Université de Lorraine — École Nationale Supérieure des Mines de Nancy
Année universitaire 2024–2025 — Semestre 6 — Tronc commun scientifique

Inférence statistique *

par Rémi Peyre †, *maitre de conférences*

version du 8 mai 2025

∗. Réf. 6JUCSN01

†. remi.peyre@mines-nancy.univ-lorraine.fr

Comment lire ce polycopié

Comme vous pourrez le constater, ce polycopié regorge de remarques, exemples et autres commentaires, dont la fonction est d'aider à la pleine appréhension des concepts présentés, mais qui peuvent dérouter l'élève en phase d'apprentissage ou de révisions, de par le volume un peu effrayant d'information que tout cela représente... En fait, tout dans ce polycopié n'est pas à retenir par cœur, loin de là ! ◡̈ Pour s'y retrouver, les points qu'il faut vraiment connaitre (qui sont principalement les définitions essentielles de la science statistique) sont signalés dans la marge par une barre noire commençant par un point d'exclamation : on peut considérer qu'il s'agit là de ce qu'il faut *apprendre*, le reste du polycopié étant plutôt à *comprendre*. Certains passages mettant en avant des points tout particulièrement essentiels sont même marqués d'une barre noire plus épaisse, accompagnée d'un *double* point d'exclamation ! Au sein des passages non marqués dans la marge, l'importance des différents paragraphes est assez variable. J'ai typographié en petits caractères (et avec une marge agrandie) les passages dont j'estime qu'ils peuvent être totalement sautés sans nuire à la bonne acquisition du cours : je les ai même carrément exclus, ainsi que la plupart des notes de bas de page, de la version "allégée" du polycopié.

!

Exemple. La marque en marge de ce passage indiquera, dans la corps du polycopié, qu'il s'agit d'un passage qu'il faut connaitre rigoureusement, notamment en vue de l'examen. ♧

!!

Exemple. Dans le corps du polycopié, la marque en marge de ce passage signifiera qu'il s'agit d'un point du cours super-méga-important ! ♧

Exemple. À l'inverse, ce passage-ci est un complément "culturel" qui peut être sauté sans nuire à la compréhension du cours en tant que tel, mais dont j'ai estimé que vous le trouveriez peut-être intéressant, notamment si vous souhaitez aller plus loin dans la science statistique au cours des années à venir. ♧

Remarque. À l'heure où de vifs débats interpellent la société contemporaine sur la transmission, consciente ou non, des stéréotypes de genre — une question particulièrement prégnante sein de la communauté mathématique, où les femmes sont aujourd'hui largement sous-représentées —, plusieurs personnes m'ont fait remarquer que des formulations comme « le futur ingénieur » ou « le statisticien » pouvaient être perçues comme excluant implicitement les femmes du champ de l'ingénierie ou de la statistique. D'un autre côté, écrire un tel polycopié en écriture inclusive aurait été stylistiquement assez lourd... J'ai donc fait le choix de restreindre l'usage de l'écriture inclusive à cette seule préface : dans la suite du texte, lorsque je devrai désigner « le lecteur », « le statisticien » etc., j'utiliserai un masculin grammatical ; mais celui-ci devra simplement être compris comme un neutre sémantique ! ◡̈ ♧

Table des matières

Première partie

Pré-requis de probabilités

Chapitre 1

Quelques notations utilisées dans ce polycopié

1.1 Divers

— Le *pour-cent* est une unité mathématique sans dimension, notée '%', valant $\frac{1}{100}$. De même, le *pour-mille*, noté '‰', vaut $\frac{1}{1\,000}$, et le *pour-dix-mille* (rarement utilisé), noté '‱', vaut $\frac{1}{10\,000}$. Il ne faut rien entendre de plus dans les symboles '%', '‰' et '‱' qu'une simple division par resp. 100, 1 000, et 10 000 : « 3,89 ‰ », par exemple, ne signifie rien de plus que « 0,003 89 » ! L'usage est cependant de réserver l'emploi des signes '%' et compagnie à des quantités pouvant être interprétées comme des proportions ou des probabilités.

— Pour A et B deux ensembles disjoints, « $A \sqcup B$ » a la même signification que « $A \cup B$ », mais rappelle simplement que A et B sont disjoints. On utilise de même la notation « $\bigsqcup_{i\in I} A_i$ » pour l'union d'un nombre quelconque d'ensembles deux-à-deux disjoints.

— Ω étant un ensemble, pour $A \subseteq \Omega$, le complémentaire de A dans Ω est noté $\complement_\Omega A$. Bien souvent l'ensemble ambiant Ω est évident, et on note alors simplement « $\complement A$ ».

— $\bar{\mathbb{N}}$ désigne l'ensemble $\mathbb{N} \cup \{\infty\}$, muni de sa structure d'ordre et de sa topologie usuelles ; $\bar{\mathbb{R}}$ désigne l'ensemble $\mathbb{R} \cup \{-\infty, +\infty\}$, muni de sa structure d'ordre et de sa topologie usuelles.

— Pour a et b des réels, $a \wedge b$ désigne le minimum de a et b, tandis que $a \vee b$ désigne leur maximum.

— Pour a un nombre réel, a^+ et a^- désignent respectivement la partie positive et la partie négative de a, soit resp. $a \vee 0$ et $(-a) \vee 0$.

— J'utiliserai parfois la notation $\vec{x}_I$, pour 'x' un symbole et I un ensemble, comme raccourci pour $(x_i)_{i\in I}$: par exemple, le 9-uplet $(a_1, a_2, \dots, a_9)$ pourra être abrégé en « $\vec{a}_{[\![1,9]\!]}$ ».

— $\{\{a, b, \dots, z\}\}$ désigne le *multiensemble* contenant les valeurs a, b, ... et z, c'est-à-dire la liste des valeurs $a, \dots, z$ où l'*ordre* dans lequel sont présentées les valeurs n'est pas pris en compte, mais le *nombre d'occurrences* de cette valeur, si ! Par exemple, le multiensemble des lettres du mot « multiensemble », $\{\{\mathrm{m}, \mathrm{u}, \mathrm{l}, \mathrm{t}, \mathrm{i}, \mathrm{e}, \mathrm{n}, \mathrm{s}, \mathrm{e}, \mathrm{m}, \mathrm{b}, \mathrm{l}, \mathrm{e}\}\}$ peut aussi être écrit $\{\{\mathrm{b}, \mathrm{e}, \mathrm{e}, \mathrm{e}, \mathrm{i}, \mathrm{l},$

l, m, m, n, s, t, u}}, mais ce n'est par contre par la même chose que {{b, e, i, l, m, n, s, t, u}}.

— « $(X) \coloneqq (Y)$ » signifie que l'on définit (X) comme étant égal à (Y) ; de même, « $(X) =: (Y)$ » signifie que l'on introduit la notation (Y) comme valant (X). Par contre, « $(X) \underset{\text{déf}}{=} (Y)$ » signifie que l'égalité entre (X) et (Y) se *déduit* de la définition, soit de (X), soit de (Y).

— Pour $A \subseteq \mathbb{R}^d$ de forme suffisamment régulière, $\mathrm{vol}(A)$ désignera le volume (d-dimensionnel) de A (alias « longueur » en dimension 1 ou « aire » en dimension 2).

— Pour $\mathcal{X}$ un ensemble, l'ensemble des distributions de probabilité sur $\mathcal{X}$ sera noté $\mathscr{M}_1(\mathcal{X})$.

1.2 Booléens

La fonction *négation*, notée « $\neg$ » (parfois utilisée sans parenthèses en tant qu'opérateur, comme on le fait p. ex. pour la fonction sin), est la fonction de $\{\mathsf{VRAI}, \mathsf{FAUX}\}$ dans lui-même définie par $\neg\mathsf{VRAI} \coloneqq \mathsf{FAUX}$ et $\neg\mathsf{FAUX} \coloneqq \mathsf{VRAI}$. Il existe également les opérateurs « et » et « ou », de $\{\mathsf{VRAI}, \mathsf{FAUX}\}^2$ dans $\{\mathsf{VRAI}, \mathsf{FAUX}\}$, dont les définitions sont évidentes : nous n'introduisons pas de notations pour ces opérateurs, qui seront écrits en toutes lettres. La *fonction indicatrice*, notée « $\mathbf{1}_\bullet$ », est la fonction de $\{\mathsf{VRAI}, \mathsf{FAUX}\}$ dans $\{0, 1\}$ définie par $\mathbf{1}_{\mathsf{VRAI}} \coloneqq 1$ et $\mathbf{1}_{\mathsf{FAUX}} \coloneqq 0$. Nous utiliserons aussi la fonction "co-indicatrice" $\mathbf{0}_A \coloneqq \mathbf{1}_{\neg A} = 1 - \mathbf{1}_A$.

Une écriture comme « $a < b$ » ou « $f(x, y) \in A$ » est ce qu'on appelle une *proposition* : autrement dit, c'est un énoncé auquel, selon la valeur des paramètres libres de cet énoncé (lesdits paramètres étant a et b dans le premier cas, resp. x et y dans le second cas si nous supposons que f et A se réfèrent à une fonction et un ensemble fixés), on attribue un booléen qui dit si l'énoncé tient ou non, booléen qu'on qualifie de *valeur de vérité* de l'énoncé. Ainsi, une proposition peut être vue comme une application qui va de l'espace que peuvent prendre les paramètres libres dans l'ensemble $\{\mathsf{VRAI}, \mathsf{FAUX}\}$, associant à chaque valeur possible du jeu de paramètres la valeur de vérité de la proposition pour cette valeur-là. Dans le cadre de ce cours, où on se place dans un paradigme probabiliste, les variables libres des propositions seront des variables *aléatoires*, de sorte que la valeur de vérité d'une proposition sera une variable aléatoire à valeurs dans $\{\mathsf{VRAI}, \mathsf{FAUX}\}$: par exemple, si mon univers probabiliste est Ω et que $X \coloneqq X(\omega)$ est une variable aléatoires à valeurs réelles, la proposition « $X > 0$ » est une variable aléatoire qui vaut VRAI sur les $\omega \in \Omega$ tels que $X(\omega) > 0$, resp. FAUX sur les ω tels que $X(\Omega) \leqslant 0$. Et l'indicatrice $\mathbf{1}_{X > \omega}$ de cette proposition sera la v.a. à valeurs dans $\{0, 1\}$ valant resp. 1 dans le premier cas et 0 dans le second.

Dans vos cours de probabilités de lycée et de S5, on vous a défini un évènement comme un sous-ensemble de l'univers probabiliste Ω. Dans ce cours, nous verrons plutôt les évènements comme des *propositions*, ce qui correspond bien mieux à la vision intuitive que nous en avons : par exemple, « le dé est tombé sur '4' » sera un évènement (en supposant que le résultat du dé soit modélisé par une variable aléatoire). En fait, là où vous voyiez auparavant un évènement comme un sous-ensemble $A \subseteq \Omega$, la contrepartie propositionnelle pour désigner cet évènement sera « $\omega \in \Omega$ » (qui est bien une variable aléatoire à valeurs booléennes) ; et inversement, pour P une proposition dont la valeur de vérité ne dépend que de l'éventualité ω, la vision

ensembliste de l'évènement P serait l'ensemble des ω pour lesquels P est vrai. Au niveau du formalisme, voir les évènements comme des propositions fera que, si E_0 et E_1 sont des évènements, on parlera de la conjonction de ces deux évènements en écrivant « E_0 et E_1 », et de leur disjonction en écrivant « E_0 ou E_1 », alors qu'en vision ensembliste vous auriez écrit resp. « $E_0 \cap E_1$ » et « $E_0 \cup E_1$ ». Au niveau notationnel, pour éviter de manipuler des guillemets dans les énoncés mathématiques, nous délimiterons si nécessaire une proposition comme « $f(x, y) \in A$ » par des accolades plutôt que par des guillemets, ce qui pourra donner les phrases comme, par exemple, « les évènements $\{X \geqslant 0\}$ et $\{|Y| \leqslant 1\}$ sont indépendants » : le choix de cette notation étant justifié par la correspondance implicite entre propositions et évènements.

1.3 Infinitésimaux

Deux notions légèrement distinctes d'intégration sont utilisées dans ce cours, que nous qualifierons respectivement d'intégration *algébrique* et *géométrique* :

Une intégrale algébrique s'écrit sous la forme :

$$\int_{x=a}^{b} f(x)\mathrm{d}x. \tag{AA}$$

Éventuellement on pourra omettre de rappeler la variable muette au niveau du signe d'intégration (« $\int_a^b f(x)\mathrm{d}x$ »), voire se livrer à des simplification notationnelles encore plus radicales (jusqu'à « $\int_a^b f$ ») ; mais dans tous les cas, le domaine d'intégration est indiqué par des *bornes* figurant resp. en indice et en exposant du symbole d'intégration — corolairement, l'intégrale sera toujours en dimension 1 [*].

L'interprétation d'une telle intégrale est la suivante. On imagine que x se déplace de a à b par des pas infinitésimaux, se déplaçant successivement de $a =: x_0$ via x_1, x_2, ... jusqu'à $x_N := b$ [†], avec à chaque fois $|x_{i+1} - x_i| \ll 1$, le nombre N de pas étant évidemment très grand. Dans ce cas, « $\mathrm{d}x$ » représente l'*accroissement* $x_{i+1} - x_i$, qui est donc un *nombre réel* (infinitésimal) ; on note aussi cet accroissement infinitésimal δx lorsqu'on l'utilise en l'absence de symbole d'intégration associé.

À l'inverse, une intégrale géométrique s'écrit sous la forme :

$$\int_{x\in U} f(x)\,\mathrm{vol}(\mathrm{d}x). \tag{AB}$$

(Des simplications comme « $\int_U f(x)\,\mathrm{vol}(\mathrm{d}x)$ », voire « $\int_U f$ », seront tolérées). Cette fois-ci, le domaine d'intégration est écrit sous la forme d'un *ensemble* (et non de *bornes*), qui peut être de dimension quelconque. En outre, il apparait dans l'intégrale une *mesure* (en général la mesure de volume [‡]), qui est appliquée à $\mathrm{d}x$.

Cette fois-ci, « $\mathrm{d}x$ » ne représente plus un nombre mais un *voisinage infinitésimal* (dans U) de x ; on notera aussi ce voisinage infinitésimal $\mathrm{d}x$ lorsqu'on l'utilise en

[*]. On peut en fait définir un avatar de l'intégrale de Riemann en dimension plus grande que 1 dans le cadre de la théorie des formes différentielles ; mais cela sort du champ de ce cours et même des compétences d'un ingénieur, fût-il mathématicien !

[†]. Noter que rien n'oblige ici la suite $(x_i)_i$ à être monotone !

[‡]. Il sera *toléré* de sous-entendre la mesure lorsqu'il s'agit de la mesure de volume, écrivant alors « $\int_{x\in U} f(x)\mathrm{d}x$ », mais cela est fortement déconseillé !

l'absence de symbole d'intégration associé (« $\mathrm{d}x$ » est alors une *variable* — à valeurs dans les parties infinitésimales de U). L'intégrale signifie alors qu'on découpe le domaine d'intégration U selon une *partition* en ensembles infinitésimaux $\mathrm{d}x_i$, chaque $\mathrm{d}x_i$ étant repéré par un point x_i contenu dans ce voisinage (ou éventuellement à une distance infinitésimalement proche) ; et l'intégrale se réfère alors, formellement, à la quantité

$$\sum_{i=1}^{N} f(x_i)\,\mathrm{vol}(\mathrm{d}x_i). \tag{AC}$$

Plus généralement, chaque fois que nous ferons intervenir une variable $\mathrm{d}x$, il faudra comprendre que celle-ci désigne « un voisinage infinitésimal du point x » ! Par exemple, si j'écris quelque chose comme

$$\mathbb{P}(X \in \mathrm{d}x) = f(x)\,\mathrm{vol}(\mathrm{d}x), \tag{AD}$$

cela signifie que la loi de la variable aléatoire X admet, au point x, la densité $f(x)$ par rapport à la mesure de Lebesgue.

Remarque (AE). Dans le cas de la dimension 1, on se retrouve alors avec deux notions d'intégration. Celles-ci sont évidemment liées : pour $a \leqslant b$, on a $\int_{x=a}^{b} f(x)\mathrm{d}x = \int_{x\in]a,b[} f(x)\,\mathrm{vol}(\mathrm{d}x)$, et pour $a > b$, $\int_{x=a}^{b} f(x)\mathrm{d}x = -\int_{x=b}^{a} f(x)\mathrm{d}x = -\int_{x\in]b,a[} f(x)\,\mathrm{vol}(\mathrm{d}x)$. Au passage, c'est parce que la formule prend une forme différente selon le sens de comparaison entre a et b qu'une valeur absolue intervient dans la formule de changement de variables en formalisme géométrique, mais pas en formalisme algébrique (confer remarque (BA)) : en effet, dans le théorème (AW), lorsque Φ est décroissante, Φ intervertit l'ordre des bornes de l'intervalle d'intégration algébrique, et il n'y a donc pas le même signe pour passer de l'intégrale algébrique à l'intégrale géométrique, d'où le fait que ce n'est pas Φ' mais $-\Phi'$ qui intervient dans ce cas : autrement dit, que Φ soit croissante ou décroissante, c'est $|\Phi'|$ qui intervient pour le changement de variable en formalisme géométrique. ♣

Chapitre 2

Distributions de probabilité et variables aléatoires

La statistique inférentielle s'appuie de façon essentielle sur la théorie des probabilités. Vous avez déjà vu diverses notions autour de cette théorie : d'abord au lycée et en classes préparatoires, sur un certain nombre de points épars, puis avec le cours de M. Villemonais au semestre 5, grâce auquel vous avez pu regrouper et étendre ces notions au sein d'un formalisme rigoureux et cohérent.

Dans cette partie, nous allons rappeler les notions que vous avez vues, et introduire quelques extensions de celles-ci qui seront utilisées dans ce cours. Le formalisme sera moins rigoureux que dans vos cours précédents, car ici le but est de pouvoir manipuler de façon simple et intuitive les concepts nécessaires aux calculs statistiques, quitte à ce que le cadre mathématique sous-jacent comporte quelques imprécisions (en particulier : occultation des problématiques de mesurabilité, utilisation de notations à base infinitésimaux, confusion des notions d'égalité et d'équivalence presque-sure).

2.1 Distributions de probabilité

Manipulation informelle des distributions de probabilité

Dans le cadre de ce cours, nous interprèterons les distributions de probabilités à densité comme un cas-limite des distributions discrètes. Plus précisément, nous considèrerons qu'on peut découper $\mathbb{R}^n$ en un très grand nombre de zones infinitésimales de la forme $\mathrm{d}x$, où $\mathrm{d}x$ représente un « voisinage infinitésimal de x » [*] dans $\mathbb{R}^n$, qu'on peut se représenter par exemple comme étant de la forme $\prod_{i=1}^{n}]x_i - \varepsilon, x_i + \varepsilon[$ avec ε très petit [†]. Il est entendu que le découpage de $\mathbb{R}^n$ en zones est fait

[*]. La notion de « voisinage » utilisée ici n'est pas exactement la même que celle utilisée en topologie : ici, quand nous disons que $\mathrm{d}x$ est un voisinage de x, cela signifie juste que la distance entre les points de $\mathrm{d}x$ et x est (au plus) du même ordre de grandeur que la distance entre deux points de $\mathrm{d}x$. Par exemple, dans $\mathbb{R}$, lorsque $\varepsilon \searrow 0$, $[\varepsilon, 2\varepsilon]$ sera bien considéré comme un voisinage infinitésimal de 0 (même si en l'occurrence il ne contient pas 0 lui-même), mais pas $[\varepsilon^{1/2}, \varepsilon^{1/2} + \varepsilon]$ ne l'est pas, car la distance de cet intervalle à 0, qui est $\varepsilon^{1/2}$, est beaucoup plus grande que sa largeur, qui est ε.

[†]. Attention : Ici $\mathrm{d}x$ ne désigne donc pas un *accroissement infinitésimal* de x, qui serait un *nombre* (ou plus généralement un n-uplet), mais un *voisinage infinitésimal de x*, qui est donc une *zone infinitésimale* de $\mathbb{R}^n$. Accroissements et voisinages infinitésimaux sont donc notés de la même façon (ce choix est dû au fait qu'ils jouent des rôles très similaires en théorie de l'intégration),

de sorte que les zones constituent une *partition* de $\mathbb{R}^n$, c.-à-d. que tout point de $\mathbb{R}^n$ appartienne à une zone et une seule. Dans ce cas-là, on pourra considérer que, pour une zone $\mathrm{d}x$ voisinage de x, c'est comme s'il y avait au point x une masse discrète très petite, dont la valeur serait proportionnelle au volume de $\mathrm{d}x$. Ainsi, nous nous représentons notre distribution de probabilité à densité comme une myriade de minuscules masses discrètes. Le point crucial étant bien entendu que, lorsque « $\mathrm{d}x$ » devient réellement infinitésimal (c.-à-d. dans l'asymptotique idoine), l'objet qu'on obtient à la limite de dépend pas de la façon précise dont on a maillé notre espace $\mathbb{R}^n$ en zones.

Pour décrire notre distribution de probabilité à densité, il faudra alors dire quelle est la masse de chaque zone infinitésimale. Par exemple, on écrira : « La distribution de probabilité $\boldsymbol{P}$ est telle que, pour $\mathrm{d}x$ un voisinage infinitésimal de x, $\boldsymbol{P}(\mathrm{d}x) = f(x)\,\mathrm{vol}(\mathrm{d}x)$ », ce qui est simplement une façon intuitive de formaliser ce qu'on écrirait, en termes rigoureux, « $\boldsymbol{P}$ est la distribution sur $\mathbb{R}^n$ dont la densité en x (par rapport à la mesure de Lebesgue) vaut $f(x)$ ». Bien entendu, l'expression dans le membre de droite doit toujours être proportionnelle au volume de $\mathrm{d}x$.

Les grandes familles de distributions de probabilité... et les autres !

Distributions de probabilité ni discrètes, ni à densité Dans les cours de probabilité et de statistique, on manipule majoritairement des distributions de probabilité discrète et des distributions de probabilité à densité. Il pourrait dès lors être tentant, pour l'ingénieur, de croire que ce sont les seuls types de distributions de probabilité qu'on est amené à rencontrer dans un cadre industriel... Mais c'est FAUX ! Même si effectivement, certains cas très "exotiques" relèvent surtout de l'univers des mathématiciens "purs", il y a des cas très concrets qui sortent de la dichotomie entre mesures discrètes et à densité, et notamment des mesures "mixtes" qui "combinent" ces deux concepts.

Exemple (AF) (Données tronquées). Imaginez que vous êtes chargé·e de contrôler la qualité de la production d'une usine de pompes hydrauliques. Vous allez échantillonner des pompes sorties de la chaine de production et les tester selon un certain protocole jusqu'à ce qu'elles tombent en panne ; votre objectif étant de dire « lorsque je prends une pompe sortie de l'usine, la probabilité que sa durée de vie soit comprise entre telle et telle durée est égale à telle valeur ». Disons qu'en général, les pompes tiennent de l'ordre de 30 jours [symbole du jour : 'd'] dans les conditions du test. Vous pourriez bien entendu tester vos pompes jusqu'à ce qu'elles soient toutes tombées en panne. Mais cela prendrait beaucoup de temps pour pas grand-chose... En pratique, on se fixe une durée maximale pour le test, disons 60 d. La durée de vie mesurée pour une pompe testée peut alors être de deux types : soit, par exemple « Cette pompe a tenu 20,8 d » ; soit « Cette pompe a tenu 60 d, c.-à-d. qu'elle était encore en fonctionnement à la fin du test ». La durée de vie mesurée pour une pompe prend donc ses valeurs dans l'intervalle $[0, 60]$ d, mais ce sera typiquement une distribution de probabilité *mixte* ayant une densité sur $[0, 60[$, et une masse ponctuelle au point 60, au sens où on décrira alors une telle distribution de probabilité par une formulation du type suivant : « Pour $\mathrm{d}x$ un voisinage infinitésimal de x, la masse attribuée par la distribution de probabilité à un voisinage infinitésimal $\mathrm{d}x$ de x vaut

mais le contexte permettra toujours de trancher quant au concept auquel la notation « $\mathrm{d}x$ » fait allusion.

— $(3{,}333 \times 10^{-2}\ \mathrm{d}^{-1}) \exp\big(-x\,/\,(30\ \mathrm{d})\big)\ \mathrm{vol}(\mathrm{d}x)$ si $x \in]0, 60[$;
— $13{,}534\ \%$ si $x = 60$.

». Notez bien qu'il y a un facteur « vol(dx) » dans le premier cas, qui correspond à la partie à densité de notre distribution de probabilité, mais qu'il n'y en a *pas* dans le second, qui correspond à la partie discrète ! ♧

Remarque (AG). Dans un tel cas, il est essentiel de ne surtout pas chercher à comparer une valeur de *densité* (par exemple, la densité la probabilité de la durée de vie autour de 7 d vaut $2{,}574 \times 10^{-2}\ \mathrm{d}^{-1}$) à une valeur de *probabilité* (par exemple, la probabilité d'une durée de vie de 60 j vaut $14{,}435\ \%$) : cela reviendrait en substance à comparer une aire et une longueur, et d'ailleurs dans le cas présent ces deux quantités n'ont pas la même homogénéité physique ! ♧

Exemple (AH) (Données inflatées en zéro). Si on cherche la plus grande faille dans un échantillon de matériau, la taille de cette plus grande faille suit une distribution sur $\mathbb{R}_+$ qui a une densité en général ; cependant il y a aussi une probabilité non nulle qu'il n'y ait *aucune* faille, de sorte que la distribution a aussi une masse non nulle en zéro... Dans ce cas, la loi de probabilité $\boldsymbol{P}$ suivi par la longueur de la plus grande faille sera telle que, pour dx un voisinage infinitésimal de x,

$$\boldsymbol{P}(\mathrm{d}x) = \begin{cases} p_0 & \text{pour } x = 0\,; \\ f(x)\,\mathrm{vol}(\mathrm{d}x) & \text{pour } x > 0, \end{cases} \tag{AI}$$

où p_0 et f sont tels que $p_0 + \int_{x=0}^{\infty} f(x)\mathrm{d}x = 1$. ♧

Mentionnons enfin quelques cas de distributions ni discrètes ni à densité plus rares, mais qu'on peut tout de même parfois rencontrer dans les situations d'ingénierie :

Exemple (AJ) (Distribution uniforme sur la sphère). Si on cherche à décrire le lieu aléatoire d'un évènement se produisant à la surface de la Terre (par exemple, la chute d'une météorite), il est naturel de modéliser ladite surface de la Terre par une sphère. Mais alors, si on considère par exemple la densité uniforme sur cette sphère (au sens où la probabilité de tomber sur une certaine partie du globe est proportionnelle à la *surface* de celle-ci), on aura une mesure qui n'est ni discrète, ni à densité dans l'espace tridimensionnel ! (En l'occurrence on pourrait *paramétrer la surface de la Terre, par exemple par ses latitude et longitude, pour sa ramener à une distribution probabilité à densité en dimension* 2 ; mais fondamentalement la "vraie" probabilité n'est pas à densité). Dans ce cas, la probabilité d'un voisinage infinitésimal de x est nulle si x n'est pas sur la sphère ; et si x est sur la sphère, elle est proportionnelle à la *surface* infinitésimale de l'*intersection* $\mathrm{d}x \cap \boldsymbol{S}$ (où $\boldsymbol{S}$ désigne notre sphère). Sachant que définir la notion de surface n'est pas si simple qu'il n'y parait [‡]... ♧

L'exemple ci-dessus rentre plus généralement dans le cadre des distributions de probabilité à densité sur une « sous-variété » de $\mathbb{R}^d$: en quelque sorte, il s'agit de

[‡]. Pour ceux que cela intéresse, si $\boldsymbol{M}$ est une surface plongée dans l'espace tridimensionnel, x un point de $\boldsymbol{M}$, et que dx est un voisinage infinitésimal de x *dans la surface* $\boldsymbol{M}$ de forme (presque) parallépipédique, avec pour côtés $\vec{\mathrm{d}}x_1$ et $\vec{\mathrm{d}}x_2$ (autrement dit, dx est l'ensemble des $x + \alpha\vec{\mathrm{d}}x_1 + \beta\vec{\mathrm{d}}x_2$ lorsque (α, β) décrit un ensemble de la forme $]a, a+1[\times]b, b+1[$), alors la surface de dx est définie comme le *volume* tridimensionnel du parallélipipède de côtés $\vec{\mathrm{d}}x_1$, $\vec{\mathrm{d}}x_2$ et $\vec{\Delta}x_3$, où $\vec{\Delta}x_3$ est un vecteur unitaire orthogonal à la fois à $\vec{\mathrm{d}}x_1$ et $\vec{\mathrm{d}}x_2$ — et ce volume peut d'ailleurs se calculer comme la valeur absolue du déterminant du triplet $(\vec{\mathrm{d}}x_1, \vec{\mathrm{d}}x_2, \vec{\Delta}x_3)$.

distributions qui sont portées par des espaces de dimension intermédiaire entre 0 et d.

2.2 Variables aléatoires

Loi d'une variable aléatoire

Notion de mesure image L'introduction du concept de loi est l'occasion de définir celui de *mesure image* :

! **Définition (AK)** (Mesure image)**.** Soit P une distribution de probabilité sur un espace $\mathcal{X}$, $\mathcal{Y}$ un autre espace, et $f : \mathcal{X} \to \mathcal{Y}$ une fonction (déterministe). Alors, si X est une variable aléatoire sur Ω à valeurs dans $\mathcal{X}$ ayant pour loi P, la loi de $f(X)$ (qui est une variable aléatoire sur Ω à valeurs dans $\mathcal{Y}$) ne dépend que de P et de f (cf. proposition (AL)) : on appelle cette distribution de probabilité (qui porte sur $\mathcal{Y}$) la *mesure image de P par f*, et on la note $f_* P$. ♡

Proposition (AL) (Formule de la mesure image)**.** *La mesure image $f_* P$ peut être caractérisée directement par la formule suivante : pour tout $B \subseteq \mathcal{Y}$,*

$$\big(f_* P\big)(B) \coloneqq P(f^{-1}(B)). \tag{AM}$$

♢

Démonstration. Soit X une v.a. à valeurs dans X de loi P. Alors, pour $B \subseteq Y$,

$$\begin{aligned}\big(f_* P\big)(B) &\underset{\text{déf}}{=} \big(\mathrm{Loi}(f(X))\big)(B) \underset{\text{déf}}{=} \mathbb{P}\big(f(X(\omega)) \in B\big) \\ &= \mathbb{P}(X(\omega) \in f^{-1}(B)) \underset{\text{déf}}{=} \big(\mathrm{Loi}(X)\big)(f^{-1}(B)) = P(f^{-1}(B)).\end{aligned}$$

♤

Remarque (AN). La notion de mesure image permet de re-formuler la définition d'une loi de façon plus compacte : la loi de la v.a. X est simplement la mesure-image de $\mathbb{P}$ par X (vue comme une application de Ω dans $\mathcal{X}$). ♧

Remarque (AO). La notion de mesure image est un outil puissant pour définir de nouvelles distributions de probabilité. Si par exemple P est la distribution uniforme sur $[0,1]$ et $f : \mathbb{R} \to \mathbb{R}$ la fonction « carré », on peut définir en quelques mots la mesure image $f_* P$[§] : on pourra dire « la mesure image par la fonction carré de la mesure uniforme sur $[0,1]$ », ou encore « la loi de X^2 pour $X \sim \mathrm{Unif}^{\mathrm{me}}(0,1)$ » si on préfère parler en termes de variables aléatoires[¶]. ♧

Quelques abus de notations sur les lois Dans ce cours, nous utiliserons par commodité deux notations où (dans ces contextes précis) les lois sont assimilées à des variables aléatoires suivant ces lois :

[§]. Qui, en l'occurrence, n'est autre que la distribution bêta de paramètres $(1/2, 1)$.

[¶]. Parler en termes de variables aléatoires pourra s'avérer commode lorsqu'on cherche à créer une nouvelle loi à partir de *plusieurs* lois de départ : par exemple, il est plus intuitif de comprendre la définition « la loi de $(X+Y)$ pour $X \sim \mathrm{Unif}^{\mathrm{me}}(0,1)$ et $Y \sim \mathrm{Normale}(0,1)$ avec X et Y indépendantes » que « la mesure image, par la fonction somme, du produit des lois $\mathrm{Unif}^{\mathrm{me}}(0,1)$ et $\mathrm{Normale}(0,1)$ ».

— Une notation comme
$$\mathbb{P}(\text{Expon}^{\text{le}}(1) \geqslant 3) \tag{AP}$$
désignera la probabilité qu'une variable aléatoire de loi *Exponentielle*(1) soit supérieure ou égale à 3 : cette probabilité ne dépend en effet pas de la variable aléatoire précise considérée mais bien uniquement de sa distribution, puisque ce n'est autre que $\big(\text{Expon}^{\text{le}}(1)\big)(]3, \infty])$.

— Une notation comme
$$\text{Normale}\big(0, \tfrac{1}{\sqrt{2}}\big)^2 \tag{AQ}$$
désignera la loi du carré d'une variable aléatoire de loi $\text{Normale}\big(0, \frac{1}{\sqrt{2}}\big)$: là encore, cela ne dépend pas de la variable aléatoire précise choisie mais uniquement de sa loi, puisque ce n'est autre que la mesure-image, par la fonction $x \mapsto x^2$, de la loi $\text{Normale}\big(0, \frac{1}{\sqrt{2}}\big)$.

Remarque (AR). Bien sûr, on peut combiner les deux notations ci-dessus : par exemple,
$$\mathbb{P}(|\text{Normale}(0,1)| \geqslant 1{,}733) \tag{AS}$$
désigne la probabilité que la valeur absolue d'une variable aléatoire de loi Normale(0, 1) dépasse 1,733. ♧

Remarque (AT). En revanche, attention : une expression comme « $\text{Unif}^{\text{me}}(0,1) \times \text{Expon}^{\text{le}}(1)$ » n'aurait pas de sens ! En effet, de la seule information que X et Y suivent resp. les lois $\text{Unif}^{\text{me}}(0,1)$ et $\text{Expon}^{\text{le}}(1)$, on ne peut pas déduire la loi de $X \times Y$: cela dépend en effet de la loi jointe de (X, Y), comme nous le verrons plus loin. La notation assimilant une loi de probabilité à une variable aléatoire suivant cette loi ne fait sens que lorsqu'il y a *une seule* variable aléatoire impliquée ! (et donc, dans la formule, une seule distribution de probabilité). ♧

Remarque (AU). Pour continuer sur la remarque ci-dessus, en revanche, la loi de $X \times Y$ sera bien spécifiée dès lors qu'on impose que X et Y soient *indépendantes* : dans ce cas, la loi correspondante est la mesure-image, par l'application « produit (sur $\mathbb{R}$) », de la loi-produit $\text{Unif}^{\text{me}}(0,1) \otimes \text{Expon}^{\text{le}}(1)$. Dans le cas où on considère la *somme* de deux variables indépendantes, cette notion de « somme de variables aléatoires indépendantes suivant les lois de » correspond en fait à un *produit de convolution*[‖] : on peut alors noter $P * Q$ la mesure-image de $P \otimes Q$ par la fonction « somme ». ♧

Remarque (AV). Pour P une distribution de probabilité, il faut bien prendre garde à ne pas confondre une expression comme « $P + a$ », qui désigne en vertu du formalisme précédent la mesure-image de la distribution P par l'application $x \mapsto x + a$ (distribution qui n'est autre que le produit de convolution $P * \delta_a$) avec une expression comme « $P + \delta_a$ », qui désigne la distribution de masse (ici ce n'est pas une distribution de probabilité, car la masse totale vaut 2) qui, à tout ensemble X, associe la masse $P(X) + \delta_a(X) = P(X) + \mathbf{1}_{a \in X}$. ♧

[‖]. Ici la notion de « produit de convolution » s'entend sur des *distributions de masse*, et non sur des *fonctions*. On utilise néanmoins le même nom, car les deux définitions coïncident lorsqu'on assimile une fonction à la distribution de masse ayant cette fonction pour densité.

2.3 Changement de variables multidimensionnel

La formule

! **Théorème** (AW) (Formule du changement de variables multidimensionnel)**.** *Soient* $d \in \mathbb{N}$, $\boldsymbol{M} \in \mathbb{R}^d$, *et* $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ *une application de classe* C^1 *au voisinage de* $\boldsymbol{M}$. *Alors, pour* $\mathrm{d}\boldsymbol{M}$ *un voisinage infinitésimal de* $\boldsymbol{M}$ *dans* $\mathbb{R}^d$ *:*

$$\mathrm{vol}(\varphi(\mathrm{d}\boldsymbol{M})) = |J_\varphi(\boldsymbol{M})| \,\mathrm{vol}(\mathrm{d}\boldsymbol{M}), \tag{AX}$$

où $J_\varphi(\boldsymbol{M})$ *est le* jacobien *de* φ *évalué en* $\boldsymbol{M}$, *c.-à-d. le déterminant la matrice de ses dérivées partielles :*

$$J_\varphi(\boldsymbol{M}) \coloneqq \det\left(\left(\frac{\partial \varphi_i}{\partial x_j}(\boldsymbol{M})\right)\right)_{ij}. \tag{AY}$$

— Dans cette formule, on a désigné comme d'habitude par « φ_i » les composantes de la fonction φ, et par « x_j » les coordonnées canoniques de $\mathbb{R}^d$: le déterminant considéré est donc celui d'une matrice $d \times d$. ♢

Remarque (AZ). Ne pas oublier la valeur absolue...! (d'ailleurs, un volume ne peut être que positif!). Noter que, lorsque le domaine de φ est connexe et que φ est bijective (ce qui sera systématiquement le cas lorsqu'on procèdera à un changement de variables), le jacobien sera forcément de signe constant, de sorte que la valeur absolue se manifestera simplement comme un éventuel changement de signe sur le résultat. ♧

Remarque (BA). Noter qu'en dimension 1, la formule de changement de variable pour laquelle on tombe n'est *pas la même* que celle à laquelle vous êtes habitués : il y a ici une valeur absolue qui apparait... Cela provient de la distinction entre le formalisme « algébrique » et le formalisme « géométrique » pour l'intégration (le premier étant celui que vous avez vu en classes préparatoires, le second celui qu'on utilise en théorie de la mesure) : voir aussi la remarque (AE). ♧

Le théorème (AW) contient la formule essentielle — qui est, à mes yeux, la seule formule du cours qu'on soit plus ou moins obligé de retenir "bêtement" [**] — ; mais il est pratiquement impossible de comprendre comment la mettre en œuvre sans avoir vu un exemple au préalable.

Exemples de mise en œuvre pour le calcul d'une mesure-image

Maintenant nous allons montrer comment la formule de changement de variables multidimensionnel peut être utilisée pour déterminer une mesure-image. Une première possibilité est d'utiliser la méthodes des fonctions-tests, ce qui nous ramène à la situation précédente avec les intégrales. Mais en fait, le formalisme infinitésimal que nous avons introduit permet de trouver *directement* la mesure-image sans passer par des intégrales, ce qui est plus simple — et limite du coup les risques d'erreur ! ◡̈

[**]. Et encore... En fait, cette formule est plus ou moins la *définition* du déterminant : il faudrait donc dire, plutôt, que ce sont les formules pour calculer les déterminants qui sont à retenir "bêtement" ! ◡̈

Exemple (BB) (Logarithme d'une loi exponentielle). Rappelons ici que la loi $\mathrm{Expon}^{\mathrm{le}}(\lambda)$, pour $\lambda \in \mathbb{R}_+^*$, est la loi décrite par la densité suivante (où $\mathrm{d}x$ est un voisinage infinitésimal de x) :

$$\mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \in \mathrm{d}x) = \mathbf{1}_{x>0}\lambda \mathrm{e}^{-\lambda x}\,\mathrm{vol}(\mathrm{d}x). \tag{BC}$$

Cette loi ne prenant presque-surement que des valeurs strictement positives, on peut se demander quelle est sa mesure-image par la fonction $\log_b$ ($\log_b(\bullet)$ est le logarithme en base b : $\log_b(x) = \log(x)/\log(b)$). Pour cela, demandons-nous quelle est masse que la mesure-image $\log_{b\,*}\mathrm{Expon}^{\mathrm{le}}(\lambda)$ associe à un voisinage $\mathrm{d}y$ de $y \in \mathbb{R}$. C'est la probabilité que, pour une variable aléatoire X suivant la loi $\mathrm{Expon}^{\mathrm{le}}(\lambda)$, $\log_b X$ soit dans $\mathrm{d}y$; mais puisque $\log_b(\bullet)$ est une bijection de $\mathbb{R}_+^*$ dans $\mathbb{R}$ dont la bijection réciproque est l'exponentielle de base b, dire que $\log_b X \in \mathrm{d}y$ est équivalent à dire que $X \in b^{\mathrm{d}y}$:

$$\mathbb{P}(\log_{b\,*}\mathrm{Expon}^{\mathrm{le}}(\lambda) \in \mathrm{d}y) = \mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \in b^{\mathrm{d}y}). \tag{BD}$$

Or $b^{\mathrm{d}y}$ est évidemment un voisinage infinitésimal de b^y. Son volume se déduit du volume de $\mathrm{d}y$ via la formule de changement de variable, en observant que l'application $y \mapsto b^y$ a pour dérivée $y \mapsto \ln b \times b^y$:

$$\mathrm{vol}(b^{\mathrm{d}y}) = |\ln b| b^y\,\mathrm{vol}(\mathrm{d}y). \tag{BE}$$

Dès lors, la densité de la loi $\mathrm{Expon}^{\mathrm{le}}(\lambda)$ nous donne la probabilité recherchée ci-dessus : (notons que l'indicatrice n'apparait pas ici, car b^y est automatiquement positif),

$$\mathbb{P}(\log_{b\,*}\mathrm{Expon}^{\mathrm{le}}(\lambda) \in \mathrm{d}y) = \lambda \mathrm{e}^{-\lambda b^y}\,\mathrm{vol}(b^{\mathrm{d}y}) = |\ln b|\lambda b^y \mathrm{e}^{-\lambda b^y}\,\mathrm{vol}(\mathrm{d}y), \tag{BF}$$

ce qui nous donne bien la densité recherchée puisque le membre de droite est un multiple explicite de $\mathrm{vol}(\mathrm{d}y)$. ♧

Dans l'exemple suivant, nous allons regarder comment on peut procéder lorsqu'on est amené à considérer une mesure-image par une fonction *non* injective. L'idée, instanciée ci-dessous, est de séparer (localement) le changement de variables selon plusieurs sous-domaines pour nous ramener à des changements de variables (localement) bijectifs.

Exemple (BG) (Carré d'une loi normale décentrée). Dans cet exemple nous allons considérer le cas d'une mesure-image par une fonction *non* injective : notre objectif sera de déterminer la densité de la mesure-image de la loi Normale$(1,1)$ par la fonction « carré », sachant que la loi Normale$(1,1)$ a pour densité :

$$\mathbb{P}\big(\mathrm{Normale}(1,1) \in \mathrm{d}x\big) = \tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-(x-1)^2/2}\,\mathrm{vol}(\mathrm{d}x). \tag{BH}$$

Un carré étant toujours positif, il est évident que la loi Normale$(1,1)^2$ est portée par $\mathbb{R}_+$; en outre, la probabilité que Normale$(1,1)^2$ vaille 0 est nulle, puisque $\mathbb{P}\big(\mathrm{Normale}(1,1) = 0\big)$ (dans la mesure où la loi Normale$(1,1)$ est à densité et donc diffuse) : on peut donc s'intéresser à la mesure-image uniquement sur $\mathbb{R}_+^*$. Soit donc $y \in \mathbb{R}_+^*$ et $\mathrm{d}y$ un voisinage infinitésimal de y. Nous cherchons $\mathbb{P}(X^2 \in \mathrm{d}y)$, où X suit la loi Normale$(1,1)$. Or dire que $X^2 \in \mathrm{d}y$ est équivalent à dire que $X \in \sqrt{\mathrm{d}y} \cup -\sqrt{\mathrm{d}y}$,

où $\sqrt{\mathrm{d}y}$ et $-\sqrt{\mathrm{d}y}$ sont des voisinages despectifs de $\sqrt{y}$ et $-\sqrt{y}$: ces voisinages étant disjoints, on a donc $\mathbb{P}(X \in \sqrt{\mathrm{d}y} \cup -\sqrt{\mathrm{d}y}) = \mathbb{P}(X \in \sqrt{\mathrm{d}y}) + \mathbb{P}(X \in -\sqrt{\mathrm{d}y})$. Puisque la dérivée de la fonction $y \mapsto \sqrt{y}$ est $y \mapsto 1/2\sqrt{y}$, on a par la formule de changement de variable que $\mathrm{vol}(\sqrt{\mathrm{d}y}) = \mathrm{vol}(\mathrm{d}y)/2\sqrt{y}$; et puisque la dérivée de la fonction $y \mapsto -\sqrt{y}$ est $y \mapsto -1/2\sqrt{y}$ (qui est négative, donc la valeur absolue donnera un changement de signe), on a $\mathrm{vol}(-\sqrt{\mathrm{d}y}) = \mathrm{vol}(\mathrm{d}y)/2\sqrt{y}$ (c'est le même volume que pour $\sqrt{\mathrm{d}y}$, ce qui est logique dans la mesure où la réflexion par rapport à 0 est une isométrie et conserve donc le volume). Finalement en appliquant la densité de la loi Normale(1, 1), on trouve que

$$\begin{aligned}\mathbb{P}\big(\mathrm{Normale}(1,1)^2 \in \mathrm{d}y\big) &= \tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-(\sqrt{y}-1)^2/2}\frac{\mathrm{vol}(\mathrm{d}y)}{2\sqrt{y}} + \tfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-(-\sqrt{y}-1)^2/2}\frac{\mathrm{vol}(\mathrm{d}y)}{2\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi\mathrm{e}}}\cosh(\sqrt{y})y^{-1/2}\mathrm{e}^{-y/2}\,\mathrm{vol}(\mathrm{d}y),\end{aligned}$$

ce qui nous donne donc la densité de la loi Normale$(1,1)^2$ (en gardant à l'esprit que cette densité sera nulle pour $y \leqslant 0$). ♧

Maintenant nous allons regarder un cas où la mesure-image vit dans un espace de dimension strictement plus petit que la mesure de départ, ce qui pose à priori une difficulté, vu que la formule de changement de variables requiert d'avoir la même dimension des deux côtés. L'idée dans ce cas est d'introduire une ou plusieurs coordonnées supplémentaires "artificielles" à l'arrivée afin de rendre les dimensions égales, puis d'intégrer selon ces coordonnées supplémentaires : voir ci-dessous.

Exemple (BI). Supposons que nous ayons deux variables aléatoires X et Y qui soient indépendantes et que la loi de X aussi bien que la loi de Y soit $\mathrm{Unif}^{\mathrm{me}}(0, 1)$ [††] ; notre objectif est de déterminer la loi du produit $XY =: Z$.

Pour commencer, déterminons la densité de la loi jointe de $(X, Y) =: M$. À partir de la densité de la loi uniforme et de la propriété d'indépendance, on écrit que, pour $\mathrm{d}x$ un voisinage de x et $\mathrm{d}y$ un voisinage de y :

$$\begin{aligned}\mathbb{P}(X \in \mathrm{d}x \text{ et } Y \in \mathrm{d}y) &= \mathbb{P}(X \in \mathrm{d}x)\,\mathbb{P}(Y \in \mathrm{d}y) \\ &= \mathbf{1}_{x \in\,]0,1[}\,\mathrm{vol}(\mathrm{d}x)\mathbf{1}_{y \in\,]0,1[}\,\mathrm{vol}(\mathrm{d}y) = \mathbf{1}_{(x,y) \in\,]0,1[^2}\,\mathrm{vol}(\mathrm{d}x)\,\mathrm{vol}(\mathrm{d}y).\end{aligned}$$

Or dire que ($X \in \mathrm{d}x$ et $Y \in \mathrm{d}y$) signifie que $M \in \mathrm{d}x \times \mathrm{d}y$, où $\mathrm{d}x \times \mathrm{d}y$ est un voisinage infinitésimal (de forme produit) du point $(x, y) =: m$, voisinsage que nous pouvons donner $\mathrm{d}m$, et dont le volume est (vu que la mesure de Lebesgue multi-dimensionnelle est un produit de mesures de Lebesgue 1-dimensionnelles) $\mathrm{vol}(\mathrm{d}m) = \mathrm{vol}(\mathrm{d}x)\,\mathrm{vol}(\mathrm{d}y)$. Autrement dit ce que nous avons montré tout à l'heure est que, pour $\mathrm{d}m$ un voisinage de m de forme rectangulaire, on a

$$\mathbb{P}(M \in \mathrm{d}m) = \mathbf{1}_{m \in\,]0,1[^2}\,\mathrm{vol}(\mathrm{d}m) : \qquad (\mathtt{BJ})$$

et puisque nous avons une formule exprimant la densité pour les voisinages de forme rectangulaire, cette formule est également valable pour les voisinages de toute forme,

[††]. Comme nous avons précisé que X et Y étaient indépendantes, cette connaissance des lois individuelles de X et Y permet de déterminer complètement la loi jointe de (X, Y) : *confer* théorème (FE).

de sorte que nous avons identifié la densité de la loi $\boldsymbol{M}$ sur $\mathbb{R}^2$, qui est $\mathbf{1}_{\bullet \in]0,1[^2}$. (En fait, nous avons redémontré une propriété bien connue : la densité d'une mesure-produit est le produit des densités).

Maintenant, notons $\boldsymbol{P} \coloneqq (\boldsymbol{XY}, \boldsymbol{Y})$: nous avons ajouté $\boldsymbol{Y}$ comme seconde coordonnée pour que $\boldsymbol{P}$ vive en dimension 2 comme $\boldsymbol{M}$. Il est alors clair que $\boldsymbol{P}$ est la mesure-image de $\boldsymbol{M}$ par l'application $\boldsymbol{\Phi} : (x, y) \mapsto (xy, y)$. L'application $\boldsymbol{\Phi}$ réalise une bijection de $(\mathbb{R}_+^*)^2$ dans $(\mathbb{R}_+^*)^2$, avec $\boldsymbol{\Phi}^{-1} : (z, y) \mapsto (z \,/\, y, y)$. Nous pouvons calculer le jacobien de $\boldsymbol{\Phi}^{-1}$ (en notant $x \coloneqq z \,/\, y$) :

$$\begin{vmatrix} \partial x/\partial z & \partial x/\partial y \\ \partial y/\partial z & \partial y/\partial y \end{vmatrix} = \begin{vmatrix} 1/y & -z \,/\, y^2 \\ 0 & 1 \end{vmatrix} = 1/y. \tag{BK}$$

Dès lors nous pouvons écrire, pour $(x, y) =: p \in (\mathbb{R}_+^*)^2$, notant $\mathrm{d}p$ un voisinage infinitésimal de p :

$$\mathbb{P}(\boldsymbol{P} \in \mathrm{d}p) = \mathbb{P}(\boldsymbol{\Phi}(\boldsymbol{M}) \in \mathrm{d}p) = \mathbb{P}\big(\boldsymbol{M} \in \boldsymbol{\Phi}^{-1}(\mathrm{d}p)\big). \tag{BL}$$

Mais $\boldsymbol{\Phi}^{-1}(\mathrm{d}p)$ est un voisinage infinitésimal de $\boldsymbol{\Phi}^{-1}(p) =: m$, dont le volume, d'après la formule du changement de variable, est $\mathrm{vol}(\mathrm{d}p) \,/\, y$ (y étant toujours positif, on n'a pas à utiliser de valeur absolue), et donc

$$\mathbb{P}(\boldsymbol{P} \in \mathrm{d}p) = \mathbf{1}_{m \in]0,1[^2} y^{-1} \,\mathrm{vol}(\mathrm{d}p) \,/\, y = \mathbf{1}_{0 < z < y < 1} y^{-1} \,\mathrm{vol}(\mathrm{d}p) \tag{BM}$$

Nous avons donc trouvé la densité de $\boldsymbol{P}$; or $\boldsymbol{Z}$ est la première composante de $\boldsymbol{P}$, donc sa loi se trouve en intégrant selon les différentes valeurs possibles pour la seconde composante : pour $z \in \mathbb{R}_+^*$ et $\mathrm{d}z$ un voisinage infinitésimal de z,

$$\begin{aligned} \mathbb{P}(\boldsymbol{Z} \in \mathrm{d}z) &= \int_{y \in \mathbb{R}_+^*} \mathbb{P}(\boldsymbol{Z} \in \mathrm{d}z \text{ et } \boldsymbol{Y} \in \mathrm{d}y) = \int_{y \in \mathbb{R}_+^*} \mathbb{P}(\boldsymbol{P} \in \mathrm{d}z \times \mathrm{d}y) \\ &= \int_{y \in \mathbb{R}_+^*} \mathbf{1}_{0 < z < y < 1} y^{-1} \,\mathrm{vol}(\mathrm{d}z \times \mathrm{d}y) = \mathbf{1}_{z > 0} \Big(\int_{y \in]z,1[} y^{-1} \,\mathrm{vol}(\mathrm{d}y) \Big) \,\mathrm{vol}(\mathrm{d}z) \\ &= \mathbf{1}_{z \in]0,1[} \Big(\int_{y=z}^{1} y^{-1} \mathrm{d}y \Big) \,\mathrm{vol}(\mathrm{d}z) = \mathbf{1}_{z \in]0,1[} \ln(1/z) \,\mathrm{vol}(\mathrm{d}z) : \end{aligned}$$

nous avons donc trouvé la densité du produit $\boldsymbol{XY}$. ♧

Changement de variables inverse

Pour finir, mentionnons un corolaire du théorème (AW) qui peut s'avérer utile dans certaines circonstances :

Corolaire (BN) (Changement de variable inverse)**.** *Soient $\boldsymbol{U}$ et $\boldsymbol{V}$ deux ouverts de $\mathbb{R}^d$ et soit $\boldsymbol{\varphi} : \boldsymbol{U} \to \boldsymbol{V}$ une bijection de classe C^1 entre $\boldsymbol{U}$ et $\boldsymbol{V}$, dont nous supposerons (ce qui est presque automatique [‡‡]) que la bijection réciproque est elle aussi de classe C^1 [∗]. Soit $x \in \boldsymbol{U}$ et notons $\boldsymbol{\varphi}(x) =: y$, de sorte que $x = \boldsymbol{\varphi}^{-1}(y)$. Alors le jacobien de $\boldsymbol{\varphi}$ en x est égal à l'inverse du jacobien de $\boldsymbol{\varphi}$ en y :*

$$\det\left(\left(\frac{\partial \varphi_i}{\partial x_j}(x)\right)\right)_{ij} = \det{}^{-1}\left(\left(\frac{\partial \varphi_i^{-1}}{\partial y_j}(y)\right)\right)_{ij}. \tag{BO}$$

♢

[‡‡]. Il faut et il suffit pour cela que le jacobien de $\boldsymbol{\varphi}$ ne s'annule nulle part

[∗]. On dit alors que $\boldsymbol{\varphi}$ est un C^1-*difféomorphisme* entre $\boldsymbol{U}$ et $\boldsymbol{V}$.

Remarque (BP). En dimension 1, on retrouve la formule pour la dérivée de la réciproque : $\varphi'(x) = 1/\big(\varphi^{-1}\big)'(y)$. ♧

Démonstration. En fait, la formule du changement de variables ne nous permet de vérifier l'égalité que modulo valeur absolue. L'idée est tout simplement de considérer un volume infinitésimal $\mathrm{d}x$ autour du point x ; la formule du changement de variables nous assure alors qu'on a

$$\mathrm{vol}(\varphi(\mathrm{d}x)) = |J_\varphi(x)|\,\mathrm{vol}(\mathrm{d}x). \tag{BQ}$$

On pose alors $y \coloneqq \varphi(x)$ et $\mathrm{d}y \coloneqq \varphi(\mathrm{d}x)$ (qui est un voisinage infinitésimal de y) : et puisque φ est une bijection, on a alors $\varphi^{-1}(\mathrm{d}y) = \mathrm{d}x$, et la formule ci-dessus devient donc

$$\mathrm{vol}(\mathrm{d}y) = |J_\varphi(x)|\,\mathrm{vol}(\varphi^{-1}(\mathrm{d}x)). \tag{BR}$$

Mais d'autre part la formule du changement de variables appliquée à φ^{-1} donne que

$$\mathrm{vol}(\varphi^{-1}(\mathrm{d}y)) = |J_{\varphi^{-1}}(y)|\,\mathrm{vol}(\mathrm{d}x), \tag{BS}$$

d'où en combinant les deux formules :

$$\mathrm{vol}(\mathrm{d}y) = |J_\varphi(x)||J_{\varphi^{-1}}(y)|\,\mathrm{vol}(\mathrm{d}y), \tag{BT}$$

dont on tire finalement que $|J_{\varphi^{-1}}(y)| = |J_\varphi(x)|^{-1}$ après simplification par $\mathrm{vol}(\mathrm{d}y)$.

En fait, il n'est pas très difficile de démontrer directement le corolaire (BN). En effet, soient i, j deux indices ; on peut calculer la dérivée de x_i par rapport à x_j en écrivant, grâce à la règle de dérivation des fonctions composées, que

$$\frac{\partial x_i}{\partial x_j} = \frac{\partial}{\partial x_j}(\varphi_i^{-1} \circ \varphi) = \sum_k \frac{\partial \varphi_i^{-1}}{\partial y_k}\frac{\partial \varphi_k}{\partial x_j}, \tag{BU}$$

où l'on reconnait l'entrée d'indices (i, j) du produit matriciel $J_{\varphi^{-1}} \times J_\varphi$ ($J_{\varphi^{-1}}$ et J_φ étant évaluées respectivement en $\varphi(x)$ et en x). Comme d'autre part il est évident que $\partial x_i/\partial x_j \equiv \mathbf{1}_{i=j}$, on en déduit finalement que les matrices $J_{\varphi^{-1}}(y)$ et $J_\varphi(x)$ sont inverses l'une de l'autre, et donc que leurs déterminants sont eux aussi inverses l'un de l'autre. ♤

Le corolaire (BN) peut s'avérer utile lorsque l'application φ^{-1} a une expression plus sympathique que l'application φ elle-même, ce qui se produit souvent lorsqu'on considère des mesures-images, puisque dans ce cas-là c'est l'application φ^{-1} qui définit la mesure-image, alors qu'on doit utiliser l'application φ pour la formule du changement de variables.

Exemple (BV). Si nous reprenons par exemple le problème de l'exemple (BI), on avait $\Phi : (x, y) \mapsto (xy, y)$. On pouvait alors observer que le jacobien de Φ est

$$\begin{vmatrix} \partial z/\partial x & \partial z/\partial y \\ \partial y/\partial x & \partial y/\partial y \end{vmatrix} = \begin{vmatrix} y & x \\ 0 & 1 \end{vmatrix} = y, \tag{BW}$$

d'où on aurait pu déduire grâce au corolaire (BN) que le jacobien de Φ^{-1} en (z, y) valait y^{-1}—ce qui est effectivement ce que nous avons trouvé en explicitant la fonction Φ^{-1}. ♧

Chapitre 3

Statistiques descriptives des distributions de probabilité

Décrire une distribution de probabilité quelconque (disons sur un espace $\mathscr{X}$) est quelque chose de compliqué, puisqu'il faudrait donner les probabilités de *tous* les $A \subseteq \mathscr{X}$... Il est donc bien utile d'introduire des quantités synthétiques qui permettent, en quelques chiffres, de se faire une idée du comportement de cette distribution. Les quantiles, l'espérance, la variance et la covariance sont de telles quantités, que nous allons définir dans ce chapitre. Notez qu'il s'agit là de quantités concernant uniquement les distributions de probabilité sur $\mathbb{R}$ (pour les quantiles, l'espérance et la variance) ou sur $\mathbb{R}^2$ (pour la covariance), ou sur un sous-ensemble de ces espaces : il faura donc, pour utiliser ces outils, considérer des variables aléatoires exprimables *numériquement*. — Mais fort heureusement, les données se présentent le plus souvent naturellement sous forme numérique ! ◡̈

3.1 Fonction de répartition

Remarque (BX). Ci-dessous nous allons présenter les fonctions de répartitions (puis les quantiles et les intervalles de fluctuation) dans le cas de distributions de probabilités portées par $\mathbb{R}$ (quitte à ce qu'elles ne prennent leurs valeurs que dans un certain sous-ensemble de $\mathbb{R}$) ; mais en fait, on peut tout aussi bien définir ces concepts en remplaçant $\mathbb{R}$ par n'importe quel ensemble totalement ordonné. Un cas d'ensemble totalement ordonné qu'on peut rencontrer pas si rarement est celui où les valeurs $-\infty$ et/ou $+\infty$ sont autorisées pour notre distribution de probabilité, qui devient alors une distribution sur $[-\infty, +\infty]$. Un autre cas est celui où on compare des éléments selon un critère discret en éventuels les exæquos selon un autre critère (de sorte qu'on obtient alors p. ex. une distribution de probabilité portant sur $\mathbb{Z} \times \mathbb{Z}$ muni de l'ordre lexicographique). Mais en fait, de telles structures d'ordres sont isomorphes à des structures d'ordre sur des sous-ensembles de $\mathbb{R}$ (par exemple, l'ordre sur $[-\infty, +\infty]$ est isomorphe à l'ordre sur un segment), de sorte qu'on peut toujours se ramener au cas de $\mathbb{R}$ malgré tout. (Il existe certes des structures d'ordre total qu'il est impossible de plonger dans $\mathbb{R}$; mais on peut démontrer qu'il n'est pas possible de faire des probabilités de façon pertinente sur de telles structures). On peut aussi formuler les propriétés des fonctions de répartition et de quantile directement dans le cadre général d'un ensemble totalement ordonné : les résultats sont essentiellement les mêmes, modulo quelques subtilités supplémentaires. ♧

Préliminaires : Limites de fonctions

Définition (BY). Si $f\colon \mathbb{R} \to \mathbb{R}$ est une fonction telle que, pour tout $x \in \mathbb{R}$, les valeurs $\lim_{x'\to x, x'<x} f(x')$ et $\lim_{x'\to x, x'>x} f(x')$ existent dans $\mathbb{R}$ (une telle fonction est qualifiée de *làglàd*, pour « limite à gauche, limite à droite »), alors on note, pour tout $x \in \mathbb{R}$:

$$f(x-) \coloneqq \lim_{x' \overset{<}{\to} x} f(x')\,; \tag{BZ}$$

$$f(x+) \coloneqq \lim_{x' \overset{>}{\to} x} f(x')\,: \tag{CA}$$

moralement, $f(x-)$ représente donc la valeur que f a "juste avant x", et $f(x+)$ représente la valeur qu'elle a "juste après x". ♡

On notera en particulier que

Proposition (CB). *Les fonctions monotones sont làglàd.* ◇

Fonction de répartition

! **Définition (CC).** Soit P une distribution de probabilité sur $\mathbb{R}$. On appelle *fonction de répartition de P* "l'"application $F\colon \mathbb{R} \to [0,1]$ définie par :

$$\begin{cases} F(x-) \coloneqq \mathbb{P}(P < x)\,; \\ F(x+) \coloneqq \mathbb{P}(P \leqslant x). \end{cases} \tag{CD}$$

♡

Remarque (CE). La définition ci-dessus sous-entend qu'il existe bel et bien une fonction làglàd F vérifiant les propriétés énoncées, ce qui est effectivement vrai (mais demanderait tout de même à être vérifié en toute rigueur). ♧

Remarque (CF). En fait, la définition ci-dessus ne caractérise pas entièrement F, car il existe plusieurs fonctions vérifiant ces limites à gauche et à droite : si on veut vraiment voir F comme une fonction, il faudra donc la "normaliser" en fixant un choix parmi les différentes possibilités compatibles. La technique la plus courante pour normaliser une fonction définie par ses limites à gauche et à droite consiste à lui imposer des conditions de continuité, soit à gauche, soit à droite. En l'occurrence, cela aboutit aux deux possibilités de normalisation suivantes :

— Si on impose en outre à F d'être continue *à droite*, cela revient à poser $F(x) \coloneqq F(x+) \underset{\text{déf}}{=} \mathbb{P}(P \leqslant x)$: c'est le choix qui est fait dans tous les ouvrages de référence, dès lors qu'on a besoin de normaliser F ;

— À l'inverse, si on imposait à F d'être continue *à gauche*, cela reviendrait à poser $F(x) \coloneqq F(x-) \underset{\text{déf}}{=} \mathbb{P}(P < x)$. Néanmoins cette normalisation n'est jamais utilisée en pratique.

Vous vous demandez peut-être pourquoi j'ai absolument tenu à définir F par ses limites à droite et à gauche, puisque je viens d'expliquer qu'il existe une convention bien établie pour définir les valeurs de F elles-mêmes. La raison est qu'en fait, dans un sens mathématique profond, la "bonne" façon de définir une fonction de répartition F est plutôt de considérer qu'il s'agit seulement d'une "quasi-fonction" définie uniquement par ses limites à gauche et à droite ; et dès lors, il est naturel de refuser de choisir une convention !

Cette façon de procéder, quoique peu banale, a plusieurs avantages. Déjà, on n'aura plus à s'embêter à retenir si le cas d'égalité doit être regroupé avec les valeurs inférieures ou avec les valeurs supérieures ᵕ̈ Mais cela va en fait bien plus loin : en s'interdisant de donner une valeur à $F(x)$, les résultats obtenus seront beaucoup plus symétriques (et donc plus élégants et faciles à retenir) que si on avait fixé une convention arbitrairement ; et aussi, la vision de la fonction de quantile comme « réciproque » de la fonction de répartition deviendra alors pleinement rigoureuse ! (modulo adaptation de la notion de réciproque pour les « quasi-fonctions »). ♧

Remarque (CG). Au passage, en vertu du fait que seules les limites à gauche et à droite de la fonction de répartition sont réellement bien définies, lorsqu'on trace la fonction de répartition d'une distribution de probabilité discrète, il apparait plus pertinent de *tracer* des segments verticaux là où la fonction de répartition fait des sauts : l'inverse de ce qu'on vous a toujours appris à faire... ! ♧

On a les propriétés évidentes suivantes [*] :

Proposition (CH). !

(i) Une fonction de répartition est croissante.

(ii) Si F est la fonction de répartition d'une distribution de probabilité P sur $\mathbb{R}$, on a :

$$F(x) \overset{x\searrow -\infty}{\to} 0\,; \tag{CI}$$
$$F(x) \overset{x\nearrow +\infty}{\to} 1. \qquad \diamond$$

(iii) Si P est une distribution à densité, ou plus généralement, si P ne donne de masse à aucun singleton (on dit alors que P est diffuse *; dans le cas contraire, les x tels que $\mathbb{P}(P = x) > 0$ sont appelés des* atomes *de P), la fonction de répartition de P est continue.*

3.2 Quantiles et intervalles de fluctuation

Quantiles

Les quantiles sont en quelque sorte le concept réciproque de la fonction de répartition :

Définition (CJ). Soit P une distribution de probabilité sur $\mathbb{R}$, ayant pour fonction de répartition F. On définit la fonction de quantile (notée ici $Q(\bullet)$) de la distribution P comme "la" fonction làglàd ayant les limites à gauche et à droite suivantes : pour $p \in]0,1[$, ! !

$$\begin{cases} Q(p-) \coloneqq \sup\{x \in \mathbb{R} \mid F(x) < p\}\,; \\ Q(p+) \coloneqq \inf\{x \in \mathbb{R} \mid F(x) > p\}. \end{cases} \tag{CK}$$

En pratique, dans la quasi-totalité des cas rencontrés, $Q(p-)$ et $Q(p+)$ coïncideront [†] : on notera alors $Q(p)$ leur valeur commune, sans se poser plus de questions. ♡

[*]. Pour donner un sens rigoureux à la proposition (CH), il faudrait que je définisse précisément ce qu'on entend par les notions de croissance, de limites ou la continuité pour une « quasi-fonction » définie uniquement par ses limites à gauche et à droite, ce qui serait assez fastidieux... En fait, il vous suffit de retenir que le propriétés de la proposition (CH) sont vraies aussi bien pour la version continue à droite de la fonction de répartition que pour sa version continue à gauche — et plus généralement, pour tout choix de normalisation respectant la propriété de croissance.

[†]. Pour les fonctions de répartition, le fait que $F(x-)$ et $F(x+)$ puissent être différentes était vraiment important en pratique, car il se produit systématiquement dès lors que F est une distribution de probabilité discrète et que x correspond à une valeur possible de P. Mais pour les fonctions de quantile, il faudrait choisir p "exprès" à certaines valeurs très précises (dépendant de P) pour observer des valeurs distinctes pour $Q(p-)$ et $Q(p+)$: or en général les valeurs de p sont fixées de façon indépendante de la distribution P; et dès lors ces cas particuliers n'ont aucune raison de se produire.

Remarque (CL). Dans la définition ci-dessus, je n'ai pas précisé quelle normalisation de la fonction de répartition je prenais : c'est parce que ce choix n'a aucune incidence [‡] sur la définition de la fonction de quantile. ♣

Remarque (CM). Si Q est la fonction de quantile d'une distribution de probabilité portée par $\mathbb{R}$, $Q(p-)$ et $Q(p+)$ sont toujours finis pour $p \in]0,1[$. En fait, on pourrait aussi définir les quantiles pour $p = 0$ et $p = 1$: dans ce cas, les quantiles de rangs $0-$ et $1+$ valent systématiquement resp. $-\infty$ et $+\infty$, tandis que les quantiles de rangs respectifs $0+$ et $1-$ correspondent respectivement au minimum et au maximum du support de P, c'est-à-dire de la plage où P prend effectivement ses valeurs [§]. Par exemple, pour une distribution exponentielle, les quantiles de rang $0+$ et $1-$ sont respectivement 0 et $+\infty$; pour une distribution uniforme sur $]a,b[$, ce sont resp. a et b ; et pour une distribution normale (de variance non nulle), ce sont resp. $-\infty$ et $+\infty$. ♣

! *Remarque* (CN). Informellement, le quantile de rang p est l'abscisse qui sépare la masse de P entre une probabilité p à gauche, et $1-p$ à droite : plus précisément, le quantile de rang p est le moment où s'effectue la transition entre le moment où on avait une probabilité inférieure à p de tomber à gauche de x, et celui où on se met à avoir une probabilité supérieure à p de tomber à gauche de x. ♣

Remarque (CO). Dans le contexte de la remarque ci-dessus, les notions de quantiles de rang $p-$ et $p+$ correspondent aux cas exceptionnels où il existe toute une plage de valeurs de x pour lequelles on a une probabilité exactement p de tomber à gauche et une probabilité exactement $1-p$ de tomber à droite : dans ce cas, l'extrêmité gauche de cette plage correspond au quantile de rang $p-$, tandis que l'extrêmité droite de la plage correspond au quantile de rang $p+$. ♣

Remarque (CP). Contrairement au cas des fonctions de répartition, où il existe une façon conventionnelle de définir $F(x)$ lorsque $F(x-)$ et $F(x+)$ diffèrent, il n'existe pas vraiment de convention universelle pour définir $Q(p)$ lorsque $Q(p-)$ et $Q(p+)$ diffèrent... Le choix le plus souvent fait est celui consistant à imposer à Q d'être continue à gauche, ce qui revient à poser $Q(p) := Q(p-)$. Mais une autre convention courante est de poser $Q(p) := (Q(p-) + Q(p+)) / 2$; et il y a aussi des auteurs qui définissent $Q(p) := Q(p-)$, *sauf* pour $p = 0$ où ils posent $Q(0) := Q(0+)$: bref, c'est un peu la pagaille... En fait, chacune de ces conventions présente ses avantages et ses inconvénients : car, de même que dans le cas des fonctions de répartition, le choix mathématiquement le plus pertinent est de ne *pas* fixer de convention et de ne définir $Q(\bullet)$ qu'à travers ses limites à gauche et à droite... ! Dans la suite, c'est en substance le choix que nous ferons : plus précisément, les énoncés sur « la » fonction de quantile seront tournés de façon à être valables pour toutes les façons possibles de normaliser la fonction de quantile [¶]. ♣

! **Proposition** (CQ). *La fonction quantile est la « réciproque » de la fonction de répartition, au sens suivant : si on dessine le graphe de la fonction de répartition de P* en traçant des segments verticaux *pour les sauts éventuels de $F(\bullet)$, et qu'on échange les axes des abscisses et des ordonnées, alors on obtient le graphe de la fonction de quantile de P (là aussi, avec des segments verticaux pour les sauts éventuels de $Q(\bullet)$).* ♢

[‡]. Du moins sous réserve que la normalisation de $F(\bullet)$ soit prise croissante, ce qu'on supposera toujours.

[§]. Techniquement parlant, le *support* d'une distribution de probabilité sur un espace $\mathcal{X}$ muni d'une topologie (polonaise), noté $\operatorname{supp} P$, est défini comme le plus petit (au sens de l'inclusion) sous-ensemble fermé de $\mathcal{X}$ dont la probabilité vaille 1 : on peut démontrer qu'un tel plus petit fermé existe bel et bien.

[¶]. Plus précisément, pour toutes les normalisations aboutissant à une fonction de quantile croissante : mais ce sera le cas pour tout choix raisonnable, incluant en particulier les différentes conventions mentionnées plus haut.

