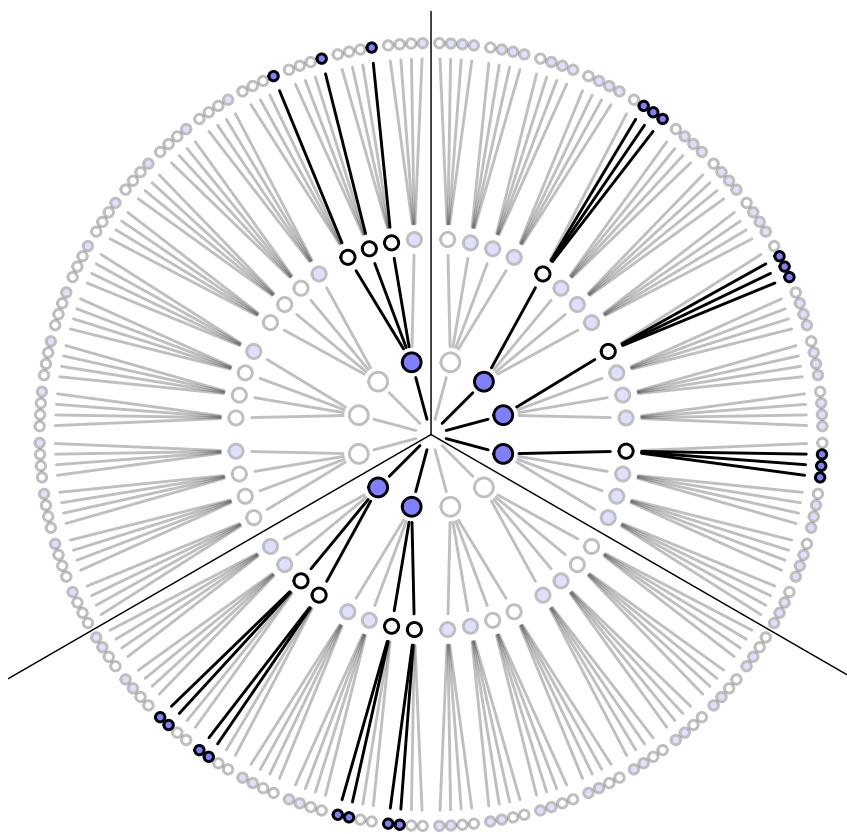


Figure 1.11 : Représentation de l'ensemble des issues possibles selon les hypothèses n°2, n°3, et n°4 sur l'ensemble des tirages. Les chemins menant aux données observées sont mis en avant.

1.3.4 Des énumérations aux probabilités

Le théorème de Bayes (que nous présenterons formellement un peu plus tard) nous dit que la probabilité d'une hypothèse après avoir observé certaines données est proportionnelle au nombre de façons qu'a cette hypothèse de produire les données observées, multiplié par sa probabilité a priori.

$$\Pr(\text{hypothèse} \mid \text{données}) \propto \Pr(\text{données} \mid \text{hypothèse}) \times \Pr(\text{hypothèse})$$

Si l'on considère toutes les hypothèses comme étant équiprobables a priori (e.g., on leur accorde toutes une probabilité de 1), la **probabilité postérieure** de l'hypothèse **sachant les données observées et le prior** est obtenue en normalisant le produit $\Pr(\text{données} \mid \text{hypothèse}) \times \Pr(\text{hypothèse})$ de la manière suivante :

$$\Pr(\text{hypothèse} \mid \text{données}) = \frac{\Pr(\text{données} \mid \text{hypothèse}) \times \Pr(\text{hypothèse})}{\text{Somme des produits}}$$

Où la somme des produits consiste à calculer le produit $\Pr(\text{données} \mid \text{hypothèse}) \times \Pr(\text{hypothèse})$ pour toutes les hypothèses considérées et à calculer la somme de ces valeurs.

Cette manipulation nous assure que la probabilité postérieure (i.e., $\Pr(\text{hypothèse} \mid \text{données})$) est bien une probabilité, c'est à dire une valeur numérique bornée entre 0 et 1.

Par exemple, si l'on définit p comme étant la proportion de billes bleues dans chaque hypothèse, alors on peut ré-exprimer chaque hypothèse en fonction de p (cf. Tableau 1.4).

Tableau 1.4 : Calcul de la probabilité postérieure d'une hypothèse.

Hypothèse	p	Manières de produire les données	Probabilité postérieure
○ ○ ○○	0	0	0
● ○ ○○	0.25	3	0.15
●● ○ ○	0.5	8	0.40
●●● ○	0.75	9	0.45
●●●●	1	0	0

Où la probabilité est calculée en divisant chaque valeur de la troisième colonne par la somme des valeurs de cette colonne. Autrement dit, en R :

```
ways <- c(0, 3, 8, 9, 0)
ways / sum(ways)
```

```
## [1] 0.00 0.15 0.40 0.45 0.00
```

Pour résumer, la probabilité postérieure représente la probabilité d'une hypothèse, **sachant** certaines données observées et certaines connaissances a priori. Cette probabilité postérieure est proportionnelle au produit de la probabilité des données sachant l'hypothèse (i.e., le nombre de manières qu'a l'hypothèse de produire les données) et de la probabilité a priori de l'hypothèse. Autrement dit et comme résumé par McElreath (2016b), “*Bayesian inference is really just counting and comparing of possibilities*”.

1.4 Rappels de théorie des probabilités

Pour rappel, une probabilité est une valeur numérique comprise entre 0 et 1 et qui respecte la règle de la somme. Ces valeurs sont assignées à des *événements* ω , étant définis comme des sous-ensembles d'un grand *ensemble* Ω . Chaque événement de cet ensemble peut se voir assigner une probabilité qui représente notre in(certitude) vis à vis de sa survenue. Ces probabilités sont assignées par des *fonctions de probabilité* qui, à chaque élément $\omega \in \Omega$, associe ou attribue une probabilité (Blitzstein & Hwang, 2019; Dekking, 2005; Noël, 2015).

Comme illustration, considérons l'exemple suivant. En postulant qu'il est impossible qu'une pièce retombe sur sa tranche, un lancer de pièce peut seulement résulter en deux issues : Pile ou Face. Autrement dit, l'ensemble des issues possibles est défini comme $\Omega = \{\text{Pile, Face}\}$. Étant donné qu'un *événement* est défini comme un sous-ensemble de Ω , Pile et Face sont donc deux événements...

Définition 1.6 (Fonction de probabilité). Une fonction de probabilité p définie sur un ensemble fini Ω assigne à chaque événement A dans Ω une valeur $\Pr(A) \in [0, 1]$ de manière à ce que :

- $\Pr(\Omega) = 1$ et
- $\Pr(A \cup B) = \Pr(A) + \Pr(B)$ si A et B sont disjoints.

La valeur $\Pr(A)$ représente la *probabilité* que A se réalise.

Dans l'exemple d'un lancer de pièce, si la pièce n'est pas truquée, alors $\Pr(\text{Pile}) = \Pr(\text{Face}) = \frac{1}{2}$...

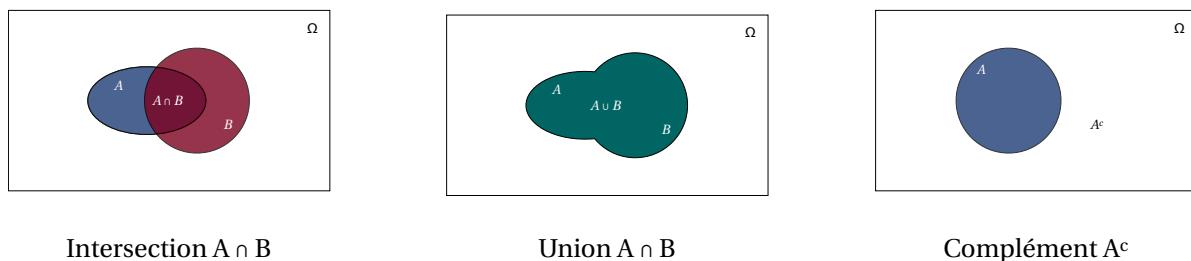


Figure 1.12 : Diagrammes représentant les notions d'intersection, d'union, et de complément.

...

1.4.1 Probabilité conjointe

La probabilité conjointe $\Pr(A, B)$ nous indique la probabilité qu'à la fois A et B se réalisent, c'est à dire la probabilité de l'union de A et B , qu'on note également $\Pr(A \cap B)$.

```
library(tidyverse)

data(HairEyeColor) # données adaptés de Snee (1974)

cont <- apply(HairEyeColor, c(1, 2), sum) %>% t
cont <- round(cont / sum(cont), 2)
cont

##          Hair
## Eye      Black Brown Red Blond
##   Brown   0.11  0.20  0.04  0.01
##   Blue    0.03  0.14  0.03  0.16
##   Hazel   0.03  0.09  0.02  0.02
##   Green   0.01  0.05  0.02  0.03
```

Dans chaque cellule du tableau de données ci-dessus, on trouve la **probabilité conjointe** d'avoir telle couleur de cheveux **ET** telle couleur d'yeux, qui s'écrit $p(c, y) = p(y, c)$.

1.4.2 Probabilité marginale

```
cont2 <- cont %>% as.data.frame %>% mutate(marginal_eye = rowSums(cont) )
rownames(cont2) <- row.names(cont)
cont2

##          Black Brown  Red Blond marginal_eye
## Brown    0.11  0.20  0.04  0.01      0.36
## Blue     0.03  0.14  0.03  0.16      0.36
## Hazel    0.03  0.09  0.02  0.02      0.16
## Green    0.01  0.05  0.02  0.03      0.11
```

On peut aussi s'intéresser à la probabilité d'avoir des yeux bleus, de manière générale. Il s'agit de la probabilité **marginale** de l'événement *yeux bleus*, qui s'obtient par la somme de toutes les probabilités jointes impliquant l'événement *yeux bleus*. Elle s'écrit $p(y) = \sum_c p(y|c)p(c)$.

```
cont3 <- rbind(cont2, colSums(cont2) )
rownames(cont3) <- c(row.names(cont2), "marginal_hair")
cont3

##          Black Brown  Red Blond marginal_eye
## Brown    0.11  0.20  0.04  0.01      0.36
## Blue     0.03  0.14  0.03  0.16      0.36
## Hazel    0.03  0.09  0.02  0.02      0.16
## Green    0.01  0.05  0.02  0.03      0.11
## marginal_hair 0.18  0.48  0.11  0.22      0.99
```

On peut bien entendu aussi s'intéresser aux probabilités des couleurs de cheveux, de manière générale. Elle s'écrit $p(c) = \sum_y p(c|y)p(y)$.

1.4.3 Probabilité conditionnelle

On pourrait aussi s'intéresser à la probabilité qu'une personne ait les cheveux blonds, **sachant** qu'elle a les yeux bleus. Il s'agit d'une probabilité **conditionnelle**, et s'écrit $p(c|y)$. Cette probabilité conditionnelle peut se ré-écrire : $p(c|y) = \frac{p(c,y)}{p(y)}$.

```
##          Black Brown  Red Blond marginal_eye
## Blue     0.03  0.14  0.03  0.16      0.36
```

Par exemple, quelle est la probabilité d'avoir des yeux bleus lorsqu'on a les cheveux blonds ?

```
cont3["Blue", "Blond"] / cont3["Blue", "marginal_eye"]
```

```
##      Blue
## 0.4444444
```

On remarque dans le cas précédent que $p(blonds|bleus)$ **n'est pas nécessairement égal à $p(bleus|blonds)$** .

Autre exemple : la probabilité de mourir sachant qu'on a été attaqué par un requin n'est pas la même que la probabilité d'avoir été attaqué par un requin, sachant qu'on est mort (*confusion of the inverse*). De la même manière, $p(data|H_0) \neq p(H_0|data)$.

À partir des axiomes de Kolmogorov (cf. début du cours), et des définitions précédentes des probabilités conjointes, marginales, et conditionnelles, découle la **règle du produit** (en multipliant chaque côté par $p(y)$) :

$$p(a, b) = p(b) \cdot p(a|b) = p(a) \cdot p(b|a)$$

...

1.4.4 Dérivation du théorème de Bayes

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

$$p(y|x)p(x) = p(x|y)p(y)$$

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

On retrouve le résultat présenté dans la section précédente, en remplaçant x par *données* et y par *hypothèse* :

$$\Pr(\text{hypothèse} | \text{données}) = \frac{\Pr(\text{données} | \text{hypothèse}) \times \Pr(\text{hypothèse})}{\text{Somme des produits}}$$

...

1.4.5 Loi de probabilité, cas discret

...

Définition 1.7 (Fonction de masse de probabilité). La fonction de masse de probabilité p d'une variable aléatoire X est la fonction $p : \mathbb{R} \rightarrow [0, 1]$, définie par :

$$p(a) = \Pr(X = a) \quad \text{for } -\infty < a < \infty$$

Une fonction de masse (*probability mass function*, ou *PMF*) est une fonction qui attribue une probabilité à chaque valeur d'une variable aléatoire. Exemple de la distribution binomiale pour une pièce non biaisée ($\theta = 0.5$), indiquant la probabilité d'obtenir N faces sur 10 lancers.

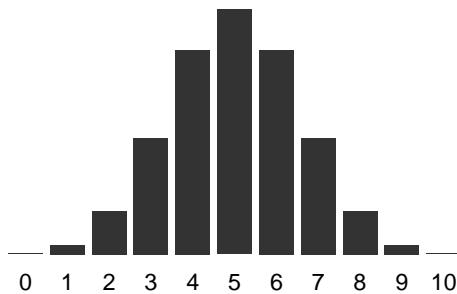


Figure 1.13 : Distribution de la probabilité d'obtenir N 'Face' sur 10 lancers de pièce.

Somme à 1...

```
# PMFs sum to 1
dbinom(x = 0:10, size = 10, prob = 0.5) %>% sum
```

```
## [1] 1
```

1.4.6 Loi de probabilité, cas continu

Définition 1.8 (Fonction de densité de probabilité). Une variable aléatoire X est dite *continue* si pour une fonction donnée $p : \mathbb{R} \rightarrow \mathbb{R}$ et pour tout nombres a et b avec $a \leq b$,

$$\Pr(a \leq X \leq b) = \int_a^b p(x)dx$$

La fonction p doit satisfaire la condition $p(x) \geq 0$ pour tout x et $\int_{-\infty}^{\infty} p(x)dx = 1$. On appelle p la fonction de densité de probabilité (ou densité de probabilité) de X .

Une fonction de densité de probabilité (*probability density function*, ou *PDF*), est une fonction qui permet de représenter une loi de probabilité sous forme d'intégrales (l'équivalent de la PMF pour des variables aléatoires strictement continues).

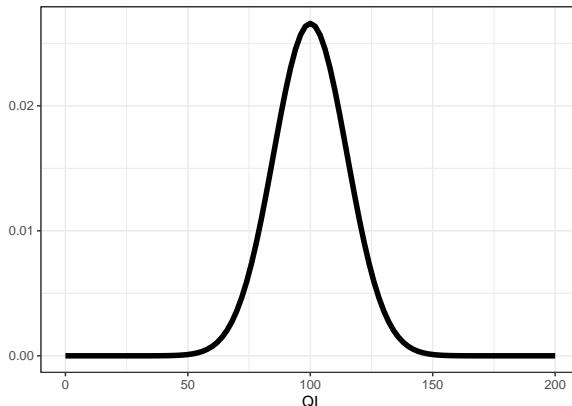


Figure 1.14 : Blah blah...

```
# PDFs integrate to 1
integrate(dnorm, -Inf, Inf, mean = 100, sd = 15)
```

```
## 1 with absolute error < 1.3e-06
```

...

Concept essentiel 1.1

Variable aléatoire continue

Une variable aléatoire continue peut prendre (littéralement) une infinité de valeurs. On se retrouve donc face un problème. Si on attribue une probabilité non-nulle à chacune de ces valeurs (à une infinité de valeurs donc), la 'somme' (l'intégrale) des probabilités de ces valeurs sera elle aussi infinie, et cette fonction ne pourra donc pas être considérée comme une fonction de probabilité. Pour pallier à ce problème, chaque valeur ponctuelle d'une variable aléatoire est assignée une probabilité nulle (i.e., $\Pr(X = x) = 0$) et uniquement des intervalles (e.g., $\Pr(a < x < b)$) peuvent se voir attribuer une probabilité.

...

1.4.7 Aparté, qu'est-ce qu'une intégrale ?

Une intégrale correspond à la **surface** (aire géométrique) délimitée par la représentation graphique d'une fonction, *l'aire sous la courbe*. Une distribution est dite **impropre** si son intégrale n'est pas égale à un nombre fini (e.g., $+\infty$) et **normalisée** si son intégrale est égale à 1.

L'intégrale de $f(x)$ sur l'intervalle $[90; 96]$ vaut : $p(90 < x < 96) = \int_{90}^{96} f(x) dx = 0.142$.

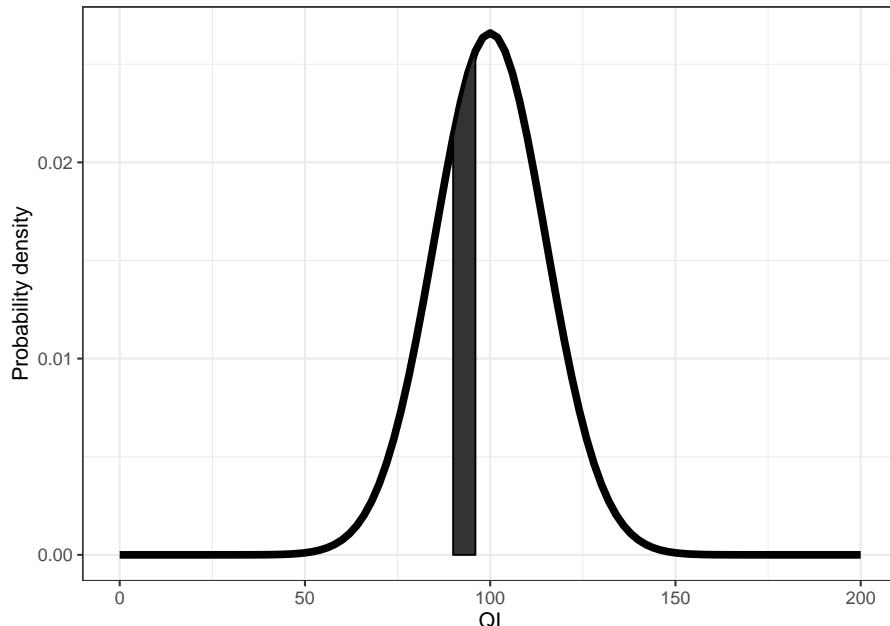


Figure 1.15 : Blah blah...

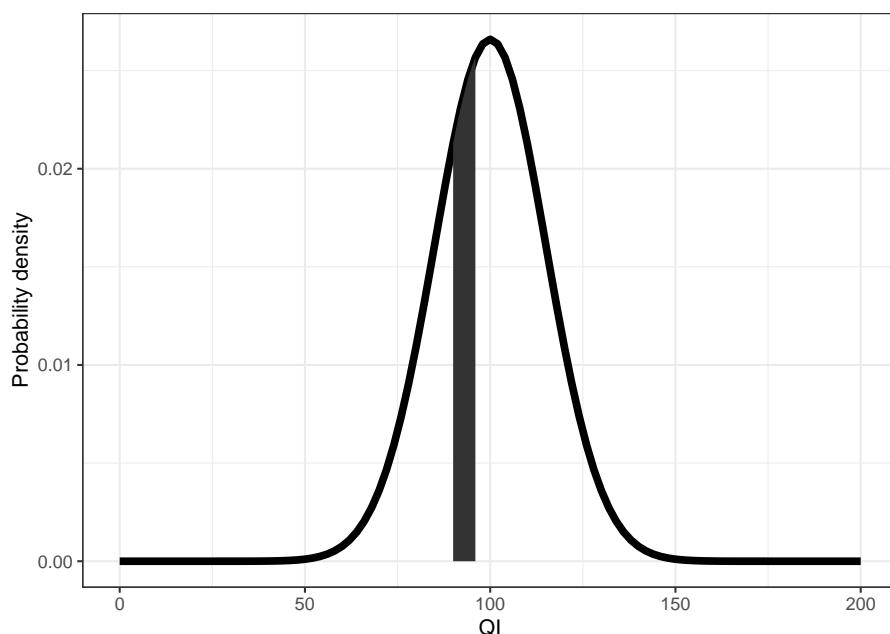


Figure 1.16 : Blah blah...

```
integrate(dnorm, 90, 96, mean = 100, sd = 15)
```

```
## 0.1423704 with absolute error < 1.6e-15
```

1.4.8 Notations, terminologie

Ayant introduit de manière intuitive les concepts centraux de l'inférence bayésienne (en particulier, la mise à jour d'un connaissance a posteriori en une connaissance a posteriori) nous allons maintenant établir la terminologie qui nous accompagner au fil de ce livre.

- θ désigne habituellement un paramètre ou un vecteur de paramètres (e.g., la proportion de billes bleues)
- $p(x|\theta)$ désigne la probabilité conditionnelle des données x sachant le paramètre θ [$p(x|\theta = \theta)$]
- $p(x|\theta)$ une fois que la valeur de x est connue, est vue comme la fonction de vraisemblance (*likelihood*) du paramètre θ . Attention, il ne s'agit pas d'une distribution de probabilité (n'intègre pas à 1). [$p(x = x|\theta)$]
- $p(\theta)$ la probabilité a priori de θ
- $p(\theta|x)$ la probabilité a posteriori de θ (sachant x)
- $p(x)$ la probabilité marginale de x (sur θ)

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\sum_{\theta} p(x|\theta)p(\theta)} = \frac{p(x|\theta)p(\theta)}{\int_{\theta} p(x|\theta)p(\theta)dx} \propto p(x|\theta)p(\theta)$$

...

Concept essentiel 1.1

La théorie des probabilités comme extension de la logique

La théorie des probabilités est parfois présentée comme une extension de la logique. En effet, elle généralise les règles de la logique qui s'appliquent à des événements discrets (vrais ou faux) à des événements continus. Ce faisant, les probabilités nous permettent de décrire et quantifier l'incertitude. Il est important de souligner que les règles du calcul probabiliste ont le même statut que les règles logiques : ces règles de base peuvent être utilisées pour déduire des conclusions qui seront garanties d'être correctes, si les prémisses sont corrects.

Dans ce cadre, l'analyse statistique bayésienne peut être conceptualisée comme une application de la théorie des probabilités à l'analyse statistique. Bien que la dépendance des conclusions de ce genre d'analyse aux a priori qu'elles rendent explicitement souvent présenté comme une faiblesse, c'est précisément ce qui les rend 'optimales' ou 'cohérentes' (au sens où elles respectent les règles du calcul probabiliste). Comme résumé par Vandekerckhove (2018), conclure que les analyses bayésiennes seraient invalidées par l'utilisation d'informations a priori serait similaire à conclure que des deductions logiques seraient invalidées par la considération de prémisses.

1.5 Quelques exemples d'application

1.5.1 Diagnostique médical (Gigerenzer, 2002)

- Chez les femmes âgées de 40-50 ans, sans antécédents familiaux et sans symptômes, la probabilité d'avoir un cancer du sein est de .008.
- Propriétés de la mammographie :
 - Si une femme a un cancer du sein, la probabilité d'avoir un résultat positif est de .90
 - Si une femme n'a pas de cancer du sein, la probabilité d'avoir un résultat positif est de .07
- Imaginons qu'une femme passe une mammographie, et que le test est positif. Que doit-on **inférer**? Quelle est la probabilité que cette femme ait un cancer du sein?

1.5.1.1 Logique du Maximum Likelihood

- Une approche générale de l'estimation de paramètre
- Les paramètres **gouvernent** les données, les données **dépendent** des paramètres
 - Sachant certaines valeurs des paramètres, nous pouvons calculer la **probabilité conditionnelle** des données observées
 - Le résultat de la mammographie (i.e., les données) dépend de la présence / absence d'un cancer du sein (i.e., le paramètre)
- L'approche par *maximum de vraisemblance* pose la question : “*Quelles sont les valeurs du paramètre qui rendent les données observées les plus probables?*”
- Spécifier la probabilité conditionnelle des données $p(x|\theta)$
- Quand on le considère comme fonction de θ , on parle de **likelihood** : $L(\theta|x) = p(X = x|\theta)$
- L'approche par maximum de vraisemblance consiste donc à maximiser cette fonction, en utilisant les valeurs (connues) de x
- Si une femme a un cancer du sein, la probabilité d'obtenir un résultat positif est de .90
 - $p(Mam = +|Cancer = +) = .90$
 - $p(Mam = -|Cancer = +) = .10$
- Si une femme n'a pas de cancer du sein, la probabilité d'obtenir un résultat positif est de .07
 - $p(Mam = +|Cancer = -) = .07$
 - $p(Mam = -|Cancer = -) = .93$

- Une femme passe une mammographie, le résultat est positif...
 - $p(Mam = + | Cancer = +) = .90$
 - $p(Mam = + | Cancer = -) = .07$
- Maximum de vraisemblance : quelle est la valeur de *Cancer* qui **maximise** $Mam = +$?
 - $p(Mam = + | Cancer = +) = .90$
 - $p(Mam = + | Cancer = -) = .07$

Wait a minute...

1.5.1.2 Diagnostique médical, fréquences naturelles

- Considérons 1000 femmes âgées de 40 à 50 ans, sans antécédents familiaux et sans symptômes de cancer
 - 8 femmes sur 1000 ont un cancer
- On réalise une mammographie
 - Sur les 8 femmes ayant un cancer, 7 auront un résultat positif
 - Sur les 992 femmes restantes, 69 auront un résultat positif
- Une femme passe une mammographie, le résultat est positif
- Que devrait-on inférer?

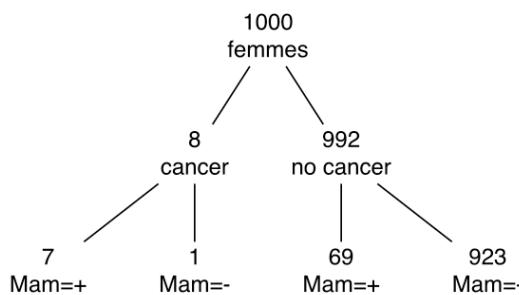


Figure 1.17 : Diagram...

$$p(Cancer = + | Mam = +) = \frac{7}{7 + 69} = \frac{7}{76} \approx .09$$

1.5.1.3 Diagnostique médical, théorème de Bayes

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{p(x)}$$

$p(\theta)$ la probabilité *a priori* de θ : tout ce qu'on sait de θ avant d'observer les données. Par exemple : $p(Cancer = +) = .008$ et $p(Cancer = -) = .992$.

```
prior <- c(0.008, 0.992)
```

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

$p(x|\theta)$ probabilité conditionnelle des données (x) sachant le paramètre (θ), qu'on appelle aussi la *likelihood* (*ou fonction de vraisemblance*) du paramètre (θ).

```
like <- rbind(c(0.9, 0.1), c(0.07, 0.93) ) %>% data.frame
```

```
colnames(like) <- c("Mam+", "Mam-")
```

```
rownames(like) <- c("Cancer+", "Cancer-")
```

```
like
```

```
##          Mam+ Mam-
```

```
## Cancer+ 0.90 0.10
```

```
## Cancer- 0.07 0.93
```

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

$p(x)$ la probabilité marginale de x (sur θ). Sert à normaliser la distribution.

$$p(x) = \sum_{\theta} p(x|\theta)p(\theta)$$

```
(marginal <- sum(like$"Mam+" * prior) )
```

```
## [1] 0.07664
```

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

$p(\theta|x)$ la probabilité a posteriori de θ sachant x , c'est à dire ce qu'on sait de θ après avoir pris connaissance de x .

```
(posterior <- (like$"Mam+" * prior) / marginal )
```

```
## [1] 0.09394572 0.90605428
```

1.5.1.4 L'inférence bayésienne comme mise à jour probabiliste des connaissances

Avant de passer le mammogramme, la probabilité qu'une femme tirée au sort ait un cancer du sein était de $p(\text{Cancer}) = .008$ (*prior*). Après un résultat positif, cette probabilité est devenue $p(\text{Cancer}|\text{Mam+}) = .09$ (*posterior*). Ces probabilités sont des expressions de nos *connaissances*. Après un mammogramme positif, on pense toujours que c'est "très improbable" d'avoir un cancer, mais cette probabilité a considérablement évolué relativement à "avant le test".

A Bayesianly justifiable analysis is one that treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes' theorem (Rubin, 1984).

