Chapitre 1

Bases probabilistes utiles en statistique bayésienne

1.1 Introduction

« *Uncertainty is everywhere and you cannot escape from it* » (Lindley, 2006).

L'incertitude est un phénomène encore trop souvent négligé par le scientifique qui préférerait de beaucoup considérer que le comportement du système qu'il étudie soit parfaitement connu. Pourtant, lorsque cette nuisance qu'on voudrait supprimer est présente, faire comme si l'incertitude inhérente à un système n'existait pas peut conduire à des résultats faux (Lindley, 2006). Reconnaître l'incertitude et la prendre en compte dans l'analyse d'un phénomène peut conduire à des résultats approximativement vrais.

Un grand nombre de phénomènes naturels sont aléatoires, c'est-à-dire non prévisibles dans notre état actuel de connaissances. En biologie, par exemple, les lois de l'hérédité suivent les lois du hasard (e.g., « Ce sera une fille ou un garçon ? »). En médecine, certaines maladies multi-factorielles (e.g., cancer) ne sont pas prévisibles. En environnement, les rendements de l'agriculture sont fortement dépendants des aléas climatiques (e.g., sécheresse). Probabilistes et statisticiens utilisent les outils mathématiques proposés par la théorie des probabilités, dont les travaux de base remontent au XVII^{e} siècle, pour modéliser efficacement les phénomènes réels dans lesquels le hasard intervient. Néanmoins, leur appréhension du hasard est bien différente en pratique. Guidé par son intuition, héritée notamment des jeux de hasard, le probabiliste utilisera plutôt les probabilités pour se fixer un cadre mathématique abstrait, appelé modèle probabiliste, lui permettant de faire certaines prévisions et d'étudier les propriétés subséquentes à ses hypothèses quant à la réalité qu'il cherche à décrire, mais sans forcément mettre son modèle à l'épreuve des faits. Il pourra notamment calculer la probabilité de certains résultats d'intérêt. Considérons, par exemple, le célèbre modèle probabiliste de Hardy-Weinberg, formulé indépendamment par G.H. Hardy et W. Weinberg en 1908, qui permet de décrire et simuler la distribution génotypique pour un gène pouvant s'exprimer sous la forme de deux allèles A et a dans une population

diploide[1] idéale[2]. Dans ce cas, les trois génotypes AA, Aa et aa sont présents dans la population avec une probabilité p^2, $2p(1-p)$ et $(1-p)^2$ respectivement, avec p un réel compris entre 0 et 1 défini comme la probabilité de l'allèle A. Le statisticien, quant à lui, utilisera les probabilités, mais travaillera toujours dans un contexte de terrain où des réalisations observées du phénomène aléatoire d'intérêt l'interrogent sur la vraisemblance des hypothèses de départ. Il pourra notamment faire des prévisions et/ou expliquer un phénomène à l'aide d'un modèle probabiliste ou encore participer à une aide à la décision. Ainsi, lorsqu'un médecin lui demande si prescrire un nouveau traitement à des patients atteints d'un cancer améliore leur survie, le statisticien pourra construire un modèle probabiliste permettant de prendre en compte les divers bénéfices et risques encourus puis s'appuiera sur les observations du médecin pour participer à la décision, la moins mauvaise possible. Il pourra également développer un modèle pour extrapoler, à la population des patients traités, les caractéristiques observées sur un échantillon tiré au hasard (e.g., espérance de vie après traitement) ou encore pour décrire la fréquence d'occurrence d'une grandeur physique au sein de la population d'étude (e.g., âge au décès des patients traités). Les lois de probabilité sont d'autant plus utiles au statisticien dit *bayésien*, conscient qu'il doit aider à la prise de décision dans un univers largement empreint d'incertitudes, qu'elles lui permettent plus généralement de quantifier son sentiment d'incertitude vis-à-vis de toute grandeur inconnue intervenant dans le système qu'il cherche à modéliser, à l'aulne d'un même outil de mesure. Ainsi, dans le cadre de l'évaluation du rendement en blé d'une parcelle agricole, le statisticien bayésien assignera, par exemple, une loi de probabilité sur la teneur en azote organique du sol ou encore sur les différentes courbes possibles de réponse du rendement à l'azote.