Proposition (CR). *La fonction de quantile est toujours croissante, et elle est strictement croissante si et seulement si la distribution P est diffuse (i.e. n'accorde de masse non nulle à aucun point). Plus généralement, les plateaux de la fonction de quantile correspondent aux sauts de la fonction de répartition, i.e. aux atomes de P, tandis que lessauts de Q correspondent aux plateaux de la fonction de répartition, c.-à-d. aux intervalles d'intérieur non vide auxquels P n'attribue aucune masse.* ♢

Proposition (CS). *Soit P une distribution de probabilité sur $\bar{\mathbb{R}}$ ayant resp. pour fonction de répartition F et pour fonction de quantile Q. Alors, pour tous $x \in \bar{\mathbb{R}}$, $p \in [0,1]$, on a les équivalences suivantes :*

$$F(x+) < p \iff x < Q(p-)\,; \tag{CT}$$
$$F(x-) \leqslant p \iff x \leqslant Q(p+). \tag{CU}$$

♢

Remarque (CV). Bien que j'aie indiqué cette proposition comme étant « à retenir », il vaut beaucoup mieux se contenter de retenir par cœur qu'il existe des relations qui expriment, moralement, le fait que se trouver avant le quantile de rang p est synonyme d'avoir une fonction de répartition plus petite que p (ce qui est justement l'idée derrière la définition des quantiles !), et s'aider de dessins pour retrouver un énoncé précis correct lorsqu'on a besoin de mettre des '+' et des '−' et/ou de savoir si les inégalités sont larges ou strictes... ♧

> *Remarque* (CW). On peut aussi essayer de retenir la proposition de la façon suivante : la première équivalence correspond aux cas où x est "carrément" avant le quantile de rang p : dans ce cas, on le compare au quantile de rang $p-$ (qui sera plus petit, en général, que le quantile de rang $p+$), et on impose que l'inégalité soit stricte ; et du côté des fonctions de répartition, on demande là encore que l'inégalité soit stricte, et on considère la valeur en $x+$ (qui est, en général, plus grande qu'en $x-$). Alors qu'à l'inverse, la seconde équivalence correspond à avoir un maximum de "tolérance" quant à l'affirmation de x est situé avant le quantile de rang p, resp. que la fonction de répartition en x est plus petite que p... ♧

Définition (CX) (q-Quantiles, Médiane). Pour d entier et $n \in]\!]0, q[\![$, ce qu'on appelle le « n-ième d-quantile » de la distribution P est son quantile de rang n/d. Pour d égal resp. à 3, 4, 5, 10 et 100, les d-quantiles sont appelés resp. « terciles », « quartiles », « quintiles », « déciles » et « centiles ». Pour d égal à 2, le quantile de rang 1/2 est appelé *médiane* de la distribution P. ♡

Théorème (CY). *Soient P une distribution de probabilité sur $\mathbb{R}$ et $p \in]0,1[$ [‖] ; et notons x le quantile de rang p de P. Soit $f\colon \mathbb{R} \to \mathbb{R}$ une fonction continue strictement croissante (resp. strictement décroissante). Alors le quantile de rang p (resp. de rang $(1-p)$ [**]) de $f_* P$ est égal à $f(x)$.* ♢

> *Remarque* (CZ). Dans certains cas, il peut être plus commode de "renverser l'ordre sur $\mathbb{R}$" : on définit alors la *fonction de répartition complémentaire* comme la probabilité d'être plus *grand* qu'une certaine valeur, resp. la *fonction de quantile complémentaire* comme la fonction qui, à p, associe un point séparant une probabilité p *à sa droite* d'une probabilité $1-p$ *à sa gauche*. Si nous notons $\check{F}(\bullet)$ et $\check{Q}(\bullet)$ ces fonctions complémentaires respectives, on a tout simplement $\check{F}(x) = 1 - F(x)$, resp. $\check{Q}(p) = \check{Q}(1-p)$. ♧

[‖]. Ici on suppose que p est en fait de la forme « $a-$ » ou « $a+$ ».
[**]. En posant $1-(a-) =: (1-a)+$ et $(1-(a+)) = (1-a)-$.

Intervalles de fluctuation

Définition (DA). Pour P une distribution de probabilité sur (un sous-ensemble de) $\mathbb{R}$, $\alpha \in]0,1[$, un intervalle $I \subseteq \mathbb{R}$ est qualifié d'*intervalle de fluctuation* au niveau de risque α (ou « au niveau de confiance $1-\alpha$ ») lorsque P donne une masse au moins égale à $1-\alpha$ à l'intervalle I :

$$\mathbb{P}(P \notin I) \leqslant \alpha. \tag{DB}$$

♡

Remarque (DC). Pour les applications pratiques, le niveau de risque α sera toujours pris $\leqslant 1/2$ [††]. C'est pourquoi il arrivera souvent qu'on dise simplement « au niveau *[tant]* », sans prendre la peine de préciser si on est en train de parler d'un niveau *de risque* ou d'un niveau *de confiance* : il suffira en effet de comparer la valeur du niveau à 1/2 pour trancher ! Par exemple, un « niveau 10 % » se réfèrera implicitement à un niveau de risque (soit $\alpha = 0{,}1$ en l'occurrence), tandis qu'un « niveau 95 % » se réfèrera à un niveau de confiance (soit $\alpha = 0{,}05$ en l'occurrence). (Et dans le cas particulier d'un niveau 1/2, « risque à 50 % » et « confiance à 50 % » sont en fait synonymes ! ☺). ♣

> *Remarque* (DD). La définition pourrait aussi s'appliquer sans exiger que I soit un intervalle : dans un tel cas, on parlerait plutôt de « zone » de fluctuation. En particulier, le concept de zone de fluctuation garde du sens même lorsque l'ensemble mesurable sur lequel P porte n'est pas totalement ordonné. Mais dans ce cours nous nous limiterons aux intervalles sur $\mathbb{R}$, qui sont à la fois le cas le plus important et le plus fréquemment rencontré. ♣

Lemme (DE). *Un intervalle de la forme* $]-\infty, q]$ *est un intervalle de fluctuation au niveau* α *pour* P *si et seulement si* $\text{Répart}(P; q+) \geqslant 1-\alpha$ *; ou, de manière équivalente, si et seulement si* $q \geqslant \text{Qtile}\big(P; (1-\alpha)-\big)$. *Par conséquent, le plus petit intervalle de fluctuation au niveau* α *pour* P *ayant cette forme (il y en a bien un) est l'intervalle* $\big]-\infty, \text{Qtile}\big(P; (1-\alpha)-\big)\big]$.

De même, le plus petit intervalle de la forme $[q, \infty[$ *qui soit un intervalle de fluctuation au niveau* α *pour* P *est l'intervalle* $[\text{Qtile}(P; \alpha+), +\infty[$. ♢

Le lemme (DE) ci-dessus suggère la définition suivante :

! **Définition (DF).** Pour P une distribution de probabilité sur $\mathbb{R}$, $\alpha \in]0, 1/2]$, on parle de « l' »intervalle de fluctuation au niveau de risque α (ou au niveau de confiance $1-\alpha$) de la distribution P pour désigner l'intervalle fermé compris entre ses quantiles de niveaux respectifs $(\alpha/2)+$ et $(1-\alpha/2)-$ [‡‡], autrement dit, l'intervalle

$$[\text{Qtile}(P; \alpha\,/\,2), \text{Qtile}(P; 1-\alpha\,/\,2)]. \tag{DG}$$

En vertu du lemme (DE), cet intervalle constitue effectivement un intervalle de fluctuation : on le qualifie de « l' »intervalle (même si ce n'est pas le seul) dans la mesure où sa construction est dans un sens "canonique". ♡

[††]. Il y aura une seule exception dans ce cours, qui sera lorsque nous introduirons le concept de *p-valeur* au chapitre 13 : et encore, l'utilisation de niveaux α plus grands que 1/2 à ce moment-là servira uniquement à *expliquer* ce que signifie la *p*-valeur ; mais dans tous les cas, vous ne serez jamais amenés à parler d'« intervalle à 75 % de risque » ! ☊

[‡‡]. En pratique, on ne s'ennuiera pas avec les fioritures '+' et '−', confer la remarque figurant à la fin de la définition (CJ).

Remarque (DH). Pour que la formule ci-dessus soit toujours valide, il est très important de prendre des bornes *fermées* à l'intervalle : en effet, dès lors que P est une mesure discrète (ou, de manière générale, qu'il y a une masse non nulle aux bornes de l'intervalle de fluctuation canonique), la formule devient le plus souvent fausse... ! C'est pourquoi, de manière générale, nous écrirons systématiquement les intervalles de fluctuation (et aussi, du même coup, les intervalles de confiance et de prédiction) sous forme fermée. ♧

Remarque (DI). En fait, du point de vue mathématique, la construction la plus pertinente pour construire un intervalle de fluctuation canonique consiste à définir « l' »intervalle de fluctuation (au niveau de risque α) pour P comme l'ensemble des x tels que

$$\alpha/2 < \text{Répart}(P; x+) \quad \text{et} \quad \text{Répart}(P; x-) < 1 - \alpha/2. \tag{DJ}$$

Cette construction, que je qualifierai au besoin d'« intervalle canonique *strict* », donne presque le même intervalle que la construction de la définition (DF), à ceci près que l'intervalle qu'elle fournit pourra, dans certains cas, être ouvert au niveau d'une ou deux de ses bornes... Du point de vue pratique, néanmoins, ce raffinement n'apporterait essentiellement rien : en effet, on pourrait montrer que, dans les cas où l'intervalle strict diffère de l'intervalle "fermé" de la définition (DF), la probabilité que la loi P tombe précisément au niveau d'un des points qui diffèrent entre ces deux intervalles... est nulle ! Dès lors, dans la mesure où le seul véritable intérêt de la construction de l'intervalle de flucutation strict est de rendre certains résultats mathématiques plus rigoureux, autant ne pas s'encombrer avec cette subtilité, et en rester à la plus simple définition (DF) : et ce, d'autant plus que prendre l'habitude d'écrire tous les intervalles de flcutuation sous forme fermée évitera de risquer de commettre certaines erreurs fâcheuses ! ◡̈ (cf. remarque (DH) ci-dessus). ♧

Remarque (DK). Notez que je n'ai pas affirmé ci-dessus que l'intervalle de fluctuation canonique serait minimal parmi les intervalles (fermés) de fluctuation de niveau α pour P, car ce n'est pas vrai en général. Par exemple, si P est la distribution de probabilité uniforme sur $\{-1, 0, +1\}$, alors l'intervalle de fluctuation (canonique) de niveau 40 % pour P est $[-1, +1]$, alors que pourtant il existe d'autres intervalles de fluctuation strictement plus petits à ce même niveau : notamment $[-1, 0]$ et $[0, +1]$. ♧

Remarque (DL). On peut se demander ce qui se passe si l'on prend $\alpha = 0$ dans la définition des intervalles de fluctuation, autrement dit, qu'on souhaite trouver un intervalle contenant *toute* la masse de P : dans ce cas-là, plutôt que de parler d'un « intervalle de fluctuation à 100 % de confiance », on dit que l'intervalle I « supporte » la distribution P. En fait, dans la cas $\alpha = 0$, toutes les formules évoquées restent parfaitement valables, avec néanmoins la possibilité que certains quantiles deviennent infinis. Ce cas n'est néanmoins pas très intéressant du point de vue de la statistique, puisque la philosophie même de la statistique est qu'on renonce à avoir des certitudes *absolues* en vue d'obtenir des informations plus intéressantes...

Il existe toujours un plus petit intervalle (non nécessairement fermé) supportant la mesure P, qu'on appelle parfois l'*étendue* de P.

Plus intéressant, si on considère plus généralement les *zones* (pas forcément des intervalles) supportant P, on peut démontrer qu'il existe toujours un plus petite zone (au sens de l'inclusion) *fermée* supportant P, qu'on appelle alors le *support* de P. Le support de P a ceci d'intéressant qu'on peut continuer à le définir y compris lorsque la mesure P porte sur un autre espace que $\mathbb{R}$: il suffit juste qu'il s'agisse d'un espace muni d'une structure topologique « polonaise », c'est-à-dire équivalente (du point de vue des ouverts qu'elle définit) à une topologie métrique complète et séparable. ♧

3.3 Mode

Définition (DM) (Mode d'une distribution de probabilité discrète)**.** Pour P une distribution de probabilité sur un espace discret $\mathcal{X}$, le *mode* de P, que je noterai !

parfois Mode(P), est la valeur de x [*] pour laquelle $P(\{x\})$ est maximale : autrement dit, la valeur qu'il y a le plus de chance qu'une v.a. de loi P prenne. ♡

Proposition (DN). *Si P est une distribution de probabilité sur l'espace* discret $\mathcal{X}$ *et que $f : \mathcal{X} \to \mathcal{Y}$ est une application injective, alors*

$$\mathrm{Mode}(f_* P) = f(\mathrm{Mode}(P)) \text{ :} \tag{DO}$$

autrement dit, le mode d'une distribution discrète est préservé par la mesure-image (injective). ♢

On parle aussi de mode pour les distributions à densité, mais c'est plutôt par abus de langage :

Définition (DP) (Mode d'une distribution de probabilité à densité)**.** Pour P une distribution de probabilité à densité sur un espace $\mathbb{R}^d$, on appelle par extension « mode » de P la valeur x pour laquelle la densité $P(\mathrm{d}x)/\mathrm{vol}_d(\mathrm{d}x)$ de P par rapport à la mesure de Lebesgue est maximale. Informellement, cela correspond à la valeur pour laquelle il y a le plus de chance qu'une v.a. de loi P tombe "juste à côté". ♡

Remarque (DQ). Attention, il n'y a pas d'analogue de la proposition (DN) pour les distributions de probabilité à densité : le mode d'une distribution à densité n'est pas préservé par reparamétrisation en général ! ♣

3.4 Quantités empiriques

! **Définition (DR)** (Distribution empirique)**.** Pour $x_1, \dots, x_n$ ($n \geqslant 1$) des données à valeurs dans un espace $\mathcal{X}$, la *distribution empirique* des x_i (ou plus exactement, du multi-ensemble $\{\{x_1, \dots, x_n\}\}$) est la distribution de probabilité P sur $\mathcal{X}$ définie par

$$\forall A \subseteq \mathcal{X} \qquad P(A) \coloneqq \frac{\sum_{i=1}^n \mathbf{1}_{x_i \in A}}{n} ; \tag{DS}$$

ou autrement dit,

$$P \coloneqq \sum_{i=1}^n \delta_{x_i}. \tag{DT}$$

♡

Remarque (DU). Il existe également une version *pondérée* des la notion de distribution empirique : si la i-ième donnée x_i est attachée au poids α_i, où les poids sont positifs et non identiquement nuls, la *distribution empirique pondérée* de $(x_i)_{1 \leqslant i \leqslant n}$ pour les poids en question sera la distribution de probabilité

$$\frac{\sum_{i=1}^n \alpha_i \delta_i}{\sum_{i=1}^n \alpha_i}. \tag{DV}$$

Cependant nous n'aurons pas besoin d'utiliser les versions pondérées dans le cadre de ce cours. ♣

! **Définition (DW)** (Quantité empirique)**.** Pour « schmilblick » un concept s'appliquant aux distributions de probabilité sur $\mathcal{X}$ (fonction de répartition, quantiles, espérance, variance, covariance, ...), et $x_1, \dots, x_n$ des valeurs de $\mathcal{X}$, le *schmilblick empirique* de $\{\{x_1, \dots, x_n\}\}$ est tout simplement le schmilblick de la distribution empirique de $\{\{x_1, \dots, x_n\}\}$. Dans le cas de l'espérance, l'espérance empirique est plus simplement appelée *moyenne*. ♡

[*]. Que nous supposerons ici unique pour simplifier, ce qui est le cas en général ☺

Définition (DX). Dans ce cours, nous noterons resp. $\mathrm{moy}(x_1, \dots, x_n)$, $\mathrm{var}_{\mathrm{emp}}(x_1, \dots, x_n)$, $\mathrm{var}^{1/2}_{\mathrm{emp}}(x_1, \dots, x_n)$, $\mathrm{cov}_{\mathrm{emp}}\big((x_1, y_1), \dots, (x_n, y_n)\big)$ les moyenne, variance empirique, écart-type empirique et covariance empirique des valeurs considérées. Notez que nous utilisons des minuscules lorsque nous nous référons à des quantités empiriques, qui s'appliquent donc à un *jeu de données*, et des majuscules lorsque nous nous référons aux concepts sous-jacents, qui s'appliquent donc à une *distribution de probabilité*. ♡

Remarque (DY). Le fait que nous mentionnions un indice « emp » pour la variance, l'écart-type et la covariance empiriques sert à ne risquer aucune confusion avec les variantes de ces concepts prenant en compte la *correction de Bessel* (cf. annexe **??**). Dans ce cours, on pourra se dispenser de cet indice, qui sera sous-entendu par défaut ; mais il faut bien garder à l'esprit que d'autres sources, et le logiciel *R* en particulier, considèrent au contraire que le concept par défaut est celui incluant la correction de Bessel. ♧

Proposition (DZ). *On a en particulier :*

$$\mathrm{moy}(x_1, \dots, x_n) \coloneqq \frac{x_1 + \dots + x_n}{n} \; ; \tag{EA}$$

$$\mathrm{var}_{\mathrm{emp}}(x_1, \dots, x_n) \coloneqq \mathrm{moy}(x_i - \mathrm{moy}(x_j)_{1 \leqslant j \leqslant n})_{1 \leqslant i \leqslant n}{}^{[\dagger]} \tag{EB}$$

$$= \mathrm{moy}(x_i^2)_{1 \leqslant i \leqslant n} - \big(\mathrm{moy}(x_i)_{1 \leqslant i \leqslant n}\big)^2. \tag{EC}$$

♢

Remarque (EE). Il est essentiel de bien comprendre que les variance empirique & C$^{\mathrm{ie}}$ ne sont *pas* des quantités aléatoires : $\mathrm{var}_{\mathrm{emp}}(x_1, \dots, x_n)$ est simplement une *fonction* des données[‡], tout ce qu'il y a de plus déterministe : par exemple, $\mathrm{var}^{1/2}_{\mathrm{emp}}(2, 4, 10, 4) = 3$, fin de la discussion ! Attention toutefois : le fait est qu'on considère souvent les variances empiriques dans le cas où $X_1, \dots, X_n$ sont des (réalisations de) variables aléatoires ; et dans ce cas évidemment $\mathrm{var}_{\mathrm{emp}}(X_1, \dots, X_n)$ sera une variable aléatoire, mais c'est simplement parce que c'est une fonction (déterministe) de variables aléatoires... ! ♧

3.5 Théorie de la décision optimale

Motivation

Le but de ce chapitre est d'introduire une notion probabiliste conceptuellement très importante en ingénierie : la prise de décision optimale en contexte incertain.

Concernant la notion de « prise de décision optimale », son lien avec l'ingénierie est évident : si la mission de votre équipe d'ingénieurs, par exemple, est de concevoir

[†]. Soit, explicitement :

$$\frac{1}{n} \sum_{i=1}^{n} \left(x_i - \frac{\sum_{j=1}^{n} x_j}{n} \right)^2. \tag{ED}$$

[‡]. Notons au passage qu'il y a un léger abus de langage à noter de la même façon la variance empirique quel que soit le nombre de données auquel elle s'applique : en réalité bien sûr, il y a une fonction $\mathbb{R} \to \mathbb{R}$ quand on a un jeu de 1 donnée, une fonction de $\mathbb{R}^2 \to \mathbb{R}$ quand on a un jeu de 2 données, &c.

un viaduc, vous allez devoir décider du type d'ouvrage, des dimensions de la voie, du nombre de piles ou de câbles, des matériaux utilisés, etc., pour optimiser un certain ensemble de critères ou de contraintes : cout de l'ouvrage, sécurité, capacité de la voie, durabilité, facilité d'entretien, considérations écologiques, etc. Vous avez déjà vu, dans vos cours du semestre 5 (et certains d'entre vous étudieront le sujet plus à fond ultérieurement), comment trouver le jeu de paramètres permettant d'optimiser une fonction donnée.

Cependant, dans de nombreux cas, la décision à prendre participe d'un pari sur l'avenir... C'est, en fait, une situation très fréquente dans la vie quotidienne, même lorsqu'on on n'est pas ingénieur !

Mettons, par exemple, que vous soyez une étudiante (hétérosexuelle, célibataire) éprise du séduisant Quentin. Pour avoir une chance de le convaincre de vous choisir pour partenaire, vous devez vous faire mieux connaitre de lui en lui consacrant du temps, par exemple en se rendant à cette soirée où vous savez qu'il sera présent. Si Quentin s'avère effectivement intéressé par vous, le gain potentiel est fabuleux, représentant des décennies de bonheur à deux ! Sauf que, si ça se trouve, même en vous présentant sous votre meilleur jour, Quentin n'aura pas ce "feeling" qui l'attirera vers vous... La seule chose qui dépende de vous, c'est le fait d'aller ou non à cette soirée ; mais le résultat de cette décision dépendra des préférences intimes de Quentin, que vous ne pouvez pas connaitre au moment de prendre votre décision ! Vous devez donc décider du plan de votre soirée sans savoir de quoi il en retourne : on est ainsi dans une situation de prise de décision en contexte incertain.

Formalisation du problème

Formellement, le question est de prendre une certaine décision d, à valeurs dans un espace noté $\mathscr{D}$, sachant que le résultat de cette décision dépendra aussi d'une certaine quantité G, à valeurs dans un espace noté $\mathscr{G}$, que l'on ne connait pas encore au moment de prendre la décision, et dont la valeur future sera aléatoire : l'enjeu étant que notre décision ayant des conséquences aussi « bonnes » que possible.

Pour pouvoir procéder à un traitement mathématique, nous avons donc besoin de *quantifier* en quoi une conséquence peut s'avérer plus ou moins « bonne » ! L'outil approprié pour ce faire est celui de *fonction de perte* :

Définition (EF). Dans le contexte de ce chapitre, une *fonction de perte* est une application $\ell : \mathscr{G} \times \mathscr{D} \to \mathbb{R}$ dont l'interprétation est la suivante :

(i) La valeur $\ell(g, \hat{d})$ nous dit à quel point est-ce que la décision $\hat{d}$ amènera des conséquences nuisibles si la quantité d'intérêt s'avère valoir g : plus cette valeur est petite, moins les conséquences sont nuisibles à nos yeux ;

(ii) En outre, entre quatre valeurs de pertes possibles $\ell_0, \ell'_0, \ell_1, \ell'_1$ avec $\ell_0 < \ell'_0$ et $\ell_1 < \ell'_1$, le ratio $(\ell'_1 - \ell_1)/(\ell'_0 - \ell_0)$ nous dit *de combien de fois pire* est le fait de passer de la situation de perte ℓ_1 à la situation de perte ℓ'_1, par rapport au fait de passer de la situation de perte ℓ_0 à la situation de perte ℓ'_0. ♡

Remarque (EG). L'opposé d'une fonction de perte est qualifié de *fonction d'utilité* (traditionnellement notée 'u') : c'est exactement le même principe, sauf que cette fois-ci nous avons une préférence pour les valeurs les plus *grandes* ! Les économistes travaillent plus volontiers avec des fonctions d'utilité ; mais dans le cadre de la statistique, il est préférable d'utiliser des fonctions de perte. ♣

Remarque (EH). Le choix d'une fonction de perte comprend une composante fondamentalement *subjective*. En effet, la notion de « conséquence nuisible » inclut la prise en compte de plusieurs aspects qui ne sont pas objectivement commensurables : considérations financières, aspects éthiques (conditions de travail des employés ; impact sur la santé des consommateurs ; possibilités de nuisances démocratiques ; ...), dégradation de l'environnement, ... La fonction de perte intégrant tous ces aspects au sein d'un même nombre, la pondération de ces différents aspects reflètera de l'échelle de valeur de celui qui choisira la fonction de perte ! ♧

Décision optimale

Une fois le contexte formalisé, la notion de décision optimale en contexte incertain devient très simple :

Principe (EI). *Dans un contexte où la quantité* $\boldsymbol{G}$*, inconnue au moment de prendre la décision, sera réalisé selon une loi aléatoire* $\boldsymbol{P}$*, une décision* $\boldsymbol{d}$ *est d'autant meilleure que l'espérance*

$$\mathbb{E}^{g\sim P}\big(\ell(d,g)\big) \tag{EJ}$$

est petite. ◇

!

Chapitre 4

Conditionnement et indépendance

4.1 Conditionnement

Définition (EK) (Conditionnement par rapport à un évènement de probabilité non nulle)**.** Si A est un évènement de probabilité non nulle, la probabilité *conditionnée à la réalisation de* A, notée $\mathbb{P}(\bullet \mid A)$, est la distribution de probabilité sur Ω telle que pour tout évènement $B \subseteq \Omega$:

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)}. \tag{EL}$$

(Théorème : il s'agit bien d'une distribution de probabilité). Les espérance, variance, loi etc. sous cette probabilité conditionnelle sont alors notées resp. $\mathbb{E}(\bullet \mid A)$, $\mathrm{Var}(\bullet \mid A)$, $\mathrm{Loi}(\bullet \mid A)$, etc. ♡

Remarque (EM). Intuitivement, $\mathbb{P}(\bullet \mid A)$ décrit comment se répartissent *entre elles* les différentes possibilités pour lesquelles A est réalisé. ♧

Définition (EN) (Conditionnement par rapport à un évènement de probabilité nulle)**.** Lorsque A est un évènement de probabilité nulle, on peut, en général, définir *quand même* $\mathbb{P}(\bullet \mid A)$ par passage à la limite [*]. Typiquement, si X est une variable aléatoire à valeurs dans $\mathbb{R}^d$ et x un élément de $\mathbb{R}^d$, on assimilera

$$\mathbb{P}(B \mid X = x) \tag{EO}$$

à

$$\mathbb{P}(B \mid X \in \mathrm{d}x), \tag{EP}$$

pour $\mathrm{d}x$ un voisinage infinitésimal de x. Sous certaines hypothèses de régularité dont nous ne préoccuperons pas ici, on peut montrer que cette distribution de probabilité a bien une limite lorsqu'on fait tendre l'intervalle infinitésimal $\mathrm{d}x$ vers le singleton $\{x\}$, et que cette limite est bien une distribution de probabilité sur Ω ; en outre, sous des hypothèses assez générales, cette distribution de probabilité conditionnelle obtenue ainsi par passage à la limite peut être définie pour (presque-)toute valeur de x. ♡

Théorème (EQ). *Soient X et Y deux variables aléatoires resp. à valeurs dans $\mathbb{R}^n$ et $\mathbb{R}^m$, dont la loi jointe a une densité $f(x,y)$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^{n+m}$ (x désignant un point de $\mathbb{R}^n$ et y un point de $\mathbb{R}^m$). Alors la loi de Y sachant « $X = x_0$ » est à densité aussi, et cette densité est donnée par :*

$$\mathbb{P}(Y \in \mathrm{d}y \mid X = x_0) = \frac{f(x_0, y)}{\int_{y' \in \mathbb{R}^m} f(x_0, y') \operatorname{vol}_m(\mathrm{d}y')} \operatorname{vol}_m(\mathrm{d}y). \tag{ER}$$

◇

Démonstration. Pour $\mathrm{d}x_0$ un voisinage infinitésimal de x_0 dans $\mathbb{R}^n$, on écrit que $\mathbb{P}(Y \in \mathrm{d}y \mid X = x_0)$ est formellement la même chose que $\mathbb{P}(Y \in \mathrm{d}y \mid X \in \mathrm{d}x_0)$, qui vaut par définition de la probabilité conditionnelle :

$$\frac{\mathbb{P}(Y \in \mathrm{d}y \text{ et } X \in \mathrm{d}x_0)}{\mathbb{P}(X \in \mathrm{d}x_0)}. \tag{ES}$$

Notons que « $Y \in \mathrm{d}y$ et $X \in \mathrm{d}x_0$ » peut se réécrire comme « $(X,Y) \in \mathrm{d}x_0 \times \mathrm{d}y$ », où $\mathrm{d}x_0 \times \mathrm{d}y$ est un voisinage infinitésimal de (x_0, y) dans $\mathbb{R}^{n+m}$. Par conséquent, par définition de la densité, le numérateur de notre fraction vaut

$$f(x_0, y) \operatorname{vol}(\mathrm{d}x_0 \times \mathrm{d}y), \tag{ET}$$

où $\operatorname{vol}(\mathrm{d}x_0 \times \mathrm{d}y) = \operatorname{vol}(\mathrm{d}x_0) \times \operatorname{vol}(\mathrm{d}y)$. Concernant le dénominateur, partitionnons $\mathbb{R}^m$ en un très grand nombre de zones de dimensions infinitésimales, de la forme $\mathrm{d}y'$, où y' décrit un maillage très dense de $\mathbb{R}^m$. On a alors

$$\mathrm{d}x_0 \times \mathbb{R}^m = \bigsqcup_{y'} x_0 \times \mathrm{d}y', \tag{EU}$$

d'où

$$\mathbb{P}(X \in \mathrm{d}x_0) = \sum_{y'} \mathbb{P}(X \in \mathrm{d}x_0 \text{ et } Y \in \mathrm{d}y'). \tag{EV}$$

Mais, d'après le travail que nous avons fait pour le numérateur,

$$\mathbb{P}(X \in \mathrm{d}x_0 \text{ et } Y \in \mathrm{d}y') = f(x_0, y') \operatorname{vol}(\mathrm{d}x_0) \operatorname{vol}(\mathrm{d}y'), \tag{EW}$$

d'où

$$\mathbb{P}(X \in \mathrm{d}x_0) = \operatorname{vol}(\mathrm{d}x_0) \times \Big(\sum_{y'} f(x_0, y') \operatorname{vol}(\mathrm{d}y') \Big) = \Big(\int_{y'} f(x_0, y') \mathrm{d}y' \Big) \operatorname{vol}(\mathrm{d}x_0), \tag{EX}$$

où la dernière égalité n'est autre que la définition informelle de l'intégrale (multi-dimensionnelle). Le résultat vient finalement après simplification du numérateur et du dénominateur par $\operatorname{vol}(\mathrm{d}x_0)$ (ce qui était attendu, puisque ce $\mathrm{d}x_0$ n'était qu'un artifice technique). ♤

Théorème (EY). *Soient μ une distribution de probabilité sur $\mathcal{X}$ et, pour tout $x \in \mathcal{X}$, ν_x une distribution de probabilité sur $\mathcal{Y}$. Alors il existe une unique distribution de probabilité π sur $\mathcal{X} \times \mathcal{Y}$ telle que, si (X,Y) a pour loi (jointe) π, X suit la loi μ et, pour tout $x \in \mathcal{X}$, Y suit la loi ν_x sous $\mathbb{P}(\bullet \mid X = x)$. En particulier, si $\mathcal{X} = \mathbb{R}^n$, $\mathcal{Y} = \mathbb{R}^m$, que μ est la distribution de densité $f(x)$ et que ν_x est la distribution de densité $g_x(y)$, π sera la distribution de densité $f(x)g_x(y)$ sur $\mathbb{R}^{n+m}$.* ◇

[∗]. La définition rigoureuse est beaucoup plus technique et subtile ; nous ne nous en soucierons pas ici, mais garderons juste à la tête que la définition ci-dessus pourrait poser problème en toute généralité, mais pas dans les cas que nous rencontrerons en pratique.

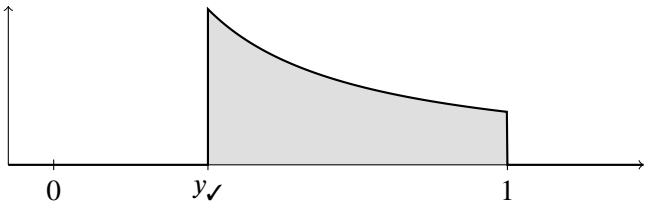

FIGURE 4.1 – Représentation de la X sachant la valeur de Y (on a pris ici $y_{\checkmark} = 0{,}34$) dans l'exemple (EZ).

Exemple (EZ). À titre d'exemple, considérons la situation suivante. On tire une variable aléatoire $X(\omega)$ uniformément entre 0 et 1, puis on tire une variable Y uniformément entre 0 et $X(\omega)$. La première question est, quelle est la loi jointe du couple (X, Y) ? Puisque nous avons décrit la loi de Y *une fois connue la valeur de* X, cela revient à dire qu'on connait la loi conditionnelle de Y sachant la valeur de X : plus précisément, pour tout $x \in [0, 1]$,

$$\mathrm{Loi}(Y \mid X = x) = \mathrm{Unif}^{\mathrm{me}}(0, x). \tag{FA}$$

Or la loi $\mathrm{Unif}^{\mathrm{me}}(0, 1)$ a pour densité $x \mapsto \mathbf{1}_{x \in [0,1]}$, tandis que la loi $\mathrm{Unif}^{\mathrm{me}}(0, x)$ a pour densité $y \mapsto \mathbf{1}_{y \in [0,x]} x^{-1}$. D'après le théorème (EY), on en déduit que la densité jointe de (X, Y) est

$$(x, y) \mapsto \mathbf{1}_{x \in [0,1]} \times \mathbf{1}_{y \in [0,x]} x^{-1} = \mathbf{1}_{0 \leqslant y \leqslant x \leqslant 1} x^{-1}. \tag{FB}$$

Continuons notre analyse et supposons qu'on nous ait révélé la valeur de Y, *mais pas celle de* X, et qu'on se demande alors ce que peut bien valoir X compte tenu de cette information : autrement dit, on cherche la loi de X *sachant que* $Y = y_{\checkmark}$, où $y_{\checkmark}$ est la valeur effectivement observée pour la variable Y ($y_{\checkmark}$ apparait donc ici comme un *paramètre* — non spécifié — et non comme une variable). Cette fois-ci, c'est donc le théorème (EQ) qu'il faut utiliser : on trouve que la loi de X sachant que $\{Y = y_{\checkmark}\}$ a pour densité

$$x \mapsto Z^{-1} \mathbf{1}_{x \in [y_{\checkmark}, 1]} x^{-1} \tag{FC}$$

(voir figure 4.1), où

$$Z = \int_{x=0}^{1} \mathbf{1}_{x \in [y_{\checkmark}, 1]} x^{-1} \mathrm{d}x \tag{FD}$$

(cette dernière valeur ne dépend pas de x — la variable 'x' dans l'expression de Z étant liée). Notez qu'en fait, l'expression de Z comme intégrale est automatique, puisque l'intégrale d'une densité de probabilité doit valoir 1. Par ailleurs, la valeur précise de Z (qui, en l'occurrence, vaut $\ln(1/y_{\checkmark})$) ne nous importe (en général) pas vraiment (et d'ailleurs, dans bien des cas elle ne serait pas calculable) : en effet, décrire une densité de probabilité *à constante multiplicative près* caractérise complètement cette densité, vu la contrainte de normalisation à 1 de la masse totale.

Remarquez au passage l'aspect contre-intuitif du résultat obtenu : alors qu'on aurait pu se dire que le fait de savoir $Y = 0{,}34$ nous permettait seulement d'en conclure de $x \geqslant 0{,}34$, on s'aperçoit qu'en fait, *du fait de l'information obtenue*, les valeurs de X les plus petites ont plus de chance d'avoir eu lieu que les plus grandes, par exemple, alors que *à priori* X avait autant de chances d'être comprise entre 0,34 et 0,67 qu'entre 0,67 et 1, *à postériori* (c'est-à-dire, *sachant* l'information sur Y) le premier cas a environ 1,77 fois plus de chances de se produire que le second ! ♧

4.2 Indépendance

Lois produits

! **Théorème** (FE) (Loi produit)**.** *Si les variables aléatoires X et Y sont indépendantes, alors la loi jointe de X et Y ne dépend que des lois de X et Y : on l'appelle la* loi produit *de* $\mathrm{Loi}(X)$ *et* $\mathrm{Loi}(Y)$*, notée* $\mathrm{Loi}(X) \otimes \mathrm{Loi}(Y)$*. De manière générale, la loi produit d'une distribution μ sur $\mathcal{X}$ et d'une distribution ν sur $\mathcal{Y}$ est caractérisée par :*

$$\forall A \subseteq \mathcal{X}, B \subseteq \mathcal{Y} \quad \big(\mu \otimes \nu\big)(A \times B) = \mu(A)\nu(B). \tag{FF}$$

◇

Remarque (FG)*.* Ce théorème n'est en fait pas valide lorsqu'on on travaille sous la théorie des probabilités finiment additives : pour le coup, on a vraiment besoin de la théorie de Kolmogorov pour parler de lois produits. ♧

Proposition (FH) (Produit de lois discrètes)**.** *Si les variables aléatoires X et Y ne prennent leurs valeurs que dans des ensembles discrets resp. $\mathcal{X}$ et $\mathcal{Y}$ et sont indépendantes, alors la variable jointe (X, Y) suit une loi discrète à valeurs dans $\mathcal{X} \times \mathcal{Y}$, avec pour tout $(x, y) \in \mathcal{X} \times \mathcal{Y}$:*

$$\mathbb{P}\big((X, Y) = (x, y)\big) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y) \tag{FI}$$

◇

Démonstration. Immédiat en observant que $\{(X, Y) = (x, y)\}$ peut se réécrire $\{X = x \text{ et } Y = y\}$. ♤

Proposition (FJ) (Produit de lois à densité)**.** *Si les variables aléatoires X et Y sont à valeurs dans resp. $\mathbb{R}^n$ et $\mathbb{R}^m$ avec des lois de densités respectives $x \mapsto f(x)$ et $y \mapsto g(y)$, et que ces variables aléatoires sont indépendantes, alors la variable jointe (X, Y) a une loi de densité $(x, y) \mapsto f(x)g(y)$.* ◇

Démonstration. Cherchons la densité de la loi en (x, y). On considère un voisinage infinitésimal de 0 dans $\mathbb{R}^{n+m}$ de forme produit $\mathrm{d}x \times \mathrm{d}y$: alors

$$\begin{aligned}\mathbb{P}\big((X, Y) \in (x, y) + \mathrm{d}x \times \mathrm{d}y\big) &= \mathbb{P}(X \in x + \mathrm{d}x \text{ et } Y \in y + \mathrm{d}y) \\ &\underset{\text{ind}^{\text{ce}}}{=} \mathbb{P}(X \in x + \mathrm{d}x)\,\mathbb{P}(Y \in y + \mathrm{d}y) = f(x)\,\mathrm{vol}(\mathrm{d}x)g(y)\,\mathrm{vol}(\mathrm{d}y) \\ &= f(x)g(y)\,\mathrm{vol}(\mathrm{d}x \times \mathrm{d}y),\end{aligned}$$

ce qui donne bien la densité recherchée. ♤

Remarque (FK)*.* Bien entendu, on peut aussi considérer des produits de lois discrètes par des lois continues, ou même des choses encore plus compliquées... L'idée de base pour les calculs des deux propositions ci-dessus continuera à être valable, mais ne conduira alors pas à une "belle" expression de loi qui serait discrète ou à densité. ♧

Chapitre 5

Théorèmes-limites

Nous présentons à présent deux théorèmes fondamentaux de la théorie des probabilités qui permettent de comprendre le comportement à la limite de la somme de variables aléatoires indépendantes : comme nous le verrons dans la suite ce cours, l'importance de ces théorèmes en modélisation statistique est primordiale ! Dans ce polycopié, j'ai opté volontairement pour des énoncés assez informels, afin de ne pas nous encombrer avec les technicités qu'on appelle « uniforme intégrabilité », « convergence en probabilité » ou « convergence en loi »... Par ailleurs, notez que les énoncés qu'on appelle « loi des grands nombres » et « théorème-limite central » *stricto sensu* concernent en fait seulement le cas où les variables aléatoires sont identiquement distribuées : les résultats ci-dessous en sont donc en toute rigueur des généralisations, auxquelles les puristes donneraient des noms différents.

5.1 La loi des grands nombres

Commençons par rappeler l'énoncé de la loi des grands nombres telle qu'on l'introduit en général dans les cours de probabilités :

Théorème (FL) (Loi (faible) dans grands nombres, version stricte)**.** *Soit* $\boldsymbol{P}$ *une distribution de probabilité* L^1 *sur* $\mathbb{R}$ *et* $X_1, X_2, \dots$ *une suite i.i.d. de variables de loi* $\boldsymbol{P}$*. Alors*

$$\mathrm{Loi}\Big(n^{-1}\sum_{i=1}^{n} X_i\Big) \overset{n\to\infty}{\to} \delta_{\mathbb{E}(P)}. \tag{FM}$$

◇

Remarque (FN). Noter que la loi qui apparait dans le membre de gauche de la relation de convergence est directement déterminée par $\boldsymbol{P}$ et par n : le fait d'introduire les X_i est juste un artifice pour exprimer cette loi plus facilement. ♧

Remarque (FO). La loi « forte » des grands nombres ajoute que la convergence de la moyenne des X_i vers la constante $\mathbb{E}(\boldsymbol{P})$ a également lieu presque-surement (et aussi dans L^1) ; mais cela ne comporte aucun intérêt du point de vue de la modélisation, donc nous en ferons abstraction. ♧

Remarque (FP). Les deux défauts de la loi des grands nombres dans sa version "classique" sont, d'une part, qu'on a besoin que les X_i soient tous de même loi ; d'autre part, que rien n'est dit sur la vitesse de convergence vers la limite, et donc du contrôle qu'on peut obtenir pour une valeur fixée de n. Or ces deux points sont

cruciaux en modélisation ; d'où le besoin d'une généralisation, que nous donnons ci-dessous. Cette généralisation sera donnée d'abord en termes informels (la seule à retenir, car de toutes façons, on n'a jamais de contrôle formel sur le qu'on fait en matière de modélisation), puis une version plus formelle sera présentée. ♣

! **Théorème** (FQ) (Loi des grands nombres, version généralisée informelle)**.** *Soit une variable aléatoire $X = X_1 + \ldots + X_n$ s'écrivant comme une somme de variables aléatoires, où :*

(i) Les X_i sont indépendantes ;

(ii) X_i a des fluctuations de l'ordre de ℓ_i ;

(iii) Les fluctuations des X_i ne sont pas trop "sauvages" (à l'échelle ℓ_i) ;

(iv) Le nombre de X_i nécessaires pour contribuer substantiellement à la somme des ℓ_i est très grand,

alors, à l'échelle $\sum_i \ell_i$, X a une très grande probabilité d'être presque égale à son espérance ; autrement dit on a une probabilité presque égale à 1 que $|X - \mathbb{E}(X)|$ soit beaucoup plus petit que $\sum_i \ell_i$. ◇

Remarque (FR). Notez que, par linéarité de l'espérance, l'espérance de X n'est autre que la somme des espérances des X_i. ♣

Remarque (FS). Par « les X_i ont des fluctuations de l'ordre de ℓ_i », on entend que "l'essentiel" des valeurs possibles que X_i peut prendre sont comprises dans une plage dont la longueur est "de l'ordre de" ℓ_i. Une façon *possible* de rendre concrète cette définition informelle pourrait être de dire que la *distance interquartiles* de $\mathrm{Loi}(X_i)$ (à savoir, la différence entre les troisième et premier quartiles de cette loi) doit être inférieure à ℓ_i. Cependant, en fonction du contexte, il peut y avoir d'autres façons pertinentes d'exprimer l'idée de « fluctuations de l'ordre de ℓ », de sorte qu'une définition floue est en fait préférable ici.

Sur la notion que les fluctuations ne soient pas trop "sauvages" : intuitivement, l'idée est de dire que, pour chaque i, la probabilité que X_i prenne des valeurs nettement plus éloignées que ℓ_i de sa valeur la plus typique (disons sa médiane, ou son espérance) est très petite. Plus précisément, la condition technique requise relève en fait que ce qu'on appelle l'*uniforme intégrabilité* : la façon décroit dont la probabilité que X_i prenne des valeurs très éloignées de ses valeurs typiques doit, d'une part, avoir toujours la même forme, et d'autre part, assurer que $|X_i - \text{médiane}(\mathrm{Loi}(X_i))|$ soit intégrable (avec, dès lors, une intégrale qui sera bornée par un multiple de ℓ_i). Le théorème **??** vu plus loin donne une façon précise de formuler cela.

Si la définition générale de « fluctuations pas trop sauvages à l'échelle ℓ_i » est compliquée, il ce trouve cependant qu'heureusement, on peut donner une condition *suffisante* pour cette notion dans le cadre du théorème (FQ), consitant à simplement contrôler l'*écart-type* de X_i par ℓ_i :

$$\forall i \in \{1, \ldots, n\} \qquad \mathrm{Var}^{1/2}(X_i) \leqslant \ell_i. \tag{FT}$$

♣

! *Point* (FU). Concernant la remarque ci-dessus, dans le cadre de ce cours, pour les exemples que vous aurez à traiter, soit cette dernière forme de la condition de contrôle uniforme des fluctuations sera vérifiée, soit on sera dans un cas où les fluctuations des X_i n'auront clairement pas de contrôle "civilisé" à l'échelle

désirée. Mais dans tous les cas, il ne vous sera jamais demandé d'expliciter ce qu'on entend par « fluctuations pas trop sauvages » : je vous demande juste de rappeler qu'il y a une condition de « contrôle des fluctuations à l'échelle ℓ_i » à vérifier, et éventuellement de justifier informellement que cela semble vrai, du point de vue de l'idée générale, pour la situation à laquelle vous êtes confrontés. ♧

Remarque (FV). Toujours au sujet de la remarque (FS), notez qu'il y a des distributions de probabilité d'écart-type infini, dont les fluctuations seraient donc d'échelle infinie au sens proposé dans cette remarque. Et de fait, il existe des lois pour lesquelles on n'observe pas ce phénomène de « concentration » vers une loi presque constante à l'échelle n : par exemple, si les X_i suivent toutes la loi Cauchy$(0,1)$, déterminée par la densité

$$\mathbb{P}\big(\text{Cauchy}(0,1) \in \mathrm{d}x\big) := \frac{1}{\pi}\frac{1}{1+x^2}\,\mathrm{vol}_1(\mathrm{d}x), \tag{FW}$$

alors on pourrait montrer que $X_1 + \cdots + X_n$ suit la loi $n \times \text{Cauchy}(0,1)$. Ainsi, alors que les X_i semblent avoir des fluctuations de l'ordre de grandeur de 1 (et toutes de la même forme, évidemment), leur somme n'est pas du tout constante à l'ordre de grandeur de n... Cependant, lorsque nous mesurons l'échelle des fluctuations au sens de la remarque (FS), on s'aperçoit que cela ne cause pas de contradiction avec la loi des grands nombres : en effet, l'écart-type d'une loi de Cauchy est infini [∗], de sorte qu'il n'existe aucune échelle qui contrôle les fluctuations des X_i ! En revanche, si nous avions mesuré les fluctuations au sens de, par exemple, la distance interquartiles, on aurait trouvé pour les X_i des fluctuations d'ordre 2, tandis que X aurait des fluctuations d'ordre 2 000, ce qui n'est évidemment pas négligeable devant $1\,000 \times 2$ [†]. Cela montre donc, d'une part, qu'on ne peut pas considérer n'importe quel concept de « fluctuations » lorsqu'on souhaite énoncer rigoureusement le théorème (FQ) ; et d'autre part, que le concept de fluctuations retenu devra nécessairement donner à certaines distributions de probabilité un « échelle de fluctuations » infinie. ♧

Remarque (FX). L'autre point un peu compliqué dans la définition ci-dessus est que « le nombre de X_i nécessaires pour contribuer substantiellement à la somme des ℓ_i est très grand ». Qu'entend-on par là ? En fait, si tous les ℓ_i étaient égaux, on pourrait simplement écrire « n est très grand ». Mais comme l'énoncé autorise les ℓ_i à être inégaux, il faut éviter un cas dégénéré comme celui où, par exemple, X_1 suivrait la loi $\text{Unif}^{\text{me}}(0,1)$, n serait égal à 10^9, et chacun des X_i pour $i > 1$ suivrait la loi $\text{Unif}^{\text{me}}(0, 10^{-19})$: dans ce cas, la somme des X_i serait *en pratique* quasiment égale à X_1, et c'est donc *comme si* on avait une unique variable aléatoire... ! En fait, ce qui est demandé ici, c'est que pour tout ensemble I dont le cardinal ne serait pas « très grand », la somme des X_i pour $i \in I$ ne représenterait qu'une fraction négligeable de la somme totale des X_i : et plus précisément (dans la mesure où la notion de « négligeable » n'est pas claire pour des sommes de variables aléatoires) on veut que la somme des ℓ_i pour $i \in I$ soit négligeable devant la somme totale des ℓ_i. En fait, une manière simple, quoique pas forcément aussi claire à première vue, d'exprimer cette idée est de demander que, pour tout i, le ratio entre ℓ_i et la

[∗]. En fait, même l'*espérance* d'une loi de Cauchy ne peut pas être définie, tellement cette loi a des queues lourdes ! On peut d'ailleurs montrer que dans le cas où les X_i sont i.i.d. et qu'on considère le comportement de leur somme à l'échelle n dans l'asymptotique $n \to \infty$, il est nécessaire et suffisant pour que la loi des grands nombres soit valide qu'on puisse définir une espérance aux X_i : c'est en fait l'énoncé classique de la loi des grands nombres tel qu'il est donné dans la plupart des cours de probabilité.

[†]. En fait, on peut même trouver des exemples similaires (avec les X_i i.i.d.) où les choses se passent d'une manière encore pire, au sens où les fluctuations de X (mesurées au sens de la distance interquartiles) croissent de façon *plus rapide* que n lorsque n augmente ! (aussi étonnant que cela paraisse à première vue). C'est par exemple le cas si les X_i suivent une loi $\text{Pareto}(1,k)$ pour $k \in]0,1[$.

somme de tous les $\ell_{i'}$ soit très grand. C'est cette formulation qui est utilisée dans le théorème **??** ci-après. ♧

Remarque (FY). Il est clair que les conditions (i) et (iv) sont nécessaires :
— En ce qui concerne la condition (i), considérons X_1 suivant n'importe quelle distribution de probabilité dont les fluctuations sont d'échelle $\ell \in]0,\infty[$, et posons $X_2 = \cdots = X_n = X_1$. (Il n'y a donc absolument pas indépendance !). Alors X sera égale à nX_1, et aura donc des fluctuations d'échelle $n\ell$; alors que la loi des grands nombres requerrait des fluctuations beaucoup plus petites que $n\ell$.
— En ce qui concerne la condition (iv), prenons $n = 1$ (l'indépendance sera alors automatiquement satisfaite), avec X_1 suivant n'importe quelle distribution de probabilité dont les fluctuations sont d'échelle $\ell \in]0,\infty[$: dans ce cas, $X = X_1$, de sorte les fluctuations de X seront d'échelle ℓ, ce qui n'est évidemment pas beaucoup plus petit que $n \times \ell = 1 \times \ell = \ell$.

♧

Exemple (FZ). Soient $X_1, \ldots, X_{999}$ 999 variables indépendantes avec $X_1 \sim \text{Bernoulli}(0{,}001), X_2 \sim \text{Bernoulli}(0{,}002), \ldots, X_{999} \sim \text{Bernoulli}(0{,}999)$; et soit $X := \sum_{i=1}^{999} X_i$. Les X_i sont indépendantes par hypothèse, et leur nombre peut manifestement être qualifié de « très grand ». Par ailleurs, puisque chaque X_i est à valeurs dans $\{0,1\}$, l'interprétation intuitive de la notion de « contrôle des fluctuations » montre que l'ordre de grandeur des fluctuations des X_i est uniformément contrôlé par 1 [‡]. Par conséquent, à l'échelle $999 \times 1 = 999$, X devrait être « presque égale à son espérance »

$$\mathbb{E}(X) = \sum_{i=1}^{999} \mathbb{E}(X_i) = \sum_{i=1}^{999} \frac{i}{1\,000} = 499\tfrac{1}{2}\,; \qquad \text{(GA)}$$

autrement dit, l'archi-majorité des valeurs de X devraient être éloignées de la valeur $499\frac{1}{2}$ d'une distance beaucoup plus petite que 999. Vérifions cela empiriquement sous R :

```
> # Pour illustration : La commande ci-dessous génère, à chaque appel,
> # un nouveau vecteur aléatoire composé de 9 réalisations de variables
> # de Bernoulli indépendantes, de paramètres respectifs 0.1, ..., 0.9.
> # (En vrai, il faudra remplacer « 9 » par « 999 », et les paramètres par
> # 0.001, ..., 0.999).
> rbinom (9, 1, seq (from=.1, to=.9, by=.1))
[1] 0 0 0 0 1 1 0 0 1
> # Nous allons maintenant faire 100 tirages de ce type,
> # calculer à chaque fois la somme, et afficher les 100 résultats obtenus.
> resultats = numeric (100)
> for (i in 1:100)
+ resultats[i] = sum (rbinom (999, 1, seq (from=.001, to=.999, by=.001)))
> resultats
  [1] 480 481 504 503 496 496 505 504 500 526 499 506 486 507 486 504 497 497
 [19] 502 516 491 495 493 498 487 516 512 495 505 541 512 512 501 488 489 508
 [37] 485 510 498 510 499 492 514 501 501 476 503 483 510 516 477 515 497 481
 [55] 501 507 500 525 476 505 507 506 504 490 517 499 493 516 502 517 500 516
 [73] 498 520 483 498 503 499 492 513 481 506 479 513 507 487 515 524 491 481
 [91] 506 528 504 500 506 490 495 530 492 501
> # Trions maintenant les résultats pour voir comment se répartissent
> # les valeurs obtenues :
> sort (resultats)
  [1] 476 476 477 479 480 481 481 481 481 483 483 485 486 486 487 487 488 489
```

[‡]. Avec la formulation rigoureuse proposée, on pourrait observer que pour tout i, on a $\text{Var}^{1/2}(X_i) \leqslant 1/2$.

```
 [19] 490 490 491 491 492 492 492 493 493 495 495 495 496 496 497 497 497 498
 [37] 498 498 498 499 499 499 499 500 500 500 500 501 501 501 501 501 502 502
 [55] 503 503 503 504 504 504 504 504 505 505 505 506 506 506 506 506 507 507
 [73] 507 507 508 510 510 510 512 512 512 513 513 514 515 515 516 516 516 516
 [91] 516 517 517 520 524 525 526 528 530 541
> # En 100 tirages, nous n'avons trouvé aucune valeur qui fût
> # plus éloignée de 499.5 que ne l'est 541, soit une distance maximale
> # de 41.5 ; ce qui est effectivement beaucoup plus petit que 999 !
```

♧

Voyons, pour information, un énoncé précis qui quantifie le théorème informel de la section 1.

Théorème (GB) (Loi des grands nombres, version généralisée formelle)**.** *Soit $k > 1$. Alors il existe une fonction (explicitable) $\varphi_k(\bullet,\bullet)\colon \mathbb{R}_+^* \times [1,\infty[\to \mathbb{R}$, telle que pour tout $\varepsilon > 0$, on ait $\varphi_k(\varepsilon,\tilde{n}) \overset{\tilde{n}\to\infty}{\to} 0$, pour laquelle on a la propriété suivante :*

Pour $n \in \mathbb{N}^*$, $X_1,\dots,X_n$ une famille de variable aléatoires indépendantes toutes intégrables[§]*, et $(\ell_i)_{1\leqslant i\leqslant n}$ des nombres réels strictement positifs tels qu'on ait*

$$\forall i \in [\![1,n]\!] \quad \forall z \in]1,\infty[\quad \mathbb{P}(|X_i - \mathbb{E}(X_i)| \geqslant z\ell_i) \leqslant \frac{1}{z^k}, \tag{GC}$$

alors pour tout $\varepsilon > 0$,

$$\mathbb{P}\Big(\Big|\sum_i X_i - \sum_i \mathbb{E}(X_i)\Big| \geqslant \varepsilon \sum_i \ell_i\Big) \leqslant \varphi_k\Big(\varepsilon, \frac{\sum_i \ell_i}{\max_i \ell_i}\Big). \tag{GD}$$

♢

Remarque (GE)*.* En fait, il n'est pas nécessaire que n soit fini pour appliquer le théorème : tout ce qui compte est que la somme des ℓ_i le soit (et que la somme des $\mathbb{E}(X_i)$ soit absolument convergente), comme on le vérifie aisément par continuité. De même, on pourrait éventuellement autoriser certains des ℓ_i à valoir 0, ce qui reviendrait simplement à ajouter des v.a. constantes. ♧

Remarque (GF)*.* Dans la condition sur le contrôle des fluctuations de X_i :

$$\forall z \in]1,\infty[\quad \mathbb{P}(|X_i - \mathbb{E}(X_i)| \geqslant z\ell_i) \leqslant \frac{1}{z^k}, \tag{GG}$$

il n'est en fait pas très importat que ce soit l'espérance de X_i qui intervienne. En effet, s'il existe un m_i tel qu'on ait le même contrôle sur $\mathbb{P}(|X_i - m_i| \geqslant z\ell_i)$, la condition d'origine est vérifiée, en remplaçant simpement ℓ_i par $(2k-1)/(k-1)\times\ell_i$. ♧

Remarque (GH)*.* Pour $k = 3$, il est facile d'expliciter une fonction φ_k convenable : dans ce cas en effet, on a $\mathrm{Var}(X_i) \leqslant 3\ell_i^2$, et donc on peut contrôler la variance de $\sum X_i$, puis appliquer l'inégalité de Bienaymé-Tchebychev pour en déduire, via l'inégaité de Hölder, le résultat, avec $\varphi_3(\varepsilon,\tilde{n}) = 1/3\varepsilon^2\tilde{n}$. (Un argument similaire fonctionne en fait dès lors que $k > 2$; par contre, pour $k \leqslant 2$ le théorème est plus difficile).

Le cas où on contrôle directement les écarts-types des X_i, quant à lui, implique les hypothèses du théorème pour $k = 2$. ♧

Remarque (GI)*.* Le théorème tel qu'il est écrit ne redonne pas la forme classique de la loi des grands nombres. Pour cela, il faut aller encore plus loin et remplacer la condition de contrôle des queues en puissance de k par une fonction de la forme

$$\forall z \in]1,\infty[\quad \mathbb{P}(|X_i - \mathbb{E}(X_i)| \geqslant z\ell_i) \leqslant \kappa(z), \tag{GJ}$$

où $\kappa(z)$ doit être une fonction (la même pour tous les i) vérifiant $\int_{z=1}^{\infty} \kappa(z)\mathrm{d}z < \infty$. (D'où le fait, quand on se restreint au cas particulier $\kappa(z) \coloneqq 1/z^k$, qu'on doive avoir $k > 1$). Dans ce cas la fonction φ_k doit être bien entendu remplacée par une certains fonction $\varphi_{\kappa(\bullet)}$. Avec cette variante, on obteint bien un résultat qui généralise complètement le théorème **??**. ♧

[§]. Cette condition est en fait automatique à cause du contrôle sur les fluctuations des X_i.

5.2 Le théorème-limite central

J'ai fait ici le choix de traiter le théorème-limite central d'une façon aussi similaire possible à la loi des grands nombres : comme vous allez le constater, les similarités sont en effet frappantes ! Commençons par l'énoncé "strict" ce théorème :

Théorème (GK) (Théorème-limite centre, version stricte)**.** *Soit $\boldsymbol{P}$ une distribution de probabilité* L^2 *sur* $\mathbb{R}$ *et* $X_1, X_2, \dots$ *une suite i.i.d. de variables de loi* $\boldsymbol{P}$*. Alors*

$$\mathrm{Loi}\left(\frac{\sum_{i=1}^n X_i - n\,\mathbb{E}(\boldsymbol{P})}{n^{1/2}} \right) \overset{n\to\infty}{\mathrm{Normale}}\,(0, \mathrm{Var}(\boldsymbol{P})). \tag{GL}$$

♢

Remarque (GM)*.* On peut également écrire le théorème-limite central dans un cadre multidimensionnel, les $\boldsymbol{X}_i$ étant alors dans $\mathbb{R}^d$, et les $\mathrm{Var}(\boldsymbol{X}_i)$ devenant alors des matrices de covariances (dans $\mathbb{R}^{d\times d}$) : tout fonctionne exactement de la même façon. ♣

> *Remarque* (GN)*.* En fait, on peut même voir le TLC multidimensionnel comme un corolaire du TLC unidimensionnel, à cause du résultat suivant : une famille de distributions de probabilité $(\boldsymbol{P}_n)_{n\in\mathbb{N}}$ sur $\mathbb{R}^d$ converge en loi vers $\boldsymbol{P}_\infty$ si et seulement si, pour toute profection linéaire $\pi\colon \mathbb{R}^d \to \mathbb{R}$, $\pi_* \boldsymbol{P}_n$ converge en loi vers $\pi_* \boldsymbol{P}_\infty$. ♣

Le théorème-limite central, dans sa version stricte, souffre exactement des deux mêmes défauts que la loi des grands nombres : il requiert que les $\boldsymbol{X}_i$ aient toutes la même loi, et ne permet pas de quantifier la vitesse de convergence. Mais, de même que pour la loi des grands nombres, on peut en donner une version généralisée, extrêmement similaire à la version généralisée de la loi des grands nombres, tant dans son énoncé formel que dans son énoncé informel :

! **Théorème** (GO) (Théorème-limite central, version généralisée informelle)**.** *Soit une variable aléatoire $\boldsymbol{X} = \boldsymbol{X}_1 + \dots + \boldsymbol{X}_n$ s'écrivant comme une somme de variables aléatoires, où :*

(i) Les $\boldsymbol{X}_i$ sont indépendantes ;

(ii) $\boldsymbol{X}_i$ a des fluctuations de l'ordre de ℓ_i ;

(iii) Les fluctuations des $\boldsymbol{X}_i$ ne sont pas trop "sauvages" (à l'échelle ℓ_i) ;

(iv) Le nombre de $\boldsymbol{X}_i$ nécessaires pour que la somme des ℓ_i^2 contribue substantiellement à la somme des $\mathrm{Var}(\boldsymbol{X}_i)$ est très grand,

alors la loi de $\boldsymbol{X}$ est presque égale à la distribution normale ayant les mêmes espérance et écart-type que $\boldsymbol{X}$; autrement dit :

$$\mathrm{Loi}(\boldsymbol{X}) \approx \mathrm{Normale}(\mathbb{E}(\boldsymbol{X}), \mathrm{Var}(\boldsymbol{X})). \tag{GP}$$

♢

Remarque (GQ)*.* Comme pour la loi des grands nombres, on a par linéarité de l'espérance que $\mathbb{E}(\boldsymbol{X}) = \sum_i \mathbb{E}(\boldsymbol{X}_i)$. En outre, puisque les $\boldsymbol{X}_i$ sont indépendantes, la variance de leur somme est égale à la somme de leurs variances :

$$\mathrm{Var}(\boldsymbol{X}) = \sum_{i=1}^n \mathrm{Var}(\boldsymbol{X}_i). \tag{GR}$$

♣

Remarque (GS). On notera que la variance de X_i est elle aussi une façon d'exprimer les fluctuations de Loi(X_i), de sorte qu'en fait, les ℓ_i ne pourront pas être n'importe quoi par rapport aux $\mathrm{Var}^{1/2}(X_i)$. Plus précisément, dans le contexte du théorème-limite central, on peut montrer que les seules façons convenables d'interpréter la notion de « fluctuations pas trop sauvages » imposent qu'on ait $\ell_i \geqslant \mathrm{Var}^{1/2}(X_i)$. ♧

Remarque (GT). Notez que, cette fois-ci, le condition qui impose qu'il y ait réellement un grand nombre de variables qui interviennent dans la somme est un peu plus compliquée, puisqu'elle compare les ℓ_i^2 à la somme totale des $\mathrm{Var}(X_i)$. En fait, on pourrait écrire le théorème en imposant que les ℓ_i soient pris égaux aux $\mathrm{Var}^{1/2}(X_i)$, ce qui rendrait alors la condition d'éparpillement des variables plus simple (et plus similaire à celle du théorème-limite central) ; mais cela empêcherait alors de considérer le cas où on a une poignée de variables dont le comportement serait très "sauvage" à l'échelle $\mathrm{Var}^{1/2}(X_i)$, mais néanmoins « gentil » à une échelle ℓ_i supérieure dont la contribution serait faible : or ce genre de situations peut réellement se rencontrer en modélisation, comme nous le verrons d'ailleurs plus loin. ♧

Remarque (GU). La condition de "gentillesse" des fluctuations dans le théorème (GO) n'est, là encore, pas évidente à expliciter : ce qui se cache derrière est le concept d'*uniforme intégrabilité quadratique*... On peut néanmoins y écrire une condition suffisante, en disant par exemple que le moment quatrième (centré) de X_i doit être majoré par ℓ_i^4 :

$$\forall i \in \{1, \dots, n\} \qquad \mathbb{E}\big((X_i - \mathbb{E}(X_i))^4\big) \leqslant \ell_i^4. \tag{GV}$$

Il serait tentant de penser que ces conditions (iii) et (iv) ne sont que des subtilités techniques qui ne poseront jamais souci en pratique ; mais malheureusement, les contre-exemples (GZ) et (HA) ci-dessous montrent qu'il existe des situations tout à fait concrètes où le théorème-limite central échoue à cause d'une de ces deux conditions ! ♧

Point (GW). Concernant la remarque précédente, pas de panique : de façon similaire au paragraphe « dans le cadre de ce cours » qui concluait la remarque (FS), dans le cadre de ce cours, vous n'êtes pas censés savoir écrire à quoi correspondent précisément les conditions (iii) et (iv) (et encore moins les vérifier) : tout ce que je vous demande, c'est de rappeler qu'il faut avoir ces deux conditions (en en donnant les énoncés heuristiques tels qu'ils apparaissent dans le théorème (GO)), et de vous convaincre à gros traits qu'elles sont "moralement" valides dans la situation qui vous intéresse. (Si jamais je devais vous présenter une situation où une de ces conditions n'était pas vérifiée, ce serait toujours de façon flagrante). ♧ !

Exemple (GX). Nous reprenons ici la situation de l'exemple (FZ) : $X := X_1 + \dots + X_{999}$, les X_i étant indépendantes et de lois respectives Bernoulli($i/1\,000$). Nous avons déjà dit dans l'exemple (FZ) que les conditions (i) et (iv), qui sont identiques dans les théorèmes (FQ) et (GO), sont vérifiées. Nous voulons maintenant vérifier la condition (??), consistant à dire qu'il n'y a pas de X_i dont l'écart-type soit nettement plus grand que pour la plupart des autres $X_{i'}$. Ici l'écart-type de X_i est $\sqrt{i(1\,000 - i)}/1\,000$, qui est maximal pour X_{500} où il vaut alors 0,5, et dont les valeurs sont $\geqslant 0{,}4$ pour tout $i \in \{200, \dots, 800\}$, ce qui représente bien une large partie des X_i : l'hypothèse (iv) est donc bien vérifiée, et les écarts-type sont pour la plupart de l'ordre de 0,5. Maintenant, l'hypothèse (iii) consiste alors à dire que la probabilité qu'un des X_i s'éloigne de son espérance de nettement plus de 0,5 devient rapidement très petite. Mais ici tous les X_i sont à valeurs dans $\{0, 1\}$, donc quel que soit i, il est strictement impossible d'avoir $|X_i - \mathbb{E}(X_i)| > 1 = 2 \times 0{,}5$: l'hypothèse

de contrôle des fluctuations est donc vérifiée, avec $\ell_i = 1$. [¶]. On s'attend donc à ce que la loi de X soit très proche d'une loi normale ayant des paramètres compatibles avec l'espérance et l'écart-type de X.

L'espérance de X, nous l'avons calculé dans l'exemple (FZ), vaut 499,5. En ce qui concerne la calcul de l'écart-type, on utilise que la variance d'une somme indépendante [‖] de variables aléatoires est la somme des variances ; et puisqu'un calcul simple montre que $\mathrm{Var}(X_i) = \frac{i}{1\,000}\left(1 - \frac{i}{1\,000}\right)$, on en déduit que

$$\mathrm{Var}(X) = \sum_{i=1}^{999} \frac{i}{1\,000}\left(1 - \frac{i}{1\,000}\right) \underset{[**]}{=} \frac{999 \times 1\,001}{6 \times 1\,000} = 166{,}666\,5, \qquad \text{(GY)}$$

d'où $\mathrm{Var}^{1/2}(X) = \sqrt{166{,}666\,5} \approx 12{,}910$: on doit donc avoir que la loi de X est très proche de la loi Normale(499,5; 12,910). Pour vérifier cela, nous voulons comparer les fonctions de répartition de la loi de X et de la loi Normale(499,5; 12,910), c.-à-d. les fonctions $x \mapsto \mathbb{P}(X \leqslant x)$ et $x \mapsto \mathbb{P}\big(\text{Normale}(499{,}5; 12{,}910) \leqslant x\big)$.

> En ce qui concerne la deuxième fonction, nous pouvons la calculer grâce à la fonction pnorm de *R*. Pour la première, malheureusement il n'y a pas d'expression analytique de la fonction de répartition de X ; nous sommes donc obligés de l'approcher numériquement par la méthode de Monte-Carlo : c.-à-d. que nous simulons un grand nombre de fois (disons 2^{16}) la loi de X de façon indépendante, et que nous considérons que la mesure empirique des valeurs ainsi obtenues est très proche de la véritable loi de X. La code *R* est donné ci-dessus.

Le résultat graphique est donné sur la figure 5.1 : effectivement, la loi de X est très bien approchée par la loi normale que nous avions prédite ! ☺

```
   > # On calcule les moyenne et écart-type de X et de la loi normale
> # censée l'approximer, appelés m et s.
> m = 499.5
> (s = sqrt (166.6665))
[1] 12.90994
> # N est le nombre de simulations pour Monte-Carlo
> (N = 2^16)
[1] 65536
> # On simule N fois la loi de X (indépendamment à chaque fois bien sûr),
> # et on stocke les résultats dans "realisations.de.X". Notez que
> # la ligne de commande pour simuler la loi de X
> # est reprise du code précédent.
> realisations.de.X = numeric (N);
> for (i in 1:N)
+ realisations.de.X[i] = sum (rbinom (999, 1,
seq (from=.001, to=.999, by=.001)))
> # À l'aide de la commande "str", on peut regarder à quoi ressemble
```

[¶]. Par contre, si on avait pris "naïvement" $\ell_i \propto \mathrm{Var}^{1/2}(X_i)$, cette fois-ci cela n'aurait pas fonctionné : en effet, si on prend par exemple $\ell_i = 8\,\mathrm{Var}^{1/2}(X_i)$, on a que $\mathbb{E}(|X_i - \mathbb{E}(X_i)|^4) \leqslant (8\,\mathrm{Var}^{1/2}(X_i))^4$ pour la grande majorité des i... mais pas dans les 30 cas (sur 999) où $i \leqslant 15$ ou $i \geqslant 985$. Et c'est là qu'on comprend l'intérêt de rendre ℓ_i potentiellement différent de $\mathrm{Var}^{1/2}(X_i)$: car certes, les variables aléatoires en question ont des fluctuations "sauvages" à l'échelle de leur écart-type ; mais elles ont malgré tout des fluctuations bien contrôlées à l'échelle 1 : or, pour 30 variables seulement, remplacer les écarts-types par 1 n'a pas grand impact sur la somme des ℓ_i^2 ! Ainsi la condition d'éparpillement est bien vérifiée.

[‖]. Ne pas oublier l'hypothèse d'indépendance ici !

[**]. On utilise ici l'identité $\sum_{i=1}^{N-1} i(N-i) = (N-1)N(N+1)/6$, que je ne démontre pas (même si c'est assez simple à vérifier), dans la mesure où cela n'est pas nécessaire à la compréhension de l'exemple.

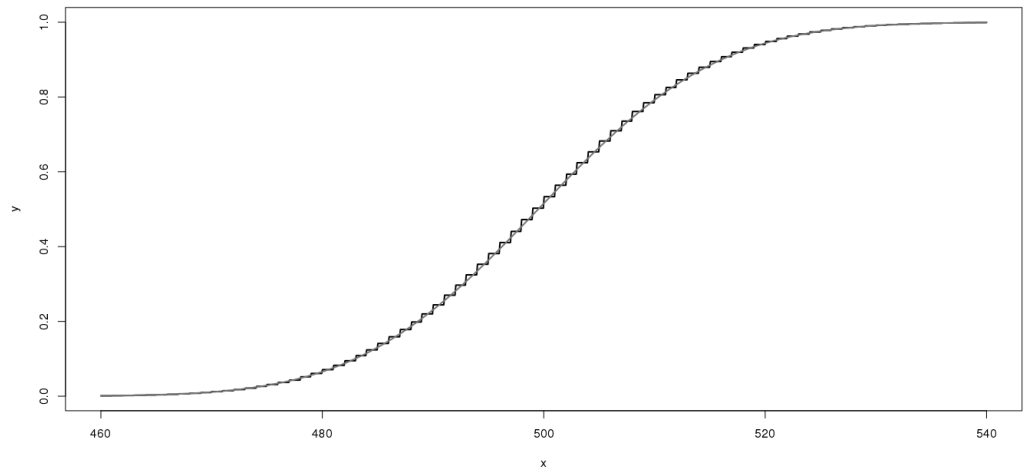

FIGURE 5.1 – La loi de X, dont la fonction de répartition est tracée en noir, est pratiquement égale à la loi normale de mêmes espérance et écart-type, tracée en gris.

```
> # "realisations.de.X".
> str (realisations.de.X)
 num [1:65536] 473 479 504 500 499 492 506 505 503 526 ...
> # Grâce aux conversions implicites entre booléens et numériques,
> # la commande « mean (simulations.de.X <= x) » nous donne la proportion
> # des simulations de X qui ont été inférieures ou égales à x, c.-
à-d.
> # (l'estimation par Monte-Carlo de) la probabilité que X soit
> # inférieure ou égale à x, alias « la valeur en x de
> # la fonction de répartition de X ».
> mean (realisations.de.X <= 470)
[1] 0.0123291
> mean (realisations.de.X <= 490)
[1] 0.2442474
> mean (realisations.de.X <= 510)
[1] 0.8062286
> # Maintenant on va tracer la fonction de répartition de X,
> # pour x variant entre 460 et 540. Pour ce faire, on appelle "x"
> # un vecteur contenant toutes les valeurs en abscisse qu'on échantillonne
> # entre 460 et 540, et "y" sera le vecteur des valeurs correspondantes
> # de la fonction de répartition.
> x = seq (from=460, to=540, length.out=2^10)
> y = numeric (length (x))
> for (i in 1 : length (x))
+ y[i] = mean (realisations.de.X <= x[i])
> # On trace la fonction de répartition de X (en noir).
> plot (x, y, type = "l", col = "black", lwd = 2)
> # On ajoute à ce tracé la courbe de la fonction de répartition
> # de notre loi normale (en gris).
> curve (pnorm (x, m, s), xname = "x", from = 460, to = 540, add = TRUE,
+ col = "gray", lwd = 2)
```

♧

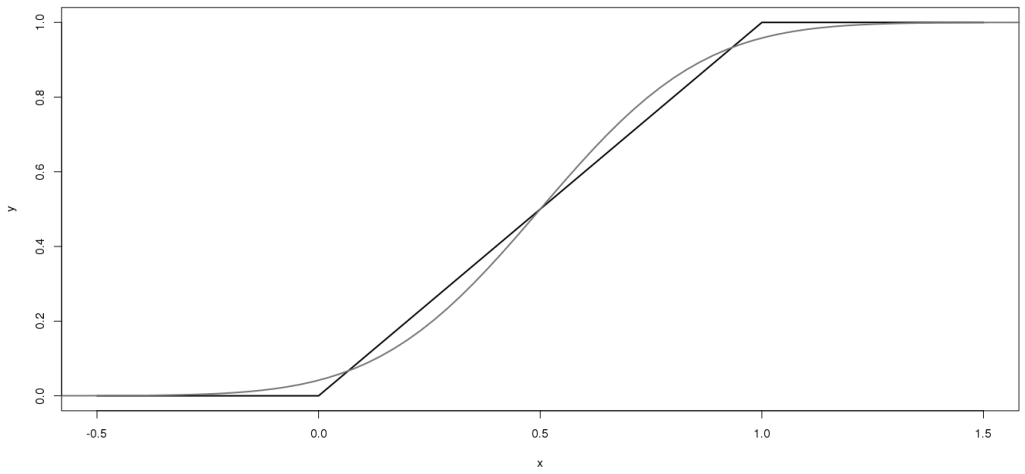

FIGURE 5.2 – Comparaison entre les fonctions de répartition d'une loi uniforme (en noir) et de la loi normale de mêmes espérance et variance (en gris) : la différence est sensible !

