

Introduction à la modélisation statistique bayésienne

Un cours en R, Stan, et brms

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

Préface 🙌 🙌

Ce cours est grandement inspiré des livres suivants :

- McElreath, R. (2016, 2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.
- Kurz, S. (2019). *Statistical Rethinking with brms, ggplot2, and the tidyverse*. Available [online](#).
- Kruschke, J. K. (2015). *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan*. Academic Press / Elsevier.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis, third edition*. London: CRC Press.
- Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. SAGE Publications Ltd.
- Nicenboim, B., Schad, D., & Vasishth, S. (2021). *An Introduction to Bayesian Data Analysis for Cognitive Science*. Available [online](#).

Les slides, données, et scripts seront disponibles juste avant chaque séance sur Github :

<https://github.com/lnalborczyk/IMSB2022>.



Objectifs

Objectifs généraux :

- Comprendre les concepts fondamentaux de la statistique bayésienne.
- Être capable d'implémenter des modèles bayésiens simples en R.
- Réaliser que l'approche bayésienne est plus intuitive que l'approche fréquentiste.

Objectifs pratiques :

- Être capable de réaliser une analyse complète (i.e., identification du modèle approprié, écriture du modèle mathématique, implémentation en R, interprétation et report des résultats) d'un jeu de données simple.



Planning

Cours n°01 : Introduction à l'inférence bayésienne

Cours n°02 : Modèle Beta-Binomial

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

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

Cours n°05 : Markov Chain Monte Carlo

Cours n°06 : Modèle linéaire généralisé

Cours n°07 : Comparaison de modèles

Cours n°08 : Modèles multi-niveaux

Cours n°09 : Modèles multi-niveaux généralisés

Cours n°10 : Data Hackathon



Axiomes des probabilités (Kolmogorov, 1933)

Une probabilité est une valeur numérique assignée à un événement $\{A\}$, compris comme une possibilité appartenant à l'univers $\{\Omega\}$ (l'ensemble de toutes les issues possibles).

Les probabilités se conforment aux axiomes suivants :

- Non-négativité : $\{\Pr(A) \geq 0\}$
- Normalisation : $\{\Pr(\Omega) = 1\}$
- Additivité (pour des événements incompatibles) : $\{\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2)\}$

Le dernier axiome est également connu comme la **règle de la somme**, et peut se généraliser à des événements non mutuellement exclusifs : $\{\Pr(A_1 \cup A_2) = \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2)\}$.



Interprétations probabilistes

Quelle est la probabilité...

- D'obtenir un chiffre pair sur un lancer de dé ?
- Que j'apprenne quelque chose pendant cette formation ?

Est-ce qu'il s'agit, pour chaque exemple, de la même **sorte** de probabilité ?



Interprétation classique (ou théorique)

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

Problème : cette définition est uniquement applicable aux situations dans lesquelles il n'y a qu'un nombre **fini** de résultats possibles **équiprobables**...

Par exemple, quelle est la probabilité qu'il pleuve demain ?

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



Interprétation fréquentiste (ou empirique)

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

Où n_x est le nombre d'occurrences de l'événement x et n_t le nombre total d'essais.

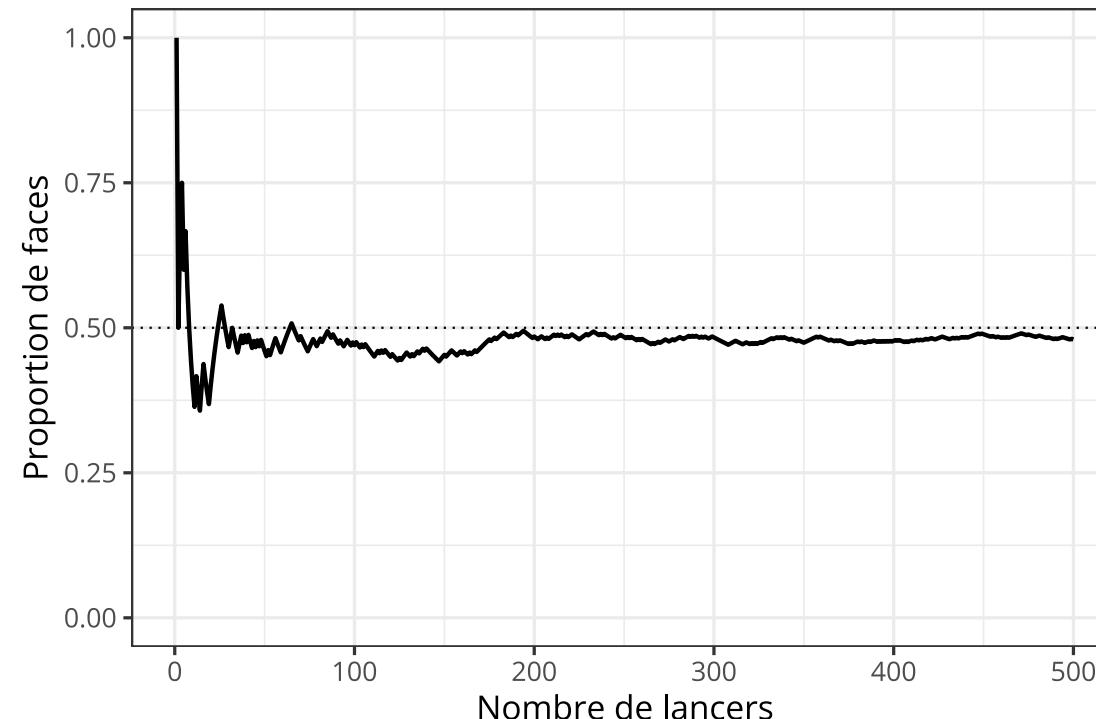
L'interprétation **fréquentiste** postule que, à long-terme (i.e., quand le nombre d'essais s'approche de l'infini), la fréquence relative va converger exactement vers ce qu'on appelle "probabilité".

Conséquence : le concept de probabilité s'applique uniquement aux **collectifs**, et non aux événements singuliers.



Interprétation fréquentiste (ou empirique)

```
1 library(tidyverse)
2
3 sample(c(0, 1), 500, replace = TRUE) %>%
4     data.frame %>%
5     mutate(x = seq_along(.), y = cumsum(.) / seq_along(.)) %>%
6     ggplot(aes(x = x, y = y), log = "y") +
7     geom_line(lwd = 1) +
8     geom_hline(yintercept = 0.5, lty = 3) +
9     xlab("Nombre de lancers") +
10    ylab("Proportion de faces") +
11    ylim(0, 1)
```



Limites de l'interprétation fréquentiste...

Quelle classe de référence ? Quelle est la probabilité que je vive jusqu'à 80 ans ? En tant qu'homme ? En tant que Français ?

Quid des événements qui ne peuvent pas se répéter ? Quelle est la probabilité que j'apprenne quelque chose pendant cette formation ?

À partir de combien de lancers (d'une pièce par exemple) a-t-on une bonne approximation de la probabilité ? Une classe finie d'événements de taille $\backslash(n\backslash)$ ne peut produire que des fréquences relatives de précision $\backslash(1/n\backslash)\dots$



Interprétation propensionniste

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) 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.

Conséquence : 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.



Interprétation logique

Il y a 10 étudiants dans cette salle

9 portent un t-shirt vert

1 porte un t-shirt rouge

Une personne est tirée au sort...

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

Conclusion n°2 : l'étudiant tiré au sort porte un t-shirt vert ✗

Conclusion n°3 : l'étudiant tiré au sort porte un t-shirt rouge ✗



Interprétation logique

L'interprétation logique du concept de probabilité essaye de généraliser la logique (vrai / faux) au monde probabiliste. La probabilité représente donc le *degré de support logique* qu'une conclusion peut avoir, relativement à un ensemble de prémisses ([Carnap, 1971](#); [Keynes, 1921](#)).

Conséquence : toute probabilité est **conditionnelle**.



Interprétation bayésienne

La probabilité est une mesure du degré 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 [...], 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 (fréquence).



Interprétations probabilistes

- Interprétation classique (Laplace, Bernouilli, Leibniz)
- **Interprétation fréquentiste** (Venn, Reichenbach, von Mises)
- Interprétation propensionniste (Popper, Miller)
- Interprétation logique (Keynes, Carnap)
- **Interprétation bayésienne** (Jeffreys, de Finetti, Savage)

[Voir plus de détails sur la Stanford Encyclopedia of Philosophy.](#)



Interprétations probabilistes - résumé

Probabilité 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, bayésienne.

Probabilité 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, fréquentiste.





Aparté - Vous avez dit “hasard” ?

Une suite de nombres “au hasard”...

```
[1] 79 27 98 28 42
```

runif, rnorm, rbinom...

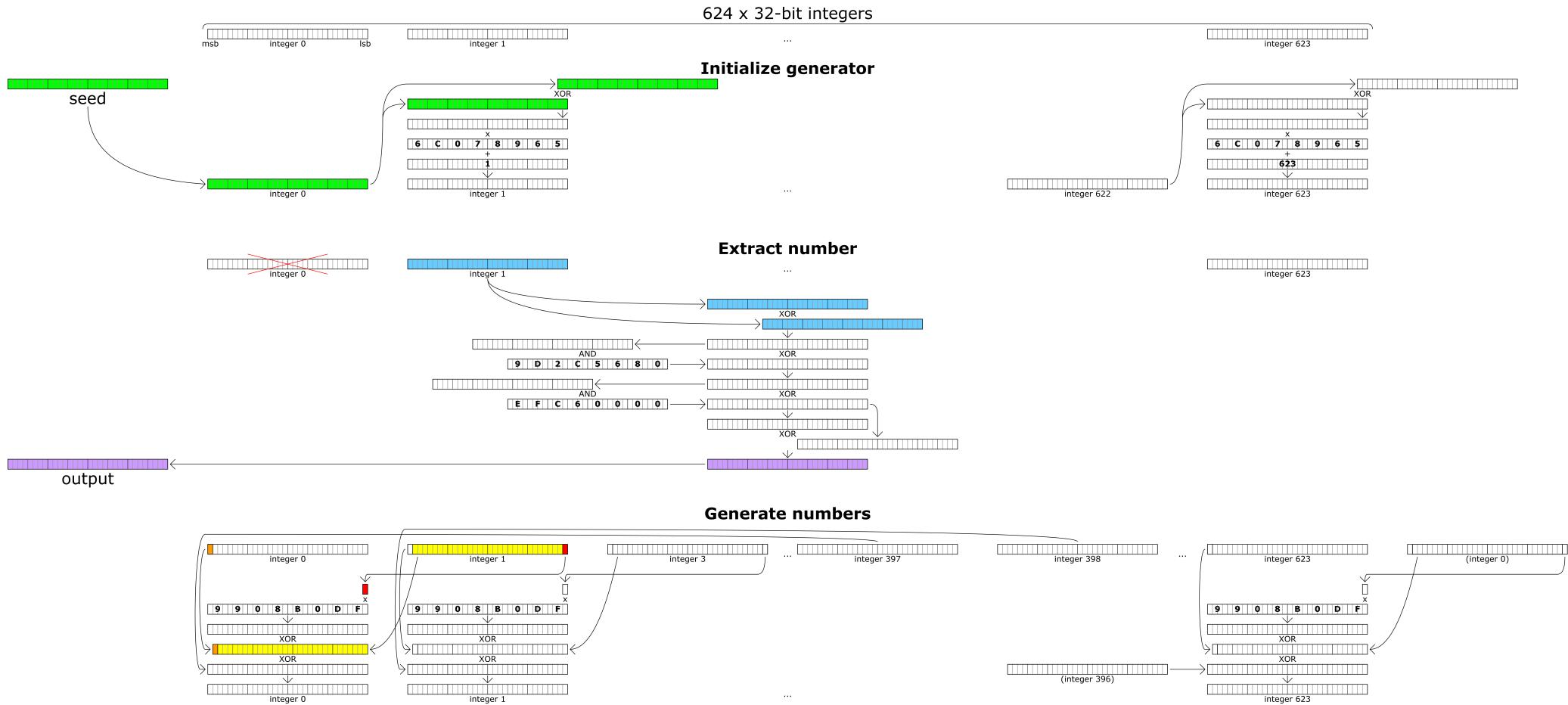
```
1 RNGkind() # default random number generator
```

```
[1] "Mersenne-Twister" "Inversion"      "Rejection"
```



Mersenne-Twister

Inventé par Makoto Matsumoto et Takuji Nishimura en 1997, le Mersenne Twister est un générateur de nombres pseudo-aléatoires. L'algorithme Mersenne-Twister est le générateur par défaut en **Python**, **Ruby**, **R**, **Matlab**...