1.5.2 Problème de Monty Hall

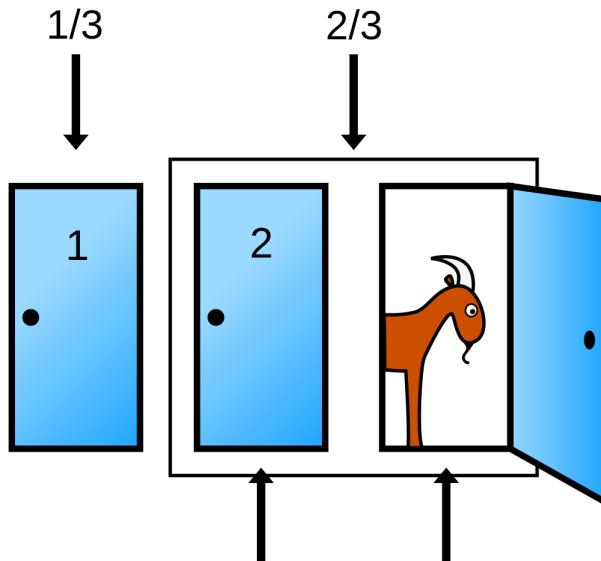


Figure 1.18 : A syllogistic penguin... Figure from...

Que-feriez-vous (intuitivement)? Analysez ensuite la situation en utilisant le théorème de Bayes.

1.5.2.1 Monty Hall - proposition de solution

Il s'agit d'un problème de probabilités conditionnelles... Définissons les événements suivants :

P1 : l'animateur ouvre la porte 1 P2 : l'animateur ouvre la porte 2 P3 : l'animateur ouvre la porte 3

V1 : la voiture se trouve derrière la porte 1 V2 : la voiture se trouve derrière la porte 2 V3 : la voiture se trouve derrière la porte 3

Si on a choisi la porte n°1 et que l'animateur a choisi la porte n°3 (*et qu'il sait où se trouve la voiture*), il s'ensuit que :

$$p(P3|V1) = \frac{1}{2}, p(P3|V2) = 1, p(P3|V3) = 0$$

On sait que $p(V3|P3) = 0$, on veut connaître $p(V1|P3)$ et $p(V2|P3)$ afin de pouvoir choisir.
Résolution par le théorème de Bayes.

$$p(V1|P3) = \frac{p(P3|V1) \times p(V1)}{p(P3)} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$p(V2|P3) = \frac{p(P3|V2) \times p(V2)}{p(P3)} = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

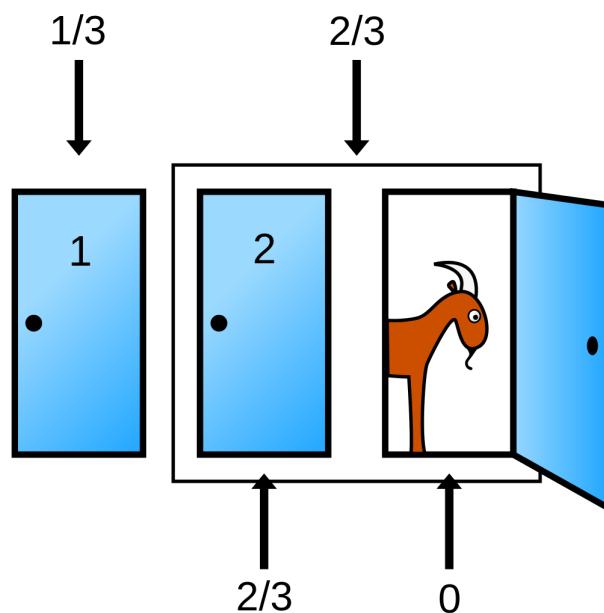


Figure 1.19 : A syllogistic penguin... Figure from...

Nos intuitions probabilistes sont, dans la grande majorité des cas, très mauvaises. Au lieu de compter sur elles, il est plus sage de se reposer sur des règles logiques (*modus ponens* et *modus tollens*) et probabilistes simples (règle du produit, règle de la somme, théorème de Bayes), nous assurant de réaliser l'inférence logique la plus juste. Autrement dit, “*Don't be clever*” (McElreath, 2016b).

Modèle beta-binomial

Introduction au chapitre blah blah...

2.1 Coefficient binomial

Définition 2.1 (Fonction factorielle). On appelle fonction factorielle la fonction qui à tout entier naturel n associe l'entier :

$$N! = N \times (N - 1) \times (N - 2) \times \cdots \times 3 \times 2 \times 1.$$

Définition 2.2 (Coefficient binomial). For any nonnegative integers k and n , the binomial coefficient...

Théorème 2.1 (Formule du coefficient binomial). *For $k \leq n$, we have :*

$$\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}$$

For $k > n$, we have $\binom{n}{k} = 0$.

Blah blah blah...

2.2 Le modèle Beta-Binomial

Pourquoi ce modèle?

Le modèle Beta-Binomial couvre un grand nombre de problèmes de la vie courante :

- Réussite / échec à un test
- Présence / absence d'effets secondaires lors du test d'un médicament
- Résultat d'un questionnaire à réponse binaire vrai / faux
- Estimation des résultats du deuxième tour de l'élection présidentielle

C'est un modèle simple

- Un seul paramètre
- Solution analytique

2.2.1 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 connaît 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 :

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

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

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

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

2.2.2 Processus 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, Face, 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 ?

2.2.3 Processus 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” :

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

Mais, jusque là on a 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., $\{0, 0, 1, 0, 1\}$)...

2.2.4 Coefficient binomial

Le **coefficients binomial** nous permet de calculer le nombre d’arrangements possibles résultant en y succès pour n lancers de la manière suivante :

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

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

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

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

```
## [1] 3
```

2.2.5 Loi binomiale

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

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)`).

2.2.6 Générer des données à partir d’une distribution binomiale

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

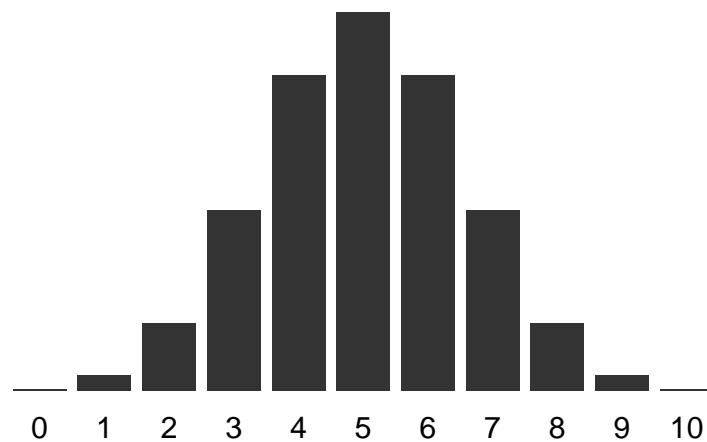


Figure 2.1 : Binomial distribution barplot...

```
rbinom(n = 500, size = 1, prob = 0.6) %>% # theta = 0.6
  data.frame %>%
  mutate(x = seq_along(.), y = cumsum(.) / seq_along(.)) %>%
  ggplot(aes(x = x, y = y), log = "y") +
  geom_line(lwd = 1) +
  geom_hline(yintercept = 0.5, lty = 3) +
  xlab("Nombre de lancers") +
  ylab("Proportion de faces") +
  ylim(0, 1) +
  theme_bw(base_size = 18)
```

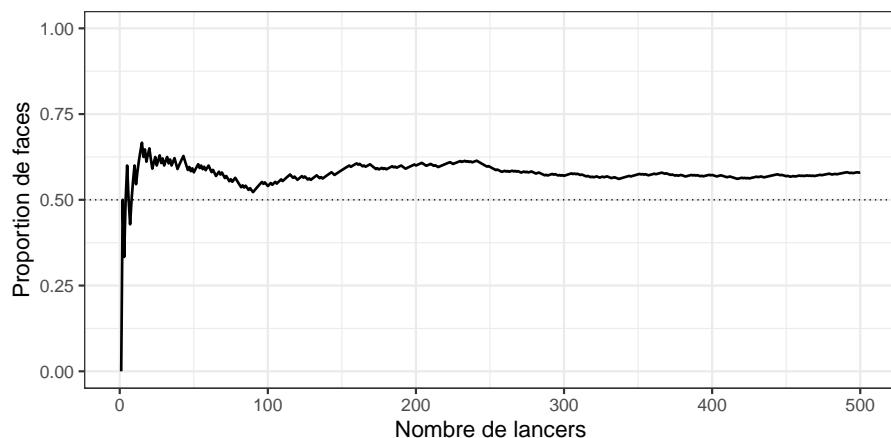


Figure 2.2 : Long-run...

2.2.7 Définition du modèle (likelihood)

2.2.7.1 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 :

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

2.2.8 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 de ces données selon (i.e., *en fonction de*) différentes valeurs de θ de la manière suivante :

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

Cette probabilité est définie pour un jeu de données fixe et une valeur de θ variable. On peut représenter cette fonction visuellement. Représentation graphique de la fonction de vraisemblance de theta pour $x = 1$ et $n = 2$...

```
y <- 1 # number of heads
n <- 2 # number of trials

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) +
  xlab(expression(paste(theta, " - Pr(face)")) ) + ylab("Likelihood") +
  theme_bw(base_size = 20)
```

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

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

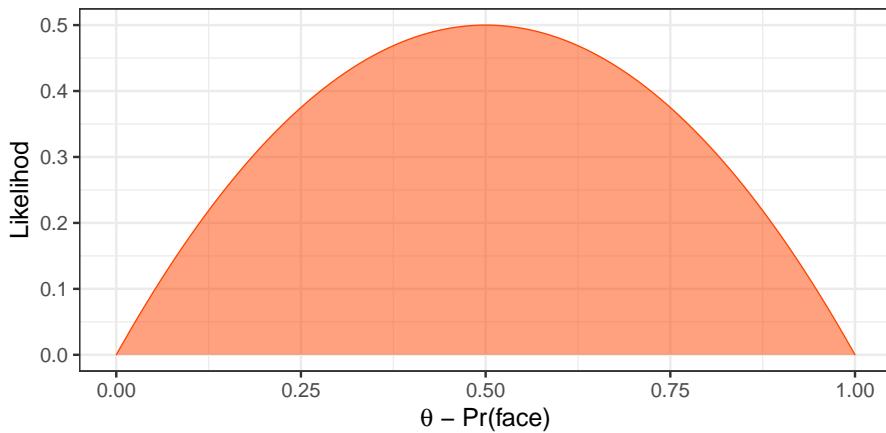


Figure 2.3 : Likelihood plot...

```
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 : $\mathcal{L}(\theta | \text{data}) = p(\text{data} | \theta)$.

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 probabilités. Dans l'analyse bayésienne, **les données sont considérées comme fixes** et la valeur de θ est considérée comme une **variable aléatoire**.

2.2.9 Définition du prior

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

Aspect sémantique → *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

2.2.10 La distribution Beta

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

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

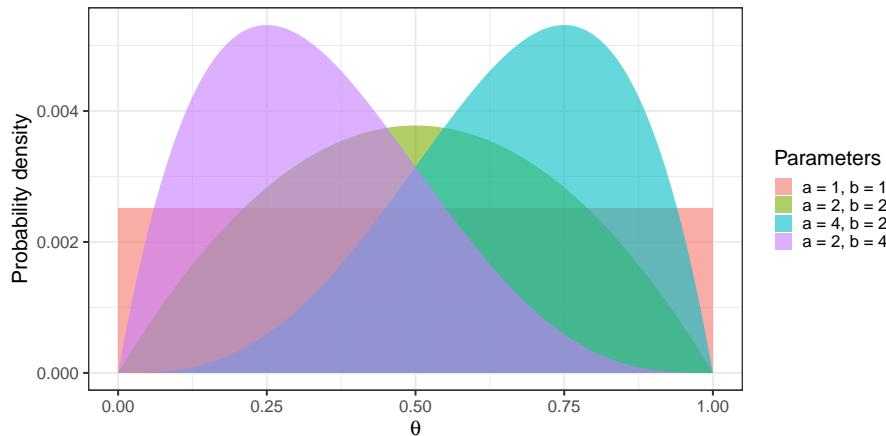


Figure 2.4 : Distribution Beta...

2.2.11 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 $a = b > 2$ (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)

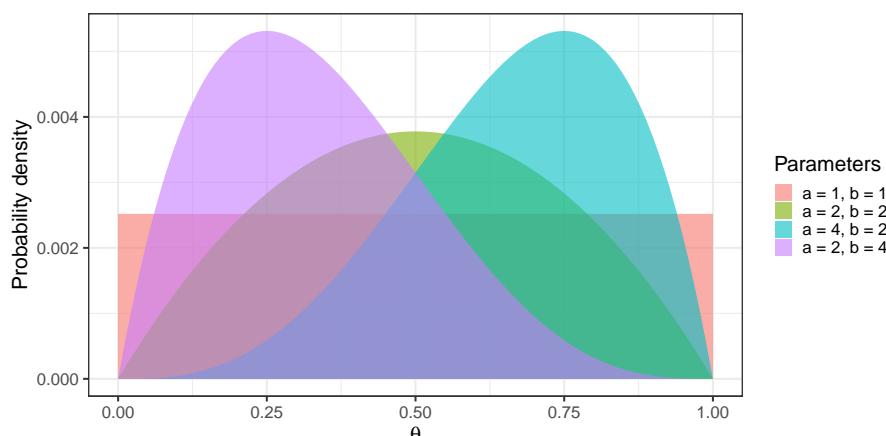


Figure 2.5 : Inteprétation des paramètres d'une distribution Beta.

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

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

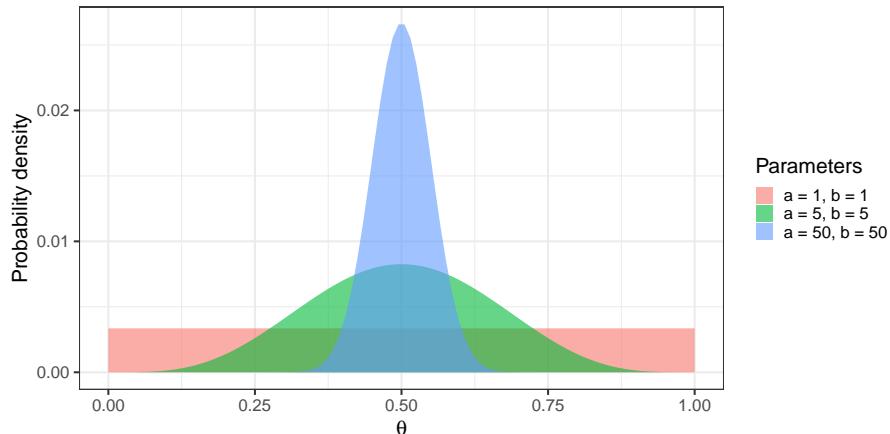


Figure 2.6 : Interprétation des paramètres d'une distribution Beta...

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 :

$$a = \omega(\kappa - 2) + 1$$

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

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

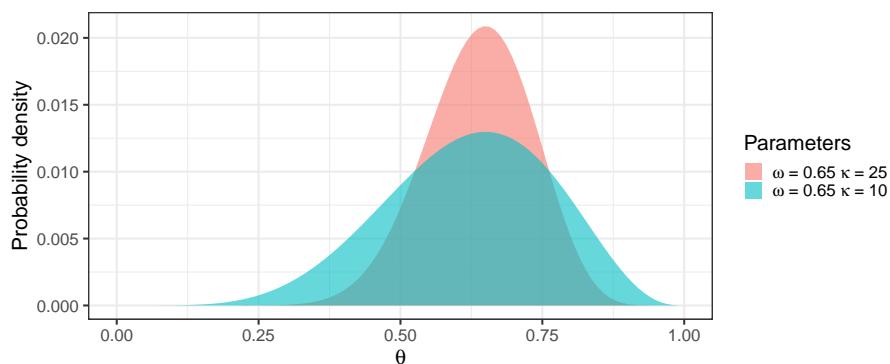


Figure 2.7 : Blah blah...

2.2.12 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 **conjugué** à \mathcal{F} si et seulement si :

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

(Gelman et al., 2013, p.35). En d'autres termes, un prior est appelé **conjugué** si, lorsqu'il est converti en une distribution a posteriori en étant multiplié par la 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 alors que la distribution Beta est **un prior conjugué** de la fonction de vraisemblance Binomiale.

2.2.13 Dérivation analytique de la distribution a posteriori

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

Soit une fonction de vraisemblance associée à y "Face" pour n lancers : $p(y | n, \theta) = \text{Bin}(y | n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$

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

2.2.14 Un exemple pour digérer

On observe $y = 7$ réponses correctes sur $n = 10$ questions. On choisit un prior 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 :

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

La moyenne de la distribution postérieure est donnée par :

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

Blah blah...

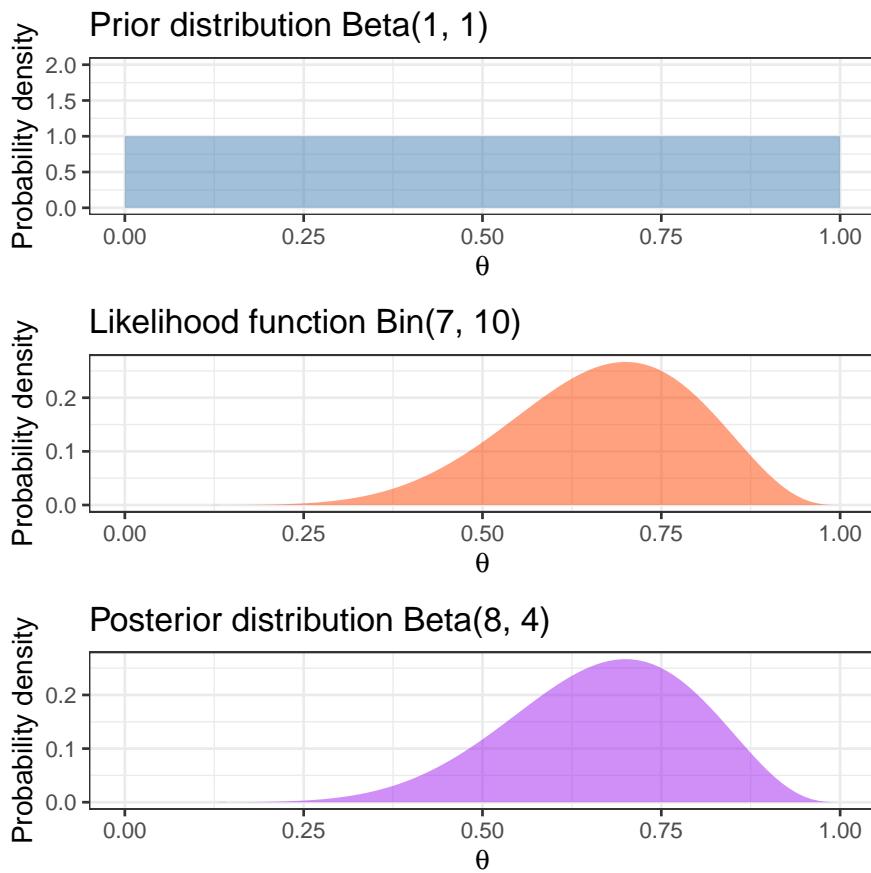


Figure 2.8 : Exemple...

2.2.15 Influence du prior sur la distribution postérieure

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

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

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

2.2.16 Ce qu'il faut retenir

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

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")...

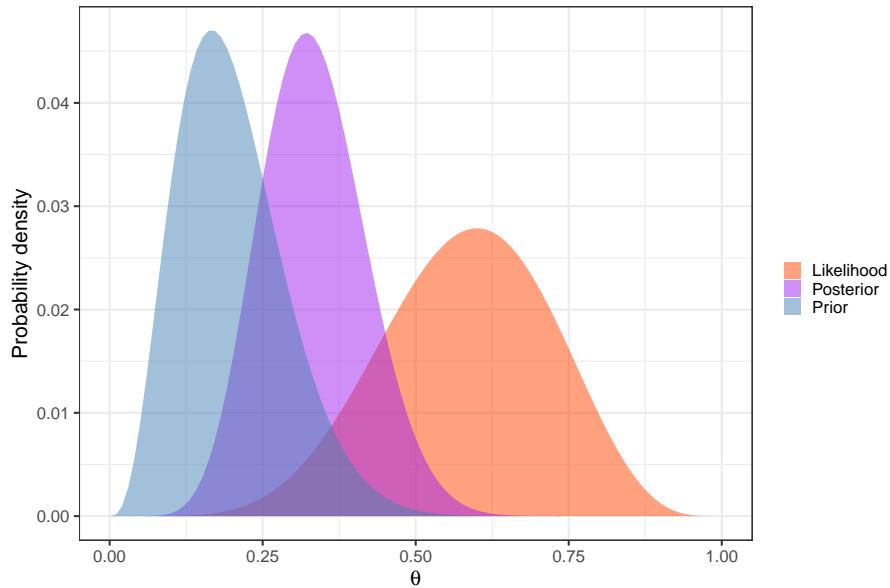


Figure 2.9 : Influence du prior...

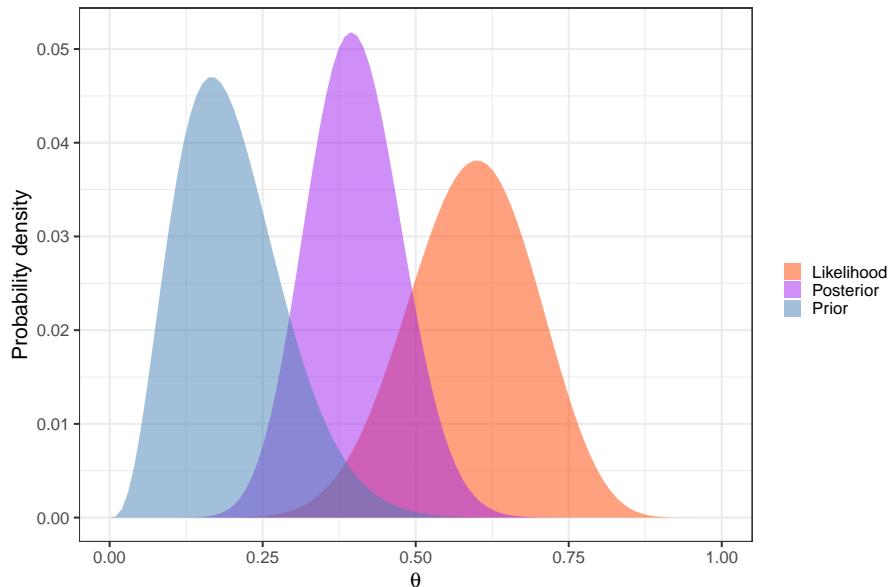


Figure 2.10 : Influence du prior...

2.2.17 La vraisemblance marginale (the devil is in the denominator)

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

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

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

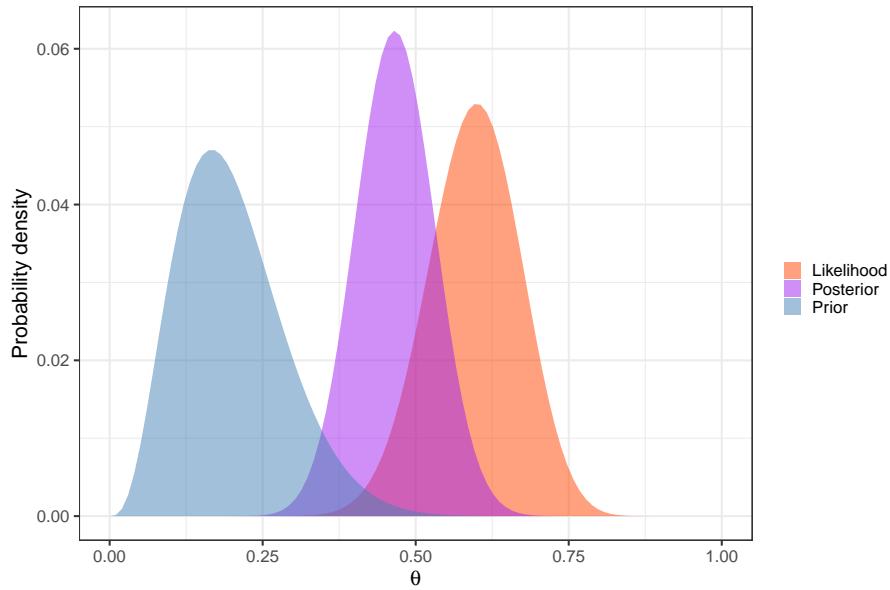


Figure 2.11 : Influence du prior...

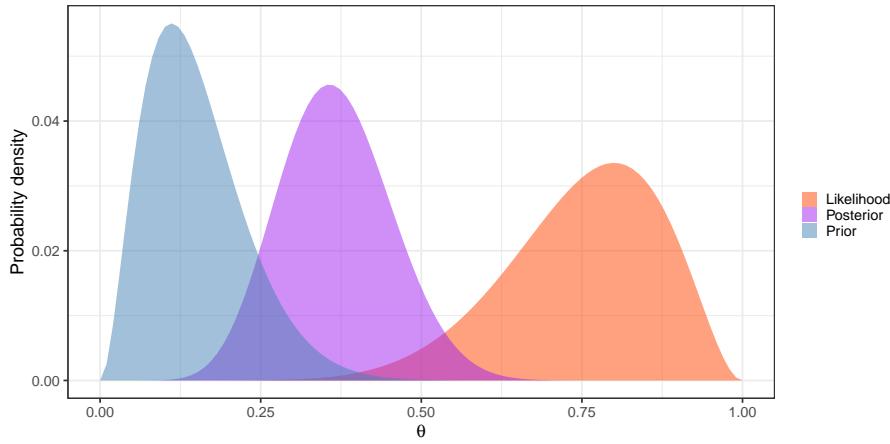


Figure 2.12 : Influence du prior...

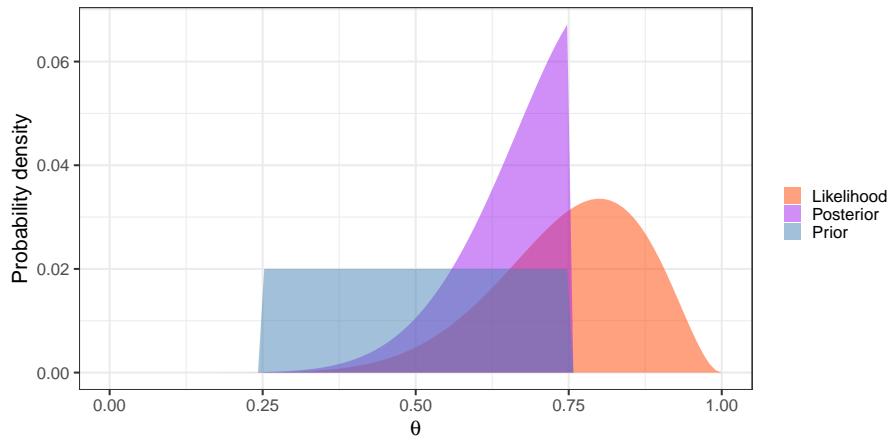


Figure 2.13 : Influence du prior...

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

$$p(\text{data}) = \int p(\text{data} | \theta) p(\theta) d\theta \quad \text{Application de la règle du produit}$$

54

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

Par exemple pour deux paramètres discrets :

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

Et pour un modèle avec deux paramètres continus :

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

Trois méthodes pour résoudre (contourner) ce problème :

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

2.2.18 La distribution postérieure, solution analytique

2.2.18.1 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]	$p(\hat{x} = 1) = \frac{\alpha'}{\alpha' + \beta'}$
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]	$\text{BetaBin}(\hat{x} \alpha', \beta')$ (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	$\text{NB}(\hat{x} k', \theta')$ (negative binomial)
			α, β ^[note 3]	$\alpha + \sum_{i=1}^n x_i, \beta + n$	α total occurrences in β intervals	$\text{NB}(\hat{x} \alpha', \frac{1}{1 + \beta'})$ (negative binomial)
Categorical	p (probability vector), k (number of categories; i.e., size of p)	Dirichlet	α	$\alpha + (c_1, \dots, c_k)$, where c_i is the number of observations in category i	$\alpha_i - 1$ occurrences of category i ^[note 1]	$p(\hat{x} = i) = \frac{\alpha_i'}{\sum_j \alpha_j'} = \frac{\alpha_i}{\alpha_i + c_i} = \frac{\alpha_i}{\sum_i \alpha_i + n}$
Multinomial	p (probability vector), k (number of categories; i.e., size of p)	Dirichlet	α	$\alpha + \sum_{i=1}^n x_i$	$\alpha_i - 1$ occurrences of category i ^[note 1]	$\text{DirMult}(\hat{x} \alpha')$ (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]	

Figure 2.14 : Illustration of the Mersenne-Twister algorithm... Figure from...

2.2.18.2 Distributions continues

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...

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	μ_0, σ_0^2	$\left(\frac{\mu_0 + \sum_{i=1}^n x_i}{\sigma_0^2 + n}, \frac{1}{\sigma_0^2 + n} \right)$	mean was estimated from observations with total precision (sum of all individual precisions) $1/\sigma_0^2$ and with sample mean μ_0	$\mathcal{N}(\bar{x} \mu_0, \sigma_0^2 + \sigma^2)^{\otimes}$
Normal with known precision τ	μ (mean)	Normal	μ_0, τ_0	$\left(\tau_0 \mu_0 + \tau \sum_{i=1}^n x_i \right) / (\tau_0 + n\tau), \tau_0 + n\tau$	mean was estimated from observations with total precision (sum of all individual precisions) τ_0 and with sample mean μ_0	$\mathcal{N}(\bar{x} \mu_0, \frac{1}{\tau_0} + \frac{1}{\tau})^{\otimes}$
Normal with known mean μ	σ^2 (variance)	Inverse gamma	α, β [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)	$t_{2\alpha'}(\bar{x} \mu, \sigma^2 = \beta'/\alpha')^{\otimes}$
Normal with known mean μ	σ^2 (variance)	Scaled inverse chi-squared	ν, σ_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	$t_{\nu'}(\bar{x} \mu, \sigma_0^2)^{\otimes}$
Normal with known mean μ	τ (precision)	Gamma	α, β [note 5]	$\alpha + \frac{n}{2}, \beta + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2}$	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)	$t_{2\alpha'}(\bar{x} \mu, \sigma^2 = \beta'/\alpha')^{\otimes}$
Normal ^[note 6]	μ and σ^2 Assuming exchangeability	Normal-inverse gamma	$\mu_0, \nu, \alpha, \beta$	$\begin{aligned} & \nu \mu_0 + n \bar{x}, \nu + n, \alpha + \frac{n}{2}, \\ & \beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n\nu}{\nu + n} \frac{(\bar{x} - \mu_0)^2}{2} \end{aligned}$	mean was estimated from n observations with sample mean μ_0 ; variance was estimated from 2α observations with sample mean μ_0 and sum of squared deviations 2β • \bar{x} is the sample mean	$t_{2\alpha'}(\bar{x} \mu', \frac{\beta'(\nu' + 1)}{\nu'\alpha'})^{\otimes}$

Figure 2.15 : Illustration of the Mersenne-Twister algorithm... Figure from...

2.2.19 La distribution postérieure, grid method

1. Définir la grille
2. Calculer la valeur du prior pour chaque valeur de la grille
3. Calculer la valeur de la vraisemblance pour chaque valeur de la grille
4. Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation du résultat

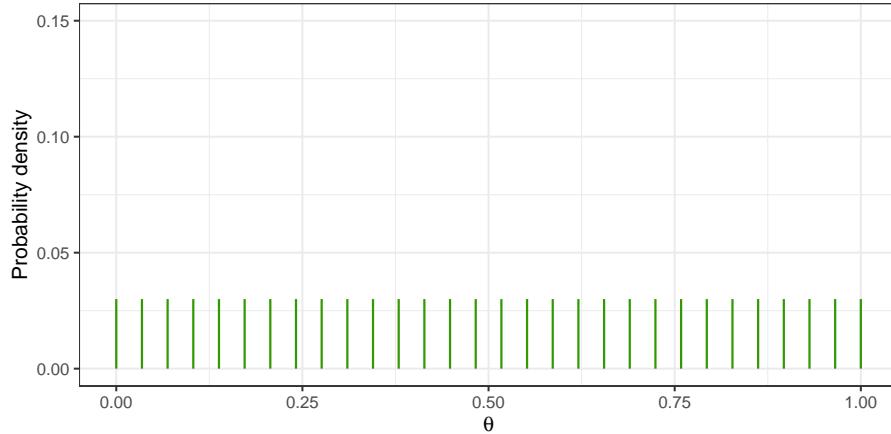


Figure 2.16 : Blah blah...

1. Définir la grille
2. Calculer la valeur du prior pour chaque valeur de la grille
3. Calculer la valeur de la vraisemblance pour chaque valeur de la grille
4. Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation du résultat

1. Définir la grille
2. Calculer la valeur du prior pour chaque valeur de la grille
3. Calculer la valeur de la vraisemblance pour chaque valeur de la grille
4. Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation du résultat

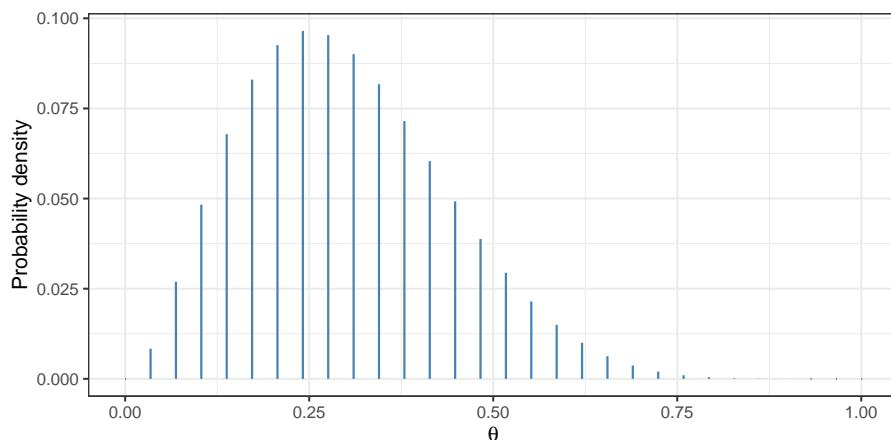


Figure 2.17 : Blah blah...

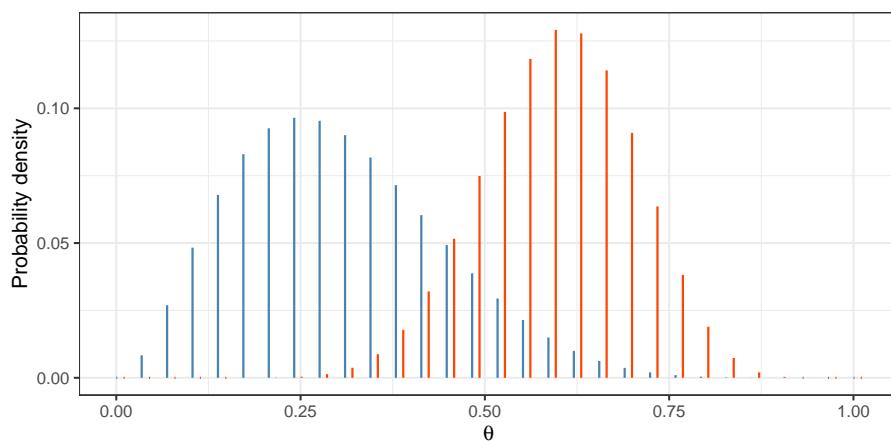


Figure 2.18 : Blah blah...

1. Définir la grille
2. Calculer la valeur du prior pour chaque valeur de la grille
3. Calculer la valeur de la vraisemblance pour chaque valeur de la grille
- 4. Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation du résultat**

1. Définir la grille
2. Calculer la valeur du prior pour chaque valeur de la grille
3. Calculer la valeur de la vraisemblance pour chaque valeur de la grille
- 4. Calculer le produit prior x vraisemblance pour chaque valeur de la grille, puis normalisation du résultat**

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

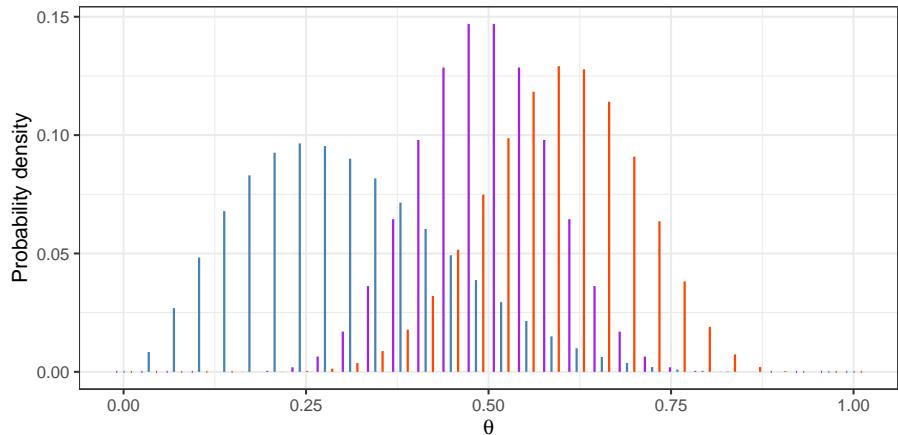


Figure 2.19 : Blah blah...

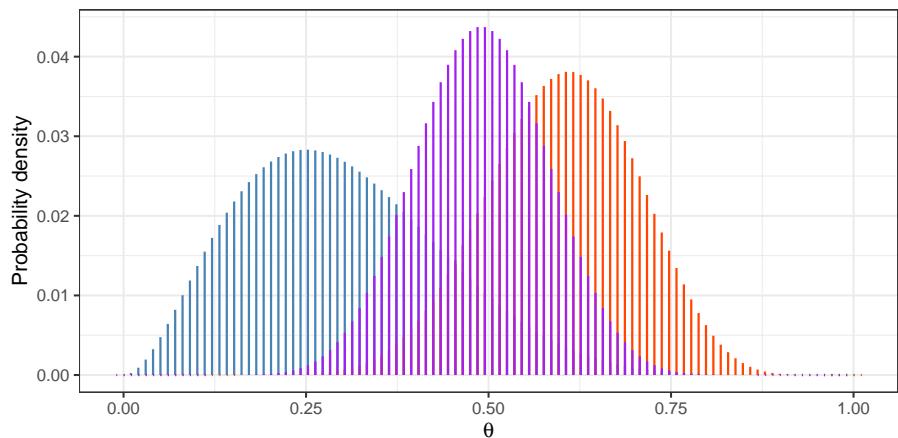


Figure 2.20 : Blah blah...

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)...

2.2.20 Échantillonner la distribution postérieure

Pour échantillonner une distribution postérieure, on peut utiliser une approximation par grille ou différentes implémentations des méthodes MCMC (e.g., Metropolis-Hastings, Gibbs, Hamilton, cf. Cours n°05).

En pratique :