Exemple (GZ). Le contre-exemple suivant montre en quoi est-ce que la condition (??), selon laquelle aucune variable ne doit prédominer devant les autres, est nécessaire. Soient $X_1, \dots, X_{1\,000}$ des variables indépendantes pour lesquelles X_i est distribuée uniformément sur $\{0, 10^{-i}, 2\cdot 10^{-i}, \dots, 9\cdot 10^{-i}\}$; et soit $X \coloneqq \sum_i X_i$. Alors X est un nombre décimal de la forme $0{,}x_1x_2\dots x_{1\,000}$ dont tous les chiffres ont été tirés indépendamment uniformément entre 0 et 9 ; autrement dit, c'est un nombre aléatoire tiré uniformément dans l'ensemble $\{0, 10^{-1\,000}, 2\cdot 10^{-1\,000}, \dots, 1-10^{-1\,000}\}$, ce qui est pratiquement la même chose que la loi $\mathrm{Unif}^{\mathrm{me}}(0,1)$: X ne suit donc pas du tout une loi normale ! (voir fig. 5.2). Quand on regarde ce qui fait que le théorème (GO) ne s'applique pas, on s'aperçoit que les conditions (i), (iv) et (iii) sont bien satisfaites, mais que la condition (??) est violée, puisque l'écart-type de X_i est proportionnel à 10^{-i}, de sorte que la v.a. X_1, notamment, a un écart-type beaucoup plus important que toutes les autres. ♧

Exemple (HA). Le contre-exemple suivant montre en quoi est-ce que la condition (iii), selon laquelle les fluctuations extrêmes des variables doivent être très rares, est nécessaire. Soient $X_1, \dots, X_{1500}$ des v.a.i.i.d. suivant la loi Bernoulli(10^{-3}). Alors la somme $X =: \sum_i X_i$ suit la loi $\mathrm{Binom}^{\mathrm{le}}(1\,500, 10^{-3})$, dont vous savez depuis les classes préparatoires qu'elle est très proche de la loi Poisson(1,5) (la valeur 1,5 provenant du calcul de $1\,500\times 10^{-3}$). Ce qui n'est absolument pas une loi normale ! (voir fig. 5.3). Cette fois-ci les conditions (i), (iv) et (??) sont bien vérifiées, mais c'est la condition (iii) qui coince : en effet, la loi Bernoulli(10^{-3}), dont l'espérance est pratiquement nulle et dont la variance vaut environ 0,03, a une probabilité de 1 ‰ de fluctuer de plus de 30 écarts-types (cela se produit lorsqu'elle prend la valeur 1) ; ce qui est bien trop élevé pour une fluctuation aussi importante, car correspondant à un kurtosis très important d'environ 998 ! ♧

De même que pour la loi des grands nombres, nous pouvons formaliser notre énoncé du théorème-limite central en un résultat plus rigoureux :

Théorème (HB) (Théorème-limite central, version généralisée formelle)**.** *Soit $k > 2$. Alors il existe une fonction (explicitable) $\psi_k(\bullet) : [1,\infty[\to \mathbb{R}$, telle que pour tout $\varepsilon > 0$, on ait $\varphi_k(\tilde{n}) \overset{\tilde{n}\to\infty}{\to} 0$, pour laquelle on a la propriété suivante :*

Pour $n \in \mathbb{N}^*$, $X_1, \dots, X_n$ une famille de variable aléatoires indépendantes toutes de carrés intégrables[††]*, et $(\ell_i)_{1\leqslant i\leqslant n}$ des nombres réels strictement positifs tels qu'on*

[††]. Cette condition est en fait automatique à cause du contrôle sur les fluctuations des X_i.

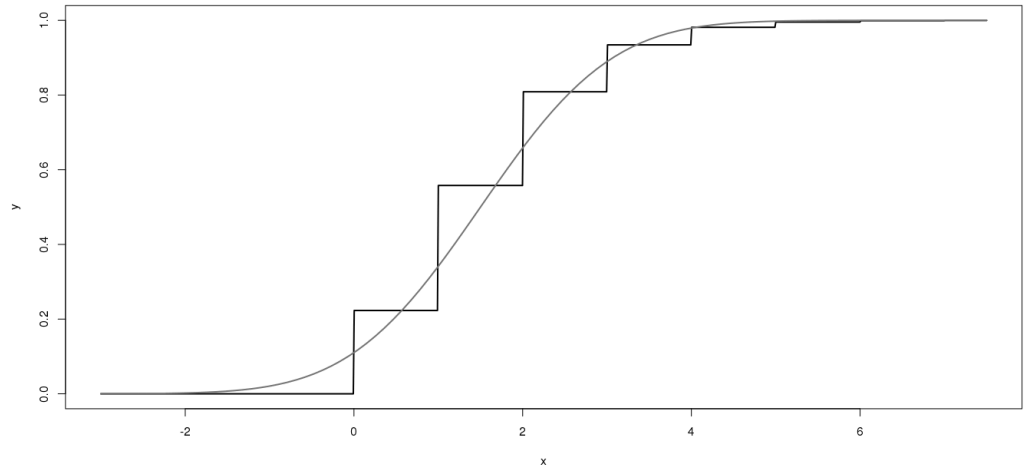

FIGURE 5.3 – Comparaison entre les fonctions de répartition d'une loi Poisson(1,5) (en noir) et de la loi normale de mêmes espérance et variance (en gris) : la différence est très sensible !

ait

$$\forall i \in [\![1,n]\!] \quad \forall z \in]1,\infty[\quad \mathbb{P}(|X_i - \mathbb{E}(X_i)| \geqslant z\ell_i) \leqslant \frac{1}{z^k}, \tag{HC}$$

$$d_{\mathrm{KS}}\Big(\mathrm{Loi}(X), \mathrm{Normale}(\mathbb{E}(X), \mathrm{Var}(X))\Big) \leqslant \psi_k\left(\frac{\sum_i \mathrm{Var}(X_i)}{\max_i \ell_i^2}\right), \tag{HD}$$

où $d_{\mathrm{KS}}(P,Q)$ *désigne la* distance de Kolmogorov-Smirnov *entre deux distributions de probabilité* P *et* Q *définies sur* $\mathbb{R}$*, définie comme la norme du supremum de la différence entre leurs fonctions de répartition :*

$$d_{\mathrm{KS}}(P,Q) \coloneqq \sup_{x\in\mathbb{R}} |\mathbb{P}(Q \leqslant x) - \mathbb{P}(P \leqslant x)| \ : \tag{HE}$$

le point ici étant en particulier que la distance de Kolmogorov-Smirnov "métrise", dans un certain sens, la convergence sur les distributions de probabilité : en particulier, une suite de lois de probabilité $(P_n)_{n\in\mathbb{N}}$ *converge vers une loi-limite* P_∞ *dès lors que* $d_{\mathrm{KS}}(P_n, P_\infty) \overset{n\to\infty}{\to} 0$*, la réciproque étant également vraie dès lors que la mesure* P_∞ *est diffuse*[‡‡]*.* ◇

Remarque (HF). On notera que, contrairement à l'énoncé analogue pour la loi des grands nombres où on requerrait d'avoir $k > 1$, ici on a besoin d'un contrôle plus fort sur les fluctuations, puisqu'on a besoin d'avoir $k > 2$: cela reflète en fait l'hypothèse, dans les énoncés classiques, qu'on avait besoin d'avoir une intégrabilité L^2 plutôt que L^1. ♧

Remarque (HG). La condition de contrôle "gentil" des fluctuations que nous avons énoncée plus haut implique les hypothèses du théorème pour $k = 4$. ♧

[‡‡]. En outre, la distance de Kolmogorov-Smirnov est invariante par les bijections croissantes, et donc en particulier par l'opération consistant à recentrer et normaliser la variance de X : par conséquent le résultat énoncé implique bien la convergence vers la loi normale standard comme on l'écrit habituellement.

Chapitre 6

Distributions de probabilité remarquables

Préambule Certaines distributions de probabilités remarquables possèdent des noms : par exemple, les lois uniformes, les lois de Poisson, les lois gamma, ... Ce chapitre en présente quelques-unes, ainsi que leurs propriétés de base. Dans le cadre de ce cours, une partie des définitions, repérées dans la marge, doivent être connues par cœur. Les propriétés de ces distributions, elles, ne sont le plus souvent pas censées être apprises à la lettre : néanmoins, dans certains cas j'attends de vous que vous soyez capables de les retrouver à partir des définitions (par exemple, retrouver la densité d'une loi exponentielle à partir de sa caractérisation). Ainsi, dans le cadre de ce chapitre, lorsqu'une propriété d'une loi est marquée dans la page comme « à retenir », cela doit être compris comme « ... ou à savoir retrouver au besoin » ; en outre, le cas échéant, les démonstrations permettant de retrouver les propriétés énoncées seront imprimées en taille normale (contrairement aux preuves fournies à pur titre de complément, pour lesquelles j'utilise les petits caractères).

Une attention plus particulière sera apportée au cas des lois normales, pour lesquelles, nonobstant ce qui précède, un certain nombre de propriétés non triviales, mais essentielles, sont à connaitre par cœur.

6.1 Familles de lois et paramétrages

Paramètres spécifiant une loi au sein d'une famille La quasi-totalité des lois que nous allons aborder ci-dessous viennent par *familles* : ainsi, même si l'on parle parfois par abus de langage de « la » loi normale (unidimensionnelle), il y a en réalité une famille de lois normales, une loi normale au sein de cette famille étant spécifiée par deux paramètres. Selon la famille de lois, il peut y avoir besoin de 0, 1, 2 ou 3 paramètres, voire plus, pour spécifier la loi.

Typologie des paramètres Pour les lois de probabilité sur $\mathbb{R}$, on a coutume de distinguer trois types de paramètres parmi ceux identifiant une distribution de probabilité au sein d'une famille donnée : les paramètres de *forme*, les paramètres d'*échelle* (ou, éventuellement, d'échelle à une certaine puissance) et les paramètres de *position* :

Définition (HH). Supposons une loi de la forme Famille($\lambda, \mu_1, \mu_2, \ldots$), où λ et les μ_i sont les paramètres de la loi. Il se peut que certaines transformations affines laissent

la famille de lois stable, au sens suivant : pour T une telle transformation affine, pour tout choix de paramètres $(\lambda, \mu_1, \mu_2, \ldots)$, la mesure-image de Famille$(\lambda, \mu_1, \mu_2, \ldots)$ par la transformation T est à nouveau une distribution de la famille Famille (mais de paramètres différents bien sûr si $T \neq \mathrm{id}$) : notons cette mesure-image Famille$(\lambda', \mu_1', \mu_2', \ldots)$. Selon la façon dont λ et λ' sont liés par la transformation T, on utilisera des adjectifs spécifiques pour qualifier le paramètre λ :

— Lorsque $\lambda' = T(\lambda)$ (pour toute transformation affine T et tout choix de paramètres $(\lambda, \mu_1, \mu_2, \ldots)$, s'entend), on dit que λ est un paramètre de *position* : informellement, λ repère une certaine position dans la distribution de probabilité $(\lambda, \mu_1, \mu_2, \ldots)$, et "bouge" donc de conserve avec la distribution de probabilité lorsqu'on lui fait subir une transformation affine.

— Lorsque, pour une transformation affine s'écrivant $T(x) = ax + b$, la relation liant λ' à λ est que $\lambda' = a\lambda$, alors λ est qualifié de « paramètre d'échelle ». Moralement, λ dit donc à quel point la distribution est étalée.

— Variante du cas précédent : lorsque la relation liant λ' à λ est de la forme « $\lambda' = a^k\lambda$ » pour un certain $k \neq 0$, λ est qualifié de « paramètre d'échelle à la puissance k ». Ainsi, pour $k = -1$, λ sera proportionnel à l'inverse de l'"étalement" de la distribution ; pour $k = 2$, il sera proportionnel au carré de son étalement ($k = 2$), etc.

— Lorsque, quelle que soit la transformation affine T, on a $\lambda' = \lambda$, alors λ est qualifié de *paramètre de forme* : moralement, cela signifie que λ ne dépend que de la forme de la distribution et reste donc invariant lorsqu'on lui applique une transformation affine. ♡

Cette définition sera éclairée par les exemples que nous verrons dans la suite du chapitre.

Attention aux choix de conventions Il est très important de savoir que le paramétrage des lois n'est pas toujours standardisé ! Un des exemples les plus marquants est celui des lois normales : certains (dont ce cours) les paramètrent en précisant leur moyenne et leur *variance*, tandis que d'autres (notamment de nombreux logiciels, comme *R*) les paramètrent en indiquant leur moyenne et leur *écart-type*. Le cas des lois exponentielles donne également lieu à plusieurs paramétrages concurrents : certains (dont ce cours) les paramètrent par leur taux, d'autres par leur demi-vie, d'autres encore par leur espérance... Le bon ingénieur, toujours soucieux d'éviter les quiproquos lors de ses collaborations [*], ne parlera donc pas de « la loi normale de paramètres $(1, 4)$ », mais plutôt de « la loi normale *d'espérance* 1 et *de variance* 4 »...! [†] Un souci similaire peut aussi être rencontré avec certains choix de conventions : par exemple, certains considèrent que les lois géométriques comptent le nombre *total* d'essais requis pour obtenir un succès (ce sont donc des lois portées par $\mathbb{N}^*$), tandis que d'autres (dont ce cours) considèrent qu'elles comptent les

[∗]. Un des exemples les plus célèbres est celui de la sonde *Mars Climate Orbiter*, qui en 1999 s'est désintégrée dans l'atmosphère martienne en raison de l'interfaçage entre les travaux d'une équipe qui travaillait en kilomètres et ceux d'une autre équipe qui travaillait en miles... Une gaffe à 600 M$! (valeur ajustée de l'inflation pour 2025).

[†]. Cela dit, au sein d'un même travail, il n'est bien entendu pas nécessaire de re-préciser les conventions employées pour une certaine famille de loi à chaque occurrence de ladite famille ! ☺ Ainsi, dans le cadre de ce cours, les conventions que j'utiliserai pour les lois classiques seront fixées une fois pour toutes au sein du présent chapitre ; et dès lors, je pourrai me contenter de parler de « la loi Expon$^{\text{le}}(\lambda)$ » sans me sentir obligé d'ajouter à chaque fois « le paramètre utilisé pour spécifier une loi exponentielle étant le taux de cette dernière »...

échecs avant un succès... Donc, là aussi, il faut préciser la convention quand il y a lieu.

Notation des familles de lois Au lycée, on vous a introduit un certain nombre de notations très abrégées pour certaines distributions de probabilité, comme '$\mathscr{G}$' pour les distributions géométriques par exemple. Je vous demande explicitement de ne pas utiliser ces abréviations dorénavant (avec une tolérance éventuelle pour la loi normale), car en statistique, le nombre de distributions "classiques" est tellement grand que cela prêterait à confusion : '$\mathscr{G}$' désigne-t-il une loi géométrique ? une loi gamma ? une loi de Gumbel ?... Pour éviter les confusions, on notera donc explicitement « Géométrique(p) », ou éventuellement une abréviation non ambigüe comme « Géom$^{\text{que}}(p)$ ».

6.2 Quelques lois remarquables discrètes

Remarque (HI). Les familles de lois portées par des structures discrètes comme $\mathbb{N}$ ne peuvent évidemment pas être invariantes par un groupe continu de transformations affines (puisqu'en général, l'image affine d'un point de $\mathbb{N}$ ne serait pas dans $\mathbb{N}$!) : dès lors, dans le cas discret il n'y a pas vraiment lieu de distinguer paramètres de position, d'échelle et de forme. ♧

Loi uniforme sur un ensemble fini

Définition (HJ) (Loi uniforme sur un ensemble fini)**.** Pour $\mathscr{X}$ un ensemble fini, la *loi uniforme* sur $\mathscr{X}$, notée $\text{Unif}^{\text{me}}(\mathscr{X})$, est la distribution de probabilité portée par $\mathscr{X}$ qui donne une masse $1/|\mathscr{X}|$ à chaque point de $\mathscr{X}$. ♡ !

> *Remarque* (HK). Nonobstant la notation utilisée, les lois uniformes sur les ensembles finis ne forment pas réellement une « famille » de lois, dans la mesure où l'ensemble $\mathscr{X}$ ne peut pas se ramener à la description d'un nombre fini fixé de réels, et n'est donc pas vraiment un « paramètre ». ♧

Un cas particulièrement fréquent est le cas où $\mathscr{X} = [\![a, b]\!]$ pour $a, b \in \mathbb{Z}$ (et $a \leqslant b$). Dans ce cas, on peut calculer explicitement l'espérance et la variance de la loi. Dans la mesure où l'espérance de la loi découle directement de la symétrie de la loi $\text{Unif}^{\text{me}}([\![a, b]\!])$ par rapport au point $(a+b)/2$, j'attends de vous que vous soyez capable de la retrouver de vous-mêmes ; pour la variance en revanche, elle vous sera redonnée le cas échéant (ou éventuellement fera l'objet d'un exercice guidé).

Théorème (HL)**.** *Pour $a, b \in \mathbb{Z}$ avec $a \leqslant b$,* !

$$\mathbb{E}\big(\text{Unif}^{\text{me}}([\![a, b]\!])\big) = \frac{a+b}{2}. \tag{HM}$$

♢

Théorème (HN)**.** *Sous les même hypothèses que le théorème (HL),*

$$\text{Var}\big(\text{Unif}^{\text{me}}([\![a, b]\!])\big) = \frac{1}{12}(b-a)^2 + \frac{1}{6}(b-a). \tag{HO}$$

♢

Remarque (HP). Attention à ne pas confondre la loi $\text{Unif}^{\text{me}}([\![a, b]\!])$, qui est la loi *discrète* sur l'ensemble d'*entiers* $\{a, a+1, \dots, b-1, b\}$, avec la loi $\text{Unif}^{\text{me}}(a, b)$, qui est la loi *continue* sur l'intervalle *réel* $]a, b[$ (confer § 6.4). ♧

Un cas particulier de loi uniforme sur un ensemble fini est lorsque cet ensemble est $\{-1,+1\}$, autrement dit, lorsqu'on tire un *signe* aléatoire uniforme. La loi $\mathrm{Unif}^{\mathrm{me}}(\{-1,+1\})$ d'un tel signe uniforme est également appelée « loi de Rademacher ».

L'espérance et la variance d'une loi de Rademacher se trouvent immédiatement par calcul direct, et doivent donc pouvoir être recalculées au besoin :

Proposition (HQ).

$$\mathbb{E}(\mathrm{Unif}^{\mathrm{me}}(\{\pm 1\})) = 0\,; \tag{HR}$$
$$\mathrm{Var}(\mathrm{Unif}^{\mathrm{me}}(\{\pm 1\})) = 1. \qquad \diamond$$

Lois de Bernoulli

! **Définition (HS)** (Loi de Bernoulli)**.** Pour $p \in [0,1]$, la *loi de Bernoulli* de paramètre p, notée $\mathrm{Bernoulli}(p)$, est la loi portée par $\{0,1\}$ qui attribue la masse p à 1 et la masse $(1-p)$ à 0. ♡

Dans la mesure où les espérance et variance d'une loi de Bernoulli s'obtiennent immédiatement par calcul direct [‡], vous êtes censés pouvoir les retrouver par vous-mêmes :

! **Proposition (HT).**

$$\mathbb{E}(\mathrm{Bernoulli}(p)) = p; \tag{HU}$$
$$\mathrm{Var}(\mathrm{Bernoulli}(p)) = p(1-p). \qquad \diamond$$

Remarque (HV). Il peut être néanmoins être pratique de retenir par cœur la formule pour la variance de la loi de Bernoulli. Un moyen mnémotechnique pour ce faire est d'observer que, puisque la loi de Bernoulli devient une masse de Dirac lorsque $p = 0$ ou $p = 1$, sa variance doit s'annuler dans ce cas, ce qui explique l'apparition du facteur « $p(1-p)$ » ; et dès lors, il faut juste retenir qu'il n'y a rien d'autre à ajouter ! ☺ ♧

Lois binomiales

! **Définition (HW)** (Loi binomiale)**.** Pour $n \in \mathbb{N}$ et $p \in [0,1]$, la *loi binomiale* de paramètres n et p, notée $\mathrm{Binom}^{\mathrm{le}}(n,p)$, est la loi de la somme de n v.a.i.i.d. $\mathrm{Bernoulli}(p)$: en d'autres termes, c'est la mesure-image par l'application « somme » (de $\mathbb{N}^n$ dans $\mathbb{N}$) de $\mathrm{Bernoulli}(p)^{\otimes n}$. Cette loi décrit donc le nombre de succès lorsqu'on procède à n expériences indépendantes ayant chacune une probabilité p de réussir. ♡

De la définition de la loi binomiale découle immédiatement la proposition suivante, très utile :

! **Proposition (HX).** *Pour $n,m \in \mathbb{N}$ et $p \in [0,1]$, si X et Y sont deux variables indépendantes avec $X \sim \mathrm{Binom}^{\mathrm{le}}(n,p)$ et $Y \sim \mathrm{Binom}^{\mathrm{le}}(m,p)$, alors $(X+Y)$ suit la loi $\mathrm{Binom}^{\mathrm{le}}(n+m,p)$. [En termes plus savants, on formule parfois cela en disant qu'on a la formule de convolution* [§] *: $\mathrm{Binom}^{\mathrm{le}}(n,p) * \mathrm{Binom}^{\mathrm{le}}(m,p) = \mathrm{Binom}^{\mathrm{le}}(n+m,p)$].* ♢

[‡]. Pour la variance, le plus simple est d'utiliser que $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, et que pour X suivant une loi de Bernoulli, $X^2 = X$: on trouve alors que la variance vaut $p - p^2$, soit $p(1-p)$.

[§]. Pour P et Q deux lois de probabilité sur $\mathbb{R}$, le *produit de convolution* de P et Q, noté $P * Q$, est la mesure-image de la loi-produit $P \otimes Q$ (qui est une loi sur $\mathbb{R}^2$) par l'application « somme » (qui va de $\mathbb{R}^2$ dans $\mathbb{R}$). Ce produit de convolution décrit donc la loi de la somme $(X+Y)$ lorsque X et Y sont deux v.a. indépendantes de lois respectives P et Q.

Remarque (HY). Ne surtout pas oublier l'hypothèse d'indépendance, qui est essentielle ! Ne pas oublier non plus que les seconds paramètres des deux lois binomiales doivent être égaux ! ♧

La fonction de masse d'une loi de Bernoulli se retrouve immédiatement à partir de sa définition (nous rappelons le raisonnement ci-dessous), et on obtient alors la proposition suivante :

Proposition (HZ). *La loi* $\mathrm{Binom}^{\mathrm{le}}(n,p)$ *est portée par* $[\![0,n]\!]$ *; et pour* $k \in [\![0,n]\!]$ *:* !

$$\mathbb{P}\big(\mathrm{Binom}^{\mathrm{le}}(n,p)=k\big) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{IA}$$

♢

Démonstration. En effet, soient $X_1, \dots, X_n$ i.i.d. Bernoulli(p) et S leur somme. La probabilité que S vaille k est la probabilité que k exactement des X_i vaillent 1 et que les $(n-k)$ autres vaillent zéro. Cet évènement se décompose en $\binom{n}{k}$ évènements disjoints, selon quels sont les indices i pour lesquels X_i vaut 1 (le nombre de façons de choisir cet ensemble d'indices correspond bien à la sélection de k éléments parmi n). Maintenant, quelle est la probabilité que les k variables X_i ayant les bons indices vaillent 1 et que les $(n-k)$ autres variables vaillent 0 ? Ces variables étant indépendantes, il faut faire le produit des probabilités pour chaque possibilité individuelle : on a ainsi k facteurs p (pour les X_i devant valoir 1) et $(n-k)$ facteurs $(1-p)$ (pour ceux devant valoir 0), ce qui conduit bien au résultat recherché (en observant que nos $\binom{n}{k}$ évènements disjoints se trouvent ici être équiprobables). ♤

> *Remarque* (IB). En revanche, la formule du coefficient binomial $\binom{n}{p} = n!\,/\,p!(n-p)!$ n'est pas exigible dans ce module, et sera rappelée au besoin. (Mais la *définition* du coefficient binomial, à savoir le nombre de façons de choisir p éléments parmi n, doit être sue !). ♧

En utilisant la définition de la loi binomiale à partir de lois de Bernoulli (et en utilisant resp. la linéarité de l'espérance et la formule d'additivité des variances pour des v.a. indépendantes), on a la proposition suivante, dont vous devez savoir retrouver les résultats :

Proposition (IC). *Pour* $n \in \mathbb{N}$*,* $p \in [0,1]]$ *:* !

$$\mathbb{E}(\mathrm{Binom}^{\mathrm{le}}(n,p)) = np\,; \tag{ID}$$
$$\mathrm{Var}(\mathrm{Binom}^{\mathrm{le}}(n,p)) = np(1-p). \qquad \diamond$$

Lois de Poisson

Définition (IE) (Loi de Poisson)**.** Pour $\theta \in \mathbb{R}_+$, la *loi de Poisson* de paramètre θ, notée Poisson(θ), est la loi du nombre d'évènements qui se produisent lorsqu'on dispose d'un très grand nombre d'évènements possibles, que ces évènements sont indépendants, que chacun de ces évènements a une probabilité individuelle très petite de se produire, et que l'espérance du nombre d'évènements se produisant vaut θ. On a donc $\mathbb{E}(\mathrm{Poisson}(\theta)) = \theta$ *par définition.* ♡ !

Remarque (IF). Autant la *formule* de la loi de Poisson, qui est compliquée et ne peut pas être retenue par un "truc" simple, n'est pas à connaitre par cœur dans le cadre de ce cours, autant j'estime qu'il est vraiment important de connaitre la définition informelle ci-dessus : en effet, la pertinence qu'il peut y avoir (ou pas) à utiliser une loi de Poisson dans une modélisation découlera du fait qu'on peut considérer qu'on se trouve (ou pas) dans la situation de la définition (IE) ! ♧

Remarque (IG). On définit parfois la loi Poisson(θ) comme la limite de la loi $\mathrm{Binom}^{\mathrm{le}}(n, n^{-1}\theta)$ lorsque n tend vers l'infini. Cette façon de voir les choses, qui a l'avantage d'être plus rigoureuse, revient en fait à considérer, dans la définition (IE) ci-dessus, le cas où les évènements individuels sont tous équiprobables. Cependant, en réalité la loi de Poisson est bien plus générale que cela, puisqu'elle apparait même en l'absence d'équiprobabilité : tout ce qui importe, c'est que tous les évènements individuels aient chacun une probabilité minuscule ! ♧

De la définition de la loi de Poisson découle l'importante propriété suivante (analogue à la proposition (HX)). (Et, comme dans le cas de la proposition (HX)), on veillera à ne pas oublier l'hypothèse d'indépendance !) :

! **Proposition** (IH). *Pour $\theta, \lambda \in \mathbb{R}_+$, si X et Y sont des variables aléatoires indépendantes de lois resp.* Poisson(θ) *et* Poisson(λ)*, alors $X+Y$ suit la loi* Poisson($\theta+\lambda$)*. Ou encore, en termes de convolution :* $\mathrm{Poisson}(\theta) * \mathrm{Poisson}(\lambda) = \mathrm{Poisson}(\theta+\lambda)$. ♢

La fonction de masse d'une loi de Poisson peut être exprimée analytiquement, même si je ne vous demande pas de retenir ce résultat :

Théorème (II). *La loi* Poisson(θ) *est portée par $\mathbb{N}$; et pour tout $n \in \mathbb{N}$:*

$$\mathbb{P}(\mathrm{Poisson}(\theta) = n) = \mathrm{e}^{-\theta}\frac{\theta^n}{n!}. \tag{IJ}$$

♢

Démonstration. Notons $(A_i)_{i\in\mathcal{I}}$ l'ensemble des évènements considérés, $(p_i)_{i\in\mathcal{I}}$ leurs probabilités respectives, et posons $\varepsilon := \max_{i\in\mathcal{I}} p_i$. (Rappelons qu'avec nos hypothèses, on a $|\mathcal{I}| \gg 1$, $p_i \ll 1$ $\forall i \in \mathcal{I}$, et $\varepsilon \ll 1$). Notons $S := \sum_{i\in\mathcal{I}} \mathbf{1}_{A_i}$ la variable aléatoire qui compte combien de A_i sont réalisés, et fixons-nous $k \in \mathbb{N}$: notre objectif est alors de calculer (aux approximations près découlant des hypothèses) la probabilité $\mathbb{P}(S = k)$.

Dans la suite de cette démonstration, nous noterons $\mathcal{P}_k(\mathcal{I})$ l'ensemble des parties de $\mathcal{I}$ de cardinal k, et similairement $\mathcal{P}_{<k}(\mathcal{I})$ l'ensemble des parties de cardinal $< k$. Avec ces notations, l'évènement $\{S = k\}$ se décompose comme l'union disjointe d'évènements de la forme

$$\{A_i \text{ pour tout } i \in J \text{ et non-}A_i \text{ pour tout } i \in \mathcal{I} \setminus J\} \tag{IK}$$

pour $J \in \mathcal{P}_k(\mathcal{I})$.

Or, par indépendance, la probabilité de l'évènement ci-dessus vaut

$$\prod_{i\in J} p_i \times \prod_{i\in\mathcal{I}\setminus J} p_i = \prod_{i\in J} \frac{p_i}{1-p_i} \times \prod_{i\in\mathcal{I}} (1-p_i), \tag{IL}$$

ce qui est (quasiment) égal à $\prod_{i\in J} p_i \times \mathrm{e}^{-\theta}$: en effet, d'une part le produit $\prod_{i\in J}(1-p_i)^{-1}$ est quasiment égal à 1 (du fait que tous les $(1-p_i)^{-1}$ sont très proches de 1, et qu'on a multiplié seulement k facteurs) ; et d'autre part, le produit $\prod_{i\in\mathcal{I}}(1-p_i)$ s'approxime par $\prod_{i\in\mathcal{I}} \mathrm{e}^{-p_i} = \exp(-\sum_{i\in\mathcal{I}} p_i) \underset{\text{déf}}{=} \mathrm{e}^{-\theta}$. Finalement, on a déduit que, aux approximations près,

$$\mathbb{P}(S = k) = \mathrm{e}^{-\theta} \sum_{J\in\mathcal{P}_k(\mathcal{I})} \prod_{i\in J} p_i. \tag{IM}$$

Ci-dessus, nous avons une somme qui porte sur des *ensembles* de taille k. Il serait plus pratique d'avoir une somme portant sur des *k-uplets* : en effet, si $(i_0, \dots, i_{k-1})$ est un k-uplet vérifiant $\{i_0, \dots, i_{k-1}\} = J$, on pourra ré-écrire $\prod_{i\in J} p_i$ comme $\prod_{q=0}^{k-1} p_{i_q}$. Mais attention, si nous faisons la somme sur tous les k-uplets de $\mathcal{I}^k$, nous allons devoir prendre en compte deux subtilités :

— D'une part, que chaque ensemble J sera encodé par $k!$ k-uplets distincts, selon l'ordre dans lequel on listera ses éléments ;
— D'autre part, que les k-uplets comportant au moins une paire de composantes identiques ne correspondront à aucun J de $\mathscr{P}_k$.

La prise en compte de ces deux subtilités nous conduit alors à la formule :

$$n!\mathrm{e}^\theta\,\mathbb{P}(S=k)=\sum_{\vec{\imath}\in\mathscr{I}^k}\prod_{q=0}^{k-1}p_{i_q}-\sum_{\vec{\imath}\in\Delta}\prod_{q=0}^{k-1}p_{i_q},\tag{IN}$$

où la notation '$\vec{\imath}$' désigne un k-uplet $(i_0,\dots,i_{k-1})$, et où Δ désigne l'ensemble des k-uplets de $\mathscr{I}^k$ ayant au moins deux composantes identiques.

Dans la formule ci-dessus, la première somme du membre de droite se factorise en

$$\prod_{q=0}^{k-1}\sum_{i_q\in\mathscr{I}}p_{i_q}\underset{\text{déf}}{=}\prod_{q=0}^{k-1}\theta=\theta^k,\tag{IO}$$

ce qui est précisément ce dont nous avons besoin pour arriver à notre formule de la loi de Poisson : ainsi, il ne reste plus qu'à démontrer que la seconde somme est négligeable pour conclure la démonstration !

Nous souhaitons donc, pour finir, contrôler $\sum_{\vec{\imath}\in\Delta}\prod_{q=0}^{k-1}p_{i_q}$. Pour $(i_0,\dots,i_{k-1})=:\vec{\imath}$ un élément de Δ, notons $E(\vec{\imath}):=\{i_0,\dots,i_{k-1}\}$: du fait que nous avons supposé $\vec{\imath}\in\Delta$, nous avons $|E(\vec{\imath})|<k$, et donc (puisque tous les p_i sont $\leqslant\varepsilon$, et que le produit en version k-uplet compte au moins un facteur en double)

$$\prod_{q=0}^{k-1}p_{i_q}\leqslant\varepsilon\prod_{j\in E(\vec{\imath})}p_j.\tag{IP}$$

En sommant terme à terme, nous obtenons ainsi que

$$\sum_{\vec{\imath}\in\Delta}\prod_{q=0}^{k-1}p_{i_q}\leqslant\varepsilon\sum_{\vec{\imath}\in\Delta}\prod_{j\in E(\vec{\imath})}p_j.\tag{IQ}$$

Dans la somme ci-dessus, nous avions sommé "$\vec{\imath}$ par $\vec{\imath}$" ; nous allons maintenant la ré-écrire pour la sommer "$E(\vec{\imath})$ par $E(\vec{\imath})$". Notons déjà que tous les $E(\vec{\imath})$ de la somme ci-dessus sont dans $\mathscr{P}_{<k}(\mathscr{I})$. Néanmoins, pour $J\in\mathscr{P}_{<k}(\mathscr{I})$, il peut y avoir plusieurs k-uplets $\vec{\imath}$ tels que $E(\vec{\imath})=J$... Mais nous savons qu'il y en a au plus $(k-1)^k$, puisqu'on tel $\vec{\imath}$ doit nécessairement avoir toutes ses composantes dans J ! On a donc la majoration suivante :

$$\sum_{\vec{\imath}\in\Delta}\prod_{j\in E(\vec{\imath})}p_j\leqslant(k-1)^k\sum_{J\in\mathscr{P}_{<k}(\mathscr{I})}\prod_{j\in J}p_j.\tag{IR}$$

Mais ici nous observons que, grâce à l'identité (IM) établie quelques lignes plus haut, la somme de produits qui apparait dans le membre de droite n'est autre que $\mathrm{e}^\theta\,\mathbb{P}(S<k)$, et se trouve donc en particulier être inférieure à e^θ ! En fin de compte, nous avons obtenu que $\sum_{\vec{\imath}\in\Delta}\prod_{q=0}^{k-1}p_{i_q}\leqslant\varepsilon(k-1)^k\mathrm{e}^\theta$, qui est bien négligeable devant 1 (dans la mesure où les valeurs k et θ, elles, sont supposées modérées) : cela conclut notre démonstration ☺ ♤

La variance d'une loi de Poisson peut être retrouvée par un argument particulièrement court quoique astucieux. Je ne vous demande pas de savoir la retrouver de but en blanc, mais éventuellement à l'aide d'une petite indication :

Théorème (IS). *Pour tout $\theta\in\mathbb{R}_+$,*

$$\mathrm{Var}(\mathrm{Poisson}(\theta))=\theta.\tag{IT}$$

♢

Démonstration. Nous avons dit qu'une v.a. de Poisson était égale à la somme d'un grand nombre de v.a. de Bernoulli indépendantes de petit paramètre. Or pour une loi de Bernoulli de paramètre p, l'espérance vaut p tandis que la variance vaut $p(1-p)$: par conséquent, pour p très proche de 0, l'espérance est pratiquement égale à la variance. Or l'espérance d'une somme de v.a. (indépendantes ou pas) est égale à la

somme de leurs espérances, tandis que la variance d'une somme de v.a. indépendantes est égale à la somme de leurs variances. Dès lors, puisqu'ici nous sommes partis de v.a. dont les espérances étaient égales aux variances respectives, et que ces variables sont indépendantes, l'espérance de la somme (donc, de la loi de Poisson) sera égale à la variance de la somme. D'où le résultat annoncé. ♤

Lois géométriques

Notation (IU). Attention ! Dans ce cours (ainsi que certains logiciels, par exemple *R*), la définition de la loi géométrique sera différente de celle que vous utilisiez jusqu'à présent : alors qu'on vous avait défini la loi géométrique de paramètre p comme le nombre d'*essais* nécessaires avant d'obtenir un succès lorsque chaque essai réussit avec probabilité p, ici $\text{Géom}^{\text{que}}(p)$ désignera le nombre d'*échecs* avant le premier succès ! Autrement dit, si l'on désigne par '$\mathscr{G}$' la loi géométrique "version prépa", nous prendrons ici $\text{Géom}^{\text{que}}(p) \coloneqq \mathscr{G}(p) - 1$. (au sens de « mesure-image de la loi $\mathscr{G}(p)$ par la fonction "soustraire 1" »).

(Cependant, dans la mesure où ce choix crée un risque d'erreur particulièrement élevé, je rappellerai systématiquement la convention retenu si des lois géométriques interviennent dans des exercices de travaux dirigés ou des rpbolèmes d'examen ☺). ♡

! **Définition (IV)** (Loi géométrique)**.** Pour $p \in]0,1]$, la *loi géométrique* de paramètre p, notée $\text{Géom}^{\text{que}}(p)$, est le nombre d'échecs avant d'obtenir un succès, lorsqu'on retente indéfiniment et indépendamment une expérience ayant une probabilité de succès p. Autrement dit, lorsque les $(X_i)_{i\in\mathbb{N}}$ sont des v.a.i.i.d. Bernoulli(p), c'est la loi de $\min\{i \in \mathbb{N} \mid X_i = 1\}$. ♡

À partir de cette définition, vous devez être capables de retrouver rapidement l'expression de la fonction de masse d'une loi géométrique :

! **Proposition (IW).** *La loi* $\text{Géom}^{\text{que}}(p)$ *est portée par* $\mathbb{N}$ *; et pour tout* $n \in \mathbb{N}$*, on a :*

$$\mathbb{P}(\text{Géom}^{\text{que}}(p) = n) = (1-p)^n p. \tag{IX}$$

♢

Démonstration. Une manière de démontrer cette formule est de commencer par observer qu'on a

$$\mathbb{P}(\text{Géom}^{\text{que}}(p) \geqslant n) = (1-p)^n : \tag{IY}$$

En effet, pour X suivant la loi géométrique, dire que $\{X \geqslant n\}$ revient à dire que les n premières expériences ont été des échecs : chacun de ces expériences étant indépendante et ayant une probabilité d'échec de $1-p$, le calcul est immédiat. On en déduit ensuite la probabilité que X vaille n en observant que $\{X = n\} = \{X \geqslant n\} \setminus \{X \geqslant n+1\}$, le second ensemble du membre de droite étant inclus dans le premier.

Une autre preuve, peut-être plus directe, consiste à remarquer que notre variable géométrique vaut n si et seulement si les n premières expériences sont ratées *et* que la $(n+1)$-ième est réussie. ♤

Remarque (IZ). Il peut être utile de retenir qu'il y a une formule particulièrement simple (et simple à démontrer) pour la probabilité qu'une loi géométrique dépasse une certaine valeur, d'autant qu'il y a une formule extrêmement similaire (et avec une preuve similaire) pour la loi exponentielle. ♧

Remarque (JA). On notera que, puisque p a été supposé strictement positif, la probabilité que les n premières expériences échouent toutes tend vers 0 lorsque n tend vers l'infini, autrement dit qu'il y aura forcément (au sens de « presque-surement ») une expérience qui finira par réussir — ce qui est au demeurant intuitivement évident — : ainsi, la loi géométrique prend toujours une valeur finie et est donc bien portée par $\mathbb{N}$. ♧

Voyons maintenant les formules sur l'espérance et la variance de la loi géométrique. Je ne vous demande pas d'être capables de remontrer directement la formule concenant l'espérance, mais éventuellement avec une petite indication. (Pour la variance, ce serait le cas échéant l'objet d'un exercice guidé).

Théorème (JB). *Pour tout $p \in]0,1]$,*

$$\mathbb{E}(\text{Géom}^{\text{que}}(p)) = (1-p)/p \tag{JC}$$
$$\text{Var}(\text{Géom}^{\text{que}}(p)) = (1-p)/p^2. \tag{JD}$$

♢

Démonstration. Commençons par la preuve concernant l'espérance. Puisque l'expérience réussit avec probabilité p et échoue avec probabilité $1-p$, il y a en moyenne $(1-p)/p$ échecs pour un succès ; or la loi géométrique compte bien le nombre d'échecs pour un succès pris au hasard, lorsqu'on associe chaque échec au succès qui le suit directement [¶]. D'où le résultat.

Pour la variance, on écrit que si X suit une loi géométrique, par la formule de transfert :

$$\mathbb{E}(X^2) = \sum_{n=0}^{\infty} n^2(1-p)^n p. \tag{JE}$$

Appelons M_2 cette espérance, et notons qu'en effectuant le changement de variable $n \leftarrow n+1$, on a aussi

$$M_2 = \sum_{n=1}^{\infty} (n-1)^2(1-p)^{n-1} p. \tag{JF}$$

Par différence terme à terme (sachant que toutes nos sommes sont absolument convergentes),

$$\begin{aligned} pM_2 = M_2 - (1-p)M_2 &= \underbrace{\left(n^2(1-p)^n p\right)_{|n=0}}_{=0} + \sum_{n=1}^{\infty} \left(n^2 - (n-1)^2\right)(1-p)^n p \\ &= \sum_{n=1}^{\infty} (2n-1)(1-p)^n p = \sum_{n=0}^{\infty} (2n-1)(1-p)^n p - (-p) = p + \mathbb{E}(2X-1), \end{aligned}$$

la dernière égalité provenant de la formule de transfert. Or $\mathbb{E}(2X-1) = 2\,\mathbb{E}(X) - 1 = (2-3p)/p$ en vertu de la formule précédemment calculée, d'où $pM_2 = (2-3p+p^2)/p$ et donc $\mathbb{E}(X^2) = (2-3p+p^2)/p^2$. Finalement, on conclut que

$$\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{2-3p+p^2}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{1-p}{p^2}. \tag{JG}$$

[¶]. Il convient cependant d'être très précautionneux avec ce genre d'argument, qui peut facilement conduire à des erreurs (qualifiées de « biais ») si on n'est pas assez attentif : par exemple, on pourrait être tenté de suivre l'argument suivant : « lorsqu'une loi géométrique vaut n, cela signifie qu'on a eu 1 succès pour n échecs, soit un taux de succès moyen pour cette tentative de $1/(n+1)$. Or nous savons qu'en moyenne, le taux de succès vaut p ; lorsqu'on prend l'espérance de $1/(n+1)$ pour la loi géométrique, on s'attend donc à trouver p ». Or en réalité $\mathbb{E}\big(1/(\text{Géom}^{\text{que}}(p)+1)\big)$ ne vaut pas p, mais $p|\ln p|/(1-p) > p$... Le problème vient de ce que, lorsqu'on parle du taux global d'échecs, les expériences où la loi géométrique a pris une grande valeur (et donc où $1/(n+1)$ a pris une petite valeur) ont un rôle plus grand que celles où elle a pris une petite valeur, puisqu'il y a eu besoin de plus d'essais dans le premier cas... Ainsi, faire la moyenne sur les essais et faire la moyenne sur les expériences n'est pas la même chose ! En ce qui concerne l'argument présenté dans la démonstration, heureusement, ce biais ne se manifeste pas : en effet, faire la moyenne par succès et faire la moyenne par expérience revient bien au même, puisqu'il y a toujours exactement un succès par expérience ☺

♤

6.3 Les lois normales

Les lois normales sont tellement importantes, notamment en modélisation statistique, qu'elles méritent une section à elles seules ! C'est donc d'elles que nous allons parler dans ce qui suit.

Lois normales unidimensionnelles

Notation (JH). Attention : Dans ce cours, conformément à ce que vous avez vu au lycée et au semestre 5, le second paramètre d'une loi normale unidimensionnelle désignera sa *variance*. Cependant, de nombreux logiciels, par exemple *R*, utilisent pour leur part un paramétrage par *écart-type* ! ♡

! **Définition (JI)** (Loi normale)**.** Pour $\mu \in \mathbb{R}$ et $\sigma \geqslant 0$, la *loi normale* d'espérance μ et de variance σ^2 [‖], notée Normale(μ, σ^2), est la loi suivie par une variable aléatoire vérifiant les propriétés suivantes :

- La variable aléatoire s'écrit comme la somme d'un très grand nombre de « sous-variables » indépendantes ;
- Chaque sous-variable est concentrée dans une très petite plage de valeurs (ce qui est plus ou moins équivalent à dire que chaque sous-variable a un très petit écart-type) ;
- Les sous-variables ne sont pas trop "sauvages", au sens où il est extrêmement rare qu'elles prennent un comportement très différent de leur comportement typique [∗∗],

lorsque cette variable aléatoire a pour espérance μ et pour variance σ^2 [††]. On a donc *par définition* $\mathbb{E}\big(\text{Normale}(\mu, \sigma^2)\big) = \mu$ et $\text{Var}\big(\text{Normale}(\mu, \sigma^2)\big) = \sigma^2$. ♡

Remarque (JJ). La loi normale peut être vue comme l'analogue, dans le cas continu, de la loi binomiale dans le cas discret [‡‡] ; en particulier, le théorème-limite central montre que pour tout $p \in]0, 1[$, lorsque n tend vers l'infini, la loi Binom$^{\text{le}}(n, p)$ devient asymptotiquement "équivalente" à la loi Normale$\big(np, np(1-p)\big)$. ♧

! **Proposition (JK).** *Si X suit la loi* Normale(μ, σ^2)*, alors pour $a, b \in \mathbb{R}$, la variable aléatoire $aX + b$ suit la loi* Normale$(a\mu + b, a^2\sigma^2)$. ♢

Remarque (JL). En fait, il y a juste besoin de retenir que $aX + b$ suit une loi normale : l'espérance et l'écart-type de celle-ci se déduisent resp. de la linéarité de l'espérance et du caractère quadratique de la variance. ♧

[‖]. Il est d'usage, lorsque cela n'alourdit pas trop la notation, d'écrire le second paramètre d'une loi normale unidimensionnelle sous la forme d'un carré. Le paramètre σ représente alors l'*écart-type* de la loi normale, qui, contrairement à la variance, a le bon gout d'avoir la même homogénéité physique que la variable aléatoire elle-même, et donc d'être plus facile à interpréter intuitivement.

[∗∗]. Je ne préciserai pas ce point ici : pour les curieux, *confer* l'énoncé du théorème-limite central dans le chapitre 5.

[††]. σ correspond donc à l'écart-type, sous réserve qu'il ait été pris positif évidemment.

[‡‡]. Par contre, le paramétrage des deux lois est différent : il n'y a pas d'homologie entre n et p d'une part, et μ et σ d'autre part.

Démonstration. Si nous notons $(X_i)_{i\in I}$ les sous-variables dont X est la somme, on peut voir $aX+b$ comme la somme des sous-variables aX_i+b. Il est alors facile de s'assurer que ces sous-variables vérifient elles aussi les hypothèses de la définition (JI), de sorte que $aX + b$ est bien aussi une loi normale. ♤

Remarque (JM). Attention, concernant les écarts-types, on a $\mathrm{Var}^{1/2}(aX + b) = (a^2\sigma^2)^{1/2} = |a|\sigma$, avec apparition d'une valeur absolue ! Pour éviter ce piège, bien garder à l'esprit qu'un écart-type est nécessairement positif. ♧

Remarque (JN). On observera en particulier que si X suit une loi Normale(μ, σ^2), alors la variable symétrique de X par rapport à μ, c.-à-d. $\mu - (X - \mu) = 2\mu - X$, suit la même loi que X. Autrement dit, la loi Normale(μ, σ^2) est *symétrique* par rapport à l'abscisse μ (cela se voit bien notamment sur le tracé de la densité), ce qui est une propriété qu'il est important de garder à l'esprit pour visualiser ce qui se passe. ♧

Corolaire (JO)**.** *Pour $\mu \in \mathbb{R}$ et $\sigma \in \mathbb{R}_+^*$, si X suit la loi* Normale$(0, 1)$*, alors $\sigma X + \mu$ suit la loi* Normale(μ, σ^2)*. "Réciproquement", si Y suit la loi* Normale(μ, σ^2)*, alors $(Y - \mu) / \sigma$ suit la loi* Normale$(0, 1)$*.* ♢ !

Remarque (JP). On voit donc que toutes les lois normales (à l'exception du cas où l'écart-type est nul, qui est dégénéré) se déduisent les unes des autres par des transformations affines : elles ont donc toutes la même *forme*. En outre, en termes de typologie des paramètres (confer § 6.1 *supra*), on peut observer que l'espérance est un paramètre de position, tandis que la variance est un paramètre d'échelle au carré (et l'écart-type un paramètre d'échelle "tout court" [∗]). ♧

Définition (JQ) (Loi normale standard)**.** La loi Normale$(0, 1)$ est appelée *loi normale standard.* ♡ !

Remarque (JR). En vertu de la Proposition (JK), il suffit de comprendre le comportement de la loi normale standard pour en déduire le comportement de toutes les autres lois normales. Par exemple, le fait qu'une loi normale standard ait 95 % de chances de tomber dans l'intervalle $[-1{,}96, 1{,}96]$ se traduit par le fait qu'une loi Normale(μ, σ^2) a 95 % de chances de tomber dans l'intervalle $[\mu - 1{,}96\,\sigma, \mu + 1{,}96\,\sigma]$. Cela était particulièrement important à l'époque où les propriétés (fonctions de répartition et quantiles) d'une loi de probabilité se retrouvaient à l'aide de *tables* : autant il aurait été impossible de tabuler toutes les lois normales envisageables, autant il suffisait de simplement tabuler la loi normale standard pour pouvoir faire tous les calculs !

Aujourd'hui, la plupart des logiciels qui permettent de faire des calculs sur la loi normale implémentent en interne la façon de passer d'une loi normale standard à une loi normale quelconque (c'est notamment le cas de *R*), de sorte qu'il est moins critique de savoir faire cette transformation ; cela dit, je ne saurais que trop vous recommander d'avoir quand même bien en tête la façon dont une loi normale quelconque correspond à une loi normale standard, parce que cela est très utile pour visualiser ce qui se passe ! ♧ !

[∗]. Attention toutefois : de manière générale, lorsqu'on applique la transformation affine $x \mapsto ax + b$ à une variable aléatoire gaussienne d'écart-type σ, l'image de cette v.a. aura pour écart-type $|a|\sigma$, ce qui n'est pas la même chose que $a\sigma$ lorsque $a < 0$; mais il n'en demeure pas moins que σ est bien un paramètre d'échelle lorsqu'on s'intéresse aux transformations affines *croissantes*.

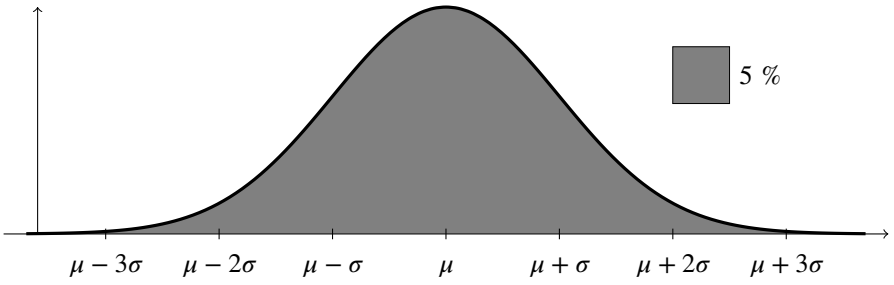

FIGURE 6.1 – Allure de la loi Normale(μ, σ^2)

Proposition (JS). *Si X et Y sont deux variables aléatoires* indépendantes *de lois resp.* Normale(μ, v) *et* Normale(ν, w)*, alors $X + Y$ suit la loi* Normale$\big(\mu + \nu, v + w\big)$*.* ♢

Remarque (JT). En fait, il y a juste besoin de retenir que $X+Y$ suit une loi normale : l'espérance et la variance de celle-ci se déduisent resp. de la linéarité de l'espérance et de l'additivité de la variance (attention : de la *variance*, pas de l'écart-type !) pour des v.a. indépendantes. ♧

Démonstration. On peut fabriquer la loi de $X+Y$ en prenant les sous-variables de X et celles de Y indépendantes ; et on se retrouve alors avec un seul grand ensemble de sous-variables qui satisfait toutes les conditions pour que sa somme, c.-à-d. $X + Y$, suive une loi normale. ♤

Théorème (JU). *Pour $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*$, la loi* Normale$(\mu, \sigma^2)$ *est à densité sur $\mathbb{R}$, avec (notant* dx *un voisinage infinitésimal de x) :*

$$\mathbb{P}\big(\text{Normale}(\mu, \sigma^2) \in \mathrm{d}x\big) = \tfrac{1}{\sqrt{2\pi}}\sigma^{-1} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{vol}_1(\mathrm{d}x). \qquad \text{(JV)}$$

♢

Remarque (JW). En revanche, il n'y a pas de formule fermée concernant la fonction de répartition de la loi normale, c.-à-d. pour calculer $\mathbb{P}\big(\text{Normale}(\mu, \sigma^2) \leqslant x\big)$. De nombreux logiciels implémentent tout de même la calcul de cette fonction à l'aide d'une fonction « spéciale » : soit la fonction de répartition de la loi normale standard, généralement notée Φ, soit la « fonction d'erreur », notée erf, définie par $\mathrm{erf}(x) \coloneqq 2\Phi\big(\sqrt{2}x\big) - 1$. ♧

La figure 6.1 vous donne l'allure des lois normales. Notez en particulier que, bien que la loi normale ne soit pas bornée *stricto sensu*, la « courbe en cloche » formée par cette loi s'atténue néanmoins extrêmement vite aux extrêmités !

Il peut être utile de retenir, au moins approximativement, les valeurs suivantes :

Proposition (JX). *On a les valeurs suivantes (arrondies par excès) :*

$$\mathbb{P}(|\text{Normale}(0,1)| > 1) = 32\ \%\ ; \qquad \text{(JY)}$$
$$\mathbb{P}(|\text{Normale}(0,1)| > 2) = 4{,}6\ \%\ ; \qquad \text{(JZ)}$$
$$\mathbb{P}(|\text{Normale}(0,1)| > 3) = 2{,}7\ ‰. \qquad \text{(KA)}$$

En pratique, on pourra retenir qu'il y a environ une chance sur 3 qu'une loi normale tombe à plus d'un écart-type de son espérance, seulement une chance sur 20 qu'elle tombe à plus de deux écarts-types, et à peine une chance sur 400 qu'elle tombe à plus de trois écarts-types ! ◇

L'intervalle de fluctuation au risque 5 % pour les lois normales est tellement utilisé en pratique que, même si je ne vous demande pas formellement de l'apprendre par cœur, on peut le considérer comme relevant des connaissances standard exigibles de tout ingénieur :

Proposition (KB). *L'intervalle de fluctuation au risque 5 % pour la loi normale standard est* $[\pm 1{,}96]$. *Corolairement, l'intervalle de fluctuation au risque 5 % pour la loi* $\mathrm{Normale}(\mu, \sigma^2)$ *est* $[\mu \pm 1{,}96\,\sigma]$. ◇

Lois normales multidimensionnelles

Définition (KC) (Loi normale multidimensionnelle)**.** Soit $d \in \mathbb{N}$, $\vec{\mu} \in \mathbb{R}^d$ et $\boldsymbol{C} \in \mathbb{R}^{d\times d}$ une matrice symétrique positive (au sens des formes quadratiques). La *loi normale d-dimensionnelle* de paramètres $\vec{\mu}$ et $\boldsymbol{C}$, notée $\mathrm{Normale}(\vec{\mu}, \boldsymbol{C})$ [†], est la loi suivie par une variable aléatoire vérifiant les propriétés suivantes :

— La variable aléatoire s'écrit comme la somme d'un très grand nombre de « sous-variables » (à valeurs dans $\mathbb{R}^d$) indépendantes ;
— Chaque sous-variable est concentrée dans une très petite plage de valeurs ;
— Les sous-variables ne sont pas trop "sauvages",

lorsque cette variable aléatoire a pour espérance (au sens vectoriel) $\vec{\mu}$ et pour matrice de covariance $\boldsymbol{C}$. On a donc par définition $\mathbb{E}\big(\mathrm{Normale}(\vec{\mu}, \boldsymbol{C})\big) = \vec{\mu}$ et $\mathrm{Var}\big(\mathrm{Normale}(\vec{\mu}, \boldsymbol{C})\big) = \boldsymbol{C}$. ♡

Remarque (KD). Notez que je note la « matrice de covariance » comme une variance, car une matrice de covariance n'est jamais que l'analogue multidimensionnel d'une variance : par exemple, si $\vec{X}$ est à valeurs vectorielles, on a, en termes matriciels, $\mathrm{Var}(\vec{X}) = \mathbb{E}(\vec{X}\vec{X}^\mathsf{T}) - \mathbb{E}(\vec{X})\,\mathbb{E}(\vec{X})^\mathsf{T}$. ♧

Théorème (KE). *Pour $\boldsymbol{A} \in \mathbb{R}^{k\times d}$ une matrice et $\vec{b} \in \mathbb{R}^d$, la mesure-image de la loi normale ci-dessus par l'application $\vec{x} \mapsto \boldsymbol{A}\vec{x} + \vec{b}$ est la loi normale de dimension k suivante :*

$$\boldsymbol{A} \times \mathrm{Normale}(\vec{\mu}, \boldsymbol{C}) + \vec{b} = \mathrm{Normale}(\boldsymbol{A}\vec{\mu} + \vec{b}, \boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^\mathsf{T}). \tag{KF}$$

◇

Remarque (KG). À nouveau, le seul point vraiment important dans ce théorème est le fait que la mesure-image soit également normale : les valeurs de ses espérance et variance respectives sont en fait des conséquences de la linéarité de l'espérance et de la bilinéarité de la variance, mais écrites ici dans un cadre multidimensionnel. ♧

Théorème (KH). *Si la matrice $\boldsymbol{C}$ est* définie *positive, alors la loi normale multidimensionnelle est à densité par rapport à la mesure de Lebesgue ; et pour* $\mathrm{d}x$ *un voisinage infinitésimal de $\vec{x} \in \mathbb{R}^d$, on a*

$$\mathbb{P}\big(\mathrm{Normale}(\vec{\mu}, \boldsymbol{C}) \in \mathrm{d}x\big) = \big((2\pi)^d \det \boldsymbol{C}\big)^{-1/2} \exp\big(-\tfrac{1}{2}(\vec{x} - \vec{\mu})^\mathsf{T}\boldsymbol{C}^{-1}(\vec{x} - \vec{\mu})\big)\,\mathrm{vol}_d(\mathrm{d}x). \tag{KI}$$

[†]. Il n'est pas nécessaire de préciser la valeur de d, celle-ci correspondant nécessairement à la dimension commune de $\vec{\mu}$ et de $\boldsymbol{C}$.

En particulier, les lignes de niveau de la densité sont des ellipsoïdes, centrées sur $\vec{\mu}$, dont les axes sont orientés selon les vecteurs propres de la matrice $\boldsymbol{C}$ et ont des dimensions proportionnelles aux racines carrées des valeurs propres respectives associées. ♢

Conditionnement gaussien

Il est important de savoir que les vecteurs gaussiens se comportent très bien par conditionnement :

Théorème (KJ). *Soit $(\vec{X};\vec{Y})$ un vecteur gaussien de dimension $n+m$, où $\vec{X}$ est le sous-vecteur constitué par les n premières coordonnées et $\vec{Y}$ le sous-vecteur constitué par les m dernières coordonnées. Alors il existe un vecteur $\vec{\mu}' \in \mathbb{R}^n$ et des matrices $\boldsymbol{L} \in \mathbb{R}^{n\times m}$ et $\boldsymbol{V}' \in \mathbb{R}^{n\times n}$ tels que, pour tout $\vec{y} \in \mathbb{R}^m$:*

$$\text{Loi}(\vec{X} \mid \vec{Y} = \vec{y}) = \text{Normale}(\boldsymbol{L}\vec{y} + \vec{\mu}', \boldsymbol{V}'). \tag{KK}$$

En outre, si la loi de $(\vec{X};\vec{Y})$ s'écrit par blocs :

$$\text{Loi}\begin{pmatrix}\vec{X}\\ \vec{Y}\end{pmatrix} = \text{Normale}\left(\begin{pmatrix}\vec{\mu}\\ \vec{\nu}\end{pmatrix}, \begin{pmatrix}\boldsymbol{V} & \boldsymbol{C}\\ \boldsymbol{C}^\mathsf{T} & \boldsymbol{W}\end{pmatrix}\right) \tag{KL}$$

et que $\boldsymbol{W}$ est de rang plein[‡]*, on a $\boldsymbol{L} = \boldsymbol{C}\boldsymbol{W}^{-1}$, $\vec{\mu}' = \vec{\mu} - \boldsymbol{C}\boldsymbol{W}^{-1}\vec{\nu}$ et $\boldsymbol{V}' = \boldsymbol{V} - \boldsymbol{C}\boldsymbol{W}^{-1}\boldsymbol{C}^\mathsf{T}$.* ♢

Remarque (KM). En fait, la seconde partie du théorème donnant les formules pour $\boldsymbol{L}$, $\vec{\mu}'$ et $\boldsymbol{V}'$ peut se retrouver à partir des formules de (bi)linéarité de l'espérance et de la covariance : en effet, si la première formule du théorème est vraie, cela signifie que $(\vec{X};\vec{Y})$ a la même loi que $(\boldsymbol{L}\vec{Y} + \vec{\mu}' + \vec{\Xi};\vec{Y})$ pour $\vec{\Xi}$ un vecteur indépendant de $\vec{Y}$ de loi Normale$(\vec{0}_n, \boldsymbol{V}')$; or on calcule par (bi)linéarité que

$$\text{Loi}\begin{pmatrix}\boldsymbol{L}\vec{Y} + \vec{\mu}' + \vec{\Xi}\\ \vec{Y}\end{pmatrix} = \text{Normale}\left(\begin{pmatrix}\boldsymbol{L}\vec{\nu} + \vec{\mu}'\\ \vec{\nu}\end{pmatrix}, \begin{pmatrix}\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T} + \boldsymbol{V}' & \boldsymbol{L}\boldsymbol{C}\\ \boldsymbol{C}\boldsymbol{L}^\mathsf{T} & \boldsymbol{W}\end{pmatrix}\right); \tag{KN}$$

et puisque cela doit correspondre à la loi de $(\vec{X};\vec{Y})$, on en déduit les formules caractérisant $\boldsymbol{L}$, $\boldsymbol{V}'$ et $\vec{\mu}'$ par identification, formules qu'il n'y a ensuite plus qu'à résoudre. ♣

6.4 Autres lois continues remarquables

Lois uniformes

! **Définition (KO)** (Loi uniforme sur un intervalle)**.** Pour $a, b \in \mathbb{R}$ avec $a < b$, la *loi uniforme sur* $]a,b[$, notée $\text{Unif}^{\text{me}}(a,b)$[§], est la distribution de probabilité portée par $]a,b[$ qui attribue à tout sous-intervalle de $]a,b[$ une masse proportionnelle à sa longueur. ♡

Remarque (KP). La loi uniforme sur un segment est clairement l'homologue, dans le cas continu, de la loi uniforme sur un intervalle d'entiers dans le cas discret : on peut même démontrer que, dans un certain sens, la loi uniforme sur $[\![a,b]\!]$ tend vers la loi uniforme sur $]a,b[$ lorsque $b-a$ tend vers l'infini. ♣

! **Proposition (KQ).** *Pour $a < b$, la loi* $\mathrm{Unif}^{\mathrm{me}}(a,b)$ *est à densité sur* $\mathbb{R}$*, avec (notant* $\mathrm{d}x$ *un voisinage infinitésimal de* x*) :*

$$\mathbb{P}\big(\mathrm{Unif}^{\mathrm{me}}(a,b) \in \mathrm{d}x\big) = \begin{cases} \frac{\mathrm{vol}_1(\mathrm{d}x)}{b-a} & \text{pour } x \in]a,b[\,; \\ 0 & \text{pour } x \notin]a,b[. \end{cases} \tag{KR}$$

On calcule également facilement la fonction de répartition de la loi :

$$\mathbb{P}\big(\mathrm{Unif}^{\mathrm{me}}(a,b) \leqslant x\big) = \begin{cases} 0 & \text{pour } x \leqslant a\,; \\ \frac{x-a}{b-a} & \text{pour } x \in [a,b]\,; \\ 1 & \text{pour } x \geqslant b. \end{cases} \tag{KS}$$

◇

Remarque (KT). Dans la formule de densité, ne vous cassez surtout pas la tête à savoir quelle est la densité pour les cas limites où x vaut a ou b : en fait, toutes les conventions sont également valables, puisque ajouter ou retrancher deux masses infinitésimales en a et b n'a aucun effet macroscopique sur la mesure ! D'ailleurs, comme vous l'avez vu au semestre précédent, la définition rigoureuse du concept de densité ne détermine la fonction de densité qu'*à égalité presque-partout près*, ce qui signifie en particulier que si on change la formule pour la densité en un nombre fini (ou même dénombrable) de points, la nouvelle densité obtenue est une formule tout aussi valable que l'ancienne. (Même si ici, cela serait évidemment assez stupide de vouloir changer la valeur de la densité ailleurs qu'en a ou b). ♧

Remarque (KU). Concernant la formule pour la fonction de répartition, lorsque x vaut a ou b, on se situe dans deux cas de la formule simultanément, mais ces deux cas donnent alors le même résultat : ce qui est logique puisque nos formules doivent décrire une fonction continue (la loi étant à densité, donc diffuse). ♧

Démonstration. On sait que pour $I \subseteq]a,b[$, $\mathbb{P}\big(\mathrm{Unif}^{\mathrm{me}}(a,b) \in I\big) = \alpha\,\mathrm{vol}_1(I)$ (où $\mathrm{vol}_1(I)$ n'est autre que la longueur de I) pour une certaine constante de proportionnalité α ; en outre, puisque la distribution de probabilité est portée par $]a,b[$, on doit avoir $\mathbb{P}\big(\mathrm{Unif}^{\mathrm{me}}(a,b) \in]a,b[\big) = 1$, d'où $\alpha = (b-a)^{-1}$. Dès lors, les deux formules de l'énoncé découlent directement de la définition de la loi uniforme, en observant, dans le second cas, que la probabilité donnée à $]-\infty,x]$ n'est autre que la probabilité donnée à $]a,x]$, vu que la mesure est portée par $]a,b[$. ♤

L'espérance de la loi uniforme intervalle s'obtient immédiatement par symétrie :

Proposition (KV). *Pour $a \leqslant b$,* !

$$\mathbb{E}\big(\mathrm{Unif}^{\mathrm{me}}(a,b)\big) = \frac{a+b}{2}. \tag{KW}$$

◇

Pour la variance, c'est un peu plus compliqué :

[‡]. Ce à quoi on peut toujours se ramener, en considérant si nécessaire un sous-vecteur $\vec{Y}'$ de $\vec{Y}$ de taille $\mathrm{rg}\,\boldsymbol{W}$ tel que $\vec{Y}$ soit presque-surement une fonction affine de $\vec{Y}'$.

[§]. À ne pas confondre avec la loi $\mathrm{Unif}^{\mathrm{me}}(\{a,b\})$, qui est la loi uniforme sur la *paire* $\{a,b\}$!

Théorème (KX). *Pour $a \leqslant b$,*

$$\mathrm{Var}\big(\mathrm{Unif}^{\mathrm{me}}(a,b)\big) = \tfrac{1}{12}(b-a)^2. \tag{KY}$$

♢

Démonstration. La façon de faire la plus directe (mais pas la plus élégante [¶]) est de calculer en utilisant la formule de la densité de la loi et la formule de transfert que

$$\mathbb{E}\big(\mathrm{Unif}^{\mathrm{me}}(a,b)^2\big) = \int_{x=a}^{b} x^2 \frac{\mathrm{d}x}{b-a} = (b-a)^{-1}\Big[\tfrac{1}{3}x^3\Big]_a^b = \tfrac{1}{3}\frac{b^3-a^3}{(b-a)} = \tfrac{1}{3}(a^2+ab+b^2), \tag{KZ}$$

et d'en déduire au vu de la formule pour l'espérance que

$$\begin{aligned}\mathrm{Var}\big(\mathrm{Unif}^{\mathrm{me}}(a,b)^2\big) &= \tfrac{1}{3}(a^2+ab+b^2) - \left(\frac{b+a}{2}\right)^2 \\ &= \big(\tfrac{1}{3}a^2+\tfrac{1}{3}ab+\tfrac{1}{3}b^2\big) - \big(\tfrac{1}{4}a^2+\tfrac{1}{2}ab+\tfrac{1}{4}b^2\big) = \tfrac{1}{12}a^2 - \tfrac{1}{6}ab + \tfrac{1}{12}a^2 = \tfrac{1}{12}(b-a)^2.\end{aligned}$$

♤

! **Proposition (LA).** *Pour X une variable suivant la loi* $\mathrm{Unif}^{\mathrm{me}}(\alpha,\beta)$*, pour $a, b \in \mathbb{R}$, la variable aléatoire $aX+b$ suit la loi :*

$$\begin{cases} \mathrm{Unif}^{\mathrm{me}}(a\alpha+b, a\beta+b) & \text{si } a>0\,; \\ \mathrm{Unif}^{\mathrm{me}}(a\beta+b, a\alpha+b) & \text{si } a<0\,; \\ \delta_b & \text{si } a=0. \end{cases} \tag{LB}$$

♢

Démonstration. Prouver rigoureusement cette proposition demande quelques calculs à l'aide de la formule de changement de variables géométrique, mais le résultat est tellement évident intuitivement que je pense qu'on peut vous demander de le retenir par cœur ☺ ♤

Lois exponentielles

! **Définition (LC)** (Loi exponentielle)**.** Pour $\lambda > 0$, la *loi exponentielle de taux λ*, notée $\mathrm{Expon}^{\mathrm{le}}(\lambda)$, est la loi de la durée de vie d'un individu qui, à tout instant, indépendamment du passé, a une probabilité λ par unité de temps de mourir ; autrement dit, sachant que l'individu est encore en vie au temps t, pour $\mathrm{d}t$ un accroissement de temps infinitésimal, l'individu a une probabilité $\lambda \mathrm{d}t$ de mourir avant le temps $(t+\mathrm{d}t)$. ♡

Remarque (LD). Attention, sauf erreur de ma part, ce n'est pas le même paramétrage que celui que vous avez vu au semestre 5 ! Une convention alternative fréquemment répandue étant en effet de paramétrer les lois exponentielles par leur *espérance de vie*, qui, avec les notations ci-dessus, correspond à $1/\lambda$. ♧

Remarque (LE). La loi exponentielle est l'homologue continu de la loi géométrique dans le cas discret, le rôle de la mort de l'individu remplaçant celui de la réussite de l'expérience. On peut même démontrer que, dans un certain sens, lorsque p tend vers 0, la loi $\mathrm{Géom}^{\mathrm{que}}(p)$ devient asymptotiquement égale à la loi $\mathrm{Expon}^{\mathrm{le}}(p)$. ♧

[¶]. Il est plus habile de commencer par observer que le résultat est nécessairement de la forme $\alpha(b-a)^2$, puis de calculer α sur un cas particulièrement simple, p. ex. pour $]a,b[=]\pm 1[$.

Les formules concernant la fonction de répartition et la densité de la loi exponentielle ne sont pas à connaitre par cœur, mais doivent pouvoir être redémontrées au besoin :

Proposition (LF). *La loi* $\mathrm{Expon}^{\mathrm{le}}(\lambda)$ *est portée par* $\mathbb{R}_+$*, avec, pour tout* $x \in \mathbb{R}_+$*,*

$$\mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \geqslant x) = \mathrm{e}^{-\lambda x}. \tag{LG}$$

◇

Démonstration. La première chose qu'il faut comprendre, c'est que, puisque l'individu a en permanence la même probabilité de mourir par unité de temps, sa probabilité d'être encore en vie au temps x va décroitre exponentiellement. Même si cela se saisit intuitivement, formalisons le raisonnement. Au vu de la définition de la façon dont notre individu meurt, il est clair que, sachant qu'il est encore en vie au temps t, son futur sera exactement le même que s'il venait de naitre ! En particulier, sachant qu'il est en vie au temps t, la probabilité qu'il vive encore pendant une durée s (et donc que sa durée de vie totale atteigne $t+s$) est égale à la probabilité, si on le considère à la naissance, qu'il vive pendant une durée s ; ce qu'on peut écrire, en termes mathématiques : pour $X \sim \mathrm{Expon}^{\mathrm{le}}(\lambda)$,

$$\mathbb{P}(X \geqslant t+s \mid X \geqslant t) = \mathbb{P}(X \geqslant s), \tag{LH}$$

ce qui donne « $\mathbb{P}(A_{t+s} \mid A_t) = \mathbb{P}(A_s)$ » en posant $\{X \geqslant u\} =: A_u$ pour alléger les notations. Utilisant la définition de l'espérance conditionnelle, on a donc

$$\frac{\mathbb{P}(A_t \text{ et } A_{t+s})}{\mathbb{P}(A_t)} = \mathbb{P}(A_s)\,; \tag{LI}$$

où nous notons que la définition des A_u fait que $A_{t+s} \implies A_t$, de sorte qu'on a en fait $\{A_t \text{ et } A_{t+s}\} = A_{t+s}$, et donc finalement

$$\mathbb{P}(A_{t+s}) = \mathbb{P}(A_t)\,\mathbb{P}(A_s). \tag{LJ}$$

Ce raisonnement peut se comprendre très simplement en disant que pour vivre pendant un temps $t+s$, il faut d'abord vivre pendant un temps t, puis à nouveau ne pas mourir pendant un temps s entre t et $t+s$, et que ces deux évènements sont indépendants [‖].

Dès lors, l'application $u \mapsto \mathbb{P}(A_u)$ est un morphisme de semigroupes de $(\mathbb{R}_+, +)$ dans $([0,1], \times)$, et comme ce morphisme est continu [**] puisqu'il n'y a aucun moment où la particule a une probabilité non nulle de mourir, c'est une application de la forme $u \mapsto \mathrm{e}^{\tau u}$: c'est pour ainsi dire la définition même de l'exponentielle ! Pour finir, il reste à identifier τ : mais, puisque la probabilité (non conditionnée) de mourir par unité de temps est la dérivée de la probabilité d'être encore en vie, on a que la probabilité par unité de temps de mourir au temps 0 vaut $-\big(u \mapsto \mathrm{e}^{\tau u}\big)'(0) = \tau$ (là encore, c'est essentiellement la définition même de l'exponentielle), et puisque par ailleurs cette probabilité vaut λ, on arrive au résultat annoncé. ♤

[‖]. À ceci près que « ne pas mourir entre t et $t+s$ » n'a pas vraiment de sens si on est déjà mort au temps t ; mais on comprend l'idée...

[**]. En fait, on n'a même pas besoin de la continuité : le fait que l'application soit décroissante — ce qui est trivial — suffit.