Hasard déterministe

$$\{Y_i\} \sim \text{Uniform}(10,100)$$

```
1 set.seed(666)
2 as.integer(runif(5, 10, 100))
```

```
[1] 79 27 98 28 42
```

Si je connais la *graine* d'origine (i.e., le “départ” de l'algorithme) et le fonctionnement précis de l'algorithme, je peux prédire les nombres qui vont être générés. Il sera alors difficile de soutenir que ces nombres auront été générés “au hasard”...

```
[1] 79 27 98 28 42 76
[1] 79 27 98 28 42 76 98
```



Le langage de l'incertain

Selon [Wikipédia](#), ‘le hasard est un concept indiquant l’aspect aléatoire des choses, des événements ou encore des décisions, signalant ainsi l’impossibilité de prévoir avec certitude un fait quelconque. Ainsi, pour éclairer le sens du mot, il peut-être dit que le hasard est synonyme d’“imprévisibilité”, ou “imprédictibilité”.

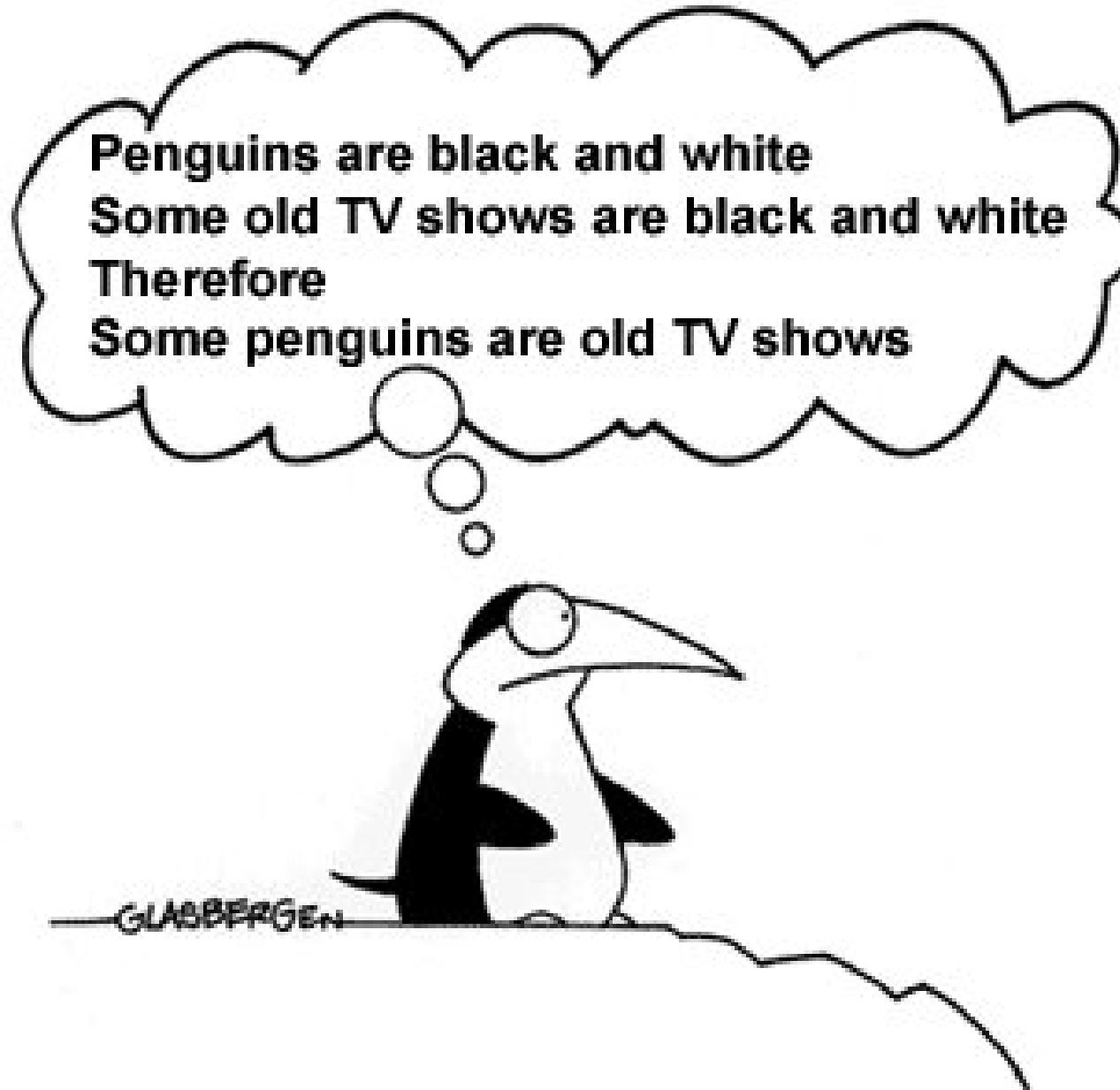
“

Ce que nous appelons hasard n'est et ne peut être que la cause ignorée d'un effet connu
(Voltaire).

Le hasard apparaît donc comme un **état subjectif d'indétermination des causes**. Ou comme dirait Richard McElreath, *randomness is a proxy for lack of knowledge*. Pour parler de l'incertain (i.e., pour quantifier l'incertitude), on utilise les probabilités comme langage.



Un peu de logique



Un peu de logique, quelques syllogismes

Exemple 1

- Si un suspect ment, il transpire. (On observe que) Ce suspect transpire.
- Par conséquent, ce suspect ment.

Exemple 2

- Si un suspect transpire, il ment. (On observe que) Ce suspect ne transpire pas.
- Par conséquent, ce suspect ne ment pas.

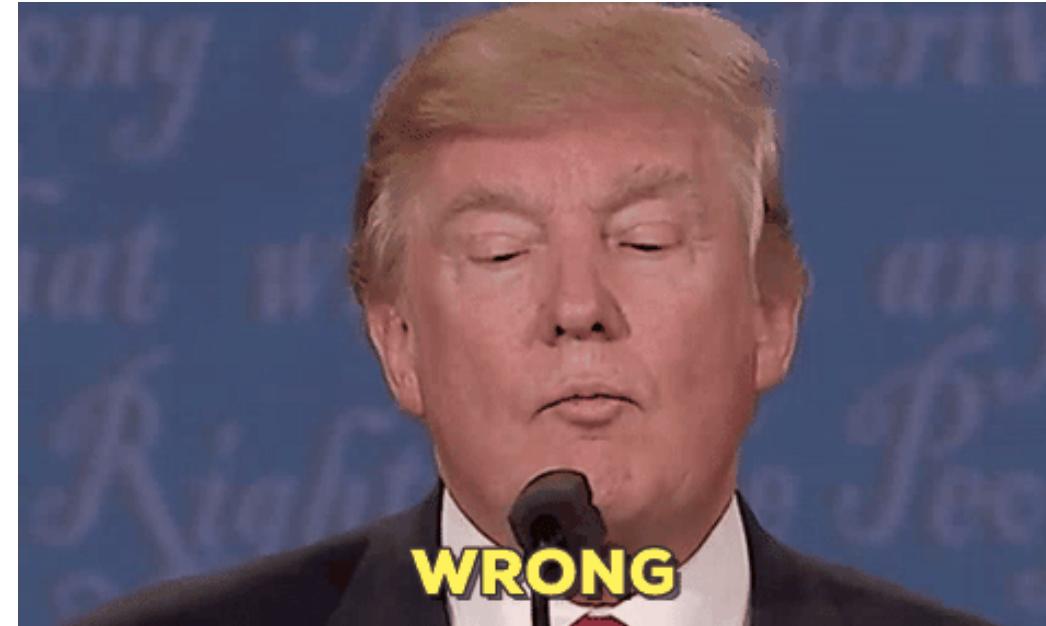
Exemple 3

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



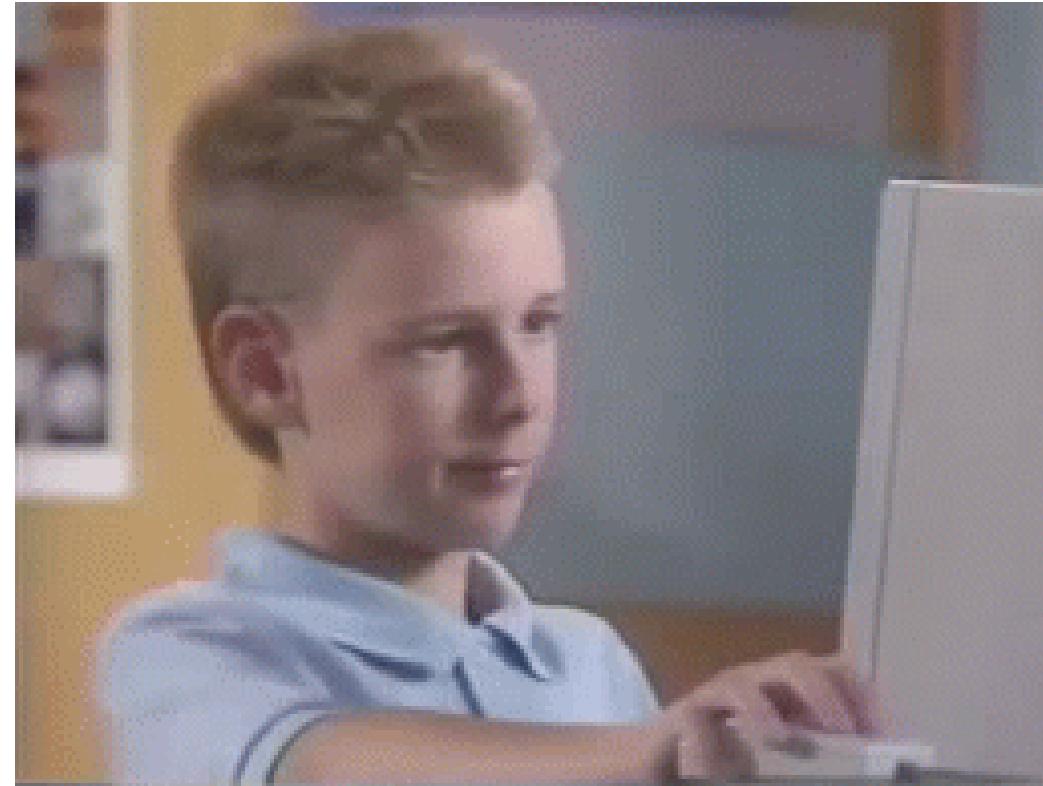
Arguments invalides

- Affirmation du conséquent : $\vdash (\frac{A \rightarrow B, \neg B}{A})$
- Si il a plu, alors le sol est mouillé (A implique B). Le sol est mouillé (B). Donc il a plu (A).
- Négation de l'antécédent : $\vdash (\frac{\neg A \rightarrow \neg B}{A \rightarrow B})$
- Si il a plu, alors le sol est mouillé (A implique B). Il n'a pas plu (non A). Donc le sol n'est pas mouillé (non B).



