Introduction à la statistique bayésienne

avec le logiciel R

Olivier Gimenez

éditions Quæ

Introduction à la statistique bayésienne

avec le logiciel R

Olivier Gimenez

éditions Quæ

Pour citer cet ouvrage :
Gimenez O., 2026. *Introduction à la statistique bayésienne avec le logiciel R*.
Versailles, éditions Quæ, 78 p.
https://doi.org/10.35690/978-2-7592-4258-0

Photo de couverture : © Yann Raulet.

Les éditions Quae réalisent une évaluation scientifique des manuscrits avant publication
dont la procédure est décrite ici :
https://www.quae.com/store/page/199/processus-d-evaluation

Le processus éditorial s'appuie également sur un logiciel de détection des similitudes
et des textes potentiellement générés par intelligence artificielle.

Cet ouvrage a bénéficié d'un financement du CNRS et de l'université de Montpellier.

Éditions Quæ
RD 10
78026 Versailles Cedex, France
www.quae.com
www.quae-open.com

ISBN papier : 978-2-7592-4257-3
ISBN PDF : 978-2-7592-4258-0
ISBN ePub : 978-2-7592-4259-7

SOMMAIRE

AVANT-PROPOS

On retrouve la statistique bayésienne un peu partout en sciences. Par exemple, en épidémiologie, pour prédire la circulation des virus, en écologie, pour expliquer l'extinction des espèces végétales et animales, ou encore en informatique, pour filtrer les courriels nuisibles. Si l'utilisation de la statistique bayésienne a explosé au cours des dernières années, c'est grâce au progrès de nos ordinateurs. C'est aussi grâce à la nature même de l'approche qui permet de coller à notre façon d'apprendre, de raisonner et d'accumuler des connaissances.

Dans ce livre, je vous propose une introduction à la statistique bayésienne. Ce livre est en français parce que c'est la langue dans laquelle je me sens le plus à l'aise pour écrire, et parce que j'aurais aimé avoir plus d'ouvrages dans ma langue maternelle lorsque j'étais étudiant.

Je me suis fixé comme objectifs de 1) synthétiser les aspects méthodologiques à bien comprendre, et 2) fournir les moyens pratiques pour utiliser vous-mêmes la statistique bayésienne. Parce que l'on comprend mieux en faisant, nous utiliserons un logiciel pour pratiquer la statistique. Ce logiciel, c'est R, c'est un logiciel libre pour faire des statistiques et de la science des données. En français, je recommande l'excellent manuel de Julien Barnier, *Introduction à R et au tidyverse*, disponible en ligne[1] et le site du projet collaboratif Analyse-R[2]. Pour la statistique bayésienne en particulier, je présente un package qui propose une syntaxe simple et familière, proche de celle utilisée pour les régressions dans R : brms. Dans la version enrichie de ce livre consultable en ligne[3], je présente aussi un package qui nécessite de programmer (écrire des boucles par exemple), mais qui offre en contrepartie une grande flexibilité : NIMBLE.

Plutôt que dans un style académique, j'ai choisi d'écrire un peu comme si nous étions ensemble dans la même pièce ou en visioconférence, et que je devais vous expliquer de vive voix la statistique bayésienne. Ainsi, je ferai parfois (souvent, en fait) des abus de langage et des approximations mathématiques pour faciliter la compréhension. Vous ne m'en voudrez pas, j'espère.

POURQUOI S'INTÉRESSER À LA STATISTIQUE BAYÉSIENNE ?