Corolaire (LK). *La loi* $\mathrm{Expon}^{\mathrm{le}}(\lambda)$ *prend presque-surement des valeurs strictement positives, et est à densité sur* $\mathbb{R}_+^*$*, avec, notant* $\mathrm{d}x$ *un voisinage infinitésimal de* x *:*

$$\mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \in \mathrm{d}x) = \lambda \mathrm{e}^{-\lambda x} \,\mathrm{vol}_1(\mathrm{d}x) \tag{LL}$$

♢

Démonstration. Le fait que la loi exponentielle ne prenne presque-jamais la valeur nulle vient de ce que la probabilité de mourir possède par définition une certaine *densité* par rapport au temps : dès lors, la probabilité de mourir à un instant bien précis est nulle.

Concernant la formule pour la densité, il y a deux façons de la prouver. La première est de prendre, pour se fixer les idées, $\mathrm{d}x = [x, x+h[$ (avec h infinitésimalement petit) et de dire que, pour mourir entre les temps x et $x+h$, il faut commencer par survivre jusqu'au temps x (ce qui, en vertu de la proposition (LF), se produit avec une probabilité $\mathrm{e}^{-\lambda x}$), puis, sachant cela, mourir pendant l'intervalle de temps infinitésimal de durée h suivant, ce qui, par définition de la loi exponentielle, se produit avec probabilité λh (on s'est placé ici dans le cas $x \geqslant 0$, le cas $x < 0$ étant trivial). En multipliant ces deux expressions, on obtient bien le résultat annoncé.

Une autre façon de prouver ce résultat est un raisonnement de probabilité très classique que vous devez absolument savoir maitriser. On fixe à nouveau $\mathrm{d}x$ de la forme $[x, x+h[$, et on observe que, pour X une variable aléatoire réelle,

$$\{X \in [x, x+h[\} = \{X \geqslant x\} \text{ et } \neg\{X \geqslant x+h\}. \tag{LM}$$

Or on a $\{X \geqslant x+h\} \implies \{X \geqslant x\}$, donc la probabilité de l'évènement de gauche s'obtient par différence :

$$\begin{aligned}\mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \in [x, x+h[) \\ = \mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \geqslant x) - \mathbb{P}(\mathrm{Expon}^{\mathrm{le}}(\lambda) \geqslant x+h) = \mathrm{e}^{-\lambda x} - \mathrm{e}^{-\lambda(x+h)}\end{aligned}$$

en vertu de la proposition (LF). On voit que cette expression est de la forme $f(x+h) - f(x)$ avec $f(u) = -\mathrm{e}^{-\lambda u}$; par développement au premier ordre, elle vaut donc $f'(x)h$, ce qui est bien, à nouveau, le résultat annoncé. ♤

Théorème (LN). *Pour* $\lambda > 0$*,*

$$\mathbb{E}(\mathrm{Expon}^{\mathrm{le}}(\lambda)) = 1/\lambda \,; \tag{LO}$$

$$\mathrm{Var}(\mathrm{Expon}^{\mathrm{le}}(\lambda)) = 1/\lambda^2. \tag{LP}$$

♢

Remarque (LQ). Ces résultats sont analogues des formules pour la loi géométrique ; et on aurait ainsi pu retrouver la formule pour l'espérance par le même genre de démonstration que pour la loi géométrique ; cependant il m'a semblé que c'était nettement plus difficile à visualiser dans le cas continu, et j'ai donc choisi de ne pas vous imposer de savoir retrouver cette formule. ♧

Remarque (LR). En termes de typologie des paramètres, il n'est pas difficile de vérifier que le paramètre de taux d'une loi exponentielle correspond à un paramètre d'échelle inverse. (Les transformations affines à considérer pour cette famille de lois étant les transformations linéaires strictement croissantes). ♧

Deuxième partie

Statistique bayésienne

Chapitre 7

Concept d'inférence statistique

7.1 Motivation : Deux exemples

Afin de motiver la notion d'inférence statistique, nous allons ci-dessous présenter deux exemples. Ces deux exemples, après formalisation, nous accompagneront tout au long du polycopié.

Exemple (LS) (Le (bon ?) chasseur). Un candidat souhaite rejoindre le club des chasseurs du Bouchonnois. Le club ayant une politique de haute exigence, il souhaite savoir avec quelle précision le candidat sait se servir de son arme en vue d'un éventuel recrutement.

La « précision » d'un tireur est évaluée via un indicateur que nous appellerons *fiabilité*, qui est défini comme la proportion de plateaux [*] que ce tireur est capable de toucher au balltrap dans certaines conditions de référence. Tout tireur possède une certaine fiabilité, qui est par construction un réel de $[0,1]$; et en l'occurrence, nous appelerons $\theta_{\checkmark}$ la fiabilité de notre candidat. Ici il est essentiel de comprendre que, bien que le nombre $\theta_{\checkmark}$ soit défini sans ambigüité, on ne peut pas déterminer sa valeur directement : en effet, quand on parle de la « proportion de plateaux que le tireur est capable de toucher », cela est à entendre dans le sens d'un nombre de plateaux *tendant vers l'infini* ; mais dans les situations pratiques, il faudrait plutôt dire que le tireur a une *probabilité* $\theta_{\checkmark}$ de toucher chaque plateau !

Au niveau de la procédure de recrutement, pour se faire une idée de la fiabilité du candidat, le club va lui demandera de tirer sur une série de $25 =: n$ plateaux. Le club regardera le nombre de plateaux qui ont été touchées au cours de ce test, nombre que nous noterons X ; avec l'espoir qu'on pourra, graĉe à la connaissance de X, se faire une "idée" assez fiable de ce que vaut $\theta_{\checkmark}$.

Par ailleurs, le club envisage, s'il recrute le candidat, de le faire participer à une compétition dans laquelle il faudra tirer sur $75 =: m$ plateaux (dans les mêmes conditions que pour le test) : ainsi, mieux cerner la fiabilité du candidat permettra également au club d'avoir une idée de score que le candidat réalisera à la compétition si on l'y inscrit ! ♧

Remarque (LT). Il importe de bien insister sur le fait que X ne permet pas de reconstituer $\theta_{\checkmark}$ à coup sûr, du fait qu'il s'agit d'une *variable aléatoire* : plus précisément, en supposant que le fait que chaque plateau de l'épreuve soit ou non touché est indépendant, X suit la loi $\mathrm{Binom}^{\mathrm{le}}(n, \theta_{\checkmark})$; or tout ce que le club voit à

[*]. Au balltrap, un *plateau* est une cible volante en argile, qui explose lorsqu'elle est touchée par une balle.

l'issue du test, c'est *une* réalisation de X ! Par exemple, mettons que le club voie que le candidat a touché $8 =: x_{\checkmark}$ plateaux. Vu que, pour toute valeur de θ entre 0 et 1 (exclus), il existe une probabilité non nulle que la loi $\mathrm{Binom}^{\mathrm{le}}(n, \theta)$ prenne la valeur 8, cette observation ne nous permet d'exclure *formellement* aucune valeur pour $\theta_{\checkmark}$: peut-être que $\theta_{\checkmark}$ vaut 20 % et que le candidat a été chanceux lors du test ; peut-être que $\theta_{\checkmark}$ vaut 30 % et que le candidat a fait un score tout à fait conforme à ce à quoi il pouvait s'attendre ; peut-être que $\theta_{\checkmark}$ vaut 75 % et que le candidat a été très malchanceux... Mais bien sûr, certaines de ces possibilités semblent plus *raisonnables* que d'autres ; et c'est bien en ce sens qu'on pourra se faire une idée assez fiable de $\theta_{\checkmark}$ malgré tout ! ♧

Exemple (LU) (Le pédagogue). Un enseignant est chargé de donner un cours de thermodynamique avancée à un groupe d'élèves ingénieurs. La méthode d'évaluation des élèves en fin de cours est standardisée selon un protocole précis, et donne pour chaque élève un « score » figuré par un nombre réel. Traditionnellement, ce cours était enseigné selon une certaine méthode pédagogique ; et la très longue expérience sur cette méthode avait montré que le score d'un élève pris au hasard formé selon la méthode traditionnelle suit une loi $\mathrm{Normale}(\mu_{\text{réf}}, \sigma^2_{\text{réf}})$ avec $\mu_{\text{réf}} = 75$ et $\sigma_{\text{réf}} = 11$, les scores des différents élèves étant indépendants [†].

Notre enseignant souhaite expérimenter une nouvelle méthode pédagogique, dont il espère évidemment qu'elle donnera de meilleurs résultats que l'ancienne. Diverses raisons l'amènent à penser que, avec la nouvelle méthode, les scores des élèves seront toujours indépendants et identiquement distribués selon une loi de la forme $\mathrm{Normale}(\mu_{\checkmark}, \sigma^2_{\checkmark})$, mais que $\mu_{\checkmark}$ et $\sigma_{\checkmark}$ seront (potentiellement) *différentes* de $\mu_{\text{réf}}$ et $\sigma_{\text{réf}}$!

De même que dans le cas du chasseur, l'enseignant ne peut pas observer directement $\mu_{\checkmark}$ et $\sigma_{\checkmark}$. Tout ce qu'il peut faire, c'est tester sa méthode sur la promotion d'élèves qu'il a cette année, que nous supposerons ici composée de $22 =: n$ élèves, et regarder les résultats : les scores obtenus seront des v.a. $X_0, \dots, X_{n-1}$ i.i.d. $\mathrm{Normale}(\mu_{\checkmark}, \sigma^2_{\checkmark})$ — ou encore, de manière équivalente, le n-uplet $(X_0, \dots, X_{n-1})$ suivra la loi $\mathrm{Normale}(\mu_{\checkmark}, \sigma^2_{\checkmark})^{\otimes n}$ sur $\mathbb{R}^n$. Et à partir de ces X_i, l'enseignant tentera de "reconstituer" les valeurs de $\mu_{\checkmark}$ et $\sigma_{\checkmark}$: là encore, il ne pourra pas le faire avec une certitude absolue, mais certaines possibilités seront plus plausibles que d'autres. ♧

7.2 Modèle d'inférence statistique

Cas d'une inférence explicative

On donne ici les définitions mathématiques de la notion de modèle d'inférence statistique. On distinguera deux situations : celle de l'inférence *explicative*, et celle de l'inférence *prédictive*.

Définition générale

!! **Définition** (LV) (Modèle d'inférence statistique explicative)**.** Un *modèle d'inférence statistique explicative* est la donnée d'un espace Θ, appelé *espace du paramètre*

[†]. Attention : Lorsque nous disons que le score d'un élève est « aléatoire », en l'occurrence cela ne veut pas dire que l'examen donne un résultat au hasard, mais que c'est l'*identité* de l'élève (est-on tombé sur une élève douée ou un élève plus en difficulté) qui est la (principale) source d'aléa.

caché, d'un espace $\mathscr{X}$, appelé *espace de l'observation*, et d'une application de $\boldsymbol{\Theta}$ dans $\mathscr{M}_1(\mathscr{X})$ [‡] qui, à chaque $\theta \in \boldsymbol{\Theta}$, associe une loi de probabilité $\boldsymbol{P}_\theta$ sur $\mathscr{X}$, appelée *loi de l'observation sachant que le paramètre caché vaut* θ. ♡

Point (LW) (Variables aléatoires associées à un modèle statistique). Chaque fois que nous introduirons un modèle statistique, nous considèrerons automatiquement un espace probabilisé sur lequel nous définirons une variable aléatoire $\boldsymbol{\theta}$ à valeurs dans $\boldsymbol{\Theta}$, appelée *paramètre caché* (ou « paramètre caché en tant que variable aléatoire » quand il y aura lieu de préciser), et une variable aléatoire $\boldsymbol{X}$ à valeurs dans $\mathscr{X}$, appelée *observation* (ou « observation en tant que variable aléatoire »), de façon que pour tout $\theta \in \boldsymbol{\Theta}$, on ait

$$\mathrm{Loi}(\boldsymbol{X} \mid \boldsymbol{\theta} = \theta) = \boldsymbol{P}_\theta. \tag{LX}$$

♧

Principe (LY) (Interprétation d'un modèle d'inférence explicative)**.**

- *Dans un modèle d'inférence statistique (explicative), l'observation représente ce qu'on est (ou sera) en mesure d'observer au moment de réaliser notre analyse. On disposera donc, au moment de l'analyse, de la* réalisation *de* $\boldsymbol{X}$*, que nous noterons* $x_{\checkmark}$*, et que nous qualifierons de « valeur effective de l'observation ».*
- *La réalisation du paramètre caché* $\boldsymbol{\theta}$*, que nous noterons* $\theta_{\checkmark}$ *et que nous appellerons « vraie valeur du paramètre caché », représente les informations qu'il pourrait être utile de connaitre (dans le cadre du modèle) sur le comportement de l'observation,* mais qu'on ne peut pas observer directement*. Dans la mesure où* $\theta_{\checkmark}$ *ne peut pas être observé directement, nous sommes obligés d'*envisager *qu'il puisse prendre toutes sortes de valeurs, d'où l'intérêt d'avoir une vision « variable aléatoire » du paramètre caché pour formaliser cette multitude de possibilités. Dans ce contexte, l'espace du paramètre caché* $\boldsymbol{\Theta}$ *est l'ensemble des valeurs qu'il est envisageable (dans le cadre de notre modèle) que* $\boldsymbol{\theta}$ *puisse valoir.*
- *La loi de l'observation sachant le paramètre caché représente l'aléa "intrinsèque" sur l'observation, celui qui demeurerait fondamentalement même si on connaissait la vraie valeur du paramètre caché.*

♢

Remarque (LZ). Bien comprendre que d'un point de vue "expérimental", quand on considère une situation donnée modélisée par un modèle d'inférence, la vraie valeur $\theta_{\checkmark}$ du paramètre caché existe "quelque part", avant même que nous observions la valeur de $\boldsymbol{X}$, et qu'il n'existe aucun aléa ni aucun contrôle sur ce qu'elle vaut ! (en tout cas, pas au moment de commencer l'expérience). Pour autant, nous n'avons aucun moyen d'y accéder directement : tout ce que nous pourrons observer, c'est la valeur effective $x_{\checkmark}$ de l'observation. ♧

Exemple (MA) (Formalisation du modèle du chasseur). Le *modèle du chasseur*, dans sa version explicative, a pour paramètre caché la variable aléatoire $\boldsymbol{\theta}$ représentant la *fiabilité* du chasseur ; l'espace du paramètre caché étant $\boldsymbol{\Theta} =:]0,1[$ [§].

[‡]. Rappelons que "$\mathscr{M}_1(\mathscr{X})$" désigne l'ensemble des lois de probabilité sur $\mathscr{X}$.

[§]. Remarque : J'ai choisi d'exclure les cas-limites $\theta = 0$ et $\theta = 1$, afin d'éviter certains soucis techniques.

L'observation, notée X, représente le nombre de plateaux touchées lors du test. L'espace de l'observation est l'intervalle d'entiers $[\![0, n]\!] =: \mathcal{X}$, où n est une valeur connue représentant le nombre d'essais lors du test, que nous prendrons égale à 25 à moins qu'il ne soit explicitement précisé autrement. La loi de l'observation sachant le paramètre caché est

$$\text{Loi}(X \mid \boldsymbol{\theta} = \theta) \coloneqq \text{Binom}^{\text{le}}(n, \theta). \tag{MB}$$

En ce qui considère le nombre $x_{\checkmark}$ de plateaux touchés par notre chasseur, nous le prendrons égal à 8 pour les applications numériques, à moins qu'il ne soit explicitement précisé autrement. ♧

Exemple (MC) (Formalisation du modèle du pédagogue). Le *modèle du pédagogue*, dans sa version explicative, a pour paramètre caché le couple de variables aléatoires $(\boldsymbol{\mu}, \boldsymbol{\sigma})$, où $\boldsymbol{\mu}$ et $\boldsymbol{\sigma}$ représentent respectivement l'*espérance* et l'*écart-type* de la distribution (sur le long terme) des scores des élèves ; l'espace du paramètre caché étant $\boldsymbol{\Theta} =: \mathbb{R} \times \mathbb{R}_+^*$ (autrement dit, $\boldsymbol{\mu}$ est à valeurs dans $\mathbb{R}$ et $\boldsymbol{\sigma}$ à valeurs dans $\mathbb{R}_+^*$). L'observation représente la liste des scores de l'ensemble des élèves (rangés, disons, dans l'ordre alphabétique) : c'est donc un n-uplet de valeurs, de sorte que nous la noterons $\vec{X}_{[\![0,n]\!]}$ (ou parfois simplement $\vec{X}$ pour alléger), ou, de façon équivalente, $(X_0, \dots, X_{n-1})$ (X_0 étant alors le score de la première élève de la liste, X_1 celui du second de la liste, etc.). La valeur n, représentant le nombre d'élèves dans la promotion expérimentant la nouvelle méthode, est connue : on la prendra égale à 22, sauf mention explicite du contraire. Les scores sont supposés être des nombres réels, de sorte que l'espace de l'observation, noté $\mathcal{X}$, correspond à $\mathbb{R}^n$. La loi de l'observation sachant le paramètre caché est

$$\text{Loi}(\vec{X} \mid \boldsymbol{\mu} = \mu \text{ et } \boldsymbol{\sigma} = \sigma) \coloneqq \text{Normale}(\mu, \sigma^2)^{\otimes n} : \tag{MD}$$

autrement dit, si le paramètre caché vaut (μ, σ), les scores des élèves sont i.i.d. de loi $\text{Normale}(\mu, \sigma^2)$. Pour les applications numériques de l'analyse, la réalisation effective de $\vec{X}_{[\![0,n]\!]}$, notée $(x_{0\checkmark}, \dots, x_{(n-1)\checkmark})$ ou $\vec{x}_{[\![0,n]\!]\checkmark}$, sera prise égale à

$$(91{,}7, 69{,}6, 81{,}5, 45{,}0, 69{,}7, 77{,}3, 42{,}1, 70{,}7, 66{,}9, 93{,}6, 75{,}7,\\ 93{,}7, 59{,}8, 75{,}2, 49{,}8, 88{,}2, 63{,}7, 90{,}0, 54{,}2, 57{,}2, 72{,}8, 58{,}3). \tag{ME}$$

♧

Cas d'une inférence prédictive

!! **Définition** (MF) (Modèle d'inférence statistique prédictive)**.** Un modèle d'inférence statistique prédictive est la donnée d'un espace du paramètre caché $\boldsymbol{\Theta}$, de *deux* espaces $\mathcal{X}$ et $\mathcal{Y}$, appelés resp. *espace de l'observation* (ou « de l'observation *passée* ») et *espace de l'observation future*, et d'une application de $\boldsymbol{\Theta}$ dans $\mathscr{M}_1(\mathcal{X} \times \mathcal{Y})$ qui, à chaque $\theta \in \boldsymbol{\Theta}$, associe une loi de probabilité $\boldsymbol{P}_\theta$, appelée *loi de l'observation complétée* sachant que le paramètre caché vaut θ. ♡

! **Définition** (MG) (Variables aléatoires associées à un modèle prédictif)**.** De même que dans le paradigme explicatif, chaque fois que nous introduirons un modèle statistique, nous considèrerons automatiquement un espace probabilisé sur lequel nous définirons une variable aléatoire « paramètre caché » $\boldsymbol{\theta}$ à valeurs dans $\boldsymbol{\Theta}$,

et *deux* variables aléatoires X et Y à valeurs dans resp. $\mathscr{X}$ et $\mathscr{Y}$, appelées resp. *observation (passée)* et *observation future*, de façon que pour tout $\theta \in \Theta$, la loi *jointe* de (X, Y) conditionnellement à $\{\theta = \theta\}$ soit P_θ :

$$\mathrm{Loi}(X, Y \mid \theta = \theta) \coloneqq P_\theta. \tag{MH}$$

♡

Principe (MI) (Interprétation d'un modèle prédictif)**.** *En termes d'interprétation, X représente l'observation à laquelle on pourra avoir accès* au moment de faire l'analyse *(et dont on connaitra donc la réalisation $x_\checkmark$), tandis que Y représente une observation à laquelle on pourra éventuellement avoir accès dans un certain futur* après *l'analyse.* ◇

!!

Exemple (MJ)*.* Dans la version prédictive du modèle du chasseur, outre l'observation (passée) mentionnée en exemple , on introduit aussi une observation future, notée Y, qui représente le nombre de plateaux touchés par le candidat lors de la compétition à venir (s'il y participe). L'espace de l'observation future est l'intervalle d'entiers $[\![0, m]\!] =: \mathscr{Y}$, où m représente le nombre de plateaux pour la compétition, qui est connu : pour les applications numériques, nous prendrons $m \coloneqq 75$, sauf mention explicite du contraire. En version prédictive, la loi de l'observation complétée sachant le paramètre caché est :

$$\mathrm{Loi}(X, Y \mid \theta = \theta) \coloneqq \mathrm{Binom}^{\mathrm{le}}(n, \theta) \otimes \mathrm{Binom}^{\mathrm{le}}(m, \theta) : \tag{MK}$$

cela signifie concrètement que, si le paramètre caché vaut θ, X et Y sont (conditionnellement) indépendantes et suivent resp. les lois $\mathrm{Binom}^{\mathrm{le}}(n, \theta)$ et $\mathrm{Binom}^{\mathrm{le}}(m, \theta)$; en particulier, pour la vraie valeur du paramètre caché, X et Y sont indépendantes et de lois respectives $\mathrm{Binom}^{\mathrm{le}}(n, \theta_\checkmark)$ et $\mathrm{Binom}^{\mathrm{le}}(m, \theta_\checkmark)$. ♧

Exemple (ML) (Modèle du pédagogue, version prédictive)*.* Dans la version prédictive du modèle du pédagogue, outre l'observation (passée) mentionnée en exemple , on introduit aussi une observation future, notée $\vec{Y}_{[\![0,m]\!]}$, qui représente les scores obtenus par la seconde promotion à tester la méthode, à valeurs dans $\mathbb{R}^m \coloneqq \mathscr{Y}$, où m représente le nombre d'élèves de la seconde promotion, qui est connu : pour les applications numériques, nous prendrons $m \coloneqq 16$, sauf mention explicite du contraire. Dans ce cas, la loi de l'observation complétée sachant le paramètre caché est :

$$\mathrm{Loi}(\vec{X}, \vec{Y} \mid \mu = \mu \text{ et } \sigma = \sigma) \coloneqq \mathrm{Normale}(\mu, \sigma^2)^{\otimes n} \otimes \mathrm{Normale}(\mu, \sigma^2)^{\otimes m}, \tag{MM}$$

ce qui signifie concrètement que, sachant la vraie valeur du paramètre caché, tous les scores des élèves (qu'ils appartiennent à la première ou à la seconde promotion) sont (conditionnellement) i.i.d. $\mathrm{Normale}(\mu_\checkmark, \sigma^2_\checkmark)^{\otimes n}$. ♧

Remarque (MN)*.* Attention ! Dans nos deux exemples favoris, dans le cadre prédictif, l'observation (passée) et l'observation future sont, sachant le paramètre caché, indépendantes : mais cela n'a aucune raison d'être le cas en général ! ♧

Remarque (MO)*.* Attention ! Nonobstant la nomenclature utilisée, l'« observation future » n'est pas forcément quelque chose qui aura lieu plus tard dans le temps, ni même forcément quelque chose qu'on peut effectivement observer ! Ce qui compte, c'est la *structure mathématique* du modèle : l'« observation future », c'est le nom qu'on donne à une quantité sur laquelle subsiste un aléa, même sachant le paramètre caché (et l'observation passée) ! ♧

7.3 Notion d'inférence

Maintenant que nous avons formalisé la notion de modèle statistique, nous sommes prêts à expliquer plus précisément ce qu'on entend par « inférence statistique ».

Notion de quantité d'intérêt

Au préalable, il nous faut introduire le concept de *quantité d'intérêt* :

! **Définition** (MP) (Quantité d'intérêt)**.** Une quantité d'intérêt correspondra à une *fonction* des quantités introduites dans le modèle, impossible à observer directement au moment de l'analyse. Plus précisément, on introduira deux types de quantités d'intérêt :

— Une quantité d'intérêt *explicative* est une fonction du paramètre caché, autrement dit, une v.a. de la forme $\gamma(\boldsymbol{\theta})$, pour $\gamma(\bullet)$ une fonction de $\boldsymbol{\Theta}$ dans un certain espace $\mathscr{G}$.

— Un quantité d'intérêt *prédictive* est une fonction de l'observation future, autrement dit, une v.a. de la forme $g(\boldsymbol{Y})$, pour $g(\bullet)$ une fonction de $\mathscr{Y}$ dans un certain espace $\mathscr{G}$.

L'espace dans lequel une certaine quantité d'intérêt est à valeurs sera appelé « espace de la quantité d'intérêt (en question) ». ♡

Remarque (MQ)*.* L'appelation « quantité d'intérêt » provient de ce que, en pratique, le statisticien s'intéresse surtout au paramètre caché ou l'observation future au travers d'une certaine quantité d'intérêt. Cependant nous parlerons plus généralement de « quantités d'intérêt » dès lors que nous voulons parler d'une quantité de la forme $\gamma(\boldsymbol{\theta})$ ou $g(\boldsymbol{Y})$, quand bien même cette quantité ne constituerait qu'une étape intermédiaire de notre raisonnement et que nous ne sommes pas intrinsèquement « intéressés » par elle. ♧

Exemple (MR)*.* Dans le modèle du chasseur, la quantité qui intéresse le plus le club est manifestement la fiabilité du candidat : autrement dit, $\boldsymbol{\theta}$ elle-même. Il s'agit bien d'une quantité d'intérêt explicative, puisque c'est l'image du paramètre caché par l'application « identité ». ♧

Exemple (MS)*.* Toujours dans le modèle du chasseur, une autre quantité d'intérêt explicative pourrait être $2\boldsymbol{\theta} - \boldsymbol{\theta}^2$, qui représente la probabilité le toucher un plateau lorsqu'on a droit à deux essais par plateau. ♧

Exemple (MT)*.* Dans le modèle du pédagogue, une quantité particulièrement intéressante est $\boldsymbol{\mu}$, qui décrit l'efficacité globale de la nouvelle méthode pédagogique : c'est bien une quantité d'intérêt, puisque c'est l'image du paramètre caché par l'application « première composante ». ♧

Exemple (MU)*.* Dans le modèle du pédagogue, une quantité d'intérêt *prédictive* intéressante pourrait être la pire score parmi les élèves de la seconde promotion (si on continue de leur appliquer la nouvelle méthode), à savoir, $\min\{\boldsymbol{Y}_j \mid j \in [\![0, m[\![\}$. ♧

Inférence statistique

!! **Définition** (MV) (Inférence statistique)**.** Une inférence statistique est un protocole qui, étant donné un modèle d'inférence statistique pour lequel on a pris connaissance

de la valeur effective de l'observation, vise à dire des choses aussi intéressantes que possibles sur la quantité d'intérêt ; et ce, de façon *mathématiquement rationnelle* et *numériquement quantifiable.* ♡

Définition (MW). On parle d'inférence *explicative* lorsque notre modèle statistique est explicatif et que le but est de "reconstituer" une certaine quantité d'intérêt explicative, resp. d'inférence *prédictive* lorsque notre modèle statistique est prédictif et que le but est de "prédire" une certaine quantité d'intérêt prédictive. ♡ !

Dans la définition (MV), nous avons parlé de « protocole ». Ceci est étroitement lié à la notion de « statistique » :

Point (MX). Dans la définition (MV), parler de « protocole » signifie qu'on a une procédure d'analyse qui est en capacité de fournir une réponse pour *toute* valeur possible de l'observation ; même si, *en pratique*, on ne donnera cette réponse que pour sa valeur *effective* $x_{✓}$... Une inférence statistique peut donc être vue comme consistant à calculer une certaine *fonction* de l'observation $\boldsymbol{X}$. ♧

Cette idée de « fonction de l'observation » correspond, dans le jargon de la statistique, au concept de « statistique » :

Définition (MY) (Statistique)**.** On appelle *statistique* une variable aléatoire qui dépend uniquement de l'observation. ♡ !!

Principe (MZ). *Le résultat final d'une analyse statistique doit nécessairement être (la réalisation de) une (ou des) statistique(s).* ♢ !!

Remarque (NA). On dit parfois que l'inférence statistique explicative est la "démarche inverse" du calcul des probabilités : dans les deux cas en effet, un paramètre $\boldsymbol{\theta}$ influe sur la loi d'une observation aléatoire $\boldsymbol{X}$; mais, alors qu'en probabilités, on connait la vraie valeur $\theta_{✓}$ de $\boldsymbol{\theta}$ et qu'on cherche à en déduire des choses sur $\boldsymbol{X}$, en inférence statistique, on connait la réalisation $x_{✓}$ de $\boldsymbol{X}$ et on cherche à en déduire des choses sur $\boldsymbol{\theta}$!

Dans le cas de l'inférence prédictive, on mélange *deux* niveaux d'incertitude :

— D'une part, l'incertitude qui provient du fait que $\theta_{✓}$ est inconnu, qui est de nature "statistique" ;

— Et d'autre part le fait que, même si on connaissait la vraie valeur $\theta_{✓}$ du paramètre caché, il resterait un aléa sur $\boldsymbol{Y}$: cette incertitude-là étant de nature "probabiliste".

♧

7.4 Paramètres des modèles

Notion de paramètre du modèle

Lorsque j'ai formalisé le modèle du chasseur, j'ai introduit des notations littérales 'n' et 'm' pour désigner les nombres de plateaux à tirer resp. lors du test et lors de la compétition. J'ai expliqué que ces nombres étaient connus, et qu'on les prendrait égaux à resp. 25 et 75 en l'absence de mention explicite du contraire. (Et de même, pour le modèle du pédagogue, on a introduit des notations 'n' et 'm' similaires). Cependant, on pourrait choisir n'importe quelles valeurs entières naturelles pour n et m sans que la définition ne cesse d'être une description correcte de modèle

statistique... Du coup, si nous n'avions pas spécifié les valeurs de n et m dans notre définition, nous aurions en fait défini une *famille* de modèles statistiques, indexée par $\mathbb{N} \times \mathbb{N}$.

Du point de vue de la nomenclature, on a envie de considérer qu'une telle famille de modèles constitue en fait « un » modèle "flou" où les notations 'n' et 'm' sont autorisées à prendre différentes valeurs. Dans ce cas, 'n' et 'm' seront qualifiées de « paramètres du modèle » :

! **Définition** (NB) (Paramètres du modèle)**.** Un *paramètre du modèle*, dans un modèle statistique, est une quantité nommée dans la description du modèle, qu'on autorise à prendre différentes valeurs (dans un espace qu'on précisera), mais dont la valeur sera toujours connue à l'avance pour les cas qu'on sera amené à étudier effectivement. Autrement dit, un « modèle présentant des paramètres du modèle » est en fait une *famille* de modèles au sens de la définition (LV) (ou de la définition (MF)), paramétrée par les noms des paramètres du modèle. ♡

Remarque (NC)*.* Lorsque le modèle étudié comporte un ou plusieurs paramètres de modèle, il convient que l'analyse statistique soit menée d'une façon pouvant s'appliquer à toutes les valeurs possibles de ce·s paramètre(s) ; autrement dit, les statistiques constituant nos réponses devront faire appel aux paramètres du modèle sous leur forme *littérale*.

Dans une telle situation, quand bien même notre « modèle avec paramètre(s) » est en fait une famille de modèle au sens des définitions (LV) ou (MF), on continuera à utiliser le singulier pour dire qu'on fait « une » inférence sur « ce » modèle. ♣

Remarque (ND)*.* Dans le cadre de nos exercices, nous conviendrons de l'usage suivant : on considère qu'une grandeur constitue un paramètre du modèle dès lors qu'elle reçoit une appellation littérale (et qui aurait été susceptible de prendre d'autres valeurs possibles sans nuire à la cohérence de la description) ou que la description du modèle prévoit explicitement qu'il puisse y avoir d'autres valeurs pour ce paramètre. ♣

Un cas particulier de modèles avec paramètre, qu'il est utile de connaitre, est celui des *modèles d'échantillonnage* :

! **Définition** (NE) (Modèle d'échantillonnage)**.** Nous nous plaçons ici dans le cadre des modèles explicatifs. Lorsque le modèle considéré contient un paramètre du modèle n tel que (les autres paramètres éventuels du modèle et) le paramètre caché étant fixé, l'observation est constituée par un n-uplet $(X_0, \dots, X_{n-1})$ de v.a.i.i.d. dont la loi ne dépend pas de n, alors on qualifie notre modèle le *modèle d'échantillonnage*, pour lequel le paramètre n constitue la *taille d'échantillon*. ♡

Exemple (NF)*.* Le modèle du pédagogue (dans sa version explicative) est un modèle d'échantillonnage, le paramètre de taille d'échantillon correspondant à n. ♣

Exemple (NG)*.* En revanche, le modèle du chasseur n'est pas, techniquement parlant, un modèle d'échantillonnage par rapport à n, puisque l'observation X (le *nombre* de tirs réussis) ne se décompose pas, dans le formalisme que nous avons choisi, en sous-observations indépendantes. ♣

Notion de régime asymptotique

Outre le fait qu'ils permettent de traiter d'un seul coup plusieurs variantes, les paramètres des modèles permettent d'introduire des notions de *régimes asympto-*

tiques, par rapport auxquels on pourra obtenir des résultats valables *à la limite*. Cela se formalise ainsi :

Définition (NH) (Régime asymptotique)**.** Dans un modèle statistique présentant des paramètres, on peut définir un *régime asymptotique* en considérant un certain passage à la limite d'un (ou plusieurs) paramètres du modèle. ♡ !

Exemple (NI). Dans le cadre du modèle du pédagogue (en version prédictive), on peut considérer l'asymptotique suivante : « la promotion actuelle et la seconde promotion ont la même taille, et cette taille tend vers l'infini ». Il est alors possible de démontrer, par exemple, qu'à partir de l'observation (passée), on peut construire une approximation de la quantité d'intérêt « pire score obtenu par la seconde promotion » dont l'erreur tendra vers zéro dans l'asymptotique considérée. ♧

7.5 Statistiques bayésienne et fréquentiste

J'ai écrit plus haut que, lorsqu'on définit un modèle de statistique (mettons ici, pour alléger les notations, qu'il s'agit d'un modèle explicatif), on introduit automatiquement un espace probabilisé où on a des variables aléatoires $\boldsymbol{\theta}$ et $\boldsymbol{X}$, où les lois conditionnelles $\mathrm{Loi}(\boldsymbol{X} \mid \boldsymbol{\theta} = \theta)$ sont imposées par le modèle considéré.

Néanmoins, la donnée des différentes $\mathrm{Loi}(\boldsymbol{X} \mid \boldsymbol{\theta} = \theta)$ ne permet pas de connaitre la loi de $\boldsymbol{X}$ "tout court", et encore moins celle de $\boldsymbol{\theta}$! La description que nous avons donnée de nos variables aléatoires $\boldsymbol{\theta}$ et $\boldsymbol{X}$ est donc "incomplète".

Pour rendre cette description complète, il faudrait que nous précisions aussi la loi de $\boldsymbol{\theta}$. Lorsqu'on le fait, on se retrouve dans ce qu'on appelle un modèle de *statistique bayésienne*. À l'inverse, si on se refuse à le faire, on dit qu'on fait de la statistique *fréquentiste* :

Définition (NJ) (Statistiques bayésienne et fréquentiste)**.** !!

— Un modèle de statistique *bayésienne* est un modèle statistique (explicatif ou prédictif) dans lequel, outre le modèle lui-même, on donne aussi une loi pour le paramètre caché. Cette loi est qualifiée de *loi à priori* du paramètre caché ; on l'appelle aussi plus simplement la *priore*[¶].

— On dit qu'on fait de la statistique *fréquentiste* lorsqu'on se refuse à introduire une priore. Dans ce cas, on ne devra manipuler $\boldsymbol{\theta}$ et $\boldsymbol{X}$, et faire nos analyses, que d'une façon qui soit indépendante de la loi de $\boldsymbol{\theta}$, et ne dépende donc que du modèle statistique à proprement parler !

♡

Principe (NK) (Interprétation de la priore)**.** *La loi à priori de $\boldsymbol{\theta}$ représente, dans le cadre de notre modèle, à quel point nous nous attendions à ce que $\theta_{\checkmark}$ vaille tant ou tant* avant *de commencer à regarder l'observation. Certes, au moment où on commence à collecter les données, la valeur $\theta_{\checkmark}$ est déjà fixée définitivement ; mais comme nous n'y avons pas accès, nous la traitons* comme s'il *s'agissait d'une quantité aléatoire !* !!

Une autre façon de formuler les choses est la suivante : la priore exprime la connaissance que nous avons, avant l'expérience, du paramètre caché, et des valeurs que, à nos yeux, il est plus ou moins susceptible de prendre. ♢

[¶]. En anglais : *prior*.

Exemple (NL) (Priore par défaut pour le modèle du chasseur). On peut traiter le modèle du chasseur dans un cadre bayésien en utilisant la priore suivante : de par l'expérience qu'il a déjà eue avec ses candidats précédents, sans rien savoir sur le candidat actuel, le club considère qu'il est plus probable que la fiabilité du candidat soit proche de 0 ou de 1, plutôt qu'elle ne soit proche de 1/2 : plus précisément, il considère que le paramètre caché $\boldsymbol{\theta}$ est distribué à priori selon la loi suivante, dite *loi de l'arcsinus* :

$$\mathbb{P}_{\text{à priori}}(\boldsymbol{\theta} \in \mathrm{d}\theta) = \frac{1}{\pi}\theta^{-1/2}(1-\theta)^{-1/2}\,\mathrm{vol}_1(\mathrm{d}\theta). \tag{NM}$$

Dans la suite de ce polycopié, lorsque le modèle du chasseur sera traité dans le paradigme bayésien, la priore considérée sera celle de l'arcsinus, à moins qu'il ne soit explicitement précisé autrement. ♧

Exemple (NN) (Priore par défaut pour le modèle du pédagogue). Quand nous traiterons le modèle du pédagogue dans le cadre bayésien, la priore que nous utiliserons sera la suivante (sauf mention explicite du contraire) : $\boldsymbol{\mu}$ et $\boldsymbol{\sigma}$ seront prises indépendantes à priori, avec $\boldsymbol{\mu}$ suivant (à priori) la loi Normale$(\mu_{\text{réf}} - \sigma_{\text{réf}}, \sigma_{\text{réf}}^2)$ et $\boldsymbol{\sigma}$ suivant la loi Expon$^{\text{le}}(\sigma_{\text{réf}}^{-2}/2)^{-1/2}$ (càd. la mesure image, par l'application $x \mapsto x^{-1/2}$, de la loi Expon$^{\text{le}}(\sigma_{\text{réf}}^{-2})$. On peut en l'occurrence calculer explicitement la densité correspondante : pour $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_+^*$, on a ainsi

$$\mathbb{P}_{\text{pr}}(\boldsymbol{\mu} \in \mathrm{d}\mu \text{ et } \boldsymbol{\sigma} \in \mathrm{d}\sigma) = \frac{\sigma_{\text{réf}}}{\sqrt{2\pi}}\frac{1}{\sigma^3}\exp\left(-\frac{(\mu-\mu_{\text{réf}}+\sigma_{\text{réf}})^2}{2\sigma_{\text{réf}}^2} - \frac{\sigma_{\text{réf}}^2}{2\sigma^2}\right)\mathrm{vol}_1(\mathrm{d}\mu)\,\mathrm{vol}_1(\mathrm{d}\sigma). \tag{NO}$$

Une visualisation graphique de cette priore est donnée en figure 7.1. ♧

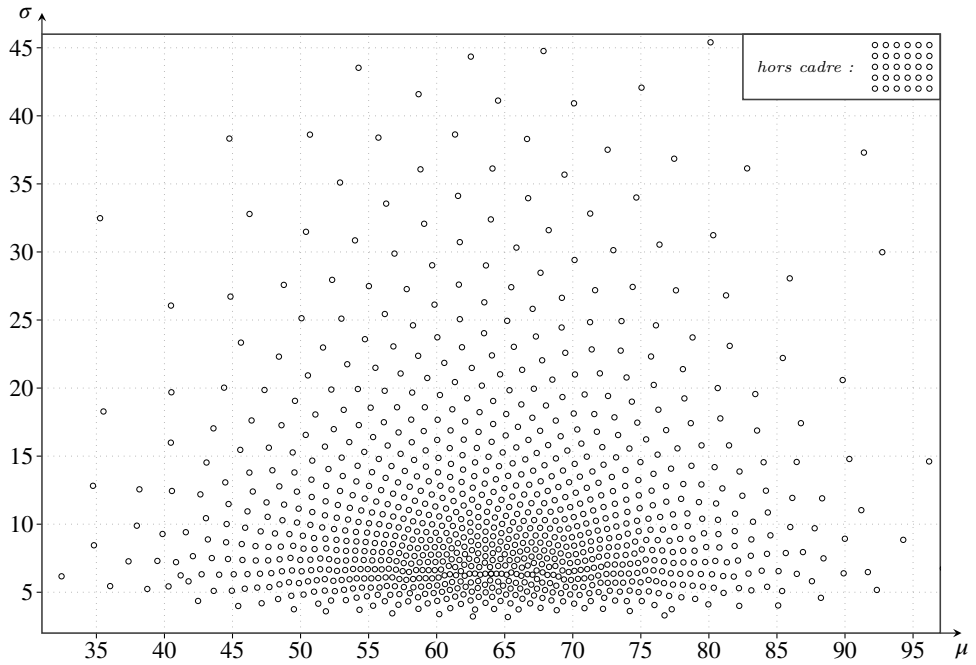

FIGURE 7.1 – Notre priore par défaut pour le modèle du pédagogue. Pour représenter cette loi sur $\mathbb{R}^2$, on a fait figurer un petit cercle par zone de probabilité 1 ‰ : ainsi, plus les cercles sont rapprochés les uns des autres, plus la densité est élevée.

Remarque (NP). Vu que la priore représente à quel point nous nous attendons à telle ou telle valeur pour le paramètre caché, elle comporte en général une part de *subjectivité* : nous verrons d'ailleurs dans la § 15 en quoi la détermination de la priore s'apparente parfois à un "art" plutôt qu'à l'application rigoureuse d'une procédure algorithmique. ♧

Remarque (NQ). Nous verrons dans le chapitre 15 comment l'idée générale de « exprimer notre connaissance du paramètre caché avant l'expérience » peut se décliner en pratique pour en arriver effectivement au choix d'une loi à priori dans une situation donnée. ♧

Dans le cas de la statistique bayésienne, deux notations très importantes interviennent : celles concernant les contextes probabilistes « à priori » et « à postériori » :

Définition (NR). Dans le cadre bayésien, on parle de « contexte probabiliste *à priori* » pour signifier que, non seulement on traite le paramètre caché comme une "authentique" variable aléatoire (i.e., on ne conditionne pas par une valeur donnée de $\boldsymbol{\theta}$), mais surtout, qu'on n'a pas conditionné par l'information sur la valeur de l'observation. Le contexte probabiliste à priori sera noté « $\mathbb{P}_{\text{pr}}(\bullet)$ » : bien que cette notation soit *techniquement* synonyme de $\mathbb{P}(\bullet)$, il est fortement préférable de souligner explicitement cette absence de conditionnement !

Par contraste, on parle de « contexte probabiliste *à postérori* » lorsque, toujours dans le cas où $\boldsymbol{\theta}$ est une authentique variable aléatoire, on raisonne conditionnellement à la valeur effectivement prise par l'observation. Le contexte à postériori sera noté "$\mathbb{P}_{\text{post}}(\bullet)$" : cette notation est techniquement synonyme de $\mathbb{P}(\bullet \mid X = x_{\checkmark})$. ♡

7.6 Description d'un modèle statistique en langage ordinaire

Point (NS). Dans les exercices, on décrira souvent les modèles de statistique à l'aide de phrases du langage ordinaire (i.e. non mathématique). Typiquement les énoncés prendront plus ou moins la forme suivante (où les passages entre crochets en italique désignent la forme générale de ce qui s'insèrera dans l'énoncé, et où les passages entre parenthèses pourront éventuellement être absents) : !

> *[Tel objet du monde réel ou abstrait]* (que nous noterons *[notation du type 'X' ou du type ('X', 'Y')]*) suit la loi *[description d'une loi pour l'objet dépendant de certains paramètres, les paramètres en question étant du type '$\theta_{\checkmark}$' (et 'λ')]*, où (*[un ou plusieurs des paramètres, ceux du type λ]* vaut *[telle valeur connue pour λ]*) et *[un ou plusieurs des paramètres, ceux du type $\theta_{\checkmark}$]* est inconnu, à valeurs dans *[un certain ensemble pour $\theta_{\checkmark}$]*. (On suppose à priori que *[$\theta_{\checkmark}$]* est la réalisation d'une variable aléatoire (notée *[une notation de type $\boldsymbol{\theta}$]*) suivant la loi *[une certaine loi pour $\boldsymbol{\theta}$]*).
>
> Dans la situation à laquelle nous sommes confrontés en pratique, on a trouvé que *[(la partie en X de) l'objet]* a pris la valeur *[telle valeur pour X]* (que nous noterons *[notation du type '$x_{\checkmark}$']*).
>
> Notre but est de mieux savoir ce que vaut *[quantité de la forme $\gamma(\theta_{\checkmark})$ ou de la forme $g(Y)$]*, en en déterminant *[non d'une méthode statistique, p. ex. « un estimateur » ou « un intervalle de prédiction »]*.

Un tel texte s'interprète de la façon suivante : *[tel objet du monde réel ou abstrait]* se réfère à l'observation complétée (s'il y a lieu de compléter, s'entend). *[La partie X de l'objet]* est l'observation. Lorsqu'il n'y a pas de *[partie Y de l'objet]*, c'est qu'on est dans un modèle d'inférence explicative ; et lorsque l'observation complétée est du type (X, Y) *[la partie Y de l'objet]* constitue l'observation future, et on est alors dans un modèle de prédiction statistique. *[Les paramètres du type $\theta_{\checkmark}$]* sont la réalisation du paramètre caché, *[les paramètres du type λ]* étant les paramètres du modèle. L'espace du paramètre caché est *[un certain ensemble pour $\theta_{\checkmark}$]* Un point délicat est l'interprétation de la description de *[une loi pour l'objet]* : en effet, l'énoncé décrit une *unique* loi dans laquelle intervient un $\theta_{\checkmark}$ *inconnu* ; mais du point de vue du modèle statistique, il faut comprendre qu'il s'agit implicitement de la description d'une *famille* de lois indexée par *[un certain ensemble pour $\theta_{\checkmark}$]*, une pour chaque valeur θ dans cet ensemble. Lorsque la phrase commençant par « on suppose à priori que » est présente, cela signifie qu'on introduit un contexte bayésien, avec pour priore *[une certaine loi pour $\boldsymbol{\theta}$]*. *[Telle valeur pour X]* est la valeur effective de l'observation (passée). *[La quantité de la forme $\gamma(\theta_{\checkmark})$ ou $g(Y)$]* désigne la quantité d'intérêt qui fait l'objet de notre étude : si celle-ci est de la forme $\gamma(\theta_{\checkmark})$, c'est qu'on s'intéresse à une question d'inférence ; si elle est de la forme $g(Y)$, c'est qu'on s'intéresse à une question de prédiction. ♧

Remarque (NT). Dans cette formulation, l'énoncé donne seulement la loi de l'observation (complétée) sous le *véritable* contexte probabiliste, autrement dit, sachant la vraie valeur $\theta_{\checkmark}$ du paramètre caché ; mais il le fait sans rien préciser sur $\theta_{\checkmark}$: c'est pourquoi on peut considérer que cette description fournit en fait une *famille* de lois indexée par $\boldsymbol{\Theta}$. Cette simplification rédactionnelle sera régulièrement utilisée dans le polycopié ainsi que dans les exercices, parfois même implicitement. Elle justifie ainsi l'introduction de deux nouvelles notations : ♧

Notation (NU). Pour $\theta \in \boldsymbol{\Theta}$ une valeur possible du paramètre caché, le contexte probabiliste $\mathbb{P}(\bullet \mid \boldsymbol{\theta} = \theta)$ — autrement dit, le contexte sachant que le paramètre caché vaut θ — pourra également être noté « $\mathbb{P}_{\theta}(\bullet)$ ».

Parmi ces contextes, on appelle « véritable contexte probabiliste », et on note "$\mathbb{P}_{\checkmark}(\bullet)$", le contexte probabiliste correspondant à la vraie valeur du paramètre caché : autrement dit, $\mathbb{P}_{\checkmark}(\bullet) \coloneqq \mathbb{P}_{\theta_{\checkmark}}$. ♡

Revenons maintenant au point (NS), en illustrant celui-ci par un exemple concret.

Exemple (NV). Voici un avatar du modèle du pédagogue qui pourrait constituer le début d'un énoncé d'exercice :

> Dans le cadre d'un nouveau cours, une enseignante s'attend à ce que les notes de ses élèves soient i.i.d. suivant une loi Normale$(\mu_{\checkmark}, \sigma^2_{\checkmark})$, pour des valeurs de $\mu_{\checkmark}$ et $\sigma_{\checkmark}$ qui sont encore complètement inconnues pour lui (puisque le cours est nouveau). L'enseignante donne le cours deux années de suite, d'abord à une promotion de n élèves, puis à une promotion de p élèves. Au moment que nous considérons, l'enseignante a gardé en mémoire les notes de la première promotion, mais la correction de la seconde promotion n'ayant pas encore eu lieu, les notes de celle-ci sont encore inconnues. Notre enseignante se demande, avant de commencer à corriger les copies, à quelles valeurs elle peut raisonnablement s'attendre pour la moyenne de la seconde promotion. Pour les

applications numériques, nous prendrons $n = 16$, $p = 21$, et les notes de la première promotion (prises dans un ordre arbitraire, p. ex. l'ordre alphabétique des noms) sont : 13,2, 9,1, 1,6, 8,6, 10,7, 12,9, 11,9, 8,9, 15,5, 13,1, 13,5, 11,6, 11,7, 11,4, 10,0, 10,9.

Dans ce cas, l'interprétation en termes de modèle est la suivante :

Observation complétée : Notes de tous les élèves des deux promotions. Cette observation complétée se divise en une observation passée, correspondant aux notes de la première promotion, et une observation future, correspondant aux notes de la seconde promotion. Si on est amené à introduire soi-même des notations pour les observation, un choix logique sera de noter l'observation (passée) comme le n-uplet $(X_0, \dots, X_{n-1})$, les X_i correspondant aux notes respectives des élèves, et l'observation future comme le p-uplet $(Y_0, \dots, Y_{p-1})$. Concernant l'espace de l'observation, les notes doivent être considérées comme réelles puisqu'elles suivent des lois normales : l'espace de l'observation passée est donc $\mathbb{R}^n$, et celui de l'observation future est $\mathbb{R}^p$. (Comme nous le verrons, n et p sont ici des paramètres du modèle, de sorte qu'ils ont parfaitement le droit d'intervenir dans la description des espaces des observations !).

Observation effective : Les 16 notes données en fin d'énoncé correspondent à la réalisation de l'observation (passée) : $x_{0\checkmark} \coloneqq 13{,}2$, $x_{1\checkmark} \coloneqq 9{,}1$, etc.

Paramètre caché : (La réalisation de) le paramètre caché est constitué du couple $(\mu_\checkmark, \sigma_\checkmark)$, puisque ce sont les quantités inconnues qui déterminent la loi de l'observation (complétée). Les variables aléatoires associées à ces réalisations pourront être notées resp. $\boldsymbol{\mu}$ et $\boldsymbol{\sigma}$. L'espace du paramètre caché n'ayant pas été précisé ici, nous allons le prendre aussi large que mathématiquement possible : $\boldsymbol{\mu}$ sera donc à valeurs dans $\mathbb{R}$ et $\boldsymbol{\sigma}$ à valeurs dans $\mathbb{R}_+$.

Loi de l'observation (complétée) sachant le paramètre caché : C'est le point le plus important dans la description d'un modèle ! L'affirmation que les notes des différents élèves sont i.i.d. de loi Normale$(\mu_\checkmark, \sigma^2_\checkmark)$ correspond à la description de la loi de l'observation complétée sachant la vraie valeur du paramètre caché. En termes plus formels, cela peut se traduire par le fait que $\mathrm{Loi}_\checkmark(X, Y) = \mathrm{Normale}(\mu_\checkmark, \sigma^2_\checkmark)^{\otimes(n+p)}$. Comme on a écrit cela sans aucune connaissance de la valeur $(\mu_\checkmark, \sigma_\checkmark)$, cela signifie en fait que, pour tout $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^*$, on a

$$\mathrm{Loi}(X, Y \mid \boldsymbol{\mu} = \mu \text{ et } \boldsymbol{\sigma} = \sigma) = \mathrm{Normale}(\mu, \sigma^2)^{\otimes(n+p)}. \qquad (\mathtt{NW})$$

Paramètres du modèle : Deux autres paramètres interviennent dans la description du modèle, qui sont n et p : ce ne sont pas des paramètres cachés dans la mesure où la taille des promotions est parfaitement connue pour notre analyse statistique (en l'occurrence, l'énoncé nous précise même que $n \coloneqq 16$ et $p \coloneqq 21$) ; ce sont donc des paramètres du modèle !

Priore : Aucune priore n'étant mentionnée pour le modèle, nous serons forcément dans un cadre fréquentiste.

Quantité d'intérêt : L'énoncé dit explicitement que la quantité d'intérêt constituant notre objectif sera la moyenne des Y_j, soit $\left(\sum_{j=1}^p Y_j\right) / p$: c'est bien une fonction de l'observation future, comme il fallait s'y attendre pour un modèle de prédiction.

♧

7.7 Récapitulatif des principales notations

Variables aléatoires, valeurs possibles et réalisations

Notation (NX). Dans tout ce polycopié, je suivrai autant que possible l'usage de désigner les *variables aléatoires* par des lettres majuscules (X, Y, M, S, ...), et les *valeurs* de ces variables aléatoires par les lettres minuscules correspondantes (x, y, m, s, ...).

Pour les variables aléatoires désignées par des lettres *grecques*, le contraste « majuscule / minuscule » sera remplacé par le contraste « gras / maigre » : les variables aléatoires seront donc désignées par $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}$, ..., et leurs valeurs par θ, α, β, λ, ... Si vous trouvez pénible de rendre le gras en écriture manuscrite, vous pourrez émuler le gras par du souligné ($\underline{\theta}$, $\underline{\alpha}$, $\underline{\beta}$, $\underline{\lambda}$, ...) ou par du double-barres ($\mathbb{\theta}$, $\mathbb{\alpha}$, $\mathbb{\beta}$, $\mathbb{\lambda}$, ...). ♡

Notation (NY). Pour distinguer la valeur *effectivement prise* par une variable aléatoire (autrement dit, sa réalisation), on utilisera le symbole '✓' (« coche ») en indice : $x_{\checkmark}$ sera ainsi la valeur effectivement prise par la variable aléatoire X. Pour désigner d'autres valeurs possibles remarquables pour X, on utilisera des notations comme x_0, x_1, $x_{\text{réf}}$, x_*, ... On utilisera enfin le simple 'x' pour désigner une valeur possible complètement générique, ou pour une variable muette. ♡

Remarque (NZ). Dans le cadre de ce cours, afin de gagner du temps, l'association entre lettres majuscules et majuscules (ou grasses et maigres), ainsi que l'usage de la coche, sera souvent implicite : si une variable aléatoire s'appelle W, on pourra manipuler directement la notation '$w_{\checkmark}$' sans préciser au préalable que cela désigne la valeur effectivement prise par W... ! ♧

Remarque (OA). Dans vos exercices et examens, une tolérance vous sera accordée concernant l'omission du symbole '✓' (sauf si celui-ci figure explicitement dans l'énoncé) ou du passage en gras, à moins que cela ne crée un souci d'intelligibilité [‖]. ♧

Il existe une convention concernant l'usage des lettres latines et grecques en statistique :

Convention (OB).
- Dans le cadre de ce polycopié, on utilisera des lettres grecques pour désigner le paramètre caché, ainsi que pour les quantités d'intérêt explicatives et pour leurs estimateurs (sur ce dernier point, confer chapitre 11).
- À l'inverse, on utilisera des lettres latines pour désigner l'observation et l'observation future, ainsi que pour les statistiques de test (sur ce point, confer chapitre 12) et les quantités d'intérêt prédictives.
- Les paramètres du modèle, quant à eux, pourront aussi bien être désignés par des lettres latines que par des lettres grecques.

♡

Remarque (OC). Attention : rien ne garantit en revanche que ces conventions soient respectées dans vos exercices ! En effet, dans une situation industrielle où vous aurez à proposer vous-mêmes un modèle statistique, il va de soi que personne n'aura pris soin de respecter les conventions en amont à votre place... ♧

[‖]. Par exemple, il faudra impérativement respecter le gras lorsqu'on écrira « $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} = \theta)$ » : sinon, on obtiendrait « $\mathbb{P}_{\text{post}}(\theta = \theta)$ », où il y aurait la même chose des deux côtés du signe '='...!

Notations raccourcies

Dans les pages précédentes, nous avons introduit différentes notations condensées pour certains contextes probabilistes que nous serons amenés à rencontrer particulièrement souvent en statistique. Récapitulons-les ci-dessous :

Notation (OD).
- Pour θ une valeur possible pour le paramètre caché, le contexte probabiliste sachant que le paramètre caché vaut θ (autrement dit, le contexte $\mathbb{P}(\bullet \mid \boldsymbol{\theta} = \theta)$) pourra être noté "$\mathbb{P}_{\theta}(\bullet)$".
- Le véritable contexte probabiliste, celui connaissant la vraie valeur du paramètre caché (autrement dit, le contexte $\mathbb{P}_{\theta_{\checkmark}}(\bullet)$), pourra être noté "$\mathbb{P}_{\checkmark}(\bullet)$".
- En statistique bayésienne, lorsqu'on ne conditionne pas par rapport à la valeur de l'observation (ni par quoi que ce soit d'autre), on note le contexte probabiliste correspondant "$\mathbb{P}_{\mathrm{pr}}(\bullet)$", histoire de bien clarifier cette absence de conditionnement. Techniquement parlant, c'est la même chose que le contexte $\mathbb{P}(\bullet)$; mais on préfère utiliser des notations qui rendent toujours explicite le contexte considéré.
- Toujours en statistique bayésienne, le contexte probabiliste à postériori (autrement dit, le contexte $\mathbb{P}(\bullet \mid X = x_{\checkmark})$, où '$X$' se réfère à l'observation et '$x_{\checkmark}$' à sa réalisation effective) pourra être noté "$\mathbb{P}_{\mathrm{post}}(\bullet)$".

♡

Un autre type de raccourci fréquemment utilisé sera le suivant :

Notation (OE). Dans l'expression d'une probabilité de la forme « $\mathbb{P}(X = x)$ », on pourra sous-entendre le "$X =$" et écrire simplement « $\mathbb{P}(x)$ ». De même, on pourra simplifier « $\mathbb{P}(X \in \mathrm{d}x)$ » en « $\mathbb{P}(\mathrm{d}x)$ ». Similairement, à droite d'une barre de conditionnement, une expression comme « $\mathrm{Loi}(Y \mid X = x)$ » pourra être raccourcie en « $\mathrm{Loi}(Y \mid x)$ ». ♡

Notations génériques

Dans le cadre de ce cours, lorsque nous voudrons introduire de nouvelles méthodes d'analyse statistique, il serait un peu pénible de devoir re-préciser à chaque fois toutes les notations permettant de se référer aux différents aspects d'un modèle statistique... Pour faciliter les choses, nous pourrons donc utiliser un certain nombre de notations par défaut, que j'appellerai « notations génériques » dans la suite de ce polycopié :

Définition (OF) (Notations génériques)**.** La phrase « *Soit un modèle statistique explicatif avec les notations génériques* » signifie : « Soit un modèle statistique dans lequel :
- L'espace du paramètre caché est noté '$\boldsymbol{\Theta}$' ;
- Le paramètre caché est noté '$\boldsymbol{\theta}$' ;
- L'espace de l'observation est noté '$\mathscr{X}$' ;
- L'observation est notée '$\boldsymbol{X}$' ;
- La notation '$\boldsymbol{\gamma}$' se réfère à une quantité d'intérêt explicative, qui peut aussi s'écrire "$\gamma(\boldsymbol{\theta})$" ;
- L'espace dans lequel la quantité d'intérêt $\boldsymbol{\gamma}$ prend ses valeurs est noté '$\mathscr{G}$' ».

La phrase « *Soit un modèle statistique prédictif avec les notations génériques* » signifie la même chose que ci-dessus, aux nuances suivantes près :

— Cette fois-ci, il y a également une observation future, notée '$\boldsymbol{Y}$' ;
— L'espace dans lequel $\boldsymbol{Y}$ prend ses valeurs est noté '$\mathscr{Y}$' ;
— La notation '$\boldsymbol{\gamma}$' n'est plus censée se référer à quelque chose en particulier, pas plus que la notation "$\gamma(\bullet)$" ;
— En revanche, la notation '$\boldsymbol{G}$' se réfère maintenant à une quantité d'intérêt prédictive, qui peut aussi s'écrire "$g(\boldsymbol{\theta})$" ;
— La notation '$\mathscr{G}$' correspond à présent à l'espace dans lequel la quantité d'intérêt $\boldsymbol{G}$ prend ses valeurs.

En outre, en cas d'utilisation des notations génériques, si nous souhaitons considérer une certaine asymptotique pour un paramètre du modèle, nous appellerons ce paramètre $\boldsymbol{\lambda}$ et désignerons l'asymptotique considérée par « $\boldsymbol{\lambda} \to \boldsymbol{\infty}$ ». ♡

Chapitre 8

Le théorème de Bayes

8.1 Notion de postériore

Contextes à postériori et à priori

(Ce qui suit est écrit en notations génériques). En statistique bayésienne, le fait de disposer d'une loi de probabilité pour $\boldsymbol{\theta}$ (la « priore ») permet d'en déduire, via les lois conditionnelles $\mathrm{Loi}(X, Y \mid \boldsymbol{\theta} = \theta)$, la loi jointe du triplet $(\boldsymbol{\theta}, X, Y)$. Par conséquent, on peut aussi de mettre à parler de la loi de $\boldsymbol{\theta}$ et de Y *conditionnellement à telle ou telle valeur de* X, et donc en particulier, de leur loi sachant la valeur qui a été *effectivement observée* pour X ! On parle alors de *probabilités à postériori* :

Définition (OG) (À postériori)**.** En statistique bayésienne, le *contexte probabiliste à postériori* désigne le contexte conditionné par rapport à $\{X = x_{\checkmark}\}$ (en notations génériques). Ainsi, pour A un évènement, la probabilité à postériori de A désigne la probabilité $\mathbb{P}(A \mid X = x_{\checkmark})$; et on alors en déduire les notions de loi à postériori, espérance à postériori, variance à postériori, etc. !!

Dans le cadre de ce cours, rappelons que, plutôt que d'écrire explicitement « $\mathbb{P}(\bullet \mid X = x_{\checkmark})$ », on pourra désigner une probabilité à postériori par la notation $\mathbb{P}_{\mathrm{post}}(\bullet)$ (ou $\mathbb{P}_{\mathrm{po}}(\bullet)$), et de même une loi à postériori par $\mathrm{Loi}_{\mathrm{post}}(\bullet)$, etc. ♡

Principe (OH). *Un modèle de statistique bayésienne donnant toute l'information suffisante au calcul de la loi jointe de* $(\boldsymbol{\theta}, X, Y)$*, on peut toujours y calculer les probabilités (resp. lois, &c.) à postériori ! À l'inverse, en statistique fréquentiste, la notion de probabilité à postériori, et tout ce qui s'y rapporte, n'a* pas de sens*, car le modèle ne donne pas suffisamment d'information pour qu'on puisse les définir… !* ◇ !!

Définition (OI). La loi à postériori du paramètre caché est appelée simplement « la postériore ». ♡

À l'inverse, dans une situation bayésienne, lorsqu'on souhaite insister sur le fait qu'on n'a pas conditionné par la valeur de l'observation, on parle de contexte probabiliste à priori :

Définition (OJ). On utilise la locution « à priori » pour souligner qu'on se place dans un contexte probabiliste où on n'a pas conditionné par la valeur de l'observation. Le contexte probabiliste à priori sera noté $\mathbb{P}_{\mathrm{pr}}(\bullet)$. !

Bien que "$\mathbb{P}_{\mathrm{pr}}(\bullet)$" soit formellement synonyme de "$\mathbb{P}(\bullet)$", on s'efforcera de préciser systématiquement l'aspect « à priort » dès lors que l'objet dont on parle dépend du choix de la priore. ♡

De la loi du paramètre caché à celle des quantités d'intérêt

Dans quelques pages, la section 8.2 nous dira comment, dans un modèle de statistique bayésienne, on peut déterminer la postériore, autrement dit la loi à postériori du paramètre caché $\boldsymbol{\theta}$. Dans la présente sous-section, nous allons voir l'argument qui justifie toute l'importance de la postériore dans nos analyses : à savoir que, à partir de la postériore, on peut déterminer les lois à postériori de nos quantités d'intérêt !

Ci-dessous, j'ai indiqué divers points comme étant à retenir : en l'occurrence, c'est la *méthode de raisonnement* qui doit être retenue, pas le résultat lui-même !

! *Point* (OK) (Loi à postériori d'une quantité d'intérêt explicative). Commençons par le cas d'une quantité d'intérêt explicative $\gamma(\boldsymbol{\theta}) =: \boldsymbol{\gamma}$, où $\gamma(\bullet)$ est une certaine fonction déterministe de $\boldsymbol{\Theta}$ dans un certain espace $\mathscr{G}$. Il est clair que, si nous connaissons la loi à postériori de $\boldsymbol{\theta}$, nous pouvons en déduire la loi de $\boldsymbol{\gamma}$ dans ce même contexte : en effet, pour $A \subseteq \mathscr{G}$, on a que

$$\mathbb{P}_{\mathrm{post}}(\boldsymbol{\gamma} \in A) = \mathbb{P}_{\mathrm{post}}(\gamma(\boldsymbol{\theta}) \in A) = \mathbb{P}_{\mathrm{post}}(\boldsymbol{\theta} \in \gamma^{-1}(A)) = \mathbb{P}(\mathrm{Loi}_{\mathrm{post}}(\boldsymbol{\theta}) \in \gamma^{-1}(A)). \tag{OL}$$

Par conséquent, si nous connaissons la postériore, nous sommes en mesure de déterminer la loi à postériori de la quantité d'intérêt $\boldsymbol{\varphi}$, et donc par ricochet ses espérance à postériori, quantiles à postériori, etc. ♧

Remarque (OM). En fait, le raisonnement ci-dessus revient à utiliser que la loi à postériori de $\gamma(\boldsymbol{\theta})$ est la mesure-image, par l'application $\gamma(\bullet)$, de la loi à postériori de $\boldsymbol{\theta}$. ♧

Remarque (ON). En fait, le raisonnement du point (OK) n'a rien de spécifique au contexte probabiliste à postériori : on pourrait donc, exactement de la même manière, déterminer la loi à priori de $\boldsymbol{\gamma}$ à partir de la priore ! Le raisonnement n'a rien non plus de spécifique au fait qu'on s'intéresse à une fonction de $\boldsymbol{\theta}$: on pourrait similairement, sous le contexte probabiliste sachant la valeur du paramètre caché, déterminer la loi $\mathrm{Loi}_{\theta}(t(X))$ d'une statistique ou celle $\mathrm{Loi}_{\theta}(g(Y))$ d'une quantité d'intérêt à partir de resp. $\mathrm{Loi}_{\theta}(X)$ et $\mathrm{Loi}_{\theta}(Y)$! ♧

! *Point* (OO) (Loi à priori de l'observation). À présent, regardons comment déterminer la loi à priori de l'observation. (On utilise ci-après les notations génériques). « Déterminer la loi de $\boldsymbol{X}$ », cela revient à calculer la probabilité d'avoir $\{\boldsymbol{X} \in \mathrm{d}x\}$ pour $\mathrm{d}x$ un voisinage infinitésimal d'un point arbitraire $x \in \mathscr{X}$: nous considérons donc dans la suite un tel voisinage infinitésimal.

Le modèle statistique nous donnant la loi de l'observation *sachant telle ou telle valeur du paramètre caché*, nous allons utiliser la formule des probabilités totales pour faire apparaitre les différentes valeurs possibles du paramètre caché :

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{X} \in \mathrm{d}x) = \int_{\theta \in \boldsymbol{\Theta}} \mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta \text{ et } \boldsymbol{X} \in \mathrm{d}x). \tag{OP}$$

De là, on fait intervenir la loi sachant la valeur de $\boldsymbol{\theta}$ en utilisant la règle de chaine :

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta \text{ et } \boldsymbol{X} \in \mathrm{d}x) = \mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta)\, \mathbb{P}_{\mathrm{pr}}(\boldsymbol{X} \in \mathrm{d}x \mid \boldsymbol{\theta} \in \mathrm{d}\theta). \tag{OQ}$$

Le membre de droite ci-dessus se simplifie à trois niveaux. Premièrement, conditionner par un évènement de précision infinitémisale comme $\{\boldsymbol{\theta} \in \mathrm{d}\theta\}$, cela signifie en fait la même chose que conditionner par $\{\boldsymbol{\theta} = \theta\}$. Deuxièmement, quand on considère le contexte $\mathbb{P}_{\mathrm{pr}}(\bullet \mid \boldsymbol{\theta} = \theta)$ (qui, formellement, est synonyme de $\mathbb{P}(\bullet \mid \boldsymbol{\theta} = \theta)$), il n'y a en fait pas lieu de préciser « à priori » : en effet, ce contexte-ci ayant été fixé par le modèle indépendamment du choix de la priore, il s'agit d'un concept fréquentiste ! Troisièmement, on se rappelle que le contexte $\mathbb{P}(\bullet \mid \boldsymbol{\theta} = \theta)$ peut aussi être noté $\mathbb{P}_\theta(\bullet)$: au final, « $\mathbb{P}_{\mathrm{pr}}(\boldsymbol{X} \in \mathrm{d}x \mid \boldsymbol{\theta} \in \mathrm{d}\theta)$ » peut donc se ré-écrire simplement « $\mathbb{P}_\theta(\boldsymbol{X} \in \mathrm{d}x)$ » !

En mettant bout à bout tout ce qui précède, on a finalement obtenu la loi à priori de $\boldsymbol{X}$:

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{X} \in \mathrm{d}x) = \int_{\theta \in \boldsymbol{\Theta}} \mathbb{P}_\theta(\boldsymbol{X} \in \mathrm{d}x)\, \mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta). \tag{OR}$$

♧

Remarque (OS). Le même raisonnement, *mutatis mutandis*, permettrait également de déterminer la loi à priori de $\boldsymbol{Y}$. ♧

Exemple (OT). Appliquons la méthodologie ci-dessus en calculant, dans le modèle du chasseur, la probabilité à priori (en utilisant la priore de l'arcsinus pour $\boldsymbol{\theta}$) que le candidat rate *tous* ses plateaux : autrement dit, il s'agit de déterminer $\mathbb{P}_{\mathrm{pr}}(\boldsymbol{X} = 0)$.

Par la formule des probabilités totales, on décompose selon la valeur de $\boldsymbol{\theta}$:

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{X} = 0) = \int_{\theta \in]0,1[} \mathbb{P}(\boldsymbol{\theta} \in \mathrm{d}\theta \text{ et } \boldsymbol{X} = 0). \tag{OU}$$

La règle de chaine, nous donne alors

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{X} = 0) = \int_{\theta \in]0,1[} \mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta)\, \mathbb{P}(\boldsymbol{X} = 0 \mid \boldsymbol{\theta} = \theta). \tag{OV}$$

Dans cette expression, nous avons alors clairement distingué le facteur $\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta)$, qui dépend uniquement de la priore et vaut $\frac{1}{\pi}\theta^{-1/2}(1-\theta)^{-1/2}\,\mathrm{vol}_1(\mathrm{d}\theta)$, et le facteur fréquentiste $\mathbb{P}(\boldsymbol{X} = 0 \mid \boldsymbol{\theta} = \theta)$, qui peut être déterminé indépendamment de la nature de la priore : dans la mesure où $\mathrm{Loi}_\theta(\boldsymbol{X}) = \mathrm{Binom}^{\mathrm{le}}(n, \theta)$, on a $\mathbb{P}_\theta(\boldsymbol{X} = 0) = \mathbb{P}(\mathrm{Binom}^{\mathrm{le}}(n, \theta) = 0 \mid =)(1-\theta)^n$. Finalement, la probabilité à priori que le candidat ne touche aucun plateau vaut

$$\tfrac{1}{\pi} \int_{\theta \in]0,1[} \theta^{-1/2}(1-\theta)^{n-1/2}\,\mathrm{vol}_1(\mathrm{d}\theta) = \frac{1}{\pi} \int_{\theta=0}^{1} \theta^{-1/2}(1-\theta)^{n-1/2}\mathrm{d}\theta, \tag{OW}$$

dont on calcule numériquement que cela vaut 11,2 %. ♧

Pour déterminer la loi *à postériori* de l'observation future, la méthode est la même, mais il y a une subtilité supplémentaire à cause du conditionnement par $\{\boldsymbol{X} = x_{\checkmark}\}$:

Point (OX) (Loi à postériori de l'observation future). Avec les notations génériques, pour $y \in \mathscr{Y}$ et $\mathrm{d}y$ un voisinage infinitésimal de y, on commence comme précédemment par découper selon la valeur de $\boldsymbol{\theta}$, puis par utiliser la règle de chaine : !

$$\begin{aligned}\mathbb{P}_{\mathrm{post}}(\boldsymbol{Y} \in \mathrm{d}y) &= \int_{\theta \in \boldsymbol{\Theta}} \mathbb{P}_{\mathrm{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta \text{ et } \boldsymbol{Y} \in \mathrm{d}y) \\ &= \int_{\theta \in \boldsymbol{\Theta}} \mathbb{P}_{\mathrm{post}}(\boldsymbol{Y} \in \mathrm{d}y \mid \boldsymbol{\theta} \in \mathrm{d}\theta)\, \mathbb{P}_{\mathrm{post}}(\boldsymbol{\theta} = \theta). \end{aligned} \tag{OY}$$

La dernière étape consiste à ré-écrire $\mathbb{P}_{\text{post}}(Y \in \mathrm{d}y \mid \boldsymbol{\theta} = \theta)$; mais cette fois-ci c'est un peu plus délicat à comprendre. Nous nous rappelons que le contexte probabiliste "$\mathbb{P}_{\text{post}}(\bullet)$" signifie en fait "$\mathbb{P}(\bullet \mid X = x_{\checkmark})$" sauf que là, nous avons un conditionnement supplémentaire par $\{\boldsymbol{\theta} = \theta\}$... Un tel "double conditionnement", en fait, signifie qu'on a conditionné par la *conjonction* des deux évènements considérés ! Par conséquent, $\mathbb{P}_{\text{post}}(Y \in \mathrm{d}y \mid \boldsymbol{\theta} = \theta)$ se ré-interprète simplement comme $\mathbb{P}(Y \in \mathrm{d}y \mid X = x_{\checkmark}$ et $\boldsymbol{\theta} = \theta)$, ou encore comme $\mathbb{P}(Y \in \mathrm{d}y \mid \boldsymbol{\theta} = \theta$ et $X = x_{\checkmark})$. Mais ici, nous pouvons décider d'interpréter le conditionnement par $\{\boldsymbol{\theta} = \theta\}$ comme relevant du *contexte* probabiliste : et finalement, $\mathbb{P}_{\text{post}}(Y \in \mathrm{d}y \mid \boldsymbol{\theta} = \theta)$ est la même chose que $\mathbb{P}_{\theta}(Y \in \mathrm{d}y \mid X = x_{\checkmark})$.