Enfin, l'utilisation de lois de probabilité conditionnelle est l'un des fondements du raisonnement bayésien et ce, aussi bien du point de vue de la modélisation – à travers le développement de structures dites hiérarchiques – que de l'inférence (Parent et Bernier, 2007). Cela permet notamment d'enrichir la panoplie des modèles probabilistes usuels souvent trop simplistes devant la complexité de certains phénomènes aléatoires d'intérêt. Ces points seront amplement développés dans l'ensemble de l'ouvrage.

Compte tenu du rôle important joué par les probabilités dans le travail du statisticien-modélisateur et avant d'entrer de plain-pied dans la pratique de la statistique bayésienne, ce premier chapitre présente avec rigueur, mais sans formalisme mathématique excessif, les concepts probabilistes de base dont le praticien pourrait être amené à se servir et qui sont utiles à la compréhension de l'ensemble de cet ouvrage. Le lecteur familier du concept de variable aléatoire pourra passer aux chapitres suivants (après avoir vérifié qu'il est à l'aise face aux problèmes de la section 1.4.6), quitte à revenir sur ce chapitre introductif s'il en ressent le besoin par la suite.

1. Une population diploide possède un double assortiment de chromosomes semblables.
2. Une population idéale est de taille infinie avec union aléatoire des individus. Pas de migration, mutation et sélection et générations séparées (i.e., pas de croisement entre générations différentes).

1.2 L'incertitude

1.2.1 Aléa et variabilité naturelle

Le terme aléa vient du latin *alea*, qui signifie jeu de dés. Rappelons la célèbre phrase prononcée par Jules César qui, au moment de traverser le fleuve Rubicon séparant la Gaule cisalpine et l'Italie au commandement de la XIIème légion, s'exclama : « *Alea jacta est* », qui signifie le sort en est jeté, pour signifier l'abandon d'individus à des événements sur lesquels ils n'ont pas prise. La notion d'aléa est à rapprocher du terme *hasard* dont l'origine est fréquemment attribuée à l'arabe *al-zahr*, signifiant également jeu de dés. Le mot *hasard* a également pris la signification de chance car il désigne aussi, par métaphore, tous les domaines relevant de la *science de la Chance*.

L'aléa peut être vu comme la cause de la part imprévisible des résultats d'une expérience qui, même dans des conditions expérimentales supposées identiques – pour autant que l'observateur puisse s'en assurer – peut donner lieu à des résultats différents. Une telle expérience sera qualifiée d'expérience aléatoire. Ainsi, l'aléa apparaît souvent comme le nom que nous donnons à notre ignorance de certaines conditions de l'expérience. Dans la littérature, on parle souvent de la *variabilité naturelle* des résultats observés. L'aspect imprévisible des résultats d'une expérience aléatoire permet d'argumenter que l'aléa est une source d'incertitude. Cette forme d'incertitude est appelée incertitude par essence (Parent et Bernier, 2007) ou encore incertitude aléatoire (Kirchner et Steiner, 2008). Elle a pour spécificité d'être *irréductible* (ou assumée telle dans le contexte où on choisit de se placer), même sous des conditions expérimentales supposées identiques, par l'apport de nouvelles connaissances et/ou données.

Voici quelques exemples, plus ou moins complexes, d'aléas et d'expériences aléatoires.

Exemple 1 *On lance une pièce de monnaie. Il est impossible de prévoir quelle face s'affichera au terme du lancer. Le résultat varie naturellement d'un lancer à l'autre car les conditions initiales du système ne sont, en réalité, jamais complètement identiques (figure 1.1 (a)).* ■

Exemple 2 *On lance un dé. Pour la même raison que dans l'exemple 1, il est impossible de prévoir laquelle des six faces s'affichera au terme du lancer (figure 1.1 (b)).* ■