La statistique bayésienne est une approche pour analyser les données et prendre des décisions en présence d'incertitude, comme lorsqu'on lance un dé ou que l'on prévoit la météo : on ne peut pas savoir exactement ce qui va se passer, mais on peut estimer les chances des différents résultats. Pourquoi adopter cette approche ? Plusieurs raisons peuvent motiver son utilisation :

– une interprétation naturelle des probabilités : en statistique bayésienne, une probabilité représente un degré de confiance dans une hypothèse ou dans un paramètre, ce qui correspond bien à notre manière intuitive de raisonner face à l'incertitude ;

– une grande flexibilité : le cadre bayésien s'adapte bien à des données incomplètes, hétérogènes ou rares, ainsi qu'à des modèles complexes (hiérarchiques, non linéaires, dynamiques, etc.) ;

– la possibilité d'intégrer des connaissances préalables : on peut capitaliser sur des résultats d'études précédentes ou des avis d'expertises de manière transparente et formalisée ;

– une gestion rigoureuse de l'incertitude : la statistique bayésienne fournit non seulement une estimation des paramètres, mais aussi une mesure directe de l'incertitude associée.

1. https://juba.github.io/tidyverse
2. https://larmarange.github.io/analyse-R/
3. https://oliviergimenez.github.io/statistique-bayes/

CE QUE NOUS ALLONS VOIR DANS CE LIVRE

J'aimerais vous guider dans l'apprentissage de la statistique bayésienne. J'ai rassemblé le matériel qui m'a paru essentiel pour la comprendre et l'appliquer. L'objectif est que vous soyez à l'aise avec l'approche bayésienne et que vous puissiez l'appliquer à vos propres données. Les objectifs sont de :

– démythifier la statistique bayésienne et les méthodes de Monte Carlo par chaînes de Markov (MCMC) ;

– comprendre les différences entre approche bayésienne et approche fréquentiste ;

– lire et comprendre les sections « méthodes » des articles scientifiques utilisant l'approche bayésienne ;

– savoir mettre en œuvre vos analyses avec la statistique bayésienne dans R.

Le chapitre 1 pose les bases en revenant sur quelques rappels de probabilité utiles pour la suite. Ce sera aussi l'occasion d'introduire les notions clés de la statistique bayésienne, à travers un exemple simple qui permettra de fixer les idées.

Dans le chapitre 2, nous passerons dans les coulisses de la statistique bayésienne, avec les méthodes MCMC, qui rendent l'inférence possible en pratique. Nous mettrons un peu la main à la pâte en codant nous-mêmes une analyse bayésienne.

Le chapitre 3 présentera brms, un outil très utile pour faire de la statistique bayésienne sans trop d'efforts. Dans la version enrichie du livre[3], je présenterai aussi NIMBLE. Grâce à ces outils, nul besoin de tout faire soi-même.

Le chapitre 4 sera consacré aux distributions a priori. Nous verrons comment les choisir avec pertinence, comment traduire de l'information existante sous forme de prior, et les pièges à éviter.

Dans le chapitre 5, nous verrons comment faire une régression linéaire en statistique bayésienne. Nous en profiterons pour illustrer la comparaison et la validation des modèles. Nous utiliserons brms (et NIMBLE) et le comparerons à l'approche fréquentiste.

Le chapitre 6 nous emmènera vers les modèles linéaires généralisés, avec ou sans effets aléatoires – des modèles très utilisés en pratique. Nous nous appuierons sur la simulation de données, un outil précieux pour bien comprendre ce que fait un modèle. Je vous montrerai comment faire ces analyses avec brms (et avec NIMBLE) et nous les comparerons aux analyses fréquentistes.

Enfin, un dernier chapitre viendra résumer les messages clés du livre et proposer quelques conseils pour appliquer la statistique bayésienne.

COMMENT LIRE CE LIVRE ?

Je n'ai pas vraiment de conseil à vous donner sur la meilleure manière de lire ce livre. Personnellement, je trouve toujours difficile d'absorber toute l'information contenue dans un ouvrage. Vous pouvez lire en continu ou bien grappiller des éléments de-ci de-là.

Dans chacun des chapitres, le code R est fourni, je l'ai aussi hébergé sur mon site[4] et le mettrai à jour. S'exercer permet de mieux comprendre et de vérifier que l'on a bien assimilé. Si vous lisez la version électronique de cet ouvrage, vous pouvez copier les lignes de code puis les coller dans R pour les exécuter. Pour gagner un peu de place, et éviter de perturber trop la lecture, certains codes ne sont pas donnés, en particulier ceux qui permettent de produire les figures, mais ils sont disponibles en ligne[4].