Une fois arrivé à cette écriture, on comprend que $\mathbb{P}_{\theta}(Y \in \mathrm{d}y \mid X = x_{\checkmark})$ est en fait un objet fréquentiste, dans la mesure où la loi jointe $\mathrm{Loi}_{\theta}(X, Y)$ est entièrement déterminée par le modèle lui-même, sans tenir compte de la priore. On a ainsi obtenu une expression de la loi à postériori de Y à partir de la postériore et de quantités fréquentistes :

$$\mathbb{P}_{\text{post}}(Y \in \mathrm{d}y) = \int_{\theta \in \Theta} \mathbb{P}_{\theta}(Y \in \mathrm{d}y \mid X = x_{\checkmark})\, \mathbb{P}_{\text{post}}(\boldsymbol{\theta} = \theta). \tag{OZ}$$

♧

L'ultime point serait de savoir comment déterminer la loi à priori ou à postériori d'une *quantité d'intérêt prédictive* à partir de resp. la priore et la postériore. Mais en fait, pour cela, il suffit de combiner les points précédents avec l'idée de mesure-image abordée dans le point (OK) : on détermine d'abord la loi à postériori de Y; puis on en déduit celle de $g(Y)$ par mesure-image ! ☺

8.2 La formule de Bayes

Dérivation de la formule

☛ *Le but de la présente sous-section est uniquement d'expliquer d'où provient la formule de Bayes que nous allons présenter dans la suite ; le lecteur peu curieux de comprendre le pourquoi des choses peut donc la sauter sans dommage.*

Supposons donné un modèle de statistique bayésienne (en notations génériques) ; nous aimerions calculer la postériore à partir de la priore et de la description du modèle. « Calculer la postériore » revient à déterminer $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta)$ pour $\mathrm{d}\theta$ un voisinage infinitésimal d'un point arbitraire $\theta \in \Theta$; soit donc un tel $\mathrm{d}\theta$.

Par du contexte à postériori, $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta)$ désigne la probabilité conditionnelle $\mathbb{P}(\boldsymbol{\theta} \in \mathrm{d}\theta \mid X = x_{\checkmark})$; et par définition du conditionnement, cette probabilité conditionnelle vaut

$$\frac{\mathbb{P}_{\text{pr}}(X \in \mathrm{d}x_{\checkmark} \text{ et } \boldsymbol{\theta} \in \mathrm{d}\theta)}{\mathbb{P}_{\text{pr}}(X \in \mathrm{d}\checkmark)}, \tag{PA}$$

pour $\mathrm{d}x_{\checkmark}$ un voisinage infinitésimal de $x_{\checkmark}$. [L'introduction de ce voisinage servant à éviter d'avoir un quotient dégénré $0/0$ dans le cas où les $\mathrm{Loi}_{\theta}(X)$ seraient diffuses en $x_{\checkmark}$].

On applique alors la règle de chaine au numérateur (en utilisant à nouveau l'équivalence entre « conditionner par $\mathrm{d}\theta$ » et « conditionner par θ ») :

$$\mathbb{P}_{\text{pr}}(X \in \mathrm{d}x_{\checkmark} \text{ et } \boldsymbol{\theta} \in \mathrm{d}\theta) = \mathbb{P}_{\text{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta)\, \mathbb{P}_{\theta}(X \in \mathrm{d}x_{\checkmark})\,; \tag{PB}$$

quant au dénominateur, on va simplement utiliser que c'est une constante (vu qu'on raisonne à d✓ fixé). Finalement, on a obtenu que

$$\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) \propto \mathbb{P}_{\theta}(X \in \mathrm{d}x_{\checkmark}) \times \mathbb{P}_{\text{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta), \tag{PC}$$

où il n'y a pas besoin de préciser la constante de proportionnalité, celle-ci pouvant être retrouvée par la condition de normalisation d'une loi de probabilité : à savoir, en l'occurrence, que $\int_{\theta \in \Theta} \mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) = 1$.

Dans cette formule, le facteur $\mathbb{P}_{\theta}(X \in \mathrm{d}x_{\checkmark})$ est en général infinitésimalement petit, ce qui n'est pas commode. Néanmoins, dans un tel cas, on peut trouver (normalement) une mesure de référence $\widetilde{\mathrm{vol}}(\bullet)$ sur $\mathcal{X}$ et une fonction $\ell_{x_{\checkmark}}(\bullet)$ sur θ, telles que, indépendamment du choix du voisinage $\mathrm{d}x_{\checkmark}$, on ait toujours $\mathbb{P}_{\theta}(X \in \mathrm{d}x_{\checkmark}) = \ell_{x_{\checkmark}}(\theta)\widetilde{\mathrm{vol}}(\mathrm{d}x_{\checkmark})$: dès lors, on peut faire rentrer le facteur $\widetilde{\mathrm{vol}}(\mathrm{d}\checkmark)$ dans la constante de proportionnalité pour donner à la formule la formule suivante, sans infinitésimaux superflus :

$$\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) \propto \ell_{x_{\checkmark}}(\theta) \times \mathbb{P}_{\text{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta). \tag{PD}$$

La quantité $\ell_{x_{\checkmark}}(\theta)$ sera, dans la suite, appelée *vraisemblance que le paramètre caché vaille $\boldsymbol{\theta}$ (sachant que l'observation vaut $x_{\checkmark}$)* ; et la formule à laquelle nous sommes arrivés sera appelée *formule de Bayes*.

La formule

La formule de Bayes reponse sur le concept de *vraisemblance*, qui sera développé en détail dans le chapitre 10. La définition générale de la vraisemblance (qui n'est pas à connaitre, car en pratique on n'aura pas besoin d'y revenir pour faire nos calculs) est la suivante :

Définition (PE) (Vraisemblance)**.** On utilise ici les notations génériques. La fonction de vraisemblance du paramètre caché au vu de l'observation effectivement réalisée, notée $\theta \mapsto \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark})$ — ou simplement $\mathscr{L}(\boldsymbol{\theta} = \theta)$, voire $\mathscr{L}(\theta)$, lorsqu'il n'y a pas d'ambigüité —, est l'application qui, à $\theta \in \boldsymbol{\Theta}$, associe la probabilité $\mathbb{P}_{\theta}(X \in \mathrm{d}x_{\checkmark})$ (où $\mathrm{d}x_{\checkmark}$ désigne un voisinage infinitésimal arbitraire de $x_{\checkmark}$), *cette fonction devant néanmoins être considérée à constante multiplicative près*, où la constante (dont l'expression fera généralement intervenir le choix de $x_{\checkmark}$) doit avoir un ordre de grandeur tel que la fonction de vraisemblance ne soit ni infinitésimalement petite, ni infinitésimalement grande. ♡

Théorème (PF) (Formule de Bayes)**.** *On utilise les notations génériques. Pour une observation effective $x_{\checkmark}$, la postériore sur $\boldsymbol{\theta}$ se déduit de la priore en multipliant par la vraisemblance, à constante multiplicative près :* !!

$$\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) \propto \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark}) \times \mathbb{P}_{\text{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta). \tag{PG}$$

♢

Remarque (PH)*.* Dans le cas où la priore est une loi discrète, les infinitésimaux ne sont plus nécessaires : la formule devient alors simplement

$$\mathbb{P}_{\text{post}}(\boldsymbol{\theta} = \theta) \propto \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark}) \times \mathbb{P}_{\text{pr}}(\boldsymbol{\theta} = \theta). \tag{PI}$$

♧

!!

Remarque (PJ). La formule de Bayes ne donne la postériore qu'à constante multiplicative près ; mais on peut toujours retrouver cette constante sans avoir à refaire les calculs depuis le début, en utilisant la condition de normalisation des lois de probabilité !

Concrètement, si on a obtenu une expression de la forme « $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) \propto f(\theta)\mathrm{vol}\tilde{}(\mathrm{d}\theta)$ », alors on va chercher la constante $Z \in \mathbb{R}_+^*$ telle que $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) = Z^{-1} f(\theta)\mathrm{vol}\tilde{}(\mathrm{d}\theta)$ en écrivant que

$$1 = \int_{\theta\in\Theta} \mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) = \int_{\theta\in\Theta} Z^{-1} f(\theta)\mathrm{vol}\tilde{}(\mathrm{d}\theta) = \frac{\int_{\theta\in\Theta} f(\theta)\mathrm{vol}\tilde{}(\mathrm{d}\theta)}{Z}, \tag{PK}$$

d'où $Z = \int_{\theta\in\Theta} f(\theta)\mathrm{vol}\tilde{}(\mathrm{d}\theta)$, et finalement

$$\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) = \frac{f(\theta)}{\int_{\theta'\in\Theta} f(\theta')\mathrm{vol}\tilde{}(\mathrm{d}\theta')}\mathrm{vol}\tilde{}(\mathrm{d}\theta). \tag{PL}$$

♧

8.3 Notion de priore impropre

En analyse bayésienne, on sera régulièrement confronté à une situation qu'on appelle la *priore impropre*.

!

Définition (PM). Une priore impropre, c'est l'utilisation comme priore une loi de probabilité définie à constante multiplicative près... mais où la constante en question serait formellement infinie ! ♡

Exemple (PN). Par exemple, la « loi de probabilité uniforme sur $\mathbb{R}$ », définie formellement par

$$\boldsymbol{P}(\mathrm{d}x) \propto \mathrm{vol}_1(\mathrm{d}x), \tag{PO}$$

autrement dit « $\boldsymbol{P}(\mathrm{d}x) = Z^{-1}\,\mathrm{vol}_1(\mathrm{d}x)$ » pour une constante de normalisation Z appropriée, n'est pas une véritable loi de probabilité, car il faudrait prendre $Z = \int_{x\in\mathbb{R}} \mathrm{vol}_1(\mathrm{d}x) = \infty$... Cependant cette expression n'est pas totalement dénuée de sens, car, par exemple, elle reflète l'idée selon laquelle il est deux fois plus probable de se trouver entre 3 et 5 qu'entre 8 et 9. ♧

Le point essentiel est que, dans la plupart des cas, le fait d'avoir pris une priore impropre n'empêche pas d'avoir une posétriore bien définie :

!

Proposition (PP). *Lorsqu'on a une modèle avec une priore impropre, il faut faire tous les calculs en traitant la constante de normalisation dégénérée comme si c'était une véritable constante (mais sans jamais l'expliciter, évidemment) : cela n'empêche en rien de faire les calculs ; et la plupart du temps, on arrivera à la fin à une postériore qui sera une véritable loi de probabilité !* ♢

Exemple (PQ). Dans le cas du modèle du pédagogue, on pourrait envisager d'utiliser la « priore de Haar » [*], qui consisterait en l'occurrence à prendre

$$\mathbb{P}(\boldsymbol{\mu} \in \mathrm{d}\mu \text{ et } \boldsymbol{\sigma} \in \mathrm{d}\sigma) \propto \sigma^{-2}\,\mathrm{vol}_1(\mathrm{d}\mu)\,\mathrm{vol}_1(\mathrm{d}\sigma). \tag{PR}$$

[*]. Il s'agit d'un type de priore non informative, confer § 15.5.

Il s'agit là d'une priore impropre, car l'intégrale de $\sigma^{-2}\,\mathrm{vol}_1(\mathrm{d}\mu)\,\mathrm{vol}_1(\mathrm{d}\sigma)$ sur $\boldsymbol{\Theta}$ est infinie... Cependant, comme nous le verrons dans la § 8.4 ci-dessous, l'application de la formule de Bayes conduit alors à trouver pour postériore

$$\mathbb{P}(\boldsymbol{\mu} \in \mathrm{d}\mu \text{ et } \boldsymbol{\sigma} \in \mathrm{d}\sigma \mid X = x_{\checkmark}) \propto \sigma^{-24} \exp\left(-\frac{11 \times (\mu - 70{,}3)^2 + 2{,}54 \times 10^3}{\sigma^2}\right) \mathrm{vol}_2(\mathrm{d}\mu \times \mathrm{d}\sigma), \quad (\mathsf{PS})$$

qui cette fois-ci décrit bien une loi propre, vu que l'intégrale de $\sigma^{-24}\exp\bigl(-\bigl(11(\mu - 70{,}3)^2 + 2\,540\bigr)\sigma^{-2}\bigr)$ est convergente sur $\mathbb{R} \times \mathbb{R}_+^*$! ♧

8.4 Exemples de calculs

Calcul de postériore pour le chasseur

Quelle est la fonction de vraisemblance pour le modèle du chasseur ? $\mathscr{X}$ étant discret, nous pouvons appliquer le théorème (TY) (confer chapitre 10) :

$$\mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark}) = \mathbb{P}(X = x_{\checkmark} \mid \boldsymbol{\theta} = \theta) = \mathbb{P}\bigl(\mathrm{Binom}^{\mathrm{le}}(n, \theta) = x_{\checkmark}\bigr) = \binom{n}{x_{\checkmark}} \theta^{x_{\checkmark}} (1-\theta)^{n - x_{\checkmark}}.$$

Cependant, nous rappelant que la vraisemblance n'est définie, en tant que fonction de θ, qu'à constante multiplicative près, il s'avère judicieux de simplifier le facteur constant $\binom{n}{x_{\checkmark}}$ pour alléger les calculs, et de prendre ainsi

$$\mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark}) = \theta^{x_{\checkmark}} (1-\theta)^{n - x_{\checkmark}}. \quad (\mathsf{PT})$$

Pour calculer la loi à postériori, nous avons également besoin de nous rappeler la priore sur $\boldsymbol{\theta}$: en l'occurrence, comme expliqué en page 82, il s'agit de la loi arcsinus, caractérisée par :

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta) \coloneqq Z_{\mathrm{pr}}^{-1} \theta^{-1/2} (1-\theta)^{-1/2}\, \mathrm{vol}_1(\mathrm{d}\theta), \quad (\mathsf{PU})$$

où Z_{pr} est la constante de normalisation faisant de cette mesure une loi de probabilité. (Nous savons en fait que celle-ci vaut en l'occurrence π ; mais je souhaite insister sur le fait que la détermination cette valeur est sans importance pour notre raisonnement).

En vertu de la formule de Bayes, on trouve alors que, pour $\mathrm{d}\theta$ un voisinage infinitésimal de θ :

$$\begin{aligned}\mathbb{P}(\boldsymbol{\theta} \in \mathrm{d}\theta \mid X = x_{\checkmark}) &\propto \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark}) \times \mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta) \\ &= \theta^{x_{\checkmark}} (1-\theta)^{n - x_{\checkmark}} \times Z_{\mathrm{pr}}^{-1} \theta^{-1/2} (1-\theta)^{-1/2}\, \mathrm{vol}_1(\mathrm{d}\theta) \\ &\propto \theta^{x_{\checkmark} - 1/2} (1-\theta)^{n - x_{\checkmark} - 1/2}\, \mathrm{vol}_1(\mathrm{d}\theta) = \theta^{7\frac{1}{2}} (1-\theta)^{16\frac{1}{2}}\, \mathrm{vol}_1(\mathrm{d}\theta).\end{aligned}$$

Nous avons ainsi déterminé la postériore pour $\boldsymbol{\theta}$... mais seulement à constante multiplicative près. Cependant, si le but est uniquement d'*identifier* la postériore, il est inutile d'aller plus loin : en effet, cette constante est forcément égale à l'unique valeur qui donne à la postériore une masse totale égale à 1 ! Or, en l'occurrence, un

statisticien expert connait justement une loi classique, appelée la loi Bêta($8\frac{1}{2}, 17\frac{1}{2}$), qui est une loi portée par $]0,1[$ et telle que

$$\mathbb{P}\big(\text{Bêta}(8\frac{1}{2}, 17\frac{1}{2}) \in \mathrm{d}x\big) \propto x^{7\frac{1}{2}}(1-x)^{16\frac{1}{2}}\,\mathrm{vol}_1(\mathrm{d}x)\,: \tag{PV}$$

la postériore coïncidant avec la loi Bêta($8\frac{1}{2}, 17\frac{1}{2}$) à constante multiplicative près, elle y est nécessairement *égale*, étant donné que postériore comme loi bêta sont des lois de probabilité, et qu'elles doivent donc avoir la même masse totale (à savoir, 1).

Noter par ailleurs que, si on avait quand même besoin de déterminer la constante de proportionnalité pour la postériore (cela arrive parfois), le calcul se ferait simplement en écrivant qu'on soit satisfaire la condition de normalisation d'une mesure de probabilité : on a donc

$$\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta) = Z_{\text{post}}^{-1}\theta^{7\frac{1}{2}}(1-\theta)^{16\frac{1}{2}}\,\mathrm{vol}_1(\mathrm{d}\theta) \tag{PW}$$

avec

$$Z_{\text{post}} = \int_{\theta\in\Theta} \theta^{7\frac{1}{2}}(1-\theta)^{16\frac{1}{2}}\,\mathrm{vol}_1(\mathrm{d}\theta) = \int_{\theta=0}^{1} \theta^{7\frac{1}{2}}(1-\theta)^{16\frac{1}{2}}\mathrm{d}\theta = 7{,}748\times 10^{-8}. \tag{PX}$$

Notez en particulier qu'on n'a pas besoin, pour faire ce calcul, de connaitre la valeur Z_{pr} de la constante de normalisation de la priore ! ☺

Nous avons donc déterminé la loi à postériori de la fiabilité du chasseur. Concrètement, cela a (notamment) l'interprétation suivante : si on définit un « mauvais chasseur » comme un chasseur dont la fiabilité est inférieure à 1/4, alors, *au vu du résultat de son test*, le club estime que la probabilité que son candidat soit un mauvais chasseur vaut $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} < 1/4) = \mathbb{P}(\text{Bêta}(8\frac{1}{2}, 17\frac{1}{2}) < 1/4) = 20{,}6\ \%$; alors que, si on avait demandé au club ce qu'il pensait du candidat *avant* le test, il lui aurait donné une probabilité d'être un mauvais chasseur égale à $\mathbb{P}_{\text{pr}}(\boldsymbol{\theta} < 1/4) = \mathbb{P}(\text{LoiArcsin} < 1/4) = 33{,}3\ \%$. Le test a donc un peu rassuré le club quant au risque que son candidat soit un mauvais chasseur, mais il demeure un doute substantiel !

Loi à postériori d'une quantité d'intérêt explicative pour le chasseur

Si la club ne n'intéresse pas à $\boldsymbol{\theta}$ elle-même, mais à la quantité d'intérêt explicative $\boldsymbol{\tau} \coloneqq 2\boldsymbol{\theta} - \boldsymbol{\theta}^2 =: \boldsymbol{\varphi}(\boldsymbol{\theta})$ (laquelle quantité correspond à la probabilité de toucher un plateau quand on a droit à *deux* essais), il va tout simplement dire que la loi (à postériori) de $\boldsymbol{\tau}$ est la *mesure-image* de celle de $\boldsymbol{\theta}$ par l'application $\boldsymbol{\varphi}(\bullet)$: en effet, pour $\tau \in]0,1[$ et $\mathrm{d}\tau$ un voisinage infinitésimal de τ, dire que $\boldsymbol{\tau}$ (qui est $\boldsymbol{\varphi}(\boldsymbol{\theta})$) est dans $\mathrm{d}\tau$, c'est la même chose que dire que $\boldsymbol{\theta}$ est dans l'image réciproque $\boldsymbol{\varphi}^{-1}(\mathrm{d}\tau)$!

En l'occurrence, l'application $\boldsymbol{\varphi}(\bullet)$ étant bijective de $]0,1[$ dans $]0,1[$, le calcul va correspondre à un simple changement de variable. Plus précisément, par bijectivité, l'image réciproque de $\mathrm{d}\tau$ par $\boldsymbol{\varphi}(\bullet)$ est aussi son image directe par la bijection réciproque $\boldsymbol{\varphi}^{-1}(\bullet)$, dont on détermine facilement la formule :

$$\boldsymbol{\varphi}^{-1}(\tau) = 1 - (1-\tau)^{1/2}. \tag{PY}$$

Puisque $\mathrm{d}\tau$ est un voisinage infinitésimal de τ, $\boldsymbol{\varphi}^{-1}(\mathrm{d}\tau)$ est un voisinage infinitésimal de $\mathrm{d}\tau$. En outre, en vertu de la formule de changement de variable, la longueur de $\boldsymbol{\varphi}^{-1}(\mathrm{d}\tau))$ est $|[\boldsymbol{\varphi}^{-1}]'(\tau)|$ fois plus grande que celle de $\mathrm{d}\tau$: d'où, après calcul de dérivée,

$$\mathrm{vol}_1(\boldsymbol{\varphi}^{-1}(\mathrm{d}\tau)) = \frac{1}{2(1-\tau)^{1/2}}\,\mathrm{vol}_1(\mathrm{d}\tau)\,{}^{[\dagger]}. \tag{PZ}$$

L'un dans l'autre, on a donc

$$\mathbb{P}_{\text{post.}}(\boldsymbol{\tau} \in \mathrm{d}\tau) = \mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \varphi^{-1}(\mathrm{d}\tau)) = Z_{\text{post}}^{-1}\varphi^{-1}(\tau)^{7\frac{1}{2}}(1-\varphi^{-1}(\tau))^{16\frac{1}{2}}\,\mathrm{vol}_1(\varphi^{-1}(\mathrm{d}\tau))$$
$$= \frac{1}{2Z_{\text{post}}}(1-\tau)^{7\frac{3}{4}}\big(1-(1-\tau)^{1/2}\big)^{7\frac{1}{2}}\,\mathrm{vol}_1(\mathrm{d}\tau)\,: \quad (\mathsf{QA})$$

on a donc déterminé l'expression explicite (à constante de normalisation près) de la loi à postériori de $\boldsymbol{\tau}$!

Loi à postériori de la quantité d'intérêt prédictive pour le chasseur

Maintenant, voyons comment on peut passer de la postériore (concernant le paramètre caché lui-même) à la loi à postériori de la quantité d'intérêt prédictive. Nous utilisons le raisonnement présenté dans le point (OX) : pour $y \in [\![0, m]\!]$, on a par formule des probabilités totales sur θ, puis par rgèle de chaine, que

$$\mathbb{P}_{\text{post}}(\boldsymbol{Y} = y) = \int_{\theta\in\Theta} \mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta \text{ et } \boldsymbol{Y} = y)$$
$$= \int_{\theta\in\Theta} \mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta)\,\mathbb{P}_{\text{post}}(\boldsymbol{Y} = y \mid \boldsymbol{\theta} = \theta)\,; \quad (\mathsf{QB})$$

puis, dans l'expression « $\mathbb{P}_{\text{post}}(\boldsymbol{Y} = y \mid \boldsymbol{\theta} = \theta)$ », qui fait intervenir à la fois le conditionnement par $\{X = x_{\checkmark}\}$ (en tant que contexte probabiliste) et le conditionnement par $\{\boldsymbol{\theta} = \theta\}$ (en tant que contexte explicite), on échange les rôles que ces deux conditionnement jouent, obtenant ainsi que cette expression peut aussi s'écrire $\mathbb{P}_\theta(\boldsymbol{Y} = y \mid X = x_{\checkmark})$: ce qui, pour ce modèle spécifique où X et Y sont indépendantes sous tous les contextes $\mathbb{P}_\theta(\bullet)$, se simplifie encore en $\mathbb{P}_\theta(\boldsymbol{Y} = y)$.

On utilise alors les formules pour resp. la postériore et la loi de l'observation sachant le paramètre caché pour obtenir que

$$\mathbb{P}_{\text{post}}(\boldsymbol{Y} = y) = \int_{\theta\in\Theta} \mathbb{P}_\theta(\boldsymbol{Y} = y)\,\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta)$$
$$\propto \int_{\theta\in]0,1[} \mathbb{P}\big(\mathrm{Binom}^{\text{le}}(m,\theta) = y\big)\theta^{7\frac{1}{2}}(1-\theta)^{16\frac{1}{2}}\,\mathrm{vol}_1(\mathrm{d}\theta), \quad (\mathsf{QC})$$

et on utilise enfin que

$$\mathbb{P}\big(\mathrm{Binom}^{\text{le}}(m,\theta) = y\big) = \binom{m}{y}\theta^y(1-\theta)^{m-y} = \frac{75!}{y!(75-y)!}\theta^y(1-\theta)^{75-y} \quad (\mathsf{QD})$$

pour arriver à l'expression de la loi à postériori pour y, à constante de normalisation près :

$$\mathbb{P}_{\text{post}}(\boldsymbol{Y} = y) \propto \frac{75!}{y!(75-y)!}\int_{\theta=0}^{1} \theta^{y+7\frac{1}{2}}(1-\theta)^{91\frac{1}{2}-y}\,\mathrm{vol}_1(\mathrm{d}\theta)$$
$$= \frac{75!}{y!(75-y)!}\,\frac{(y+7\frac{1}{2})!(91\frac{1}{2}-y)!}{100!} \propto \frac{(y+7\frac{1}{2})!}{y!}\,\frac{(91\frac{1}{2}-y)!}{(75-y)!}, \quad (\mathsf{QE})$$

[†]. En l'occurrence la valeur absolue n'apparait pas, car la dérivée est toujours positive.

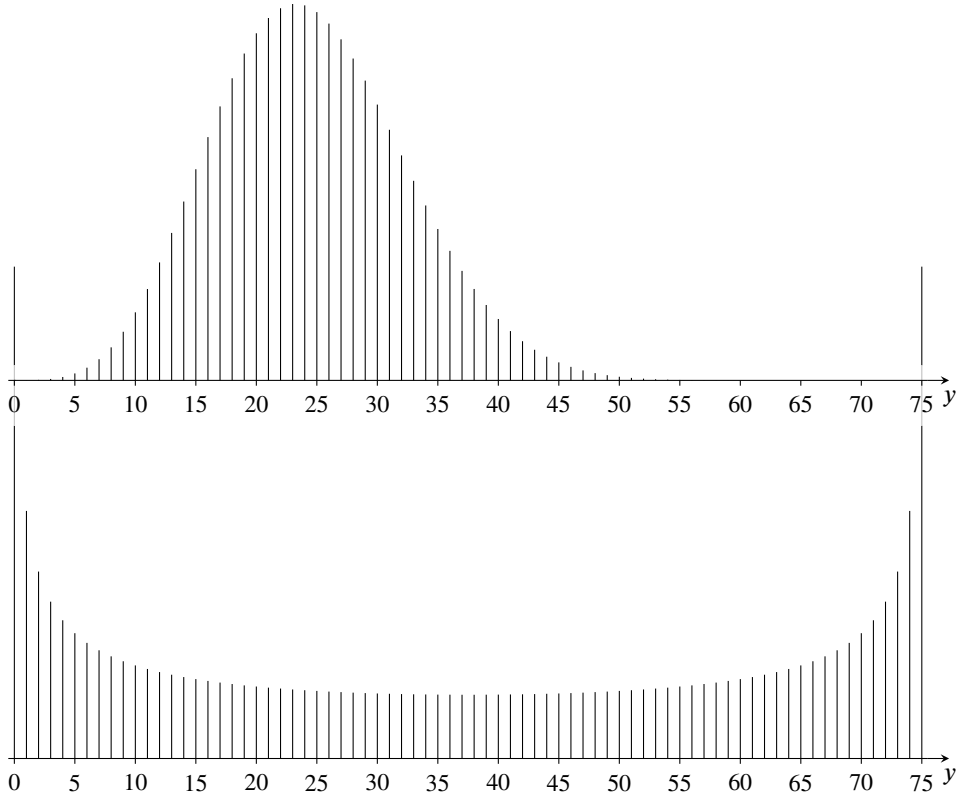

FIGURE 8.1 – En haut, la loi à postériori de Y dans le modèle du chasseur. En bas, à titre de comparaison, la loi à priori de Y, tracée à la même échelle.

où on a utilisé la formule de l'« intégrale bêta » (que vous n'êtes pas supposés connaitre) pour l'égalité centrale [‡].

On peut enfin récupérer la constante de proportionnalité des cette loi en écrivant la condition de normalisation : ce qui donne finalement que

$$\mathbb{P}_{\text{post}}(Y = y) = Z_Y^{-1} \frac{(y + 7½)!}{y!} \frac{(91½ - y)!}{(75 - y)!}, \tag{QF}$$

avec

$$Z_Y = \sum_{y=0}^{75} \frac{(y + 7½)!}{y!} \frac{(91½ - y)!}{(75 - y)!}. \tag{QG}$$

Cette loi peut se déterminer numériquement : vous en trouverez le tracé en figure 8.1, accompagné du tracé de la loi à priori de Y à titre de comparaison. On voit que, alors qu'à priori le club n'avait pour ainsi dire aucune idée du score que son candidat ferait en compétition, grâce au résultat du test, le club considère à présent qu'il peut être quasi-certain (à 97,8 % de probabilité, plus précisément) que le candidat atteindra entre 8 et 42 cibles à la compétition !

Calcul de postériore pour le pédagogue

Comme dans l'exemple précédent, nous commençons par calculer la vraisemblance pour l'observation effective $(x_{0\checkmark}, \dots, x_{(n-1)\checkmark})$. Comme l'observation $(X_{1\checkmark}, \dots,$

[‡]. La formule fait intervenir des factorielles d'arguments non entiers : mais les ordinateurs savent bien calculer ces dernières.

$X_{n\checkmark}$) se décompose en n sous-observations indépendantes, nous pouvons appliquer le théorème (UD) :

$$\mathscr{L}\big(\boldsymbol{\theta} = (\mu, \sigma) \;\big|\; X = (x_{0\checkmark}, \dots, x_{(n-1)\checkmark})\big) = \prod_{i=0}^{n-1} \mathscr{L}\big(\boldsymbol{\theta} = (\mu, \sigma) \;\big|\; X_i = x_{i\checkmark}\big), \tag{QH}$$

où X_i désigne le score obtenu par l'élève numéro i. Maintenant, quelle est la vraisemblance $\mathscr{L}(\boldsymbol{\theta} \mid X_i = x_{i\checkmark})$ du paramètre caché pour la "sous-expérience" consistant à regarder seulement le score de l'élève i ? On remarque que, sachant que $\boldsymbol{\theta}$ vaut (μ, σ), la loi de X_i est la loi Normale(μ, σ^2), caractérisée par

$$\mathbb{P}\big(X_i \in \mathrm{d}x \;\big|\; \boldsymbol{\theta} = (\mu, \sigma)\big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{vol}_1(\mathrm{d}x) : \tag{QI}$$

vu qu'il s'agit d'une loi à densité sur $\mathbb{R}$, on peut appliquer le théorème (UA) pour en déduire que

$$\mathscr{L}\big(\boldsymbol{\theta} = (\mu, \sigma) \;\big|\; X_i = x_{i\checkmark}\big) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_{i\checkmark}-\mu)^2}{2\sigma^2}\right), \tag{QJ}$$

expression dans laquelle nous nous empresserons évidemment de simplifier le facteur constant (par rapport à θ) $1/\sqrt{2\pi}$. En appliquant le théorème (UD), on obtient alors :

$$\mathscr{L}\big(\boldsymbol{\theta} = (\mu, \sigma) \;\big|\; X = (x_{0\checkmark}, \dots, x_{(n-1)\checkmark})\big) = \sigma^{-n} \exp\left(-\frac{\sum_{i=0}^{n-1}(x_i-\mu)^2}{2\sigma^2}\right). \tag{QK}$$

Il s'avèrera commode de ré-organiser encore cette expression, en particulier la quantité intervenant dans l'exponentielle. Pour cela, nous introduisons la *moyenne empirique* des X_i, soit $M \coloneqq \big(\sum_{i=0}^{n-1} X_i\big) / n$, ainsi que leur *écart-type empirique*, soit $S \coloneqq \big(\sum_{i=0}^{n-1}(X_i - M)^2 / n\big)^{1/2}$ (statistiques dont, suivant nos conventions, les réalisations seront notées resp. $m_\checkmark$ et $s_\checkmark$) : on vérifie aisément que

$$\sum_{i=0}^{n-1}(X_i-\mu)^2 = n\big((M-\mu)^2 + S^2\big), \tag{QL}$$

d'où :

$$\mathscr{L}\big(\boldsymbol{\theta} = (\mu, \sigma) \;\big|\; X = (x_{0\checkmark}, \dots, x_{(n-1)\checkmark})\big) = \sigma^{-n} \exp\big(-\tfrac{1}{2} n\sigma^{-2}\big((m_\checkmark - \mu)^2 + s_\checkmark^2\big)\big). \tag{QM}$$

Au vu de la formule pour notre priore (confer p. **??**), la formule de Bayes nous donne finalement que

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{\theta} \in \mathrm{d}\mu \times \mathrm{d}\sigma \mid X = x_\checkmark) \propto \\
&\quad \sigma^{-(n+3)} \exp\left(-\frac{n(\mu - m_\checkmark)^2 + n s_\checkmark^2 + \sigma_{\text{réf}}^2}{2\sigma^2} - \frac{(\mu - \mu_{\text{réf}} + \sigma_{\text{réf}})^2}{2\sigma_{\text{réf}}^2}\right) \mathrm{vol}_1(\mathrm{d}\mu)\,\mathrm{vol}_1(\mathrm{d}\sigma) \\
&\qquad = \sigma^{-25} \exp\left(-\frac{11 \times (\mu - 70{,}3)^2 + 2{,}77 \times 10^3}{\sigma^2} - \frac{(\mu-64)^2}{242}\right) \mathrm{vol}_2(\mathrm{d}\mu \times \mathrm{d}\sigma).
\end{aligned}$$

On notera que l'expression de cette postériore est assez compliquée ; en particulier, $\boldsymbol{\mu}$ et $\boldsymbol{\sigma}$ ne sont plus indépendantes pour la loi à postériori.

Le figure 8.2 donne l'allure de cette postériore, obtenue numériquement à partir de la formule ci-dessus. On remarque que l'observation des notes de la première promotion a permis à l'enseignant d'affiner considérablement sa connaissance du paramètre caché : à présent, il est en mesure de dire que $\mu_\checkmark$ se situe probablement entre 65 et 75, et que $\sigma_\checkmark$ se situe probablement entre 12 et 20 !

Loi à postériori de $\boldsymbol{\mu}$ pour le pédagogue

Si ce qui intéresse spécifiquement notre enseignant est de connaitre la valeur de la quantité d'intérêt $\boldsymbol{\mu}$, il peut obtenir la loi à postériori de $\boldsymbol{\mu}$ à partir de la loi à postériori de $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ par mesure-image : ce qui, en l'occurrence, revient simplement à intégrer selon les différentes possibilités pour σ :

$$\begin{aligned}\mathbb{P}_{\text{post}}(\boldsymbol{\mu} \in \mathrm{d}\mu) = \\ \int_{\sigma\in\mathbb{R}_+^*} \mathbb{P}_{\text{post}}(\boldsymbol{\mu} \in \mathrm{d}\mu \text{ et } \boldsymbol{\sigma} \in \mathrm{d}\sigma) \propto \exp\left(-\frac{(\mu-\mu_{\text{réf}}+\sigma_{\text{réf}})^2}{2\sigma_{\text{réf}}^2}\right) \mathrm{vol}_1(\mathrm{d}\mu) \times \\ \int_{\sigma\in\mathbb{R}_+^*} \sigma^{-(n+3)} \exp\left(-\frac{n(\mu-m_{\checkmark})^2 + n s_{\checkmark}^2 + \sigma_{\text{réf}}^2}{2\sigma^2}\right) \mathrm{vol}_1(\mathrm{d}\sigma) \quad (\mathtt{QN})\end{aligned}$$

Ici on utilise que, pour $k > 1$, pour $a \in \mathbb{R}_+^*$, on a par la changement de variable $y \leftarrow a^{-1/2}x$ que

$$\int_{x=0}^{\infty} x^{-k} \mathrm{e}^{-a/x^2} \mathrm{d}x = \int_{y=0}^{\infty} (a^{1/2}y)^{-k} \mathrm{e}^{-1/y^2} a^{1/2} \mathrm{d}y \propto a^{-(k-1)/2} \quad (\mathtt{QO})$$

(la proportionnalité s'entendant à valeur de k fixée) pour en déduire que

$$\mathbb{P}_{\text{post}}(\boldsymbol{\mu} \in \mathrm{d}\mu) \propto \exp\left(-\frac{(\mu-\mu_{\text{réf}}+\sigma_{\text{réf}})^2}{2\sigma_{\text{réf}}^2}\right) \times \frac{1}{\left(n(\mu-m_{\checkmark})^2 + n s_{\checkmark}^2 + \sigma_{\text{réf}}^2\right)^{n/2+1}} \mathrm{vol}_1(\mathrm{d}\mu). \quad (\mathtt{QP})$$

Variante avec la priore de Haar

Si, au lieu de prendre la priore par défaut pour notre modèle de pédagogue, nous avions pris la priore de Haar (cf. exemple (PQ)), tout se serait passé de la même manière (à condition de ne jamais chercher à expliciter les constantes de normalisation !), à ceci près que nous aurions dû remplacer le facteur « $\sigma^{-3}\exp(-(\mu-\mu_{\text{réf}}+\sigma_{\text{réf}})^2/2\sigma_{\text{réf}}^2 - \sigma_{\text{réf}}^2/2\sigma^2)$ » qui apparait dans la priore par défaut par le facteur « σ^{-2} » qui en est l'analogue pour la priore de Haar : on serait alors arrivé à

$$\begin{aligned}\mathbb{P}(\boldsymbol{\theta} \in \mathrm{d}\mu \times \mathrm{d}\sigma \mid X = x_{\checkmark}) &\propto \sigma^{-(n+2)} \exp\left(-\frac{n(\mu-m_{\checkmark})^2 + n s_{\checkmark}^2}{2\sigma^2}\right) \mathrm{vol}_1(\mathrm{d}\mu)\,\mathrm{vol}_1(\mathrm{d}\sigma) \\ &= \sigma^{-24} \exp\left(-\frac{11\times(\mu-70{,}3)^2 + 2{,}54\times 10^3}{\sigma^2}\right) \mathrm{vol}_2(\mathrm{d}\mu\times\mathrm{d}\sigma)\,; \quad (\mathtt{QQ})\end{aligned}$$

et si nous avions voulu nous intéresser plus spécifiquement à la quantité d'intérêt $\boldsymbol{\mu}$, le même argument dans le cas de la priore par défaut nous aurait permis d'obtenir que, cette fois-ci,

$$\mathbb{P}_{\text{post}}(\boldsymbol{\mu} \in \mathrm{d}\mu) \propto \frac{\mathrm{vol}_1(\mathrm{d}\mu)}{\left((\mu-m_{\checkmark})^2 + s_{\checkmark}^2\right)^{(n+1)/2}}. \quad (\mathtt{QR})$$

On observe ainsi que le résultat de l'analyse dépend du choix de la priore ! Voir une illustration de ce phénomène en figure 8.3.

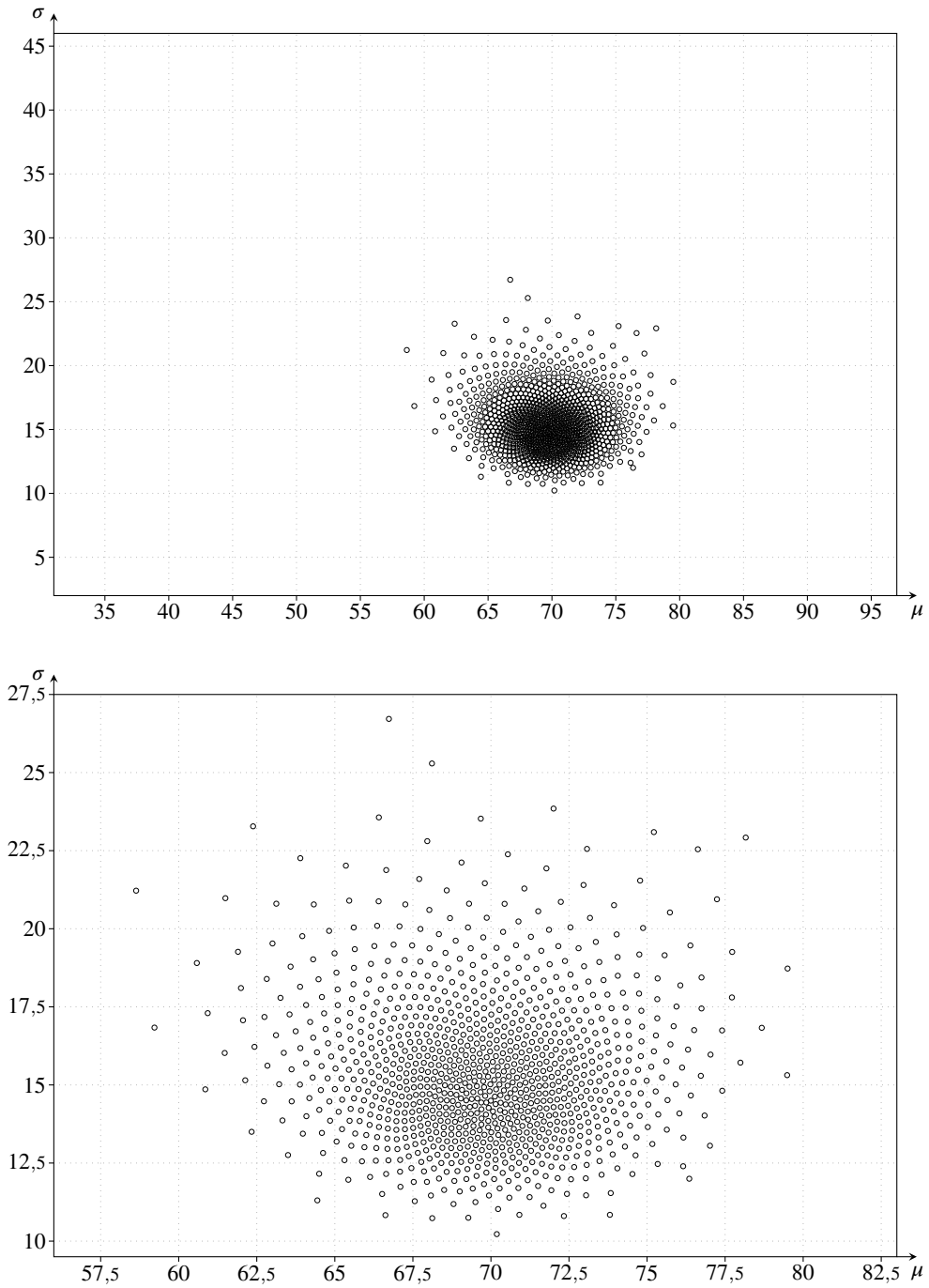

FIGURE 8.2 – Postériore pour le modèle du pédagogue, avec un petit cercle par zone de probabilité 1 ‰ : en haut, avec la même échelle que pour la priore ; en bas, zoom sur la zone la plus chargée.

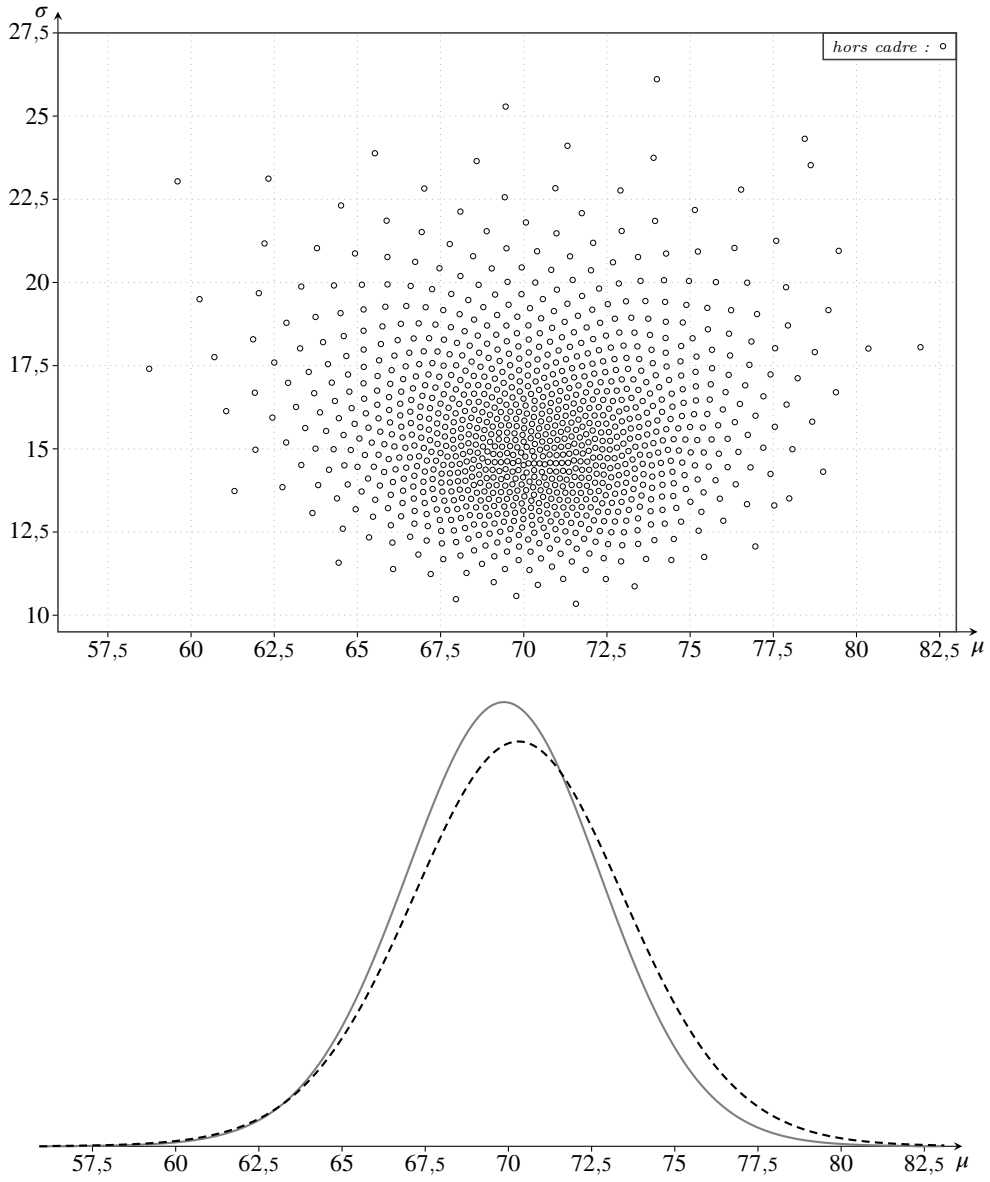

FIGURE 8.3 – En haut : Postériore pour le modèle du pédagogue avec priore de Haar, à la même échelle qu'en figure 8.2. En bas : Densité de la loi à postériori de $\boldsymbol{\mu}$ dans les cas respectifs de la priore de référence (trait plein gris) et de la priore de Haar (trait pointillé noir).

Chapitre 9

Méthodes statistiques bayésiennes

9.1 Importance des lois à postériori

À l'issue du chapitre précédent, nous avons appris à exploiter le théorème de Bayes pour, une certaine priore étant donnée, déduire de notre observation la postériore, puis, par ricochet, les *lois à postériori* de quantités d'intérêt $\gamma(\boldsymbol{\theta})$ (cas explicatif) ou $g(\boldsymbol{Y})$ (cas prédictif). Il est important d'observer que lois à postériori sont des *statistiques*, et par conséquent, constituent des réponses techniquement acceptables pour une analyse statistique, de même que toute quantité s'en déduisant :

Remarque (QS). La postériore ne dépendant que [*] de l'observation, et pas du paramètre caché, il s'agit de la réalisation d'une statistique. Du point de vue technique, la statistique en question peut s'écrire comme l'image de la v.a. $\boldsymbol{X}$ par la fonction déterministe [†]

$$\begin{aligned} \mathscr{X} &\to \mathscr{M}_1(\Theta) \\ x &\mapsto \mathrm{Loi}(\boldsymbol{\theta} \mid \boldsymbol{X} = x) : \end{aligned} \tag{QT}$$

Il s'agit ainsi d'une variable aléatoire (en tant que fonction de $\boldsymbol{X}$) à valeurs dans les *lois de probabilité* sur $\boldsymbol{\Omega}$: qu'une variable aléatoire soit à valeurs « lois » est certes un peu paradoxal, avec un aspect "méta" ; mais cela ne pose aucune, objection du point de vue technique ! ⌒

Corolairement, dans la mesure où nous avons vu que la loi à postériori d'une quantité d'intérêt $\gamma(\boldsymbol{\theta})$ ou $g(\boldsymbol{Y})$ pouvait se calculer directement à partir de la postériore (et, éventuellement, de la valeur de l'observation passée), ces lois à postériori sont elles aussi des statistiques. ♧

En fait, comme nous le verrons dans la section suivante, non seulement la loi à postériori de $g(\boldsymbol{Y})$ (mettons ici qu'on est dans un contexte prédictif) est une statistique, mais elle est, dans un sens, "exhaustive", en ce qu'elle décrit l'*intégralité* de l'information (et de l'incertitude) dont nous disposons sur $g(\boldsymbol{Y})$ à l'issue de notre analyse ! C'est donc un jalon absolument fondamental dans l'exploitation de l'observation dont nous disposons.

Cependant, du point de vue pratique, la loi à postériori d'une quantité d'intérêt comporte deux écueils :

[*]. Outre, s'entend, du modèle statistique et de la priore choisie.

[†]. Regardez attentivement la formule : même si le symbole '$\boldsymbol{X}$' intervient, c'est bien une fonction *déterministe* qui est décrite !

— Tout d'abord, elle ne dit dit pas directement *quoi faire* : car en fin de compte, notre but ne sera pas de déterminer une loi, mais de prendre une décision... !
— D'autre part, à supposer que notre but soit simplement de transmettre l'information obtenue sur la quantité d'intérêt au commanditaire de notre étude, l'objet « loi de probabilité à postériori », qui est à valeurs dans l'espace de dimension (en général) infinie $\mathscr{M}_1(\mathscr{G})$, est particulièrement compliqué à décrire complètement, et sa description complète risque donc d'être particulièrement compliquée à appréhender par lesdits commanditaires... Il peut donc s'avérer utile de "résumer" cette distribution à postériori par simplement une poignée de nombres réels.

La suite du présent chapitre vise à traiter ces deux problématiques. La § 9.2 s'attaquera à la question très générale de « quoi faire » ; tandis que les § **??** à **??** viseront plus modestement à résumer la loi à postériori en quelques nombres-clés.

9.2 Décision optimale en contexte bayésien

Loi à postériori et décision optimale

Dans cette section, nous nous intéresserons uniquement au cas d'une quantité d'intérêt *prédictive* $g(Y) =: G$: ce cas permet en effet de traiter aussi le cas explicatif, quitte à adjoindre à l'observation future une copie du paramètre caché.

Nous supposons choisie une certaine fonction de perte $\ell(\bullet, \bullet)$ pour notre problème, le premier argument de cette fonction de perte se référant à une quantité d'intérêt prédictive $g(Y) =: G$, et le second à une décision à prendre au sein d'un certain espace $\mathscr{D}$. Cette décision ne pouvant être prise qu'en fonction des informations dont nous disposons déjà au moment de l'analyse, cela devra être une fonction de l'observation, autrement dit, une *statistique* : nous la noterons donc $\hat{d}(X)$.

Notre but est alors de choisir une $\hat{d}(X) =: \hat{D}$ de façon à minimiser l'espérance de notre perte (confer § 3.5) :

$$\text{On cherche } \hat{d} : \mathscr{X} \to \mathscr{D} \text{ minimisant } \mathbb{E}\big(\ell(g(Y), \hat{d}(X))\big). \qquad (\mathrm{Opt}_{\mathbb{E}})$$

Le point essentiel est alors le suivant :

!! **Principe** (QU) (Décision optimale en contexte bayésien, principe général)**.** *La réalisation* $\hat{\hat{d}}(x_{\checkmark})$ *de la décision optimale se trouve en résolvant le problème* probabiliste *où on doit prendre une décision déterministe dans un contexte où* G *suit la loi* $\mathrm{Loi}_{\mathrm{post}}(G)$ *! Un énoncé plus formel est donné ci-après.* ◇

Dit en termes plus formels :

Théorème (QV) (Décision optimale en contexte bayésien, version formelle)**.** *Pour tout* $x \in \mathscr{X}$*, notons*

$$P_x := \mathrm{Loi}(g(Y) \mid X = x), \qquad (\mathtt{QW})$$

et notons $\mathscr{S}_x$ *l'ensemble des solutions du problème d'optimisation « particularisé » suivant :*

$$\text{On cherche } \hat{d} \in \mathscr{D} \text{ maximisant } \mathbb{E}^{G\sim P_x}\big(u(G, \hat{d})\big). \qquad (\mathrm{Opt}_x)$$

Alors[‡] *une application* $\hat{d} : \mathscr{X} \to \mathscr{D}$ *constitue une solution optimale au problème* $(\mathrm{Opt}_{\mathbb{E}})$ *si et seulement si on a* $\hat{d}(x) \in \mathscr{S}_x$ *pour (presque-)tout* x*.* ◇

! *Remarque* (QX). Le principe (QU) montre donc que, dans le paradigme bayésien, seule la *réalisation* de la postériore importe pour prendre une décision optimale : c'est extrêmement commode ! ⊍ En statistique fréquentiste, par contre, cette propriété ne sera plus valable... ⁚́ ♧

Remarque (QY). En général, dans la pratique, le but n'est pas tant de déterminer l'optimum exact (à supposer qu'il existe) que de s'approche très près de l'optimum (ce qui est toujours possible, quand bien même l'optimum lui-même n'est pas atteignable). Mais cela ne change rien au principe (QU) : il faut juste remplacer l'optimisation exacte par une optimisation *approchée* ! ⌒ ♧

Remarque (QZ). Nous comprenons donc pourquoi, dans la section précédente, nous avions dit que la loi à postériori de $g(Y)$ fournissait *l'intégralité* de l'information dont nous disposons sur la quantité d'intérêt au vu de l'observation : de fait, il suffit de connaitre cette loi à postériori pour prendre optimalement n'importe quelle décision concernant $g(Y)$! ♧

Notions de risques

La thématique de la prise de décision en contexte bayésien est l'occasion d'introduire le vocabulaire lié aux différentes notions de *risque*. En statistique, « risque » singifie simplement « espérance de la perte » :

! **Définition** (RA) (Marco-définition du « rsique »)**.** En statistique, une quantité d'intérêt $g(Y)$ et une stratégie de décision $d(X)$ étant données, ainsi qu'une fonction de perte $\ell(\bullet, \bullet)$ relative à l'inadéquation entre telle valeur de la quantité d'intérêt et telle possibilité de décision, on parle de *risque* associé à cette stratégie de décision pour désigner, de manière générale, « l'espérance de la perte $\ell(g(Y), \hat{d}(X))$ ». ♡

Dans la mesure où, en statistique, la notion d'« espérance » se rapport en fait à un certain *contexte probabiliste*, à chaque contexte probabiliste de trouve associé une notion de « risque » différente :

! **Définition** (RB)**.** Une stratégie de décision $\hat{d}(X) =: \hat{D}$ étant donnée, dans un contexte associé à une quantité d'intérêt $g(Y)$ et une fonction de perte $(g, \hat{d}) \mapsto \ell(g, \hat{d})$, on appelle *risque intégré* de cette stratégie, qu'on pourra noter par exemple $R^{\text{int}}_{\hat{D}}$, l'espérance *à priori* de la perte engendrée par cette stratégie :

$$R^{\text{int}}_{\hat{D}} := \mathbb{E}_{\text{pr}}\big(\ell(g(Y), \hat{d}(X))\big). \tag{RC}$$

♡

! **Définition** (RD)**.** Dans le même cadre, la *fonction de risque bayésien*[§] associée à la stratégie $\hat{d}(X)$, qu'on pourra noter par exemple $R^{\text{bay}}_{\hat{D}}(\bullet)$, est l'application qui, à une valeur donnée pour l'observation passée, associe la perte moyenne *conditionnellement à cette valeur de l'observation* :

$$R^{\text{bay}}_{\hat{D}}(x) := \mathbb{E}\big(\ell(g(Y), \hat{d}(x)) \;\big|\; X = x\big). \tag{RE}$$

♡

Définition (RF). Toujours dans le même cadre, le *risque à postériori*, qu'on pourra noter par exemple $R^{\text{post}}_{\hat{D}}$, et l'espérance *à postériori* $\mathbb{E}_{\text{post}}\big(\ell(g(Y), \hat{d}(x_{\checkmark}))\big)$ de la perte engendrée par la stratégie de décision ; en d'autre termes, c'est la valeur $R^{\text{bay}}_{\hat{D}}(x_{\checkmark})$ de la fonction de risque bayésien en l'observation effective. ♡ !

Avec le vocabulaire ci-dessus, le début de cette section peut se reformuler en disant que pour minimiser le risque intégrer, il faut chercher à minimiser la fonction de risque bayésien, ce qui peut se faire « x par x » : en particulier, trouver la décision optimale pour la valeur effective $x_{\checkmark}$ de l'observation revient à minimiser le risque à postériori.

Dans le cadre de la statistique bayésienne, nous pourrions nous arrêter ici ; cependant, tant que nous y sommes, introduisons aussi le vocabulaire du risque fréquentiste, dont nous nous re-servirons dans le chapitre 11 :

! **Définition (RG).** Toujours dans le même cadre que pour les définitions (RB) à (RF), la *fonction de risque (fréquentiste)*, qu'on pourra noter par exemple $R^{\text{fréq}}_{\hat{D}}(\bullet)$, est l'application qui, à une valeur donnée pour le paramètre caché, associe la perte moyenne *lorsque le paramètre caché prend cette valeur* :

$$R^{\text{fréq}}_{\hat{D}}(\theta) \coloneqq \mathbb{E}_{\theta}\big(\ell(g(Y), \hat{d}(x))\big). \tag{RH}$$

On parle parfois aussi de *risque véritable*, qu'on peut noter $R^{\checkmark}_{\hat{D}}$, pour parler de la valeur $R^{\text{fréq}}_{\hat{D}}(\theta_{\checkmark})$ prise par la fonction de risque fréquentiste en la vraie valeur du paramètre caché : il s'agit alors de l'espérance de la perte sous la loi véritable. ♡

Remarque (RI). Conformément à ce que son nom suggère, la fonction de risque fréquentiste est un concept fréquentiste : elle ne dépend que du modèle statistique lui-même (plus, bien sûr, de la fonction de perte choisie, et de la décision considérée), pas de la priore ! ♧

Remarque (RJ). Noter que les fonctions de risques resp. bayésien et fréquentiste n'ont pas la même domaine : l'argument de la première vit dans l'espace de l'observation, tandis que l'argument de la seconde vit dans l'espace du paramètre caché ! ♧

On peut enfin mentionner le lien suivant entre fonction de risque fréquentiste et risque intégré :

Théorème (RK). *Toujours avec les mêmes notation que ci-dessus, on a la relation :*

$$R^{\text{int}}_{\hat{D}} = \int_{\theta\in\Theta} R^{\text{fréq}}_{\hat{D}}(\theta)\, \mathbb{P}_{\text{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta). \tag{RL}$$

♢

On déduit de ce théorème que « trouver la stratégie de décision qui minimise le risque intégré », c'est la même chose que « trouver la stratégie de décision qui minimise la fonction de risque fréquentiste pour tous les θ à la fois ».

[‡]. Sous réserve de considérations techniques d'intégrabilité que je passe sous silence ici.

[§]. Bien préciser « bayésien » !

9.3 Probabilités à postériori

Enjeu de résumer les lois à postériori

Dans les dernières sections de ce chapitre, nous ne nous intéressons pas à la façon de prendre une décision à partir d'une distribution de probabilité à postériori, mais plus simplement à la façon de *communiquer* efficacement ladite distribution de probabilité à postériori aux commanditaires de notre analyse statistique : ce sera à eux ensuite de prendre leurs décisions à partir de celle-ci, que ce soit en utilisant la méthode du chapitre précédent ou des techniques plus "artisanales" !

En effet, décrire une distribution de probabilité à postériori, c'est compliqué, dans la mesure où celle-ci vit généralement dans un espace de dimension infinie... En pratique, l'ingénieur désireux de transmettre le résultat de son analyse statistique optera donc en général pour une réponse synthétique résumant ce résultat en un ou deux nombres. Il existe trois principales façons de "synthétiser" ainsi un résultat, qui constitueront autant de sections de ce chapitre :

— La première idée est de dire que, pour résumer la loi à postériori de notre quantité d'intérêt, on peut tout simplement donner les probabilités à postériori que ladite quantité d'intérêt tombe dans telle ou telle région de l'espace dans lequel elle vit ! Cette idée sera explorée dans la suite de cette section-ci. À vrai dire, il s'agit là d'une notion essentiellement triviale : cependant, nous profiterons de l'étude des probabilités à postériori pour introduire le vocabulaire des *tests*, préparant ce faisant le terrain à l'étude des tests fréquentistes dans la partie III du cours.

— La seconde idée est de dire que, décrire une loi de probabilité, cela revient en première approche à dire dans quelle région de l'espace on doit s'attendre à tomber. Cela nous conduire à la notion d'*intervalle de confiance bayésien*, que nous étudierons dans la § 9.4.

— La troisième idée est d'essayer de compacter au maximum l'information sur la loi à postériori de la quantité d'intérêt en la ramenant à une valeur, censée être "la plus représentative", pour guider simplement la prise de décision : cela conduira aux notions d'*estimateur* et de *prédicteur* (en l'occurrence, dans le cadre bayésien), qui feront l'objet de la § 9.5.

Les trois façons évoquées ci-dessus de synthétiser la loi à postériori de la quantité d'intérêt ont chacune son utilité : selon le contexte, il sera plus pertinent d'utiliser l'un ou l'autre de ces trois types de résultats. Dans la suite cette section, nous allons commencer par nous focaliser sur la première de ces façons.

Probabilité à postériori d'un évènement

La façon la plus évidente de résumer en quelques nombres réels les calculs de lois à postériori auxquels conduit une analyse bayséienne est tout simplement de dire la probabilité que telle ou telle chose se passe. Ainsi, si $\{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1\}$ (resp. $\{Y \in \mathscr{Y}_1\}$) est une quantité d'intérêt explicative (resp. prédictive) à valeurs booléennes — autrement dit, que c'est un évènement ne dépendant que du paramètre caché (resp. de l'observation future) —, une information pertinente sur cette quantité d'intérêt booléenne est tout simplement sa probabilité à postériori : $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \boldsymbol{\Theta}_1) \underset{\text{déf}}{=} \mathbb{P}(\boldsymbol{\theta} \in \boldsymbol{\Theta}_1 \mid X = x_{\checkmark})$, resp. $\mathbb{P}_{\text{post}}(Y \in \mathscr{Y}_1) \underset{\text{déf}}{=} \mathbb{P}(Y \in \mathscr{Y}_1 \mid X = x_{\checkmark})$, où $x_{\checkmark}$ est la valeur effective de l'observation.

L'exemple ci-dessous montre sur quel genre de quantités d'intérêt booléennes ont peut être amené à se pencher :

Exemple (RM).

— Plaçons-nous dans le modèle du chasseur. Mettons que, aux compétitions de tir, un club de chasse fasse forte impression sur les autres clubs si lorsque son compétiteur atteint au moins 46 cibles, et qu'à l'inverse il se ridiculise lorsque son compétiteur n'atteint pas plus de 18 cibles. Si le club du Bouchonnois se demande « quelle est la probabilité, au vu de ce qu'il a montré lors de son test, que le candidat impressionne les autres clubs lors de la compétition ? (si nous le sélectionnons) », ou « quelle est la probabilité que le candidat nous ridiculise ? », cela revient à regarder les probabilités à postériori des évènements resp. $\{Y \geqslant 46\}$ et $\{Y \leqslant 18\}$, lesquels sont des quantités d'intérêt prédictives, puisque ne dépendant que de Y (et à valeurs booléennes, puisqu'étant des évènements).

— Plaçons-nous dans le modèle du pédagogue. Les évènements $\{\boldsymbol{\mu} < \mu_{\text{réf}} - \sigma_{\text{réf}}\}$, $\{\boldsymbol{\mu} \in [\mu_{\text{réf}} - \sigma_{\text{réf}}, \mu_{\text{réf}} - \sigma_{\text{réf}} / 3[\}$, $\{|\boldsymbol{\mu} - \mu_{\text{réf}}| \leqslant \sigma_{\text{réf}} / 3\}$, $\{\boldsymbol{\mu} \in]\mu_{\text{réf}} + \sigma_{\text{réf}} / 2, \mu_{\text{réf}} + \sigma_{\text{réf}}]\}$, $\{\boldsymbol{\mu} > \mu_{\text{réf}} + \sigma_{\text{réf}}\}$ sont des quantités d'intérêt (booléennes) explicatives qui expriment informellement les idées respectives que « la nouvelle méthode conduit à des résultats catastrophiques », « la nouvelle méthode conduit à des résultats nettement moins bons que l'ancienne », « ... à peu près comparables à l'ancienne », « ... nettement meilleurs que l'ancienne », ou « ... extraordinaires ». Et du point de vue pratique, il est clair que calculer les probabilités à postériori des ces cinq évènements nous donnera une information intéressante sur ce qu'on peut dire de la nouvelle méthode au vu des résultats de la première promotion.

♧

Les quantités d'intérêt booléennes de type explicatif ont reçu, dans le jargon statistique, un nom plus spécifique : on les appelle des « hypothèses ». Cette nomenclature sera fréquemment ré-utilisée dans la suite, notamment au chapitre 12. Retenons donc :

! **Définition** (RN) (Hypothèse)**.** En statistique, une *hypothèse* est une quantité d'intérêt explicative à valeurs booléennes. Ce qu'on peut reformuler de multiples façons :

— En d'autres termes, une hypothèse est un évènement $\boldsymbol{\theta}$-mesurable [¶] ;

— En d'autres termes, une hypothèse est un évènement de la forme $\{\boldsymbol{\theta} \in \boldsymbol{\Theta}_1\}$ pour un certain $\boldsymbol{\Theta}_1 \subseteq \boldsymbol{\Theta}$: pour cette raison, nous parlerons parfois par abus de langage de « l'hypothèse $\boldsymbol{\Theta}_1$ ».

— En d'autres termes, une hypothèse une proposition dont la valeur de vérité ne dépend que du paramètre caché $\boldsymbol{\theta}$.

On utilisera souvent la lettre '$\mathscr{H}$' pour désigner les hypothèses. ♡

Exemple (RO).

— Dans le modèle du chasseur, la proposition « le candidat est capable de toucher au moins un plateau sur quatre sur le long terme », autrement dit l'évènement $\{\boldsymbol{\theta} \geqslant 1/4\}$, est une hypothèse.

— Dans le modèle du pédagogue, la proposition « la dispersion des notes est la même avec la nouvelle méthode qu'avec l'ancienne », autrement dit l'évè-

[¶]. J'entends par là que c'est un évènement mesurable par rapport à la tribu $\sigma(\boldsymbol{\theta})$ engendrée par $\boldsymbol{\theta}$.

nement $\{\boldsymbol{\sigma} = \sigma_{\text{réf}}\}$ (qu'on pourrait aussi écrire « $\boldsymbol{\theta} \in \mathbb{R} \times \{\sigma_{\text{réf}}\}$ »), est une hypothèse.

— Par contre, dans le modèle du chasseur, la proposition « le chasseur atteindra au moins 19 plateaux lors de la compétition » n'est pas une hypothèse, car elle se réfère à l'observation future, et non pas au paramètre caché ! En revanche, il s'agit bien d'une quantité d'intérêt boolénne (de nature prédictive), et déterminer sa probabilité à postériori serait également tout à fait intéressant en l'occurrence !

♧

Le calcul de la probabilité à postériori d'une hypothèse, ou plus généralement d'une quantité d'intérêt booléenne, revient alors à dire quelle masse la loi à postériori donne à cette quantité d'intérêt :

Exemple (RP). Dans le modèle (explicatif) du chasseur, avec les valeurs standards ($n = 25$, priore arcsinus, $x_{\checkmark} = 8$), supposons que le club cherche à savoir si le niveau du candidat est suffisant pour toucher au moins un plateau sur quatre sur le long terme : autrement dit, on cherche à connaitre la probabilité à postériori de l'hypothèse $\{\boldsymbol{\theta} \geqslant 1/4\}$. Vu que la loi à postériori de $\boldsymbol{\theta}$ est alors une distribution $\text{Bêta}(8\frac{1}{2}, 17\frac{1}{2})$ (cf. § 8.4), cette probabilité est tout simplement $\mathbb{P}\big(\text{Bêta}(8\frac{1}{2}, 17\frac{1}{2}) \geqslant 1/4\big)$, soit $1 - \texttt{répartBêta}(8\frac{1}{2}, 17\frac{1}{2}; (1/4)-)$, ce qu'on peut calculer à l'aide par exemple du logiciel R : [‖]

```
> 1 - pbeta(1 / 4, 8 + 1 / 2, 17 + 1 / 2)
[1] 0.7944297
```

Ainsi, au vu de son examen, le club estime qu'il a 79,4 % de chances que le candidat ait le niveau pour toucher au moins un plateau sur quatre. (Notons qu'on aurait pu calculer que, à priori, la probabilité qu'il eût ce niveau avait été estimée à 66,7 %). ♧

Profitons au passage de ces considérations sur les probabilités à postériori pour attirer l'attention sur un écueil crucial de la statistique bayésienne :

Proposition (RQ). *Si la probabilité à priori d'un évènement d'intérêt donnée est nulle (resp. égale à 1), alors, quelles que soient les données, sa probabilité à postériori sera nulle (resp. égale à 1) aussi. Par conséquent, lorsqu'on souhaite étudier la probabilité à postériori d'un évènement d'intérêt, pour que notre analyse soit pertinente, il faut absolument qu'on ait donné à cet évènement une probabilité à priori non nulle !* ♢ **!!**

Démonstration. En fait, de manière générale, si $\boldsymbol{A}$ et $\boldsymbol{B}$ sont des évènements, on a $\mathbb{P}(\boldsymbol{A}) = 0 \implies \mathbb{P}(\boldsymbol{A} \mid \boldsymbol{B}) = 0$: en effet, si $\boldsymbol{A}$ est de probabilité nulle, {$\boldsymbol{A}$ et $\boldsymbol{B}$} est également de probabilité nulle en tant que sous-évènement de $\boldsymbol{A}$; et dès lors, $\mathbb{P}(\boldsymbol{A} \mid \boldsymbol{B})$, qui par définition vaut $\mathbb{P}(\boldsymbol{A}$ et $\boldsymbol{B}) / \mathbb{P}(\boldsymbol{B})$, est nulle aussi. Par passage au complémentaire, on a de même que $\mathbb{P}(\boldsymbol{A}) = 1 \implies \mathbb{P}(\boldsymbol{A} \mid \boldsymbol{B}) = 1$.

Quand on y réfléchit bien, cela est très intuitif : si on est *absolument sûr* dès le début d'une certaine hypothèse, rien de ce qui va se passer après ne peut ébranler une telle certitude absolue ! ♤

[‖]. Les lois bêta étant diffuses, le fait de prendre la fonction de répartition en version continue à gauche ou continue à droite est sans importance ici.

Remarque (RR). Dans la proposition (RQ), la formulation « il faut qu'on ait donné à cet évènement une probabilité à priori non nulle » provient de ce que, dans certains cas, on dispose d'une certaine latéralité sur le choix de la priore (confer chap. 15) : dans ce cas, il faut absolument éviter de choisir une priore qui ne permettrait pas d'étudier de façon non triviale la question qui nous intéresse ! ♧

Exemple (RS). Dans le cas du pédagogue, par exemple, la priore que nous avons prise en page 82 attribue une probabilité nulle à l'hypothèse $\{\boldsymbol{\mu} = \mu_{\text{réf}}\}$ que la nouvelle méthode soit rigoureusement aussi efficace que l'ancienne. Par conséquent, quels que soient les observations, la probabilité à postériori de l'hypothèse $\{\boldsymbol{\mu} = \mu_{\text{réf}}\}$ sera nulle... Si on souhaite envisager la possibilité que les deux méthodes pédagogiques soient rigoureusement équivalentes en moyenne, cette modélisation ne convient donc pas : il faudrait, par exemple, proposer comme priore un mélange de notre priore initiale et d'une priore réservée à l'hypothèse $\{\boldsymbol{\mu} = \mu_{\text{réf}}\}$. ♧

Vision bayésienne des tests d'hypothèse

Avoir parlé de probabilités à postériori va nous servir de prétexte pour introduire un vocabulaire et des idées que nous retrouverons lorsque nous étudierons la notion de *tests* en statistique fréquentiste (cf. chap. 12). Il sera alors utile de comparer les notions fréquentistes que nous développerons avec leurs pendants bayésiens : c'est pourquoi nous allons maintenant parler des pendants bayésiens en question, quand bien même leur intérêt intrinsèque est assez limité !

Dans la suite de cette sous-section, nous allons plus spécifiquement nous intéresser à une alternative entre deux hypothèses complémentaires l'une à l'autre, auxquelles nous ferons jouer des rôles *différents* dans nos analyses, de sorte que nous leur donnerons aussi des appellations différentes. (La raison d'être de ce vocabulaire n'a essentiellement aucun intérêt dans le cadre bayésien, mais sera en revanche cruciale dans le chapitre 12).

! **Définition** (RT). Lorsqu'une hypothèse $\mathscr{H}_0$ est qualifiée de « nulle », l'hypothèse complémentaire (autrement dit, l'évènement $\neg\mathscr{H}_0$) est appelée *hypothèse alternative*. Réciproquement, lorsqu'un hypothèse $\mathscr{H}_1$ est qualifiée d'« alternative », l'hypothèse complémentaire est appelée *hypothèse nulle*.

On utilise conventionnellement les indices respectifs '0' et '1' pour étiqueter les hypothèses resp. nulle et alternative. ♡

Définition (RU) ("p-valeur bayésienne"). Deux hypothèses complémentaires $\mathscr{H}_0$ (nulle) et $\mathscr{H}_1$ (alternative) étant données, dans le cadre d'une analyse bayésienne, la "p-valeur bayésienne" est la probabilité à postériori que l'hypothèse nulle soit vraie, autrement dit que l'hypothèse alternative soit fausse. Il s'agit évidemment de la réalisation d'une certaine statistique bayésienne, statistique qu'on appellera également « la "p-valeur bayésienne" » quand on la verra en tant que variable aléatoire (le contexte permettant de trancher). ♡

Remarque (RV). Vous aurez noté que j'ai mis des guillemets autour de l'expression « p-valeur bayésienne » : c'est parce que cette expression n'est pas standard dans le vocabulaire de la statistique... En fait, je vous *déconseille* même de l'employer : parlez simplement de « probabilité à postériori que $\mathscr{H}_1$ soit fausse » le cas échéant !