Arguments valides

- Modus ponens : $\vdash (\frac{A \rightarrow B, A}{B})$
- Si on est lundi, alors John ira au travail (A implique B). On est lundi (A). Donc John ira au travail (B).
- Modus tollens : $\vdash (\frac{\neg B, A \rightarrow B}{\neg A})$
- Si mon chien détecte un intru, alors il aboie (A implique B). Mon chien n'a pas aboyé (non B). Donc il n'a pas détecté d'intrus (non A).



Logique, fréquentisme et raisonnement probabiliste

Le *modus tollens* est un des raisonnements logiques les plus importants et les plus performants. Dans le cadre de l'inférence statistique, il s'applique parfaitement au cas suivant : “Si $\backslash(H_{\{0\}})$ est vraie, alors $\backslash(x)$ ne devrait pas se produire. On observe $\backslash(x)$. Alors $\backslash(H_{\{0\}})$ est fausse”.

Mais nous avons le plus souvent affaire à des hypothèses “continues”, probabilistes.

L'inférence fréquentiste (fishérienne) est elle aussi probabiliste, de la forme “Si $\backslash(H_{\{0\}})$ est vraie, alors $\backslash(x)$ est peu probable. On observe $\backslash(x)$. Alors $\backslash(H_{\{0\}})$ est peu probable.”

Or cet argument est invalide, le *modus tollens* ne s'applique pas au monde probabiliste (e.g., [Pollard & Richardson, 1987](#); [Jeffrey N. Rouder et al., 2016a](#)).

Par exemple : *si un individu est un homme, alors il est peu probable qu'il soit pape. François est pape. François n'est donc certainement pas un homme...*



L'échec de la falsification

Poppérisme naïf: la science progresse par falsification logique, donc la statistique devrait viser la falsification. Mais...

- Les hypothèses théoriques ne sont pas les modèles (hypothèses statistiques).

“

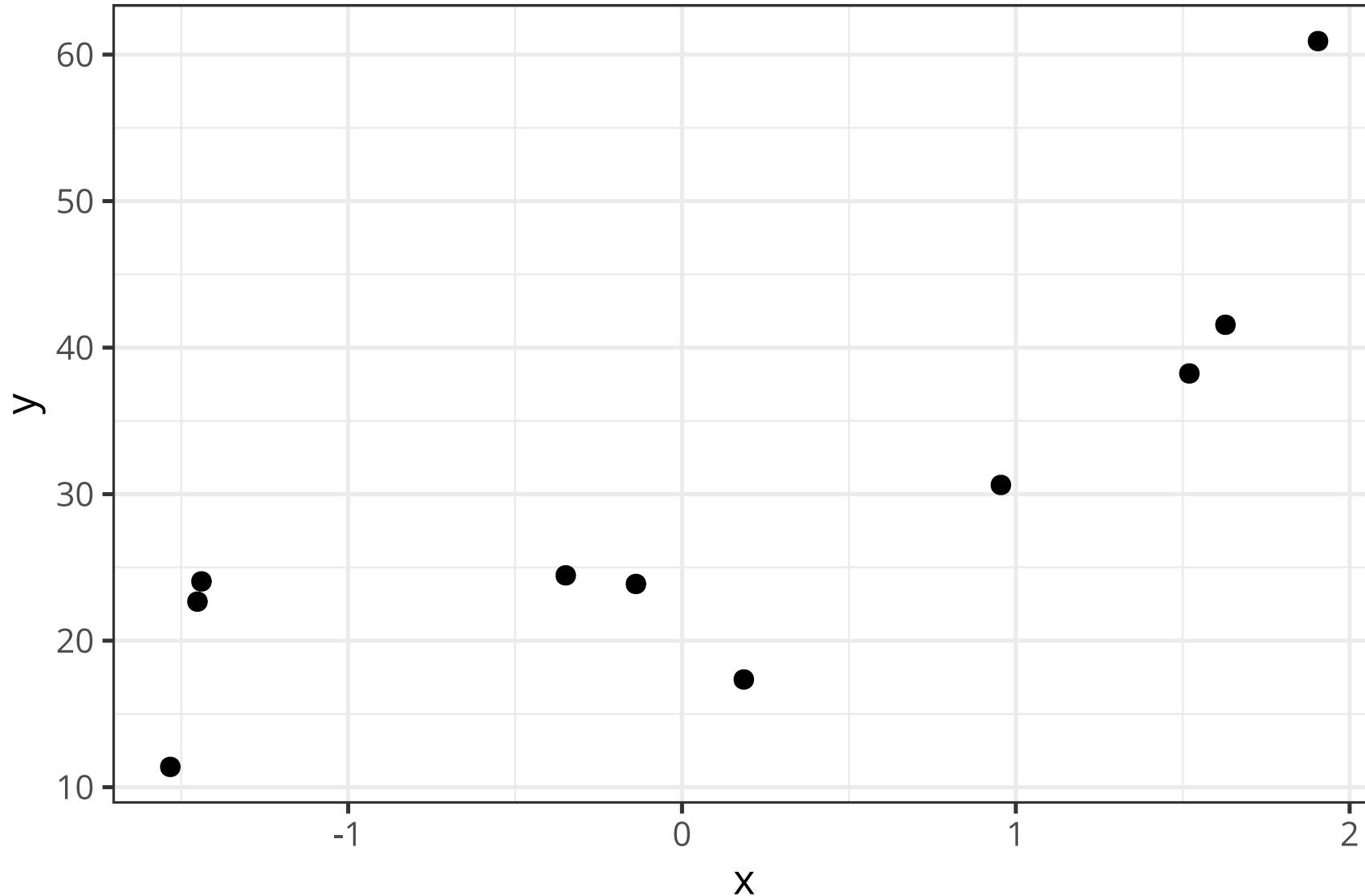
“Models are devices that connect theories to data. A model is an instantiation of a theory as a set of probabilistic statements” .([Jeffrey N. Rouder et al., 2016b](#)).

- Nos hypothèses sont souvent probabilistes.
- Erreurs de mesure.
- 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.



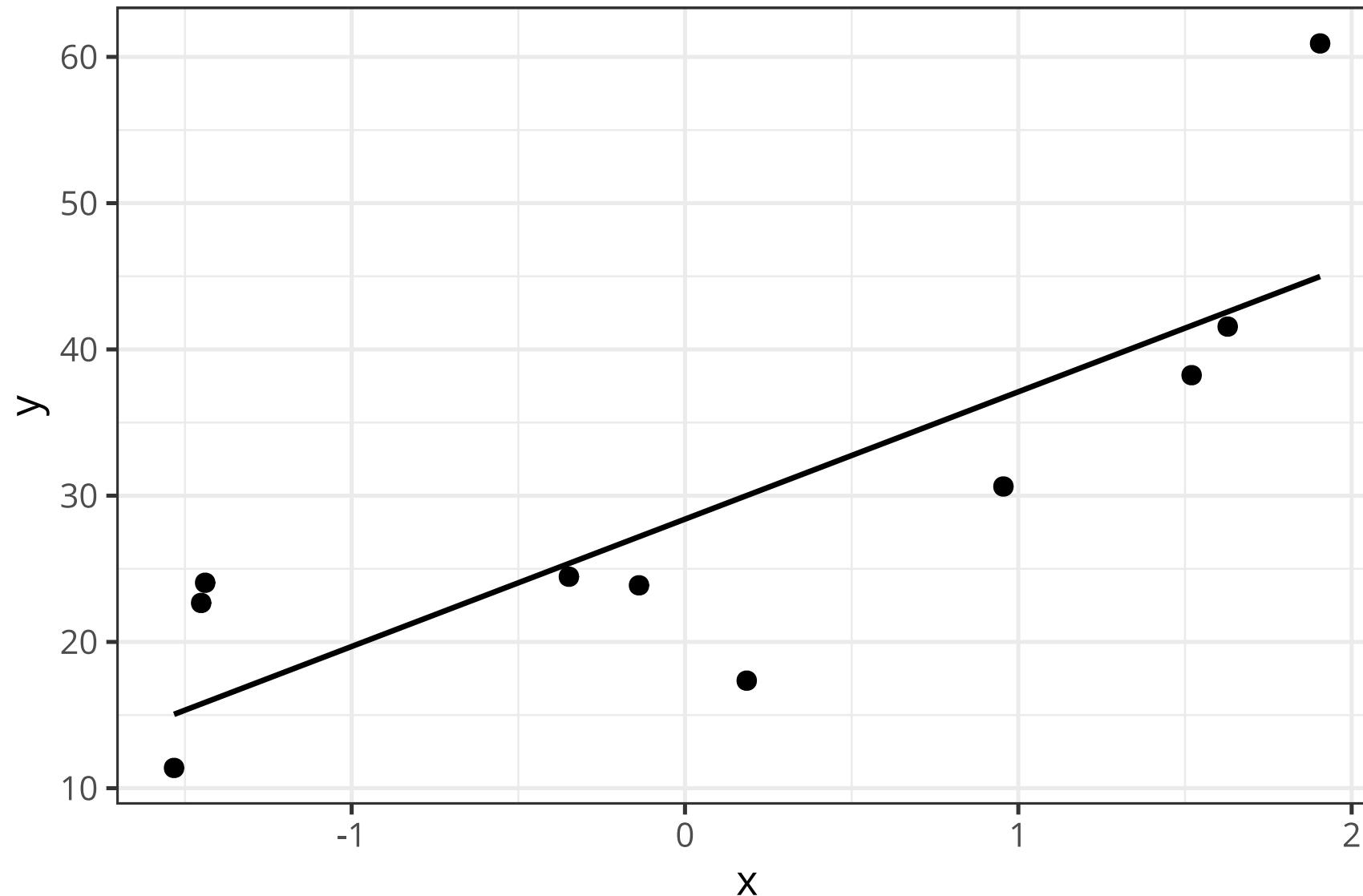
Comparaison de modèles

On s'intéresse au lien entre deux variables aléatoires continues, $\langle x \rangle$ et $\langle y \rangle$.