```
p_grid <- seq(from = 0, to = 1, length.out = 1000) # creating a grid
prior <- rep(1, 1000) # uniform prior
likelihood <- dbinom(y, size = n, prob = p_grid) # computes likelihood
posterior <- (likelihood * prior) / sum(likelihood * prior) # computes posterior
samples <- sample(posterior, size = 1e3, prob = posterior, replace = TRUE) # sampling
hist(samples, main = "", xlab = expression(theta), cex.axis = 1, cex.lab = 1.5) # histogram
```

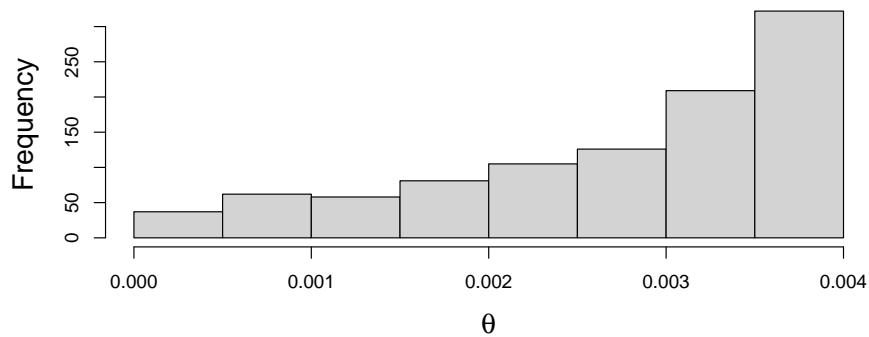


Figure 2.21 : Blah blah...

La précision dépend de la taille de l'échantillon...

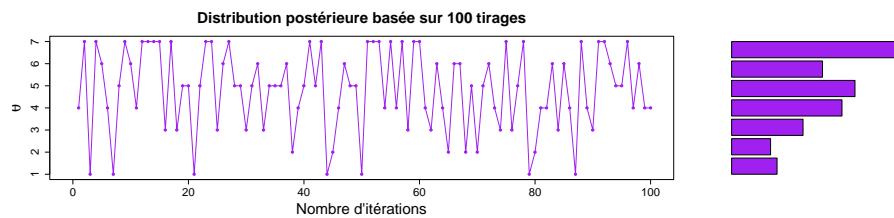


Figure 2.22 : Blah blah...

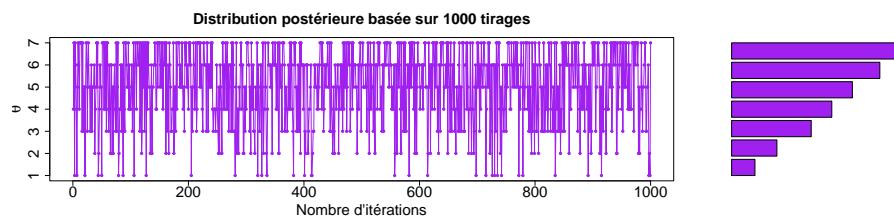


Figure 2.23 : Blah blah...

2.2.21 La distribution postérieure, résumé

- Cas analytique :

```
p_grid <- seq(from = 0, to = 1, length.out = 1000)
a <- b <- 1 # parameters of the Beta prior
n <- 9 # number of observations
y <- 6 # number of successes
posterior <- dbeta(p_grid, z + a - 1, N - z + b - 1)
```

- Grid method :

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

- Échantillonner la distribution postérieure :

```
sample(data, size = trajLength, prob = prob, replace = TRUE)
```

Méthode analytique * La distribution postérieure est décrite explicitement * Le modèle est fortement contraint

Méthode Grid * 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

2.2.22 Utiliser les échantillons pour résumer la distribution postérieure

2.2.22.1 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.

```
mode_posterior <- find_mode(samples) # in blue
mean_posterior <- mean(samples) # in orange
median_posterior <- median(samples) # in green
```

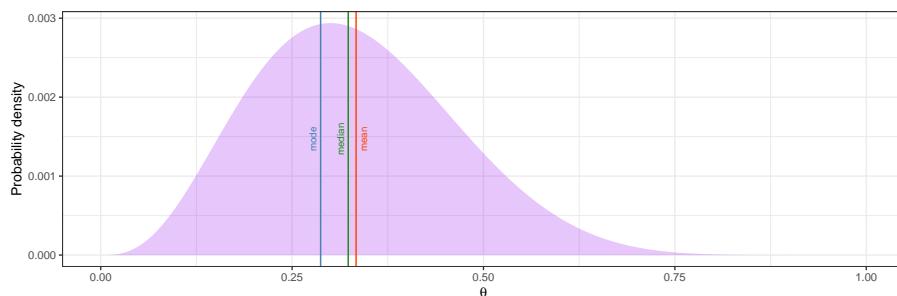


Figure 2.24 : Blah blah...

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

```
sum(samples > 0.5) / length(samples) # length(samples) is the number of samples
```

```
## [1] 0.112
```

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

```
sum(samples > 0.2 & samples < 0.4) / 1e4 # length(samples) is the number of samples
## [1] 0.5482
```

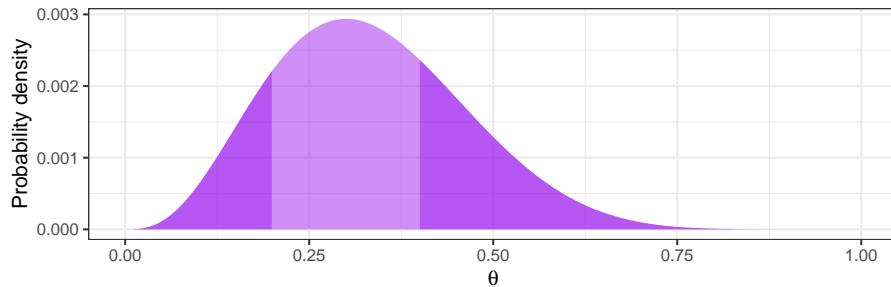


Figure 2.25 : Blah blah...

2.2.23 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 :

$$\int_{\theta : p(\theta) > W} p(\theta) d\theta = 0.89.$$

```
library(BEST)

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(p_grid, size = 1e4, replace = TRUE, prob = pTheta)

plotPost(samples, credMass = 0.89, cex = 1.5, xlab = expression(theta), xlim = c(0, 1))
```

2.2.24 Region of practical equivalence (ROPE)

On l'utilise pour tester une hypothèse :

- La valeur du paramètre (e.g., $\theta = 0.5$) est rejetée si le HDI est entièrement hors de la ROPE

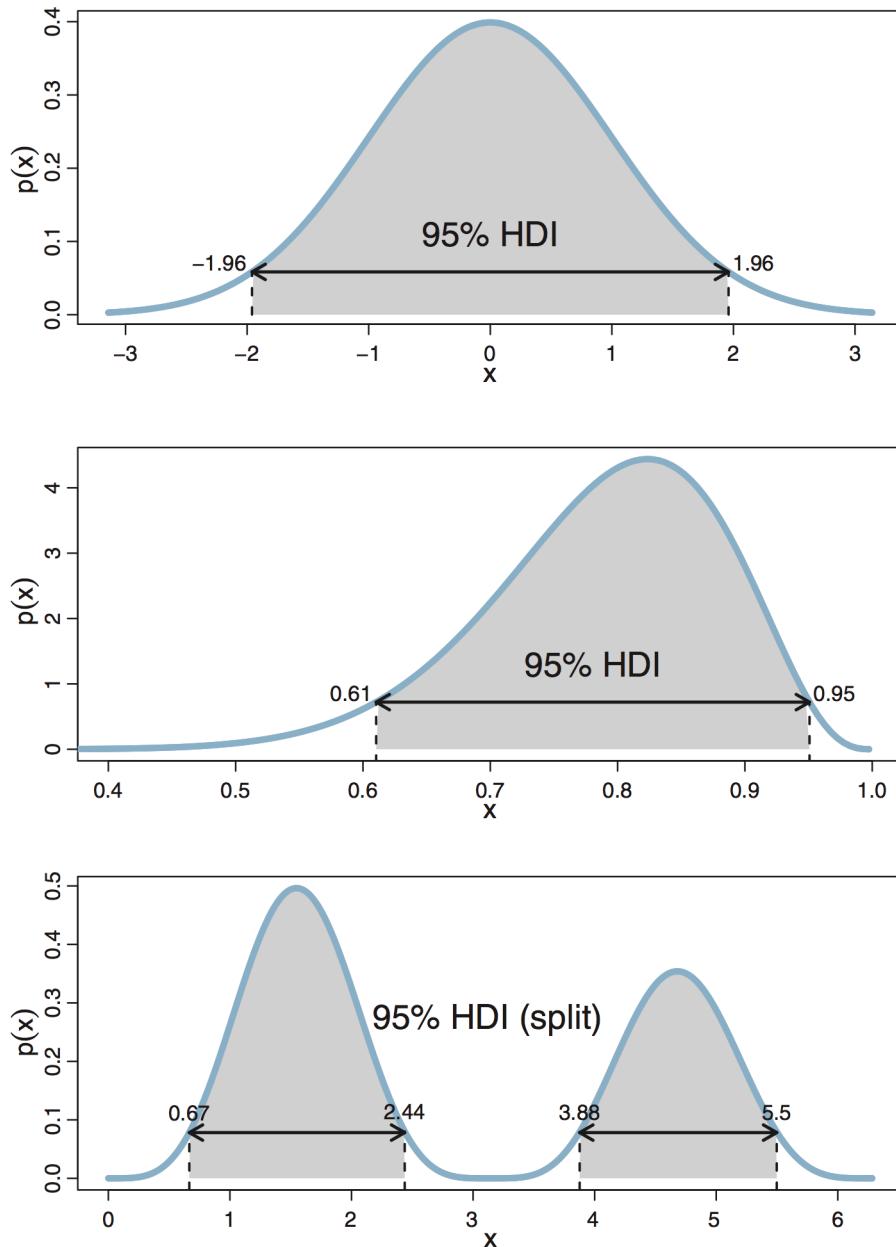


Figure 2.26 : Illustration of the Mersenne-Twister algorithm... Figure from...

- 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...

2.2.25 Model checking

Les deux rôles de la fonction de vraisemblance :

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

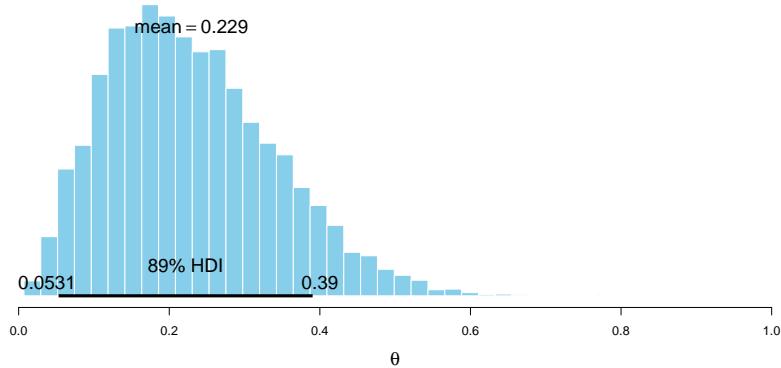


Figure 2.27 : Blah blah...

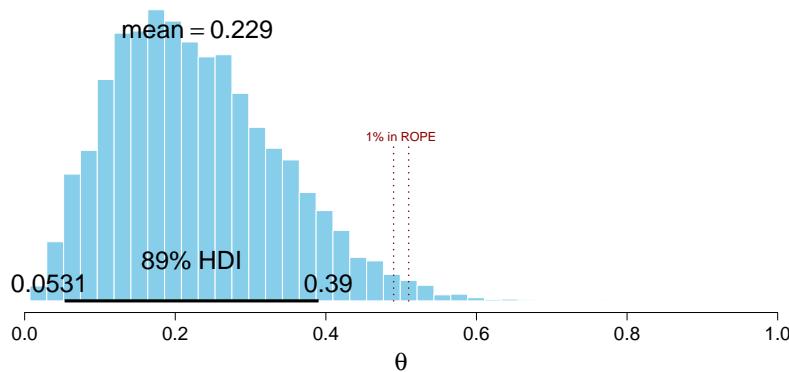


Figure 2.28 : Blah blah...

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

Par exemple : Générer 10000 valeurs à partir d'une loi binomiale basée sur 9 lancers et une probabilité de Face de 0.6 :

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

Deux sources d'incertitude dans ces prédictions :

- Incertitude liée au processus d'échantillonnage -> Chaque valeur apparaît avec une probabilité θ
- Incertitude sur la valeur de θ elle-même -> Pour chaque valeur de θ on peut calculer une distribution implicite

Par exemple : Générer 10000 valeurs à partir d'une loi binomiale basé sur 9 lancers et une probabilité de Face décrite par la distribution postérieure de θ :

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

2.2.26 Posterior predictive checking

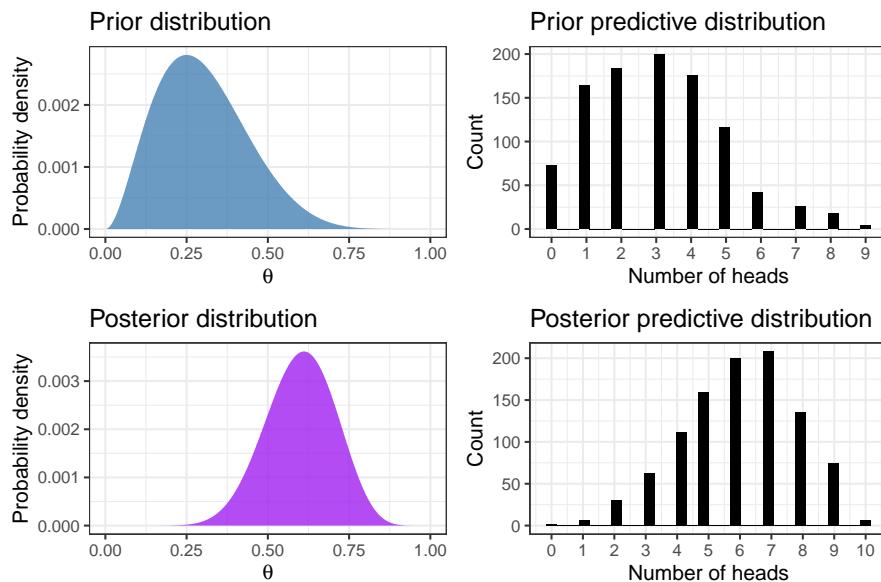


Figure 2.29 : Blah blah...

Blah blah...

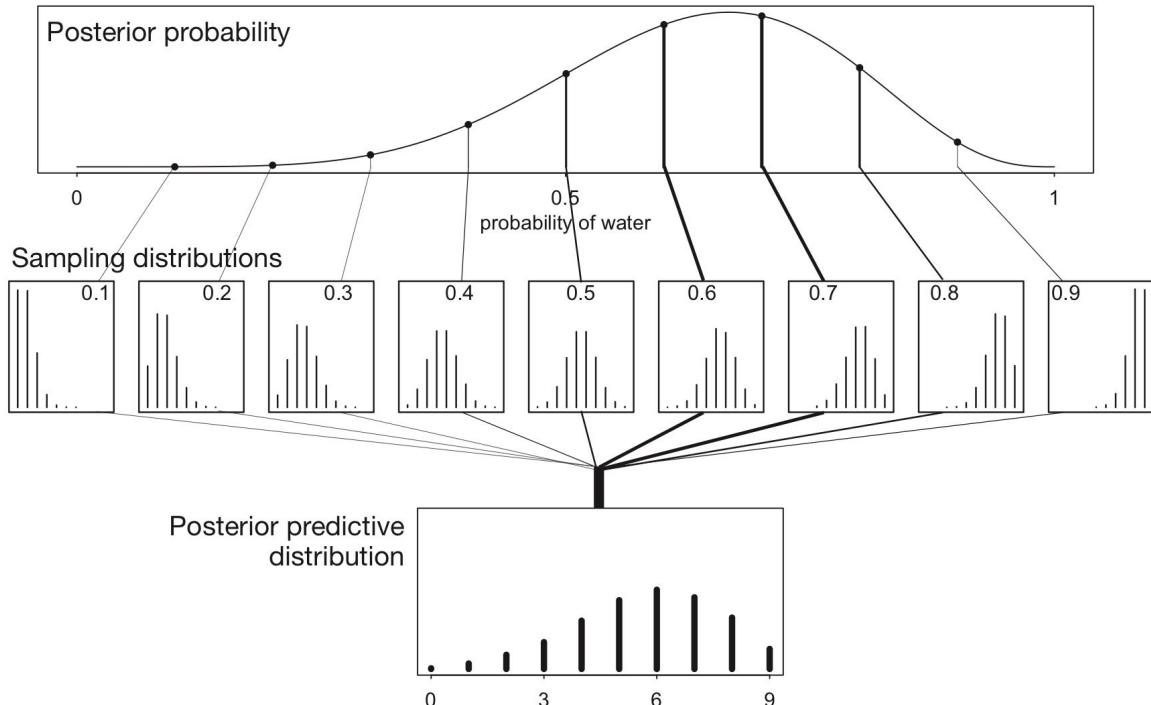


Figure 2.30 : Illustration of the posterior predictive checking procedure. Figure from McElreath (2016).

2.3 Conclusions

...

Modèle de régression linéaire

Introduction au chapitre blah blah...

3.1 Langage de la modélisation

$$\begin{aligned}y_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta x_i \\ \alpha &\sim \text{Normal}(60, 10) \\ \beta &\sim \text{Normal}(0, 10) \\ \sigma &\sim \text{HalfCauchy}(0, 1)\end{aligned}$$

Objectif de la séance : comprendre ce type de modèle.

Les constituants de nos modèles seront toujours les mêmes et nous suivrons les deux mêmes étapes :

- Construire le modèle (*likelihood + priors*).
- Mettre à jour grâce aux données (*updating*), afin de calculer la distribution postérieure.

3.2 Un premier modèle