Néanmoins, l'expression « p-valeur booléenne » demeure utile sur le plan *pédagogique*, car elle montre que l'idée de « probabilité à postériori que $\mathscr{H}_1$ soit fausse » est l'analogue, dans le cadre bayésien, de la notion fréquentiste de « p-valeur » (tout

court) que nous verrons au chapitre 13. Or, cette notion de p-valeur (fréquentiste) est assez subtile à saisir : et en pratique, les débutants ont tendance à la confondre avec la fameuse "p-valeur bayésienne" de la définition (RU) ci-dessus... En donnant explicitement les deux définitions et en soulignant que l'une est le pendant bayséien de l'autre, j'espère attirer votre attention sur la nuance entre ces deux concepts ! ☺ ♧

Voyons maintenant la notion de ce que j'appellerai, dans ce cours, un « test booléen » [**]. Ce concept sera ré-utilisé dans le cadre fréquentiste ; vous pouvez donc d'ores et déjà le retenir :

Définition (RW) (Test booléen)**.** Un *test (booléen)* est une statistique à valeurs booléennes, autrement dit une fonction (déterministe) de l'observation (passée) qui ne peut valoir que VRAI ou FAUX. !

Un test booléen T est qualifié de « test de l'hypothèse nulle $\mathscr{H}_0$ » (ou de l'hypothèse alternative $\mathscr{H}_1$) lorsqu'il "essaye de deviner" qui de $\mathscr{H}_0$ ou de $\mathscr{H}_1$ est vraie : dans ce cas, lorsque $t_{\checkmark}$ vaut VRAI, on penche en faveur de l'hypothèse alternative tandis que lorsque $t_{\checkmark}$ vaut FAUX, on penche du côté de l'hypothèse nulle. ♡

Remarque (RX)*.* Notez qu'il faut toujours préciser (éventuellement en le sous-entendant par l'indice '0' ou '1') si l'hypothèse dont on est en train de parler joue le rôle d'hypothèse nulle ou d'hypothèse alternative ! Ainsi on n'écrira jamais « T est un test de l'hypothèse $\{\theta \geqslant 1/4\}$ » : on écrira soit « T est un test de l'hypothèse *nulle* $\{\theta \geqslant 1/4\}$ », soit « T est un test de l'hypothèse *alternative* $\{\theta \geqslant 1/4\}$ ». ♧

Un principe très important, qui prendra toutes son importance dans le paradigme fréquentiste, est que la notion de « test » sous-entend une *dissymétrie* entre la façon de traiter l'hypothèse alternative et celle de traiter l'hypothèse nulle :

Principe (RY) (Dissymétrie de l'interprétation d'un test booléen)**.** *Lorsqu'on parle de tests booléns, l'usage est de faire jouer des rôles* différents *à $\mathscr{H}_0$ et $\mathscr{H}_1$! Le principe étant que, lorsque le test vaudra* VRAI*, cela signifiera qu'on a un* haut niveau *de confiance en valeur de l'hypothèse alternative (autrement dit, qu'on la considère comme « très probablement vraie ») ; tandis que lorsque le test vaut* FAUX*, cela signifie simplement qu'on considère l'hypothèse $\mathscr{H}_0$ comme* plausible*, mais sans que cela implique nécessairement qu'elle soit particulièrement* probable *!* ◇ !!

Définition (RZ) ("Test bayésien" de niveau α)**.** Pour $\alpha \in]0,1[$ (sachant qu'on s'intéressera en général à des cas où $\alpha \leqslant 1/2$), un test $T = t(X)$ de l'hypothèse nulle $\mathscr{H}_0$ (ou de l'hypothèse alternative $\mathscr{H}_1$) est dit *de "niveau de risque bayésien" α* (ou « de "niveau de confiance" bayésien $1-\alpha$ ») lorsque la véracité du test entraine que la probabilité à postériori de l'hypothèse nulle est inférieure ou égale à α, ce qui est équivalent à dire que la probabilité à postériori de l'hypothèse alternative est supérieure ou égale à $1-\alpha$:

$$\forall x \in \mathscr{X} \qquad t(x) \quad \Longrightarrow \quad \mathbb{P}(\mathscr{H}_0 \mid X = x) \leqslant \alpha. \tag{SA}$$

♡

Remarque (SB)*.* Notez l'usage des guillemets, confer remarque (RV) : j'ai introduit ce vocabulaire à des fins purement pédagogiques, et je vous déconseille de l'utiliser en pratique ! ♧

[**]. La plupart des références parlent simplement de « test », sans préciser « booléen » : dans ce cours néanmoins, je parlerai aussi de « tests par p-valeur », d'où cette précision.

Remarque (SC). Notez que le fait d'« être un test de "niveau bayésien" α de l'hypothèse alternative $\mathscr{H}_1$ » ne concerne que ce qui se passe lorsque le test vaut VRAI : à l'inverse, lorsque la réalisation d'un test booléen vaut FAUX, savoir que ce test est de "niveau bayésien" α ne nous apporte aucune information supplémentaire... Cela est cohérent avec le principe (RY), qui veut que de toutes façons, c'est seulement dans le cas où un test booléen vaudra VRAI qu'il fournira une conclusion réellement intéressante.

Cette remarque vaudra également dans le paradigme fréquentiste. ♧

En vertu de la remarque (SC), un test de "niveau bayésien" α d'une hypothèse donnée sera d'autant plus intéressant qu'il répondra souvent VRAI. Il est donc naturel de se demander, parmi tous les tests d'une hypothèse donnée ayant un certain "niveau bayésien", s'il en existe un qui répond VRAI plus souvent que les autres. C'est effectivement le cas : parmi tous les tests (booléens) de "niveau bayésien" α d'une hypothèse nulle $\mathscr{H}_0$ donnée, le test optimal [††] est le test $t(X)$ défini par

$$t(x) := \big(\mathbb{P}(\mathscr{H}_0 \mid X = x) \leqslant \alpha\big). \tag{SD}$$

De cela, on en déduit la connexion suivante entre les "p-valeurs bayésiennes" et la façon de construire un test booléen de "niveau bayésien" donné. J'ai noté ces points comme importants, car ils continueront d'êtres valables dans le cadre fréquentiste :

!! *Point* (SE). Ainsi, le "bon" moyen de construire un test booléen de "niveau bayésien" α pour une hypothèse donnée est de répondre en disant si la "p-valeur bayésienne" de cette hypothèse est, ou pas, inférieure ou égale à α. ♧

! **Proposition** (SF). *Réciproquement, si une statistique $p(X)$ est une* [‡‡] *"p-valeur bayésienne" pour une hypothèse donnée, alors le test $\{p(X) \leqslant \alpha\}$ de cette hypothèse est de "niveau bayésien" α.* ♢

9.4 Intervalles de confiance et de prédiction bayésiens

! **Définition** (SG) (Intervalle de confiance bayésien)**.** Dans le cas où φ (ou g, dans le cas d'une modèle d'inférence prédictive) est à valeurs réelles, un *intervalle de confiance bayésien* au niveau de risque α (ou « au niveau de confiance $1-\alpha$ ») [∗] est un intervalle réel $I(X)$ ne dépendant que de l'observation (autrement dit, une statistique à valeurs « intervalles ») tel que, pour (presque-)tout $x \in \mathscr{X}$, on ait

$$\mathbb{P}(\varphi(\theta) \notin I(x) \mid X = x) \leqslant \alpha : \tag{SH}$$

autrement dit, l'intervalle $I(X)$ comprend une proportion au moins $1-\alpha$ de la loi de $\varphi(\theta)$ conditionnellement à l'observation de X. ♡

!

[††]. Du point de vue technique, l'« optimalité » prend le sens suivant : pour tout autre test $t'(X)$ de niveau bayésien α de notre hypothèse, on aura que $t'(X) \implies t(X)$ presque-surement (c'est-à-dire que l'ensemble des x pour lesquels on a $t'(x)$ est inclus dans l'ensemble des x pour lesquels on a $t(x)$).

[‡‡]. On dit que $p(X)$ est « une » "p-valeur bayésienne" pour l'hypothèse nulle $\mathscr{H}_0$ lorsque (avec les notations génériques), pour tout $x \in \mathscr{X}$, on a $\mathbb{P}(\mathscr{H}_0 \mid X = x) \leqslant p(x)$. (La réalisation du cas d'égalité correspondant à « la » p-valeur).

[∗]. En pratique, comme α est toujours choisi $\leqslant 1/2$, on ne précise pas si le niveau dont on parle est un niveau de risque ou de confiance : ainsi, un niveau de 5 % désigne implicitement un niveau de risque, et un niveau de 80 % désigne implicitement un niveau de confiance.

Définition (SI) (Intervalle de prédiction bayésien)**.** Dans le cas où on a un modèle de prévision statistique, avec g à valeurs réelles, un *intervalle de prédiction bayésien* au risque α pour $g(Y)$ est une statistique $I(X)$ à valeurs intervalles telle que, pour (presque-)tout $x \in \mathcal{X}$, on ait

$$\mathbb{P}(g(Y) \notin I(x) \mid X = x) \leqslant \alpha : \tag{SJ}$$

autrement dit, l'intervalle $I(X)$ comprend une proportion au moins $1 - \alpha$ de la loi de $g(Y)$ conditionnellement à l'observation de X. ♡

Remarque (SK). Il est absolument essentiel de bien préciser « bayésien » dans les locutions « intervalle de confiance bayésien » et « intervalle de prédiction » : si vous l'omettez en effet, vous obtenez la locution « intervalle de confiance » (resp. « intervalle de prédiction »), qui se réfère par défaut à un intervalle de confiance (resp. prédiction) *fréquentiste* [†] : or la définition des intervalles de confiance (resp. prédiction) fréquentistes est fondamentalement *différente* de celle de leurs contreparties bayésiennes ! La même remarque vaut pour les intervalles de prédiction. ♧ !

Remarque (SL). Attention, la nomenclature « intervalle de confiance bayésien » peut varier suivant les sources. En particulier, certains auteurs désignent par « intervalle de confiance bayésien » un type d'objet différent (à savoir, un intervalle *construit* selon une procédure bayésienne, mais dont les propriétés requises sont néanmoins *fréquentiste*) : dans ce cours néanmoins nous n'introduirons pas de tels objets, de sorte qu'il n'y aura pas ce risque d'ambigüité. (Et dès lors, l'appellation « intervalle de confiance bayésien » m'a semblé préférable pour bien montrer le parallélisme entre bayésianisme et fréquentisme). Si vous craignez l'ambigüité, vous pouvez préférer la locution « intervalle de croyance » pour désigner les intervalles de confiance bayésiens.

En anglais, les intervalles de confiance bayésiens sont souvent appelés « *credible intervals* ». En français, cela est parfois rendu par « intervalles de crédibilité »... Cette dernière traduction est néanmoins EXTRÊMEMENT malheureuse, car elle laisse entendre que l'intervalle de confiance bayésien serait l'ensemble des choses qu'il n'est "pas déraisonnable" de croire, alors que ce sont au contraire les intervalles de confiance *fréquentistes* qui correspondent à cette idée-là ! L'intervalle de confiance bayésien, lui, correspond à une description de l'état actuel de notre croyance, pas seulement en tant compte de la crédibilité vis-à-vis de l'observation, mais aussi de nos renseignements à priori reflétés par la priore : ce qui peut conduire certaines hypothèses "crédibles" à être rejetées (par que notre priore les pénalise), ou au contraire certaines hypothèses "peu crédibles" à être acceptées (parce que notre priore les favorise). Je vous interdis donc catégoriquement, dans le cadre de ce cours, d'employer la locution « intervalle de crédibilité » pour parler d'intervalles de de confiance bayésien : le cas échéant, cela sera comptabilisé comme une faute ! ♧

Remarque (SM). En fait, rien dans les définitions précédentes n'oblige à ce que $I(X)$ soit un *intervalle* à proprement parler : on pourrait plus généralement définir une *zone* de confiance (resp. prédiction), ce qui aurait l'avantage de se généraliser immédiatement au cas où $\varphi(\bullet)$ est à valeurs dans un espace de dimension > 1. Cependant les intervalles présentent l'avantage pratique de pouvoir se résumer à deux nombres : c'est pourquoi, dans les usages courants de la statistique, c'est quasiment toujours en termes d'intervalles qu'on exprimera de telles zones ; et c'est donc pourquoi nous nous limiterons à ce cas ici. ♧

Définition (SN) (Intervalle de confiance bayésien, bis)**.** En pratique, on parle souvent de « l' »intervalle de confiance (resp. prédiction) bayésien au niveau α, qui correspond alors à la construction standard de l'intervalle de fluctuation pour la loi à postériori de φ : autrement dit, la réalisation de cet intervalle de confiance (resp. !

[†]. Et le risque de confusion est d'autant plus grand que, contrairement au cas des tests (cf. remarque (RV)), la façon de rédiger les phrases de conclusion concernant les intervalles de confiance ou de prédiction est très similaire entre les cas bayésien et fréquentiste !

prédiction) $[\mathrm{Qtile}(\mathrm{Loi}_{\mathrm{post}}(\boldsymbol{\varphi}); \alpha/2), \mathrm{Qtile}(\mathrm{Loi}_{\mathrm{post}}(\boldsymbol{\varphi}); 1-\alpha/2)]$, où $\mathrm{Qtile}(\boldsymbol{P}; \beta)$ désigne le quantile au niveau β d'une distribution de probabilité $\boldsymbol{P}$[‡]. ♡

Remarque (SO). Il peut aussi arriver qu'on ait besoin d'un *minorant (resp. majorant) bayésien au risque* α, à savoir d'une statistique $m(\boldsymbol{X})$ telle que, pour (presque-)tout $x \in \mathcal{X}$, on ait $\mathbb{P}(\varphi(\boldsymbol{\theta}) < m(x) \mid \boldsymbol{X} = x) \leqslant \alpha$ (resp. $\mathbb{P}(\varphi(\boldsymbol{\theta}) > m(x) \mid \boldsymbol{X} = x) \leqslant \alpha$). Il s'agit en fait d'un cas particulier d'intervalle de confiance (resp. prédiction) bayésien où on impose à la borne de droite (resp. de gauche) d'être infinie. Dans ce cas, le minorant (resp. majorant) bayésien optimal au risque α est évidemment $\mathrm{Qtile}(\mathrm{Loi}_{\mathrm{post}}(\boldsymbol{\varphi}); \alpha)$, resp. $\mathrm{Qtile}(\boldsymbol{\varphi}); 1-\alpha)$. ♧

> *Remarque* (SP). Certains ouvrages construisent « l' »intervalle de confiance (resp. prédiction) bayésien au niveau α en faisant en sorte qu'il soit aussi court que possible, ce qui revient à chercher un seuil sur la densité à postériori, aussi haut que possible, tel que la probabilité à postériori de tomber au-dessus de ce seuil soit au moins égale à $1-\alpha$. Nous ne suivrons pas cette approche qui est plus compliquée à calculer (et qui, par ailleurs, est susceptible de conduire à des zones de confiance (resp. prédiction) qui ne sont *pas* des intervalles) ; mais il est utile que vous sachiez qu'une telle convention existe si jamais vous la rencontrez. Une troisième convention envisageable, mais plus rare, serait enfin de fixer un seuil sur la *vraisemblance* du paramètre caché (ce qui aurait l'avantage d'être invariant par reparamétrisation, mais ne donnerait pas forcément un intervalle aussi court—et pas toujours un intervalle non plus, d'ailleurs). ♧

Remarque (SQ). Lorsqu'on souhaite arrondir les valeurs numériques calculées pour un intervalle de confiance ou de prédiction (qu'ils soient bayésiens ou fréquentistes, d'ailleurs), il convient toujours de procéder à l'arrondi en *élargissant* l'intervalle, afin pour rester cohérent avec la définition d'intervalle de confiance : on arrondit donc les bornes inférieures par dessous et les bornes supérieures par dessus. ♧

Exemple (SR). Calculons l'intervalle de confiance bayésien au risque 5 % (c'est une valeur que l'usage a consacrée comme standard en l'absence de contexte particulier) pour le cas du chasseur, avec la priore arcsinus, pour $x_{\checkmark} = 8$. Il nous faut donc trouver les quantiles aux niveaux 2,5 % et 97,5 % de la loi à postériori $\mathrm{Bêta}(8\frac{1}{2}, 17\frac{1}{2})$ de $\boldsymbol{\theta}$. Cela peut se calculer sous R :

```
> qbeta(0.025, 8 + 1 / 2, 17 + 1 / 2)
[1] 0.1644348
0.1721441
> qbeta(0.975, 8 + 1 / 2, 17 + 1 / 2)
[1] 0.5146123
```

On trouve donc que (la réalisation de) l'intervalle de confiance bayésien au risque 5 % vaut $[0{,}164, 0{,}515]$ (concernant les arrondis, confer remarque (SQ) ci-dessus). Autrement dit, au vu de l'épreuve passée par le tireur, il y a 95 % de chances que la fiabilité $\boldsymbol{\theta}$ de celui-ci soit comprise entre 0,172 et 0,515. Ce qui peut paraitre étonnamment large : alors que notre tireur a subi un test sur pas moins 25 items et qu'il a touché un chouïa moins d'un plateau sur trois, l'intervalle probable pour la valeur de son niveau réel inclut des niveaux supérieurs à 50 % ! On a là un premier

[‡]. Dans ce cours, j'écrirai systématiquement l'intervalle sous forme fermée. Les pointilleux remarqueront toutefois que dans certains cas, on pourrait en fait ouvrir une ou deux des bornes de l'intervalle : la définition la plus générale consiste en effet à prendre pour $\boldsymbol{I}$ le plus petit intervalle tel que $\mathbb{P}_{\mathrm{post}}(\forall z \in \boldsymbol{I}\ \varphi(\boldsymbol{\theta}) < z), \mathbb{P}_{\mathrm{post}}(\forall z \in \boldsymbol{I}\ \varphi(\boldsymbol{\theta}) > z) \leqslant \alpha/2$. Cela dit, ouvrir les bornes de l'intervalle n'a aucun intérêt du point de vue pratique, puisque cela ne fait rendre l'intervalle plus étroit que d'*un seul* point, autrement dit rien du tout par rapport à la taille de l'intervalle, et même à la précision des bornes...

exemple où l'analyse statistique contredit notre intuition, qui aurait eu tendance à nous faire sur-interpréter le résultat du test de tir passé par le chasseur.

Dans le contexte du club de chasse, on peut aussi imaginer que ce qui intéresse le club, c'est d'avoir une valeur telle qu'on est quasi-sûr que le niveau du tireur est supérieur à cette valeur (pour avoir une garantie sur son niveau), ou au contraire telle qu'on est quasi-sûr que le niveau du tireur est inférieur à cette valeur (dans le contexte où ce qu'on voudrait serait avant tout ne pas risquer de se priver d'une bonne recrue). On pourrait alors chercher les minorant et majorant bayésiens au risque 5 %, qu'on calcule similairement par :

```
> qbeta(0.05, 8 + 1 / 2, 17 + 1 / 2)
[1] 0.186193
> qbeta(0.95, 8 + 1 / 2, 17 + 1 / 2)
[1] 0.483074
```

Cela donne, après arrondi (cf. remarque (SQ)), un minorant de 18,6 %, resp. un majorant de 48,4 %. (Ces valeurs étant aussi les bornes de l'intervalle de confiance bayésien au risque 10 %).

Si on avait plutôt pris la priore uniforme $\mathrm{Unif}^{\mathrm{me}}(0,1)$ pour $\boldsymbol{\theta}$, on aurait trouvé un intervalle de confiance bayésien au risque 5 % valant $[17{,}2\ \%, 51{,}8\ \%]$.

Remarque (SS). On voit ici que les résultats obtenus sont très similaires quelle que soit la priore utilisée. C'est là un phénomène général, que nous constaterons également lorsque nous parlerons d'estimation : dès lors qu'on d'un (assez) grand nombre d'observations comme ici, l'importance du choix de la priore devient minime, toutes les priores raisonnables donnant alors des résultats comparables. (Par contre, le choix de la priore s'avère souvent crucial lorsque les données sont moins nombreuses !). ♧

Pour compléter l'exemple ci-dessus, voyons comment évolue l'intervalle de confiance bayésien au risque 10 % (et la médiane de la loi à postériori, qui représente en quelque sorte le "centre" des intervalles de confiance, et dont nous verrons plus loin qu'elle correspond aussi à un *estimateur* pertinent de $\boldsymbol{\theta}$) à mesure que le nombre de plateaux lancés au cours du test passé par notre chasseur progresse, en supposant que le test va même jusqu'à 125 plateaux (autrement dit, ici la valeur de n utilisée pour l'analyse évolue avec le nombre total de plateaux lancés). Ici le niveau réel de notre tireur a été pris à $\theta_{\checkmark} = 40\ \%$ pour la simulation, et ses résultats successifs ont été les suivants, où 0 indique un échec et 1 un succès [§] :

```
  0 1 0 0 1 1 0 0 0 1 0 0 0 0 1 1 0 1 1 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0
  0 0 ↵
... 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 1 1 1 1 0 0 0 0 1 0 1 0 1 0 1 0 0 0
  1 1 ↵
... 1 1 0 0 0 0 1 1 1 1 0 0 1 1 0 0 1 1 0 0 0 1 1 0 0 0 0 1 1 1 0 1 1 0
  0 0 ↵
... 0 0 0 0 0 0 0 1 1 1 1 0 0 1 1 1 0 1
```

Le tracé de la façon dont le club évalue le niveau du tireur est donné par la figure 9.1. Cette figure nous montre plusieurs points qui satisfont notre bon sens : notamment, après un tir réussi, l'estimateur et les deux bornes de l'intervalle de confiance bayésien pour $\boldsymbol{\theta}$ se déplacent vers le haut, alors qu'après un tir raté, c'est l'inverse. On voit aussi que, à mesure qu'on collecte des données sur la performance du chasseur, l'intervalle de confiance bayésien sur $\boldsymbol{\theta}$ est de plus en plus étroit, ce qui signifie que notre incertitude sur son niveau réel est de plus en plus faible. Par

[§]. Pour information, chaque ligne comporte 36 essais.

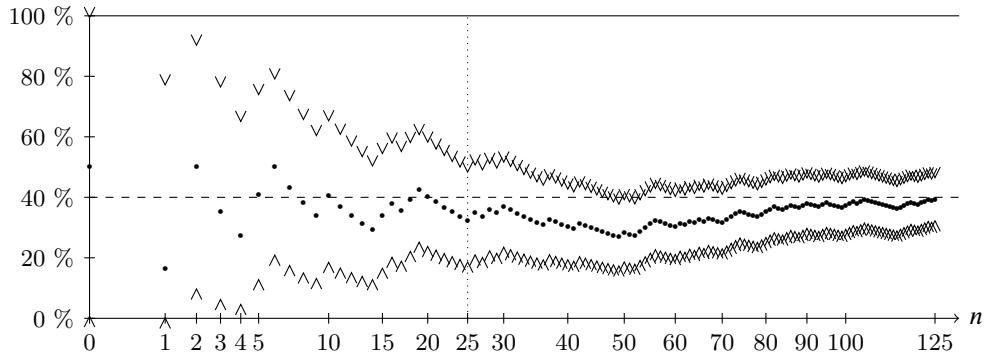

FIGURE 9.1 – Évolution de la médiane de la loi à postériori de $\boldsymbol{\theta}$ (marquée par le symbole $\cdot$) et de l'intervalle de confiance bayésien (entre les symboles $\wedge$ et $\vee$) obtenus par la méthode bayésienne à mesure du nombre de tirs tentés par le chasseur. L'axe des abscisses a été gradué irrégulièrement pour faciliter la visualisation. La ligne tiretée horizontale indique le niveau réel de réussite du chasseur utilisé pour la simulation, qui était de 40 % ; la ligne pointillée verticale correspond à $n = 25$, pour lequel, conformément aux valeurs utilisées précédemment, le chasseur a réussi 8 tirs.

ailleurs, sauf sur une courte période, la véritable valeur $\theta_{\checkmark}$ est bien contenue dans l'intervalle de confiance. On pourrait démontrer que si on continuait à tracer ce graphique à l'infini en faisant tirer le chasseur encore et encore, les deux bornes de l'intervalle de confiance (et à fortiori la médiane à postériori) convergeraient vers $\theta_{\checkmark}$.

Notez qu'on voit clairement que l'intervalle de confiance bayésien ne nous donne qu'une indication solide sur la valeur de $\boldsymbol{\theta}$ (fiable à 90 %, très précisément ◡̈), mais pas de certitude absolue : par exemple, entre $n = 46$ et $n = 53$, il se trouve que, par chance, le tireur a largement sous-performé jusqu'à présent, de sorte que le club tend alors à nettement sous-estimer son niveau réel et que $\theta_{\checkmark}$ n'est *pas* dans l'intervalle de confiance. Au demeurant, si on devait raisonner en termes de *certitudes*, il est clair que ce qui est *rigoureusement* impossible au vu d'un certain ensemble d'informations ne peut pas redevenir possible avec des informations supplémentaires : ainsi, si on devait raisonner en termes de certitudes, il est clair que l'intervalle où l'on *sait* que $\boldsymbol{\theta}$ se trouve [¶] ne pourrait faire que devenir de plus en plus petit [‖] à mesure qu'on récolte des données. Or ce n'est pas le cas ici : par exemple, après 4 tirs, l'intervalle de confiance $[4{,}6\ \%, 65{,}1\ \%]$ ne contient pas la valeur 70 %, alors qu'après le succès au 5e tir, cette valeur revient à nouveau dans l'intervalle ! (qui devient alors $[12{,}7\ \%, 74{,}0\ \%]$). ♧

9.5 Décisions et estimateurs

Introduction aux notions d'estimation et de prédiction

La théorie de la décision optimale peut s'appliquer au cas où on considère une certaine quantité d'intérêt *explicative* $\boldsymbol{\varphi}(\boldsymbol{\theta}) =: \boldsymbol{\varphi}$ et que notre "décision" consiste à

[¶]. Dans le cas du modèle probabiliste considéré, en fait, dans la mesure où on a exclu 0 et 1 se l'espace du paramètre caché, *toute* valeur de $\boldsymbol{\theta}$ est *théoriquement* compatible avec l'observation !...

[‖]. « De plus en plus petit » au sens ensembliste du terme, s'entend, pas seulement en termes de largeur !

essayer deviner quelle est la valeur de $\boldsymbol{\varphi}$ au moyen d'une statistique $\hat{\varphi}(X) =: \hat{\boldsymbol{\varphi}}$. Dans ce cas, on parle plus spécifiquement d'*estimation* :

Définition (ST) (Estimateur)**.** Un *estimateur* d'une quantité d'intérêt explicative $\varphi(\boldsymbol{\theta}) =: \boldsymbol{\varphi}$ est une statistique $\hat{\varphi}(X) =: \hat{\boldsymbol{\varphi}}$ qui "essaye de ressembler" à $\boldsymbol{\varphi}$. La réalisation $\hat{\varphi}(x_{\checkmark})$ d'un estimateur est qualifiée d'*estimation* de la quantité d'intérêt $\varphi(\theta_{\checkmark})$. !!

La convention est d'utiliser le même symbole pour désigner l'estimateur que celui pour désigner la quantité qu'il estime, mais agrémenté d'un chapeau.

L'*estimation statistique* est la branche de la statistique qui s'intéresse à la façon d'obtenir de (bons) estimateurs des quantités d'intérêt explicatives. ♡

Remarque (SU)*.* Dans la définition ci-dessus, la locution « essaye de ressembler » a été mise entre guillemets, car ce n'est pas un concept mathématique : ce qui fait qu'on qualifie $\hat{\varphi}(X)$ d'estimateur de $\varphi(\boldsymbol{\theta})$, c'est qu'on a décidé d'introduire cette quantité avec l'*intention* d'approcher $\varphi(\boldsymbol{\theta})$ par un statistique ! Mais, formellement, n'importe quelle statistique peut être considérée comme un estimateur de n'importe quelle quantité d'intérêt (sous réserve, tout de même, qu'elles soient à valeurs dans le même espace). En fait, c'est au niveau de l'*usage* qu'on va chercher à ne s'intéresser, autant que possible, qu'à des estimateurs ayant certaines "bonnes" propriétés. ♧

Le pendant de la notion d'estimation dans le cadre prédictif est qualifié de *prédiction* :

Définition (SV) (Prédicteur)**.** Un *prédicteur* d'une quantité d'intérêt explicative $g(\boldsymbol{Y}) =: \boldsymbol{G}$ est une statistique $\hat{g}(\boldsymbol{X}) =: \hat{\boldsymbol{G}}$ qui "essaye de ressembler" à $\boldsymbol{G}$. La réalisation $\hat{g}(x_{\checkmark})$ d'un prédicteur est qualifiée de *prédiction* de la quantité d'intérêt $g(y_{\checkmark})$. La *prédiction statistique* est la branche de la statistique qui s'intéresse à la façon d'obtenir de (bons) prédicteurs des quantités d'intérêt prédictives. ♡ !

Remarque (SW)*.* Dans le cadre bayésien, estimation est prédiction se traiteront exactement de la même façon, *mutatis mutandis*. Si nous avons introduit des définitions clairement séparées pour ces deux concepts, c'est parce que nous verrons que dans le cadre fréquentiste, par contre, il y a des techniques bien spécifiques à l'estimation qui n'ont pas leur pendant dans le cadre prédictif. ♧

Remarque (SX)*.* Dans la mesure où un estimateur (resp. un prédicteur) essaye de ressembler à la quantité d'intérêt qu'il estime (prédit), il va de soit qu'il faudra que les espaces dans lesquels vivent les v.a. $\boldsymbol{\varphi}$ et $\hat{\boldsymbol{\varphi}}$ (resp. $\boldsymbol{G}$ et $\hat{\boldsymbol{G}}$) soient les mêmes.

Néanmoins, il pourra arriver qu'on se permette, dans certains cas, d'"étendre" l'espace $\mathcal{G}$ dans lequel vit $\boldsymbol{\varphi}$ (ou $\boldsymbol{G}$) : par exemple, dans certains cas, il peut être intéressant d'estimer un nombre entier par un nombre réel. Cela se fait d'ailleurs plus ou moins dans la vie de tous les jours : si je dis « vu l'épidémie de gastro-entérite, il devrait y avoir environ cinq ou six élèves absents cette semaine », on peut considérer que « cinq ou six » est une locution signifiant en fait « 5,5 » [**], et donc que j'énonce une prédiction du nombre d'élèves absents (qui sera, évidemment, un entier !) comme valant 5,5, afin de rester assez bien compatible avec tous les scénarios. (C'est la même idée que pour la remarque digressive **??** un peu plus haut). Mais dans tous les cas on peut considérer que quantité d'intérêt et prédiction vivent dans le même espace : pour cela, il me suffit de décréter que la quantité d'intérêt

[**]. En tout cas, si je dis qu'un sac pèse « cinq ou six kilos », je veux bien dire qu'il pèse environ 5,5 kg, et pas qu'il pèse *soit* 5 kg, *soit* 6 kg ! (Même si on peut considérer que, dans un tel cadre, cette locution indique *aussi* l'ordre de grandeur de notre incertitude sur l'estimation du poids).

« nombre d'élèves absents » est à valeurs dans $\mathbb{R}_+$: ce qui est techniquement tout à fait correct, même si, *en pratique*, seules des valeurs dans $\mathbb{N}$ pourront effectivement être prises ! ♧

Dans le contexte de l'estimation ou de la prédiction, on peut, exactement comme dans le cadre général de la théorie de la décision, introduire une fonction de perte. Cette fonction de perte $\ell(g, \hat{g})$ sera néanmoins un peu particulière, puisque ses deux arguments g et $\hat{g}$ vivront dans le même espace $\mathcal{G}$. En outre, notre but étant ici que notre estimation (ou prédiction) approche au mieux la quantité d'intérêt, on veut que, pour tout $g \in \mathcal{G}$, l'application $\hat{g} \mapsto u(g, \hat{g})$ prenne son maximum pour $\hat{g} = g$. On choisit alors de normaliser "verticalement" la fonction de perte de sorte que $\ell(g, g) = 0$ pour tout $g \in \mathcal{G}$[††]. De ce fait, les pertes seront positives : c'est pourquoi on tend à préférer ce formalisme sur celui de l'utilité, où on se retrouverait à manipuler des quantités *négatives* ☹

Convention (SY). Dans le cadre de l'estimation et de la prédiction, les fonctions de perte $(g, \hat{g}) \mapsto \ell(g, \hat{g})$ ne prennent que des valeurs positives ou nulles, et vérifient $\ell(g, g) = 0$ pour tout g[‡‡]. ♡

> *Remarque* (SZ). La convention de normalisation à zéro peut aboutir à des fonctions de pertes assez difficiles à appréhender dans le cadre industriel, que je voudrais essayer d'expliquer ici. Mettons que vous soyez un éditeur et que vous veniez d'accepter le roman d'une nouvelle autrice : vous aimeriez savoir combien de lecteurs seront intéressés par acheter ce roman pour ajuster au mieux votre tirage : l'idée étant que, si vous saviez qu'il y a exactement φ lecteurs intéressés, vous imprimeriez précisément φ exemplaires afin de rentrer au mieux dans vos frais. Nous prenons ici le modèle économique simplifié suivant (où nous assimilons gain financier et utilité) : l'impression de chaque exemplaire nous coute a, et la vente de chaque examplaire nous rapporte $b > a$; en outre, tous les lecteurs intéressés achèteront un exemplaire dans la limite des stocks disponibles. Dès lors, si le nombre de lecteurs réels est de φ et que nous l'avons estimé à $\hat{\varphi}$ (et donc que nous avons imprimé $\hat{\varphi}$ exemplaires), nous aurons engagé des frais $a\hat{\varphi}$ pour l'impression et récupèrerons $b(\varphi \wedge \hat{\varphi})$ à la vente, ce qui suggère de prendre comme perte $\ell_{\text{brut}}(\varphi, \hat{\varphi}) := a\varphi - b(\varphi \wedge \hat{\varphi})$. Néanmoins, on voit que cette perte n'est pas normalisée : lorsque $\hat{\varphi} = \varphi$, $\ell_{\text{brut}}(\varphi, \varphi)$ ne vaut pas du tout zéro... L'idée derrière la normalisation est de prendre
>
> $$\ell_{\text{norm}}(\varphi, \hat{\varphi}) := \ell_{\text{brut}}(\varphi, \hat{\varphi}) - \ell_{\text{brut}}(\varphi, \varphi)\ ^{[*]} : \qquad \text{(TA)}$$
>
> ainsi, la fonction de perte décrit un *manque à gagner* lorsqu'on croit que le nombre de lecteurs intéressés est $\hat{\varphi}$, *par rapport à si on avait su* que ce nombre était φ : mais ce n'est pas une perte *au sens propre* : il se peut très bien que $\ell(\varphi, \hat{\varphi})$ soit positif mais qu'en réalité l'entreprise fasse un bénéfice en ayant estimé φ par $\hat{\varphi}$!
>
> Au passage, noter que, ici, notre fonction de perte normalisée a une forme plutôt simple, mais non symétrique : on a $\ell(\varphi, \hat{\varphi}) = (b - a)(\varphi - \hat{\varphi})$ pour $\hat{\varphi} \leqslant \varphi$ (on aurait pu imprimer $(\varphi - \hat{\varphi})$ livres supplémentaires et faire un bénéfice de $b - a$ sur chacun d'eux) ; et de $\ell(\varphi, \hat{\varphi}) = a(\hat{\varphi} - \varphi)$ pour $\hat{\varphi} \geqslant \varphi$ (on a gâché la dépense provenant de l'impression de $(\hat{\varphi} - \varphi)$ livres qui ne seront pas achetés). ♧

Remarque (TB). Attention, une fonction de perte n'est pas nécessairement symétrique quand on échange les variables : par exemple, si vous cherchez à estimer

[††]. Cela ne perd pas en généralité, car cette translation de la perte ne changera rien à la comparaison des risques intégrés des différentes décisions, et donc au choix de la décision : confer remarque (HU').

[‡‡]. La plupart des fonctions de pertes seront en outre telles que $\ell(g, \hat{g}) > 0$ dès que $\hat{g} \neq g$, mais il peut néanmoins y avoir des exceptions.

[*]. Attention ! La normalisation doit se faire en retranchant $\ell(\varphi, \varphi)$, surtout pas $\ell(\hat{\varphi}, \hat{\varphi})$: en effet, c'est quand on translate la perte par une fonction de φ qu'on laisse le problème physiquement invariant ; par contre, translater par une fonction de $\hat{\varphi}$ changerait radicalement la modélisation... !

la résistance d'un acier pour construire un pont, estimer la limite d'élasticité de votre acier à 300 MPa alors que la réalité est à 500 MPa sera certes problématique (car vous utiliserez alors plus d'acier que nécessaire, entrainant des surcouts), mais moins qu'estimer que la limite d'élasticité est de 500 MPa alors que la réalité est de 300 MPa, où votre pont risque de se détériorer rapidement après sa construction, voire de s'effronder... ! ☹ ♧

Remarque (TC). Dans de nombreux ouvrages de statistique, écrits par des mathématiciens, on ne s'intéresse pas vraiment à déterminer la fonction de perte, puisque le contexte industriel n'est pas présent à l'esprit [†]. De ce fait, on utilise souvent des fonctions de perte génériques. En particulier, une fonction de perte très courante consiste à prendre $\ell(\varphi_0, \varphi_1) = (\varphi_1 - \varphi_0)^2$, qui a l'avantage d'être strictement convexe et très simple analytiquement : c'est ce qu'on appelle la *fonction de perte quadratique*. Cependant cette fonction de perte n'est pas une panacée dans les situations pratiques : et bien trop souvent, au motif que c'est elle qui est mentionnée dans les ouvrages de référence, on l'utilise dans des contextes où elle est tout-à-fait inadaptée...! ♧

Lorsque la quantité d'intérêt est à valeurs réelles, il y a trois fonctions de perte particulièrement classiques qu'il est utile de connaitre :

Définition (TD)**.**

— La fonction de perte $(g, \hat{g}) \mapsto \mathbf{1}_{\hat{g} \neq g}$ est appelée « fonction de perte de Hamming ». Le risque associé (autrement dit, l'espérance de la fonction de perte) est appelé *taux d'erreur*, abrégé en anglais « ER » (*error rate*). Cette fonction de perte peut en fait être définie sur n'importe quel espace, et en général on l'utilise surtout sur des espaces discrets, même s'ils ne se plongent pas naturellement dans $\mathbb{R}$.

— La fonction de perte $(g, \hat{g}) \mapsto |\hat{g} - g|$ est appelée *erreur absolue*, ou *erreur* L^1. Le risque associé est appelé *erreur absolue moyenne*, abrégé en anglais « MAE » (*mean absolute error*).

— La fonction de perte $(g, \hat{g}) \mapsto (\hat{g} - g)^2$ est appelée *erreur quadratique*. Le risque associé est appelé *erreur quadratique moyenne*, abrégé en anglais « MSE » (*mean squared error*). Pour des raisons d'homogénéité physique, on préfère généralement afficher la *racine carrée* de l'erreur quadratique moyenne, abrégée « RMSD » ou « RMSE » en anglais (*root-mean-square deviation* [‡], resp. *root mean square error*).

♡

Pour les trois fonctions de perte ci-dessus, la solution du problème d'optimisation particularisé auquel nous conduit le théorème **??** a une forme particulièrement élégante :

Théorème (TE)**.** *Soit $\boldsymbol{G}$ une quantité d'intérêt à valeurs réelles que nous cherchons à estimer ou prédire dans le cadre bayésien.*

(i) Si nous utilisons la fonction de perte de Hamming, la réalisation de l'estimateur (ou du prédicteur) optimal sera la valeur de g pour laquelle $\mathbb{P}_{\mathrm{post}}(\boldsymbol{G} = g)$ sera maximal (en supposant ici que $\boldsymbol{G}$ est à valeurs discrètes) : autrement

[†]. Non tant parce que les mathématiciens se fichent des applications, mais tout simplement parce qu'ils développent des méthodes *générales* : or, il peut y avoir une foultitude de contextes industriels correspondants, avec des fonctions de pertes très différentes !

[‡]. En fait, cette idée de prendre la racine de la moyenne d'un carré est apparentée à la notion d'*écart-type* ; or « écart-type » se dit « standard deviation » en anglais, d'où le fait qu'on utilise régulièrement le sigle « RMSD ».

dit, c'est le mode *de la loi à postériori de* $\boldsymbol{G}$ [§] [¶]*. (Cela reste en l'occurrence valable pour* $\boldsymbol{G}$ *à valeurs dans un espace quelconque).*

(ii) Si nous utilisons l'erreur absolue comme fonction de perte, la réalisation de l'estimateur/prédicteur optimal sera la médiane *de la loi à postériori de* $\boldsymbol{G}$*.*

*(iii) Si nous utilisons l'erreur quadratique comme fonction de perte, la réalisation de l'estimateur/prédicteur optimal sera l'*espérance *de la loi à postériori de* $\boldsymbol{G}$*.*

◇

Remarque (TF). Si $\boldsymbol{G}$ est à valeurs dans $\mathbb{R}^d$, l'espérance à postériori de $\boldsymbol{G}$ peut également être définie : dans ce cas, elle correspond à optimiser l'espérance de la fonction de perte $\ell(\varphi, \hat{\varphi}) \coloneqq Q(\hat{\varphi} - \varphi)$, pour $Q(\bullet)$ une forme quadratique définie positive [‖] quelconque sur $\mathbb{R}^d$. (Quelle que soit la forme quadratique définie positive choisie, la solution au problème d'optimisation sera l'espérance à postériori). ♧

! *Remarque* (TG). Ce qui est remarquable, c'est qu'avec cette approche fondée sur la notion de décision optimale, on finit par retomber sur une idée toute simple : pour estimer ou prédire notre quantité d'intérêt $\boldsymbol{G}$, on va regarder sa distribution à postériori et essayer de "résumer" celle-ci par une valeur "typique". Les trois indicateurs les plus classiques pour résumer une distribution de probabilité à valeurs réelles sont l'espérance, la médiane et le mode : et ce que montre le théorème ci-dessus, c'est que chacun de ces choix peut en fait être considéré comme optimal par rapport à une fonction de perte appropriée ! ☺ ♧

Remarque (TH). À noter que dans le cadre où $\boldsymbol{G}$ est une variable à densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^d$, on peut également définir son « mode à postériori » comme le point où la densité à postériori par rapport à la mesure de Lebesgue atteint son maximum. Il convient néanmoins de prendre garde à ce que, malgré la similarité des noms, le « mode » d'une distribution de probabilité à densité n'a pas les mêmes propriétés d'invariance que le « mode » d'une distribution discrète. ♧

Remarque (TI). Dans le cas d'une distribution à densité, le mode à postériori peut aussi être vu comme résultant de l'optimisation de l'espérance d'une certaine perte, mais la formalisation est un peu plus subtile. Soit c une fonction continue croissante de $\mathbb{R}_+$ dans $[0,1]$ vérifiant $c(0) = 0$ et $c(d) \equiv 1$ pour d suffisamment grand (mais pour le reste, la forme précise de $c(\bullet)$ importe peu) ; et pour $\varepsilon > 0$, condisérons la fonction de perte $\ell^{(\varepsilon)}(\varphi, \hat{\varphi}) \coloneqq c(|\hat{\varphi} - \varphi| \,/\, \varepsilon)$, où $|\bullet|$ se réfère à une norme quelconque sur $\mathbb{R}^d$. Alors l'estimateur ou prédicteur qui optimise le risque associé à la perte $\ell^{(\varepsilon)}$ va tendre, lorsque $\varepsilon \to 0$, vers le mode de la loi à postériori de $\boldsymbol{G}$ (qu'on suppose ici exister et être unique). ♧

Exemple (TJ). Voyons ce que donnent les trois estimateurs classiques de $\boldsymbol{\theta}$ dans pour le modèle du chasseur, avec les valeurs par défaut (en particulier $x_{\checkmark} = 8$). Dans ce cas, comme nous l'avons vu, la loi à postériori de $\boldsymbol{\theta}$ est la loi Bêta$(8\frac{1}{2}, 17\frac{1}{2})$. Les mathématiciens ont établi une formule exacte pour l'espérance, resp. le mode, des lois bêta : pour $\alpha, \beta > 0$, on a

$$\mathbb{E}\big(\text{Bêta}(\alpha, \beta)\big) = \frac{\alpha}{\alpha + \beta}\,; \tag{TK}$$

[§]. On peut dire aussi « le mode à postériori de $\boldsymbol{G}$ » ; la même remarque vaut aussi pour la médiane, l'espérance, &c.

[¶]. Dans de nombreuses références, le mode à postériori de $\boldsymbol{G}$ est également appelé (par abus de langage) *maximum à postériori* de $\boldsymbol{G}$, ce qu'on abrège communément en « MAP ».

[‖]. Une *forme quadratique définie positive* sur $\mathbb{R}^d$ est une application de la forme $x \mapsto \vec{x}^{\mathsf{T}} \boldsymbol{M} \vec{x}$, où $\vec{x}$ est le vecteur des coordonnées de x dans la base canonique de $\mathbb{R}^d$ est $\boldsymbol{M}$ une matrice symétrique définie positive. (Pour information, il existe aussi une définition directe ne passant pas par les matrices).

et pour $\alpha, \beta \geqslant 1$:

$$\arg\max_{x\in]0,1[} \frac{\mathbb{P}\big(\text{Bêta}(\alpha,\beta) \in \mathrm{d}x\big)}{\mathrm{vol}_1(\mathrm{d}x)} = \frac{\alpha-1}{\alpha+\beta-2}. \tag{TL}$$

On en déduit, en l'occurrence, que l'espérance à postériori de $\boldsymbol{\theta}$ vaut 0,327, et que son mode à postériori vaut 0,312. Quant à la médiane à postériori, il n'y a pas de formule simple, mais on peut la calculer par exemple par la commande suivante sous R :

```
> qbeta(1 / 2, 8 + 1 / 2, 17 + 1 / 2)
[1] 0.3224241
```

Par ailleurs, si, à la place de la priore arcsinus, on avait utilisé la priore uniforme sur $]0,1[$, on aurait obtenu comme postériore la loi $\text{Bêta}(9,18)$; et les trois estimateurs ci-dessus auraient été remplacés respectivement par 0,333, 0,320 et 0,329.

On voit sur cet exemple qu'il n'y a pas forcément *un* estimateur parfait pour la quantité d'intérêt : selon les hypothèses de modélisation (bayésienne) et le critère de qualité visé (confer fonction de perte), on trouve des estimateurs différents ! Cependant il est aussi rassurant de constater que tous ces estimateurs donnent des résultats très proches, et cohérents avec l'idée intuitive que notre tireur devrait être en mesure de toucher à peu près 8 plateaux sur 25 (puisque c'est la performance qu'il a eue lors du test) : cela était attendu, puisqu'on est en train d'exploiter les mêmes données, et que celles-ci sont suffisamment nombreuses pour qu'on commence à se faire une idée honnête de la vraie valeur de $\boldsymbol{\theta}$... Confer aussi remarque (SS) *supra*. ♧

Remarque (TM). Il peut aussi être naturel de d'intéresser à des fonctions de perte qui croissent très vite lorsque la distance entre g et $\hat{g}$ augmente, ce qui revient en substance à vouloir contrôler le pire cas possible concernant cette distance : par exemple, on peut instancier cette idée en considérant la fonction de perte $\ell(g,\hat{g}) \coloneqq |\hat{g}-g|^p$ pour $p \to \infty$. Dans ce cas encore, (l'asymptotique de) l'estimateur optimal est un indicateur de tendance centrale assez classique : il s'agit du milieu de la plage de valeurs que la loi à postériori de G est susceptible de prendre (plus précisément, c'est $\big(\text{quantile}(\text{Loi}_{\text{post}}(G);0+)+\text{quantile}(\text{Loi}_{\text{post}}(G);1-)\big)/2$), ce qu'on appelle parfois la « mi-gamme » [**] de cette loi. Néanmoins, cet indicateur ne présente guère d'intérêt dans les situations qu'on rencontre en pratique, car la valeur centrale de $\text{Loi}(G \mid X = x)$ est très souvent indéfinie et/ou indépendante de x, de sorte qu'elle ne présente guère d'intérêt pour l'analyse statistique ! ⚟ ♧

9.6 Conclusion sur l'analyse bayésienne

Le contexte bayésien présente l'avantage considérable qu'on a une réponse précise aux questions qu'on se pose sur $\varphi(\boldsymbol{\theta})$ (ou $g(Y)$), qu'on peut calculer d'une façon presque "mécanique". Les concepts de probabilité à postériori, de test, d'intervalle de croyance et d'estimateur (dont nous retrouverons des avatars dans la partie III) viennent automatiquement accompagnés d'une façon de les calculer qui est optimale dans un certain sens. En outre, on comprend très bien la mécanique du raisonnement ; bref, c'est parfait...

Le souci cependant, c'est que dans de très nombreuses situations, il n'y a pas de manière vraiment *objective*, voire pas même *neutre*, de déterminer la distribution de probabilité à priori pour $\boldsymbol{\theta}$. Prenons l'exemple de notre chasseur : nous avons

[**]. Également appelée : milieu de gamme, valeur centrale, mid-extrême, *midrange* (anglicisme).

dit que nous considérions que son taux de succès suivait à priori une loi Bêta$\left(\frac{1}{2}, \frac{1}{2}\right)$, mais... il n'y a pas de raison que ce soit vraiment le cas ! En effet, comme nous le développerons dans la § 15, sauf cas bien particulier, il n'y a pas de façon objective de choisir la priore, pour deux raisons essentielles :

— D'une part, la priore reflète l'état des informations, formelles et informelles, dont on dispose sur $\boldsymbol{\theta}$. Deux individus disposant d'informations différentes, ou même tout simplement d'intuitions différentes sur l'état du monde, risquent donc de se retrouver en désaccord sur le choix de la priore approprié pour un problème donné, sans qu'il y ait de façon incontestable de des départager...

— D'autre part, même lorsqu'on est d'accord sur les informations dont on dispose, il n'y a pas véritablement de méthode parfaite pour transformer celles-ci en priore : une forme de "tour de main" s'avère alors indispensable, sur lequel deux statisticiens pourront être en désaccord, et ce désaccord peut dans certains cas (en particulier lorsque la dimension de $\boldsymbol{\Theta}$ est grande, voir ci-après) conduire à des conclusions fortement différentes... !

Il est certes vrai que, dans un certain nombre de cas, le choix de la priore n'est pas *si* important que cela, car les données dont on dispose sont suffisamment nombreuses pour "écraser" l'impact du choix de la priore : c'est notamment ce que nous avons observé dans les exemples `(SR)` et `(TJ)`. Néanmoins, d'une part, cela requiert que les données soient suffisamment nombreuses (or dans de nombreuses situations pratiques, on est intrinsèquement limité quant aux données qu'on peut collecter) ; et d'autre part, ce phénomène ne se produit que très laborieusement (au sens où il faut vraiment *énormément* de données) lorsque la dimension de l'espace du paramètre caché $\boldsymbol{\Theta}$ devient grande [††] [‡‡] ! Et il y a même des cas où $\boldsymbol{\Theta}$ est de dimension infinie, par exemple si le paramètre caché est une fonction continue de $\mathbb{R}$ dans $\mathbb{R}$... La problématique du choix de la priore est donc véritablement une épine dans le pied à subir dès lors qu'on fait de la statistique bayésienne, dont il est impossible de se débarrasser complètement...

... En revanche, il y a une chose qui est parfaitement connue *sans* avoir à formuler aucun jugement subjectif concernant la priore : la probabilité que telle ou telle observation se produise selon que le paramètre caché vaille tant ou tant ! (P. ex., dans le cas du chasseur, la probabilité, pour un niveau de fiabilité donné, de toucher tant ou tant de plateaux lors du test). C'est pourquoi, dans un grand nombre de situations, on va préférer recourir à des techniques *fréquentistes*, dans lesquelles on se passera complètement de l'information sur la loi à priori !

[††]. La raison en est un peu compliquée à expliquer ; mais intuitivement, l'idée est que chaque dimension supplémentaire de l'espace du paramètre caché $\boldsymbol{\Theta}$ peut représenter un paramètre sur lequel il y a une ambigüité quant au choix de priore à prendre, et que quand on cumule les effets d'un grand nombre de paramètres, le grand nombre de choix différents qu'ont pu faire deux statisticiens finit par avoir un effet très important, qui ne pourrait être pallié que par un nombre gigantesque de données...

[‡‡]. Un cas extrêmement important où la dimension de $\boldsymbol{\Theta}$ peut devenir *très* grande est celui des *réseaux de neurones* utilisés en intelligence artificielle. L'apprentissage automatique à l'aide d'un tel réseau de neurones consiste en substance à considérer que les données qu'on observe (par exemple, dans le cas d'un modèle de langage, il s'agit des exemples de textes récoltés pour entrainer le modèle) provient d'un certain réglage du réseau de neurones (ou du moins, peut être très bien imité par un tel réglage), dont le paramètre $\boldsymbol{\theta}_{\checkmark}$, décrivant *l'ensemble* des poids à utiliser pour connecter les différents neurones, est inconnu. Dans cette optique, l'apprentissage automatique peut être vu comme une *inférence statistique* sur $\boldsymbol{\theta}$, où la dimension de l'espace du paramètre caché, pour les modèles les plus avancés comme *GPT*, peut atteindre plusieurs *milliards*... !

Dans la partie suivante de ce polycopié, nous allons présenter les méthodes d'analyses qu'on peut utiliser dans un tel cadre fréquentiste.

Troisième partie

Méthodes fréquentistes

Chapitre 10

Notion de vraisemblance

Lors de la partie de ce cours consacrée à l'approche bayésienne de l'inférence statistique, nous avons mis en évidence un concept essentiel intervenant au cœur du théorème de Bayes : à savoir, la *fonction de vraisemblance* associée l'observation qui a été réalisée (définition (PE)). Cependant, nonobstant son rôle crucial en statistique bayésienne, la vraisemblance est un concept fondamentalement *fréquentiste*, puisque pouvant être définie sans faire référence à quelque choix de priore que ce soit... À vrai dire, la vraisemblance est même un des rares outils qui restent encore à notre disposition dans le paradigme fréquentiste : et pour cette raison, elle joue en analyse fréquentiste un rôle encore plus important que dans le paradigme bayésien ! C'est pourquoi, plutôt que de reléguer le concept de vraisemblance à une simple section du chapitre sur le théorème de Bayes, il m'a semblé approprié de lui consacrer un chapitre à part entière, et de placer ce chapitre au sein de la partie fréquentiste du cours. C'est ce chapitre que vous vous apprêtez à lire : parlons donc, en toute généralité, de la notion de vraisemblance ! ◡̈

10.1 Définition générale et type d'objet

Définition et notation

Pour commencer, rappelons la définition générale de la vraisemblance, telle qu'elle nous est apparue dans le chapitre 8. Il est à noter que, pour les applications qu'un ingénieur pourra être amené à rencontrer, il n'est pas nécessaire de connaitre cette définition générale : je le mentionne donc uniquement par souci de cohérence interne ! ◠̈

Définition (TN) (Vraisemblance)**.** Considérons un modèle de statistique avec les notations génériques. On appelle « fonction de vraisemblance du paramètre caché au vu de l'observation effectivement réalisée », et on note $\theta \mapsto \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark})$ [∗] (ou plus simplement $\mathscr{L}(\boldsymbol{\theta} = \theta)$, voire $\mathscr{L}(\theta)$, lorsqu'il n'y a pas d'ambigüité), toute application de $\boldsymbol{\Theta}$ dans $\mathbb{R}_+$ qui est proportionnelle à l'application

$$\theta \mapsto \mathbb{P}(X \in \mathrm{d}x_{\checkmark} \mid \boldsymbol{\theta} = \theta), \tag{TO}$$

où $\mathrm{d}x_{\checkmark}$ désigne un voisinage infinitésimal arbitraire de $x_{\checkmark}$, à condition que cette application prenne des valeurs qui ne soient ni infinitésimalement petites, ni infini-

[∗]. Le symbole '$\mathscr{L}$', universellement employé, provient de l'anglais *likelihood*.

tésimalement grandes. (Autrement dit, il doit s'agir d'une application qu'on peut évaluer numériquement).

Il s'avère que le choix du voisinage infinitésimal $\mathrm{d}x_{\checkmark}$ retenu est sans incidence sur la définition : en effet, si $\mathrm{d}x_{\checkmark}'$ est un autre voisinage infinitésimal de $x_{\checkmark}$, la fonction $\theta \mapsto \mathbb{P}_\theta(X \in \mathrm{d}x_{\checkmark}')$ sera proportionnelle à la fonction $\theta \mapsto \mathbb{P}_\theta(X \in \mathrm{d}x_{\checkmark})$, de sorte que l'ensemble des fonctions qui sont proportionnelles à l'une ou à l'autre sera le même ! ♡

Remarque (TP). La barre verticale dans la notation de la vraisemblance se lit « sachant que », comme pour les probabilités conditionnelles : ce qu'il y a derrière la barre indique un *contexte* (on écrit notre fonction de vraisemblance *au vu de l'observation* $x_{\checkmark}$), tandis que ce qu'il y a devant la barre indique *l'objet* sur lequel porte la vraisemblance (la fonction de vraisemblance ayant pour domaine Θ, $\mathscr{L}(\theta \mid x_{\checkmark})$ est la vraisemblance *de la valeur* θ dans le contexte $x_{\checkmark}$, pas « la vraisemblance du couple $(\theta, x_{\checkmark})$ » : confer aussi remarque (TS) *infra*).

Néanmoins, malgré la similitude de notation et de sens, la barre verticale intervenant dans l'écriture de la vraisemblance ne constitue pas pour autant un « conditionnement » au sens probabiliste du terme : c'est juste une autre forme de contexte ! ⌒ ♧

> *Remarque* (TQ). Dans ce cours, tous les modèles que nous manipulerons admettront toujours des fonctions de vraisemblance, pour toutes les valeurs possibles de l'observation : des tels modèles sont dits « dominés ». Cependant, il existe aussi certains modèles "pathologiques" pour lesquels la vraisemblance ne peut pas être définie... Heureusement, l'existence de tels modèles « non dominés » n'est pas très gênante en pratique, dans la mesure où ils peuvent toujours être approchés par des modèles dominés : on pourra donc faire si la vraisemblance était toujours définie ! ◡ ♧

Comme nous l'explique la proposition ci-dessous, toutes les fonctions de vraisemblance se déduisent les unes des autres par proportionnalité ; c'est pourquoi, en pratique, on parle de « la » fonction de vraisemblance : même si, en réalité, il ne s'agit pas d'une fonction à proprement parler mais plutôt d'une *classe d'équivalence* de fonctions pour la relation de proportionnalité, classe d'équivalence que, par abus de langage, on pourra désigner par n'importe laquelle de ses représentants [†].

! **Proposition** (TR). *La fonction de vraisemblance est définie de façon unique à constante multiplicative près : si $\theta \mapsto L(\theta)$ est une fonction de vraisemblance de notre modèle (associée à l'observation $x_{\checkmark}$), alors une fonction $\theta \mapsto \tilde{L}(\theta)$ sera une fonction de vraisemblance si et seulement si il existe une constante $\alpha \in \mathbb{R}_+^*$ (ni infinitésimalement grande, ni infinitésimalement petite) telle que $\tilde{L}(\theta) = \alpha L(\theta)$ pour tout θ.*

Par conséquent, nous parlerons par abus de langage de « la » vraisemblance ; mais il faut bien comprendre que, quand on écrit « $\mathscr{L}(\theta = \theta \mid X = x_{\checkmark}) = L(\theta)$ », cela signifie en fait « l'application $\theta \mapsto L(\theta)$ est une fonction de vraisemblance associée à l'observation $x_{\checkmark}$; et plus généralement, l'ensemble des fonctions de vraisemblance associées à cette observation est donné par les applications de la forme $\theta \mapsto \alpha L(\theta)$ pour $\alpha \in \mathbb{R}_+^*$ ». ◇

[†]. En fait, vous avez déjà rencontré ce principe dans le passé, lorsqu'on vous a parlé des congruences modulo n en arithmétique. En effet, quand on écrit que, « modulo 7, on a $5+4=2$ », du fait qu'on se place modulo 7, les notations '5', '4' et '2' dans l'égalité se réfèrent en fait aux *classes de congruences* (dans $\mathbb{Z}/7\mathbb{Z}$) des (véritables) entiers resp. 5, 4 et 2 [véritables entiers pour lesquels on a évidemment $5+4 \neq 2$...], càd. aux ensembles $\{5+7k \mid k \in \mathbb{Z}\}$, etc. Et la notion d'addition *entre classes de congruences*, elle, vérifie bien que la somme de la classe de 5 et de la classe de 4 est égale (*rigoureusement* égale, en tant que classe de congruence) à la classe de 2 !

Type d'objet

La définition (TN) et la proposition (TR) ci-dessus nous ont dit que « la vraisemblance associée à l'observation x » est un objet de type « fonction de $\boldsymbol{\Theta}$ dans $\mathbb{R}_+$, vue à constante multiplicative près ». Ce type d'objet mathématique étant quelque peu inhabituel, nous allons dans cette sous-section souligner les conséquences que la nature de la vraisemblance entraine sur la façon de la manipuler.

Un premier point important à conserver à l'esprit est la valeur numérique $\mathscr{L}(\theta \mid x_{\checkmark})$ prise par la vraisemblance en une valeur possible du paramètre caché, dès lors qu'elle est non nulle, n'a aucun sens dans l'absolu. Ainsi une vraisemblance n'est pas petite ou grande *en soi*, mais seulement *en comparaison* avec la vraisemblance associée à une autre valeur du paramètre caché :

Remarque (TS). !

(i) La fonction de vraisemblance n'étant définie qu'à proportionnalité près, la seule chose objective (i.e., indépendante d'un choix de normalisation) qu'on puisse dire sur une valeur spécifique $\mathscr{L}(\theta \mid x_{\checkmark})$ de la vraisemblance est le fait qu'elle soit nulle ou non nulle.

(ii) En revanche, on peut donner un sens non ambigu (j'entends par là, indépendant du choix de la constante de proportionnalité) au *rapport* entre les vraisemblances de deux valeurs possibles du paramètre caché associés à une même valeur de l'observation, càd. aux expressions de la forme $\mathscr{L}(\theta_1 \mid x) \,/\, \mathscr{L}(\theta_0 \mid x)$. ♧

Attention cependant : certes, on peut comparer $\mathscr{L}(\theta_0 \mid x)$ et $\mathscr{L}(\theta_1 \mid x)$; mais par contre, il n'y a pas de sens à comparer des vraisemblances associées à deux valeurs *distinctes* de l'observation :

Remarque (TT). Lorsqu'on choisit une normalisation de la fonction de vraisemblance (càd. qu'on choisit un représentant dans la classe des fonctions vues à proportionnalité près), il n'y a pas de moyen canonique de lier les choix de représentants correspondant à deux valeurs *distinctes* de l'observation. Par conséquent, pour $x_1 \neq x_0$, les valeurs prises par des expressions comme « $\mathscr{L}(\theta \mid x_1) \,/\, \mathscr{L}(\theta \mid x_0)$ » ou « $\mathscr{L}(\theta_1 \mid x_1) \,/\, \mathscr{L}(\theta_1 \mid x_0)$ » dépendront du choix de normalisation : ces expressions n'ont donc *aucun sens objectif* ! ♧

Bien voir aussi que la vraisemblance se rapporte toujours (quoique parfois indirectement, confer définition ?? *infra*) au *paramètre caché* : c'est un concept fondamentalement *statistique*, qui n'a pas d'analogue en théories des probabilités "pures" ! Ainsi :

Remarque (TU). La vraisemblance est un concept qui concerne uniquement *le paramètre caché*, dans un contexte relatif à l'*observation* : mais à part cela, on ne peut pas parler de « vraisemblance d'une variable aléatoire sachant une autre » en général ! Notamment, il n'y aurait *pas de sens* à écrire quelque chose comme « $\mathscr{L}(Y = y \mid X = x_{\checkmark})$ » ou « $\mathscr{L}(X = x \mid \boldsymbol{\theta} = \theta_{\checkmark})$ »...

Un peu plus loin, vous verrez néanmoins une « exception » apparent à cette règle : en effet, les définitions ?? et ?? *infra* étendront la notion de vraisemblance aux *quantités d'intérêt explicatives* [expressions du type « $\mathscr{L}(\gamma(\boldsymbol{\theta} = \gamma_{\bullet} \mid X = x_{\checkmark})$ »], resp. aux *hypothèses* [« $\mathscr{L}(\boldsymbol{\theta} \in \boldsymbol{\Theta}' \mid X = x_{\checkmark})$ »]. Cependant, l'exception ne sera qu'apparente : en effet, ces notions de vraisemblance étendue n'auront de sens que dans la mesure où quantités d'intérêt explicatives et hypothèses sont elles-mêmes

définies *à partir du paramètre caché* ! Il ne s'agira donc pas d'autres instances d'un inexistant concept plus général de soi-disant « vraisemblance d'une v.a. par rapport à une autre », mais simplement de "raccourcis notationnels", qui exprimeront en fait des propriétés de la fonction de vraisemblance *du paramètre caché* lui-même ! ♧

Nuances entre vraisemblance et conditionnement

Après avoir bien précisé quel type d'objet mathématique est la fonction de vraisemblance, je voudrais aussi souligner les différences subtiles, mais néanmoins importantes, que ce concept présente avec deux autres notions proches : à savoir, la probabilité fréquentiste et la probabilité à postériori.

En effet, à la lecture de la définition générale de la vraisemblance (et ce sera encore plus marqué avec le théorème **??** *infra*), on est tenté de se dire : « ah mais, en fait, "$\mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x)$, ça veut juste dire la même chose que $\mathbb{P}(X = x \mid \boldsymbol{\theta} = \theta)$", à ceci près qu'on a inversé les expressions qui apparaissaient de part et d'autres de la barre ! ». Mais si tel était le cas, il n'y aurait pas lieu de consacrer un chapitre complet à ce concept... Soulignons donc bien les nuances qu'il y a entre vraisemblance et probabilité fréquentiste :

Remarque (TV) (Différences entre vraisemblance et probabilité fréquentiste).

(i) Alors que la fonction de masse $\mathbb{P}(X = x \mid \boldsymbol{\theta} = \theta)$ est une fonction de x (qu'on peut considérer pour différentes valeurs de θ), la fonction de vraisemblance $\mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x)$ est au contraire une fonction de θ (qu'on peut considérer pour différentes valeurs de x) !

(ii) En outre, il n'y a que dans le cas où $\mathscr{X}$ est discret que la loi fréquentiste $\mathrm{Loi}(X \mid \boldsymbol{\theta} = \theta)$ peut être assimilée à sa fonction de masse : dans le cas continu, l'expression « $\mathbb{P}(X = x \mid \boldsymbol{\theta} = \theta)$ » vaut systématiquement zéro, et n'a donc aucun intérêt... En réalité, les lois fréquentistes sont fondamentalement des *mesures* sur $\mathscr{X}$, de sorte que les seules expressions véritablement pertinentes les concernant sont celles de la forme « $\mathbb{P}(X \in \mathrm{d}x \mid \boldsymbol{\theta} = \theta)$ ». Rien de tel pour la vraisemblance, où c'est véritablement une *égalité* qu'on considère lorsqu'on écrit le contexte « $X = x$ » à droite de la barre verticale : il n'y a pas une « mesure de contextes » sur $\mathscr{X}$, mais bien un contexte différent pour *chaque* valeur possible de l'observation !

(iii) On notera par ailleurs que l'expression « $\mathbb{P}(X \in \mathrm{d}x \mid \boldsymbol{\theta} = \theta)$ » se réfère le plus souvent à une valeur infinitésimalement petite ; alors que la vraisemblance $\mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x)$, elle, est un "véritable" nombre réel...

(iv) Enfin, n'oublions que, alors que l'expression $\mathbb{P}(X \in \mathrm{d}x \mid \boldsymbol{\theta} = \theta)$ se réfère à une valeur *précise*, définie de façon non ambigüe, la fonction de vraisemblance $\theta \mapsto \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x)$, elle, n'est définie qu'à proportionnalité près... Expliquer le pourquoi du comment de l'importance de cette définition à proportionnalité près serait malheureusement un peu trop ambitieux dans le cadre de ce cours ; mais sachez en tout cas qu'il ne s'agit pas d'un simple *choix* définitionnel qu'on aurait fait "par commodité", mais bien d'une véritable *contrainte*, nécessaire pour que la vraisemblance puisse se manipuler convenablement ! ♧

Un autre concept qui ressemble fortement à la vraisemblance est celui de *postériore*. (Et d'ailleurs, en l'occurrence, même la position des expressions de part et d'autre de la barre verticale est identique...). Mais là encore, il ne faut pas confondre

les deux concepts ! Les différences sont même encore plus fondamentales qu'avec les probabilités fréquentistes, comme l'explique la remarque suivante :

Remarque (TW) (Différences entre vraisemblance et postériore).

(i) La postériore est un concept *bayésien* (on a besoin de spécifier une loi à priori pour lui donner un sens), tandis que la vraisemblance est un concept *fréquentiste* (elle est définie indépendamment du choix d'une priore) !

(ii) Postériore comme vraisemblance portent toutes les deux sur l'espace du paramètre caché $\boldsymbol{\Theta}$: néanmoins, la postériore est une *mesure de probabilité*, tandis que la vraisemblance est une *fonction*. En particulier, lorsque $\boldsymbol{\Theta}$ est continu (avec une priore diffuse), $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} = \theta)$ vaudra systématiquement zéro, et même l'expression « $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta)$ » se réfèrera à une quantité infinitésimalement petite ; tandis que $\mathscr{L}(\boldsymbol{\theta} = \theta)$ sera (en général) non nulle...

(iii) Nous verrons ultérieurement comment on peut étendre la vraisemblance pour donner un sens à des expressions de la forme « $\mathscr{L}(\boldsymbol{\theta} \in \boldsymbol{\Theta}')$ » pour $\boldsymbol{\Theta}' \subseteq \boldsymbol{\Theta}$. Mais attention : du fait que la vraisemblance est une *fonction* et non pas une *mesure* la façon de passer de la vraisemblance "ponctuelle" à la vraisemblance d'une hypothèse n'aura rien à voir avec ce qui se passe pour la postériore : dans le cas de la postériore, la probabilité à postériori $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \boldsymbol{\Theta}')$ s'obtient par *somme* (intégrale), puisqu'elle est égale à $\int_{\theta \in \boldsymbol{\Theta}'} \mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta)$; tandis que pour la vraisemblance, la vraisemblance $\mathscr{L}(\boldsymbol{\theta} \in \boldsymbol{\Theta}')$ d'une hypothèse sera définie comme le *supremum* $\sup_{\theta \in \boldsymbol{\Theta}'} \mathscr{L}(\boldsymbol{\theta} = \theta)$ des sous-hypothèses ponctuelles qui la composent !

(iv) En outre, comme nous l'avons déjà signalé en remarque **??**, la notion de vraisemblance est fondamentalement définie relativement au paramètre caché $\boldsymbol{\theta}$ *lui-même* (par exemple, il n'y a pas de sens à parler de vraisemblance de l'observation future). La postériore, en revanche, n'est qu'une instance de la notion plus générale de « la loi à postériori », dans le cas particulier où on applique cette notion au paramètre caché : mais on peut tout aussi bien considérer la loi à postériori de l'observation future, ou d'une quantité d'intérêt prédictive !

(v) Pour finir, même remarque qu'en (TV)-(iv) : constrairement à l'expression « $\mathbb{P}_{\text{post}}(\boldsymbol{\theta} \in \mathrm{d}\theta)$ », qui se réfère à une valeur bien précise, la fonction de vraisemblance $\theta \mapsto \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x)$, elle, n'est, rappelons-le, définie qu'à proportionnalité près... ♧

Dans l'item (i) de la remarque ci-dessus, nous avons souligné à nouveau un point déjà mentionné dans l'introduction de ce chapitre : à savoir, que la vraisemblance est un concept fondamentalement fréquentiste. Ce point est tellement important que je voudrais le mettre en avant encore une fois, sous la forme d'une remarque spécifiquement dédie :

Remarque (TX). Il résulte directement de la définition (TN) que la vraisemblance d'un modèle statistique dépend uniquement de la donnée de la famille de lois $\theta \mapsto \mathrm{Loi}(X \mid \boldsymbol{\theta} = \theta)$, mais *pas* de la priore $\mathrm{Loi}(\boldsymbol{\theta})$: par conséquent, *la vraisemblance est un contexte fréquentiste*, qui concerve tout son sens même en-dehors du paradigme bayésien ! ♧ **!!**

10.2 Comment calculer la vraisemblance

Comme dit au début de la § 10.1, vous n'êtes pas censés retenir la définition générale (TN) : en pratique, seule la façon de calculer la vraisemblance dans les cas les plus classiques mérite d'être retenue ! Il y a trois résultats à connaitre : celui qui donne la vraisemblance dans le cas discret ; celui qui donne la vraisemblance dans le cadre à densité ; et celui qui permet de combiner des vraisemblances partielles dans le cas indépendant.

! **Théorème (TY)** (Vraisemblance dans le cas discret)**.** *On considère les notations génériques. Supposons que l'espace de l'observation $\mathcal{X}$ soit discret. Alors une fonction de vraisemblance est donnée par*

$$\mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark}) = \mathbb{P}(X = x_{\checkmark} \mid \boldsymbol{\theta} = \theta). \tag{TZ}$$

◇

Démonstration. En effet, dans le cas discret, pour $\mathrm{d}x_{\checkmark}$ un voisinage infinitésimal de $x_{\checkmark}$, $\mathbb{P}_\theta(X \in \mathrm{d}x_{\checkmark})$ est précisément égale à $\mathbb{P}_\theta(X = \mathrm{d}x_{\checkmark})$. ♤

! **Théorème (UA)** (Vraisemblance dans le cas à densité)**.** *On considère les notations génériques. Supposons que l'espace de l'observation $\mathcal{X}$ soit de la forme $\mathbb{R}^n$; et supposons en outre que, pour tout $\theta \in \boldsymbol{\Theta}$, la loi $\mathrm{Loi}_\theta(X \mid \boldsymbol{\theta} = \theta)$ possède une densité par rapport à la mesure de Lebesgue vol_n, densité notée $f_\theta(\bullet) : \mathcal{X} \to \mathbb{R}_+$.*

Alors une fonction de vraisemblance du modèle est donnée par

$$\mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark}) = f_\theta(x_{\checkmark}). \tag{UB}$$

◇

Démonstration. La propriété de densité énonce que, pour toute zone infinitésimale $\mathrm{d}x$ de $\mathcal{X}$ (au niveau du point x), on a

$$\mathbb{P}(X \in \mathrm{d}x \mid \boldsymbol{\theta} = \theta) = f_\theta(x)\,\mathrm{vol}_n(\mathrm{d}x). \tag{UC}$$

Par conséquent, si nous appliquons la définition générale (TN) de la vraisemblance à un voisinage infintésimal $\mathrm{d}x_{\checkmark}$ de $x_{\checkmark}$, on trouve que notre fonction de vraisemblance doit être proportionnelle à $\theta \mapsto f_\theta(x_{\checkmark})\,\mathrm{vol}_n(\mathrm{d}x_{\checkmark})$: la fonction $\theta \mapsto f_\theta(x_{\checkmark})$ convient donc, puisque le facteur « $\mathrm{vol}_n(\mathrm{d}x_{\checkmark})$ » ne dépend pas de θ. [Noter, au passage, que la fonction $\theta \mapsto f_\theta(x_{\checkmark})\,\mathrm{vol}_n(\mathrm{d}x_{\checkmark})$ n'aurait par contre pas été une vraisemblance acceptable, dans la mesure où les valeurs qu'elle prend sont infinitésimalement petites : ce qui est interdit par la définition !]. ♤

! **Théorème (UD)** (Vraisemblance pour un modèle à sous-observations indépendantes)**.** *Soit un modèle statistique (fréquentiste) dont l'espace du paramètre caché est $\boldsymbol{\Theta}$ et dont l'espace de l'observation est de la forme $\mathcal{X}_0 \times \mathcal{X}_1 \cdots \times \mathcal{X}_{n-1} =: \mathcal{X}$, l'observation s'écrivant alors $(X_0, X_1, \ldots, X_{k-1}) =: \vec{X}$: autrement dit, les X_i sont les « sous-observations » dont l'observation est constituée. Supposons que, dans notre modèle, sous les différents contextes fréquentistes $\mathbb{P}_\theta(\bullet)$, les sous-observations soient toujours indépendantes :*

$$\forall \theta \in \boldsymbol{\Theta} \quad \mathrm{Loi}_\theta(\vec{X}) = \mathrm{Loi}_\theta(X_0) \otimes \mathrm{Loi}_\theta(X_1) \otimes \cdots \otimes \mathrm{Loi}_\theta(X_{n-1}). \tag{UE}$$

Supposons enfin que pour tout i, dans le modèle $\theta \mapsto \mathrm{Loi}_\theta(X_i)$ (d'espace du paramètre caché $\boldsymbol{\Theta}$ et d'espace de l'observation $\mathcal{X}_i$) consistant à "ne regarder que la sous-observation X_i", on dispose d'une fonction de vraisemblance associée à $x_{i\checkmark}$, que que nous noterons $\mathscr{L}(\theta \mid X_i = x_{i\checkmark})$.