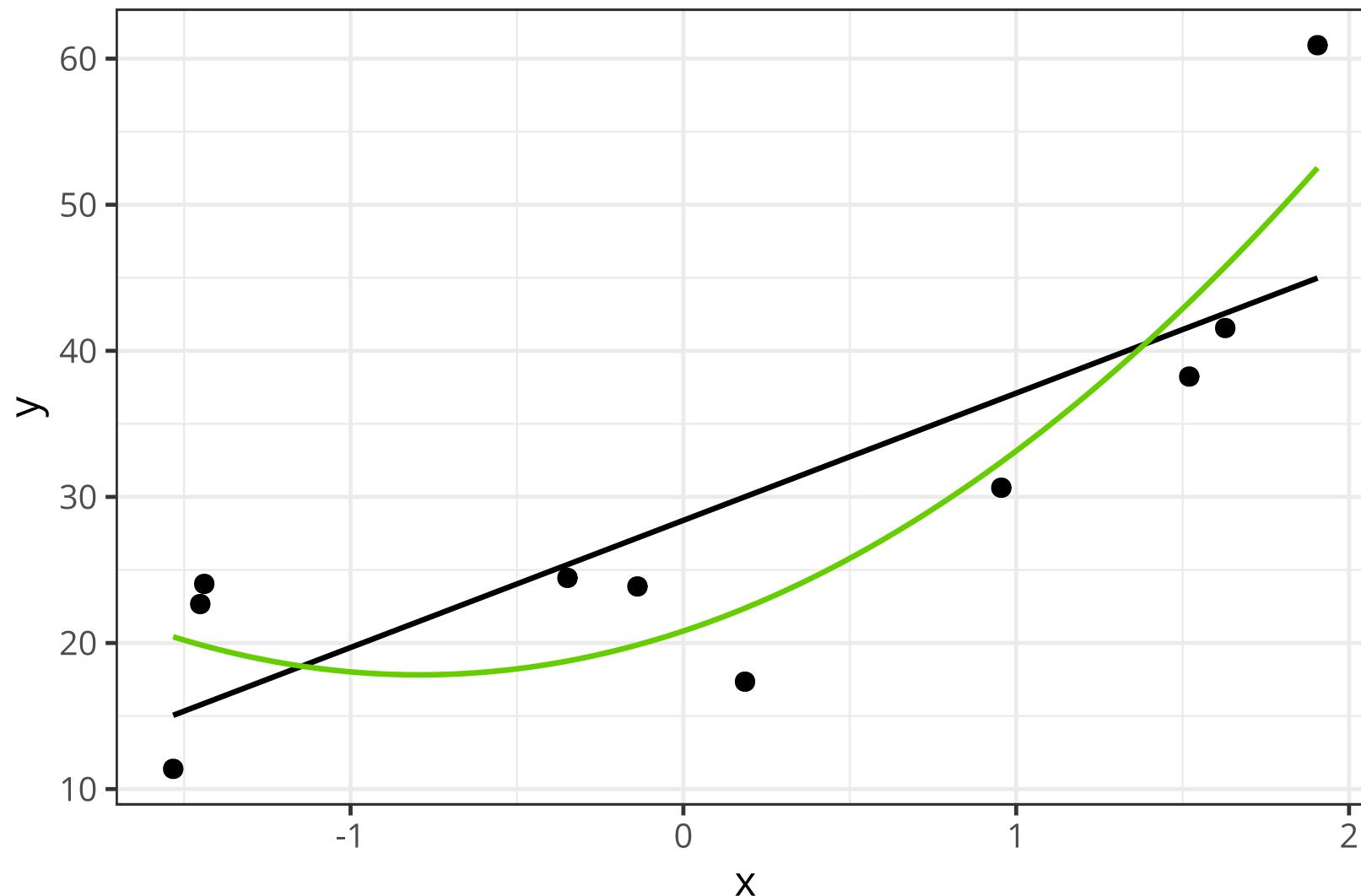
Comparaison de modèles

L'hypothèse de modélisation la plus simple est de postuler une relation linéaire.



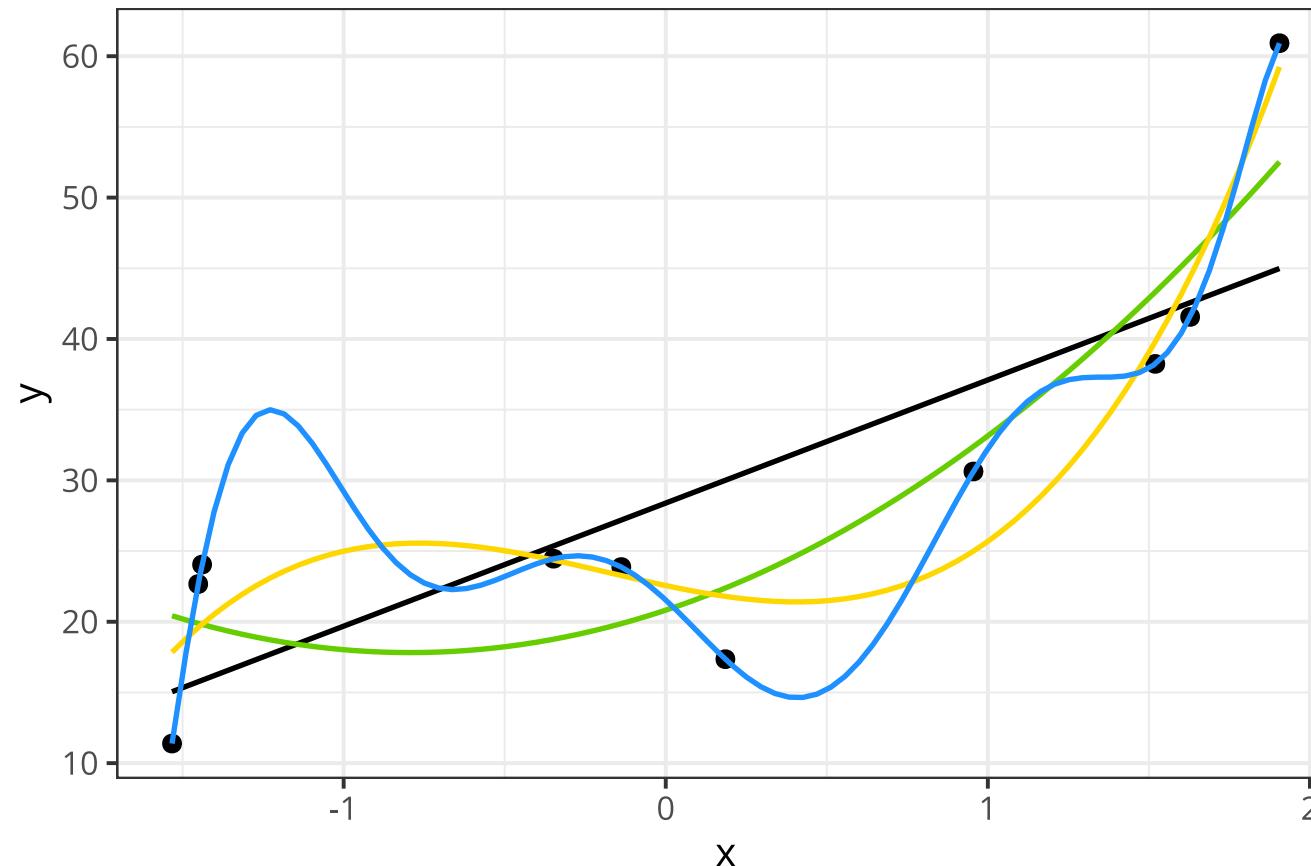
Comparaison de modèles

Cette description peut-être *améliorée* pour mieux prendre en compte les données qui s'écartent de la prédiction linéaire.



Comparaison de modèles

Un ensemble de $\backslash(N\backslash)$ points peut être exhaustivement (i.e., sans erreur) décrit par une fonction polynomiale d'ordre $\backslash(N-1\backslash)$. Augmenter la complexité du modèle améliore donc la précision de notre description des données mais réduit la généralisabilité de ses prédictions (*bias-variance tradeoff*).



Nous avons besoin d'outils qui prennent en compte le rapport qualité de la description / complexité, c'est à dire la **parcimonie** des modèles (e.g., AIC, WAIC).



Notre stratégie

Besoin d'un cadre pour développer des modèles cohérents. Nos outils :

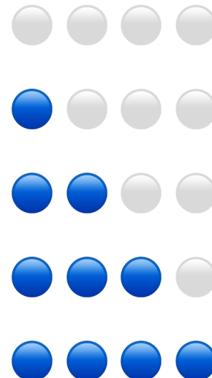
- **Bayesian data analysis** : utiliser les probabilités pour décrire l'incertitude.
- **Multilevel modelling** : des modèles à multiples niveaux d'incertitude.
- Approche par **comparaison de modèles** : au lieu d'essayer de falsifier un “null model”, on va comparer des modèles intéressants (AIC, WAIC).



Exercice - Problème du sac de billes

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*) :

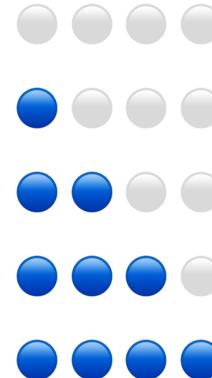


Exercice - Problème du sac de billes

Le but est de déterminer quelle combinaison serait 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 là, quelle **inférence** peut-on faire sur le contenu du sac ? En d'autres termes, que peut-on dire de la probabilité de chaque hypothèse ?