Exemple 3 *On considère une urne contenant x boules blanches et y boules rouges ($x \geq 1$, $y \geq 1$) telles que $x + y = 100$. On suppose que le contenu de l'urne n'est pas visible de l'extérieur. On tire au hasard et avec remise deux boules dans l'urne. Il est impossible de prévoir le couple de couleurs de boules qui va être pioché. Si l'expérience est répétée plusieurs fois, les couples de couleurs obtenus seront naturellement variables d'une expérience à l'autre. Cet aléa sera, entre autres, dû au tirage et au caractère, homogène ou non, de la répartition des boules dans l'urne (figure 1.1 (c)).* ■

Exemple 4 *On lâche une bille depuis le haut de la planche de Galton (figure 1.1 (d)), dispositif expérimental constitué d'une planche inclinée sur laquelle sont disposés*

des clous en quinconce. Chaque bille roule alors à la surface de la planche, passe aléatoirement d'un côté ou de l'autre des clous pour finir sa course dans l'une des boîtes (i.e., A, B, C, D, E) situées en bas de la planche. Même en lâchant la même bille depuis la même position initiale, il est impossible de prévoir dans quelle boîte elle finira sa course. Cet aléa résulte non seulement du fait que les conditions initiales ne sont pas strictement identiques mais aussi de toutes les déviations que la bille a subies en tombant sur les différents clous de la planche. À noter que chacune de ces déviations est elle-même une expérience aléatoire. ■

Exemple 5 *On mélange 25 g de purée de courgettes contaminée à 1 ufc/g*[3] *par Bacillus cereus à 225 ml de bouillon de culture puis on pasteurise le mélange obtenu à la température cible de 90 degrés Celsius. On cherche à quantifier le nombre d'ufc dans le mélange pasteurisé obtenu. On prélève au hasard 1 ml du mélange et on compte le nombre d'ufc présentes. Si l'expérience est répétée plusieurs fois, le nombre d'ufc obtenu sera naturellement variable d'une expérience à l'autre. Cet aléa sera, entre autres, dû à la variabilité d'échantillonnage qui, elle-même, dépendra notamment de la dispersion spatiale de la bactérie dans le mélange et de sa dynamique temporelle de croissance. Le nombre d'ufc observé sera également variable d'une souche de bactérie Bacillus cereus à l'autre du fait des caractéristiques spécifiques de chaque souche en matière de thermosensibilité.* ■

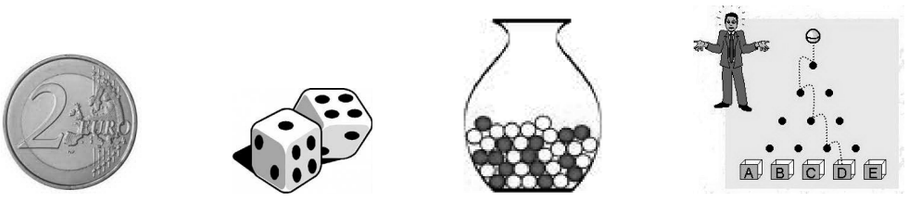
Exemple 6 *Les rayonnements ionisants peuvent porter atteinte à la forme des chromosomes par suite de perturbations locales de la structure moléculaire de l'ADN. Malgré des mécanismes de réparation efficaces, un nombre réduit de dommages peut subsister et entraîner l'apparition d'aberrations chromosomiques observables au sein des lymphocytes sanguins. Suite à l'irradiation in-vitro d'un échantillon sanguin humain exposé à une dose fixée de Cobalt 60, deux observateurs comptent visuellement le nombre de chromosomes dicentriques dans les lymphocytes circulants. Ce nombre sera probablement différent entre les deux observateurs car ceux-ci n'auront pas nécessairement la même rigueur d'observation. Par ailleurs, si l'expérience est répétée plusieurs fois (à la même dose), le nombre de chromosomes dicentriques obtenu sera naturellement variable d'une expérience à l'autre. Cet aléa sera, entre autres, dú à la variabilité d'échantillonnage des cellules, à l'hétérogénéité plus ou moins forte de dispersion de la dose à travers les cellules et à des dynamiques temporelles différentes de réparation des cellules.* ■