Alors, dans notre modèle de départ (celui où on considère l'ensemble de toutes les sous-observations), une fonction de vraisemblance associée à l'observation $(x_{0\checkmark}, x_{1\checkmark}, \dots, x_{(n-1)\checkmark}) =: \vec{x}$ *s'obtient en faisant le produit des vraisemblances "partielles" associées à chaque sous-observation :*

$$\mathscr{L}(\theta \mid \vec{X} = \vec{x}_{\checkmark}) = \mathscr{L}(\theta \mid X_0 = x_{0\checkmark}) \times \mathscr{L}(\theta \mid X_1 = x_{1\checkmark}) \times \cdots \times \mathscr{L}(\theta \mid X_{n-1} = x_{(n-1)\checkmark}). \tag{UF}$$

◇

Démonstration. Fixons-nous des voisinages infinitésimaux $\mathrm{d}x_{0\checkmark}, \dots, \mathrm{d}x_{(n-1)\checkmark}$ des différents $x_{i\checkmark}$. Le fait que les $\mathscr{L}(\bullet \mid X_i = \mathrm{d}x_{i\checkmark})$ soient des vraisemblances signifie alors qu'on a, pour tout $\theta \in \boldsymbol{\theta}$,

$$\mathbb{P}(X_i \in \mathrm{d}x_i \mid \boldsymbol{\theta} = \theta) = \alpha_i(\mathrm{d}x_{i\checkmark}) \times \mathscr{L}(\bullet \mid X_i = \mathrm{d}x_{i\checkmark}) \tag{UG}$$

pour un certain $\alpha_i(\mathrm{d}x_{i\checkmark}) \in \mathbb{R}_+^*$ (possiblement infinitésimalement petit). Mais alors on a

$$\begin{aligned}
\mathbb{P}\big((X_0, \dots, X_{n-1}) \in \mathrm{d}x_{0\checkmark} \times \cdots \times \mathrm{d}x_{(n-1)} \;\big|\; \boldsymbol{\theta} = \theta\big) &\underset{\text{déf}}{=} \\
\mathbb{P}_\theta(X_0 \in \mathrm{d}x_{0\checkmark} \text{ et } \dots \text{ et } X_{n-1} &\in \mathrm{d}x_{(n-1)\checkmark}) \\
\underset{\text{ind}^{\text{ce}}}{=} \mathbb{P}_\theta(X_0 \in \mathrm{d}x_{0\checkmark}) \times \dots \times \mathbb{P}_\theta(X_{n-1} &\in \mathrm{d}x_{(n-1)\checkmark}) \\
= \alpha(\mathrm{d}x_{0\checkmark})\, \mathscr{L}(\theta \mid X_0 = x_{0\checkmark}) \times \cdots \times \alpha(\mathrm{d}x_{(n-1)\checkmark})\, &\mathscr{L}(\theta \mid X_0 = x_{(n-1)\checkmark}) \\
= \prod_{i=0}^{n-1} \alpha(\mathrm{d}x_{0\checkmark}) \times \mathscr{L}(\theta \mid X_0 = x_{0\checkmark}) \times \cdots \times &\mathscr{L}(\theta \mid X_0 = x_{(n-1)\checkmark}).
\end{aligned}$$

Cela montre bien que le produit $\mathscr{L}(\theta \mid X_0 = x_{0\checkmark}) \times \cdots \times \mathscr{L}(\theta \mid X_0 = x_{(n-1)\checkmark})$ est une fonction de vraisemblance pour le modèle de départ (au vu de l'observation $\vec{x}_{\checkmark}$) : en effet, $\mathrm{d}x_{0\checkmark} \times \cdots \mathrm{d}x_{(n-1)\checkmark}$ est bien un voisinage infinitésimal de $\vec{x}_{\checkmark}$; le facteur $\prod_{i=0}^{n-1}$ est bien indépendant de θ ; et les valeurs prises par le produit $\mathscr{L}(\theta \mid X_0 = x_{0\checkmark}) \times \cdots \times \mathscr{L}(\theta \mid X_0 = x_{(n-1)\checkmark})$ ne sont effectivement ni infinitésimalement petites, ni infinitésimalement grandes (puisqu'aucune des $\mathscr{L}(\theta \mid X_i = x_{i\checkmark})$ ne l'est). ♤

Remarque (UH). En pratique, il est très courant de se retrouver dans cette situation où l'observation se décompose en plusieurs sous-observations qui sont indépendantes sous les différents contextes fréquentistes : cela recouvre notamment le cas important des modèles d'échantillonnage, confer définition (NE). ♧

Remarque (UI). Dans le cas où les « sous-modèles » $\theta \mapsto \mathrm{Loi}_\theta(X_i)$ sont tous discrets, resp. tous à densité, on peut écrire explicitement la fonction de masse (resp. de densité) de la loi de l'observation gloable comme le produit des fonctions de masse (resp. de densité) des lois de chaque sous-observation ; et alors on peut calculer la vraisemblance en utilisant uniquement le théorème (TY) (resp. le théorème (UA)). Néanmoins, le théorème (UD) s'avère incontournable si *certains* des sous-modèles sont discrets tandis que *certains* autres sont à densité : car dans ce cas, le modèle global n'est ni discret, ni à densité ! ♧

10.3 Extensions du concept de vraisemblance

Vraisemblance d'une hypothèse

L'importance du concept de vraisemblance dans le cadre fréquentiste justifie d'en introduire diverses variantes et extensions.

Bien que, comme nous l'avons souligné en remarque (TU), la définition de la vraisemblance porte fondamentalement sur le paramètre caché *lui-même* (et non pas sur la quantité d'intérêt — y compris en inférence explicative) ; il peut être

utile[‡] de définir un concept de « vraisemblance d'une hypothèse ». Il s'avère qu'en l'occurrence, la façon appropriée de définir la vraisemblance d'une hypothèse est de prendre le *supremum* de la fonction de vraisemblance sur l'ensemble des valeurs individuelles du paramètre caché composant cette hypothèse :

! **Définition (UJ).** On prend les notations génériques. Pour $\Theta' \subseteq \Theta$, on définit la vraisemblance de l'hypothèse $\{\theta \in \Theta'\}$[§] comme

$$\mathscr{L}(\theta \in \Theta' \mid X = x_{\checkmark}) \coloneqq \sup_{\theta \in \Theta'} \mathscr{L}(\theta = \theta \mid X = x_{\checkmark}). \tag{UK}$$

♡

Remarque (UL). Notez qu'il s'agit d'une *définition*, pas d'un théorème : jusqu'ici, la notion de « vraisemblance d'une hypothèse » n'avait en effet pas encore de sens... ! Expliquer pourquoi c'est bien un supremum qu'il convient de considérer en l'occurrence, plutôt qu'autre chose, dépasserait malheureusement de cadre pédagogique ce ce cours : mais en tout cas, sachez qu'il y a une bonne raison à ce que la vraisemblance d'une hypothèse soit définie ainsi ! ☺ ♧

Remarque (UM). De même que pour la vraisemblance du paramètre caché, la vraisemblance d'une hypothèse n'est définie qu'à constante multiplicative près : seul le *rappor* de vraisemblance entre deux hypothèses, pour une *même* valeur de l'observation, est défini de façon univoque ! ♧

On pourra remarquer que, via la notion de vraisemblance pour une hypothèse, on obtient même une notion de fonction de vraisemblance pour les quantités d'intérêt explicatives :

Définition (UN). Pour $\gamma =: \gamma(\theta)$ une quantité d'intérêt explicative à valeurs dans $\mathscr{G}$, on pourra appeler *fonction de vraisemblance de* γ (pour l'observation $x_{\checkmark}$) l'application :

$$\begin{aligned} &\mathscr{G} \to \mathbb{R}_+ \\ &\gamma_* \mapsto \mathscr{L}(\gamma = \gamma_* \mid X = x_{\checkmark}) : \end{aligned} \tag{UO}$$

dans cette formule, "$\mathscr{L}(\gamma = \gamma_*)$" se refère à la vraisemblance de l'hypothèse $\{\gamma = \gamma_*\}$, elle-même définie via la définition (UJ) : l'évènement $\{\gamma = \gamma_*\}$ étant bien une hypothèse, puisque γ désigne en fait $\gamma(\theta)$, de sorte que $\{\gamma = \gamma_*\}$ se produit si et seulement si θ appartient au sous-ensemble $\{\theta \in \Theta \mid \gamma(\theta) = \gamma_*\}$.

De même que pour la fonction de vraisemblance du paramètre caché lui-même, la fonction de vraisemblance d'une quantité d'intérêt explicative n'est définie qu'à proportionnalité près, et ne permet pas de comparaison entre deux cas relatifs à des valeurs différentes de l'observation passée. ♡

Attention néanmoins à un piège : bien que nous ayons étendu le vocabulaire de la « vraisemblance » pour l'appliquer à des hypothèses ou des quantités d'intérêt explicatives, le théorème de Bayes, lui, ne se généralise pas dans ce cadre :

Remarque (UP). Attention ! La formule de Bayes ne marche pas avec des hypothèses : dans le cadre bayésien, pour $\Theta', \Theta'' \subseteq \Theta$, on aura en général

$$\frac{\mathbb{P}_{\text{post}}(\theta \in \Theta'')}{\mathbb{P}_{\text{post}}(\theta \in \Theta')} \neq \frac{\mathscr{L}(\theta \in \Theta'')\,\mathbb{P}_{\text{pr}}(\theta \in \Theta')}{\mathscr{L}(\theta \in \Theta'')\,\mathbb{P}_{\text{pr}}(\theta \in \Theta')} \,! \tag{UQ}$$

[‡]. La notion de vraisemblance d'une hypothèse sera en particulier utile pour construire des *statistiques de test* : confer chapitre 12 à venir.

[§]. Rappelons ici que n'importe quelle hypothèse peut de mettre sous la forme $\{\theta \in \Theta\}$: confer définition (RN).

Et cela ne fonctionne pas non plus avec les quantités d'intérêt explicatives : en général,

$$\frac{\mathbb{P}_{\text{post}}(\boldsymbol{\gamma} \in \mathrm{d}\gamma_1)}{\mathbb{P}_{\text{post}}(\boldsymbol{\gamma} \in \mathrm{d}\gamma_0)} \neq \frac{\mathscr{L}(\boldsymbol{\gamma} = \gamma_1)\, \mathbb{P}_{\text{pr}}(\boldsymbol{\gamma} \in \mathrm{d}\gamma_1)}{\mathscr{L}(\boldsymbol{\gamma} \in \mathrm{d}\gamma_0)\, \mathbb{P}_{\text{pr}}(\boldsymbol{\gamma} \in \mathrm{d}\gamma_0)}. \tag{UR}$$

♧

Log-vraisemblance

Dans un certain nombre de cas (voir remarque (UW) ci-dessous), plutôt que de manipuler la vraisemblance elle-même, il est préférable de manipuler son *logarithme*. On parle alors de *log-vraisemblance* :

Définition (US) (Log-vraisemblance)**.** Prenons les notations génériques. On appelle *fonction de log-vraisemblance* (associée à l'observation effective $x_{\checkmark}$) la fonction : !

$$\begin{aligned} \boldsymbol{\Theta} &\to \mathbb{R} \sqcup \{-\infty\} \\ \theta &\mapsto \log \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x_{\checkmark}). \end{aligned} \tag{UT}$$

On parlera plus précisément de ln-vraisemblance, de lg-vraisemblance, etc. pour signifier que le logarithme est considéré en base e, en base 10, etc.

La définition s'étend aussi, *mutatis mutandis*, en une notion de log-vraisemblance pour une hypothèse ou une quantité d'intérêt explicative. ♡

Remarque (UU). Les vraisemblances étant définies à proportionnalité près, les log-vraisemblances sont définies *à translation verticale près* : seule la *différence* entre deux log-vraisemblances (associées à la *même* valeur du paramètre caché) possède une valeur numérique non ambigüe ! ♧

Remarque (UV). On notera que, lors du passage au logarithme, le théorème (UD) (sur la multiplicativité de la vraisemblance dans le cas de sous-observations indépendantes) se traduit par une propriété d'*additivité* concernant les log-vraisemblances dans ce cas. ♧

Remarque (UW). Parler en termes de log-vraisemblance, plutôt que de vraisemblance elle-même, s'avère particulièrement utile dans les cas suivants :

1. Lorsque les jeux de données considérés commencent à être grand, les formules numériques pour va vraisemblance vont facilement fournir des nombres absoluement minuscules (comme $10^{-1\,000}$ ou gigantesques (comme $10^{1\,000}$)... Or, du point de vue de l'implémentation numérique, on est alors confronté à un grave souci, car il y a *dépassement de capacité* de la façon standard dont les ordinateurs représentent les nombres réels : lorsque l'ordinateur trouve un nombre inférieur à 10^{-324} en valeur absolue, il le représente en interne comme valant *exactement* zéro ; tandis que les valeurs supérieures à 10^{309} sont considérées comme infinies [¶]... Travailler en logarithmes évite ces soucis.
2. Plus généralement, les variations de $\mathscr{L}(\theta)$ lorsque θ se déplace dans $\boldsymbol{\Theta}$ ont tendance à couvrir facilement plusieurs ordres de grandeur (voire des centaines d'ordre de grandeur) : la log-vraisemblance est alors largement préférable pour comparer efficacement les valeurs — en particulier dans une perspective de visualisation, d'une part, pour pouvoir distinguer les vraisemblances valant 10^{-4} de celles valant 10^{-8} !
3. Puisque, dans les modèles à sous-observations indépendantes (qui sont fréquents en pratique), la log-vraisemblance se comporte de façon additive, cela rend alors les calculs bien plus faciles à exprimer en termes de log-vraisemblances : du point de vue numérique, d'une part, comme évoqué dans

[¶]. Ce qui peut même faire planter la machine !

les items précédents, mais aussi pour les calculs théoriques, par exemple lorsqu'on souhaitera dériver la fonction de (log-)vraisemblance pour trouver en quel point elle est maximale ! (confer § **??**).

4. Enfin, à un niveau encore plus fondamental, de nombreuses notions de théorie de l'information font apparaitre naturellement la *log*-vraisemblance plutôt que la vraisemblance elle-même : c'est en particulier le cas de l'*information de Fisher*, très utile en statistique asymptotique, qui se définit comme la dérivée seconde de la ln-vraisemblance. (Et on pourrait citer aussi les critères de sélection de modèles, comme AIC et BIC — que vous verrez l'année prochaine en analyse de données —, qui se formulent naturellement en termes de log-vraisemblances également).

♧

10.4 Deux propriétés remarquables de la vraisemblance

Pour terminer ce chapitre, je voudrais présenter deux propriétés que je trouve particulièrement intéressantes concernant la vraisemblance : s'il ne sera pas nécessaire de la connaitre ou de les utiliser dans le cadre de ce cours, je trouve néanmoins qu'elles montrent particulièrement bien toute la puissance du concept de vraisemblance ! ◡̈

Invariance par reparamétrisation

Proposition (UX). *Soit* $\big(\Theta, \mathcal{X}, \theta \mapsto \mathrm{Loi}(X \mid \boldsymbol{\theta} = \theta)\big)$ *un modèle statistique, et soit* $f\colon \Xi \to \Theta$ *une application. On peut alors déduire de notre modèle un autre modèle statistique dont l'espace du paramètre caché serait* Ξ *(le paramètre caché lui-même étant alors noté* $\boldsymbol{\xi}$*), ayant également* $\mathcal{X}$ *pour espace de l'observation, par*

$$\mathrm{Loi}(X \mid \boldsymbol{\xi} = \xi) \coloneqq \mathrm{Loi}(X \mid \boldsymbol{\theta} = f(\xi)) \text{ :} \tag{UY}$$

cela revient en substance à re-paramétrer le modèle par ξ *plutôt que par* θ*, les paramètres* ξ *et* θ *étant liés par la relation* $\theta = f(\xi)$*. Dans ce cas, la fonction de vraisemblance pour le modèle re-paramétré se déduit immédiatement de la fonction de vraisemblance pour le modèle de base : pour tout* $x \in \mathcal{X}$*,*

$$\mathscr{L}(\boldsymbol{\xi} = \xi \mid X = x) = \mathscr{L}(\boldsymbol{\theta} = f(\xi) \mid X = x). \tag{UZ}$$

♢

Cette proposition, en dépit de son allure technique, découle immédiatement de la définition de la vraisemblance. La stabilité par re-paramétrisation de l'observation, en revanche, est plus inattendue ; c'est la proposition suivante :

Proposition (VA). *Soit* $\big(\Theta, \mathcal{X}, \theta \mapsto \mathrm{Loi}(X \mid \boldsymbol{\theta} = \theta)\big)$ *un modèle statistique, et soit* $f\colon \mathcal{X} \to \tilde{\mathcal{X}}$ *une application. On peut alors déduire de notre modèle un autre modèle statistique dont l'espace de l'observation serait* $\tilde{\mathcal{X}}$ *(l'observation elle-même étant alors notée* $\tilde{X}$*), ayant également* Θ *pour espace du paramètre caché, par*

$$\mathrm{Loi}(\tilde{X} \mid \boldsymbol{\theta} = \theta) \coloneqq f_* \,\mathrm{Loi}(X \mid \boldsymbol{\theta} = \theta) \underset{\text{déf}}{=} \mathrm{Loi}(f(X) \mid \boldsymbol{\theta} = \theta) \text{ :} \tag{VB}$$

cela revient en substance à re-paramétrer l'observation x *en* $\tilde{x}$*,* x *et* $\tilde{x}$ *étant liées par la relation* $\tilde{x} = f(x)$*. Supposons que* f *soit injective ; elle admet alors sur* $f(\mathcal{X})$ *une unique rétraction* f^{-1}*. Dans ce cas, la fonction de vraisemblance pour le modèle re-paramétré se déduit de la fonction de vraisemblance du modèle de base par la relation :*

$$\mathscr{L}(\boldsymbol{\theta} = \theta \mid \tilde{X} = \tilde{x}) = \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = f^{-1}(\tilde{x})), \tag{VC}$$

valable pour tout $\tilde{x} \in f(\mathcal{X})$ [‖]*.* ♢

[‖]. La condition « $\tilde{x} \in f(\mathcal{X})$ » n'est en réalité pas restrictive, car, quelle que soit la valeur de $\theta_\checkmark$, on aura presque-surement $\tilde{X} \in f(\mathcal{X})$ sous la loi $\mathbb{P}(\bullet \mid \boldsymbol{\theta} = \theta_\checkmark)$.

Remarque (VD). Attention, l'injectivité de f est essentielle pour que ce théorème soit valide ! ♧

Démonstration. Moralement, l'idée de la preuve est de dire que, notant $\mathrm{d}\tilde{x}$ un voisinage infinitésimal de $\tilde{x}$, l'évènement $\{\tilde{X} \in \mathrm{d}\tilde{x}\}$ est égal à l'évènement $\{f^{-1}(\tilde{X}) \in f^{-1}(\mathrm{d}\tilde{x})\}$, et que l'injectivité de f assure, d'une part que $f^{-1}(\tilde{X})$ suit la loi de X, et d'autre part que $f^{-1}(\mathrm{d}\tilde{x})$ est un voisinage infinitésimal de $f^{-1}(\tilde{x})$... ♤

Non-nullité de la vraisemblance

En toute généralité, une vraisemblance peut être positive *ou nulle*. Des situations de vraisemblance nulle peuvent se rencontrer, par exemple, lorsqu'on a un modèle dans lequel l'erreur est bornée : ainsi, dans le modèle dont le paramètre caché est $(\boldsymbol{\mu}, \boldsymbol{\varepsilon}) \in \mathbb{R}\times\mathbb{R}_+^*$ et dont l'observation, à valeur dans $\mathbb{R}$, suit (sous $\mathbb{P}_{\checkmark}$) la loi $\mathrm{Unif}^{\mathrm{me}}(\mu_{\checkmark}-\varepsilon_{\checkmark}, \mu_{\checkmark}+\varepsilon_{\checkmark})$, alors une fonction devraisemblance est donnée par

$$\mathscr{L}(\mu, \varepsilon \mid x_{\checkmark}) = \frac{\mathbf{1}_{\mu-\varepsilon<x}\mathbf{1}_{\mu+\varepsilon>x}}{\varepsilon} \ : \tag{VE}$$

il y a donc tout une partie de l'espace du paramètre caché qui aura une vraisemblance nulle !

On peut se demander si « vraisemblance nulle » implique « valeur *rigoureusement* impossible pour le paramètre caché » (et pas juste « très improbable »). Bien que cela semble assez intuitif, cela n'est pas si évident, après tout. En effet, pensez au paradoxe suivant, qu'on rencontre en théorie des probabilité : chaque fois qu'on tire une loi diffuse (une loi $\mathrm{Unif}^{\mathrm{me}}(0,1)$, par exemple), la valeur qu'on obtiendra sera un certain nombre réel ; et ce nombre réel, en tant que valeur bien spécifique, avait *forcément* une probabilité rigoureusement nulle de sortir lorsqu'on a procédé à notre tirage...

Les vraisemblances, néanmoins, ne présentent pas ce paradoxe : lorsqu'une vraisemblance est nulle, cela veut dire qu'on peut *complètement* exclure la valeur correspondante du paramètre caché. C'est ce qu'énonce le théorème suivant :

Théorème (VF). *Considérons un modèle statistique avec les notations génériques, pour lequel, pour tout $x \in \mathcal{X}$, nous disposons d'une fonction de vraisemblance $\theta \mapsto \mathscr{L}(\boldsymbol{\theta} = \theta \mid X = x)$. (Cette dernière expression sera notée plus simplement « $\mathscr{L}(\theta \mid x)$ » dans la suite de l'énoncé).*

Alors, pour tout $\theta \in \boldsymbol{\Theta}$, $\mathbb{P}_\theta$-presque-surement, on aura une observation pour laquelle θ sera de vraisemblance non nulle :

$$\forall \theta \in \boldsymbol{\Theta} \quad \mathbb{P}^{x\sim\mathrm{Loi}_\theta(X)}\big(\mathscr{L}(\theta \mid x) > 0\big) = 1 \ [**]. \tag{VG}$$

♢

Démonstration. Si x et θ sont des valeurs pour lesquelles $\mathscr{L}(\theta \mid x) = 0$, alors le théorème de Bayes nous assure que, dans une version bayésienne de notre modèle, si l'observation effective vaut x, la probabilité à postériori que $\boldsymbol{\theta}$ vaille θ sera nulle (et ce, peu importe le choix de la priore), ce qu'on peut encore écrire en disant que, dans ce cas, il sera presque-certain à postériori que $\{\boldsymbol{\theta} \neq \theta\}$:

$$\mathscr{L}(\theta \mid x) = 0 \quad \Longrightarrow \quad \mathbb{P}(\boldsymbol{\theta} \neq \theta \mid X = x) = 1. \tag{VH}$$

Maintenant, pour $\theta \in \boldsymbol{\Theta}$, considérons une version bayésienne de notre modèle statistique où la priore est prise égale à δ_θ : autrement dit, dès la phase à priori, nous avons la *certitude* que $\{\boldsymbol{\theta} = \theta\}$. Il est intuitivement clair que cette certitude "devrait" alors rester variable à postériori. Pour le voir formellement, on va décomposer l'évènement $\{\boldsymbol{\theta} \neq \theta\}$ selon les valeurs possibles de x : on obtient que

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \neq \theta) = \int_{x\in\mathcal{X}} \mathbb{P}_{\mathrm{pr}}(X \in \mathrm{d}x)\, \mathbb{P}(\boldsymbol{\theta} \neq \theta \mid X = x). \tag{VI}$$

[**]. Le formalisme de cette formule est un peu compliqué à comprendre... En fait, ce que (VG) veut dire, c'est la chose suivante : un véritable contexte probabiliste $\mathbb{P}_{\checkmark}$ ayant été fixé (correspondant à une vraie valeur $\theta_{\checkmark}$ du paramètre caché), si dans un premier temps on tire une observation x sous la loi $\mathrm{Loi}_{\theta_{\checkmark}}(X)$, puis que dans un second temps on regarde ce que vaut, *au point* $\theta_{\checkmark}$, la fonction de vraisemblance pour le x que nous venons de tirer, alors on obtiendra ($\mathbb{P}_{\checkmark}$-presque)-toujours une valeur non nulle.

Appelant $N_\theta \coloneqq \{x \in \mathcal{X} \mid \mathscr{L}(\theta \mid x) = 0\}$ l'ensemble des observations qui donnent une vraisemblance nulle à θ, on en déduit en particulier que

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \neq \theta) \geqslant \int_{x \in N_\theta} \mathbb{P}_{\mathrm{pr}}(X \in \mathrm{d}x) \underbrace{\mathbb{P}(\boldsymbol{\theta} \neq \theta \mid X = x)}_{=1}$$
$$= \int_{x \in N_\theta} \mathbb{P}_{\mathrm{pr}}(X \in \mathrm{d}x) = \mathbb{P}_{\mathrm{pr}}(X \in N_\theta). \quad (\mathtt{VJ})$$

Mais si nous prenons $\mathrm{Loi}_{\mathrm{pr}}(\boldsymbol{\theta}) \leftarrow \delta_\theta$, le contexte probabiliste à priori correspond par construction au contexte probabiliste $\mathbb{P}_\theta$! Le membre de gauche de notre inégalité vaut alors 0, tandis que le membre de droite vaut $\mathbb{P}_\theta(X \in N_\theta)$. On a donc $\mathbb{P}_\theta(X \in N_\theta) \leqslant 0$, d'où en fait $\mathbb{P}_\theta(X \in N_\theta) = 0$: ce qui est bien le résultat annoncé ! ♤

Chapitre 11

Estimation et prédiction en statistique fréquentiste

11.1 Prolégomènes

Notations pour la statistique fréquentiste

Le présent chapitre s'inscrivant de plain-pied dans le paradigme fréquentiste de la statistique inférentielle, il est utile de rappeler au préalable quelques usages notationnels relatifs à ce paradigme, notamment concernant la notion de « véritable contexte probabiliste » :

Remarque (VK). Dans le cadre fréquentiste, il n'y a plus de notions de contextes probabilistes à priori ou à postériori : seuls les contextes fréquentistes (i.e. les contextes sachant une valeur donnée du paramètre caché) restent bien définis. Pour cette raison, il n'y aura plus vraiment d'intérêt à utiliser les notations conditionnelles « $\mathbb{P}(\bullet \mid \boldsymbol{\theta} = \theta)$ », « $\mathbb{E}(Z \mid \boldsymbol{\theta} = \theta)$ », etc. : à la place, nous privilégierons les notations à indices « $\mathbb{P}_\theta(\bullet)$ », « $\mathbb{E}_\theta(Z)$ », etc. ♧

Remarque (VL). Rappelons que, parmi les contextes probabilistes fréquentistes, le contexte $\mathbb{P}_{\theta_\checkmark}(\bullet)$ associé à la *vraie* valeur du paramètre caché sera appelé « véritable contexte probabiliste » : on pourra également l'abréger par la notation '$\mathbb{P}_\checkmark(\bullet)$' [∗]. ♧

La notation '$\mathbb{P}_\checkmark$' sera en particulier utilisée pour un "truc" rédactionnel que j'utiliserai très régulièrement dans mes corrigés d'exercices, et qu'il est donc important de connaitre :

Point (VM). Souvent nous serons amené à établir une certaine propriété relative au véritable contexte probabiliste $\mathbb{P}_\checkmark(\bullet)$, puis à faire remarquer que, dans la mesure où cette propriété a été établie sans utiliser aucune connaissance spécifique sur la valeur $\theta_\checkmark$, cela signifie qu'elle est en fait valable en remplaçant $\theta_\checkmark$ par n'importe quelle valeur $\theta \in \Theta$. Par exemple, pour $\boldsymbol{\gamma} =: \gamma(\boldsymbol{\theta})$ une certaine quantité d'intérêt, si on parvient à établir, sans rien savoir spécifiquement sur la valeur $\theta_\checkmark$, qu'on a $\mathbb{E}_\checkmark(\hat{\gamma}(X)) = \gamma_\checkmark$, cela signifiera en fait qu'on a montré l'égalité $\mathbb{E}_\theta(\hat{\gamma}(X)) = \gamma(\theta)$ [†] !

[∗]. Attention, le véritable contexte probabiliste $\mathbb{P}_\checkmark(\bullet)$ signifie donc « $\mathbb{P}(\bullet \mid \boldsymbol{\theta} = \theta_\checkmark)$ » : ce n'est pas la même chose que « $\mathbb{P}(\bullet \mid X = x_\checkmark)$ », qui correspondrait pour sa part au contexte à postériori !

[†]. En utilisant que, puisque la v.a. $\boldsymbol{\gamma}$ désigne en fait $\gamma(\boldsymbol{\theta})$, sa réalisation $\gamma_\checkmark$ correspond en l'occurrence à $\gamma(\theta_\checkmark)$.

pour tout θ. (Pour un cas concret de mise en application de cette astuce rédactionnelle, voir p.ex. la remarque (VW) *infra*). ♧

Enjeux de l'estimation en contexte fréquentiste

Les notions d'estimation et de prédiction ont déjà été rencontrées dans la § 9.5 du polycopié, où le but était de les investiguer dans le cadre bayésien. Rappelons donc les définitions de ces concepts (définitions (ST) et (SV)) :

! **Définition (VN).**

(i) Pour $\theta \mapsto \mathrm{Loi}_\theta(X)$ un modèle statistique explicatif avec les notations génériques, pour $\gamma(\boldsymbol{\theta}) =: \boldsymbol{\gamma}$ une quantité d'intérêt explicative à valeurs dans un espace $\mathcal{G}$, un *estimateur* de $\gamma(\boldsymbol{\theta})$ est une statistique $\hat{\gamma}(X) =: \hat{\boldsymbol{\gamma}}$ à valeurs dans $\mathcal{G}$, visant à "reconstruire" $\gamma(\theta_{\checkmark})$ à partir de l'observation $x_{\checkmark}$. On parle d'*estimation* (de $\gamma(\boldsymbol{\theta})$) pour désigner la réalisation d'un estimateur (autrement dit, la valeur $\hat{\gamma}(x_{\checkmark})$).

(ii) Pour $\theta \mapsto \mathrm{Loi}_\theta(X, Y)$ un modèle statistique prédictif avec les notations génériques, pour $g(Y) =: \boldsymbol{G}$ une quantité d'intérêt prédictive à valeurs dans un espace $\mathcal{G}$, un *prédicteur* de $g(Y)$ est une statistique $\hat{g}(X) =: \hat{\boldsymbol{G}}$ à valeurs dans $\mathcal{G}$, visant à "deviner" $g(y_{\checkmark})$ à partir de l'observation passée $x_{\checkmark}$. On parle de *prédiction* (de $g(Y)$) pour désigner la réalisation d'un estimateur (autrement dit, la valeur $\hat{g}(x_{\checkmark})$). ♡

Dans le paradigme bayésien, les développements sur les estimateurs et prédicteurs se sont avérés particulièrement simples : en effet, on disposait d'une façon "balisée" de construire des estimateurs ou prédicteurs, qui permettait de trouver les statistiques optimales par rapport à des fonctions de perte données. Dans le paradigme fréquentiste, en revanche, la situation va s'avérer sensiblement plus complexe : car les critères de qualité des estimateurs et prédicteurs seront nettement plus flous ; et cela nous amènera en outre à envisager des techniques bien plus diverses pour construire des statistiques pertinentes...

C'est donc à cette approche fréquentiste de l'estimation et de la prédiction que nous allons consacrer la suite de ce chapitre. Nous présenterons tout d'abord des critères pour jauger de la "qualité" d'un estimateur ou d'un prédicteur (§§ **??** et **??**) ; puis nous aborderons les grands principes qui peuvent être suivis pour trouver des estimateurs ou des prédicteurs satisfaisant de telles « bonnes » propriétés (§§ **??** à **??**).

On notera au passage que, alors qu'il n'y avait aucune différence essentielle quant à la façon de définir et d'utiliser les estimateurs et les prédicteurs dans le paradigme bayésien, il n'en ira pas de même dans le cadre fréquentiste : de nombreuses méthodes de construction, en particulier, concerneront spécifiquement les estimateurs, tandis que les prédicteurs ne pourront le plus souvent être construits qu'indirectement, au travers d'estimateurs ! (confer § **??**).

11.2 Fonctions de risque et de biais

Fonction de risque

Dans la § 9.5, nous avons introduit la notion de *risque* pour jauger la qualité d'un estimateur ou d'un prédicteur : le risque étant défini comme l'espérance (sous

tel ou tel contexte probabiliste, en fonction du type de risque considéré) d'une certaine *perte* quantifiant l'inadéquation entre la statistique destinée à approcher la quantité d'intérêt considérée et la quantité d'intérêt elle-même.

En l'occurrence, puisque le présent chapitre s'inscrit dans le paradigme fréquentiste, « risque » appropriée en l'occurrence sera celle de *risque fréquentiste*, qui consiste en une certaine *fonction* définie sur l'espace du paramètre caché. Rappelons la définition de cette fameuse « fonction de risque fréquentiste » :

Définition (VO) (Fonction de risque fréquentiste)**.** Considérons un modèle avec les notations génériques. Soit $g(Y) =: G$ une quantité d'intérêt prédictive à valeurs dans un certain espace $\mathcal{G}$; et soit $\ell : \mathcal{G} \times \mathcal{G} \to \mathbb{R}_+$ une fonction de perte, $\ell(g, \hat{g})$ quantifiant à quel point il est "grave" de prédire que la quantité d'intérêt va valoir $\hat{g}$ lorsqu'elle s'avèrera en réalité valoir g.

Dans ce contexte, pour $\hat{g}(X) =: \hat{G}$ un prédicteur visant à approcher la quantité d'intérêt G, la *fonction de risque fréquentiste* du prédicteur $\hat{G}$ est la fonction $R^{\text{fréq}}_{\hat{G}} : \Theta \to \mathbb{R}_+$ [on pourra omettre l'exposant "fréq" en l'absence d'ambigüité] qui donne l'espérance de la perte sous les différents contextes fréquentistes :

$$R^{\text{fréq}}_{\hat{G}}(\theta) \coloneqq \mathbb{E}_\theta\big(\ell(G, \hat{G})\big) \underset{\text{déf}}{=} \mathbb{E}_\theta\big(\ell(g(Y), \hat{g}(X))\big). \tag{VP}$$

♡

J'ai donné la définition ci-dessus dans le cadre d'un prédicteur ; mais on a la même définition, *mutatis mutandis*, pour un estimateur :

Définition (VQ). Toujours avec les notations génériques, si $\hat{\gamma}(X) =: \hat{\boldsymbol{\gamma}}$ est un estimateur de la quantité d'intérêt explicative $\gamma(\theta) \coloneqq \boldsymbol{\gamma}$ (l'espace dans lequel vivent $\boldsymbol{\gamma}$ et $\hat{\boldsymbol{\gamma}}$ étant noté $\mathcal{G}$), alors la fonction de risque (fréquentiste) de cet estimateur (par rapport à la perte $\ell(\bullet, \bullet)$ sur $\mathcal{G} \times \mathcal{G}$) est

$$\begin{aligned} R^{\text{fréq}}_{\hat{\boldsymbol{\gamma}}} &: \Theta \to \mathbb{R}_+ \\ &\theta \mapsto \mathbb{E}_\theta\big(\ell(\boldsymbol{\gamma}, \hat{\boldsymbol{\gamma}})\big) \underset{\text{déf}}{=} \mathbb{E}_\theta\big(\ell(\gamma(\theta), \hat{\gamma}(X))\big). \end{aligned} \tag{VR}$$

♡

Exemple (VS). Considérons le modèle du chasseur avec $n = 1$ (!), dans lequel nous souhaitons estimer la fiabilité θ. Dans ce modèle très simple, l'observation X ne peut prendre que deux valeurs, à savoir 0 ou 1) : par conséquent, pour décrire un estimateur, il suffit de préciser les valeurs qu'il prend resp. lorsqu'on observe $\{X = 0\}$ et lorsqu'on observe $\{X = 1\}$. Ci-après, pour $\hat{\theta}_0, \hat{\theta}_1 \in [0, 1]$, j'utiliserai la notation $\hat{\boldsymbol{\theta}}^{\hat{\theta}_0, \hat{\theta}_1}$ pour désigner la statistique qui vaut resp. $\hat{\theta}_0$ lorsque $\{X = 0\}$ et $\hat{\theta}_1$ lorsque $\{X = 1\}$. Je considèrerai plus spécifiquement trois estimateurs particuliers : $\hat{\boldsymbol{\theta}}^{0,1}$, qui n'est autre que X elle-même [‡] ; $\hat{\boldsymbol{\theta}}^{1/4, 3/4)}$, qui peut aussi s'écrire $(X + 1/2)/2$; et $\hat{\boldsymbol{\theta}}^{1/8, 5/8}$.

Pour jauger de la qualité de tels estimateurs, il me faut choisir une fonction de perte $\ell(\bullet, \bullet)$ sur $]0, 1[\times [0, 1]$, qui dira à quel point je considère comme "grave" d'estimer une véritable fiabilité θ par une estimation $\hat{\theta}$ qui s'en éloigne plus ou moins : en l'occurrence, je choisirai pour perte l'« erreur absolue » $\ell(\theta, \hat{\theta}) \coloneqq |\hat{\theta} - \theta|$.

[‡]. Dans ce cas, l'estimateur vit dans l'intervalle *fermé* $[0, 1]$, qui est légèrement plus gros que l'espace $]0, 1[$ dans lequel vit la quantité d'intérêt θ. Cela arrive parfois : mais cela ne change fondamentalement rien à la théorie, à part qu'à certains endroits on doit remplacer l'espace $\mathcal{G}$ de la quantité d'intérêt par le sur-espace $\bar{\mathcal{G}}$ dans lequel vit l'estimateur ! ☊

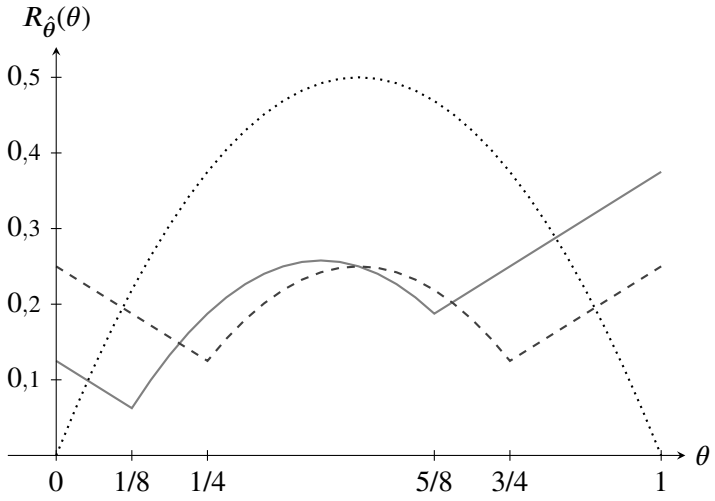

FIGURE 11.1 – Fonctions de risques de trois estimateurs de $\boldsymbol{\theta}$ pour le modèle du chasseur avec un seul tir, pour la fonction de perte « erreur absolue ». La courbe en ligne continue correspond à l'estimateur $\hat{\boldsymbol{\theta}}^{1/8,5/8}$ (qui, par définition, vaut 1/8 quand le chasseur a manqué son plateau, resp. 5/8 quand il l'a touché) ; la courbe tiretée correspond à l'estimateur $\hat{\boldsymbol{\theta}}^{1/4,3/4}$; et la courbe en pointillés correspond à l'estimateur $\hat{\boldsymbol{\theta}}^{0,1}$.

Cette fonction de perte étant choisie, calculons la fonction de risque de l'estimateur $\hat{\boldsymbol{\theta}}^{\hat{\theta}_0,\hat{\theta}_1}$. Nous allons utiliser l'astuce de rédaction consistant à faire les calculs sous la vraie probabilité $\mathbb{P}_{\checkmark}$, mais sans préciser qui est $\theta_{\checkmark}$, pour sous-entendre implicitement toutes les valeurs possibles de θ (cf. point (VM) dans la § 11.1). Sous la loi $\mathbb{P}_{\checkmark}$, il y a une probabilité $\theta_{\checkmark}$ qu'on observe $X = 1$ et qu'on ait donc $\hat{\boldsymbol{\theta}} = \hat{\theta}_1$, générant dans ce cas une perte de $|\hat{\theta}_1 - \theta_{\checkmark}|$), et une probabilité qu'on observe $X = 0$, générant alors de même une perte de $|\hat{\theta}_0 - \theta_{\checkmark}|$. En sommant ces deux cas, on en conclut que la perte moyenne de notre estimateur sous la vraie loi vaut $\theta_{\checkmark}|\hat{\theta}_1 - \theta_{\checkmark}| + (1 - \theta_{\checkmark})|\hat{\theta}_0 - \theta_{\checkmark}|$; et comme nous avons établi cela sans connaissance spécifique sur la valeur $\theta_{\checkmark}$, cela signifie en fait que la fonction de risque de notre estimateur $\hat{\boldsymbol{\theta}}^{\hat{\theta}_0,\hat{\theta}_1}$ [abrégé en '$\hat{\boldsymbol{\theta}}$' dans la formule ci-dessous] vaut

$$R_{\hat{\boldsymbol{\theta}}}(\theta) = \theta|\hat{\theta}_1 - \theta| + (1 - \theta)|\hat{\theta}_0 - \theta|. \qquad \text{(VT)}$$

La figure 11.1 montre ce que vaut cette fonction de risque pour les trois estimateurs particuliers évoqués un peu plus haut. ♧

Uniforme méliorité

Maintenant que nous disposons de la notion de fonction de risque pour jauger la qualité d'un estimateur, nous aimerions nous en servir pour *comparer* deux estimateurs (ou prédicteurs). Mais, autant on sait dire, entre deux *nombres* disctincts, lequel est le plus petit, autant il n'y a pas d'ordre total sur les *fonctions*...

Une idée qui semblerait naturelle ici serait de *moyenner* les fonctions de risque pour les transformer en un nombre unique, et comparer ensuite les estimateurs via les moyennes de leurs fonctions de risque. Cependant, pour procéder à une telle moyennisation, il faudrait disposer d'une mesure d'intégration sur $\boldsymbol{\Theta}$... Ce qui est précisément ce que le fréquentisme cherche à éviter, puisqu'il part du postulat qu'on

n'a *aucune* idée de la façon dont il convient de pondérer les différentes zones de Θ entre elles ![§]

Nous sommes donc plus ou moins contraints de nous contenter de l'ordre *partiel* usuel entre fonctions pour comparer les fonctions de risque : si un estimateur (ou un prédicteur) a une fonction de risque en tout point inférieure à la fonction de risque d'un autre estimateur (resp. prédicteur), *alors* on pourra dire que le premier estimateur est meilleur que le second... et sinon, eh bien, nos deux estimateurs ne seront pas comparables de façon claire au sens fréquentiste du terme !

Dans le jargon, on dit qu'un estimateur (ou prédicteur) est *uniformément moins risqué* (ou « uniformément meilleur ») qu'un autre lorsqu'on a une telle comparaison entre fonctions de risque :

Définition (VU) (Estimateur ou prédicteur uniformément moins risqué)**.** Soit un modèle statistique explicatif (resp. prédictif) avec les notations génériques et $\gamma(\theta) =: \boldsymbol{\gamma}$ (resp. $g(Y) =: G$) une quantité d'intérêt à valeurs dans $\mathcal{G}$; soient $\hat{\gamma}_0(X)$ et $\hat{\gamma}_1(X)$ deux estimateurs concurrents de $\boldsymbol{\gamma}$ (resp. $\hat{g}_0(X)$ et $\hat{g}_1(X)$ deux prédicteurs...) ; et soit $\ell(\bullet, \bullet)$ une fonction de perte sur $\mathcal{G}$. On dit que l'estimateur $\hat{\boldsymbol{\gamma}}_0$ est *uniformément moins risqué* que l'estimateur $\hat{\boldsymbol{\gamma}}_1$ (pour la fonction de perte ℓ) lorsque la fonction de risque du premier est majorée (sur tout Θ) par la fonction de risque du second : !

$$\forall \theta \in \Theta \quad R^{\text{fréq}}_{\hat{\boldsymbol{\gamma}}_0}(\theta) \leqslant R^{\text{fréq}}_{\hat{\boldsymbol{\gamma}}_1}(\theta) \tag{VV}$$

(et on a la même définition, *mutatis mutandis*, pour les prédicteurs). ♡

Remarque (VW). En pratique, quand nous voudrons montrer qu'on estimateur (resp. prédicteur...) est uniformément moins risqué qu'un autre, nous utiliserons l'astuce de rédaction évoquée en remarque (VM) : à savoir, nous étudierons ce qui se passe sous le véritable contexte probabiliste, i.e. au niveau de la vraie valeur du paramètre caché, mais sans rien supposer sur ce que vaut cette valeur au sein de Θ ; et notre but sera alors d'établir, sans avoir à faire aucune hypothèse sur $\theta_{\checkmark}$ (j'insiste !), que

$$\mathbb{E}_{\checkmark}\big(\ell(\gamma(\theta_{\checkmark}), \hat{\gamma}_0(X))\big) \leqslant \mathbb{E}_{\checkmark}\big(\ell(\gamma(\theta_{\checkmark}), \hat{\gamma}_1(X))\big). \tag{VX}$$

♧

Exemple (VY). Si nous reprenons l'exemple (VS), parmi les trois estimateurs dont les fonctions de risque sont tracées sur la figure 11.1, on voit aucun n'est uniformément meilleur qu'aucun autre : par exemple, entre les estimateurs $\hat{\theta}^{1/8,5/8}$ (ligne continue) et $\hat{\theta}^{1/4,3/4}$ (ligne tiretée), on voit que le premier a un meilleur risque lorsque $\theta < 20\,\%$ et que $\theta \in]50\,\%, 65\,\%[$, mais que c'est le second qui a le risque le plus faible lorsque $\theta \in]21\,\%, 50\,\%[$ ou que $\theta > 66\,\%$... (Et de même, il n'y a pas de comparaison uniforme entre $\hat{\theta}^{1/8,5/8}$ et $\hat{\theta}^{0,1}$, ni entre $\hat{\theta}^{1/4,3/4}$ et $\hat{\theta}^{0,1}$). ♧

Exemple (VZ). Considérons cette fois-ci le modèle du pédagogue, en version prédictive, où notre objectif est de prédire la variance empirique [¶] $\text{var}_{\text{emp}}(Y_0, \dots, Y_{m-1})$

[§]. Et de fait, on pourrait montrer que comparer deux estimateurs (ou prédicteurs) via l'intégrale de leurs fonctions de risques respectives contre une mesure $\mu \in \mathscr{M}(\Theta)$ serait rigoureusement équivalent à comparer les risques intégrés de ces estimateurs dans l'approche bayésienne où on prendrait pour priore sur Θ la loi de probabilité proportionnelle à μ...

[¶]. Rappelons que, par définition de la notion de « empirique », la variance empirique d'un jeu de données $(y_0, \dots, y_{m-1})$ est égale à $m^{-1} \sum_{j=0}^{m-1} y_j^2 - \big(m^{-1} \sum_{j=0}^{m-1} y_j\big)^2$.

des notes de la seconde promotion. Un prédicteur classique pour cela est le prédicteur de Bessel, qui vaut

$$\hat{G}^{\mathrm{B}} \coloneqq \frac{(m-1)n}{m(n-1)} \operatorname{var}_{\mathrm{emp}}((X_i)_{i\in[\![0,n[\![}). \tag{WA}$$

Un prédicteur concurrent est le prédicteur dit « rétréci », défini cette fois-ci par

$$\hat{G}^{\mathrm{R}} \coloneqq \frac{(m-1)n}{m(n+1)} \operatorname{var}_{\mathrm{emp}}((X_i)_{i\in[\![0,n[\![}) \underset{\text{déf}}{=} \frac{n-1}{n+1}\hat{G}^{\mathrm{B}}. \tag{WB}$$

Les fonctions de risques de ces deux prédicteurs sont respectivement

$$R^{\text{fréq}}_{\hat{G}^{\mathrm{B}}}(\mu,\sigma) = 2(1-1/m)^2\big(1/(n-1)+1/(m-1)\big)\sigma^4\,; \tag{WC}$$

$$R^{\text{fréq}}_{\hat{G}^{\mathrm{R}}}(\mu,\sigma) = 2(1-1/m)^2\big(1/(n+1)+1/(m-1)\big)\sigma^4. \tag{WD}$$

On voit ainsi que, pour tout $(\mu,\sigma)\in\Theta$, le risque du prédicteur rétréci est strictement inférieur au risque correspondant du prédicteur de Bessel : le prédicteur rétréci est donc uniformément moins risqué que le prédicteur de Bessel ! ♧

Remarque (WE). En pratique, il est assez rare que le critère d'uniforme méliorité permette de trancher entre deux estimateurs ou prédicteurs : notamment, on peut montrer que, sauf cas trivial, il n'existe aucun estimateur (ou prédicteur) qui soit uniformément moins risqué que tous les autres ! ♧

Optimalité minimax

Comme nous l'avons vu ci-dessus, la notion d'uniforme méliorité ne permet que rarement de trancher entre deux estimateurs (ou prédicteurs) concurrents, ce qui est assez frustrant. On aimerait donc résumer les fonctions de risques en des nombres réels, qu'on pourrait comparer entre eux plus efficacement... Certes, comme nous l'avons souligné plus haut, il serait inapproprié, si on tient au paradigme fréquentiste, de le faire via l'introduction d'une mesure sur l'espace du paramètre caché. Néanmoins l'intégration n'est pas la seule fonctionnelle permettant de résumer une fonction à un nombre unique : puisqu'on peut aussi, par exemple, considérer l'*infimum* ou le *supremum* de cette fonction : ce qui, pour le coup, peut être défini sans avoir à introduire aucune mesure arbitraire sur le domaine de la fonction ! ◡̈

Pour ce qui est de jauger la qualité d'un estimateur ou d'un prédicteur, prendre l'infimum de la fonction de risque fréquentiste n'aurait pas grand sens, dans la mesure où cela reviendrait à regarder ce que devient le risque si on a beaucoup de chance : alors que, justement, le point est que nous ne savons pas ce que vaut la quantité d'intérêt, et que nous ne voulons pas compter sur la chance pour la deviner ! En revanche, le supremum de la fonction de risque présente un véritable intérêt en termes d'ingénierie, puisqu'il nous fournir une *garantie* sur le valeur que le risque véritable est assuré de ne pas dépasser. Ce supremum est appelé, tout simplement, « risque maximal » :

! **Définition** (WF)**.** Avec les notations génériques, on appelle *risque maximal* d'un estimateur $\hat{\gamma}(X)$ (resp. d'un prédicteur $\hat{g}(X)$) le supremum de sa fonction de risque sur Θ. ♡

Exemple (WG). Si nous reprenons l'exemple (VS) (confer en particulier figure 11.1), le risque maximal pour l'estimateur $\hat{\boldsymbol{\theta}}^{0,1}$ vaut 1/2, atteint en $\theta = 1/2$; le risque maximal pour l'estimateur $\hat{\boldsymbol{\theta}}^{1/8,5/8}$ vaut 3/8, atteint asymptotiquement pour $\theta \to 1$, et le risque maximal pour l'estimateur $\hat{\boldsymbol{\theta}}^{1/4,3/4}$ vaut 1/4, atteint à la fois pour $\theta \to 0$, $\theta = 1/2$ et $\theta \to 1$. Au sens du risque maximal, le meilleur de ces trois estimateurs est donc $\hat{\boldsymbol{\theta}}^{1/4,3/4}$, suivi par $\hat{\boldsymbol{\theta}}^{1/8,5/8}$ puis par $\hat{\boldsymbol{\theta}}^{0,1}$.

Et si nous reprenons les prédicteurs de l'exemple (VZ), cette fois-ci, nous voyons qu'aussi bien le prédicteur de Bessel que le prédicteur rétréci ont un risque maximal infini : ces deux prédicteurs sont donc à égalité de ce point de vue-là ! ♧

La notion de risque maximal nous donnant un (pré)ordre sur l'ensemble des estimateurs (ou prédicteurs) d'une quantité d'intérêt donnée, on va cette fois-ci pouvoir trouver des estimateurs qui soient *optimaux* de ce point de vue. On parler alors d'estimateur (resp. prédicteur) « minimax » :

Définition (WH). Étant donnés un modèle statistique, une quantité d'intérêt et une fonction de perte, un estimateur (resp. prédicteur) de la quantité d'intérêt est dit *minimax* lorsque son risque maximal est aussi petit que possible, au sens où aucun autre estimateur (resp. prédicteur) de cette quantité d'intérêt, quel qu'il soit, n'a de risque maximal qui soit strictement plus faible. ♡ !

Exemple (WI). Dans le cadre de l'exemple (VS), on peut montrer que l'estimateur $\hat{\boldsymbol{\theta}}^{1/4,3/4}$ est l'(unique) estimateur minimax de $\boldsymbol{\theta}$ (pour la fonction de perte considérée). ♧

> *Remarque* (WJ). Attention néanmoins, il existe d'assez nombreuses situations où aucun estimateur (ou prédicteur) n'a de risque maximal fini [de manière générale, c'est susceptible de se produire dès lors que la fonction de perte n'est pas bornée] : dans ce cas, tous les estimateurs sont minimax, ce qui rend alors la notion sans intérêt...! ¨˘ C'est notamment le cas pour la problématique considérée dans l'exemple (VZ) (à savoir : dans le modèle du pédagogue, prédiction de la variance empirique de la seconde promotion, avec perte quadratique). ♧

Fonctions de biais

La fonction de risque nous explique, en substance, *de combien* un estimateur ou un prédicteur s'écarte de sa valeur cible. Dans le cas où la cible en question vit dans $\mathbb{R}$, il y a deux directions opposées dans lesquelles on peut se tromper : "vers le haut" ou "vers le bas". Il est alors naturel de se demander si l'erreur de l'estimateur (resp. prédicteur) est "équilibrée" entre ces deux façons de se tromper, ou s'il existe un *biais* systématique vers le haut ou vers le bas : dans ce dernier cas, en effet, cela pourrait suggérer une façon de "corriger" l'estimateur en le "tirant" vers le bas ou vers le haut selon le cas ! ˘¨

L'objet qui décrit la déviation moyenne dans un tel cas est ce qu'on appelle la *fonction de biais* :

Définition (WK). Soit un modèle statistique avec les notations génériques, une quantité d'intérêt à valeurs réelles, et un estimateur (ou prédicteur) de cette quantité d'intérêt. La *fonction de biais* associée à cet estimateur (resp. prédicteur) est alors la fonction de $\boldsymbol{\Theta}$ dans $\mathbb{R}$ qui nous donne, en fonction de θ, l'erreur *signée* moyenne commise par l'estimateur sous le contexte fréquentiste $\mathbb{P}_\theta(\bullet)$. !

Autrement dit, pour un estimateur $\hat{\gamma}(X)$ de $\gamma(\theta)$, c'est

$$\theta \mapsto \mathbb{E}_\theta(\hat{\gamma}(X)) - \gamma(\theta), \tag{WL}$$

et pour un prédicteur $\hat{g}(X)$ de $g(Y)$, c'est

$$\theta \mapsto \mathbb{E}_\theta(\hat{g}(X) - g(Y)) \quad (= \mathbb{E}_\theta(\hat{g}(X) - \mathbb{E}_\theta(g(Y)). \tag{WM}$$

♡

Selon le comportement de la fonction de biais, on dit que les estimateurs (resp. prédicteurs) sont *biaisés* (ou pas) de telle ou telle façon :

! **Définition (WN).** Nous utilisons les notions génériques. Un estimateur ou prédicteur d'une quantité d'intérêt réelle est dit :
- — *Sans biais* lorsque sa fonction de biais est identiquement nulle.
- — *Biaisé* lorsqu'il n'est pas sans biais, autrement dit s'il existe au moins une valeur de θ pour laquelle le biais est non nul. *Biaisé vers le haut* (resp. *vers le bas*), lorsqu'il est biaisé et que sa fonction de biais est partout positive ou nulle (resp. négative ou nulle).

♡

Remarque (WO). Il est tout à fait possible qu'un estimateur biaisé ne soit ni biaisé vers le bas, ni biaisé vers le haut ; en fait, c'est même une situation assez courante : pour certaines valeurs de θ, on aura alors $\mathbb{E}_\theta(\hat{\gamma}(X)) < \gamma(\theta)$, mais pour d'autres, on aura l'inégalité inverse. ♣

Exemple (WP). Pour le modèle du chasseur, nous avions considéré dans l'exemple (TJ) l'estimateur de $\boldsymbol{\theta}$ défini par $\hat{\boldsymbol{\theta}}^{\mathrm{B}} \coloneqq (X + 1/2) / (n + 1)$, que nous avions introduit comme l'espérance à postériori de $\boldsymbol{\theta}$ lorsqu'on prend la priore arcsinus. Nous allons maintenant "oublier" les considérations bayésiennes impliquées dans la dérivation de cet estimateur et analyser les propriétés de biais ce dernier dans une perspective purement fréquentiste.

Il est facile de vérifier que l'espérance (sous le véritable contexte probabiliste) de l'estimateur $\hat{\boldsymbol{\theta}}^{\mathrm{B}}$ vaut $(\theta_{\checkmark} + 1/2) / (n + 1)$, qui n'est pas égal à $\theta_{\checkmark}$ en général : ainsi cet estimateur est biaisé. On observe en outre que l'espérance de notre estimateur sera strictement supérieure à $\theta_{\checkmark}$ dans le cas où $\theta_{\checkmark} < 1/2$, resp. strictement inférieure à $\theta_{\checkmark}$ dans le cas où $\theta_{\checkmark} > 1/2$: ainsi l'estimateur n'est ni biaisé vers le bas ni biaisé vers le haut ; par contre, on pourrait dire à la rigueur qu'il est "biaisé vers 1/2".

Un autre estimateur possible de $\boldsymbol{\theta}$ est l'estimateur "naïf" $\hat{\boldsymbol{\theta}}^{\mathrm{naïf}} \coloneqq X / n$. Cet estimateur-là est sans biais, puisque sous la loi $\mathbb{P}_{\checkmark}$, l'espérance de $\hat{\boldsymbol{\theta}}^{\mathrm{naïf}}$ vaut, par linéarité, n^{-1} fois l'espérance d'une loi Bernoulli$(n, \theta_{\checkmark})$, ce qui vaut bien $\theta_{\checkmark}$ (et que nous avons établi ce résultat sans aucune connaissance spécifique sur la valeur $\theta_{\checkmark}$). ♣

Remarque (WQ). Attention, la notion d'« être sans biais » n'est pas stable par mesure-image par une fonction continue. Par exemple, nous avons vu ci-dessus que, pour le modèle du chasseur, $\hat{\boldsymbol{\theta}}^{\mathrm{naïf}}$ était un estimateur sans biais de $\boldsymbol{\theta}$; néanmoins, l'estimateur de $2\boldsymbol{\theta} - \boldsymbol{\theta}^2$ défini par $2\hat{\boldsymbol{\theta}}^{\mathrm{naïf}} - (\hat{\boldsymbol{\theta}}^{\mathrm{naïf}})^2$ est quant à lui biaisé vers le bas (son espérance sous la vraie loi vaut $2\theta_{\checkmark} - \theta_{\checkmark}^2 - n^{-1}\theta_{\checkmark}(1 - \theta_{\checkmark})$). ♣

Remarque (WR). Attention, un estimateur (ou prédicteur) peut être sans biais tout en étant de qualité excécrable... Par exemple, dans le modèle du pédagoque, l'estimateur de $\boldsymbol{\mu}$ fourni par X_0 (autrement dit, on regarde juste la note de la première copie corrigée) peut facilement être montré comme étant sans biais ; pour autant, on se doute bien qu'un tel estimateur ne sera pas satisfaisant du tout [‖] ! ♧

Remarque (WS). Dans la littérature classique sur la statistique fréquentiste, beaucoup d'auteurs "s'excitent" sur le fait d'être sans biais, comme si cela était un critère de qualité essentiel des estimateurs et prédicteurs. Mais cela est doublement exagéré :

— D'une part, parce qu'il peut y avoir des estimateurs ou prédicteurs biaisés qui sont en fait *meilleurs* que les estimateurs "corrigés" qu'on obtiendrait en les débiaisant. C'est par exemple le cas pour les prédicteurs de l'exemple (VZ) : en effet, l'estimateur de Bessel est sans biais ; le prédicteur rétréci, qui lu est proportionnel, est biaisé vers le bas ; et pourtant le second prédicteur est uniformément moins risqué (en risque quadratique) que le premier !

— D'autre part, parce qu'il est très fréquent de rencontrer des quantités d'intérêt pour lesquelles il est *impossible* de trouver un estimateur (resp. prédicteur) non biaisé (alors qu'on peut néanmoins en trouver des estimateurs d'excellente qualité !). Pour ne citer qu'un exemple, dans le modèle du chasseur, on pourrait montrer qu'il n'existe aucun estimateur sans biais de la quantité d'intérêt $2/(1+\theta)$.

Il n'en demeure pas moins que, conformément à l'intuition, un estimateur ou prédicteur qui aurait un biais "vraiment" trop fort ne sera pas pertinent : on peut d'ailleurs montrer que, dans le cas de la fonction de perte quadratique, le risque est toujours au moins égal au carré du biais ! ♧

En conclusion de cette section, les fonctions de risque et de biais nous fournissent, dans le cadre fréquentiste, des critères numériques pour jauger la qualité d'un estimateur (ou d'un prédicteur). Cependant, on est loin d'avoir une "pierre de touche" aussi décisive que ne l'était la notion de risque intégré dans le cadre bayésien...

C'est pourquoi il est fréquent de recourir aussi à des critères *qualitatifs* pour jauger la qualité des estimateurs et des prédicteurs. Dans la section suivante, nous allons parler du plus important de ces critères : à savoir, la notion de *convergence* des estimateurs en statistique asymptotique.

11.3 Convergence des estimateurs

La présente section traite de la notion de *convergence* des estimateurs. Il s'agit là d'un critère qualitatif asymptotique extrêmement important pour jauger si un estimateur est raisonnablement pertinent. La définition est la suivante :

Définition (WT). Considérons un modèle statistique explicative avec les notations génériques, dans lequel il y a un paramètre du modèle λ, pour lequel on considère un certain régime asymptotique noté « $\lambda \to \infty$ ». (Lorsque nous voudrons souligner la dépendance d'un certain objet en le paramètre du modèle, nous l'indiquerons par un exposant « $^{(\lambda)}$ »). Soient, pour un tel modèle, une quantité d'intérêt $\gamma(\boldsymbol{\theta})$ et un estimateur $\hat{\gamma}^{(\lambda)}(\boldsymbol{X}) =: \hat{\boldsymbol{\gamma}}^{(\lambda)}$ de cette quantité d'intérêt. !

Dans un tel contexte, on dit que « l' »estimateur $\hat{\boldsymbol{\gamma}}$ [ou, plus exactement, la *famille* d'estimateurs $(\hat{\boldsymbol{\gamma}}^{(\lambda)})_\lambda$ [**]] est *convergent* lorsque, pour tout $\theta \in \boldsymbol{\Theta}$, la loi

[‖]. En particulier, cet estimateur n'est pas « convergent » lorsque $n \to \infty$: confer le vocabulaire introduit dans la section suivante.

$\mathrm{Loi}_\theta(\hat\gamma^{(\lambda)}(X^{(\lambda)}))$ (autrement dit, la loi de l'estimateur $\hat{\boldsymbol\gamma}^{(\lambda)}$ dans le contexte probabiliste fréquentiste $\mathbb{P}_\theta(\bullet)$) converge, lorsque $\lambda \to \infty$, vers $\delta_{\gamma(\theta)}$ (autrement dit, vers la masse de Dirac en la valeur constante prise par $\boldsymbol\gamma$ dans le contexte probabiliste $\mathbb{P}_\theta(\bullet)$). ♡

Remarque (WU). Vu que la convergence en loi vers une masse de Dirac est équivalente à la convergence en probabilité vers la constante correspondante, dans la définition précédente, si nous supposons $\mathcal{G}$ muni d'une structure métrique, la propriété de convergence de l'estimateur $\hat{\boldsymbol\gamma}$ revient à dire que :

$$\forall\theta\in\Theta \quad \forall\varepsilon>0 \quad \mathbb{P}_\theta\big(dist(\gamma(\theta),\hat\gamma^{(\lambda)}(X^{(\lambda)}))>\varepsilon\big) \overset{\lambda\to\infty}{\to} 0. \tag{WV}$$

♧

Notez au passage que la notion de convergence s'applique *uniquement* aux estimateurs, pas aux prédicteurs :

Remarque (WW). Il n'y a pas de notion de convergence pour les prédicteurs ; ou plus exactement, si on essayait de transposer la notion de convergence des estimateurs aux prédicteurs, on trouverait qu'aucun prédicteur n'est jamais convergent (sauf cas trivial)... En effet, autant se placer dans l'asymptotique d'une observation de plus en plus riche peut permettre d'obtenir une information arbitrairement fine sur la valeur du paramètre caché, autant l'aléa inhérent à l'observation future, lui, ne pourra pas être éliminé par l'accumulation d'observations ! ♧

Voyons à présent quelques exemples d'estimateurs convergents (ou pas) :

Exemple (WX).

(i) Dans le modèle du chasseur, considérons l'estimateur « naïf » de $\boldsymbol\theta$ donné par $\hat{\boldsymbol\theta}^{\text{naïf}} := X\,/\,n$. Notez déjà que, bien que je parle de « l' »estimateur, il s'agit en réalité d'une *famille* l'estimateurs, puisque l'application $x \mapsto x\,/\,n$ qui dit comment on crée l'estimateur à partir de l'observation dépend du paramètre du modèle n ! Considérons maintenant l'asymptotique $n\to\infty$. Sous le véritable contexte probabiliste $\mathbb{P}_{\text{✓}}(\bullet)$ [††], $\hat{\boldsymbol\theta}^{\text{naïf}(n)}$ apparait comme la moyenne de n v.a.i.i.d. Bernoulli($\theta_{\text{✓}}$) : il s'ensuit, par la loi des grands nombres (au sens de votre cours du semestre 5), que la loi de cet estimateur tend vers $\delta_{\theta_{\text{✓}}}$ lorsque $n\to\infty$. Et puisque nous avons écrit cela sans rien spécifier sur $\theta_{\text{✓}}$, on a en fait que $\mathrm{Loi}_\theta(\hat{\boldsymbol\theta}^{\text{naïf}(n)}) \overset{n\to\infty}{\to} \delta_\theta$ pour tout $\theta\in\,]0,1[$: ainsi l'estimateur $\hat{\boldsymbol\theta}^{\text{naïf}}$ est convergent ! ☺

(ii) Toujours dans le modèle du chasseur, considérons maintenant l'estimateur $\hat{\boldsymbol\theta}^{\text{Bay}} := (X+1/2)\,/\,(n+1/2)$ [‡‡]. Des calculs élémentaires nous montrent qu'on a toujours $|\hat{\boldsymbol\theta}^{\text{Bay}} - \hat{\boldsymbol\theta}^{\text{naïf}}| \leqslant 1/n$. Or nous avons vu au point précédent que, sous le contexte probabiliste $\mathbb{P}_{\text{✓}}(\bullet)$, l'estimateur $\hat{\boldsymbol\theta}^{\text{naïf}}$ converge en loi vers la constante $\theta_{\text{✓}}$. Par conséquent, l'estimateur converge aussi au sens de la convergence en probabilité : en particulier, des seuils arbitrairement petits $\varepsilon,\varepsilon'>0$ étant supposés fixés, on aura pour n suffisamment grand que $\mathbb{P}_{\text{✓}}(|\hat{\boldsymbol\theta}^{\text{naïf}} - \theta_{\text{✓}}| \geqslant \varepsilon/2) \leqslant \varepsilon'$. Dans ce régime, il s'ensuit, par l'inégalité triangulaire, qu'on aura

[∗∗]. En pratique néanmoins, on dira toujours « l'estimateur » plutôt que « la famille d'estimateurs », la dépendance en le paramètre du modèle étant évidente.

[††]. En fait, là aussi, on devrait parler, en toute rigueur, de la *famille* de contextes probabilistes $\mathbb{P}^{(n)}_{\text{✓}}(\bullet)$ associés à la vraie valeur $\theta_{\text{✓}}$ du paramètre caché.

[‡‡]. Cet estimateur correspond à l'espérance à postériori de $\boldsymbol\theta$ dans le paradigme bayésien ; mais ici nous allons le regarder en termes purement fréquentistes.

$\mathbb{P}_{\checkmark}(|\hat{\theta}^{\mathrm{Bay}} - \theta_{\checkmark}| \geqslant \varepsilon/2 + 1/n) \leqslant \varepsilon'$. Or, si n est suffisamment grand, la borne $\varepsilon/2+1/n$ devient $\leqslant \varepsilon$: par conséquent, on a alors que $\mathbb{P}_{\checkmark}(|\hat{\theta}^{\mathrm{Bay}} - \theta_{\checkmark}| \geqslant \varepsilon) \leqslant \varepsilon'$, ce qui montre que $\hat{\theta}^{\mathrm{Bay}}$ converge vers $\theta_{\checkmark}$ au sens de la convergence en probabilité. Dès lors, $\mathrm{Loi}_{\checkmark}(\hat{\theta}^{\mathrm{Bay}})$ converge vers $\delta_{\theta_{\checkmark}}$; et puisque nous avons établi tout cela sans rien dire sur $\theta_{\checkmark}$, on en déduit que l'estimateur $\hat{\theta}^{\mathrm{Bay}}$ est convergent lui aussi !

(iii) Encore et toujours dans le modèle du chasseur, considérons à présent la quantité d'intérêt explicative $\tau := 2\theta - \theta^2$; et proposons-nous d'estimer celle-ci par $\hat{\tau}^{\mathrm{naïf}} := 2\hat{\theta}^{\mathrm{naïf}} - (\hat{\theta}^{\mathrm{naïf}})^2$. Dans la suite, pour simplifier les notations, nous poserons $2\theta - \theta^2 =: \tau(\theta)$. Pour établir que l'estimateur $\hat{\tau}^{\mathrm{naïf}}$ est convergent, on voudrait montrer que, sous $\mathbb{P}_{\checkmark}(\bullet)$, on ait convergence de $\hat{\tau}^{\mathrm{naïf}}$ vers $\tau_{\checkmark}$ au sens de la convergence en probabilité, càd. établir que, pour tous $\varepsilon, \varepsilon' > 0$, on ait $\mathbb{P}_{\checkmark}(|\hat{\tau}^{\mathrm{naïf}} - \tau_{\checkmark}| \geqslant \varepsilon) \leqslant \varepsilon'$ pour n suffisamment grand. Or l'évènement $\{|\hat{\tau}^{\mathrm{naïf}} - \tau_{\checkmark}| \geqslant \varepsilon\}$ peut se ré-écrire comme $\{|\tau(\hat{\theta}^{\mathrm{naïf}}) - \tau(\theta_{\checkmark})| \geqslant \varepsilon\}$; et puisque la fonction $\tau(\bullet)$ est continue en $\theta_{\checkmark}$, cet évènement implique que $\{|\hat{\theta}^{\mathrm{naïf}} - \theta_{\checkmark}| \geqslant \delta\}$ pour peu que nous choisissions $\delta > 0$ suffisamment petit... Et comme la convergence de l'estimateur $\hat{\theta}^{\mathrm{naïf}}$ vers θ nous assure que, quel que soit $\delta > 0$, on a $\mathbb{P}_{\checkmark}(|\hat{\theta}^{\mathrm{naïf}} - \theta_{\checkmark}| \geqslant \delta) \leqslant \varepsilon'$ pour n suffisamment grand, on obtient bien le contrôle dont nous avions besoin pour établir la convergence de l'estimateur $\hat{\tau}^{\mathrm{naïf}}$! ♧

Exemple (WY).

(i) Dans le modèle du pédagogue, considérons l'estimateur de μ consistant à prendre, tout simplement, la moyenne de la promotion, autrement dit, $\hat{\mu}^{\mathrm{naïf}} := (X_0 + \cdots + X_{n-1}) / n$. (Notez bien qu'il s'agit en fait, là encore, d'une *famille* n'estimateurs : non seulement parce que n intervient dans la formule définissant $\hat{\mu}^{\mathrm{naïf}}$, mais aussi parce que le *domaine* de l'application $(x_0, \ldots, x_{n-1}) \mapsto (x_0 + \cdots + x_{n-1}) / n$ qui permet de déduire $\hat{\mu}^{\mathrm{naïf}}$ de l'observation dépend lui-même de n !). Sous le véritable contexte probabiliste $\mathbb{P}_{\checkmark}(\bullet)$, les X_i sont i.i.d. Normale$(\mu_{\checkmark}, \sigma^2_{\checkmark})$, de sorte que $\hat{\mu}^{\mathrm{naïf}}$ suit la loi Normale$(\mu_{\checkmark}, \sigma^2_{\checkmark} / n)$ (en vertu des propriétés des lois normales) : laquelle loi converge bien vers $\delta_{\mu_{\checkmark}}$ lorsque $n \to \infty$. Cela ayant été établi sans rien spécifier sur $\theta_{\checkmark}$, on a bien montré la convergence de l'estimateur $\hat{\mu}^{\mathrm{naïf}}$.