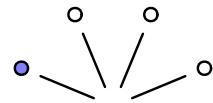
02:00
Zadisław Nalborszyk - IT-IB2022



Énumérer les possibilités

Hypothèse : ● ● ○ ○

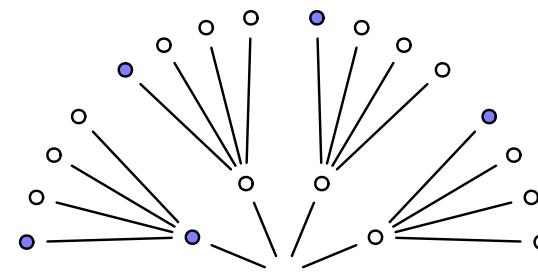
Données : ●



Énumérer les possibilités

Hypothèse : ● ● ○ ○

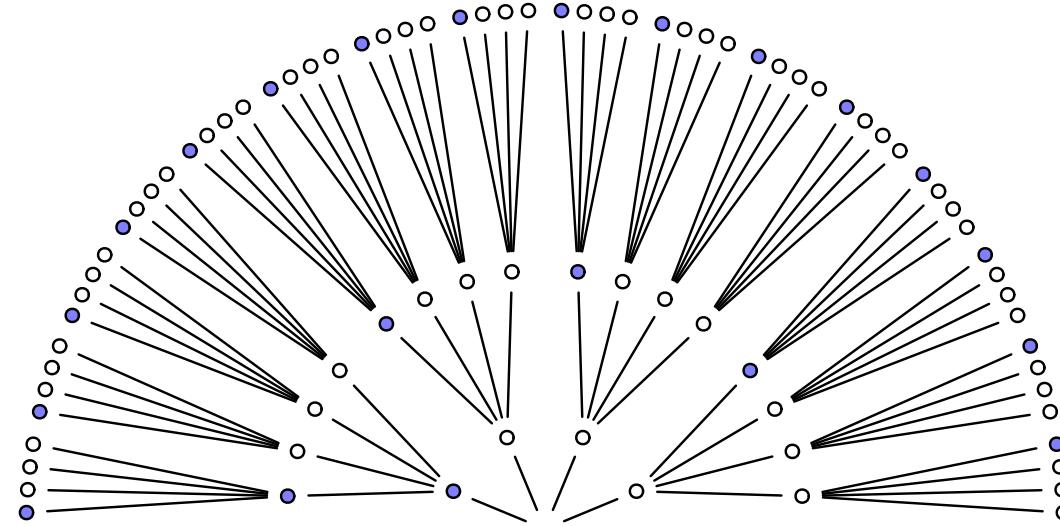
Données : ● ○



Énumérer les possibilités

Hypothèse : ● ● ○ ○

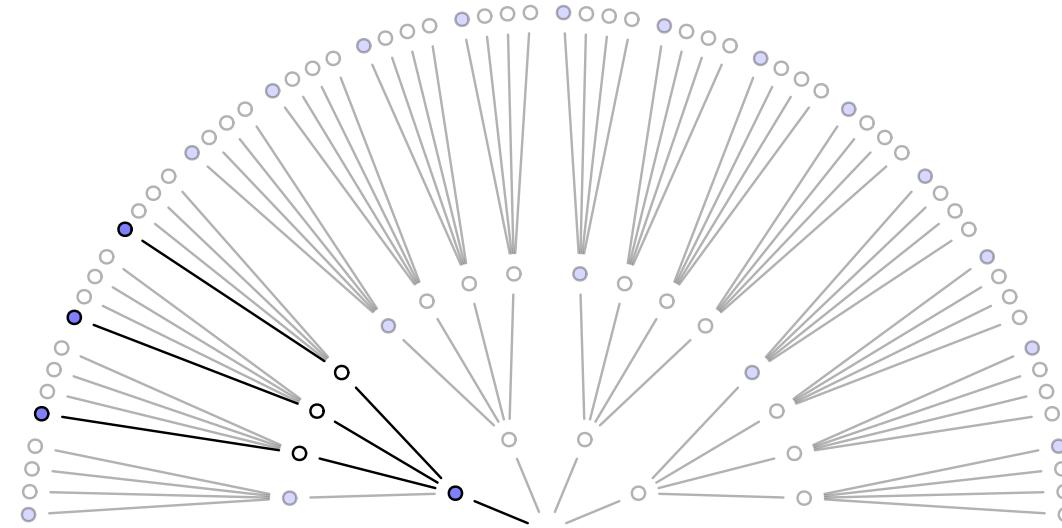
Données : ● ○ ●



Énumérer les possibilités

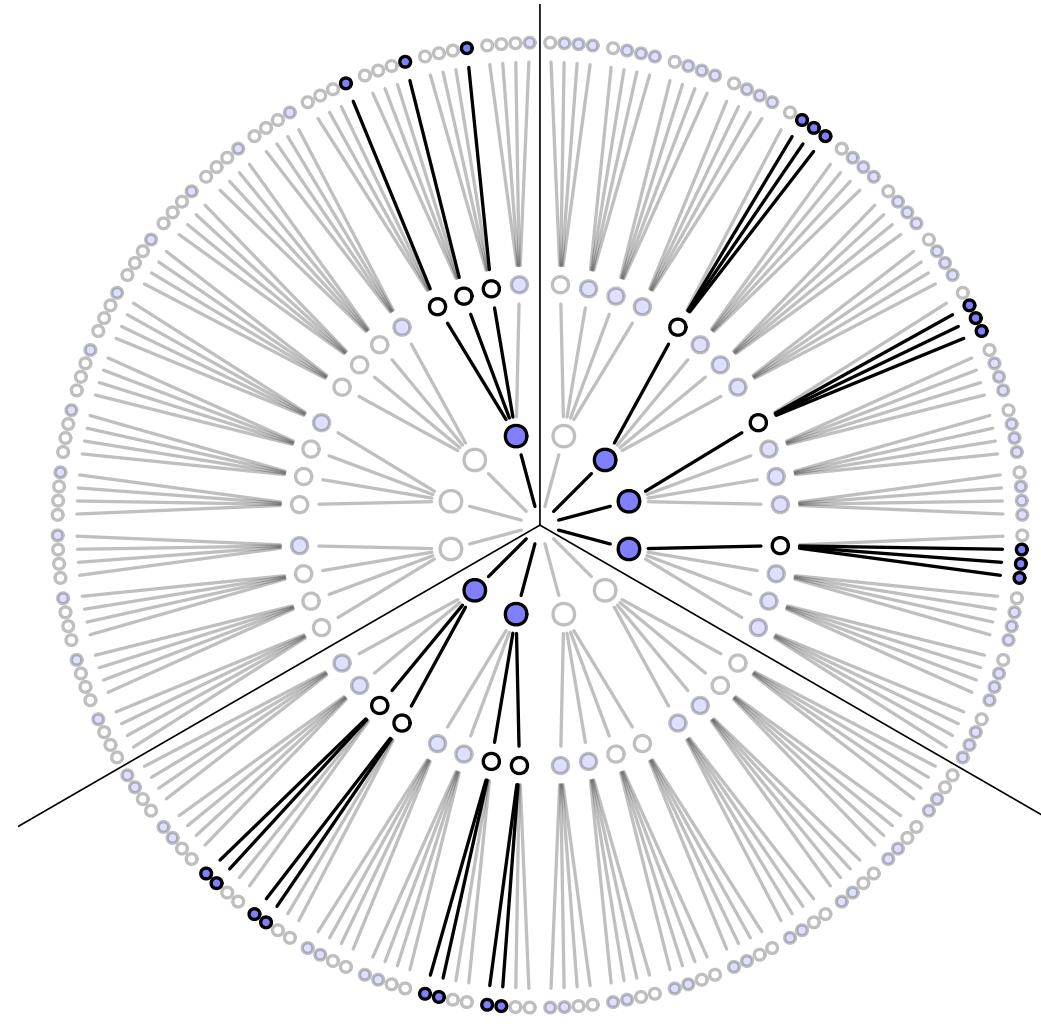
Hypothèse : ● ● ○ ○

Données : ● ○ ●



Énumérer les possibilités

Sous cette hypothèse, $\binom{3}{3}$ chemins (sur $4^3 = 64$) conduisent au résultat obtenu. Qu'en est-il des autres hypothèses ?



Comparer les hypothèses

Au vu des données, l'hypothèse la plus *probable* est celle qui **maximise le nombre de manières possibles** d'obtenir les données obtenues.

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



Accumulation d'évidence

Dans le cas précédent, 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 (des particularités des sacs de billes par exemple) ou de données antérieures.

Imaginons que nous tirions une nouvelle bille du sac, comment incorporer cette nouvelle donnée ?



Accumulation d'évidence

Il suffit d'appliquer la même stratégie que précédemment, et de mettre à jour le dernier compte en le multipliant par ces nouvelles données. *Yesterday's posterior is today's prior* ([Lindley, 2000](#)).

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	$0 \times 4 = 0$



Incorporer un prior

Supposons maintenant qu'un employé de l'usine de fabrication des billes nous dise que les billes bleues sont rares... Cet employé nous dit que pour chaque sac contenant 3 billes bleues, ils fabriquent deux sacs en contenant seulement deux, et trois sacs en contenant seulement une. Il nous apprend également que tous les sacs contiennent au moins une bille bleue et une bille blanche...

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	$0 \times 0 = 0$



Des énumérations aux probabilités

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.

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

Pour passer des *plausibilités* aux *probabilités*, il suffit de standardiser ces plausibilités pour que la somme des plausibilités de toutes les hypothèses possibles soit égale à 1.

$$\Pr(\text{hypothèse} \mid \text{données}) = \frac{\Pr(\text{données} \mid \text{hypothèse}) \times \Pr(\text{hypothèse})}{\sum \Pr(\text{hypothèse})}$$



Des énumérations aux probabilités

Définissons $\backslash(p\backslash)$ comme la proportion de billes bleues.

Hypothèse	$\backslash(p\backslash)$	Manières de produire les données	Probabilité
	0	0	0
	0.25	3	0.15
	0.5	8	0.40
	0.75	9	0.45
	1	0	0

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

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



Notations, terminologie

- θ est un paramètre ou vecteur de paramètres (e.g., la proportion de billes bleues).
- $p(x|\theta)$ 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 valide $[p(x = x | \theta)]$.
- $p(\theta)$ 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) ou “vraisemblance marginale”, “vraisemblance intégrée”.

$$\begin{aligned}
 p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\
 p(x) &= \sum_{\theta} p(x|\theta)p(\theta) \\
 &= \int_{\theta} p(x|\theta)p(\theta)d\theta
 \end{aligned}$$



Inférence bayésienne

Dans ce cadre, pour chaque problème, nous allons suivre 3 étapes :

- Construire le modèle (l'histoire des données): *likelihood + prior*
- Mettre à jour grâce aux données (*updating*), calculer la probabilité *a posteriori*
- Evaluer le modèle, *fit*, sensibilité, résumer les résultats, ré-ajuster

“

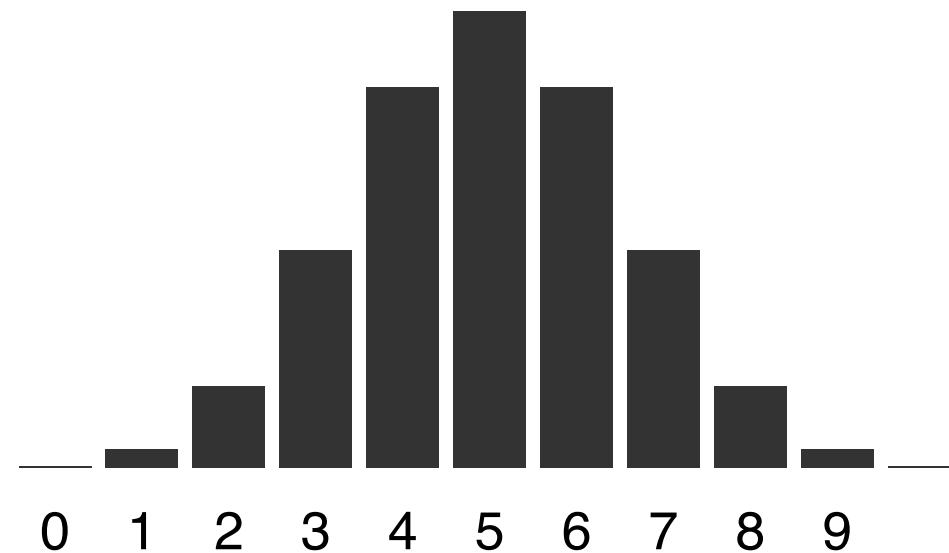
Bayesian inference is really just counting and comparing of possibilities [...] in order to make good inference about what actually happened, it helps to consider everything that could have happened (McElreath, 2015).



Rappels : Théorie des probabilités

Loi de probabilité, cas discret

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 ($\theta = 0.5$), probabilité d'obtenir N faces sur 10 lancers.