Exemple 7 *On réalise plusieurs mesures du diamètre d'une même pastèque à dix secondes d'intervalle. Les mesures obtenues donneront lieu à des résultats différents bien que le vrai diamètre de la pastèque n'aura pas varié pendant ce petit intervalle de temps. Cet aléa est dû, entre autres, à la qualité de l'appareil de mesure (i.e., justesse, fiabilité, résolution,* etc.*) mais aussi, dans le cas où les mesures ont été réalisées par plusieurs observateurs, à la variabilité d'attention de ces derniers ainsi qu'au choix de l'endroit où réaliser la mesure. De même, les mesures du diamètre de plusieurs pastèques issues d'un même champ sont distinctes du fait, entre autres, de la variabilité naturelle du diamêtre d'une pastèque.* ■

3. Une ufc signifie unité formant colonie.

Exemple 8 *On chronomètre le temps mis par un Parisien pour se rendre à son travail le matin. Ce temps varie généralement d'une journée à l'autre. Cet aléa est induit par l'aléa lié à l'occurrence ou non d'embouteillages, d'incidents de voyageurs, de grèves...* ■

Exemple 9 *Le radon est un gaz radioactif reconnu comme cancérigène pulmonaire depuis 1988. Dans le cadre de leur activité professionnelle, les mineurs d'uranium sont exposés au gaz radon. Le délai de survie d'un mineur jusqu'à l'occurrence d'un décès par cancer du poumon est une grandeur d'intérêt dans les études épidémiologiques. Il varie d'un individu à l'autre. Cet aléa est induit, entre autres, par la variabilité inter-individuelle des concentrations en radon inhalées mais aussi par le caractère multifactoriel du développement d'un cancer et la variabilité inter-individuelle des prédispositions.* ■



(a) (b) (c) (d)

FIGURE 1.1 – Quelques expériences aléatoires : (a) le lancer d'une pièce de monnaie, (b) le jet de dés, (c) le tirage dans une urne, et (d) la planche de Galton.

Comme illustré dans les neuf exemples précédents, l'aléa ou variabilité naturelle est inévitable et inhérent à de nombreux systèmes (e.g., biologiques, physiques) du fait de fluctuations souvent complexes et non maitrisées du système et de son environnement. Il peut s'agir de fluctuations temporelles, spatiales ou encore inter-individuelles. Il peut également s'agir d'erreurs de mesure. En effet, au cours de n'importe quel processus de collecte de données de terrain ou expérimentales, des erreurs de mesure sont généralement induites par un manque de précision (i.e., justesse, fiabilité, résolution, *etc.*) des appareils de mesure utilisés et/ou par la variabilité d'attention des observateurs.

Par la suite, nous utiliserons systématiquement le terme variabilité pour référer à la notion de variabilité naturelle et d'aléa et donc, plus généralement, à la notion d'incertitude par essence.

1.2.2 Incertitude par ignorance

« *Statements of uncertainty are personalistic, they belong to the person making them and express a relationship between that person and the real world about a statement which is being made* » (Lindley, 2006).

En pratique, le statisticien-modélisateur caractérise entièrement le fonctionnement du système aléatoire ou non qu'il cherche à décrire à l'aide d'un ensemble de grandeurs inconnues notées θ et appelées paramètres (au sens statistique du terme) ou encore état de la nature. En accord avec cette interprétation phénoménologique subjective d'un système, l'état inconnu de la nature θ ne peut prendre plusieurs valeurs distinctes, la nature ne pouvant se trouver dans plusieurs états à la fois. L'incertitude qui affecte un état de la nature est qualifiée d'*incertitude par ignorance* (Parent et Bernier, 2007) ou *incertitude épistémique* (Keller *et al.*, 2011). Elle est dûe à un manque de connaissance du système par le modélisateur qui se traduit non seulement par une connaissance imparfaite de la valeur de l'état de la nature considéré mais aussi souvent par une connaissance imparfaite du modèle mathématique à utiliser pour décrire le fonctionnement du système.

L'incertitude par ignorance est réductible par l'apport de nouvelles connaissances et/ou de données complémentaires. En particulier, la théorie des statistiques asymptotiques (Van Der Vaart, 2000) étudie les conditions sous lesquelles on peut faire diminuer arbitrairement l'incertitude par ignorance sur les paramètres inconnus d'un modèle probabiliste en augmentant la taille de l'échantillon observé.