```
library(rethinking)
library(tidyverse)

data(Howell1)
d <- Howell1
str(d)
```

```
## 'data.frame': 544 obs. of 4 variables:  
## $ height: num 152 140 137 157 145 ...  
## $ weight: num 47.8 36.5 31.9 53 41.3 ...  
## $ age   : num 63 63 65 41 51 35 32 27 19 54 ...  
## $ male  : int 1 0 0 1 0 1 0 1 0 1 ...
```

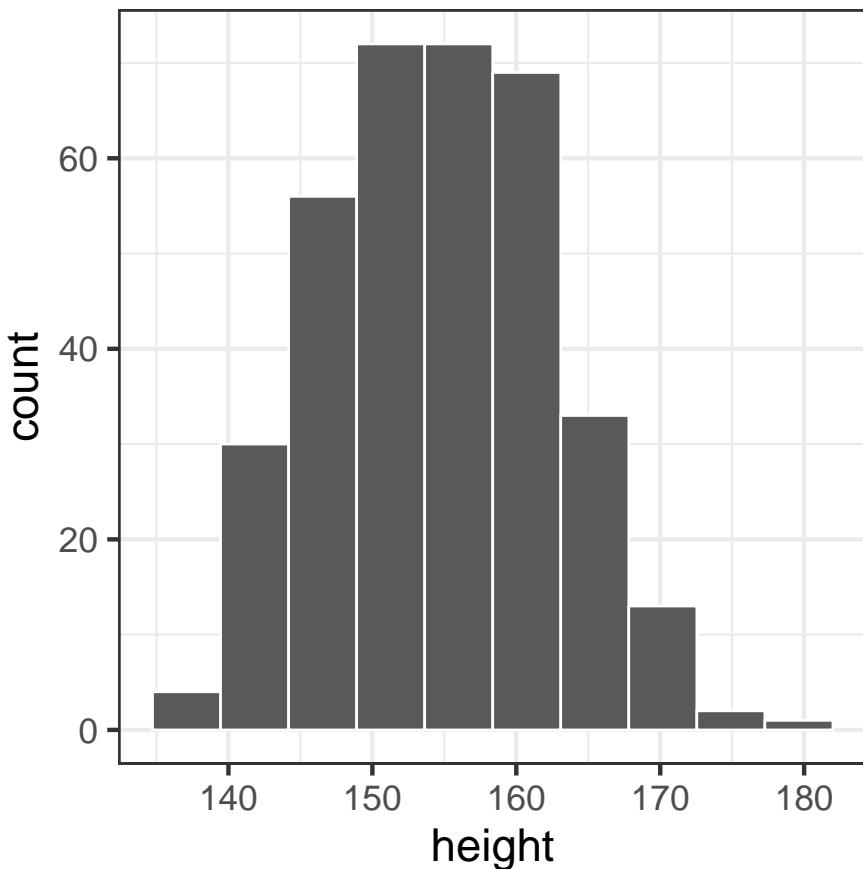
```
d2 <- d %>% filter(age >= 18)  
head(d2)
```

```
##      height    weight  age male  
## 1 151.765 47.82561 63    1  
## 2 139.700 36.48581 63    0  
## 3 136.525 31.86484 65    0  
## 4 156.845 53.04191 41    1  
## 5 145.415 41.27687 51    0  
## 6 163.830 62.99259 35    1
```

...

$$h_i \sim \text{Normal}(\mu, \sigma)$$

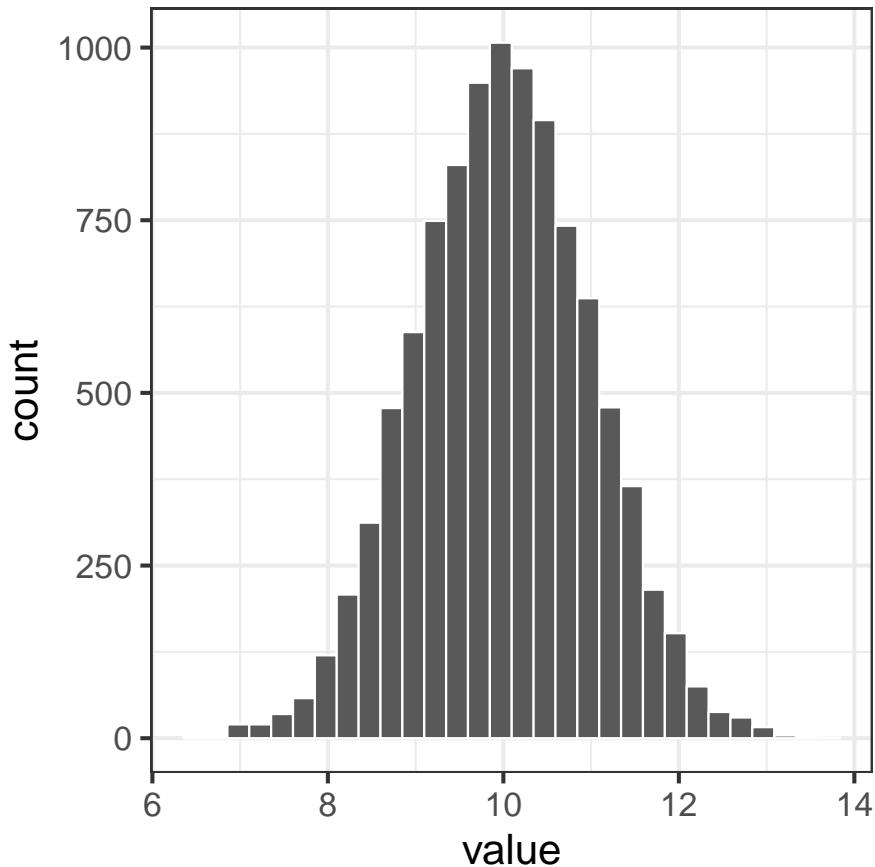
```
d2 %>%  
  ggplot(aes(x = height) ) +  
  geom_histogram(bins = 10, col = "white") +  
  theme_bw(base_size = 18)
```



3.3 Loi normale

$$p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2}(\mu - x)^2 \right]$$

```
data.frame(value = rnorm(1e4, 10, 1)) %>% # 10000 samples from Normal(10, 1)
  ggplot(aes(x = value)) +
  geom_histogram(col = "white") +
  theme_bw(base_size = 20)
```



3.3.1 D'où vient la loi normale?

Certaines valeurs sont fortement probables (autour de la moyenne μ). Plus on s'éloigne, moins les valeurs sont probables (en suivant une décroissance exponentielle).

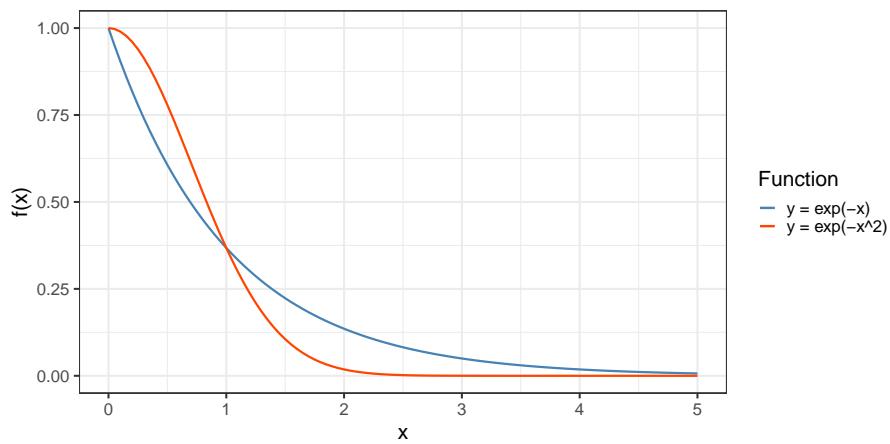


Figure 3.1 : blah blah...

$$y = \exp[-x^2]$$

On étend notre fonction aux valeurs négatives.

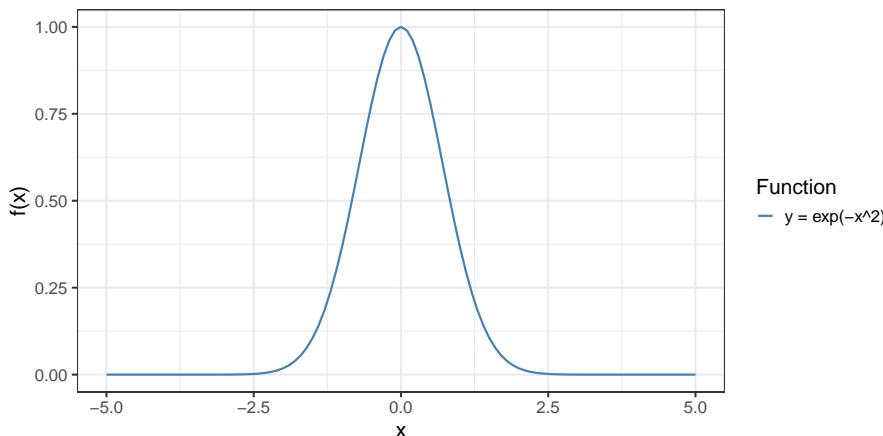


Figure 3.2 : blah blah...

$$y = \exp[-x^2]$$

Les points d'inflection nous donnent une bonne indication de là où la plupart des valeurs se trouvent (i.e., entre les points d'inflection). Les pics de la dérivée nous montrent les points d'inflection.

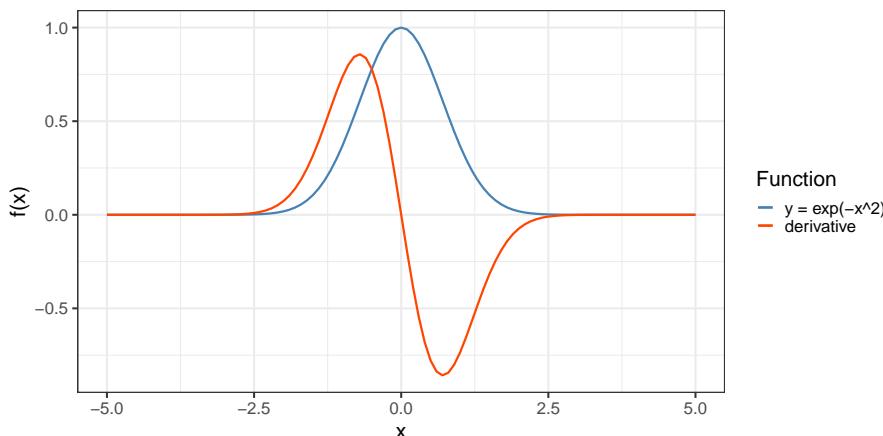


Figure 3.3 : blah blah...

$$y = \exp\left[-\frac{1}{2}x^2\right]$$

Ensuite on standardise la distribution de manière à ce que les deux points d'inflection se trouvent à $x = -1$ et $x = 1$.

$$y = \exp\left[-\frac{1}{2\sigma^2}x^2\right]$$

On insère un paramètre σ^2 pour contrôler la distance entre les points d'inflection.

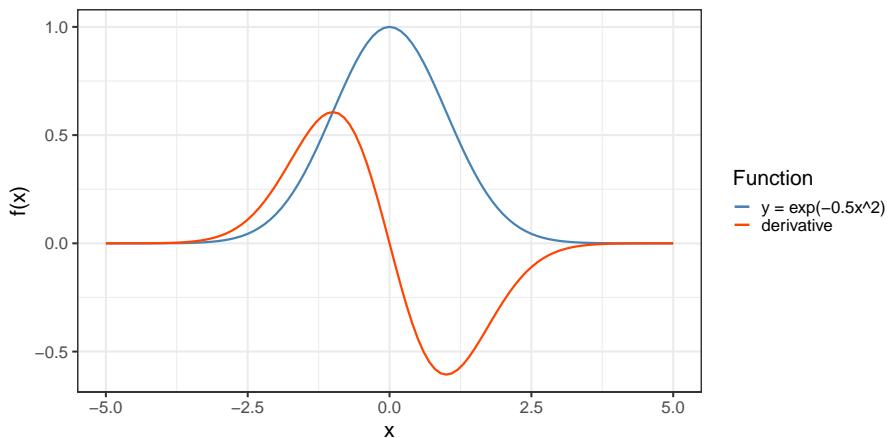


Figure 3.4 : blah blah...

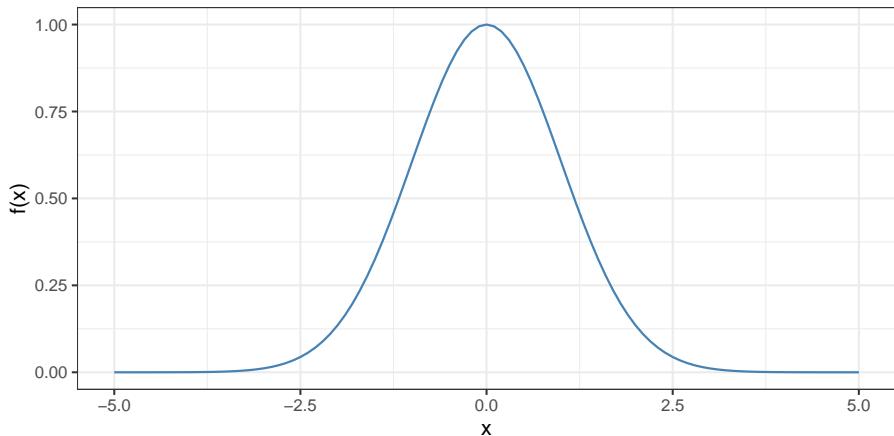


Figure 3.5 : blah blah...

$$y = \exp \left[-\frac{1}{2\sigma^2} (x - \mu)^2 \right]$$

On insère ensuite un paramètre μ afin de pouvoir contrôler la position (la tendance centrale) de la distribution.

$$y = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2} (\mu - x)^2 \right]$$

Mais... cette distribution n'intègre pas à 1. On divise donc par une constante de normalisation (la partie gauche), afin d'obtenir une distribution de probabilité.

3.4 Modèle gaussien

Nous allons construire un modèle de régression, mais avant d'ajouter un prédicteur, essayons de modéliser la distribution des tailles.

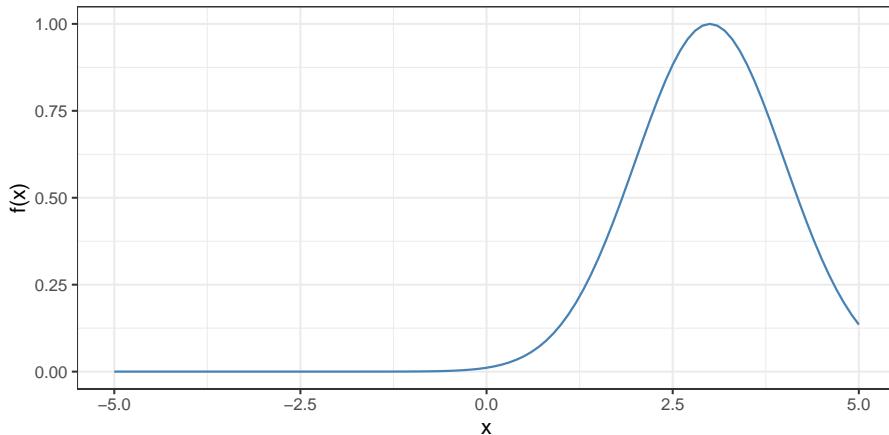


Figure 3.6 : blah blah...

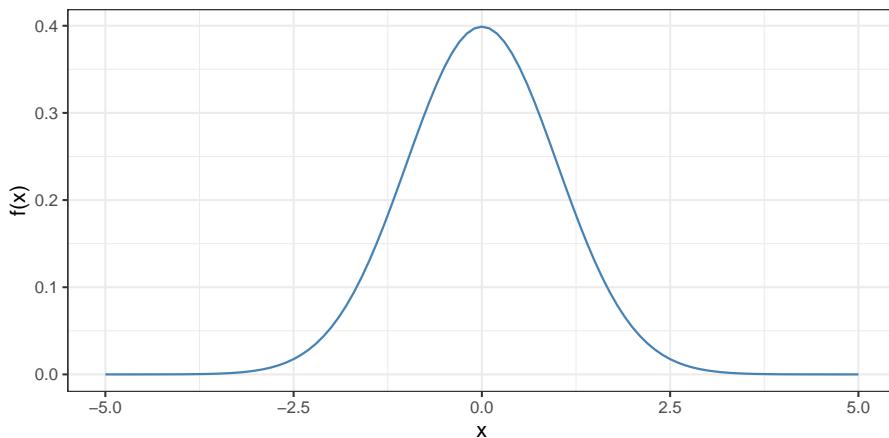


Figure 3.7 : blah blah...

On cherche à savoir quel est le modèle (la distribution) qui décrit le mieux la répartition des tailles. On va donc explorer toutes les combinaisons possibles de μ et σ et les classer par leurs probabilités respectives.

Notre but, une fois encore, est de décrire **la distribution postérieure**, qui sera donc d'une certaine manière **une distribution de distributions**.

On définit ensuite $p(\mu, \sigma)$, la distribution a priori conjointe de tous les paramètres du modèle. On peut spécifier ces priors indépendamment pour chaque paramètre, sachant que $p(\mu, \sigma) = p(\mu)p(\sigma)$.

$$\mu \sim \text{Normal}(178, 20)$$

On définit ensuite $p(\mu, \sigma)$, la distribution a priori conjointe de tous les paramètres du modèle. On peut spécifier ces priors indépendamment pour chaque paramètre, sachant que $p(\mu, \sigma) = p(\mu)p(\sigma)$.

$$\sigma \sim \text{Uniform}(0, 50)$$

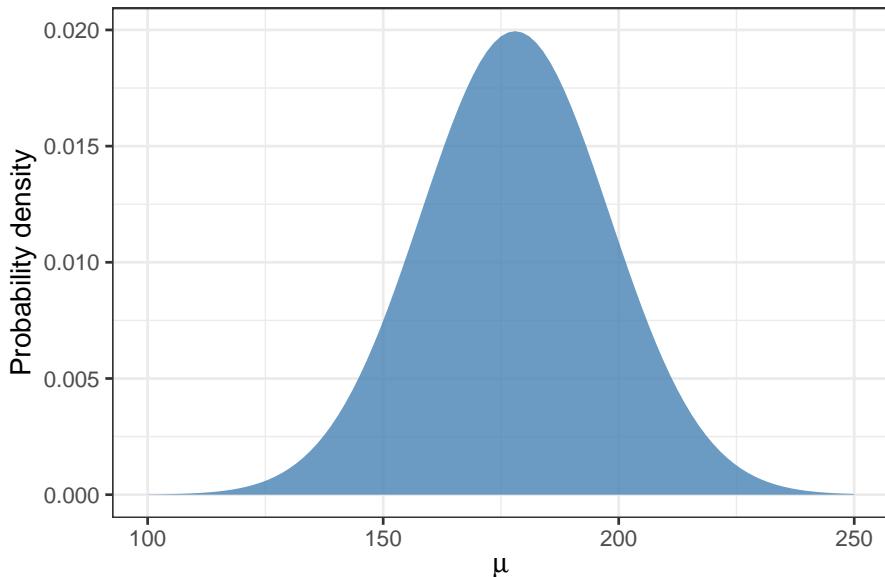


Figure 3.8 : blah blah...

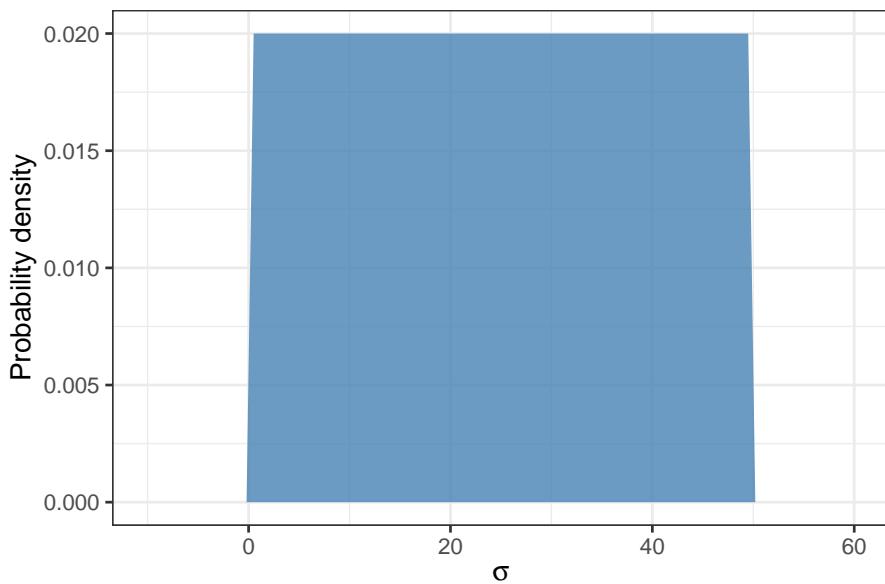


Figure 3.9 : blah blah...

3.5 Visualiser le prior

```
library(ks)
sample_mu <- rnorm(1e4, 178, 20) # prior on mu
sample_sigma <- runif(1e4, 0, 50) # prior on sigma
prior <- data.frame(cbind(sample_mu, sample_sigma)) # multivariate prior
H.scv <- Hscv(x = prior, verbose = TRUE)
fhat_prior <- kde(x = prior, H = H.scv, compute.cont = TRUE)
```

```

plot(
  fhat_prior, display = "persp", col = "steelblue", border = NA,
  xlab = "\nmu", ylab = "\nsigma", zlab = "\n\np(mu, sigma)",
  shade = 0.8, phi = 30, ticktype = "detailed",
  cex.lab = 1.2, family = "Helvetica")

knitr::include_graphics("figures/prior.png")

```

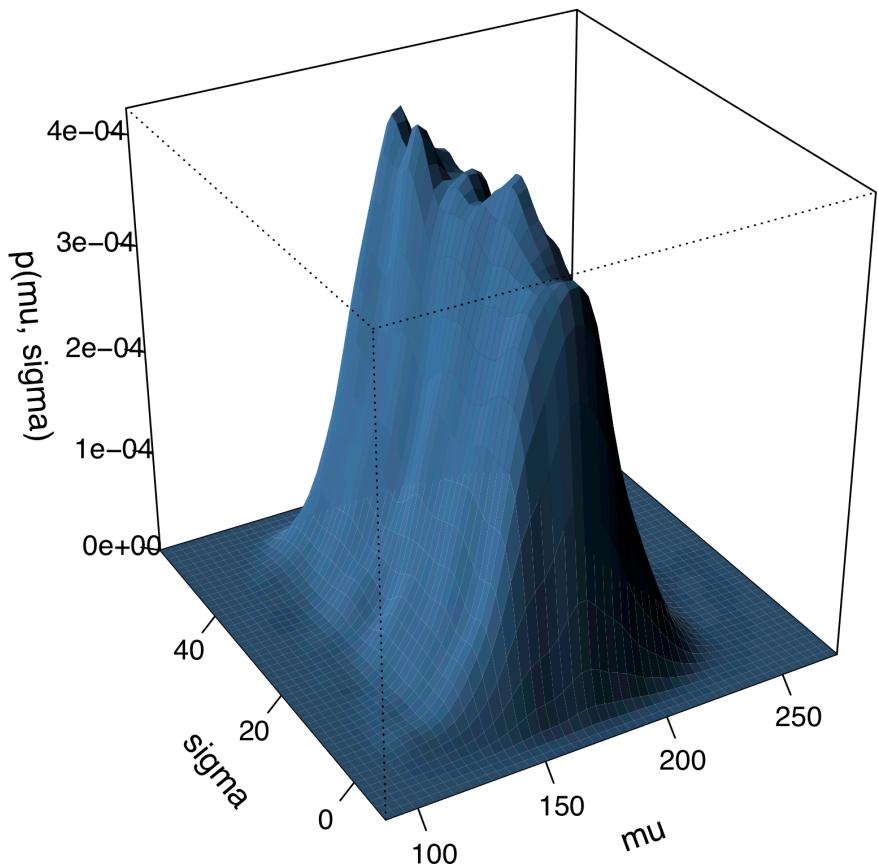


Figure 3.10 : blah blah...

3.6 Échantillonner à partir du prior

```

sample_mu <- rnorm(1000, 178, 20)
sample_sigma <- runif(1000, 0, 50)

data.frame(x = rnorm(1000, sample_mu, sample_sigma) ) %>%
  ggplot(aes(x) ) +
  geom_histogram() +
  xlab(expression(y[i])) +
  theme_bw(base_size = 20)

```

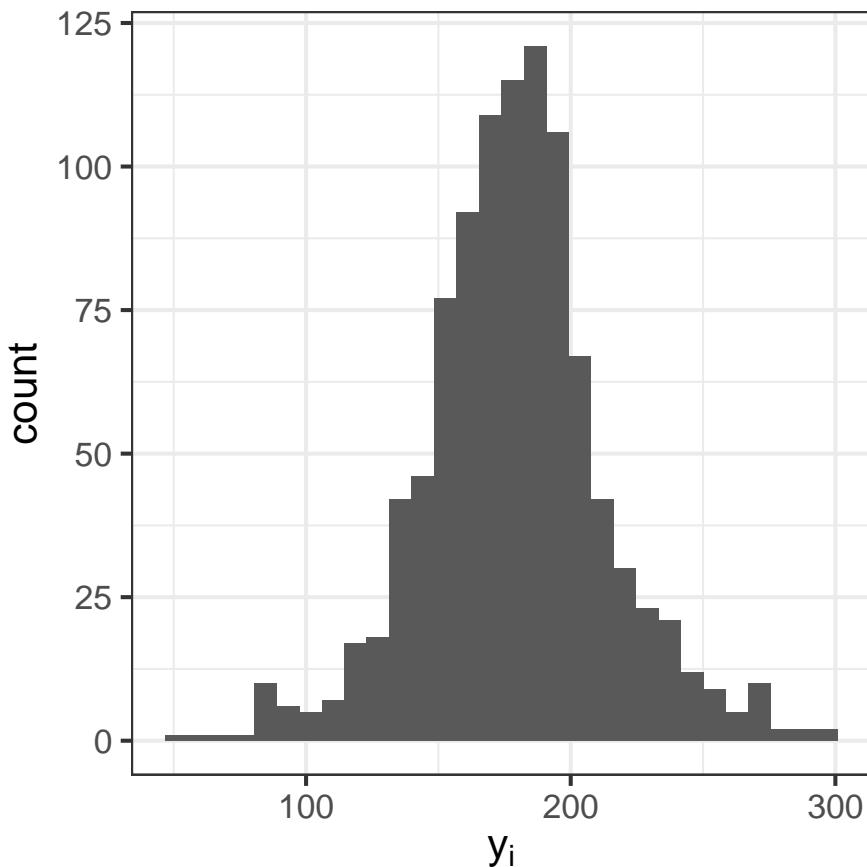


Figure 3.11 : blah blah...

3.7 Fonction de vraisemblance

```
mu_exemple <- 151.23
sigma_exemple <- 23.42

d2$height[34] # one observation

## [1] 162.8648
```

On veut calculer la probabilité d'observer une certaine valeur de taille, sachant certaines valeurs de μ et σ , c'est à dire :

$$p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2}(\mu - x)^2 \right]$$

On peut calculer cette *densité de probabilité* à l'aide des fonctions `dnorm`, `dbeta`, `dt`, `dexp`, `dgamma`, etc.

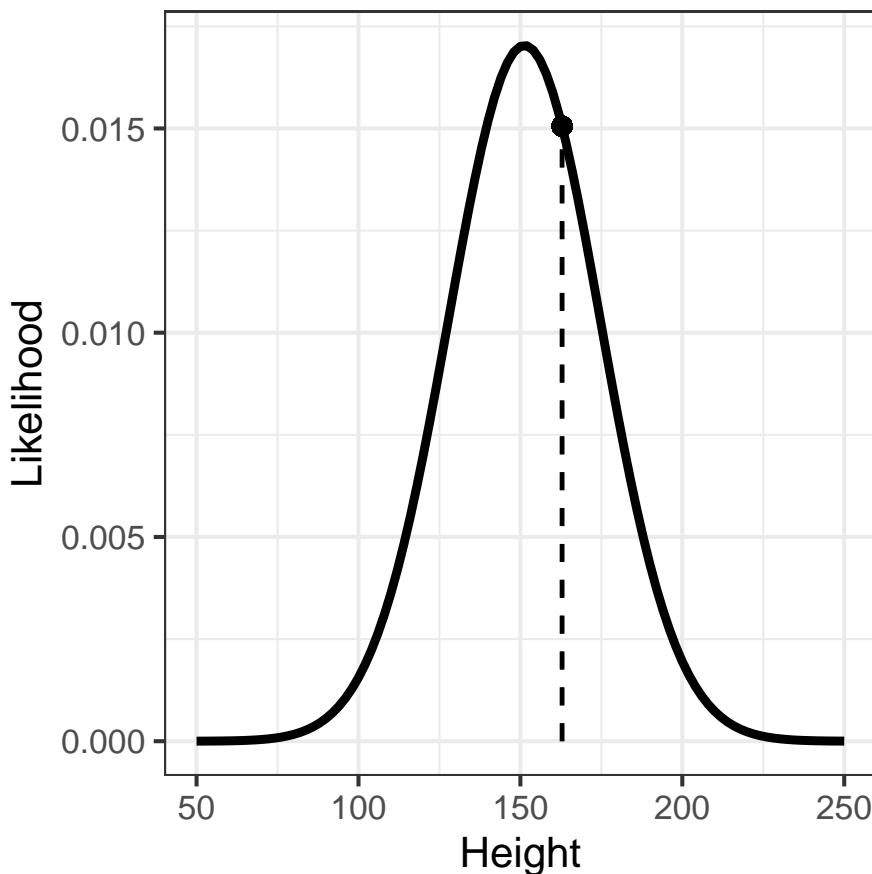


Figure 3.12 : blah blah...