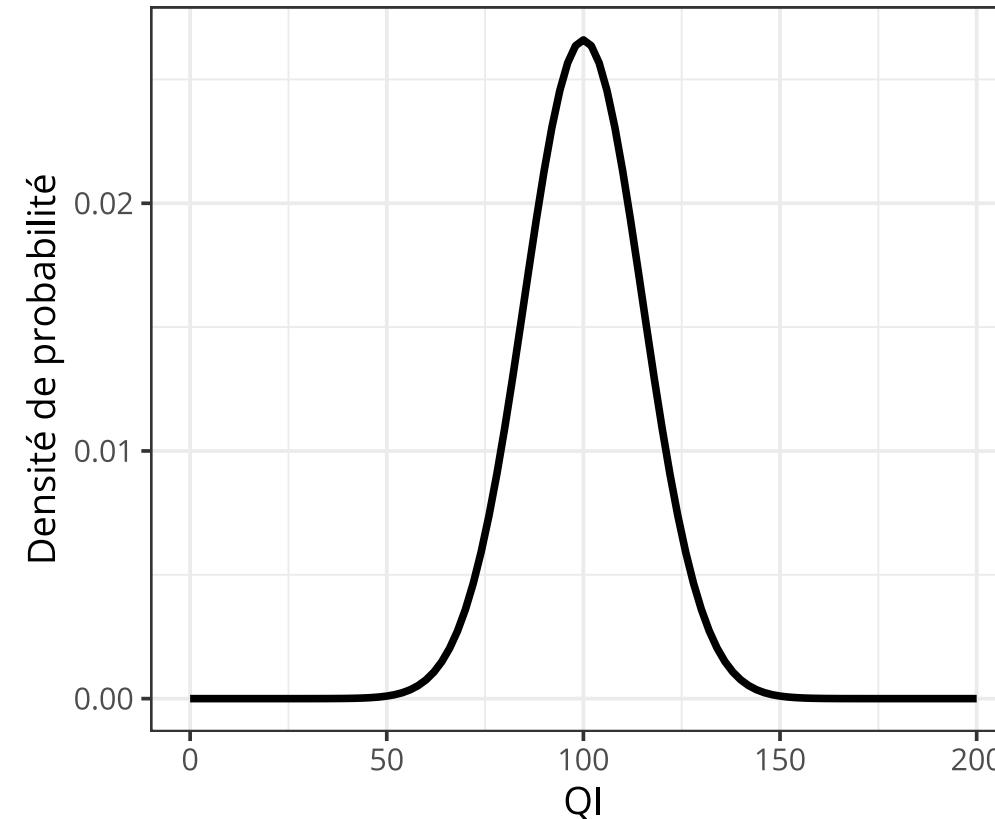
```
1 # PMFs sum to 1
2 dbinom(x = 0:10, size = 10, prob = 0.5) %>% sum
```

```
[1] 1
```



Loi de probabilité, cas continu

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).



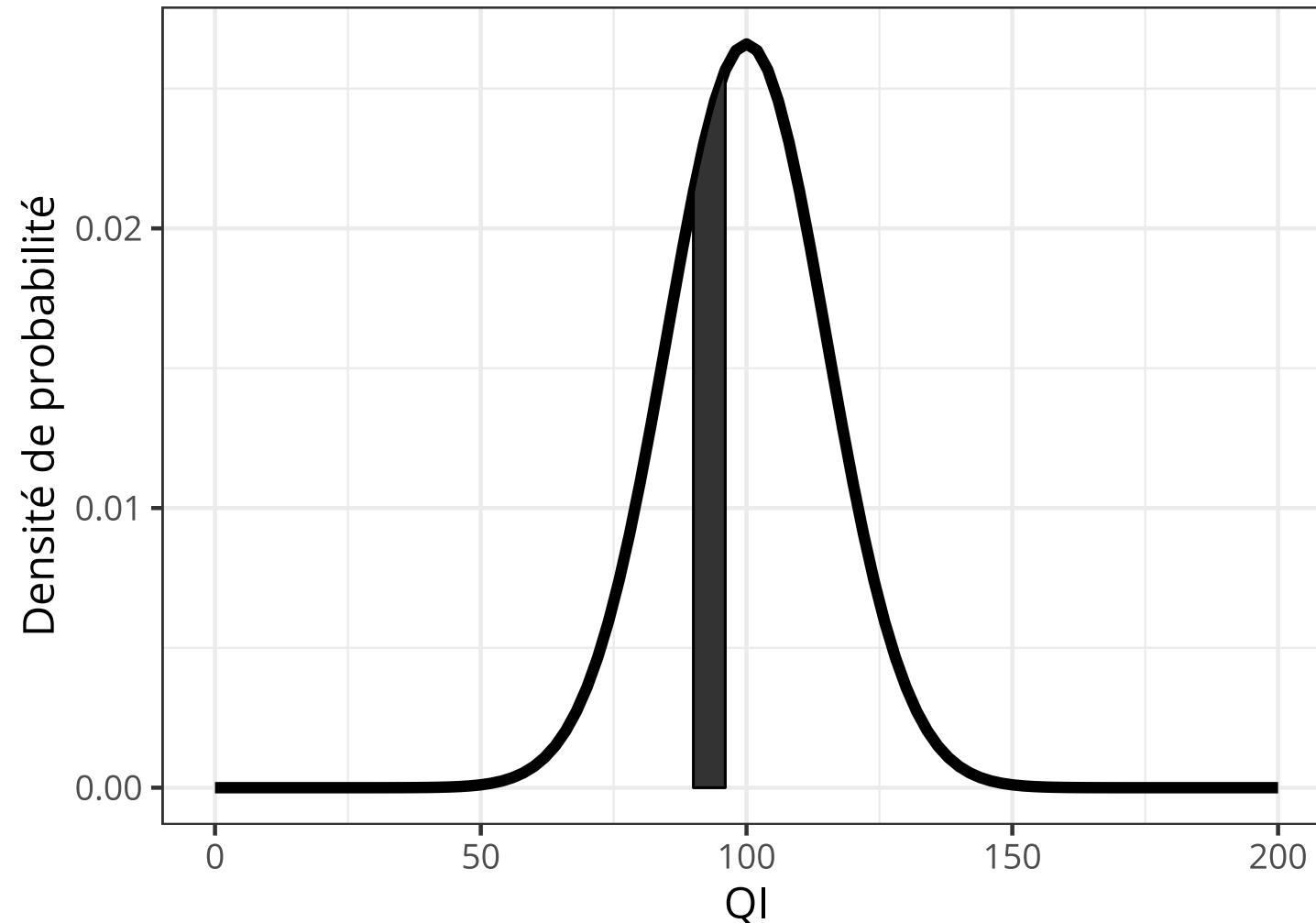
```
1 # PDFs integrate to 1
2 integrate(dnorm, -Inf, Inf, mean = 100, sd = 15)
```

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

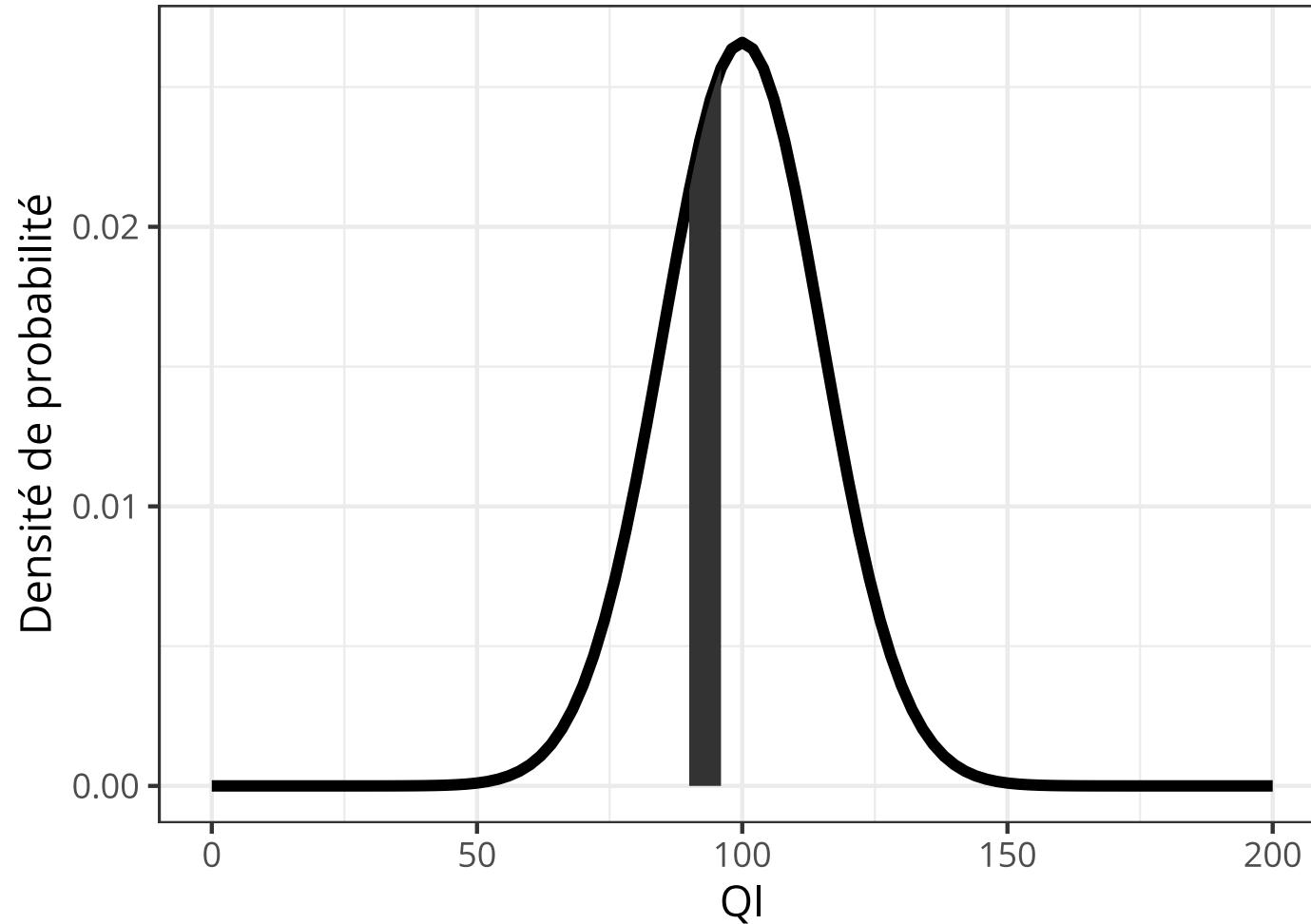


Aparté, qu'est-ce qu'une intégrale ?

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.



Aparté, qu'est-ce qu'une intégrale ?



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

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

0.1423704 with absolute error < 1.6e-15

Ladislas Nalborczyk - IMSB2022



Probabilité conjointe

```
1 library(tidyverse)
2
3 data(HairEyeColor) # data adapted from Snee (1974)
4
5 cont <- apply(HairEyeColor, c(1, 2), sum) %>% t
6 cont <- round(cont / sum(cont), 2)
7 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, on trouve la **probabilité conjointe** d'avoir telle couleur de cheveux **ET** telle couleur d'yeux, qui s'écrit $\{p(c,y)=p(y,c)\}$.



Probabilité marginale

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

	Black	Brown	Red	Blond	marginal_eye
Black	0.11	0.20	0.04	0.01	0.36
Brown	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 $(p(y)=\sum\limits_{\{c\}} p(y|c)p(c))$.



Probabilité marginale

```
1 cont3 <- rbind(cont2, colSums(cont2) )
2 rownames(cont3) <- c(row.names(cont2), "marginal_hair")
3 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 $p(c) = \sum \lim_{y} p(c|y)p(y)$.