Si vous voulez aller plus loin, vous trouverez une liste bibliographique enrichie à la fin de ce livre. Les ouvrages Biobayes collectif (2015), McCarthy (2007), Kéry et Kellner (2010), Johnson *et al.* (2022), Kruschke (2010), Gelman *et al.* (2013) et McElreath (2020) ont été une source d'inspiration dans la rédaction de ce livre. J'ai hésité à donner plus de références et à citer (beaucoup) d'articles scientifiques, mais je ne le ferai pas. Les ouvrages cités sont largement suffisants.

4. https://github.com/oliviergimenez/statistique-bayes-quae

1. L'approche bayésienne

Dans ce chapitre, nous posons les bases en faisant quelques rappels de probabilité utiles pour la suite. J'introduis les notions clés de la statistique bayésienne à travers un exemple simple qui permet de fixer les idées, et que nous utiliserons souvent dans le livre. Nous ferons aussi des parallèles entre la statistique classique ou fréquentiste et la statistique bayésienne.

1.1. LE THÉORÈME DE BAYES

Ne tardons plus et entrons dans le vif du sujet. La statistique bayésienne repose sur le théorème de Bayes (ou formule de Bayes), dont on doit la première formulation au mathématicien et révérend Thomas Bayes. Ce théorème a été publié en 1763, deux ans après la mort de Bayes, grâce aux efforts de son ami Richard Price. Il a par ailleurs été découvert indépendamment par Pierre-Simon Laplace.

Le théorème de Bayes porte sur les probabilités conditionnelles, qui sont parfois un peu délicates à comprendre. La probabilité conditionnelle d'un événement A sachant un événement B, que l'on note Pr($A \mid B$), est la probabilité que A se produise, révisée en tenant compte de l'information supplémentaire que l'événement B s'est produit. Par exemple, imaginez qu'un de vos amis lance un dé (équilibré) et vous demande la probabilité que le résultat soit un six (A). Votre réponse est 1/6, car chaque face du dé a la même chance d'apparaître. Maintenant, imaginez que l'on vous dise que le nombre obtenu est pair (B) avant de répondre. Comme il n'y a que trois nombres pairs, dont un seul est six, vous pouvez réviser votre réponse : Pr($A \mid B$) = 1/3.

Vous voyez comment l'information supplémentaire (ici, savoir que le nombre est pair) change l'estimation ? C'est exactement ce type de raisonnement que le théorème de Bayes formalise et généralise : il permet de calculer la probabilité d'un événement A sachant qu'un autre événement B s'est produit. Plus précisément, le théorème de Bayes vous donne Pr($A \mid B$) en utilisant les probabilités marginales Pr(A) et Pr(B) et la probabilité Pr($B \mid A$) :

$$\Pr\left(A \mid B\right) = \frac{\Pr\left(B \mid A\right)\Pr\left(A\right)}{\Pr\left(B\right)}$$

On parle de probabilité marginale quand on s'intéresse à la probabilité d'un événement « tout seul », sans condition particulière. Par exemple, Pr(A) ou Pr(B), ce sont les chances globales de A ou de B, sans tenir compte d'autre chose. On dit « marginale » parce que, si vous faisiez un tableau avec toutes les combinaisons possibles (par exemple les résultats d'un dé classés en pair/impair et en six/pas six), Pr(A) ou Pr(B) s'obtiendraient alors en additionnant les cases d'une ligne ou d'une colonne, autrement dit ce qu'on lit en marge du tableau.