Exemple 10 *Reprenons l'exemple 5 concernant le comptage du nombre d'ufc de Bacillus cereus dans un mélange pasteurisé d'1 ml de purée de courgettes. Certaines grandeurs caractérisant la thermosensibilité ou encore la croissance microbiologique des différentes souches Bacillus cereus peuvent être considérées comme fixes mais inconnues. Il peut s'agir, par exemple, du temps t_s de première réduction décimale de la population de bactéries, pour une souche donnée s à une température de pasteurisation cible de 90 degrés Celsius ; de la température cardinale minimale T_s^{min} et maximale T_s^{max} moyenne de croissance pour une souche donnée s ; du nombre maximal N possible en bactéries dans un mélange pasteurisé d'1 ml. Le tableau 1.1 donne des exemples de grandeurs fixes mais inconnues possibles caractérisant les résultats des expériences aléatoires décrites dans les exemples 1 à 9.* ■

1.2.3 Compléments

Dans la littérature, les définitions associées aux notions d'incertitude et de variabilité varient subtilement d'une terminologie à une autre (Parent et Bernier (2007), Keller *et al.* (2011), Cullen et Frey (1999), Kirchner et Steiner (2008)). Par ailleurs, des divergences existent sur la manière effective d'aborder la notion de variabilité dans les analyses. Certains auteurs, comme Cullen et Frey (1999), ne la considèrent pas comme contribuant à l'incertitude inhérente au système d'intérêt alors que d'autres auteurs, comme Lindley (2006) et Parent et Bernier (2007), la considèrent comme une source d'incertitude à part entière. Dans ce livre, nous adoptons le point de vue de ces derniers en considérant que la variabilité n'est qu'une source particulière d'incertitude. Néanmoins, nous sommes conscients, et le lecteur doit l'être également, qu'il s'agit avant tout d'un choix subjectif du modélisateur, dépendant de sa vision du monde réel.

Il appartient en effet au modélisateur de décider si une dispersion reflète une incertitude par ignorance ou de la variabilité tant la différence entre ces deux concepts

Ex.	Une grandeur fixe mais inconnue
1	Probabilité d'obtenir pile
2	Probabilité d'obtenir 2
3	Nombre y de boules rouges dans l'urne
4	Probabilité de tomber d'un côté ou de l'autre d'un clou
5	Température cardinale minimale moyenne de croissance d'une souche *Bacillus cereus*
6	Vitesse de réparation moyenne d'une cellule porteuse de chromosomes dicentriques
7	Mesure moyenne du diamètre d'une pastèque
8	Probabilité d'occurrence d'un embouteillage sur le trajet maison-travail
9	Dose reçue au poumon par un mineur exposé à une concentration ambiante de radon

Tableau 1.1 – Une grandeur fixe mais inconnue caractérisant les résultats de chaque expérience aléatoire associée aux exemples (notés Ex.) 1 à 9.

est subjective et ténue. Prenons l'exemple 2 relatif au lancer d'un dé. Un premier modélisateur pourra considérer que seules les probabilités d'obtenir chaque face du dé sont entâchées d'une incertitude par ignorance et que, même si on connaissait parfaitement ces valeurs, il resterait une dispersion résiduelle des résultats d'un lancer. Cette dispersion serait le reflet d'une variabilité dûe aux conditions initiales inconnues du système physique et de son environnement et à des mouvements non maitrisés du dé au cours du lancer. Toute incertitude dans notre connaissance des conditions initiales se trouverait amplifiée au cours du mouvement : un dé roulant sur un tapis conduirait à une amplification extraordinairement rapide des incertitudes initiales interdisant toute prévision. Un second modélisateur pourra considérer que la dispersion des résultats du lancer d'un dé n'est que le fruit d'une incertitude par ignorance. Si toutes les conditions initiales du système physique et de son environnement ainsi que toutes les caractéristiques du mouvement du dé étaient connues avec certitude alors le résultat du lancer d'un dé serait lui-même connu avec certitude.