```
dnorm(d2$height[34], mu_exemple, sigma_exemple)
```

```
## [1] 0.01505675
```

$$p(x | \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{1}{2\sigma^2}(\mu - x)^2 \right]$$

Ou à la main...

```
normal_likelihood <- function (x, mu, sigma) {

  bell <- exp( (- 1 / (2 * sigma^2)) * (mu - x)^2 )
  norm <- sqrt(2 * pi * sigma^2)

  return(bell / norm)
}
```

```
normal_likelihood(d2$height[34], mu_exemple, sigma_exemple)  
  
## [1] 0.01505675
```

3.8 Distribution postérieure

$$p(\mu, \sigma | h) = \frac{\prod_i \text{Normal}(h_i | \mu, \sigma) \text{Normal}(\mu | 178, 20) \text{Uniform}(\sigma | 0, 50)}{\int \int \prod_i \text{Normal}(h_i | \mu, \sigma) \text{Normal}(\mu | 178, 20) \text{Uniform}(\sigma | 0, 50) d\mu d\sigma}$$

$$p(\mu, \sigma | h) \propto \prod_i \text{Normal}(h_i | \mu, \sigma) \text{Normal}(\mu | 178, 20) \text{Uniform}(\sigma | 0, 50)$$

Il s'agit de la même formule vue lors des cours 1 et 2, mais cette fois en considérant qu'il existe plusieurs observations de taille (h_i), et deux paramètres à estimer μ et σ .

Pour calculer la **vraisemblance marginale** (en vert), il faut donc intégrer sur deux paramètres : μ et σ .

On réalise ici encore que la probabilité a posteriori est proportionnelle au produit de la vraisemblance et du prior.

3.8.1 Distribution postérieure - grid approximation

```
# définit une grille de valeurs possibles pour mu et sigma  
mu.list <- seq(from = 140, to = 160, length.out = 200)  
sigma.list <- seq(from = 4, to = 9, length.out = 200)  
  
# étend la grille en deux dimensions (chaque combinaison de mu et sigma)  
post <- expand.grid(mu = mu.list, sigma = sigma.list)  
  
# calcul de la log-vraisemblance (pour chaque couple de mu et sigma)  
post$LL <-  
  sapply(  
    1:nrow(post),  
    function(i) sum(dnorm(  
      d2$height,  
      mean = post$mu[i],  
      sd = post$sigma[i],  
      log = TRUE)  
    ))
```

```

)
# calcul de la probabilité a posteriori (non normalisée)
post$prod <-
  post$LL +
  dnorm(post$mu, 178, 20, log = TRUE) +
  dunif(post$sigma, 0, 50, log = TRUE)

# on "annule" le log en avec exp() et on standardise par la valeur maximale
post$prob <- exp(post$prod - max(post$prod) )

sample.rows <- sample(1:nrow(post), size = 1e4, replace = TRUE, prob = post$prob)

```

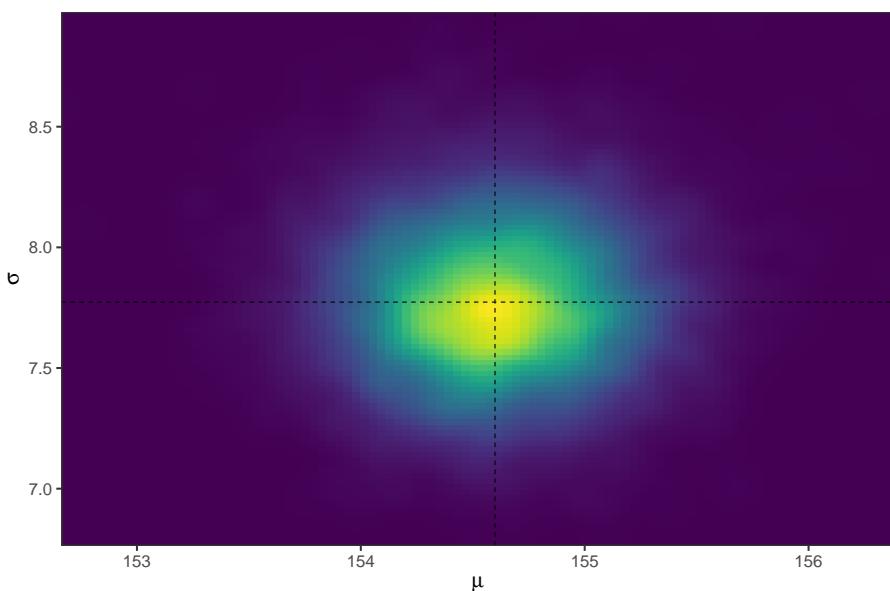


Figure 3.13 : blah blah...

3.8.2 Distribution postérieure - distributions marginales

```

BEST::plotPost(
  sample.mu, breaks = 40, xlab = expression(mu)
)

```

```

BEST::plotPost(
  sample.sigma, breaks = 40, xlab = expression(sigma)
)

```

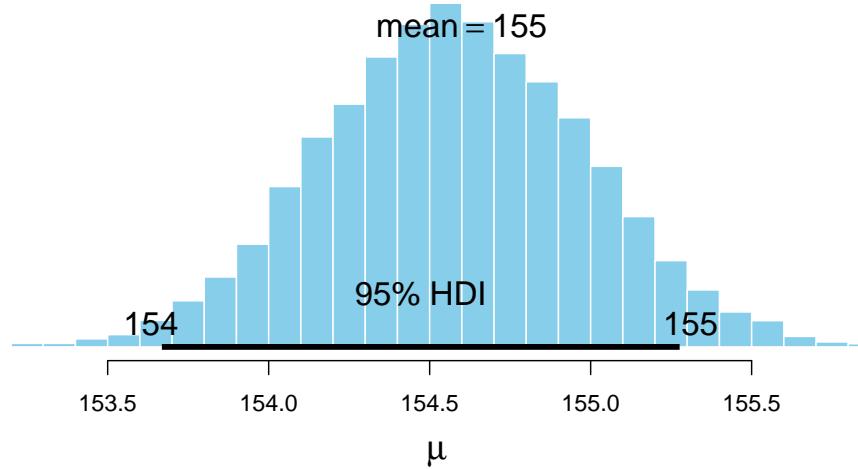


Figure 3.14 : blah blah...

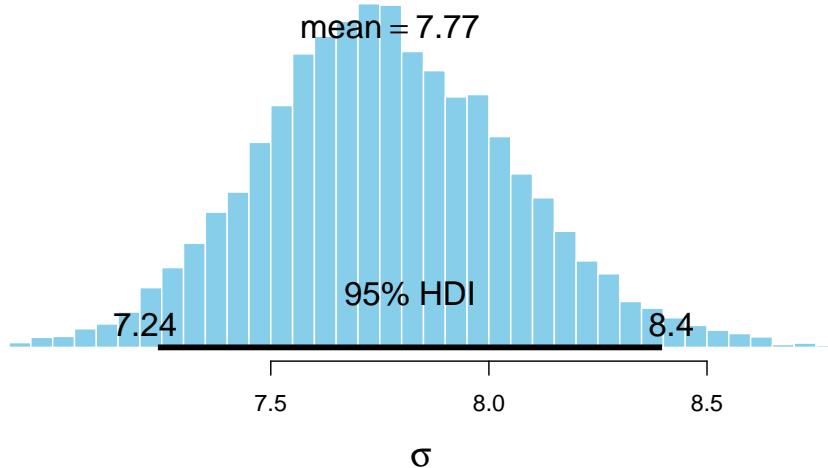


Figure 3.15 : blah blah...

3.9 Introduction à brms

Under the hood : Stan est un langage de programmation probabiliste écrit en C++, et qui implémente plusieurs algorithmes de MCMC : HMC, NUTS, L-BFGS...

```
data {  
  int<lower=0> J; // number of schools  
  real y[J]; // estimated treatment effects  
  real<lower=0> sigma[J]; // s.e. of effect estimates  
}
```

```

parameters {
  real mu;
  real<lower=0> tau;
  real eta[J];
}

transformed parameters {
  real theta[J];
  for (j in 1:J)
    theta[j] = mu + tau * eta[j];
}

model {
  target += normal_lpdf(eta | 0, 1);
  target += normal_lpdf(y | theta, sigma);
}

```

Le package `brms` (Bürkner, 2017) permet de fitter des modèles multi-niveaux (ou pas) linéaires (ou pas) bayésiens en Stan mais en utilisant la syntaxe de `lme4`.

Par exemple, le modèle suivant :

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \alpha_{\text{subject}[i]} + \alpha_{\text{item}[i]} + \beta x_i$$

se spécifie avec `brms` (comme avec `lme4`) de la manière suivante :

```
brm(y ~ x + (1 | subject) + (1 | item), data = d, family = gaussian() )
```

3.9.1 Rappels de syntaxe

Le package `brms` utilise la même syntaxe que les fonctions de base R (comme `lm`) ou que le package `lme4`.

```
Reaction ~ Days + (1 + Days | Subject)
```

La partie gauche représente notre variable dépendante (ou *outcome*, i.e., ce qu'on essaye de prédire). Le package `brms` permet également de fitter des modèles multivariés (plusieurs outcomes) en les combinant avec `mvbind()` :

```
mvbind(Reaction, Memory) ~ Days + (1 + Days | Subject)
```

La partie droite permet de définir les prédicteurs. L'intercept est généralement implicite, de sorte que les deux écritures ci-dessous sont équivalentes.

```
mvbind(Reaction, Memory) ~ Days + (1 + Days | Subject)  
mvbind(Reaction, Memory) ~ 1 + Days + (1 + Days | Subject)
```

Si l'on veut fitter un modèle sans intercept (why not), il faut le spécifier explicitement comme ci-dessous.

```
mvbind(Reaction, Memory) ~ 0 + Days + (1 + Days | Subject)
```

Par défaut `brms` postule une vraisemblance gaussienne. Ce postulat peut être changé facilement en spécifiant la vraisemblance souhaitée via l'argument `family`.

```
brm(Reaction ~ 1 + Days + (1 + Days | Subject), family = lognormal() )
```

Lisez la documentation (c'est très enthousiasmant à lire) accessible via `?brm`.

3.9.2 Quelques fonctions utiles

```
# Generate the Stan code:  
make_stancode(formula, ...)  
stancode(fit)  
  
# Generate the data passed to Stan:  
make_standata(formula, ...)  
standata(fit)  
  
# Handle priors:  
get_prior(formula, ...)  
set_prior(prior, ...)  
  
# Generate expected values and predictions:  
fitted(fit, ...)  
predict(fit, ...)  
marginal_effects(fit, ...)  
  
# Model comparison:  
loo(fit1, fit2, ...)  
bayes_factor(fit1, fit2, ...)  
model_weights(fit1, fit2, ...)
```

```
# Hypothesis testing:
hypothesis(fit, hypothesis, ...)
```

3.9.3 Un premier exemple

```
library(brms)
mod1 <- brm(height ~ 1, data = d2)

rbind(summary(mod1)$fixed, summary(mod1)$spec_pars )

##           Estimate Est.Error   1-95% CI   u-95% CI     Rhat Bulk_ESS Tail_ESS
## Intercept 154.599646 0.4264638 153.758096 155.447157 1.001954 3354.586 2603.114
## sigma      7.759921 0.2968472   7.221742   8.365626 1.000931 3661.881 2319.010
```

Ces données représentent les distributions marginales de chaque paramètre. En d'autres termes, la *probabilité* de chaque valeur de μ , après avoir *moyenné* sur toutes les valeurs possible de σ , est décrite par une distribution gaussienne avec une moyenne de 154.61 et un écart type de 0.41. L'intervalle de crédibilité (\neq intervalle de confiance) nous indique les 95% valeurs de μ ou σ les plus probables (sachant les données et les priors).

3.9.4 En utilisant notre prior

Par défaut `brms` utilise un prior très peu informatif centré sur la valeur moyenne de la variable mesurée. On peut donc affiner l'estimation réalisée par ce modèle en utilisant nos connaissances sur la distribution habituelle des tailles chez les humains.

La fonction `get_prior()` permet de visualiser une liste des priors par défaut ainsi que de tous les prios qu'on peut spécifier, sachant une certaine formule (i.e., une manière d'écrire notre modèle) et un jeu de données.

```
get_prior(height ~ 1, data = d2)

##           prior    class  coef group resp dpar nlpar lb ub  source
## student_t(3, 154.3, 8.5) Intercept                         default
## student_t(3, 0, 8.5)    sigma                           0  default
...
priors <- c(
  prior(normal(178, 20), class = Intercept),
  prior(exponential(0.01), class = sigma)
)
```

```
mod2 <- brm(  
  height ~ 1,  
  prior = priors,  
  family = gaussian(),  
  data = d2  
)
```

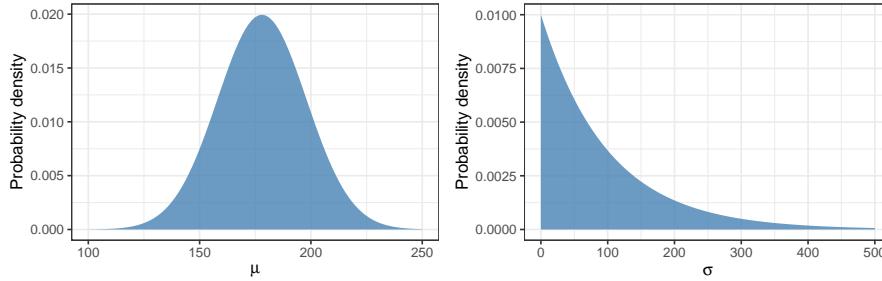


Figure 3.16 : blah blah...

...

```
summary(mod2)
```

```
## Family: gaussian  
## Links: mu = identity; sigma = identity  
## Formula: height ~ 1  
## Data: d2 (Number of observations: 352)  
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
##          total post-warmup draws = 4000  
##  
## Population-Level Effects:  
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
## Intercept    154.60      0.41   153.78   155.40 1.00     2902     2067  
##  
## Family Specific Parameters:  
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
## sigma       7.77      0.29     7.23     8.34 1.00     2996     2564  
##  
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS  
## and Tail_ESS are effective sample size measures, and Rhat is the potential  
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

3.9.5 En utilisant un prior plus informatif

```
priors <- c(
  prior(normal(178, 0.1), class = Intercept),
  prior(exponential(0.01), class = sigma)
)

mod3 <- brm(
  height ~ 1,
  prior = priors,
  family = gaussian(),
  data = d2
)
```

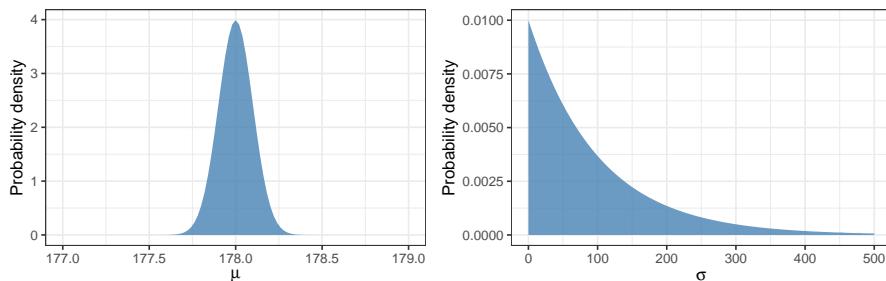


Figure 3.17 : blah blah...

...

```
summary(mod3)
```

```
## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: height ~ 1
## Data: d2 (Number of observations: 352)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##        total post-warmup draws = 4000
##
## Population-Level Effects:
##                               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept      177.86       0.10    177.67    178.06 1.00     3212     2603
## 
## Family Specific Parameters:
##                               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma      24.58       0.92    22.89    26.46 1.00     3687     2453
## 
```

```
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS  
## and Tail_ESS are effective sample size measures, and Rhat is the potential  
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

On remarque que la valeur estimée pour μ n'a presque pas "bougée" du prior...mais on remarque également que la valeur estimée pour σ a largement augmentée. Nous avons dit au modèle que nous étions assez certain de notre valeur de μ , le modèle s'est ensuite "adapté", ce qui explique la valeur de σ ...

3.9.6 Précision du prior (heuristique)

Le prior peut généralement être considéré comme un posterior obtenu sur des données antérieures.

On sait que le σ d'un posterior gaussien nous est donné par la formule :

$$\sigma_{post} = 1/\sqrt{n}$$

Qui implique une *quantité de données* $n = 1/\sigma_{post}^2$. Notre prior avait un $\sigma = 0.1$, ce qui donne $n = 1/0.1^2 = 100$.

Donc, on peut considérer que le prior $\mu \sim \text{Normal}(178, 0.1)$ est équivalent au cas dans lequel nous aurions observé 100 tailles de moyenne 178.

3.9.7 Récupérer et visualiser les échantillons de la distribution postérieure

```
post <- posterior_samples(mod2) %>%  
  mutate(density = get_density(b_Intercept, sigma, n = 1e2) )  
  
ggplot(post, aes(x = b_Intercept, y = sigma, color = density) ) +  
  geom_point(size = 2, alpha = 0.5, show.legend = FALSE) +  
  theme_bw(base_size = 20) +  
  labs(x = expression(mu), y = expression(sigma) ) +  
  viridis::scale_color_viridis()
```

3.9.8 Récupérer les échantillons de la distribution postérieure

```
# gets the first 6 samples  
head(post)
```

```
##   b_Intercept     sigma    lprior     lp__   density  
## 1    154.4015 8.200152 -9.297956 -1227.873 0.3142422  
## 2    155.1757 7.532505 -9.246354 -1227.914 0.4316129
```

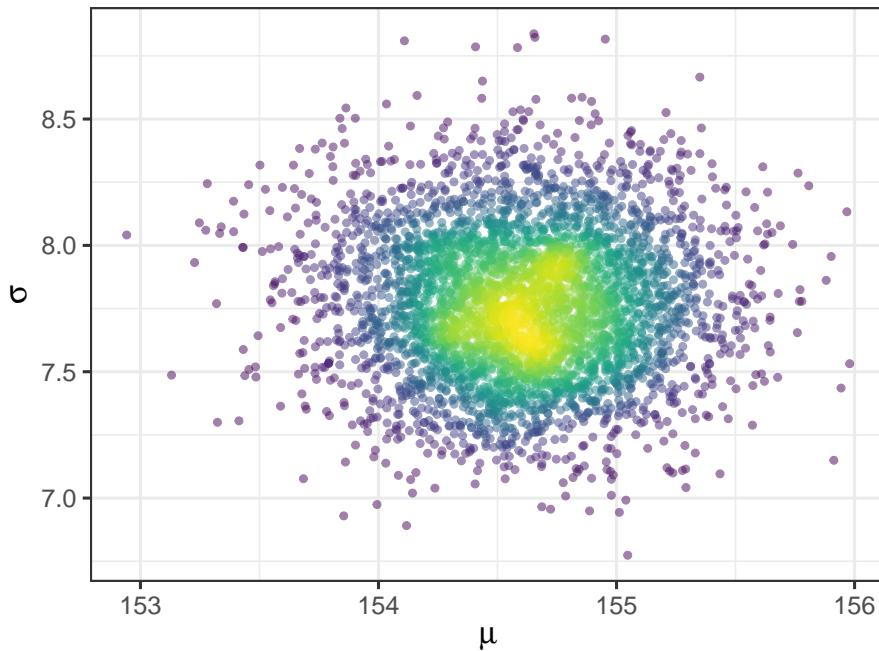


Figure 3.18 : blah blah...

```
## 3    154.9014 7.510597 -9.261880 -1227.241 0.7801206
## 4    154.4747 8.189997 -9.293538 -1227.761 0.3675453
## 5    154.7345 7.629289 -9.272738 -1226.766 1.1603879
## 6    155.0476 7.905011 -9.257408 -1227.338 0.6852718
```

```
# gets the median and the 95% credible interval
t(sapply(post[, 1:2], quantile, probs = c(0.025, 0.5, 0.975) ) )
```

```
##                      2.5%          50%         97.5%
## b_Intercept 153.782212 154.605778 155.40092
## sigma        7.231156   7.759307   8.33988
```

3.9.9 Visualiser la distribution postérieure

```
H.scv <- Hscv(post[, 1:2])
fhat_post <- kde(x = post[, 1:2], H = H.scv, compute.cont = TRUE)

plot(fhat_post, display = "persp", col = "purple", border = NA,
      xlab = "\nmu", ylab = "\nsigma", zlab = "\nnp(mu, sigma)",
      shade = 0.8, phi = 30, ticktype = "detailed",
      cex.lab = 1.2, family = "Helvetica")
```

```
knitr::include_graphics("figures/posterior.png")
```

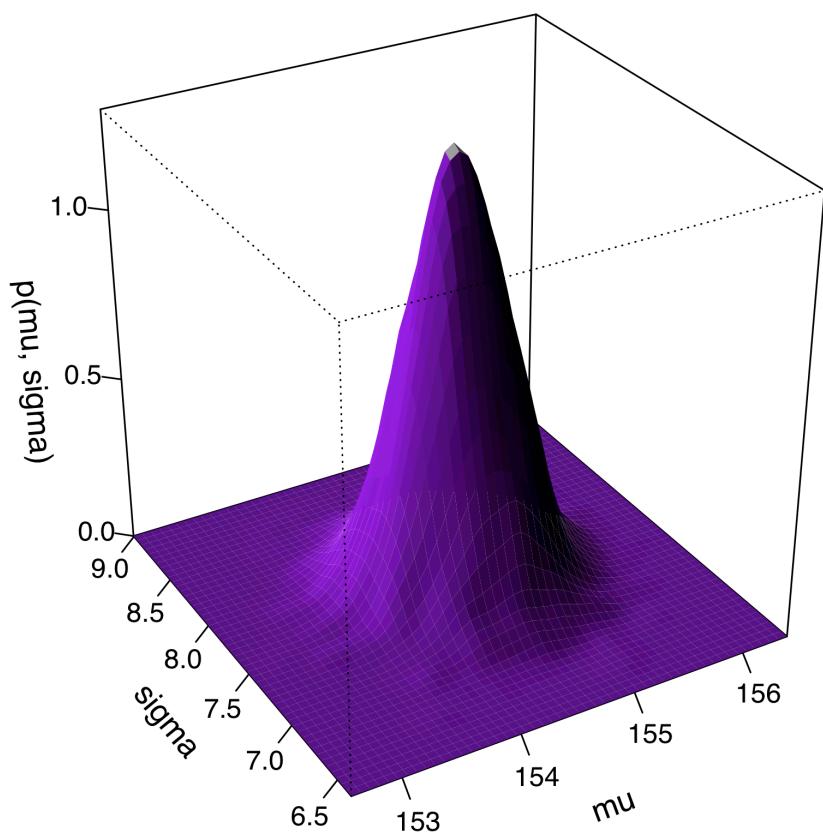


Figure 3.19 : blah blah...

3.9.10 Visualiser la distribution postérieure

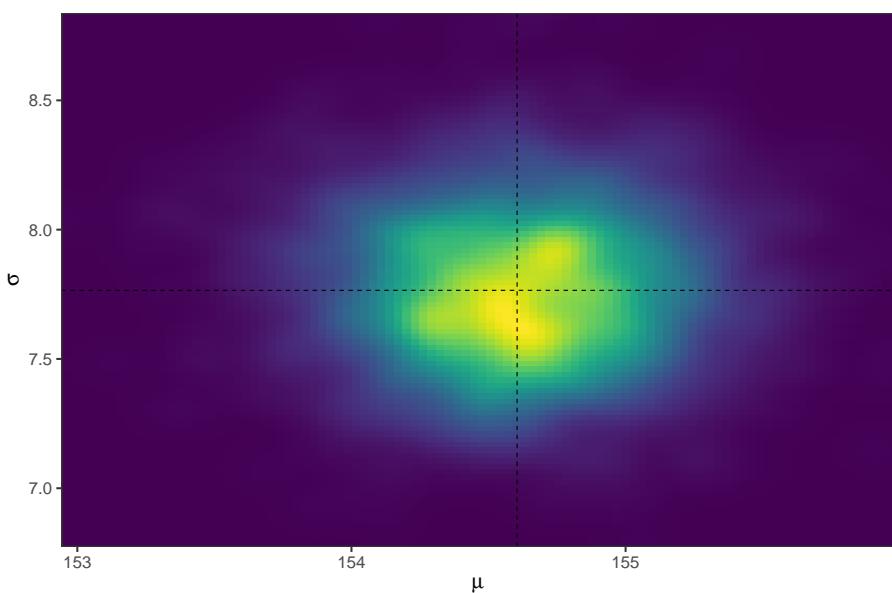


Figure 3.20 : blah blah...

3.9.11 Ajouter un prédicteur

Comment est-ce que la taille co-varie avec le poids ?

```
d2 %>%
  ggplot(aes(x = weight, y = height) ) +
  geom_point(colour = "white", fill = "black", pch = 21, size = 3, alpha = 0.8) +
  theme_bw(base_size = 20)
```

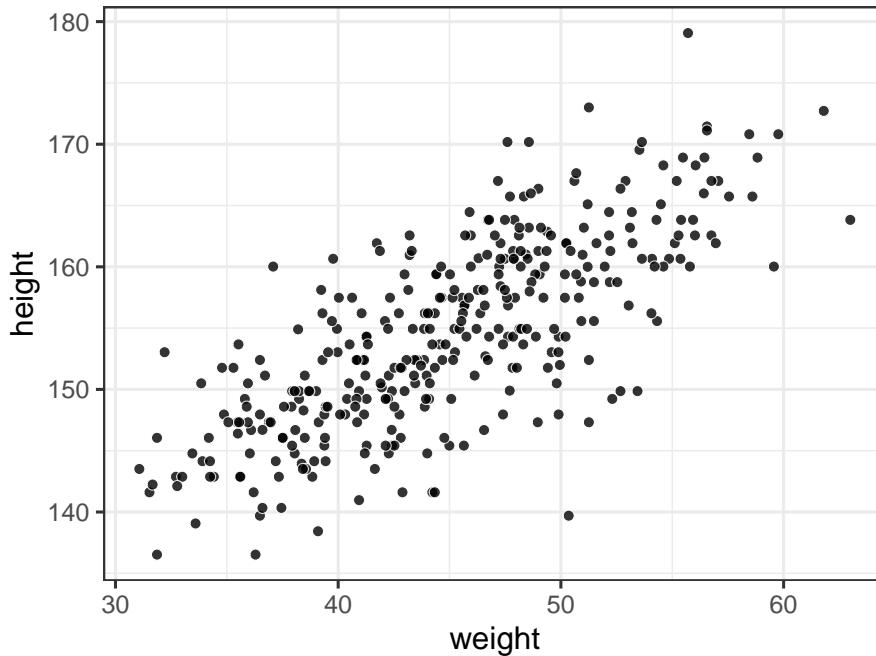


Figure 3.21 : blah blah...

3.10 Régression linéaire à un prédicteur continu

$$h_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta x_i$$

```
linear_model <- lm(height ~ weight, data = d2)
precis(linear_model, prob = 0.95)
```

	mean	sd	2.5%	97.5%
## (Intercept)	113.8793936	1.91106523	110.1337746	117.6250126
## weight	0.9050291	0.04204752	0.8226175	0.9874407

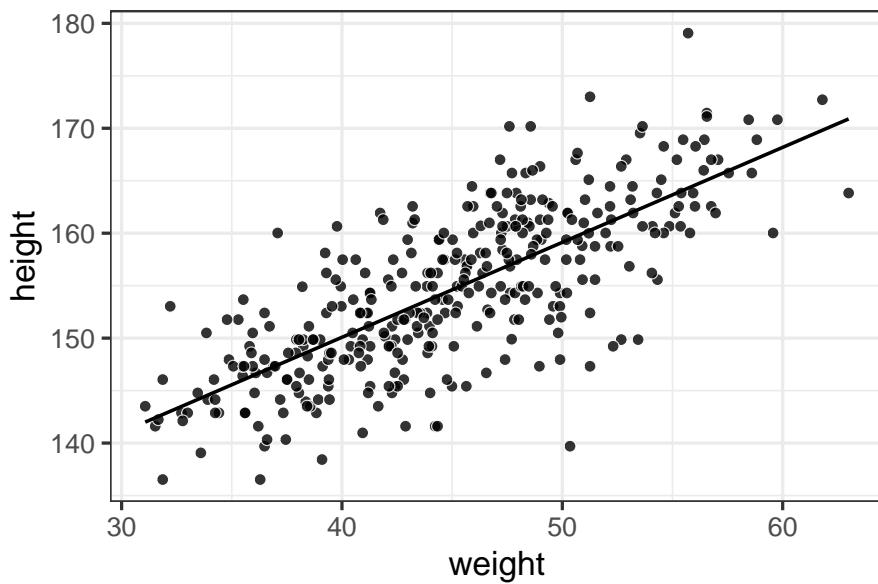


Figure 3.22 : blah blah...

3.10.1 Différentes notations équivalentes

On considère un modèle de régression linéaire avec un seul prédicteur, une pente, un intercept, et des résidus distribués selon une loi normale. La notation :

$$h_i = \alpha + \beta x_i + \epsilon_i \quad \text{avec} \quad \epsilon_i \sim \text{Normal}(0, \sigma)$$

est équivalente à :

$$h_i - (\alpha + \beta x_i) \sim \text{Normal}(0, \sigma)$$

et si on réduit encore un peu :

$$h_i \sim \text{Normal}(\alpha + \beta x_i, \sigma).$$

Les notations ci-dessus sont équivalentes, mais la dernière est plus flexible, et nous permettra par la suite de l'étendre plus simplement aux modèles multi-niveaux.

$$\begin{aligned} h_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta x_i \\ \alpha &\sim \text{Normal}(178, 20) \\ \beta &\sim \text{Normal}(0, 10) \\ \sigma &\sim \text{Exponential}(0.01) \end{aligned}$$

Dans ce modèle μ n'est plus un paramètre à estimer (car μ est déterminé par α et β). À la place, nous allons estimer α et β .

Rappels : α est l'*intercept*, c'est à dire la taille attendue, lorsque le poids est égal à 0. β est la pente, c'est à dire le changement de taille attendu quand le poids augmente d'une unité.