Le théorème de Bayes est souvent considéré comme un moyen de remonter d'un effet B à une cause A inconnue, en connaissant la probabilité de l'effet B sachant la cause A. Pensez, par exemple, à une situation où un diagnostic médical est requis, avec A, une maladie inconnue et B, des symptômes ; la médecin connaît les risques d'avoir certains symptômes en fonction de plusieurs maladies ou Pr(symptômes | maladie), et souhaite déduire la probabilité d'avoir une maladie connaissant les symptômes ou Pr(maladie | symptômes). Cette manière de renverser Pr($B \mid A$) en Pr($A \mid B$) explique pourquoi le raisonnement bayésien est parfois appelé « probabilité inverse ».

Plutôt que d'utiliser des lettres au risque de s'embrouiller, je trouve plus facile de retenir le théorème de Bayes écrit ainsi :

$$\Pr\left(\text{hypothèse} \mid \text{données}\right) = \frac{\Pr\left(\text{données} \mid \text{hypothèse}\right)\Pr\left(\text{hypothèse}\right)}{\Pr\left(\text{données}\right)}$$

L'hypothèse peut être un paramètre, comme la probabilité d'apparition d'une maladie, ou des coefficients de régression liant cette probabilité à des facteurs de risque (par exemple le lieu de vie, le tabagisme). Le théorème de Bayes nous dit comment obtenir la probabilité d'une hypothèse à partir des données disponibles.

C'est pertinent car, réfléchissez-y, c'est exactement ce que fait la méthode scientifique. Nous voulons savoir à quel point une hypothèse est plausible à partir de données que nous avons collectées, et peut-être comparer plusieurs hypothèses entre elles. De ce point de vue, le raisonnement bayésien s'accorde avec le raisonnement scientifique, ce qui explique probablement pourquoi le cadre bayésien est si naturel pour faire et comprendre la statistique.

Vous pourriez alors demander pourquoi la statistique bayésienne n'est-elle pas la norme ? Pendant longtemps, la mise en œuvre du théorème de Bayes a été limitée par des difficultés de calcul, comme nous le verrons au chapitre suivant. Fort heureusement, l'amélioration de la puissance de calcul de nos ordinateurs et le développement de nouveaux algorithmes ont entraîné un essor marqué de l'approche bayésienne au cours des trente dernières années.

1.2. QU'EST-CE QUE LA STATISTIQUE BAYÉSIENNE ?

Les problèmes statistiques typiques consistent à estimer un paramètre (ou plusieurs) à partir de données disponibles. On va noter ce ou ces paramètres de manière générique, disons θ. Pour estimer θ, vous êtes sûrement plus familier de l'approche fréquentiste que de l'approche bayésienne. L'approche fréquentiste, en particulier l'estimation par maximum de vraisemblance, suppose que les paramètres sont fixes et de valeurs inconnues. Les estimateurs classiques sont donc en général des valeurs ponctuelles ; par exemple, un estimateur de la probabilité d'obtenir une face d'un dé est le nombre de fois qu'on a obtenu cette face, divisé par le nombre de fois qu'on a lancé ce dé. L'approche bayésienne suppose que les paramètres ne sont pas fixes et suivent une distribution inconnue. Une distribution de probabilité est une expression mathématique qui donne la probabilité qu'une variable aléatoire prenne certaines valeurs. Elle peut être discrète (par exemple la loi de Bernoulli, la loi binomiale ou la loi de Poisson) ou continue (comme la loi normale ou gaussienne).

L'approche bayésienne repose sur l'idée que vous commencez avec certaines connaissances sur le système avant même de l'étudier vous-même. Ensuite, vous collectez des données et mettez à jour ces connaissances a priori en fonction des observations. Ce processus de mise à jour repose sur le théorème de Bayes. De façon simplifiée, si l'on prend $A = \theta$ et $B =$ données, alors le théorème de Bayes permet d'estimer le paramètre θ à partir des données comme suit :

$$\Pr\left(\theta \mid \text{données}\right) = \frac{\Pr\left(\text{données} \mid \theta\right) \times \Pr\left(\theta\right)}{\Pr\left(\text{données}\right)}$$

Prenons un peu de temps pour passer en revue chaque terme de cette formule.