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 $\langle p(c|y) \rangle$. Cette probabilité conditionnelle peut se ré-écrire : $\langle p(c | y) = \frac{p(c, y)}{p(y)} \rangle$.

	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 ?

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

```
Blue
0.4444444
```



Probabilité conditionnelle

On remarque dans le cas précédent que $\text{p}(\text{cheveux} = \text{blonds} \mid \text{yeux} = \text{bleus})$ **n'est pas nécessairement égal** à $\text{p}(\text{yeux} = \text{bleus} \mid \text{cheveux} = \text{blonds})$.

Twitframe

@BillGates

Loading tweet...

[Unknown date](#)

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, $\text{p}(\text{data} \mid H_0) \neq p(H_0 \mid \text{data})$.



Dérivation du théorème de Bayes

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** :

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

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

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

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

$$\{ \Pr(\text{hypothèse} \mid \text{données}) = \frac{\Pr(\text{données} \mid \text{hypothèse}) \times \Pr(\text{hypothèse})}{\Pr(\text{hypothèse}) + \Pr(\text{non-hypothèse})} \}$$



Exemple d'application

Diagnostique médical (Gigerenzer et al., 2007)

- Chez les femmes âgées de 40-50 ans, sans antécédents familiaux et sans symptômes, la probabilité d'avoir un cancer du sein est de 0.008.
- Propriétés de la mammographie :
 - Si une femme a un cancer du sein, la probabilité d'avoir un résultat positif est de 0.90.
 - Si une femme n'a pas de cancer du sein, la probabilité d'avoir un résultat positif est de 0.07.
- 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 ?



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 $\{\theta\}$, 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 $\{x\}$.



Probabilité conditionnelle

- Si une femme a un cancer du sein, la probabilité d'obtenir un résultat positif est de .90.
 - $\Pr(\text{Mam=+} | \text{Cancer=+}) = 0.90$
 - $\Pr(\text{Mam=-} | \text{Cancer=+}) = 0.10$
- Si une femme n'a pas de cancer du sein, la probabilité d'obtenir un résultat positif est de .07.
 - $\Pr(\text{Mam=+} | \text{Cancer=-}) = 0.07$
 - $\Pr(\text{Mam=-} | \text{Cancer=-}) = 0.93$



Diagnostique médical, maximum likelihood

- Une femme passe une mammographie, le résultat est positif...
 - $\Pr(\text{Mam}=+ | \text{Cancer}=+) = 0.90$
 - $\Pr(\text{Mam}=+ | \text{Cancer}=-) = 0.07$
- Maximum de vraisemblance : quelle est la valeur de *Cancer* qui **maximise** $\Pr(\text{Mam}=+)$?
 - $\Pr(\text{Mam}=+ | \text{Cancer}=+) = 0.90$
 - $\Pr(\text{Mam}=+ | \text{Cancer}=-) = 0.07$



Wait a minute...

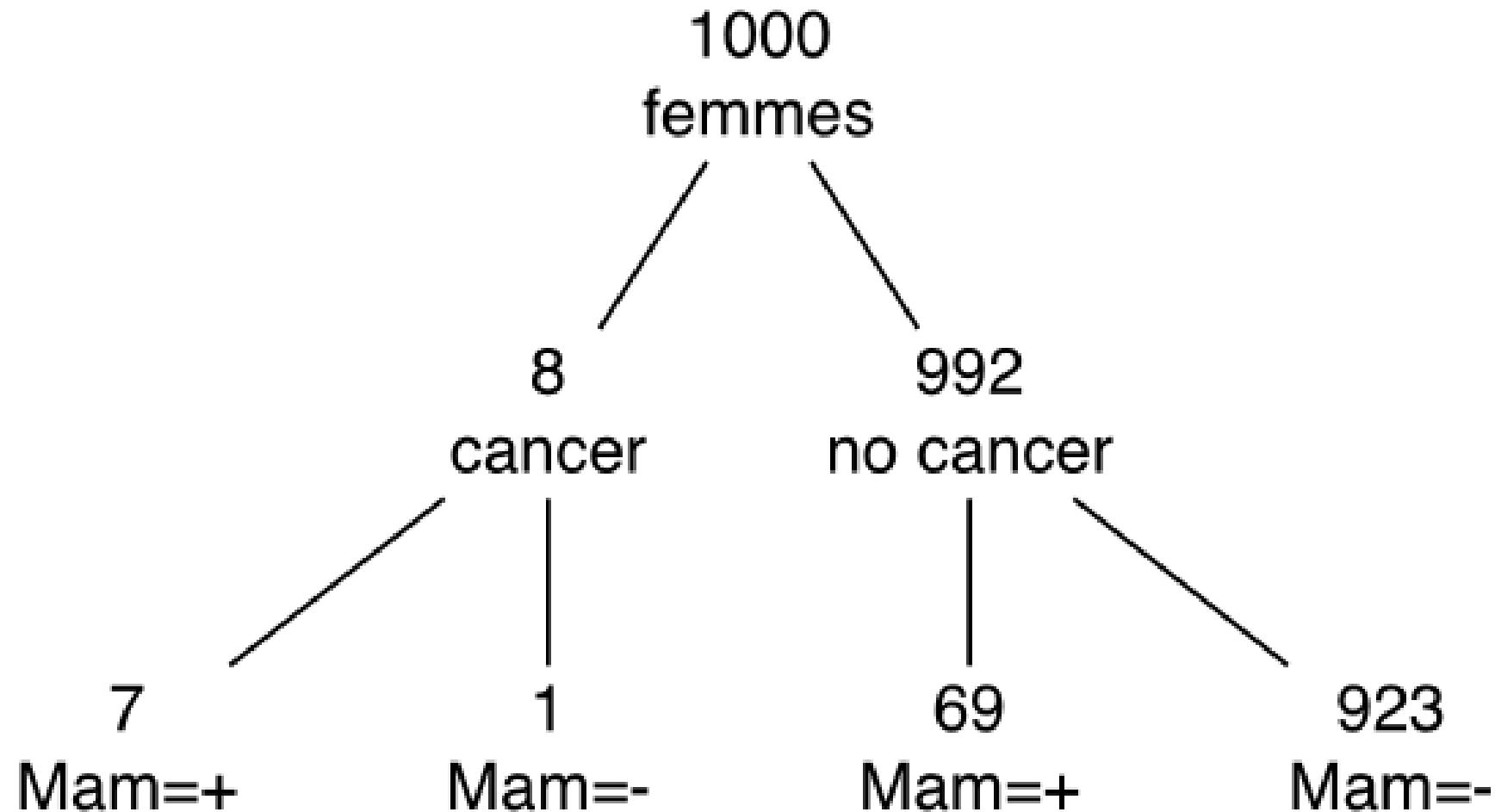


Diagnostique médical, fréquences naturelles

- 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 ?



Diagnostique médical, fréquences naturelles



$$\Pr(\text{Cancer}=+ | \text{Mam}=+) = \frac{7}{7 + 69} = \frac{7}{76} \approx 0.09$$



Diagnostique médical, théorème de Bayes

$$[\color{purple}{p(\theta | x)} = \frac{\color{orange}{p(x|\theta)}}{\color{steelblue}{p(\theta)}} \color{green}{p(x)}]$$

$\color{steelblue}{p(\theta)}$ représente la probabilité *a priori* de θ : tout ce qu'on sait de θ avant d'observer les données. En l'occurrence : $\Pr(\text{Cancer}+) = 0.008$ et $\Pr(\text{Cancer}-) = 0.992$.

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



Diagnostique médical, théorème de Bayes

$$[\color{purple}{p(\theta | x)} = \frac{\color{orange}{p(x|\theta)}}{\color{steelblue}{p(\theta)}} \color{green}{p(x)}]$$

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

```
1 like <- rbind(c(0.9, 0.1), c(0.07, 0.93) ) %>% data.frame
2 colnames(like) <- c("Mam+", "Mam-")
3 rownames(like) <- c("Cancer+", "Cancer-")
4 like
```

	Mam+	Mam-
Cancer+	0.90	0.10
Cancer-	0.07	0.93



Diagnostique médical, théorème de Bayes

$$\color{purple} p(\theta | x) = \frac{\color{orange} p(x|\theta)}{\color{steelblue} p(\theta)} \color{green} p(x)$$

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

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

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

```
[1] 0.07664
```



Diagnostique médical, théorème de Bayes

$$\color{purple} p(\theta | x) = \frac{\color{orange} p(x|\theta) \color{steelblue} p(\theta)}{\color{green} p(x)}$$

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

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

```
[1] 0.09394572 0.90605428
```