```
priors <- c(
  prior(normal(178, 20), class = Intercept),
  prior(normal(0, 10), class = b),
  prior(exponential(0.01), class = sigma)
)

mod4 <- brm(
  height ~ 1 + weight,
  prior = priors,
  family = gaussian(),
  data = d2
)

summary(mod4)

## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: height ~ 1 + weight
## Data: d2 (Number of observations: 352)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##         total post-warmup draws = 4000
##
## Population-Level Effects:
##             Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept    113.95     1.90   110.23   117.70 1.00    3941    2787
## weight        0.90      0.04     0.82     0.99 1.00    3866    2713
##
## Family Specific Parameters:
##             Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma       5.11      0.20     4.74     5.51 1.00    3946    2967
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

- $\beta = 0.90, 95\% \text{ CrI } [0.82, 0.99]$ nous indique qu'une augmentation de 1kg entraîne une augmentation de 0.90cm.
- $\alpha = 113.91, 95\% \text{ CrI } [110.12, 117.59]$ représente la taille moyenne quand le poids est égal à 0kg...

...

```
d2$weight.c <- d2$weight - mean(d2$weight)

mod5 <- brm(
  height ~ 1 + weight.c,
  prior = priors,
  family = gaussian(),
  data = d2
)

fixef(mod5) # retrieves the fixed effects estimates

##           Estimate   Est.Error      Q2.5      Q97.5
## Intercept 154.6047599 0.27418324 154.071037 155.1471805
## weight.c    0.9040667 0.04174453  0.822152  0.9844688
```

- Après avoir centré la réponse, l'intercept représente la valeur attendue de *taille* lorsque le poids est à sa valeur moyenne.

3.10.2 Représenter les prédictions du modèle

```
d2 %>%
  ggplot(aes(x = weight, y = height) ) +
  geom_point(colour = "white", fill = "black", pch = 21, size = 3, alpha = 0.8) +
  geom_abline(intercept = fixef(mod4)[1], slope = fixef(mod4)[2], lwd = 1) +
  theme_bw(base_size = 20)
```

3.10.3 Représenter l'incertitude sur μ via fitted()

```
# on crée un vecteur de valeurs possibles pour "weight"
weight.seq <- data.frame(weight = seq(from = 25, to = 70, by = 1) )

# on récupère les prédictions du modèle pour ces valeurs de poids
mu <- data.frame(fitted(mod4, newdata = weight.seq) ) %>% bind_cols(weight.seq)

# on affiche les 10 premières lignes de mu
head(mu, 10)
```

```
##           Estimate   Est.Error      Q2.5      Q97.5 weight
## 1     136.5378 0.8810557 134.8053 138.3012     25
```

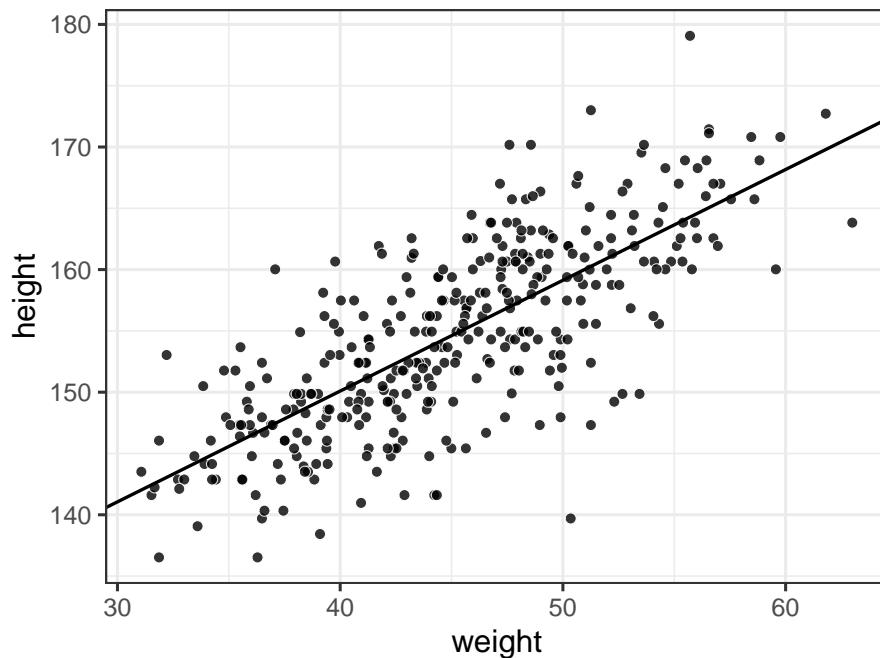


Figure 3.23 : blah blah...

```
## 2 137.4414 0.8415178 135.7820 139.1252      26
## 3 138.3450 0.8022026 136.7638 139.9587      27
## 4 139.2486 0.7631446 137.7317 140.7903      28
## 5 140.1523 0.7243854 138.7222 141.6077      29
## 6 141.0559 0.6859756 139.6953 142.4340      30
## 7 141.9595 0.6479773 140.6758 143.2636      31
## 8 142.8631 0.6104675 141.6608 144.0954      32
## 9 143.7667 0.5735418 142.6334 144.9188      33
## 10 144.6704 0.5373209 143.6020 145.7463      34
```

...

```
d2 %>%
  ggplot(aes(x = weight, y = height) ) +
  geom_point(colour = "white", fill = "black", pch = 21, size = 3, alpha = 0.8) +
  geom_smooth(
    data = mu, aes(y = Estimate, ymin = Q2.5, ymax = Q97.5),
    stat = "identity",
    color = "black", alpha = 0.8, size = 1
  ) +
  theme_bw(base_size = 20)
```

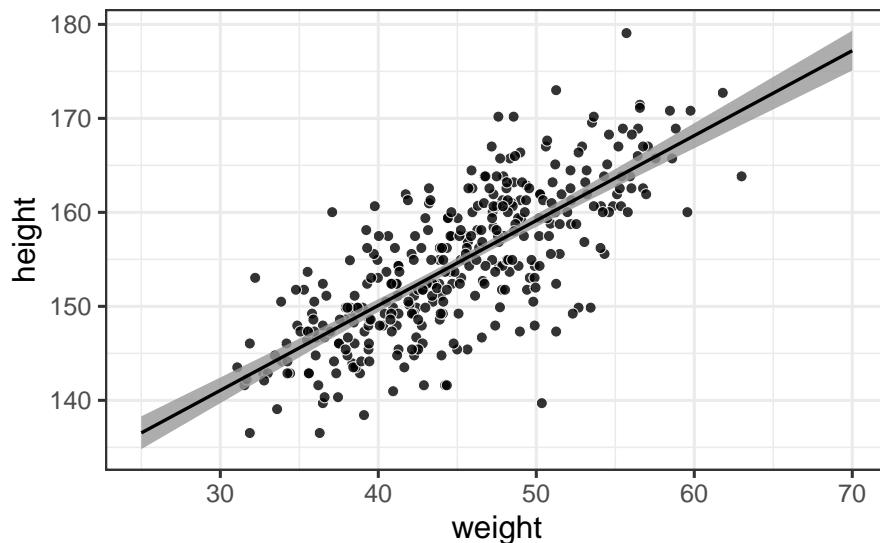


Figure 3.24 : blah blah...

3.10.4 Intervalles de prédition (incorporer σ)

Pour rappel, voici notre modèle : $h_i \sim \text{Normal}(\alpha + \beta x_i, \sigma)$. Pour l'instant, on a seulement représenté les prédictions pour μ . Comment incorporer σ dans nos prédictions ?

```
# on crée un vecteur de valeurs possibles pour "weight"
weight.seq <- data.frame(weight = seq(from = 25, to = 70, by = 1) )

# on récupère les prédictions du modèle pour ces valeurs de poids
pred_height <- data.frame(predict(mod4, newdata = weight.seq) ) %>% bind_cols(weight.seq)

# on affiche les 10 premières lignes de pred_height
head(pred_height, 10)
```

```
##      Estimate Est.Error    Q2.5    Q97.5 weight
## 1 136.5810  5.132924 126.3841 146.7682    25
## 2 137.5784  5.193987 127.7218 147.6821    26
## 3 138.3971  5.233652 128.2104 148.5753    27
## 4 139.4044  5.212859 129.5219 150.0878    28
## 5 140.2113  5.082795 130.4095 150.2197    29
## 6 141.0264  5.146524 130.9380 151.0698    30
## 7 142.0185  5.152701 131.8745 152.2535    31
## 8 142.8792  5.165274 132.8146 153.1452    32
## 9 143.7302  5.157094 133.3480 153.8876    33
## 10 144.7373  5.130561 134.4967 154.7412    34
```

```
d2 %>%
  ggplot(aes(x = weight, y = height) ) +
  geom_point(colour = "white", fill = "black", pch = 21, size = 3, alpha = 0.8) +
  geom_ribbon(
    data = pred_height, aes(x = weight, ymin = Q2.5, ymax = Q97.5),
    alpha = 0.2, inherit.aes = FALSE
  ) +
  geom_smooth(
    data = mu, aes(y = Estimate, ymin = Q2.5, ymax = Q97.5),
    stat = "identity", color = "black", alpha = 0.8, size = 1
  ) +
  theme_bw(base_size = 20)
```

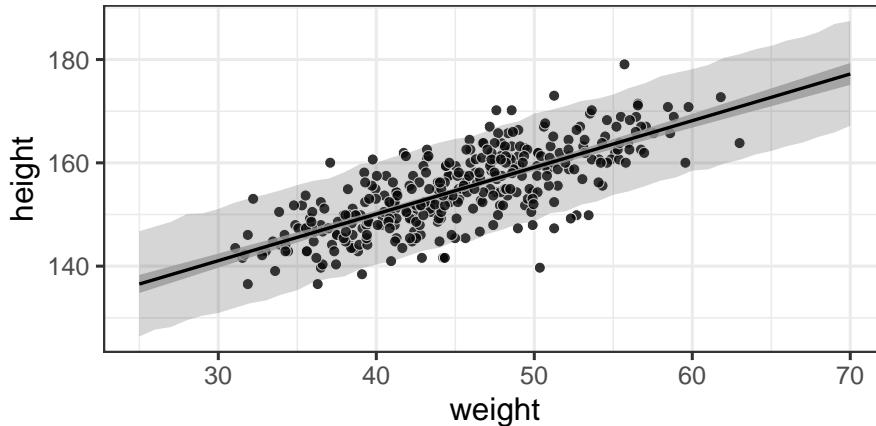


Figure 3.25 : blah blah...

3.10.5 Deux types d'incertitude

Deux sources d'incertitude dans le modèle : incertitude concernant l'estimation de la valeur des paramètres mais également concernant le processus d'échantillonnage.

Incertitude épistémique : La distribution a posteriori ordonne toutes les combinaisons possibles des valeurs des paramètres selon leurs plausibilités relatives.

Incertitude aléatoire : La distribution des données simulées est elle, une distribution qui contient de l'incertitude liée à un processus d'échantillonnage (i.e., générer des données à partir d'une gaussienne).

Voir aussi ce [court article](#) par O'Hagan (2012).

3.11 Régression polynomiale

```
d %>% # on utilise d au lieu de d2
  ggplot(aes(x = weight, y = height) ) +
  geom_point(colour = "white", fill = "black", pch = 21, size = 3, alpha = 0.8) +
  theme_bw(base_size = 20)
```

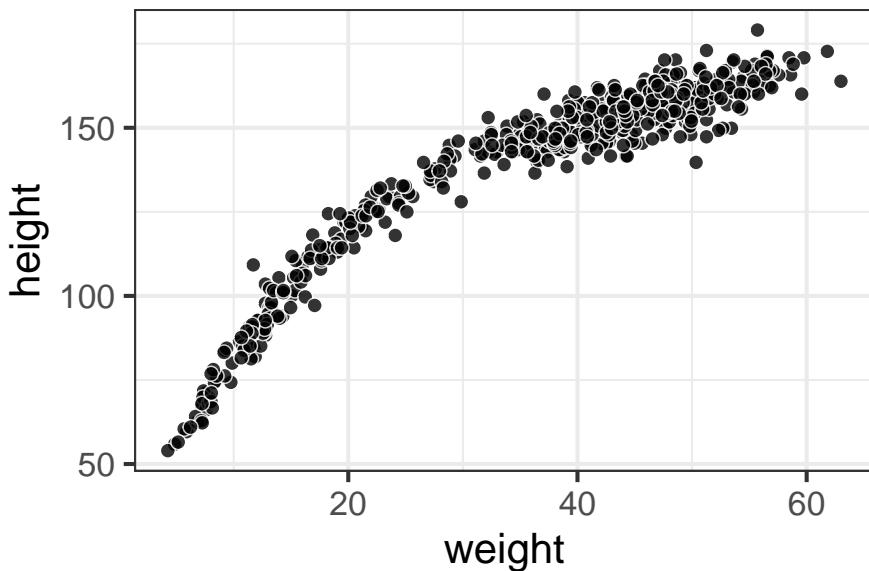


Figure 3.26 : blah blah...

Si on considère tout l'échantillon (pas seulement les adultes), la relation entre taille et poids semble incurvée...

```
d <- d %>% mutate(weight.s = (weight - mean(weight) ) / sd(weight) )
d %>%
  ggplot(aes(x = weight.s, y = height) ) +
  geom_point(colour = "white", fill = "black", pch = 21, size = 3, alpha = 0.8) +
  theme_bw(base_size = 20)

c(mean(d$weight.s), sd(d$weight.s) )

## [1] -2.712698e-18  1.000000e+00
```

Pourquoi standardiser les prédicteurs?

- **Interprétation.** Un changement d'une unité du prédicteur correspond à un changement d'un écart-type sur la réponse. Permet de comparer les coefficients de plusieurs prédicteurs.
- **Fitting.** Quand les prédicteurs contiennent de grandes valeurs, cela peut poser des problèmes...

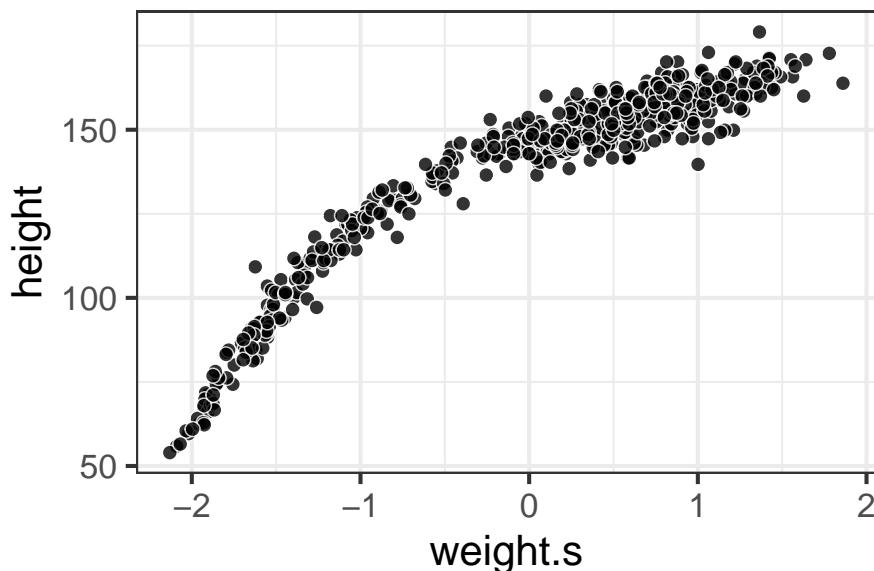


Figure 3.27 : blah blah...

3.11.1 Modèle de régression polynomiale

$$\begin{aligned}
 h_i &\sim \text{Normal}(\mu_i, \sigma) \\
 \mu_i &= \alpha + \beta_1 x_i + \beta_2 x_i^2 \\
 \alpha &\sim \text{Normal}(156, 100) \\
 \beta_1 &\sim \text{Normal}(0, 10) \\
 \beta_2 &\sim \text{Normal}(0, 10) \\
 \sigma &\sim \text{Exponential}(0.01)
 \end{aligned}$$

À vous de construire ce modèle en utilisant `brms::brm()` ...

```

priors <- c(
  prior(normal(156, 100), class = Intercept),
  prior(normal(0, 10), class = b),
  prior(exponential(0.01), class = sigma)
)

mod6 <- brm(
  # NB: polynomials should be written with the I() function...
  height ~ 1 + weight.s + I(weight.s^2),
  prior = priors,
  family = gaussian(),
  data = d
)

```

```
summary(mod6)

## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: height ~ 1 + weight.s + I(weight.s^2)
## Data: d (Number of observations: 544)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##         total post-warmup draws = 4000
##
## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     146.67      0.37   145.95   147.40 1.00     3200    2938
## weight.s       21.40      0.29   20.82    21.95 1.00     3669    3307
## Iweight.sE2   -8.42      0.28   -8.97    -7.88 1.00     3445    3086
##
## Family Specific Parameters:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma       5.78      0.17     5.46     6.12 1.00     3864    2511
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

...

3.11.2 Représenter les prédictions du modèle

```
# on crée un vecteur de valeurs possibles pour "weight"
weight.seq <- data.frame(weight.s = seq(from = -2.5, to = 2.5, length.out = 50) )

# on récupère les prédictions du modèle pour ces valeurs de poids
mu <- data.frame(fitted(mod6, newdata = weight.seq) ) %>% bind_cols(weight.seq)
pred_height <- data.frame(predict(mod6, newdata = weight.seq) ) %>% bind_cols(weight.seq)

# on affiche les 10 premières lignes de pred_height
head(pred_height, 10)

##   Estimate Est.Error   Q2.5   Q97.5 weight.s
## 1  40.40390  5.986537 28.31332  52.08907 -2.500000
## 2  46.83829  5.968128 35.05460  58.59766 -2.397959
## 3  53.02497  5.860617 41.64739  64.10652 -2.295918
```

```
## 4 59.15998 5.869889 47.32456 70.24130 -2.193878
## 5 65.14660 5.766219 53.96058 76.36846 -2.091837
## 6 70.61914 5.755012 59.06915 82.05318 -1.989796
## 7 76.27009 5.845372 64.95498 87.65043 -1.887755
## 8 81.59569 5.704064 70.59550 92.86416 -1.785714
## 9 86.67676 5.821030 75.53530 97.99047 -1.683673
## 10 91.74770 5.875861 80.04455 102.84168 -1.581633
```

...

```
d %>%
  ggplot(aes(x = weight.s, y = height) ) +
  geom_point(colour = "white", fill = "black", pch = 21, size = 3, alpha = 0.8) +
  geom_ribbon(
    data = pred_height, aes(x = weight.s, ymin = Q2.5, ymax = Q97.5),
    alpha = 0.2, inherit.aes = FALSE
  ) +
  geom_smooth(
    data = mu, aes(y = Estimate, ymin = Q2.5, ymax = Q97.5),
    stat = "identity", color = "black", alpha = 0.8, size = 1
  ) +
  theme_bw(base_size = 20)
```

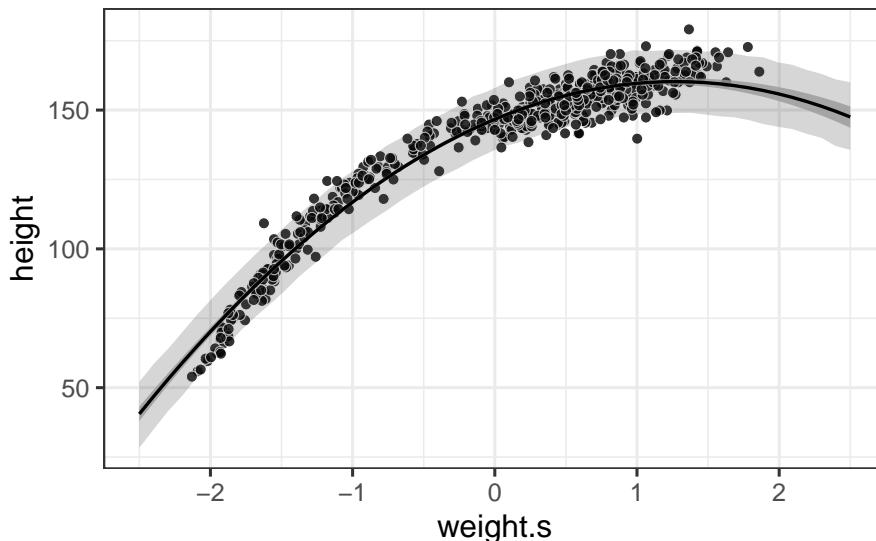


Figure 3.28 : blah blah...

3.12 Modèle de régression, taille d'effet

Il existe plusieurs méthodes pour calculer les tailles d'effet dans les modèles bayésiens. Gelman & Pardoe (2006) proposent une méthode pour calculer un R^2 basé sur l'échantillon.

Marsman et al. (2017), Marsman et al. (2019) généralisent des méthodes existantes pour calculer un ρ^2 pour les designs de type ANOVA (i.e., avec prédicteurs catégoriels), qui représente une estimation de la taille d'effet *dans la population*, et non basé sur l'échantillon.

“Similar to most of the ES measures that have been proposed for the ANOVA model, the squared multiple correlation coefficient ρ^2 [...] is a so-called proportional reduction in error measure (PRE; Reynolds, 1977). In general, a PRE measure expresses the proportion of the variance in an outcome y that is attributed to the independent variables x ” (Marsman et al., 2019).

$$\begin{aligned}\rho^2 &= \frac{\sum_{i=1}^n \pi_i (\beta_i - \beta)^2}{\sigma^2 + \sum_{i=1}^n \pi_i (\beta_i - \beta)^2} \\ \rho^2 &= \frac{\frac{1}{n} \sum_{i=1}^n \beta_i^2}{\sigma^2 + \frac{1}{n} \sum_{i=1}^n \beta_i^2} \\ \rho^2 &= \frac{\beta^2 \tau^2}{\sigma^2 + \beta^2 \tau^2}\end{aligned}$$

```
post <- posterior_samples(mod4)
beta <- post$b_weight
sigma <- post$sigma

f1 <- beta^2 * var(d2$weight)
rho <- f1 / (f1 + sigma^2)
```

Attention, si plusieurs prédicteurs, dépend de la structure de covariance...

```
BEST::plotPost(rho, showMode = TRUE, xlab = expression(rho) )
```

```
summary(lm(height ~ weight, data = d2) )$r.squared
## [1] 0.5696444
```

3.13 Conclusions

On a présenté un nouveau modèle à deux puis trois paramètres : le modèle gaussien, puis la régression linéaire gaussienne, permettant de mettre en relation deux variables continues.

Comme précédemment, le théorème de Bayes est utilisé pour mettre à jour nos connaissances a priori quant à la valeur des paramètres en une connaissance a posteriori, synthèse entre nos priors et l'information contenue dans les données.

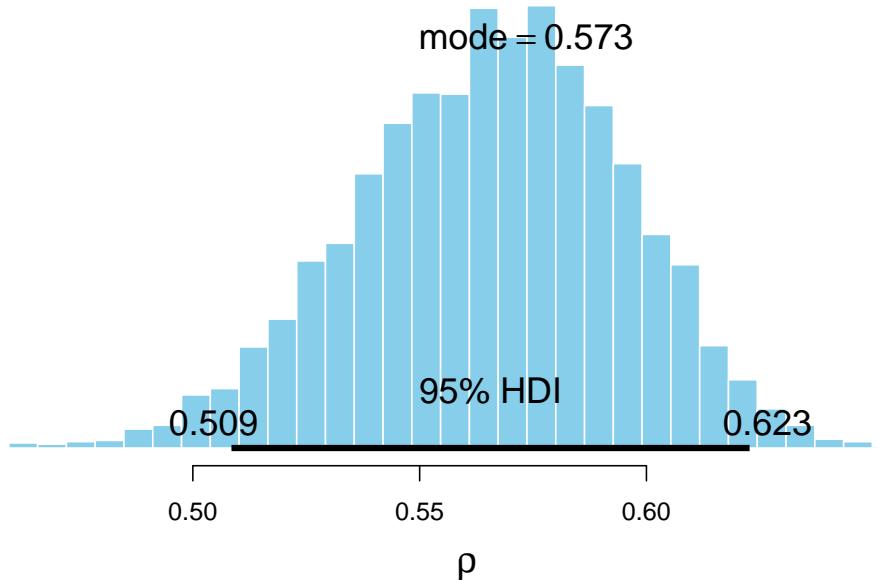


Figure 3.29 : blah blah...

Le package `brms` permet de fitter toutes sortes de modèles avec une syntaxe similaire à celle utilisée par `lm()`.

La fonction `fitted()` permet de récupérer les prédictions d'un modèle fitté avec `brms` (i.e., un modèle de classe `brmsfit`).

La fonction `predict()` permet de simuler des données à partir d'un modèle fitté avec `brms`.

Modèle de régression linéaire, suite

Introduction au chapitre blah blah...

4.1 Régression multiple

On va étendre le modèle précédent en ajoutant plusieurs prédicteurs, continus et/ou catégoriels. Pourquoi faire?

- *Contrôle* des facteurs de confusion (e.g., [spurious correlations](#), [simpson's paradox](#)). Un facteur de confusion est une variable aléatoire qui influence à la fois la variable dépendante et les variables explicatives. Une approche multivariée peut nous aider à démêler les influences causales de différents prédicteurs.
- Multiples causes : un phénomène peut émerger sous l'influence de multiples causes.
- Interactions : l'influence d'un prédicteur sur la variable observée peut dépendre de la valeur d'un autre prédicteur.

4.1.1 Associations fortuites

```
library(rethinking)
library(tidyverse)

data(WaffleDivorce) # import des données
df1 <- WaffleDivorce # import dans une dataframe nommée df1

str(df1) # structure des données

## 'data.frame':    50 obs. of  13 variables:
```

```

## $ Location      : Factor w/ 50 levels "Alabama","Alaska",...: 1 2 3 4 5 6 7 8 9 10 ...
## $ Loc           : Factor w/ 50 levels "AK","AL","AR",...: 2 1 4 3 5 6 7 9 8 10 ...
## $ Population    : num  4.78 0.71 6.33 2.92 37.25 ...
## $ MedianAgeMarriage: num  25.3 25.2 25.8 24.3 26.8 25.7 27.6 26.6 29.7 26.4 ...
## $ Marriage       : num  20.2 26 20.3 26.4 19.1 23.5 17.1 23.1 17.7 17 ...
## $ Marriage.SE    : num  1.27 2.93 0.98 1.7 0.39 1.24 1.06 2.89 2.53 0.58 ...
## $ Divorce        : num  12.7 12.5 10.8 13.5 8 11.6 6.7 8.9 6.3 8.5 ...
## $ Divorce.SE     : num  0.79 2.05 0.74 1.22 0.24 0.94 0.77 1.39 1.89 0.32 ...
## $ WaffleHouses   : int  128 0 18 41 0 11 0 3 0 133 ...
## $ South          : int  1 0 0 1 0 0 0 0 0 1 ...
## $ Slaves1860     : int  435080 0 0 111115 0 0 0 1798 0 61745 ...
## $ Population1860 : int  964201 0 0 435450 379994 34277 460147 112216 75080 140424 ...
## $ PropSlaves1860 : num  0.45 0 0 0.26 0 0 0 0.016 0 0.44 ...