À gauche, nous avons Pr(θ | données), la distribution a posteriori. La probabilité de θ sachant les données. Elle représente ce que vous savez de θ après avoir vu les données. C'est la base de l'inférence et c'est précisément ce que vous cherchez : une distribution, possiblement multivariée si vous avez plusieurs paramètres.

À droite, on trouve Pr(données | θ), la vraisemblance (ou *likelihood* en anglais). La probabilité des données sachant θ. Cette quantité est la même que dans l'approche classique ou fréquentiste. Car oui, les approches bayésienne et fréquentiste partagent la même composante, la vraisemblance, ce qui explique pourquoi leurs résultats sont souvent proches. La vraisemblance exprime l'information que contiennent vos données, étant donné un modèle paramétré par θ. On en reparle à la section 1.4.

Ensuite, nous avons Pr(θ) la distribution a priori. Cette quantité représente ce que vous savez sur θ avant de voir les données. Cette distribution a priori ne doit pas dépendre des données, autrement dit, on ne devrait pas utiliser les données pour la construire. Elle peut être vague ou non informative si vous ne savez rien sur θ. Souvent, vous ne partez jamais de zéro, et dans l'idéal vous souhaiteriez que votre a priori reflète les connaissances existantes. Je vous parlerai plus en détail des priors au chapitre 4.

Enfin, il y a le dénominateur Pr(données), parfois appelé vraisemblance moyenne (*average likelihood*), moyenne par rapport au prior, car il s'obtient en intégrant la vraisemblance selon la loi a priori Pr(données) = $\int$ Pr(données | θ) × Pr(θ) $d\theta$. Cette quantité permet de normaliser la distribution a posteriori pour qu'elle intègre à 1. Autrement dit, comme $\int$ Pr(θ | données) $d\theta$ = 1 puisque l'intégrale d'une densité de probabilité vaut 1, on a $\int \frac{\Pr(\text{données} \mid \theta) \times \Pr(\theta)}{\Pr(\text{données})} d\theta = 1$.

Et puisque Pr(données) ne dépend pas de θ, on a Pr(données) = $\int$ Pr(données | θ) × Pr(θ) $d\theta$. C'est une intégrale de dimension égale au nombre de paramètres θ à estimer : pour deux paramètres, une intégrale double, pour trois paramètres, une intégrale triple, etc. Or, au-delà de trois dimensions, il devient difficile, voire impossible, de calculer cette intégrale. C'est l'une des raisons qui expliquent que l'approche bayésienne n'a pas été utilisée plus tôt et que l'on a besoin d'algorithmes pour estimer les distributions a posteriori, comme je l'explique dans le chapitre 2. En attendant, on va dérouler un exemple relativement simple dans lequel la distribution a posteriori a une forme explicite.

1.3. UN EXEMPLE FIL ROUGE

Prenons un exemple concret pour fixer les idées. Je travaille sur le ragondin (*Myocastor coypus*) (**figure 1.1**), un rongeur semi-aquatique originaire d'Amérique du Sud, introduit en Europe pour l'élevage de fourrure. Il est aujourd'hui considéré comme une espèce exotique envahissante, en raison des dégâts qu'il occasionne dans les milieux humides (érosion des berges, destruction de la végétation) et de son rôle possible dans la transmission à l'humain de la leptospirose, une infection bactérienne potentiellement sévère, transmise *via* l'eau. Grâce à sa forte fécondité et à sa bonne adaptation aux climats tempérés, le ragondin a proliféré rapidement.

Une des questions qui m'intéressent est d'estimer la probabilité qu'ils survivent à l'hiver, les ragondins étant particulièrement sensibles au froid. Pour cela, on équipe plusieurs individus d'une balise GPS au début de l'hiver, disons ici n = 57. À la fin de l'hiver, on observe que y = 19 ragondins sont encore vivants. L'objectif est d'estimer la probabilité de survie hivernale que l'on notera θ. Voici les données :

```
y <- 19 # nombre d'individus ayant survécu à l'hiver
n <- 57 # nombre d'individus suivis au début de l'hiver
```

Vous vous dites sûrement qu'avec ces éléments on peut déjà estimer une probabilité de survie. Intuitivement, on pense à la proportion d'individus ayant survécu, soit 19/57. Et vous n'avez pas tort. C'est une estimation raisonnable de θ, la probabilité de survie hivernale. Essayons maintenant de formaliser cette intuition, afin de mieux comprendre ce qu'elle représente, et ce qu'elle suppose.