Ainsi, en pratique, les modèles probabilistes incluent à la fois des grandeurs mathématiques dont le rôle est de décrire une variabilité – elles seront appelées variables aléatoires (cf. section 1.5) – et des grandeurs entâchées d'une incertitude par ignorance – appelées paramètres ou états de la nature. Reprenons l'exemple 9. Un problème d'intérêt en épidémiologie des rayonnements ionisants est d'estimer l'excès de risque de décès par cancer du poumon chez les mineurs d'uranium qui, dans le cadre de leur activité professionnelle, sont exposés de façon chronique au gaz radon. Une approche possible est de passer par la spécification d'un modèle probabiliste permettant de décrire la relation dose-risque d'intérêt. Les modèles classiquement proposés tiennent compte de la variabilité naturelle liée à l'hétérogénéité inter-individuelle, à risque fixé, de décéder par cancer du poumon ainsi qu'aux erreurs, inévitables en pratique, de mesure de l'exposition des mineurs au gaz radon. Par ailleurs, ces modèles tiennent compte de l'incertitude par ignorance relative à certains paramètres intervenant dans

le calcul de la dose reçue au poumon par un mineur après inhalation d'une certaine concentration de gaz radon : débit respiratoire moyen d'un mineur en activité, facteur d'équilibre du radon... Vient encore s'ajouter l'incertitude du modélisateur quant aux modèles à utiliser pour décrire au mieux la dose reçue au poumon à partir d'une concentration de radon inhalée ou encore pour décrire la relation entre la dose au poumon et le risque de décès par cancer du poumon.

La probabilité apparaît comme un outil mathématique pertinent permettant de modéliser non seulement la variabilité de grandeurs observables, liées à une expérience quantitative, mais aussi l'incertitude par ignorance associée à toute grandeur fixe mais inconnue. La représentation de ces deux formes d'incertitude par le même outil mathématique est essentielle à la cohérence profonde de l'approche bayésienne.

1.2.4 Événements incertains

Bien que le résultat d'une expérience aléatoire ne soit pas connu par avance (cf. section 1.2.1), l'ensemble de tous les résultats possibles est, lui, supposé parfaitement identifié. Il s'appelle l'*ensemble fondamental* ou *univers des possibles*. Il est traditionnel de représenter cet ensemble par la lettre Ω. Cet ensemble peut être fini, infini dénombrable ou infini non dénombrable. Une même expérience pourra être décrite de façons différentes selon l'objectif de l'étude. Si deux joueurs lancent chacun un dé, on pourra prendre pour Ω aussi bien l'ensemble des couples de résultats possibles (i, j) (avec i, j dans $\{1, \ldots, 6\}$) que la somme entre 2 et 12 des résultats obtenus. Le tableau 1.2 donne un univers Ω possible, associé à chacun des neuf exemples d'expérience aléatoire donnés dans la section 1.2.1.

Ex.	Expérience aléatoire	Ω
1	Lancer d'une pièce de monnaie	{pile, face}
2	Jet d'un dé	$\{1, 2, 3, 4, 5, 6\}$
3	Tirage avec remise dans l'urne	$\{(B, B), (R, R), (B, R)\}$
4	Lancer d'une bille du haut de la planche de Galton	{A, B, C, D, E}
5	Comptage du nombre d'ufc de *Bacillus cereus* dans un mélange de purée de courgettes	$\mathbb{N}$
6	Comptage du nombre de chromosomes dicentriques dans une cellule irradiée	$\{0, 1, \ldots, 46\}$
7	Mesure du diamètre d'une pastèque	$\mathbb{R}^+$
8	Temps de trajet maison-travail	$\mathbb{R}^+$
9	Délai entre exposition au radon et cancer du poumon	$\mathbb{R}^+$

Tableau 1.2 – Exemples d'univers des possibles Ω caractérisant les résultats de chaque expérience aléatoire associée aux exemples (notés Ex.) 1 à 9. Dans l'exemple 3, B=« Blanche » et R=« Rouge ». $\mathbb{N}$ désigne l'ensemble des entiers naturels. $\mathbb{R}^+$ désigne l'ensemble des nombres réels strictement positifs.

Un événement aléatoire est une assertion ou proposition logique relative au ré-