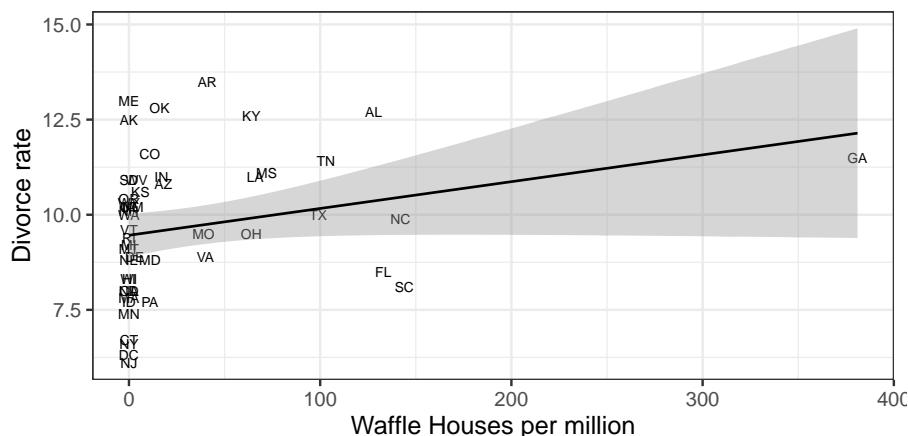
```

On observe un lien positif entre le nombre de “waffle houses” et le taux de divorce...

```

df1 %>%
  ggplot(aes(x = WaffleHouses, y = Divorce) ) +
  geom_text(aes(label = Loc) ) +
  geom_smooth(method = "lm", color = "black", se = TRUE) +
  theme_bw(base_size = 20) +
  labs(x = "Waffle Houses per million", y = "Divorce rate")

```



On observe un lien positif entre le taux de mariage et le taux de divorce, mais est-ce qu'on peut vraiment dire que le mariage “cause” le divorce ?

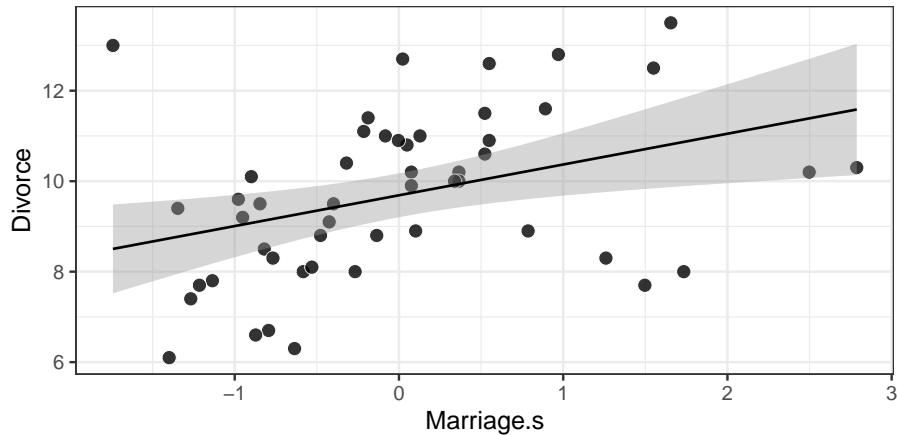
```
df1$Marriage.s <- (df1$Marriage - mean(df1$Marriage) ) / sd(df1$Marriage)
```

```

df1 %>%
  ggplot(aes(x = Marriage.s, y = Divorce) ) +
  geom_point(pch = 21, color = "white", fill = "black", size = 5, alpha = 0.8) +

```

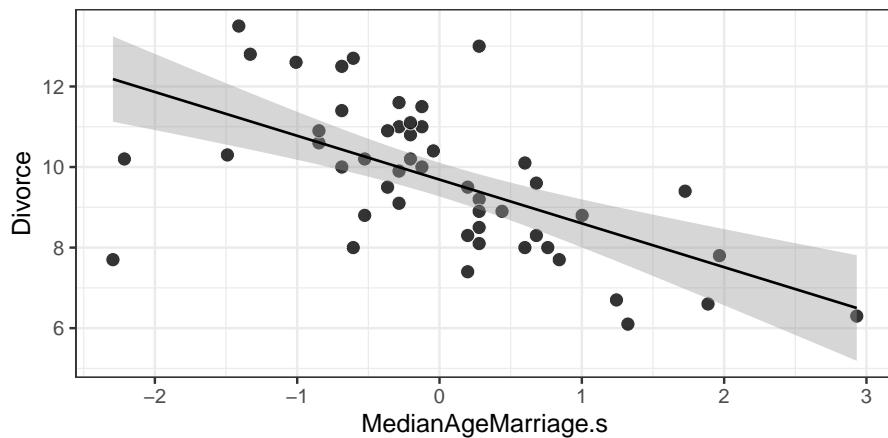
```
geom_smooth(method = "lm", color = "black", se = TRUE) +
theme_bw(base_size = 20)
```



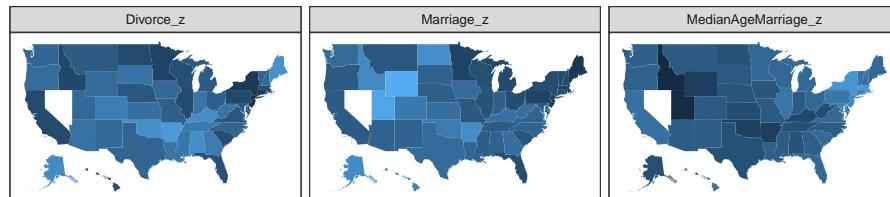
On observe l'association inverse entre le taux de divorce et l'âge médian de mariage.

```
df1$MedianAgeMarriage.s <- (df1$MedianAgeMarriage - mean(df1$MedianAgeMarriage) ) /
sd(df1$MedianAgeMarriage)

df1 %>%
ggplot(aes(x = MedianAgeMarriage.s, y = Divorce) ) +
geom_point(pch = 21, color = "white", fill = "black", size = 5, alpha = 0.8) +
geom_smooth(method = "lm", color = "black", se = TRUE) +
theme_bw(base_size = 20)
```



On peut représenter nos trois variables principales sur une carte des 50 états.



4.1.1.1 Influence du taux de mariage

$$\begin{aligned} D_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta_R R_i \\ \alpha &\sim \text{Normal}(10, 10) \\ \beta_R &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{Exponential}(0.01) \end{aligned}$$

```
priors <- c(
  prior(normal(10, 10), class = Intercept),
  prior(normal(0, 1), class = b),
  prior(exponential(0.01), class = sigma)
)
```

```
mod1 <- brm(
  Divorce ~ 1 + Marriage.s,
  family = gaussian(),
  prior = priors,
  data = df1
)
```

...

```
summary(mod1)
```

```
## Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Divorce ~ 1 + Marriage.s
##   Data: df1 (Number of observations: 50)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##                               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
```

```

## Intercept      9.69      0.25     9.19    10.19 1.00      4473     3023
## Marriage.s    0.64      0.24     0.15     1.11 1.00      3697     2407
##
## Family Specific Parameters:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma      1.76      0.18     1.44     2.15 1.00      4016     3017
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).

```

4.1.1.2 Influence de l'âge médian de mariage

$$\begin{aligned}
D_i &\sim \text{Normal}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_A A_i \\
\alpha &\sim \text{Normal}(10, 10) \\
\beta_A &\sim \text{Normal}(0, 1) \\
\sigma &\sim \text{Exponential}(0.01)
\end{aligned}$$

```

priors <- c(
  prior(normal(10, 10), class = Intercept),
  prior(normal(0, 1), class = b),
  prior(exponential(0.01), class = sigma)
)

```

```

mod2 <- brm(
  Divorce ~ 1 + MedianAgeMarriage.s,
  family = gaussian(),
  prior = priors,
  data = df1
)

```

...

```
summary(mod2)
```

```

## Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: Divorce ~ 1 + MedianAgeMarriage.s
##   Data: df1 (Number of observations: 50)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000

```

```
##  
## Population-Level Effects:  
##           Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
## Intercept      9.69     0.22    9.25   10.11 1.00     4251    2843  
## MedianAgeMarriage.s -1.04     0.22   -1.47   -0.61 1.00     3832    2722  
##  
## Family Specific Parameters:  
##           Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
## sigma      1.52     0.16    1.24    1.88 1.00     3845    2725  
##  
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS  
## and Tail_ESS are effective sample size measures, and Rhat is the potential  
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

4.1.2 Régression multiple

Quelle est la valeur prédictive d'une variable, une fois que je connais tous les autres prédicteurs?

$$\begin{aligned} D_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta_R R_i + \beta_A A_i \\ \alpha &\sim \text{Normal}(10, 10) \\ \beta_R, \beta_A &\sim \text{Normal}(0, 1) \\ \sigma &\sim \text{Exponential}(0.01) \end{aligned}$$

Ce modèle répond à deux questions :

- Une fois connue le taux de mariage, quelle valeur ajoutée apporte la connaissance de l'âge médian du mariage?
- Une fois connu l'âge médian du mariage, quelle valeur ajoutée apporte la connaissance de le taux de mariage?

```
priors <- c(  
  prior(normal(10, 10), class = Intercept),  
  prior(normal(0, 1), class = b),  
  prior(exponential(0.01), class = sigma)  
)  
  
mod3 <- brm(  
  Divorce ~ 1 + Marriage.s + MedianAgeMarriage.s,  
  family = gaussian(),  
  prior = priors,
```

```
data = df1
)
```

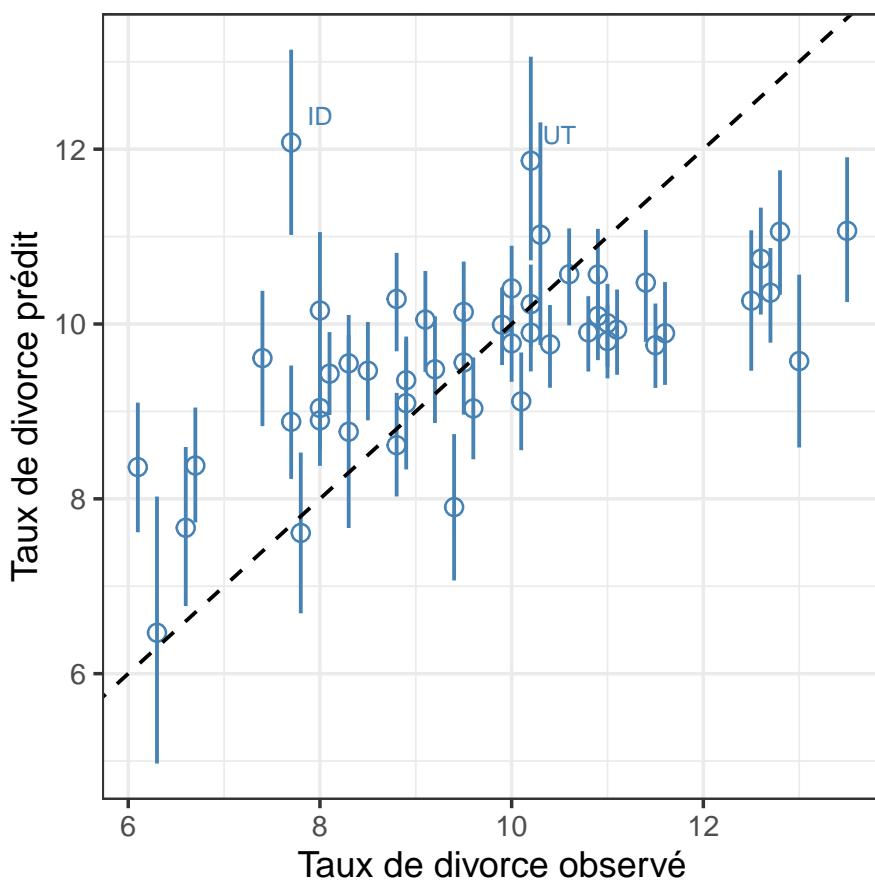
Interprétation : Une fois qu'on connaît l'âge median de mariage dans un état, connaître le taux de mariage de cet état n'apporte pas vraiment d'information supplémentaire...

```
summary(mod3)
```

```
## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: Divorce ~ 1 + Marriage.s + MedianAgeMarriage.s
## Data: df1 (Number of observations: 50)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##         total post-warmup draws = 4000
##
## Population-Level Effects:
##                               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept                  9.68     0.21    9.26   10.09 1.00    3067    2325
## Marriage.s                 -0.12     0.30   -0.68    0.45 1.00    2504    2783
## MedianAgeMarriage.s       -1.12     0.30   -1.70   -0.52 1.00    2490    2915
##
## Family Specific Parameters:
##                               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma          1.53      0.16    1.24     1.89 1.00    3085    2387
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

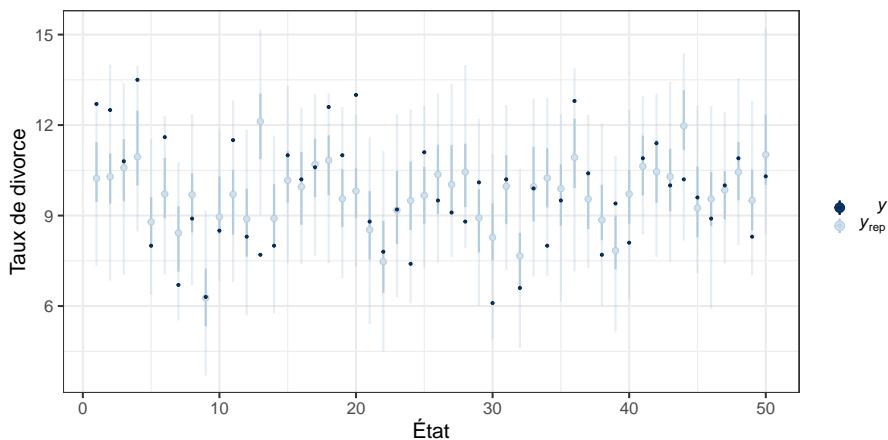
4.1.2.1 Visualiser les prédictions du modèle

On peut comparer le taux de divorce observé dans chaque état au taux de divorce prédit par notre modèle (la ligne diagonale représente une prédition parfaite).

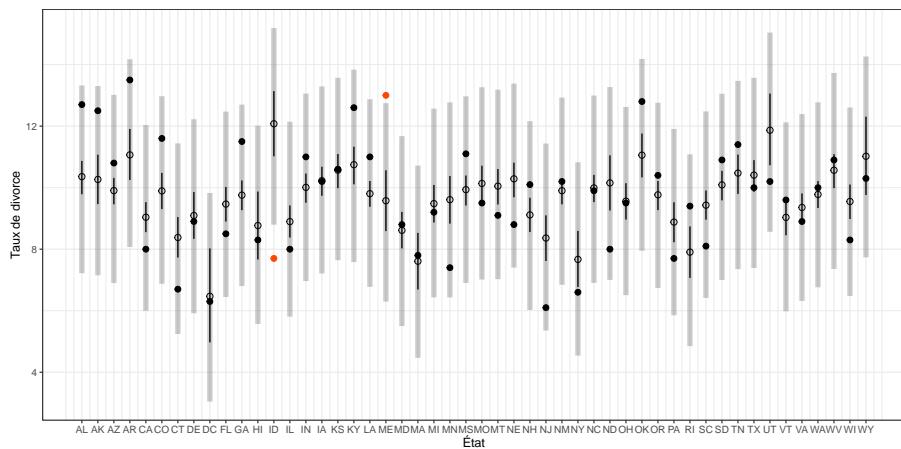


En plus de l'interprétation des paramètres, il est important d'évaluer les prédictions du modèle en les comparant aux données observées. Cela nous permet de savoir si le modèle rend bien compte des données et (surtout) où est-ce que le modèle échoue.

```
pp_check(mod3, type = "intervals", nsamples = 1e2, prob = 0.5, prob_outer = 0.95) +
  theme_bw(base_size = 20) + labs(x = "État", y = "Taux de divorce")
```



...



4.1.3 Toujours plus de prédicteurs

Pourquoi ne pas simplement construire un modèle incluant tous les prédicteurs et regarder ce qu'il se passe?

- Raison n°1 : Multicollinearité
- Raison n°2 : Post-treatment bias
- Raison n°3 : Overfitting (cf. Cours n°07)

4.1.3.1 Multicollinearité

Situation dans laquelle certains prédicteurs sont très fortement corrélés. Par exemple, essayons de prédire la taille d'un individu par la taille de ses jambes.

```
set.seed(666) # afin de pouvoir reproduire les résultats
```

```
N <- 100 # nombre d'individus
height <- rnorm(N, 179, 5) # génère N observations
leg_prop <- runif(N, 0.4, 0.5) # taille des jambes (proportion taille totale)
leg_left <- leg_prop * height + rnorm(N, 0, 0.5) # taille jambe gauche (+ erreur)
leg_right <- leg_prop * height + rnorm(N, 0, 0.5) # taille jambe droite (+ erreur)
df2 <- data.frame(height, leg_left, leg_right) # création d'une dataframe

head(df2) # affiche les six première lignes
```

```
##      height leg_left leg_right
## 1 182.7666 75.50967 76.00645
## 2 189.0718 81.10741 82.18046
## 3 177.2243 71.43856 71.49741
## 4 189.1408 82.81510 82.54405
## 5 167.9156 82.70860 84.00048
## 6 182.7920 84.86230 84.19933
```

On fit un modèle avec deux prédicteurs : un pour la taille de chaque jambe.

```
priors <- c(
  prior(normal(174, 10), class = Intercept),
  prior(normal(0, 10), class = b),
  prior(exponential(0.01), class = sigma)
)

mod4 <- brm(
  height ~ 1 + leg_left + leg_right,
  prior = priors,
  family = gaussian,
  data = df2
)
```

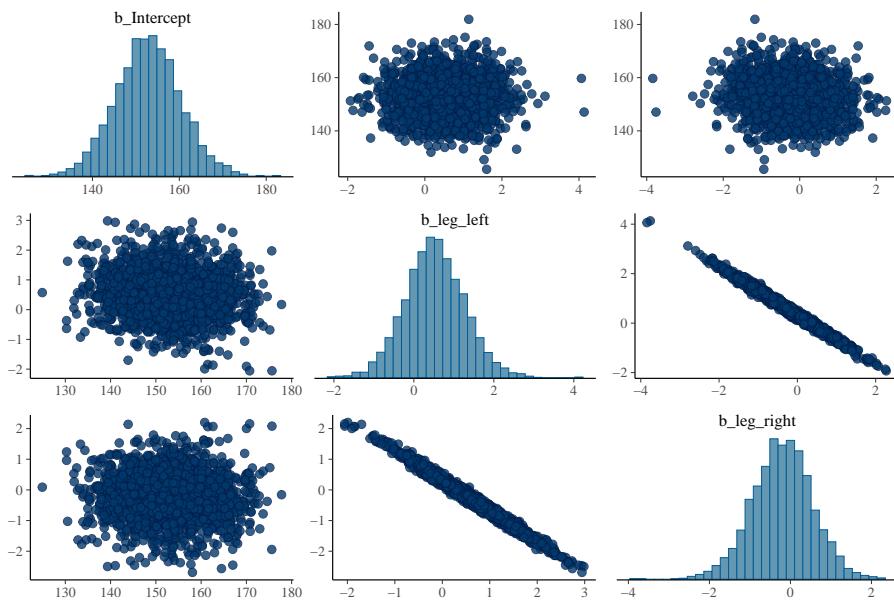
Les estimations semblent étranges... mais le modèle ne fait que répondre à la question qu'on lui pose : Une fois que je connais la taille de la jambe gauche, quelle est la valeur prédictive de la taille de la jambe droite (et vice versa) ?

```
summary(mod4) # look at the SE...
```

```
## Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: height ~ 1 + leg_left + leg_right
##   Data: df2 (Number of observations: 100)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Population-Level Effects:
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     152.82      7.44   138.29   167.72 1.00    3828    2264
## leg_left       0.52      0.75    -0.93    2.02 1.00    1543    1591
## leg_right     -0.20      0.75    -1.71    1.24 1.00    1558    1604
##
## Family Specific Parameters:
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma        4.94      0.35     4.30     5.67 1.00    2462    2260
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

Comment traquer la colinéarité de deux prédicteurs? En représentant la distribution postérieure de ces deux paramètres.

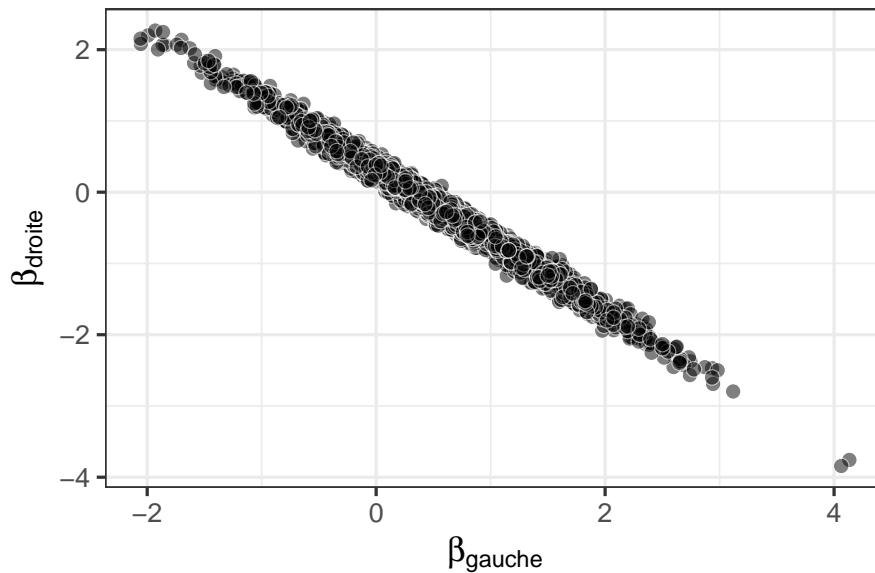
```
pairs(mod4, pars = parnames(mod4)[1:3])
```



Comment traquer la colinéarité de deux prédicteurs? En représentant la distribution postérieure de ces deux paramètres.

```
post <- posterior_samples(mod4)
```

```
post %>%
  ggplot(aes(x = b_leg_left, y = b_leg_right)) +
  geom_point(pch = 21, size = 4, color = "white", fill = "black", alpha = 0.5) +
  theme_bw(base_size = 20) +
  labs(x = expression(beta[gauche]), y = expression(beta[droite]))
```

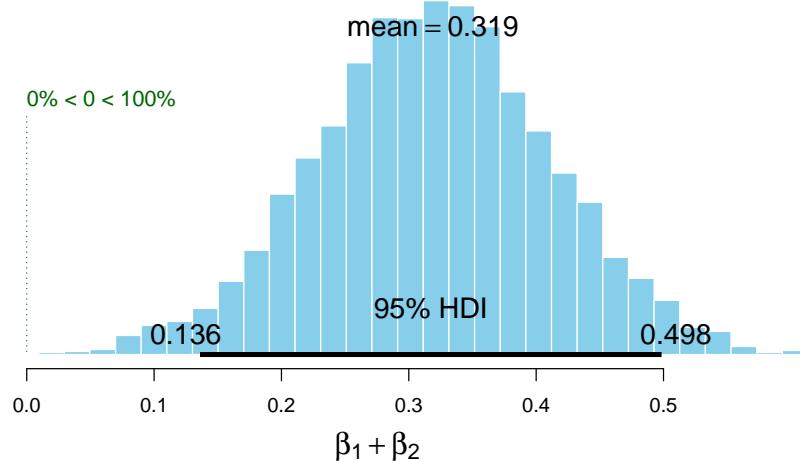


Le modèle précédent peut se réécrire en faisant apparaître la somme des deux prédicteurs β_1 et β_2 .

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + (\beta_1 + \beta_2)x_i$$

```
library(BEST)
sum_legs <- post$b_leg_left + post$b_leg_right
plotPost(sum_legs, xlab = expression(beta[1] + beta[2]), compVal = 0)
```



On crée un nouveau modèle avec seulement une jambe.

```
priors <- c(
  prior(normal(174, 10), class = Intercept),
  prior(normal(0, 10), class = b),
  prior(exponential(0.01), class = sigma)
)

mod5 <- brm(
  height ~ 1 + leg_left,
  prior = priors,
  family = gaussian,
  data = df2
)
```

En utilisant comme prédicteur une seule jambe, on retrouve l'estimation qui correspondait à la somme des deux pentes dans le modèle précédent.

```
summary(mod5)
```

```
## Family: gaussian
##  Links: mu = identity; sigma = identity
## Formula: height ~ 1 + leg_left
##  Data: df2 (Number of observations: 100)
##  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##        total post-warmup draws = 4000
##
## Population-Level Effects:
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     152.70      7.34   138.01   166.67 1.00     4158     2882
## leg_left       0.32      0.09     0.15     0.50 1.00     4144     3016
##
## Family Specific Parameters:
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma        4.93      0.37     4.29     5.76 1.00     3810     2551
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

4.1.3.2 Post-treatment bias

Problèmes qui arrivent lorsqu'on inclut des prédicteurs qui sont eux-mêmes définis directement ou indirectement par d'autres prédicteurs inclus dans le modèle.

Supposons par exemple qu'on s'intéresse à la pousse des plantes en serre. On voudrait savoir quel traitement permettant de réduire la présence de champignons améliore la pousse des plantes.

On commence donc par planter et laisser germer des graines, mesurer la taille initiale des pousses, puis appliquer différents traitements.

Enfin, on mesure à la fin de l'expérience la taille finale de chaque plante et la présence de champignons.

```
# nombre de plantes
N <- 100

# on simule différentes tailles à l'origine
h0 <- rnorm(N, mean = 10, sd = 2)

# on assigne différents traitements et on
# simule la présence de fungus et la pousse des plantes
treatment <- rep(0:1, each = N / 2)
fungus <- rbinom(N, size = 1, prob = 0.5 - treatment * 0.4)
h1 <- h0 + rnorm(N, mean = 5 - 3 * fungus)

# on rassemble les données dans une dataframe
df3 <- data.frame(h0, h1, treatment, fungus)

head(df3)
```

```
##          h0        h1 treatment fungus
## 1  8.842591 13.820383       0       0
## 2  5.094913  7.844256       0       1
## 3  9.423155 10.763637       0       1
## 4 13.008697 17.141846       0       0
## 5 11.566223 17.161368       0       0
## 6  9.520248 16.648277       0       0
```

...

$$\begin{aligned}h_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta_1 h_0 + \beta_2 T_i + \beta_3 F_i \\ \alpha &\sim \text{Normal}(0, 10) \\ \beta_1, \beta_2, \beta_3 &\sim \text{Normal}(0, 10) \\ \sigma &\sim \text{Exponential}(0.01)\end{aligned}$$

```

priors <- c(
  prior(normal(0, 10), class = Intercept),
  prior(normal(0, 10), class = b),
  prior(exponential(0.01), class = sigma)
)

mod6 <- brm(
  h1 ~ 1 + h0 + treatment + fungus,
  prior = priors,
  family = gaussian,
  data = df3
)

```

On remarque que l'effet du traitement est négligeable. La présence des champignons (`fungus`) est une conséquence de l'application du `treatment`. On demande au modèle si le traitement a une influence sachant que la plante a (ou n'a pas) développé de champignons...

```
summary(mod6)
```

```

## Family: gaussian
##  Links: mu = identity; sigma = identity
## Formula: h1 ~ 1 + h0 + treatment + fungus
##  Data: df3 (Number of observations: 100)
##  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##        total post-warmup draws = 4000
##
## Population-Level Effects:
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept      4.32     0.49    3.37   5.28 1.00    4412    2880
## h0             1.07     0.04    0.99   1.16 1.00    4667    2955
## treatment     -0.09     0.20   -0.49   0.31 1.00    4529    2326
## fungus        -2.65     0.23   -3.09  -2.18 1.00    4241    2800
##
## Family Specific Parameters:
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma       0.91     0.07    0.79    1.05 1.00    4198    3139
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).

```

Nous nous intéressons plutôt à l'influence du traitement sur la pousse. Il suffit de fitter un modèle sans la variable `fungus`. Remarque : il fait sens de prendre en compte h_0 , la taille initiale, car les différences observées pourraient masquer l'effet du traitement.

```
mod7 <- brm(  
  h1 ~ 1 + h0 + treatment,  
  prior = priors,  
  family = gaussian,  
  data = df3  
)
```

Note : on pourrait également utiliser la méthode `update()`...