Comme mentionné plus haut, la vraisemblance est une idée centrale que l'on retrouve dans les approches fréquentiste et bayésienne. Commençons donc par construire cette vraisemblance. Pour cela, il nous faut poser quelques hypothèses.

Premièrement, nous supposons que les individus sont indépendants, c'est-à-dire que la survie d'un ragondin n'influence pas la survie des autres ragondins. C'est une hypothèse forte, surtout quand on sait qu'une femelle peut se reproduire deux à trois fois par an et donner naissance jusqu'à dix petits dépendants d'elle au début de leur vie. Mais, en modélisation, il vaut souvent mieux commencer simplement.

Figure 1.1. Cliché de ragondins (*Myocastor coypus*) pris sur le bassin-versant du Lez dans les environs de Montpellier. Crédits : Yann Raulet.

Deuxièmement, nous supposons que tous les individus ont la même probabilité de survie. Là encore, c'est une simplification : on sait, par exemple, que la mortalité des jeunes est plus élevée que celle des adultes.

Sous ces deux hypothèses, le nombre y d'animaux encore vivants à la fin de l'hiver suit une loi binomiale, avec θ comme probabilité de succès (survie), et n comme nombre d'essais (individus suivis). On notera $y \sim \text{Bin}(n,\theta)$. La loi binomiale est en fait la somme de plusieurs épreuves de Bernoulli indépendantes, comme dans l'exemple classique du « pile ou face ». À chaque lancé, ici le relâché d'un ragondin équipé d'un GPS en début d'hiver, on suppose une probabilité θ de succès, c'est-à-dire de survivre à l'hiver, et d'échec, c'est-à-dire de mourir de froid. Si toutes ces épreuves sont indépendantes et ont la même probabilité de succès (nos hypothèses), alors le nombre de succès ou de ragondins vivants en sortie d'hiver suit une loi binomiale (voir aussi le chapitre 6). Je donne des exemples de tirages Bernoulli et binomiaux dans la **figure 1.2**[5].

Au passage, il est facile de s'embrouiller entre tous les termes utilisés pour décrire une Bernoulli et une binomiale (et la normale) : vous pouvez retenir qu'une probabilité est un nombre, une distribution est une loi, une densité est la fonction qui la représente.

1.4. LE MAXIMUM DE VRAISEMBLANCE

Dans l'approche classique (ou fréquentiste), on estime la probabilité de survie θ en utilisant la méthode du maximum de vraisemblance. Mais qu'est-ce que cela signifie concrètement ? Il s'agit de trouver la valeur de θ qui rend les données observées les plus probables. Autrement dit, puisque les données sont ce qu'elles sont – elles ont été observées – on cherche la valeur de θ qui maximise la probabilité que cet ensemble de données ait été produit.

Comment justifie-t-on mathématiquement cette idée plutôt intuitive ? Relisez bien la fin du paragraphe précédent. L'idée de chercher la valeur qui donne la plus grande probabilité revient à maximiser quelque chose. Mais quoi exactement ? La probabilité des données, étant donné un certain modèle paramétré par θ, autrement dit la vraisemblance ou $\Pr(\text{données} \mid \theta)$, vue à la

5. Dans tout l'ouvrage, pour une question de cohérence entre les lignes de code, les figures (dans R) et le texte, les parties entières et décimales des nombres seront séparées par un point et non une virgule.