L'inférence bayésienne comme mise à jour probabiliste des connaissances

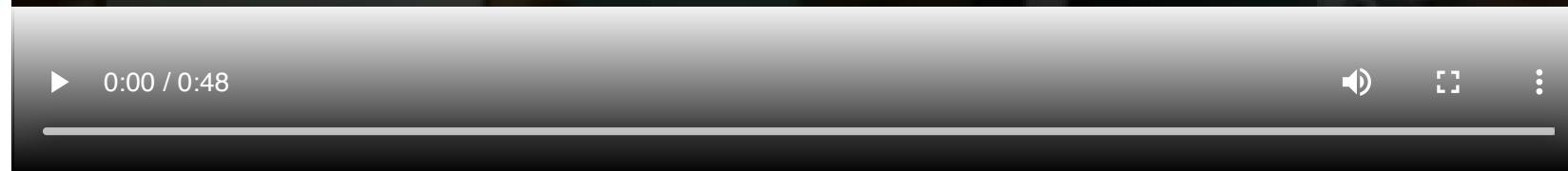
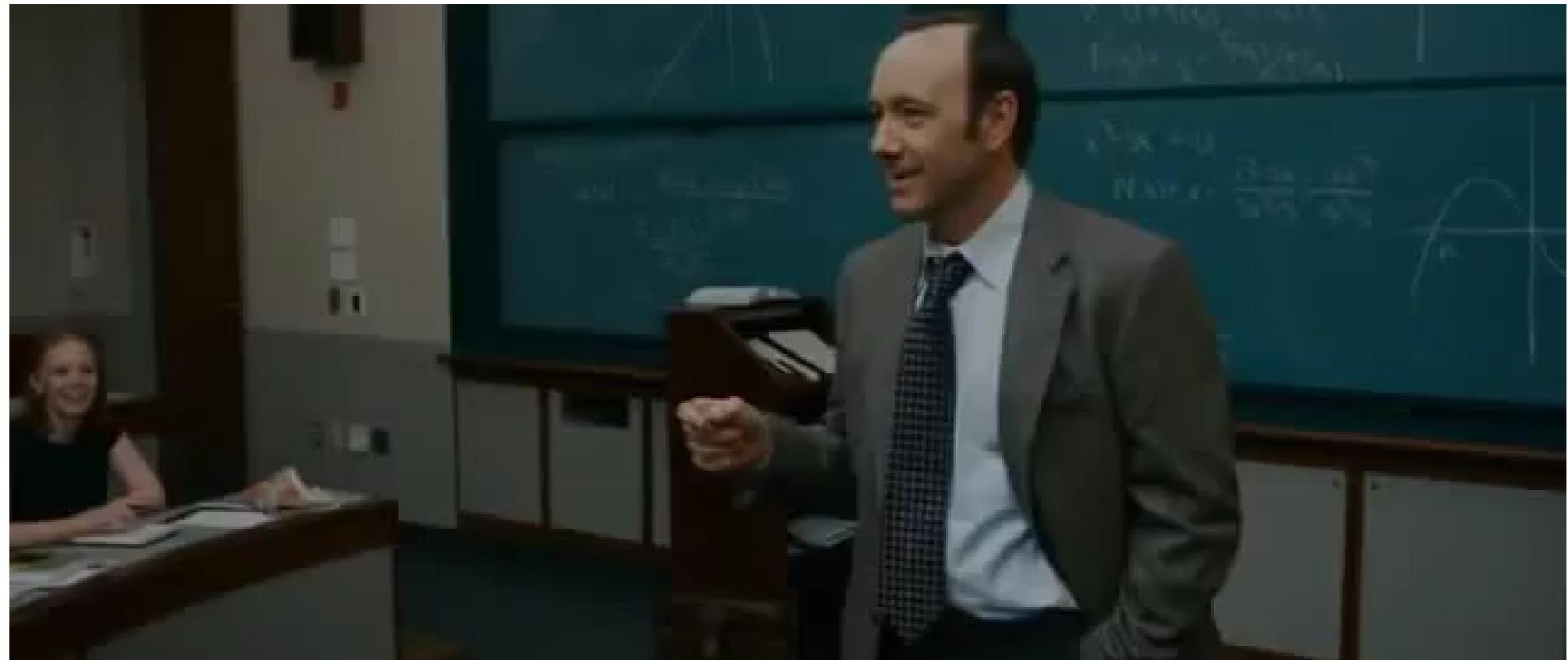
Avant de passer le mammogramme, la probabilité qu'une femme tirée au sort ait un cancer du sein était de $\Pr(\text{Cancer}=+) = 0.008$ (prior). Après un résultat positif, cette probabilité est devenue $\Pr(\text{Cancer}=+ | \text{Mam}=+) = 0.09$ (posterior). Ces probabilités sont des expressions de 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".

“

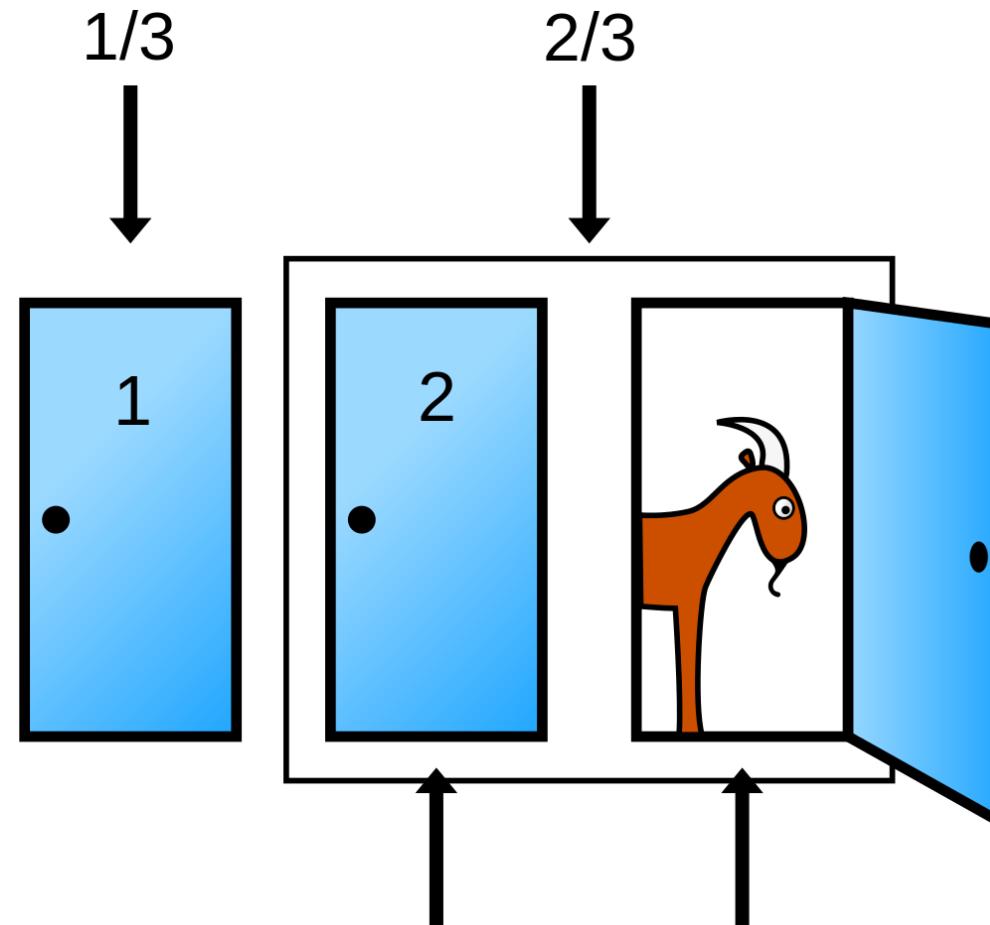
A Bayesianly justifiable analysis is one that treats known values as observed values of random variables, treats unknown values as unobserved random variables, and calculates the conditional distribution of unknowns given knowns and model specifications using Bayes' theorem ([Rubin, 1984](#)).



Monty Hall



Monty Hall



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



Monty Hall

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 :

$$\Pr(\text{P3} | \text{V1}) = \frac{1}{2}, \quad \Pr(\text{P3} | \text{V2}) = 1, \quad \Pr(\text{P3} | \text{V3}) = 0.$$



Monty Hall

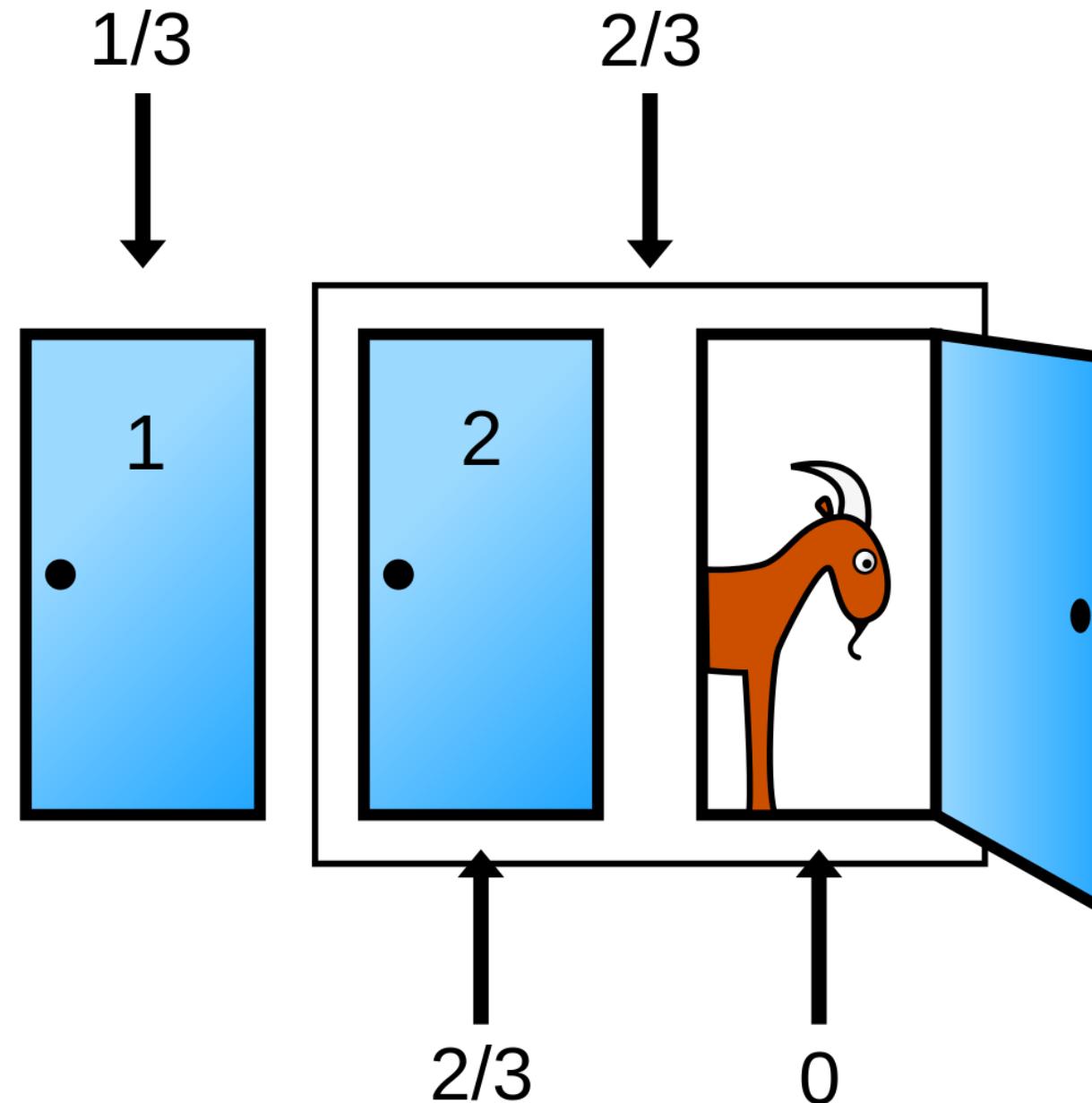
On sait que $\Pr(\text{V3} | \text{P3}) = 0$, on veut connaître $\Pr(\text{V1} | \text{P3})$ et $\Pr(\text{V2} | \text{P3})$ afin de pouvoir choisir. Résolution par le théorème de Bayes.

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

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



Monty Hall



Take-home message

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 (e.g., modus ponens et modus tollens) et probabilistes simples (e.g., 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, 2020](#)).



Références

- Carnap, R. (1971). *Logical foundations of probability* (4. impr.). Univ. of Chicago Press [u.a].
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest*, 8(2), 53–96.
<https://doi.org/10.1111/j.1539-6053.2008.00033.x>
- Jaynes, E. T. (1986). *Bayesian methods: General background*.
- 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.
- McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan* (2nd ed.). Taylor; Francis, CRC Press.
- 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>
- Rouder, Jeffrey N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016a). Is There a Free Lunch in Inference? *Topics in Cognitive Science*, 8(3), 520–547. <https://doi.org/10.1111/tops.12214>
- Rouder, Jeffrey N., Morey, R. D., & Wagenmakers, E.-J. (2016b). The Interplay between Subjectivity, Statistical Practice, and Psychological Science. *Collabra*, 2(1), 6. <https://doi.org/10.1525/collabra.28>
- Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4). <https://doi.org/10.1214/aos/1176346785>