```
mod7 <- update(mod6, formula = h1 ~ 1 + h0 + treatment)
```

```
summary(mod7)
```

```
## Family: gaussian  
## Links: mu = identity; sigma = identity  
## Formula: h1 ~ 1 + h0 + treatment  
## Data: df3 (Number of observations: 100)  
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
##          total post-warmup draws = 4000  
##  
## Population-Level Effects:  
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
## Intercept     2.24      0.70     0.80     3.59 1.00    4401    3233  
## h0            1.17      0.07     1.04     1.31 1.00    4277    3165  
## treatment     0.74      0.28     0.19     1.29 1.00    4764    3296  
##  
## Family Specific Parameters:  
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS  
## sigma       1.42      0.10     1.24     1.63 1.00    4106    2988  
##  
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS  
## and Tail_ESS are effective sample size measures, and Rhat is the potential  
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

L'influence du traitement est maintenant forte et positive...

4.1.4 Prédicteurs catégoriels

```
data(Howell1)
df4 <- Howell1

str(df4)

## 'data.frame':    544 obs. of  4 variables:
## $ height: num  152 140 137 157 145 ...
## $ weight: num  47.8 36.5 31.9 53 41.3 ...
## $ age   : num  63 63 65 41 51 35 32 27 19 54 ...
## $ male  : int  1 0 0 1 0 1 0 1 0 1 ...
```

Le **genre** est codé comme une **dummy variable**, c'est à dire une variable où chaque modalité est représentée soit par 0 soit par 1. On peut imaginer que cette nouvelle variable *active* le paramètre uniquement pour la catégorie codée 1, et le *désactive* pour la catégorie codée 0.

$$\begin{aligned} h_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha + \beta_m m_i \\ \alpha &\sim \text{Normal}(178, 100) \\ \beta_m &\sim \text{Normal}(0, 10) \\ \sigma &\sim \text{Uniform}(0, 50) \end{aligned}$$

```
priors <- c(
  prior(normal(178, 100), class = Intercept),
  prior(normal(0, 10), class = b),
  prior(exponential(0.01), class = sigma)
)

mod8 <- brm(
  height ~ 1 + male,
  prior = priors,
  family = gaussian,
  data = df4
)
```

L'intercept α représente la taille moyenne des femmes (car $\mu_i = \beta_m(0) = \alpha$).

```
fixef(mod8) # retrieves fixed effects
```

##	Estimate	Est.Error	Q2.5	Q97.5
----	----------	-----------	------	-------

```
## Intercept 134.828068 1.633549 131.704599 138.07054
## male      7.284156 2.301133 2.746952 11.67705
```

La pente β nous indique la différence de taille moyenne entre les hommes et les femmes. Pour obtenir la taille moyenne des hommes, il suffit donc d'ajouter α et β .

```
post <- posterior_samples(mod8)
mu.male <- post$b_Intercept + post$b_male
quantile(x = mu.male, probs = c(0.025, 0.5, 0.975) )

##      2.5%      50%     97.5%
## 138.7874 142.1054 145.4737
```

Au lieu d'utiliser un paramètre pour la différence entre les deux catégories, on pourrait estimer un paramètre par catégorie...

$$\begin{aligned} h_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha_f(1 - m_i) + \alpha_h m_i \end{aligned}$$

Cette formulation est strictement équivalente à la précédente car :

$$\begin{aligned} \mu_i &= \alpha_f(1 - m_i) + \alpha_h m_i \\ &= \alpha_f + (\alpha_m - \alpha_f)m_i \end{aligned}$$

où $(\alpha_m - \alpha_f)$ est égal à la différence entre la moyenne des hommes et la moyenne des femmes (i.e., β_m).

```
# on crée une nouvelle colonne pour les femmes
df4 <- df4 %>% mutate(female = 1 - male)

priors <- c(
  # il n'y a plus d'intercept dans ce modèle
  # prior(normal(178, 100), class = Intercept),
  prior(normal(0, 10), class = b),
  prior(exponential(0.01), class = sigma)
)

mod9 <- brm(
  height ~ 0 + female + male,
  prior = priors,
  family = gaussian,
  data = df4
)
```

```
summary(mod9)

## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: height ~ 0 + female + male
## Data: df4 (Number of observations: 544)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##         total post-warmup draws = 4000
##
## Population-Level Effects:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## female     131.17      1.63   127.96   134.31 1.00      4102     2760
## male       138.20      1.73   134.73   141.60 1.00      4251     3212
##
## Family Specific Parameters:
##             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma      27.67      0.86    26.05   29.35 1.00      3521     3110
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

4.1.4.1 Prédicteurs catégoriels, taille d'effet

$$\rho^2 = \frac{\sum_{i=1}^n \pi_i (\beta_i - \beta)^2}{\sigma^2 + \sum_{i=1}^n \pi_i (\beta_i - \beta)^2}$$

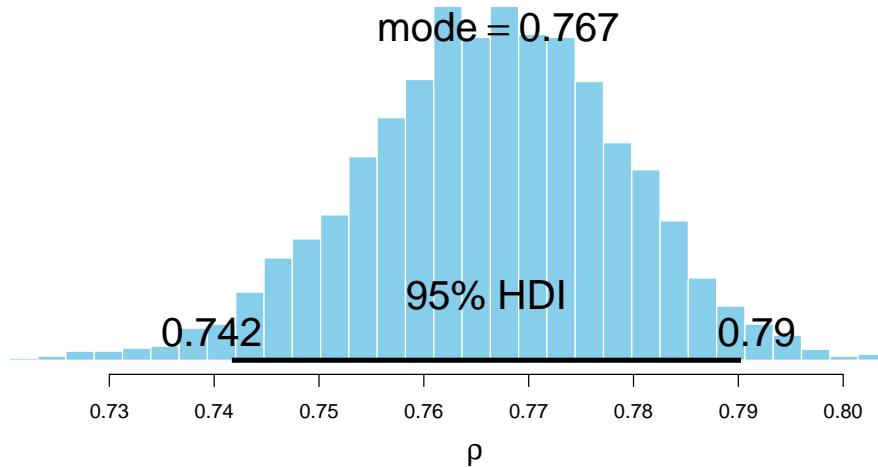
```
post <- posterior_samples(mod9)
pi <- sum(df4$male) / length(df4$male) # proportion of males

beta <- post$b_male # posterior samples for beta
sigma <- post$sigma # posterior samples for sigma

f1 <- pi * (beta - beta * pi)^2
rho <- f1 / (f1 + sigma^2)
```

$$\rho^2 = \frac{\sum_{i=1}^n \pi_i (\beta_i - \beta)^2}{\sigma^2 + \sum_{i=1}^n \pi_i (\beta_i - \beta)^2}$$

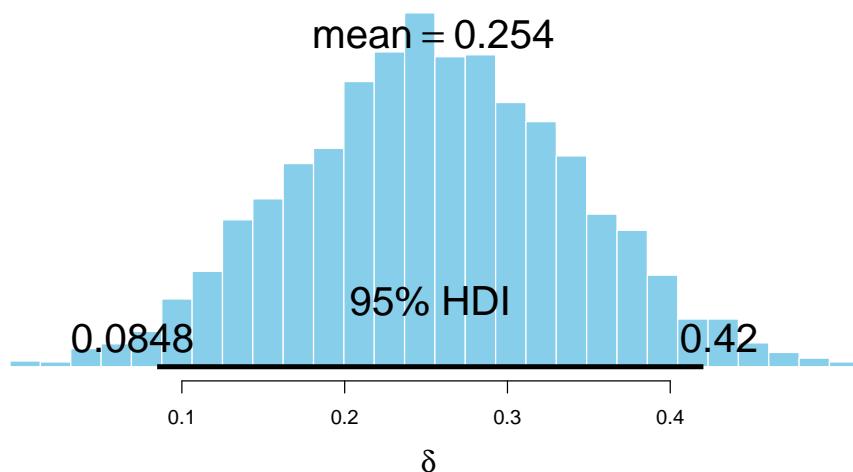
```
plotPost(rho, showMode = TRUE, cex = 2, xlab = expression(rho) )
```



...

$$\text{Cohen's } d = \frac{\text{différence des moyennes}}{\text{écart-type}}$$

```
plotPost((post$b_male - post$b_female) / post$sigma, cex = 2, xlab = expression(delta) )
```



4.1.5 Prédicteurs catégoriels, nombre de catégories > 3

```
data(milk)
df5 <- milk
str(df5)
```

```
## 'data.frame': 29 obs. of 8 variables:
## $ clade      : Factor w/ 4 levels "Ape","New World Monkey",...: 4 4 4 4 4 2 2 2 2 ...
## $ species    : Factor w/ 29 levels "A palliata","Alouatta seniculus",...: 11 8 9 10 16 ...
## $ kcal.per.g  : num  0.49 0.51 0.46 0.48 0.6 0.47 0.56 0.89 0.91 0.92 ...
## $ perc.fat   : num  16.6 19.3 14.1 14.9 27.3 ...
## $ perc.protein: num  15.4 16.9 16.9 13.2 19.5 ...
## $ perc.lactose: num  68 63.8 69 71.9 53.2 ...
## $ mass        : num  1.95 2.09 2.51 1.62 2.19 5.25 5.37 2.51 0.71 0.68 ...
## $ neocortex.perc: num  55.2 NA NA NA NA ...
```

Règle : pour k catégories, nous aurons besoin de $k - 1$ *dummy variables*. Pas la peine de créer une variable pour `ape`, qui sera notre *intercept*.

```
df5$clade.NWM <- ifelse(df5$clade == "New World Monkey", 1, 0)
df5$clade.OWM <- ifelse(df5$clade == "Old World Monkey", 1, 0)
df5$clade.S <- ifelse(df5$clade == "Strepsirrhine", 1, 0)
```

...

$$k_i \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_{NWM} NWM_i + \beta_{OWM} OWM_i + \beta_S S_i$$

$$\alpha \sim \text{Normal}(0.6, 10)$$

$$\beta_{NWM}, \beta_{OWM}, \beta_S \sim \text{Normal}(0, 1)$$

$$\sigma \sim \text{Exponential}(0.01)$$

...

Category	NWM_i	OWM_i	S_i	μ_i
Ape	0	0	0	$\mu_i = \alpha$
New World monkey	1	0	0	$\mu_i = \alpha + \beta_{NWM}$
Old World monkey	0	1	0	$\mu_i = \alpha + \beta_{OWM}$
Strepsirrhine	0	0	1	$\mu_i = \alpha + \beta_S$

...

```
priors <- c(
  prior(normal(0.6, 10), class = Intercept),
  prior(normal(0, 1), class = b),
  prior(exponential(0.01), class = sigma)
)

mod10 <- brm(
```

```
kcal.per.g ~ 1 + clade.NWM + clade.OWM + clade.S,
prior = priors,
family = gaussian,
data = df5
)

summary(mod10)

## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: kcal.per.g ~ 1 + clade.NWM + clade.OWM + clade.S
## Data: df5 (Number of observations: 29)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##         total post-warmup draws = 4000
##
## Population-Level Effects:
##             Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.55      0.04    0.46    0.63 1.00    3283    2606
## clade.NWM    0.17      0.06    0.04    0.29 1.00    3340    2561
## clade.OWM    0.24      0.07    0.10    0.37 1.00    3397    2943
## clade.S     -0.04      0.07   -0.18    0.11 1.00    3400    2890
##
## Family Specific Parameters:
##             Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma       0.13      0.02    0.10    0.18 1.00    3226    2843
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).

...
# retrieves posterior samples
post <- posterior_samples(mod10)

# retrieves posterior samples for each category
mu.ape <- post$b_Intercept
mu.NWM <- post$b_Intercept + post$b_clade.NWM
mu.OWM <- post$b_Intercept + post$b_clade.OWM
mu.S <- post$b_Intercept + post$b_clade.S
```

```
precis(data.frame(mu.ape, mu.NWM, mu.OWM, mu.S), prob = 0.95)
```

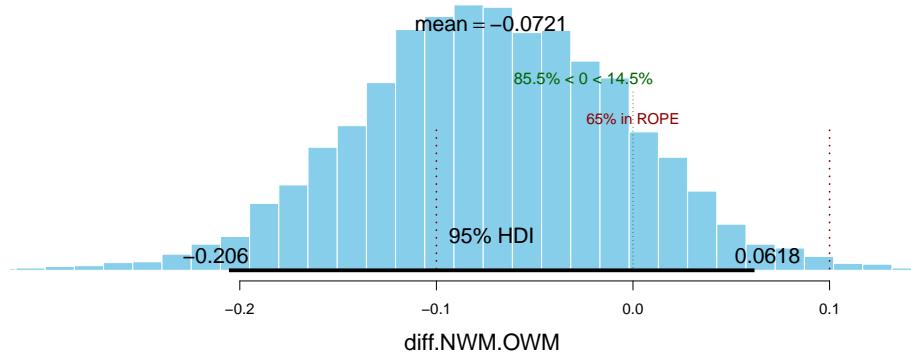
```
##               mean           sd      2.5%     97.5%   histogram
## mu.ape  0.5472530 0.04416200 0.4632161 0.6345129
## mu.NWM  0.7145580 0.04381130 0.6283760 0.8001637
## mu.OWM  0.7866443 0.05449798 0.6804187 0.8938789
## mu.S    0.5092347 0.05951719 0.3911451 0.6296908
```

Si on s'intéresse à la différence entre deux groupes, on peut calculer la distribution postérieure de cette différence.

```
diff.NWM.OWM <- mu.NWM - mu.OWM
quantile(diff.NWM.OWM, probs = c(0.025, 0.5, 0.975) )
```

```
##      2.5%      50%     97.5%
## -0.20781665 -0.07280640  0.05937144
```

```
plotPost(diff.NWM.OWM, compVal = 0, ROPE = c(-0.1, 0.1) )
```



Une autre manière de considérer les variables catégorielles consiste à construire un vecteur d'intercepts, avec un intercept par catégorie.

$$\begin{aligned} k_i &\sim \text{Normal}(\mu_i, \sigma) \\ \mu_i &= \alpha_{\text{clade}[i]} \\ \alpha_{\text{clade}[i]} &\sim \text{Normal}(0.6, 10) \\ \sigma &\sim \text{Exponential}(0.01) \end{aligned}$$

Comme on a vu avec l'exemple du genre, brms “comprend” automatiquement que c'est ce qu'on veut faire lorsqu'on fit un modèle sans intercept et avec un prédicteur catégoriel (codé en facteur).

```
priors <- c(
  prior(normal(0.6, 10), class = b),
  prior(exponential(0.01), class = sigma)
)

mod11 <- brm(
  # modèle sans intercept avec seulement un prédicteur catégoriel (facteur)
  kcal.per.g ~ 0 + clade,
  prior = priors,
  family = gaussian,
  data = df5
)

summary(mod11)

## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: kcal.per.g ~ 0 + clade
## Data: df5 (Number of observations: 29)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##         total post-warmup draws = 4000
##
## Population-Level Effects:
##                               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## cladeApe                  0.54     0.04    0.46    0.63 1.00    4789    2503
## cladeNewWorldMonkey        0.72     0.04    0.63    0.81 1.00    5125    2714
## cladeOldWorldMonkey        0.79     0.05    0.69    0.89 1.00    4834    2979
## cladeStrepsirrhine        0.51     0.06    0.40    0.63 1.00    4667    2685
##
## Family Specific Parameters:
##                               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma      0.13     0.02     0.10     0.17 1.00    3367    3005
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```

4.1.6 Interaction

Jusque là, les prédicteurs du modèle entretenaient des relations mutuellement indépendantes. Et si nous souhaitions que ces relations soient **conditionnelles**, ou **dépendantes** les unes des autres ?

Par exemple : on s'intéresse à la pousse des tulipes selon la quantité de lumière reçue et l'humidité du sol. Il se pourrait que la relation entre quantité de lumière reçue et pousse des tulipes soit différente selon l'humidité du sol. En d'autres termes, il se pourrait que la relation entre quantité de lumière reçue et pousse des tulipe soit **conditionnelle** à l'humidité du sol...

```
data(tulips)
df6 <- tulips

head(df6, 10)

##   bed water shade blooms
## 1   a     1     1    0.00
## 2   a     1     2    0.00
## 3   a     1     3 111.04
## 4   a     2     1 183.47
## 5   a     2     2  59.16
## 6   a     2     3  76.75
## 7   a     3     1 224.97
## 8   a     3     2  83.77
## 9   a     3     3 134.95
## 10  b     1     1  80.10
```

Modèle sans interaction :

$$\begin{aligned} B_i &\sim \text{Normal}(\mu, \sigma) \\ \mu_i &= \alpha + \beta_W W_i + \beta_S S_i \end{aligned}$$

Modèle avec interaction :

$$\begin{aligned} B_i &\sim \text{Normal}(\mu, \sigma) \\ \mu_i &= \alpha + \beta_W W_i + \beta_S S_i + \beta_{WS} W_i S_i \end{aligned}$$

On centre les prédicteurs (pour faciliter l'interprétation des paramètres).

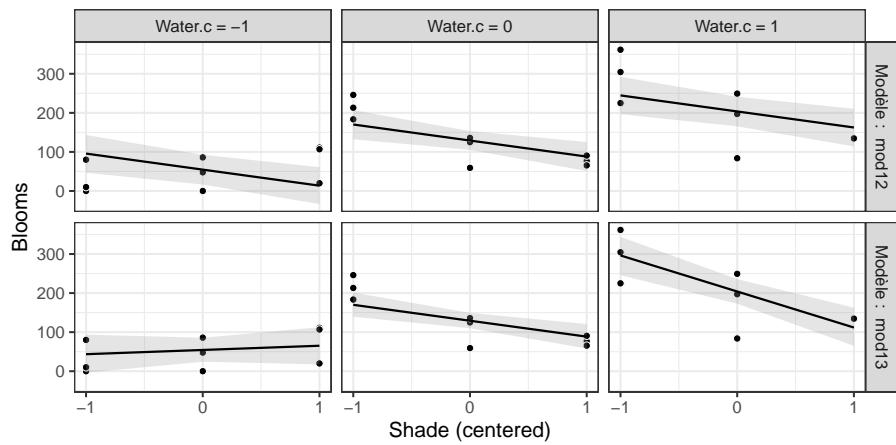
```
df6$shade.c <- df6$shade - mean(df6$shade)
df6$water.c <- df6$water - mean(df6$water)
```

```
priors <- c(
  prior(normal(130, 100), class = Intercept),
  prior(normal(0, 100), class = b),
  prior(exponential(0.01), class = sigma)
)
```

```
mod12 <- brm(  
  blooms ~ 1 + water.c + shade.c,  
  prior = priors,  
  family = gaussian,  
  data = df6  
)  
  
mod13 <- brm(  
  blooms ~ 1 + water.c * shade.c,  
  # equivalent to blooms ~ 1 + water.c + shade.c + water.c:shade.c  
  prior = priors,  
  family = gaussian,  
  data = df6  
)  
  
...  
  
##          term      mod12      mod13  
## 1    b_Intercept 129.12581 129.24273  
## 2    b_water.c   74.47215  74.84815  
## 3    b_shade.c  -41.06619 -40.72284  
## 4      sigma    63.28024  51.15140  
## 5     lprior  -22.20149 -27.73942  
## 6 b_water.c:shade.c       NA -51.55219
```

- L'intercept α représente la valeur attendue de `blooms` quand `water` et `shade` sont à 0 (i.e., la moyenne générale de la variable dépendante).
- La pente β_W nous donne la valeur attendue de changement de `blooms` quand `water` augmente d'une unité et `shade` est à sa valeur moyenne. On voit qu'augmenter la quantité d'eau est très bénéfique.
- La pente β_S nous donne la valeur attendue de changement de `blooms` quand `shade` augmente d'une unité et `water` est à sa valeur moyenne. On voit qu'augmenter la "quantité d'ombre" (diminuer l'exposition à la lumière) est plutôt délétère.
- La pente β_{WS} nous renseigne sur l'effet attendu de `water` sur `blooms` quand `shade` augmente d'une unité (et réciproquement).

Dans un modèle qui inclut un effet d'interaction, l'effet d'un prédicteur sur la mesure va dépendre de la valeur de l'autre prédicteur. La meilleure manière de représenter cette dépendance est de plotter la relation entre un prédicteur et la mesure, à différentes valeurs de l'autre prédicteur.



L'effet d'interaction nous indique que les tulipes ont besoin à la fois d'eau et de lumière pour pousser, mais aussi qu'à de faibles niveaux d'humidité, la luminosité a peu d'effet, tandis que cet effet est plus important à haut niveau d'humidité.

Cette explication vaut de manière **symétrique** pour l'effet de l'humidité sur la relation entre la luminosité et la pousse des plantes.



Glossaire

Affirmation du conséquent : Raisonnement fallacieux qui consiste à inférer la réalisation d'un antécédent sur la base de la réalisation du conséquent.

Facteur de Bayes (Bayes factor) : Support relatif fourni par un jeu de données pour deux hypothèses alternatives. Si on considère deux hypothèses \mathcal{H}_1 et \mathcal{H}_2 et des données x , alors le facteur de Bayes est $p(x | \mathcal{H}_1)/p(x | \mathcal{H}_2)$. Ce dernier peut-être interprété comme un “facteur de mise à jour” du rapport des chances a priori (i.e., avant de prendre connaissance des données) vers le rapport des chances a posteriori (i.e., après avoir pris connaissance des données).

Négation de l'antécédent : Raisonnement fallacieux qui consiste à affirmer une négation du conséquent sur la base d'une négation de l'antécédent.

Probabilité : Une valeur numérique comprise entre 0 et 1 respectant les règles du calcul probabiliste (cf. Chapitre 1).

Règle de la somme : Pour deux événements A et B , la règle de la somme nous dit que : $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$.

Règle du produit : Pour deux événements A et B , la règle du produit nous dit que : $\Pr(A, B) = \Pr(B) \cdot \Pr(A | B) = \Pr(A) \cdot \Pr(B | A)$.

Vraisemblance (Likelihood) : Une mesure du support fourni par les données pour certaines valeurs de paramètres. Après avoir observé des données x , la vraisemblance d'un paramètre θ est proportionnelle à $p(x | \theta)$.

Notations**Événements et ensembles**

- A ou B Les lettres majuscules représentent des *événements*.
 $\Pr(A)$ La probabilité de l'événement A .
 $\Pr(A|B)$ La probabilité conditionnelle de l'événement A sachant B .
 $\binom{n}{k}$ Coefficient binomial.

Variables aléatoires

- X ou Y Les lettres majuscules représentent des variables aléatoires.
 $E[X]$ Espérance de X .
 $Var[X]$ Variance de X .
 $p(x)$ Fonction de masse / densité de la variable aléatoire X .

Bibliographie

- Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2020). *Rmarkdown : Dynamic documents for r*. <https://CRAN.R-project.org/package=rmarkdown>
- Aust, F., & Barth, M. (2020). *Papaja : Prepare reproducible APA journal articles with r markdown*. <https://github.com/crsh/papaja>
- Blitzstein, J. K., & Hwang, J. (2019). *Introduction to probability* (Second edition). Taylor & Francis.
- Bürkner, P.-C. (2020). *Brms : Bayesian regression models using 'stan'*. <https://CRAN.R-project.org/package=brms>
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference : Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2), 261–304. <https://doi.org/10.1177/0049124104268644>
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference : A practical information-theoretic approach* (2nd ed). Springer.
- Campbell, D. T. (1990). The Meehlian Corroboration-Verisimilitude Theory of Science. *Psychological Inquiry*, 1(2), 142–147. <https://www.jstor.org/stable/1448769>
- Carnap, R. (1950). *Logical Foundations of Probability*. Chicago]University of Chicago Press.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997–1003. <https://doi.org/10.1037/0003-066X.49.12.997>
- Cumming, G. (2014). The new statistics : Why and how. *Psychological Science*, 25(1), 7–29. <https://doi.org/10.1177/0956797613504966>
- Cumming, G. (2012). *Understanding the new statistics : Effect sizes, confidence intervals, and meta-analysis*. (New York :).
- Dekking, M. (Ed.). (2005). *A modern introduction to probability and statistics : Understanding why and how*. Springer.
- Fidler, F., Thorn, F. S., Barnett, A., Kambouris, S., & Kruger, A. (2018). The Epistemic Importance of Establishing the Absence of an Effect : *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/2515245918770407>
- 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.
- Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.
- Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics : Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1), 8–38. <https://doi.org/10.1111/j.2044-8317.2011.02037.x>

- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In *A handbook for data analysis in the behavioral sciences : Methodological issues* (pp. 311–339). Lawrence Erlbaum Associates, Inc.
- Hájek, A. (2019). Interpretations of probability. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Fall 2019). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2019/entries/probability-interpret/>
- Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2018). *Introduction to Econometrics with R*. <https://www.econometrics-with-r.org/>
- Jaynes, E. T. (1986, November). Bayesian Methods : General Background. *Maximum Entropy and Bayesian Methods in Applied Statistics : Proceedings of the Fourth Maximum Entropy Workshop University of Calgary, 1984*. <https://doi.org/10.1017/CBO9780511569678.003>
- Judd, C. M., McClelland, G. H., & Ryan, C. S. (2009). *Data analysis : A model comparison approach, 2nd ed.* Routledge/Taylor & Francis Group.
- Keynes, J. M. (1921). *A Treatise On Probability*. Macmillan And Co.,. <http://archive.org/details/treatiseonprobab007528mbp>
- Kolmogorov, A. N. (1933). *Foundations of the theory of probability*. New York, USA : Chelsea Publishing Company.
- Kruschke, J. K. (2015). *Doing Bayesian data analysis : A tutorial with R, JAGS, and Stan* (2nd Edition). Academic Press.
- Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1), 155–177. <https://doi.org/10.3758/s13423-017-1272-1>
- Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian New Statistics : Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178–206. <https://doi.org/10.3758/s13423-016-1221-4>
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research : A Tutorial : *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/2515245918770963>
- Lindley, D. V. (2001). The Philosophy of Statistics. *Journal of the Royal Statistical Society : Series D (The Statistician)*, 49(3), 293–337. <https://doi.org/10.1111/1467-9884.00238>
- McElreath, R. (2016a). *Rethinking : Statistical rethinking book package*.
- McElreath, R. (2016b). *Statistical rethinking : A Bayesian course with examples in R and Stan*. CRC Press/Taylor & Francis Group.
- McElreath, R. (2020). *Statistical rethinking : A Bayesian course with examples in R and Stan* (2nd ed.). Taylor; Francis, CRC Press.
- Meehl, Paul E. (1990a). Appraising and amending theories : The strategy of Lakatosian defense and two principles that warrant It. *Psychological Inquiry*, 1(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1
- Meehl, Paul E. (1967). Theory-testing in Psychology and Physics : A methodological paradox. *Philosophy of Science*, 34(2), 103–115. <https://doi.org/10.1086/288135>
- Meehl, Paul E. (1978). Theoretical risks and tabular asterisks : Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834. <https://doi.org/10.1037/0022-006X.46.4.806>
- Meehl, P. E. (1986). What Social Scientists Don't Understand. In D. W. Fiske & R. A. Shweder

- (Eds.), *Metatheory in social science : Pluralisms and subjectivities* (p. 24). Chicago : University of Chicago Press.
- Meehl, Paul E. (1990b). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*. <https://doi.org/10.2466/pr0.1990.66.1.195>
- Meehl, Paul E. (1997). The problem is epistemology, not statistics : Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. *What If There Were No Significance Tests?*, 393–425. <http://citeserx.ist.psu.edu/viewdoc/citations;jsessionid=72FF987997EFB5F0602B02E1A2E04E40?doi=10.1.1.693.9583>
- Morey, R. D., Homer, S., & Proulx, T. (2018). Beyond Statistics : Accepting the Null Hypothesis in Mature Sciences : *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/2515245918776023>
- Noël, Y. (2015). *Psychologie statistique avec R*. EDP SCIENCES.
- Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin*, 102(1), 159–163. <https://doi.org/10.1037/0033-2909.102.1.159>
- R Core Team. (2019). *R : A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, 113(3), 553–565. <https://doi.org/10.1037/0033-2909.113.3.553>
- Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Collabra*, 2(1), 6. <https://doi.org/10.1525/collabra.28>
- Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3), e2000797. <https://doi.org/10.1371/journal.pbio.2000797>
- Talbot, M. (2015). *Critical Reasoning : A Romp through the Foothills of Logic for the Complete Beginner* (C. Wood, Ed.; 1 edition). CreateSpace Independent Publishing Platform.
- Wickham, H. (2019). *Tidyverse : Easily install and load the 'tidyverse'*. <https://CRAN.R-project.org/package=tidyverse>
- Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C., Woo, K., & Yutani, H. (2019). *ggplot2 : Create elegant data visualisations using the grammar of graphics*. <https://CRAN.R-project.org/package=ggplot2>
- Xie, Y. (2020a). *Bookdown : Authoring books and technical documents with r markdown*. <https://CRAN.R-project.org/package=bookdown>
- Xie, Y. (2020b). *Knitr : A general-purpose package for dynamic report generation in r*. <https://CRAN.R-project.org/package=knitr>