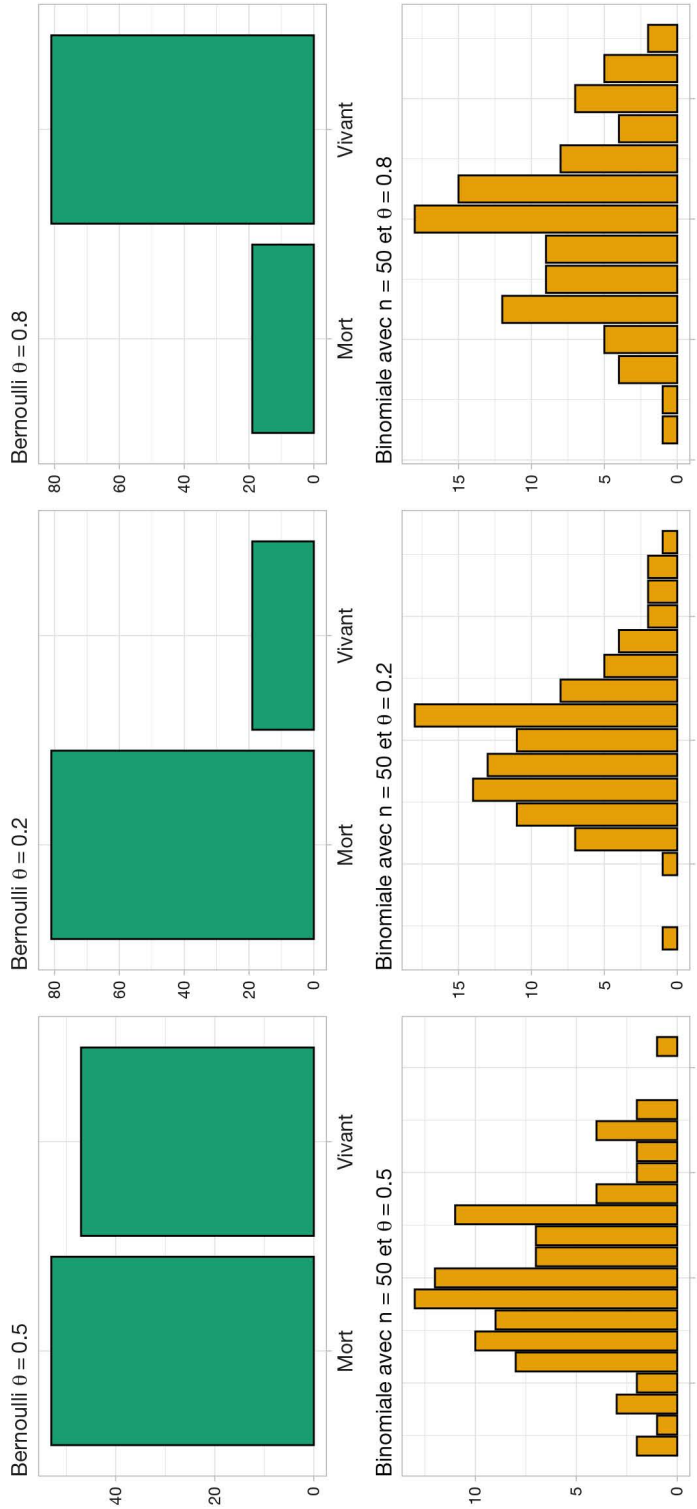

Figure 1.2. Distributions de probabilité discrètes, Bernoulli et binomiale, illustrées avec 100 simulations (des tirages aléatoires générés par ordinateur). On représente, sur la ligne du haut, la fréquence observée d'un tirage Bernoulli (mort ou vivant) pour différentes valeurs de probabilité de survie θ (0.5, 0.2 et 0.8) ; et sur la ligne du bas, les histogrammes pour un tirage binomial (nombre de survivants) avec 50 tentatives et différentes valeurs de probabilité de survie θ (0.5, 0.2 et 0.8).
Les codes pour reproduire les figures sont disponibles en ligne : https://github.com/oliviergimenez/statistique-bayes-quae