(ii) Bien entendu, dans ce modèle, $\hat{\mu}^{\mathrm{naïf}}$ n'est pas le seul estimateur convergent (dans l'asymptotique $n \to \infty$) de μ... Par exemple, l'enseignant pourrait considérer un estimateur $\hat{\mu}^{\mathrm{pond}}$ dans laquelle les notes de ses ouailles seraient *pondérées* de sorte à donner plus d'importance aux étudiants de milieu de classement qu'à ceux des extrêmités, de la façon suivante : à partir des notes $x_{0\checkmark}, \ldots, x_{(n-1)\checkmark}$, il commencerait par les trier de la plus petite à la plus grande, obtenant ainsi des notes ordonnées $r_{0\checkmark}, \ldots, r_{(n-1)\checkmark}$ (avec $\{\{r_{0\checkmark}, \ldots, r_{(n-1)\checkmark}\}\} = \{\{x_{0\checkmark}, \ldots, x_{(n-1)\checkmark}\}\}$ et $r_{0\checkmark} \leqslant r_{1\checkmark} \leqslant \cdots \leqslant r_{(n-1)\checkmark}$) ; puis il définirait

$$\hat{\mu}^{\mathrm{pond}} := \frac{x_{0\checkmark} + 2x_{1\checkmark} + 3x_{2\checkmark} + \cdots + 3x_{(n-3)\checkmark} + 2x_{(n-2)\checkmark} + x_{(n-1)\checkmark}}{1 + 2 + 3 + \cdots + 3 + 2 + 1}. \qquad (\mathrm{WZ})$$

La v.a. $\hat{\mu}^{\mathrm{pond}}$ ainsi obtenue est alors bien une statistique [ou plus exactement une *famille* de statistiques indexée par n], puisque l'opération pour l'obtenir à partir de l'observation, bien que compliquée, est entièrement déterministe ; et on pourrait démontrer [*] que sa loi sous $\mathbb{P}_{\checkmark}(\bullet)$ converge, lorsque $n \to \infty$, vers

[*]. Ce serait cependant sensiblement plus difficile que dans le cas de $\hat{\mu}^{\mathrm{naïf}}$... La démonstration

la masse de Dirac $\delta_{\mu_\checkmark}$: il s'agit donc bien, là aussi, d'un estimateur convergent de $\boldsymbol{\mu}$.

(iii) Voyons enfin, dans le cas du pédagogue, un exemple d'estimateur de $\boldsymbol{\mu}$ qui ne soit *pas* convergents. Si l'enseignant est paresseux, il peut considérer l'estimateur $\hat{\boldsymbol{\mu}}^{\text{par}}$ de $\boldsymbol{\mu}$ obtenu en faisant la moyenne de ses copies *en s'arrêtant à* 25 *élèves s'ils sont trop nombreux* : $\hat{\boldsymbol{\mu}}^{\text{par}}$ coïncide donc avec $\hat{\boldsymbol{\mu}}^{\text{naïf}}$ pour $n \leqslant 25$, et vaut $(X_0 + X_1 + \cdots + X_{23})/25$ pour $n > 25$ [†]. Dans ce cas, dans l'asymptotique $n \to \infty$, on a à partir d'un certain rang que $\text{Loi}_\checkmark(\hat{\boldsymbol{\mu}}^{\text{par}}) = \text{Normale}(\mu_\checkmark, \sigma^2_\checkmark/25)$: cette fois-ci on n'a pas convergence vers $\delta_{\mu_\checkmark}$, donc l'estimateur $\hat{\boldsymbol{\mu}}^{\text{par}}$ n'est pas convergent. (Même si, en pratique, cela demeure un assez bon estimateur malgré tout ☹). ♣

Remarque (XA). La comparaison entre les items numéros (i) et (iii) de l'exemple (WY) ci-dessus met en valeur le fait que c'est fondamentalement à une *famille* d'estimateurs que la notion de convergence s'applique : en effet, si dans le modèle du pédagogue on ne regarde que le cas $n = 22$ (par exemple), on voit que les deux estimateurs $\hat{\boldsymbol{\mu}}^{\text{naïf}}$ et $\hat{\boldsymbol{\mu}}^{\text{par}}$ sont alors rigoureusement identiques : il n'y aurait donc aucun sens, dans une telle situation, à dire que l'un est convergent, mais pas l'autre...! ♣

Remarque (XB). En pratique, le critère de convergence des estimateurs est plutôt "facile" à satisfaire. Plus précisément : soit on sera dans une asymptotique qui ne fournit pas suffisamment d'information sur la quantité d'intérêt à la limite, et alors il n'existera aucun estimateur convergent de toutes façons ; soit l'asymptotique nous permet de construire des estimateurs convergents, et alors essentiellement tous les estimateurs "intelligents" auxquels on pourra penser seront effectivement convergents ! ☺ ♣

À l'issue de ces trois premières sections, nous comprenons à présent bien ce que signifie « être un bon estimateur (ou prédicteur) ». Néanmoins, cela ne dit pas pour autant *comment* trouver de tels bons estimateurs ou prédicteurs... C'est la question à laquelle nous allons maintenant nous attaquer ; dans la suite de ce chapitre, nous allons présenter différentes méthodes permettant de trouver des estimateurs ou prédicteurs intelligents — lesquels s'avèreront effectivement, en général, être « bons » au sens des critères présentés ci-devant ! ☺

11.4 Estimateur du maximum de vraisemblance

La méthode d'estimation par maximum de vraisemblance, que cette section va présenter, est la "star" des méthodes d'estimation : elle a en effet le triple mérite d'être définie sans ambigüité, de fonctionner pour essentiellement tous les types de modèles, et d'avoir de très bonnes propriétés d'optimalité ! Elle présente néanmoins un inconvénient, c'est que les calculs qu'elle requiert sont compliqués, et pas toujours menables à bien : d'où l'intérêt des autres méthodes d'estimations présentées par ailleurs dans ce chapitre.

Définissons donc la méthode :

! serait à votre portée sous forme d'exercice avec questions intermédiaires ; mais sans aide, elle reste trop ardue pour des élèves-ingénieurs, je pense ! ☹

[†]. Notons que dans ce cas, bien que la formule définissant l'estimateur à partir des X_i soit toujours la même, il s'agit *quand même*, *techniquement*, d'estimateurs *différents* pour des valeurs de n différentes, dans la mesure où il s'appliquent à des *modèles* différents !

Définition (XC) (Estimateur du maximum de vraisemblance)**.** Soit $\Theta \ni \theta \mapsto \mathrm{Loi}_\theta(X) \in \mathscr{M}_1(\mathscr{X})$ un modèle statistique explicatif pour lequel on peut définir une fonction de vraisemblance $\mathscr{L}(\theta \mid x)$. Alors :

(i) $\hat{\theta}(X)$ est qualifié d'*estimateur du maximum de vraisemblance* pour le paramètre caché $\boldsymbol{\theta}$ lorsque, pour tout $x \in \mathscr{X}$, $\hat{\theta}(x)$ est le (ou un) point où $\theta \mapsto \mathscr{L}(\theta \mid x)$ atteint son maximum ;

(ii) Pour $\gamma : \Theta \to \mathscr{G}$ une fonction, on dit que $\gamma(\hat{\boldsymbol{\theta}})$ est un estimateur du maximum de vraisemblance pour la quantité d'intérêt $\gamma(\boldsymbol{\theta})$. ♡

Autrement dit, l'estimation par maximum de vraisemblance consiste à estimer $\boldsymbol{\theta}$ par la valeur qui nous donnait la plus grande chance de tomber sur les données que nous avons effectivement observées ; et on en déduit, par compostion, l'estimation par maximum de vraisemblance de n'importe quelle quantité d'intérêt explicative.

Remarque (XD)*.* Dans le cas de l'estimation par maximum de vraisemblance, on remarquera que la *réalisation* de l'estimateur du maximum de vraisemblance (ce que j'appelle l'« estimation » du maximum de vraisemblance) se détermine plus directement que l'estimateur lui-même : en effet, il suffit alors de considérer la fonction de vraisemblance *pour l'observation qu'on a effectivement obtenue*, et à regarder où elle atteint son maximum. Et dans nombre de situations pratiques, en fait, seule la *réalisation* de l'estimateur nous intéressera : dans de tel cas, on se s'embêtera évidemment pas à écrire l'estimateur du maximum de vraisemblance en version « variable aléatoire » ! ◡̈ ♧

Remarque (XE)*.* Attention, le vocabulaire est un peu glissant : nonobstant son nom, l'estimation du maximum de vraisemblance n'est pas le maximum de la vraisemblance, mais le lieu où ce maximum est atteint : autrement dit, c'est en fait *l'argument du maximum* de la fonction de vraisemblance... ♧

Remarque (XF)*.* On voit que la méthode d'estimation par maximum de vraisemblance ne laisse pas de place à l'arbitraire dans la façon de le construire (en effet, le statisticien n'a rien à choisir), ce qui est évidemment un point très intéressant ! ◡̈ ♧

> *Remarque* (XG)*.* L'estimateur du maximum de vraisemblance n'existe pas forcément (le supremum de la fonction de vraisemblance pouvant n'être atteint qu'asymptotiquement) ; et quand il existe, il n'est pas forcément unique (car le maximum de la fonction de vraisemblance peut, à l'inverse, être atteint en plusieurs valeurs). Heureusement, on aura quand même unique existence la plupart du temps ◡̈ ♧

On peut montrer que, dans les modèles avec paramètre du modèle disposant d'un régime asymptotique, les estimateurs par maximum de vraisemblance sont en général convergents. On a notamment des résultats précis dans ce sens pour les modèles d'échantillonnage [‡], comme le théorème ci-dessous, que je mentionne pour la culture :

> **Théorème** (XH)**.** *Dans un modèle d'échantillonnage, sous certaines hypothèses de régularité toujours vérifiées en pratique, l'estimateur du maximum de vraisemblance est convergent.* ◇

Remarque (XI)*.* On notera que la méthode du maximum de vraisemblance relève fondamentalement de l'inférence *explicative* : il n'existe pas de notion de soi-disant "prédicteur par maximum de vraisemblance" ! De manière générale, en statistique

[‡]. Ceux d'entre vous qui suivront le cours électif « Problèmes inverses » au semestre 8 (réf. 8KUCEN07) auront l'occasion d'en apprendre davantage à ce sujet lors de la partie statistique de ce cours-là (dispensée par M$^{\text{me}}$ Fristch).

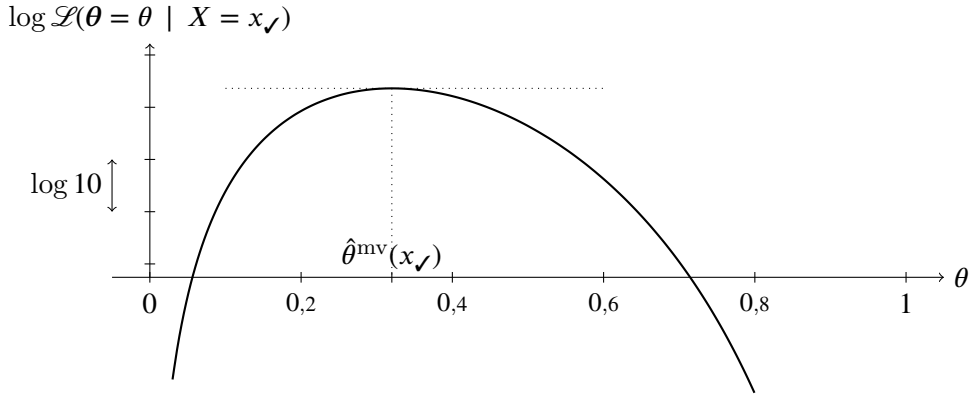

FIGURE 11.2 – Détermination du maximum de vraisemblance pour le modèle du chasseur.

fréquentiste, la plupart des méthodes de ce chapitre ne concerneront que des quantités d'intérêt explicatives : pour ce qui est de la prédiction, on procèdera généralement en deux temps, confer § ?? [§]. ♣

Maintenant que nous avons expliqué *ce qu'est* l'estimation par maximum de vraisemblance, regardons quelques exemples ! Ci-dessous nous allons calculer l'estimateur par maximum de vraisemblance pour nos deux exemples favoris ☺

Exemple (XJ) (Maximum de vraisemblance pour le chasseur). Nous avons vu que la log-vraisemblance du modèle du chasseur est

$$\ln \mathscr{L}(\theta) = x_{\checkmark} \ln(\theta) + (n - x_{\checkmark}) \ln(1-\theta) : \tag{XK}$$

on voit sur la figure 11.2 que, lorsque θ varie (on a pris ici $n = 25$ et $x_{\checkmark} = 8$), cette fonction de vraisemblance semble bien prendre des valeurs très faibles pour des valeurs "invraisemblables" de $\boldsymbol{\theta}$ (car il apparait invraisemblable qu'un tireur ayant touché 8 cibles et en ayant manqué 17 ait un taux de succès véritable inférieur à 1 % ou supérieur à 90 %, par exemple), et passer par un maximum autour de 0,3. Calculons plus précisément où ce maximum est atteint, en annulant le dérivée de la fonction de log-vraisemblance : on calcule que

$$\frac{\partial}{\partial\theta} \ln \mathscr{L}(\boldsymbol{\theta} = \theta) = \frac{x_{\checkmark}}{\theta} - \frac{n - x_{\checkmark}}{1-\theta}, \tag{XL}$$

qui s'annule lorsque $\theta = n^{-1} x_{\checkmark}$: ainsi, l'estimation du maximum de vraisemblance pour $\boldsymbol{\theta}$ est $n^{-1} x_{\checkmark}$.

Et si on cherche l'estimateur du maximum de vraisemblance en tant que *variable aléatoire*, il faut alors simplement donner la v.a. dont $n^{-1} x_{\checkmark}$ est la réalisation : en l'occurrence, c'est simplement $n^{-1} \boldsymbol{X}$. Nous pourrons noter cette statistique '$\hat{\boldsymbol{\theta}}^{\mathrm{mv}}$' pour indiquer qu'il s'agit de l'estimateur du maximum de vraisemblance. ♣

Exemple (XM) (Maximum de vraisemblance pour le pédagogue). Nous avons calculé que la log-vraisemblance, dans l'exemple du pédagogue, était

$$\ln \mathscr{L}(\boldsymbol{\theta} = (\mu, \sigma) \mid \boldsymbol{X} = (x_{1\checkmark}, \dots, x_{n\checkmark})) = \\ - n \ln \sigma - \tfrac{1}{2} n \sigma^{-2} \big(\big(\mu - \mathrm{moy}(x_{1\checkmark}, \dots, x_{n\checkmark}) \big)^2 + \mathrm{var}_{\mathrm{emp}}(x_{1\checkmark}, \dots, x_{n\checkmark}) \big) :$$

[§]. Section non encore rédigée à l'heure actuelle

il s'agit d'ajuster μ et σ pour maximiser cette quantité. En fait, on s'aperçoit qu'à σ fixé, la vraisemblance est maximisée pour $\mu = \mathrm{moy}(x_{1\checkmark}, \dots, x_{n\checkmark})$; de sorte que, si on admet qu'il existe un maximum de vraisemblance, il est nécessairement atteint en un point $(\hat{\mu}_\checkmark, \hat{\sigma}_\checkmark)$ tel que $\hat{\mu}_\checkmark = \mathrm{moy}(x_{1\checkmark}, \dots, x_{n\checkmark})$: ainsi, sans même connaitre le maximum de vraisemblance $\hat{\theta}(x_\checkmark)$ pour $\boldsymbol{\theta}$, nous pouvons déjà conclure que (la réalisation de) le maximum de vraisemblance pour $\boldsymbol{\mu}$ est égal à $\mathrm{moy}(x_{1\checkmark}, \dots, x_{n\checkmark})$.

Avec un peu de travail supplémentaire, on montrerait qu'il y a effectivement un (unique) maximum de vraisemblance pour $\boldsymbol{\theta}$, et que (la réalisation de) celui-ci est égal à $\big(\mathrm{moy}(x_{1\checkmark}, \dots, x_{n\checkmark}), \mathrm{var}^{1/2}_{\mathrm{emp}}(x_{1\checkmark}, \dots, x_{n\checkmark})\big)$: en l'occurrence donc, l'estimateur du maximum de vraisemblance pour $(\mu_\checkmark, \sigma_\checkmark)$ coïncide avec les moyenne et écart-type empiriques ! (confer § 11.5 *infra*). ♧

11.5 Estimateurs empiriques

La notion d'*estimation empirique* nous allons présenter dans cette section est une des notions d'estimation plus simples à saisir et à appliquer ; néanmoins son usage sera assez restreint, car on ne peut l'appliquer qu'à certains modèles particuliers... Cela englobera néanmoins une des classes de modèles les plus importantes, à savoir, celle des modèles d'échantillonnage.

Les modèles où nous nous intéressons dans le cadre de cette section sont ceux relevant du point ci-dessous :

Point (XN). Dans cette section, nous nous intéresserons à des modèles dont l'observation se décompose en $n > 1$ sous-observations (n étant un paramètre du modèle) qui sont toutes de même loi lorsqu'on se place sous un contexte probabiliste fréquentiste : autrement dit, avec les notations génériques, on a n sous-observations $X_0, \dots, X_{n-1}$ qui vérifient, pour tout $\theta \in \Theta$, $\mathrm{Loi}_\theta(X_i) = \mathrm{Loi}_\theta(X_j)$ $\forall i, j \in [\![0, n[\![$. On pourra alors aussi noter l'observation globale $\vec{X}$ pour rappeler qu'il s'agit en fait d'un "vecteur" de sous-observations. L'espace dans lequel vivent chacune de ces sous-observations sera alors noté $\mathscr{X}^{(1)}$. (L'espace de l'observation globale $\mathscr{X}$ étant alors $(\mathscr{X}^{(1)})^n$, qu'on pourra aussi noter '$\mathscr{X}^{(n)}$').

La situation décrite par ce point englobe notamment tous les modèles d'échantillonnage, qui correspondent au cas où, en plus de ce qui précède, les X_i sont indépendants sous les contextes $\mathbb{P}_\theta(\bullet)$. Le modèle du pédagogue, par exemple, rentre dans un tel cadre. ♧

Dans un tel modèle, sous le véritable contexte probabiliste, les sous-observations $x_{0\checkmark}, \dots, x_{(n-1)\checkmark}$ constituent un *échantillon* de tirages de la loi $\mathrm{Loi}_\checkmark(X_0)$. Si, en outre, on est dans un modèle d'échantillonnage, ces tirages sont indépendants : par conséquent, si n est suffisamment grand, on s'attend d'après la loi des grands nombres à ce que la distribution de cet échantillon ressemble à celle de la loi dont ils sont tirés ! On peut même espérer que cela reste vrai pour des modèles généraux où il n'y aurait pas indépendance *stricto sensu*, dès lors que les réalisations X_i de la loi $\mathrm{Loi}_\checkmark(X_0)$ soient suffisamment bien « mélangées ».

Par exemple, dans le modèle du pédagogue, on s'attend à ce que pour n grand, la loi (sous le véritable contexte probabiliste) de la variable aléatoire $n^{-1}\,\mathrm{card}\{i \mid X_i \geqslant 86\}$ soit très concentrée autour de la valeur constante $\mathbb{P}\big(\mathrm{Normale}(\mu_\checkmark, \sigma^2_\checkmark) \geqslant$

$86\big)$ [¶] ; de même, $n^{-1}\sum_i X_i$ devrait être proche de la constante $\mathbb{E}\big(\text{Normale}(\mu_\checkmark, \sigma_\checkmark^2)\big)$ (qui, en l'occurrence, n'est autre que $\mu_\checkmark$) ; et le quantile de niveau 80 % de l'(multi)ensemble des X_i (i.e. la valeur telle que 80 % des notes des élèves sont tombées en-dessus et 20 % au-dessus : par exemple, si $n = 22$, c'est la note de l'élève classé 5e) devrait être proche du quantile de niveau 80 % de la loi Normale($\mu_\checkmark$, $\sigma_\checkmark^2$) [‖].

En fait, ces trois cas peuvent tous être reformulés dans un même cadre, en disant qu'une certaine fonction de la *loi empirique* des (sous-)observations converge vers la valeur correspondante de la véritable distribution $\text{Loi}_\checkmark(X_0)$ de ces observations. Rappelons ci-dessous le concept de « loi empirique » (déjà introduit en définition (DR)) :

! **Définition (XO)** (Loi empirique)**.** Soient $x_0, \dots, x_{n-1}$ un jeu de n valeurs appartenant à un même espace $\mathcal{X}^{(1)}$; alors la *loi empirique* des x_i est la loi de probabilité qui donne le poids $1/n$ à chaque x_i (en comptant plusieurs fois les éventuelles valeurs identiques), soit

$$\frac{1}{n}\sum_{i=0}^{n-1}\delta_{x_i}\ : \tag{XP}$$

en d'autres termes, c'est la distribution qui donne à $A \subseteq \mathcal{X}^{(1)}$ une probabilité égale à la *proportion* des $x_{i\checkmark}$ qui sont tombés dans A.

Dans le cas d'un modèle relevant du point (XN), on appelle *loi empirique des (sous-)observations* la variable aléatoire à valeurs dans $\mathscr{M}_1(\mathcal{X}^{(1)})$ dont la réalisation est la loi empirique des $x_{i\checkmark}$. ♡

Remarque (XQ)*.* Bien comprendre que la loi empirique des observations est une loi de probabilité, mais que cette loi dépend directement de l'observation, et constitue donc (si on se place avant de faire l'expérience) un objet *aléatoire* ! Autrement dit, la loi empirique des observations est une *variable aléatoire à valeurs dans les distributions de probabilité* (outch !). ♧

Avec ce vocabulaire, on voit bien que, dans les trois exemples cités avant la définition (XO), on était à chaque fois en train de comparer une certaine propriété de la loi empirique des observations avec la propriété correspondante de la loi véritable. Par exemple, $n^{-1}\,\text{card}\{i \mid x_{i\checkmark} \geqslant 86\}$ n'était autre que la masse que la distribution empirique attribuait aux valeurs $\geqslant 86$, et nous disions que cela devait tendre, pour n grand, vers la masse que la loi des observations Normale($\mu_\checkmark, \sigma_\checkmark^2$) attribue aux valeurs $\geqslant 86$: rien de plus logique, puisque la distribution empirique des observations est censée ressembler à leur véritable loi ! Et les deux autres exemples relevaient du même schéma, en remplaçant « masse attribuée aux valeurs $\geqslant 86$ » par resp. « espérance » et « quantile de niveau 80 % ».

On peut généraliser cette idée *ad libitum* : c'est en cela que consiste la notion d'*estimation empirique* :

! **Définition (XR)** (Estimateur empirique)**.** Plaçons-nous dans le cadre d'un modèle relevant du point (XN), et supposons qu'il existe une relation fonctionnelle

[¶]. Qui, en l'occurrence, peut être exprimée comme valant $1/2 + \frac{1}{2}\,\text{erf}\big((\mu_\checkmark - 86)\,/\,\sqrt{2}\sigma_\checkmark\big)$, où 'erf' désigne la fonction spéciale appelée « fonction d'erreur ».

[‖]. Qui, en l'occurrence, vaut $\mu_\checkmark + \text{erf}^{-1}(0{,}6) \times \sqrt{2}\sigma_\checkmark$.

entre $\mathrm{Loi}_\theta(X_1)$ et la quantité d'intérêt $\gamma(\theta) \in \mathcal{G}$, exprimée par une certaine fonctionnelle [**] $F(\bullet)$

$$\forall \theta \in \Theta \qquad F(\mathrm{Loi}_\theta(X_1)) = \gamma(\theta), \tag{XS}$$

où l'expression « $F(P)$ » a du sens pour n'importe quelle distribution de probabilité P sur $\mathcal{X}^{(1)}$ (y compris celles qui ne sont pas de la forme $\mathrm{Loi}_\theta(X_0)$. Alors une estimation de $\gamma(\theta)$ est donnée par

$$F\left(\frac{1}{n}\sum_{i=0}^{n-1}\delta_{x_{i\checkmark}}\right) : \tag{XT}$$

dans ce cas, on parle d'*estimation empirique* (et on appelle « estimateur empirique » l'estimateur dont notre estimation est la réalisation).

En d'autre termes, pour « bidule » un concept pouvant s'appliquer à des distributions de probabilité, une procédure d'estimation empirique consiste à estimer le *bidule* de la distribution des observations par le *bidule* empirique (cf. définition (DW)) des observations. ♡

Remarque (XU). Il peut tout à fait exister plusieurs fonctionnelles différentes $F_0, F_1, \ldots$ vérifiant chacune la relation $\gamma(\theta) = F_i(\mathrm{Loi}_\theta(X_0))$, menant à autant d'estimateurs empiriques différents. Par exemple, dans le cas du pédagogue, μ peut être certes être vue comme la moyenne de $\mathrm{Normale}(\mu, \sigma^2)$, mais c'est aussi la médiane de cette loi : si on préfère utiliser cette fonctionnelle-là, on obtient donc un autre estimateur empirique de μ ! Par exemple, dans le cas où $n = 5$ et $(x_{0\checkmark}, \ldots, x_{4\checkmark}) = (52, 79, 88, 91, 70)$, l'estimation empirique de μ par moyenne vaut 76, tandis que l'estimation empirique de μ par médiane vaut 79... ♧

On l'aura compris, les estimateurs empiriques sont très simples à fabriquer ! En outre, pour peu que le modèle présente de suffisamment bonnes propriétés de « mélange », ils auront en général le bon gout d'être convergents lorsque le nombre d'observations tend vers l'infini. Citons ainsi, pour la culture, le résultat suivant :

> **Théorème** (XV). *Lorsqu'on est dans le cas d'un modèle d'échantillonnage, sous des hypothèses assez souples de régularité de la fonctionnelle $F(\bullet)$* [††]*, l'estimateur empirique associé à la fonctionnelle F est convergent lorsque la taille de l'échantillon tend vers l'infini.* ◇

Remarque (XW). De même que pour le maximum de vraisemblance, la méthode empirique est une technique spécifique aux questions d'*estimation* : il n'existe pas de notion de soi-disant « prédicteur empirique » ! ♧

[**]. Note de vocabulaire : Lorsqu'on considère une fonction $F(\bullet)$ dont l'espace de définition est lui-même un espace de fonctions (ou quelque chose d'apparenté, comme un espace de lois de probabilités en l'occurrence), l'usage des mathématiciens est de parler de « fonctionnelle », plutôt que de « fonction », pour désigner $F(\bullet)$, afin de limiter les risques de confusion. (Par exemple, une fonction $\varphi : \mathcal{X} \to \mathbb{R}$ étant donnée, l'application « espérance contre φ », définie de $\mathcal{M}_1(\mathcal{X})$ dans $\mathbb{R}$ par $P \mapsto \mathbb{E}^{X\sim P}(\varphi(X))$, sera volontiers qualifiée de « fonctionnelle », afin d'éviter les confusions avec la « fonction » $\varphi(\bullet)$). En général, quand on parle de « fonctionnelle », on sous-entend en outre que l'espace d'arrivée (qui, dans le cadre de la présente définition, est l'espace $\mathcal{G}$ de la quantité d'intérêt) est quelque chose de "simple", comme $\mathbb{R}$.

[††]. Mais qu'il serait fort compliqué d'expliciter ici... ⚟

11.6 Méthode des moindres carrés

Nous allons maintenant présenter la méthode dite « des moindres carrés ». Bien que cette méthode ne soit pas aussi universelle que l'estimation par maximum de vraisemblance, elle peut tout de même être appliquée à une très grande variété de modèles. La méthode des moindres carrés sera tout particulièrement adaptée à tous les modèles où l'observation est vue comme la somme d'une « tendance » déterministe et d'un « bruit » centré aléatoire, ce qui la rend très appréciée dans le domaine de l'*analyse de données*, qui vous sera enseigné au semestre 7.

La catégorie de modèles à laquelle s'applique la méthode des moindres carrés est décrite dans le point ci-dessous :

Point (XX). Dans cette section, nous considèrerons un modèle statistique dont nous noterons l'espace du paramètre caché Θ, et dont l'observation X se décompose en n sous-observations réelles $X_0, \dots, X_{n-1}$, qui sont supposées d'intégrabilité L^1 sous les différents contextes probabilistes (autrement dit, on peut donner du sens aux $\mathbb{E}_\theta(X_i)$).

On notera que, par contre, on ne requiert absolument pas que les X_i aient la même loi sous $\mathbb{P}_\theta(\bullet)$, ni qu'ils soient indépendants ! ♧

Au fondement de la méthode des moindres carrés est l'observation suivante : parmi les différentes valeurs possibles θ du paramètre caché, c'est lorsqu'on aura $\theta = \theta_\checkmark$ que les $\mathbb{E}_\theta(X_i)$ seront, *en moyenne* (sous le véritable contexte probabiliste), les plus proches des X_i au sens de l'erreur quadratique ! Formellement, il s'agit de la proposition suivante :

Proposition (XY). *Considérons un modèle statistique relevant du point (XX) ; supposons en outre que pour tout $\theta \in \Theta$, les X_i soient toutes d'intégrabilité L^2 sous $\mathbb{P}_\theta(\bullet)$. Introduisons alors la fonction suivante, dite* fonction de contraste quadratique *:*

$$\begin{aligned} c(\bullet,\bullet) : \mathbb{R}^n &\to \mathbb{R}_+ \\ \big((x_0,\dots,x_{n-1}),\theta\big) &\mapsto \sum_{i=0}^{n-1} (x_i - \mathbb{E}_\theta(X_i))^2. \end{aligned} \tag{XZ}$$

alors, pour une valeur $\theta_\checkmark$ définissant un véritable contexte probabiliste $\mathbb{P}_\checkmark(\bullet)$, la fonction

$$\Theta \ni \theta \mapsto \mathbb{E}_\checkmark\big(c(\vec{X},\theta)\big) \tag{YA}$$

atteint son minimum en $\theta_\checkmark$. ♢

Démonstration. En effet, pour $\theta \in \Theta$,

$$\begin{aligned} \mathbb{E}_\checkmark\big((X_i - \mathbb{E}_\theta(X_i))^2\big) &= \mathbb{E}_\checkmark(X_i - \mathbb{E}_\theta(X_i))^2 + \mathrm{Var}_\checkmark(X_i - \mathbb{E}_\theta(X_i))^2 \\ &= (\mathbb{E}_\checkmark(X_i) - \mathbb{E}_\theta(X_i))^2 + \mathrm{Var}_\checkmark(X_i), \end{aligned}$$

de sorte que

$$\mathbb{E}_\checkmark\Big(\sum_{i=1}^n (X_i - \mathbb{E}_\theta(X_i))^2\Big) = \sum_{i=1}^n \mathrm{Var}_\checkmark(X_i) + \sum_{i=1}^n (\mathbb{E}_\checkmark(X_i) - \mathbb{E}_\theta(X_i))^2, \tag{YB}$$

où le premier terme est indépendant de θ tandis que le second, qui est évidemment toujours positif, est nul pour $\theta = \theta_\checkmark$, de sorte que globalement l'expression est bien minimisée pour $\theta = \theta_\checkmark$. ♤

La proposition (XY) suggère alors l'idée suivante : puisque $\theta_{\checkmark}$ est la valeur qui minimise la fonction de constraste quadratique *en moyenne*, alors elle ne devrait pas se situer trop loin de la valeur qui minimisera la fonction de constraste quatratique *pour l'observation effective* ! La valeur $\hat{\theta}^{\mathrm{mc}}_{\checkmark}$ réalisant cette minimisation est appelée *estimateur des moindres carrés* de notre modèle :

Définition (YC) (Estimateur des moindres carrés)**.** Pour un modèle statistique relevant du point (XX), l'*estimation des moindres carrés* pour $\boldsymbol{\theta}$ associée à l'observation effective $(x_{0\checkmark}, \dots, x_{(n-1)\checkmark})$ est définie comme la valeur de θ minimisant la fonction de contraste des moindres carrés pour l'observation effective :

$$\hat{\theta}^{\mathrm{mc}}_{\checkmark} \coloneqq \arg\min_{\theta} \sum_{i=1}^{n} (x_{i\checkmark} - \mathbb{E}_{\theta}(X_i))^2. \tag{YD}$$

♡

On peut ensuite éventuellement en déduire, par composition, une estimation de $\gamma(\hat{\theta}^{\mathrm{mc}}_{\checkmark})$ d'une quantité d'intérêt de la forme $\gamma(\boldsymbol{\theta})$, qu'on qualifiera là encore d'« estimation des moindres carrés » le cas échéant.

Remarque (YE). En fait, il ne sera pas rare que le minimum de le fonction de constraste soit atteint sur un ensemble de *plusieurs* valeurs θ, mais que toutes ces valeurs correspondent à la même valeur $\gamma(\theta)$ en ce qui concerne la quantité d'intérêt considérée : auquel cas, on considèrera qu'on a bien un estimateur des moindres carrés défini de façon non ambigüe pour cette quantité d'intérêt malgré tout ! C'est notamment ce qui se produira dans le modèle de la *régression linéaire* (confer chapitre **??**), ou les paramètres de régression $\boldsymbol{\alpha}$ et $\boldsymbol{\beta}$ auront des estimateurs de moindres carrés bien définis, mais pas le paremère de bruit $\boldsymbol{\sigma}$. ♧

Remarque (YF). L'estimateur des moindres carrés est très apprécié en analyse des données, car il ne fait intervenir que les « valeurs de régression » $\mathbb{E}_{\theta}(X_i)$, peu importe la loi précise suivie par les X_i : or, dans la mesure où il est difficile de modéliser pertinemment les détails de la loi des observations pour des données "de la vraie vie" un peu compliquées, cela constitue un avantage considérable ! ◡̈ ♧

Remarque (YG). En fait, l'estimateur des moindres carrés n'est qu'un cas particulier d'une méthode beaucoup plus générale, appelée « méthode de contraste », où le principe consistera à trouver d'autres « fonctions de contraste » vérifiant la propriété de la proposition (XY), puis à optimiser ces fonctions de contraste de la meêm façon que nous l'avions fait pour le contraste quadratique. Cependant, ces méthodes dépassent le bagage qu'on est raisonnablement en droit d'attendre d'un ingénieur généraliste, et nous ne les développerons donc pas dans ce cours ◡̈ ♧

Finissons cette section avec un exemple, en calculant l'estimateur des moindres carrés pour le modèle du chasseur :

Exemple (YH) (Moindres carrés pour le chasseur). Bien que l'observation du modèle du chasseur ne se décompose pas en sous-observations, les valeurs qu'elle prend sont réelles : on peut donc essayer d'appliquer la définition (YC) avec $n \leftarrow 1$ [attention, le "n" de la définition (YC) n'est pas le même que le "n" du modèle du chasseur : en l'occurrence, c'est le *premier* de ces deux "n" que nous prenons égal à 1 ici !]. On calcule facilement qu'on a $\mathbb{E}_{\theta}(X) = n\theta$ [‡‡] ; par conséquent, la fonction de contraste est $(x, \theta) \mapsto (x - n\theta)^2$; et l'estimation des moindres carrés pour $\boldsymbol{\theta}$ sera donc la valeur de θ minimisant $(x_{\checkmark} - n\theta)^2$: on tombe ainsi sur $\hat{\theta}^{\mathrm{mc}}_{\checkmark} = x_{\checkmark} / n$, ce qui coïncide en l'occurrence avec l'estimateur "naïf" dont nous avions déjà parlé ! ◡̈ ♧

[‡‡]. Où, cette fois-ci, "n" se réfère au nombre de tirs dans le test du candidat...

11.7 Estimation par tendance

Dans cette section nous allons présenter une méthode d'estimation très générale, que j'appellerai *méthode par tendance*. De manière assez étrange, cette méthode ne semble pas avoir de nom ni même être identifiée comme telle dans les ouvrages statistiques de référence ; elle est cependant au cœur de la très classique « méthode des moments » que nous présenterons un peu plus loin.

L'idée est la suivante : on va chercher une statistique $\hat{\gamma}(X)$ qui "ressemble" à la quantité d'intérêt $\gamma(\theta_✓)$, au sens où, peu importe la valeur $\theta_✓$, son espérance (sous la véritable probabilité $\mathbb{P}_✓$) vaille $\gamma(\theta_✓)$ (et, de préférence, que sa loi soit concentrée autant que possible autour de $\gamma(\theta_✓)$) ; et on parlera alors d'« estimateur par tendance ».

! **Définition** (YI) (Estimateur par tendance)**.** Soit un modèle statistique (fréquentiste) avec les notations habituelles, et soit $\hat{\gamma}(X)$ une statistique à valeurs réelles telle que

$$\forall \theta \in \Theta \quad \mathbb{E}_\theta(\hat{\gamma}(X)) = \gamma(\theta) \tag{YJ}$$

pour une certaine fonction $\gamma : \Theta \to \mathbb{R}$. Alors $\hat{\gamma}(X)$ est qualifié d'*estimateur par tendance* de la quantité d'intérêt $\gamma(\theta)$. ♡

Remarque (YK). On pourrait, pour le coup, étendre la notion d'« estimateur par tendance » en une notion de « prédicteur par tendance » ; cependant, ce ne serait pas très intéressant en pratique ∴ ♧

Remarque (YL). En fait, dire que $\hat{\gamma}(X)$ est un estimateur par tendance de $\gamma(\theta)$ est rigoureusement synonyme à dire qu'il s'en agit d'un estimateur *sans biais*... (confer définition (WN)). Par rapport à la section 11.2, j'ai juste changé l'ordre dans lequel on raisonne : alors qu'à l'époque, l'idée était de dire « considérons un certain estimateur obtenu par telle ou telle méthode ; calculons sa fonction de biais : oh, l'estimateur est sans biais, c'est une bonne nouvelle ! », cette fois-ci, la démarche est : « considérons une quantité d'intérêt qu'on voudrait estimer ; si on arrive à trouver une statistique qui n'a pas de biais par rapport à cette quantité d'intérêt, alors il sera légitime de considérer qu'elle en constitue un estimateur ! » ⁔ ♧

Remarque (YM). En pratique, toute la difficulté est que la quantité d'intérêt $\gamma(\theta)$ est imposée par le contexte de notre analyse, et que nous devons trouver $\hat{\gamma}$ pour retomber précisément sur cette quantité d'intérêt ! Comme cela est particulièrement difficile à faire, nous devrons souvent partir d'une statistique "inadaptée" $\hat{\psi}$, en déduire un estimateur par tendance d'une *autre* quantité d'intérêt $\psi(\theta)$, et utiliser ensuite l'idée de *substitution* que nous présenterons plus loin. ♧

Exemple (YN). Voyons un exemple très simple d'estimateur par tendance pour le modèle du chasseur : le nombre de tirs X réussis lors du test est censé avoir pour moyenne $\mathbb{E}\big(\mathrm{Binom}^{\mathrm{le}}(n, \theta_✓)\big) = \theta_✓ n$; on a donc $\theta_✓ = \mathbb{E}(n^{-1}X)$, ce qui suggère l'estimateur par tendance $n^{-1}X$ pour θ. Avec les paramètres pris en exemple, on estimera donc θ par 0,32. (Incidemment, on retombe en l'occurrence sur l'estimateur du maximum de vraisemblance). ♧

Exemple (YO). Dans la modèle du pédagogue, on calcule que la variance empirique $\mathrm{var}_{\mathrm{emp}}(X_1, \dots, X_n)$ des observations (autrement dit, la quantité $n^{-1}\sum_{i=1}^n X_i^2 - \big(n^{-1}\sum_{i=1}^n X_i\big)^2$) a pour espérance $\frac{n-1}{n}\sigma_✓^2$; par conséquent, la variance bessélisée $\mathrm{var}_{\mathrm{B}}(X_1, \dots, X_n) \coloneqq \frac{n}{n-1}\mathrm{var}_{\mathrm{emp}}(X_1, \dots, X_n)$ a pour espérance $\sigma_✓^2$: comme cette variance bessélisée est une statistique, elle constitue donc un estimateur par tendance de la variance σ^2 intrinsèque à la nouvelle méthode. ♧

11.8 Technique de substitution

Principe de substitution

Nous venons de présenter un certain nombre de méthodes d'estimation ; mais en pratique, dès que le modèle devient un peu subtil, il est rare qu'aucune de ces méthodes nous donne directement (et de façon simple à calculer) accès à un estimateur de la quantité d'intérêt qui nous intéresse... C'est ici qu'intervient une idée aussi simple que puissante, appelée *substitution* :

Définition (YP) (Obtention de nouveaux estimateurs par substitution)**.** Soit un modèle statistique avec les notations habituelles, pour lequel nous nous intéressons à la quantité d'intérêt $\gamma(\boldsymbol{\theta}) \in \mathcal{G}$. Soient $(\hat{\psi}_i(X))_{i\in I}$ des estimateurs (qui peuvent avoir été obtenus par n'importe quelles techniques) des quantités d'intérêt respectives $(\psi_i(\boldsymbol{\theta}))_{i\in I}$ (nous appellerons $\mathcal{G}_i$ l'espace où vit la quantité d'intérêt $\psi_i(\boldsymbol{\theta})$) ; et supposons maintenant qu'il existe une fonction (pas trop biscornue) $F\colon \prod_{i\in I} \mathcal{G}_i \to \mathcal{G}$ telle que, pour tout $\theta \in \boldsymbol{\Theta}$, on ait

$$\gamma(\theta) = F\big((\psi_i(\theta))_{i\in I}\big). \tag{YQ}$$

On qualifie alors la statistique

$$\hat{\gamma}(X) \coloneqq F\big((\hat{\psi}_i(X))_{i\in I}\big) \tag{YR}$$

d'*estimateur par substitution* de $\gamma(\boldsymbol{\theta})$. ♡

Remarque (YS). J'ai écrit ce qui précède dans le cadre de l'estimation, mais le principe de substitution s'étend directement au cas des prédicteurs ! ♧

Remarque (YT). Ce qui rend la technique de substitution tellement puissante est que, au lieu d'être obligés de partir de notre quantité d'intérêt $\gamma(\boldsymbol{\theta})$ et de devoir "cuisiner" pour trouver *le* bon estimateur $\hat{\gamma}(X)$ qui se trouvera par miracle estimer précisément notre quantité d'intérêt, nous pouvons procéder "dans l'autre sens" en choisissant plus ou moins arbitrairement des statistiques $\hat{\psi}_1(X), \hat{\psi}_2(X), \dots$ qui sont faciles à calculer, en regardant de quelles quantités d'intérêt $\psi_1(\boldsymbol{\theta}), \psi_2(\boldsymbol{\theta}), \dots$ ces statistiques se trouvent être des estimateurs respectifs (en particulier, si on veut que $\hat{\psi}_i(X)$ soit un estimateur par tendance, la fonction ψ_i sera déterminée par l'égalité $\psi_i(\theta) = \mathbb{E}_\theta(\hat{\psi}_i(X))$; et, *seulement alors*, on regardera s'il y a un moyen de combiner les $\psi_i(\boldsymbol{\theta})$ pour en déduire notre quantité d'intérêt $\gamma(\boldsymbol{\theta})$ (via une relation du type $F(\psi_1(\theta), \psi_2(\theta), \dots) = \gamma(\theta)$), ce qui nous donnera alors par substitution l'estimateur $F(\hat{\psi}_1(X), \hat{\psi}_2(X), \dots)$ pour $\gamma(\boldsymbol{\theta})$. ♧

Exemple (YU). Voici un exemple d'utilisation de la substitution pour le modèle du pédagogue. On prend ici $n = 4$. (On a alors $\mathcal{X} = \mathbb{R}^4$). Nous allons considérer les deux fonctions suivantes sur $\mathcal{X}$:

- $\hat{\psi}_1(x_1, x_2, x_3, x_4) \coloneqq \max_{i\in\{1,\dots,4\}} x_i$, autrement dit la meilleure note obtenue parmi les quatre élèves ;
- $\hat{\psi}_2(x_1, x_2, x_3, x_4) \coloneqq \max(\{\{x_1, x_2, x_3, x_4\}\} - \{\{\max_{i\in\{1,\dots,4\}} x_i\}\})$, autrement dit la seconde meilleure note (qui peut éventuellement être identique à la première s'il y a ex-æquo).

On peut montrer que, dans le cas de notre modèle, lorsque X suit la loi Normale$(\mu_\checkmark, \sigma_\checkmark^2)^{\otimes 4}$, les espérances respectives de $\hat{\psi}_1(X)$ et $\hat{\psi}_2(X)$ sont

$$\mathbb{E}_\checkmark(\hat{\psi}_1(X)) = \mu_\checkmark + C_1\sigma_\checkmark\,;$$
$$\mathbb{E}_\checkmark(\hat{\psi}_2(X)) = \mu_\checkmark + C_2\sigma_\checkmark,$$

où C_1 et C_2 sont des constantes numériques qu'on peut calculer et dont les valeurs respectives sont 1,029 375 et 0,297 011. Par conséquent, $\hat{\psi}_1(X)$ et $\hat{\psi}_2(X)$ sont des estimateurs par tendance resp. $\psi_1 \coloneqq \mu + C_1\sigma$ et $\psi_2 \coloneqq \mu + C_2\sigma$. Maintenant, supposons que notre quantité d'intérêt soit μ. On observe qu'on peut récupérer $\mu_\checkmark$ à partir de $\psi_{1\checkmark}$ et $\psi_{2\checkmark}$ par la formule

$$\mu_\checkmark = \frac{C_1}{C_1 - C_2}\psi_{2\checkmark} - \frac{C_2}{C_1 - C_2}\psi_{1\checkmark}. \tag{YV}$$

Or, puisque nous savons estimer $\psi_{1\checkmark}$ et $\psi_{2\checkmark}$ par resp. $\hat{\psi}_1(X)$ et $\hat{\psi}_2(X)$, nous en déduisons par substitution l'estimateur suivant pour μ :

$$\hat{\mu}(X) \coloneqq \frac{C_1}{C_1 - C_2}\hat{\psi}_2(X) - \frac{C_2}{C_1 - C_2}\hat{\psi}_1(X). \tag{YW}$$

Cet estimateur illustre à merveille l'idée derrière la substitution, puisque nous avons *décidé* de la construire à partir de $\hat{\psi}_1(X)$ et $\hat{\psi}_2(X)$ (qui sont ce qu'on appelle les *statistiques d'ordre*, de co-rangs respectifs 1 et 2, de notre observation) n'est à vrai dire pas très performant ; mais il a l'avantage de ne demander que la connaissance des deux meilleurs notes et pas des autres ! ♧

Remarque (YX). L'estimateur présenté dans l'exemple ci-dessus semble assez grossièrement sous-optimal, dans la mesure où il n'utilise que les deux meilleures notes de la promotion : et de fait, ce n'est pas un estimateur ayant de très bonnes performances ! ∴ Néanmoins, ce genre d'estimateurs peut s'avérer très utiles dans certains domaines (je pense par exemple aux questions de performance sportive) où il est facile de trouver les données concernant les champions, mais difficile de trouver des données précises concernant la population générale... ♧

Méthode des moments

Un cas particulièrement important de technique de substitution, qui permet d'obtenir des estimateurs faciles à calculer dans un grand nombre de cas, est la *méthode des moments*, exposée ci-dessous :

Définition (YY). Supposons que nous soyons en présence d'un modèle relevant du point (XN) (càd. où l'observation est constituée de sous-observations qui sont de même loi sous les contextes fréquentistes), et pour lequel les sous-observations X_i sont en outre à valeurs réelles. Appelons k la *dimension* de l'espace Θ du paramètre caché de ce modèle, càd. le nombre de coordonnées dont on a besoin pour se repérer dans Θ [∗]. Soit $\gamma \coloneqq \gamma(\theta)$ une quantité d'intérêt : la *méthode des moments* consiste alors à estimer γ en procédant par substitution à partir des k premiers *moments empiriques* $\hat{\psi}_1, \dots, \hat{\psi}_k$ des observations, lesquels sont définis par

$$\hat{\psi}_r(\vec{X}_{[\![0,n]\!]}) \coloneqq \frac{1}{n}\sum_{i=0}^{n-1} X_i^r \ : \tag{YZ}$$

le moment empirique $\hat{\psi}_r(\vec{x}_\checkmark)$ pouvant être vu comme une estimation [†] de (la variable aléatoire dont la réalisation est) $\mathbb{E}_\checkmark(X_0^r)$, qui est ce qu'on appelle le *r-ième moment* de la loi $\mathrm{Loi}_\checkmark(X_0)$. ♡

[∗]. Par exemple, la surface de la Terre est de dimension 2 puisqu'on peut s'y repérer par les latitude et longitude, et ce, malgré que la Terre n'est pas plane.

[†]. En l'occurrence, il s'agit à la fois d'une estimation empirique et d'une estimation par tendance.

Remarque (AA'). Dans certains cas, il peut arriver qu'à cause des symétries du problème, un des moments de $\mathrm{Loi}_{\checkmark}(X_0)$ soit en fait une fonction des moments précédents (et ce, indépendamment de la valeur $\theta_{\checkmark}$, s'entend : dans ce cas, ce moment n'apporte aucune information pour la substitution, et il faut donc le "sauter" — et aller alors jusqu'au $(k+1)$-ième moment. ♧

Remarque (AB'). Je n'exigerai pas de vous que vous reteniez par cœur la méthode des moments : cependant, comme il s'agit d'un outil très classique en statistique, il est bon que vous l'ayez vue à l'occasion de ce cours, et que vous sachiez la reconnaitre si elle intervient dans vos exercices ! ◡̈ ♧

Exemple (AC'). Voyons ce que donne la méthode des moments dans l'exemple du pédagogue, en prenant pour quantité d'intérêt $\boldsymbol{\mu}-\boldsymbol{\sigma}$. Le premier moment empirique de nos observations est

$$\hat{\psi}_1(\vec{X}) \coloneqq \frac{1}{n}\sum_{i=0}^{n-1} X_i = \mathrm{moy}(X_0,\dots,X_{n-1}) : \qquad (\mathsf{AD}')$$

il s'agit d'un estimateur empirique de la quantité d'intérêt (dont la réalisation est) $\mathbb{E}_{\checkmark}(X_0) = \mathbb{E}\big(\mathrm{Normale}(\mu_{\checkmark},\sigma^2_{\checkmark})\big) = \mu_{\checkmark}$. (C'en serait aussi un estimateur par tendance, du reste).

Le second moment empirique des nos observations est

$$\hat{\psi}_2(X) \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathrm{var}_{\mathrm{emp}}(X_1,\dots,X_n) + \mathrm{moy}(X_1,\dots,X_n)^2 : \qquad (\mathsf{AE}')$$

il s'agit d'un estimateur empirique (et par tendance) de $\mathbb{E}_{\checkmark}(X_1^2) = \mathbb{E}\big(\mathrm{Normale}(\mu_{\checkmark},\sigma^2_{\checkmark})^2\big) = \mu_{\checkmark}+\sigma^2_{\checkmark}$.

Nous avons donc trouvé deux estimateurs des quantités d'intérêt resp. $\boldsymbol{\mu}$ et $\boldsymbol{\mu}+\boldsymbol{\sigma}^2$, à savoir, $\hat{\boldsymbol{\psi}}_1$ et $\hat{\boldsymbol{\psi}}_2$. Nous voulons maintenant procéder par *substitution* en trouvant une moyen de "combiner" ces quantités d'intérêt pour obtenir $\boldsymbol{\mu}-\boldsymbol{\sigma}$ (qui est la quantité d'intérêt qui nous intéresse vraiment). Dit en termes de réalisation, il s'agit donc de trouver un moyen de déduire $\mu_{\checkmark}-\sigma_{\checkmark}$ à partir de $\mu_{\checkmark}$ et $\mu_{\checkmark}+\sigma^2_{\checkmark}$. (Pour rendre les notations plus efficaces, dans la suite, nous poserons resp. $\boldsymbol{\mu}_{\checkmark} =: \psi_{1\checkmark}$ et $\boldsymbol{\mu}_{\checkmark} =: \psi_{2\checkmark}$).

En substance, il s'agit donc d'"inverser" la fonction $\boldsymbol{\Theta} \ni (\mu,\sigma) \mapsto (\mu,\mu+\sigma^2)$: comme aussi bien l'espace de départ que l'espace d'arrivée sont de dimension 2, cela a à priori de bonnes chances d'être possible (et, de fait, ce sera bien le cas ◡̈). On trouve immédiatement que $\sigma^2_{\checkmark} = \psi_{2\checkmark} - \psi^2_{1\checkmark}$, d'où $\sigma_{\checkmark} = (\psi_{2\checkmark}-\psi^2_{1\checkmark})^{1/2}$ (puisque $\sigma_{\checkmark}$ est positif), et finalement $\mu_{\checkmark}-\sigma_{\checkmark} = \psi_{1\checkmark} - (\psi_{2\checkmark}-\psi^2_{1\checkmark})^{1/2}$. Par substitution, on en déduit finalement notre estimateur des moments pour $\boldsymbol{\mu}-\boldsymbol{\sigma}$: c'est

$$\hat{\gamma}(X) \coloneqq \hat{\psi}_1(X) - (\hat{\psi}_2(X) - \hat{\psi}_1(X)^2)^{1/2} = \mathrm{moy}(X_1,\dots,X_n) - \mathrm{var}^{1/2}_{\mathrm{emp}}(X_1,\dots,X_n). \quad (\mathsf{AF}')$$

(Noter qu'en l'occurrence, on voit qu'il s'agit aussi de l'estimateur empirique obtenu à partir de la fonctionnelle « moyenne moins écart-type »). ♧

Chapitre 12

Tests d'hypothèses

12.1 Motivation du concept

Exemple introductif

Dans cette sous-section, nous reprenons le contexte et les notations du modèle du pédagogue : en particulier, nous nous intéressons à la nouvelle méthode pédagogique mise au point par un enseignant, et à l'impact de celle-ci sur les résultats des élèves. À ce stade, considérons que la nouvelle méthode vient juste d'être conçue, et n'a pas encore été testée en conditions réelles. Notre enseignant expose sa méthode à ses collègues, et un désaccord entre les deux camps apparait :

— D'un côté, notre enseignant pense que la nouvelle méthode va globalement faire progresser les élèves par rapport à l'ancienne (ou, au pire, ne rien changer) : autrement dit, il pense que $\mu_{\checkmark}$, la véritable espérance des scores pour sa nouvelle méthode est supérieure ou égale à $\mu_{\text{réf}}$;

— À l'inverse, les collègues estiment que la nouvelle méthode va s'avérer strictement nuisible : autrement dit, selon eux, la valeur $\mu_{\checkmark}$ est donc strictement inférieure à $\mu_{\text{réf}}$!

Par ailleurs, dans le cadre de cette section, nous allons considérer qu'aussi bien l'enseignant que ses collègues sont d'accord pour dire que la variabilité des résultats d'un élève à l'autre ne sera pas impactée par le changement pédagogique : autrement dit, tous sont d'accord pour dire que la valeur $\sigma_{\checkmark}$ est égale à $\sigma_{\text{réf}}$. De la sorte, l'espace dans lequel vit le paramètre caché $(\boldsymbol{\mu}, \boldsymbol{\sigma})$ n'est pas l'espace originel $\mathbb{R} \times \mathbb{R}_+^*$ (que nous noterons dorénavant $\boldsymbol{\Theta}^{\text{orig}}$ lorsqu'il faudra éviter toute ambigüité), mais le sous-espace $\mathbb{R} \times \{\sigma_{\text{ref}}\} =: \boldsymbol{\Theta}^{\text{hsc}}$ [∗], que nous noterons simplement $\boldsymbol{\Theta}$ dans la suite de cette sous-section. (Gardez néanmoins à l'esprit que ce n'est pas exactement le même '$\boldsymbol{\Theta}$' que dans l'exemple `(MC)` d'origine : ici, seul $\boldsymbol{\mu}$ est réellement inconnu).

La question qui se pose est : comment trancher le débat ? Par l'expérience, évidemment ! On va expérimenter la nouvelle méthode sur une promotion, et s'appuyer sur les résultats obtenus par les élèves pour dire si on a plutôt envie de pencher dans le sens de l'enseignant, ou plutôt dans celui de ses collègues. Mais comment, concrètement, procéder : quel traitement précis allons-nous appliquer aux notes pour déterminer qui a raison ?

Dans le contexte considéré, où les deux thèses soutenues ne diffèrent que par le résultat moyen espéré pour les élèves, il est intuitivement évident que la "bonne"

[∗]. 'hsc' est une abréviation pour « homoscédastique » : ce que signifie cet adjectif vous sera expliqué plus tard.

façon de procéder pour trancher entre ces deux thèses devrait être la suivante : on va regarder quelle est la moyenne $m_{\checkmark}$ effectivement obtenue par les élèves [il s'agit donc de la réalisation de la statistique $M := \left(\sum_{i=0}^{n-1} X_i\right) / n$] : plus $m_{\checkmark}$ sera élevée, plus on aura envie de trancher en faveur de l'enseignant ; et plus elle sera basse, plus on aura envie de trancher en faveur de ses collègues ! À ce stade, on a même envie d'ajouter (mais nous ne le ferons pas) qu'on voudrait trancher en faveur de l'enseignant si $m_{\checkmark}$ s'avère plus grande que $\mu_{\text{réf}}$, et en sa défaveur dans le cas contraire...

... Cela dit, autant notre enseignant est d'accord avec l'idée d'utiliser la moyenne de la promotion pour jauger l'efficacité de sa méthode, autant il considère que fixer le seuil à $\mu_{\text{réf}}$ n'est pas la bonne chose à faire. « En effet, explique-t-il, une moyenne effective de 98 serait certes plus en phase avec la théorie des collègues qu'avec la mienne ; mais après tout, elle peut aussi tout à fait s'expliquer par le fait qu'on ait en réalité $\mu_{\checkmark} = 103$, et que le hasard ait fait qu'on soit tombé sur un promotion un peu faible cette année-là : dès lors, je refuserais de remettre en cause ma méthode sur foi de ce seul résultat... ! ». Têtu, notre enseignant ? Assurément ! ◡̈ Néanmoins, ses collègues reconnaissent que, en l'occurrence, il est effectivement légitime de traiter les deux thèses de façon *dissymétrique*. En effet, *in fine*, le point n'est pas seulement de savoir qui a raison, mais de savoir quelle méthode pédagogique il convient d'utiliser. Or, à supposer que la nouvelle méthode pédagogique soit bel et bien meilleure, l'abandonner à tort porterait un préjudice à *toutes* les générations futures d'élèves qui auraient pu en bénéficier ! Alors que si la nouvelle méthode est moins bonne, mais qu'on reconduit l'expérimentation quelques années avant de l'abandonner définitivement, cela aura certes été nuisible aux quelques promotions "cobayes", mais c'est tout de même moins grave... ◠ Bref ; pour prendre en compte ces considérations, on souhaite avoir une procédure qui conclura en faveur des collègues *seulement* si on a une *forte conviction* qu'ils ont raison, et en faveur de l'enseignant dès lors qu'il reste *plausible*, au vu des notes des élèves, que ce soit lui qui ait raison ◡̈

Après discussion, l'enseignant et ses collègues choisissent de fixer leur seuil de décision à $95 =: \mu_{\text{seuil}}$: si la moyenne de la promotion s'avère supérieure ou égale cette valeur, ils laisseront l'enseignant continuer d'expérimenter sa nouvelle méthode (fût-ce "au bénéfice du doute" si la moyenne est comprise entre 95 et 100) ; mais si elle s'avère strictement inférieure, l'enseignant conviendra que c'est une preuve forte que sa méthode est en fait mauvaise, et qu'il convient donc d'arrêter là les frais... ! Pour fixer ce seuil à 95, l'enseignant et ses collègues ont raisonné ainsi. « Ce qu'on veut, c'est que si la thèse de l'enseignant est vraie, on n'ait qu'un risque très faible de se tromper en jugeant en sa défaveur. Plaçons-nous dans le cas où la thèse de l'enseignant ne serait vraie que "de justesse", avec, mettons, $\mu_{\checkmark} = 100{,}1$. Dans ce cas, la v.a. $M := (X_1 + \cdots + X_n) / n$ suivra (sous le contexte probabiliste $\mathbb{P}_{\checkmark}$ correspondant au $\mu_{\checkmark}$ en question) la loi Normale$(\mu_{\checkmark}, \sigma^2_{\text{réf}} / n)$. Or la probabilité qu'une telle v.a. soit inférieure à μ_{seuil} peut être calculée comme valant environ 5 % :

```
> n = 22
> sigma = 15
> muref = 100
> mucoche = muref + 0.1
> museuil = 95
> pnorm(museuil, mucoche, sqrt(sigma ^ 2 / n))
[1] 0.05538504
```

Ainsi, si notre enseignant a effectivement raison, il y a au pire environ 5 % de risque qu'on rejette erronément sa thèse. C'est faible, ce qui est voulu : en effet, comme nous l'avons fait remarquer ci-dessus, on veut avoir une conviction *forte* que ce sont les collègues qui ont raison avant de trancher en leur faveur ; or, on n'aurait pas pu raisonnablement qualifier notre conviction de "forte" s'il y avait eu, mettons, 15 % de risque de rejeter erronément la thèse de l'enseignant. D'un autre côté, ce seuil n'est pas non plus *trop* faible : ce qui est souhaitable aussi, car il faut tout de même donner une chance à la procédure de trancher en faveur des collègues si ce sont eux qui ont raison ! ». [Ainsi, on aurait certes pu fixer le seuil à 90, auquel cas il y aurait eu moins d'un risque sur 1 000 de rejeter à tort la thèse de l'enseignant ; mais ç'aurait été là requérir un niveau de preuve peu raisonnable : pas suffisamment raisonnable en tout cas pour justifier qu'on continue d'accorder à l'enseignant le bénéfice du doute pour une moyenne comprise entre 90 et 95, dans la mesure où une moyenne aussi faible à peine 5 % de chances d'arriver si l'enseignant a raison...].

Introduction du vocabulaire

Dans cette sous-section nous allons reprendre la situation discutée dans la sous-section précédente, mais en soulignant explicitement cette fois-ci les concepts qui y ont été utilisés sans le dire, et en introduisant le vocabulaire que les statisticiens utilisent pour désigner ces concepts :

☛ L'affirmation de l'enseignant et celle de ces collègues correspondent à des *hypothèses* (ce vocabulaire a déjà été introduit dans la § 9.3), autrement dit, à des quantités d'intérêt explicatives booléennes. Plus précisément, l'hypothèses défendue par l'enseignant est que $\{\boldsymbol{\mu} \geqslant \mu_{\text{réf}}\}$, tandis que celle défendue par ses collègues est que $\{\boldsymbol{\mu} < \mu_{\text{réf}}\}$. Dans la suite, nous noterons ces hypothèses resp. $\mathscr{H}_0$ et $\mathscr{H}_1$.

☛ On a plus précisément considéré deux hypothèses *complémentaires* l'une de l'autre : $\mathscr{H}_0$ et $\mathscr{H}_1$ sont chacune la négation de l'autre (ce qu'on note formellement « $\mathscr{H}_{1-i} = \neg\mathscr{H}_i$ »).

☛ L'hypothèse de l'enseignant est celle qu'on veut ne récuser que lorsqu'on a de *fortes* preuves en sa *défaveur* : dans ce cas, on la qualifie d'*hypothèse nulle* (d'où la notation '$\mathscr{H}_0$'). [†]

☛ Complémentairement, l'hypothèse des collègues est celle qu'on ne veut valider que lorsqu'on a de *fortes* preuves en sa *faveur* : dans ce cas, on a la qualifie d'*hypothèse alternative*. [Les hypothèses alternatives sont généralement notées à l'aide de l'indice '1'].

☛ La moyenne $\boldsymbol{M}$ des élèves (dont $m_{\checkmark}$) est la réalisation est l'indicateur qu'on a choisi de regarder pour trancher entre des hypothèses, indicateur qui correspond à une *statistique* (puisqu'il est obtenu directement à partir des données) à valeurs *réelles*. Un tel indicateur est qualifié de *statistique de test*.

☛ Ici on choisit de trancher en faveur de l'hypothèse alternative lorsqu'on se situe *en-deçà* d'un certain seuil : on dit dans ce cas qu'on procède à un « *test à gauche* ». (Si on avait au contraire utilisé un seuil pour trancher en direction de l'hypothèse alternative *au-delà* dudit seuil, on aurait parlé de « test à droite »). De

[†]. La raison que nous avons invoquée dans la § 12.1 pour traiter de façon dissymétrique les hypothèses nulle et alternative est l'idée qu'il serait plus grave de rejeter la thèse de l'enseignant à tort, que de l'accepter à tort. En réalité, c'est une raison possible au fait de traiter les hypothèses nulle et alternative de façon dissymétrique, mais en réalité la raison *fondamentale* pour laquelle la théorie des tests est fondamentalement dissymétrique n'est pas celle-là ! Cela sera discuté ultérieurement, confer § **??**.

manière équivalente, on pourra dire que notre « *critère de suspicion* » consiste à considèrer les valeurs de la statistique de test comme d'autant plus suspectes (par rapport à l'hypothèse nulle, s'entend) qu'elles sont situées vers la gauche.

☛ La procédure à laquelle on est arrivé correspond à trancher soit dans un sens, soit dans l'autre, selon ce que vaut l'observation (via la valeur de la statistique de test qui s'en déduit). On peut donc la décrire comme une statistique à valeurs dans {VRAI, FAUX}, où FAUX signifie qu'on penche vers l'hypothèse nulle, tandis que VRAI signifie qu'on penche vers l'hypothèse alternative [‡] : formellement, cette statistique s'écrirait donc $\{\boldsymbol{M} < \mu_{\text{seuil}}\}$. Une telle statistique booléenne est qualifiée de *test (booléen)*.

☛ En fait, ce qu'on qualifie de « test (booléen) », c'est plutôt la *procédure de jugement* décrite par la variable aléatoire que vous venons d'évoquer, plutôt que la variable aléatoire elle-même : autrement dit, notre « test » consiste ici à « conclure en faveur des collègues si $\boldsymbol{M} < \mu_{\text{seuil}}$, et en faveur de l'enseignant sinon ».

☛ Lorsqu'on conclut en faveur de l'hypothèse alternative, on dit qu'on *rejette l'hypothèse nulle* (dans la mesure où on a trouvé un résultat qui n'est vraiment pas compatible avec celle-ci) ; on peut aussi dire, éventuellement, qu'on *valide l'hypothèse alternative* (au sens où on a apporté une quasi-preuve de ce que c'est elle qui est correcte). Dans un tel cas, on dit que le test est *positif* (au sens où sa réalisation vaut VRAI).

☛ Lorsque, au contraire, on conclut en faveur de l'hypothèse nulle, on dit qu'on *accepte* l'hypothèse nulle. L'emploi du mot « accepter » signifie que cela n'implique pas pour autant qu'il *faille* rejeter l'hypothèse alternative, juste que l'hypothèse nulle est acceptable. Pour bien souligner cet aspect non exclusif, on peut aussi dire qu'on « ne rejette pas » l'hypothèse nulle, ou qu'on « n'est pas en mesure de valider » l'hypothèse alternative [§]. Dans ce cas, on dit que le test est *négatif* [¶].

☛ La valeur d'environ 5 % que nous avons trouvée correspond à la probabilité qu'on a de conclure positivement (donc en rejetant l'hypothèse nulle) lorsque l'hypothèse nulle est vraie, et plus précisément, dans le "pire" cas où l'hypothèse nulle est vraie [ou presque : pour nos calculs, nous avons en effet pris une valeur de 100,1, alors que le pire cas aurait été de prendre *précisément* $\mu_{\checkmark} = \mu_{\text{réf}}$]. C'est (modulo le fait que nous n'ayons pas *exactement* considéré le pire cas) il s'agit de ce que les statisticiens qualifient de *niveau* du test.

[‡]. Mnémotechniquement, rappelons qu'il est d'usage d'associer FAUX et VRAI aux valeurs respectives 0 et 1.

[§]. Attention : « Ne pas être en mesure de valider » $\mathscr{H}_1$, cela ne signifie pas pour autant qu'on *invalide* cette hypothèse : juste qu'on n'est pas en état d'en avoir une confirmation forte !

[¶]. Rappelez-vous que l'idée est de voir si on a assez d'éléments pour rejeter l'hypothèse nulle avec conviction. Vu sous cet angle, le fait de ne pas être en mesure de rejeter est une *absence* de conclusion ferme, d'où l'adjectif « négatif » ; tandis que le fait de rejeter est une affirmation désignant une véritable conclusion, d'où l'adjectif « positif ». Par ailleurs, il est logique de parler de résultat « positif » quand on penche en faveur de $\mathscr{H}_1$ (puisque 1 est un nombre strictement positif !), resp. de résultat « négatif » quand on penche en faveur de $\mathscr{H}_0$ (qui, au sens large, est bien un nombre négatif ☺)

12.2 Formalisation du concept dans le cas unilatéral

Définition formelle

Le moment est à présent venu de définir formellement la notion de test (booléen) :

Définition (AG') (Test booléen)**.** Avec les notations standard, soit $\Theta_1 \subseteq \Theta$ un sous-ensemble de l'espace du paramètre caché, permettant de définir une hypothèse associée $\{\theta \in \Theta_1\} =: \mathscr{H}_1$. Dans ces conditions, on appelle *test de l'hypothèse alternative* $\mathscr{H}_1$ une statistique à valeurs booléenne, autrement dit une v.a. *Test* de la forme $test(X)$ à valeurs dans $\{\mathsf{VRAI}, \mathsf{FAUX}\}$, qui "essaye" de dire si l'hypothèse $\mathscr{H}_1$ est correcte ou non. !

De façon équivalente, on dit aussi que *Test* est un *test de l'hypothèse nulle* $\mathscr{H}_0$, où $\mathscr{H}_0$ est l'hypothèse complémentaire de $\mathscr{H}_1$, autrement dit, $\mathscr{H}_0 := \neg\mathscr{H}_1 = \{\theta \in \Theta_0\}$, où $\Theta_0 := \Theta \setminus \Theta_1$ est le sous-ensemble complémentaire de Θ_1.

Lorsque la réalisation de $test(x_{\checkmark})$ du test vaut VRAI, on interprète ce résultat en disant que « l'hypothèse nulle est rejetée », ou que « l'hypothèse alternative est validée », ou encore que « le test est positif ». Lorsque le résultat du test vaut FAUX, on interprète ce résultat en disant que « l'hypothèse nulle est acceptée », ou que « l'hypothèse alternative demeure indécise », ou encore que « le test est négatif ».

On appelle *niveau* du test (traditionnellement noté 'α') la probabilité supremale que le test soit positif alors que l'hypothèse nulle est vraie :

$$\alpha := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(Test)\ ^{[\|]}. \qquad (\mathsf{AH}')$$

On dit que le test est « de niveau α » lorsque son niveau est $\leqslant \alpha$ [**]. ♡

Cette définition appelle un certain nombre de commentaires :

Remarque (AI')*.* Les hypothèses nulle et alternative étant par définition complémentaires l'une de l'autre, en général on ne définira qu'une des deux lorsqu'on parlera d'un test, l'autre s'en déduisant automatiquement. ♧

Remarque (AJ')*.* Comme vous le voyez, les hypothèses nulle et alternative jouent des rôles bien distincts au niveau de l'interprétation du résultat et de la définition

[‖]. Notez que, dans la mesure où v.a. *Test* est *booléenne*, on peut bien parler de la probabilité de la v.a. *elle-même*, puisque dans un tel cas l'évènement *Test* est strictement équivalent à $\{Test = \mathsf{VRAI}\}$: en effet, l'évènement $\{Test = \mathsf{VRAI}\}$ est l'évènement qui est vrai lorsque *Test* vaut VRAI, et faux sinon (sachant que, en tant que v.a. booléenne, lorsque *Test* ne vaut pas VRAI, elle vaut nécessairement FAUX) : par conséquent, cet évènement vaut VRAI lorque *Test* vaut VRAI, resp. FAUX lorsque *Test* vaut FAUX : c'est donc bien *Test* elle-même ! Bien sûr, si cela vous dérange trop d'écrire une formule avec une variable aléatoire "nue", sans signe de relation, vous *pouvez* écrire « $\mathbb{P}_\theta(Test = \mathsf{VRAI})$ » plutôt que « $\mathbb{P}_\theta(Test)$ » à l'intérieur d'une probabilité ; mais notez l'aspect pléonastique de la chose ! N'oubliez pas en effet que, même si on ne vous l'a pas dit ainsi dans vos cours précédents, un évènement n'est autre qu'une variable aléatoire booléenne : on peut certes *aussi* le voir comme un sous-ensemble de l'univers Ω ; néanmoins, il est plus cohérent pour l'intuition de dire que, quand on parle de l'évènement {la somme des deux dés vaut au moins 10}, on parle bien de quelque chose qui vaut soit VRAI, soit FAUX, selon ce que vaut la somme des deux dés : c'est là une façon de voir les choses qui est beaucoup plus naturelle que de prétendre qu'on serait en fait en train de parler de l'ensemble des éventualités ω de l'univers pour lesquelles la valeur de la variable aléatoire « somme des deux dés » atteint ou dépasse 10...

[**]. Attention donc : dire « le niveau du test est α » n'est pas exactement synonyme de « le test est de niveau α ».

du niveau. Par conséquent, veillez à ne jamais parler de « test de l'hypothèse $\mathscr{H}$ », mais à toujours préciser si $\mathscr{H}$ joue le rôle de l'hypothèse nulle ou de l'hypothèse alternative : en disant soit « test de l'hypothèse *nulle* $\mathscr{H}$ » (ce qui est équivalent à « test de l'hypothèse alternative $\neg\mathscr{H}$ »), soit « test de l'hypothèse *alternative* $\mathscr{H}$ » (ce qui est équivalent à « test de l'hypothèse nulle $\neg\mathscr{H}$ »). ♣

!! *Remarque* (AK') (Dissymétrie des conclusions d'un test). Attention ! « Accepter » l'hypothèse nulle ne signifie pas qu'on a de quoi affirmer avec confiance que l'hypothèse nulle est vraie : cela signifie juste qu'on trouve qu'il est *acceptable* de croire en l'hypothèse nulle au vu des résultats obtenus — mais peut-être que l'hypothèse alternative est tout aussi acceptable ! À l'inverse, « rejeter » l'hypothèse nulle signifie qu'on trouve que cette hypothèse n'est vraiment pas crédible au vu des données. Il y a donc une forte dissymétrie : « accepter l'hypothèse nulle » n'est absolument pas synonyme de « rejeter l'hypothèse alternative » ! Pour cette raison, certains auteurs préfèrent écrire qu'on « ne rejette pas » l'hypothèse nulle (c'est pourquoi on parle de test *négatif* dans ce cas), le terme « accepter » étant jugé trop ambigu. ♣

! *Remarque* (AL'). Notez bien que, dans la formule définissant le niveau d'un test, seule l'hypothèse *nulle* intervient ! Ce qui se passe dans le cas où $\theta_{\checkmark}$ est dans Θ_1 n'est tout simplement pas considéré ici. ♣

!! *Remarque* (AM'). En fait, la remarque précédente provient de ce que, en substance, le test répond à la question « l'observation est-elle *compatible* avec l'hypothèse nulle », *sans* se poser la question de l'hypothèse alternative ! (Ou alors, seulement de façon indirecte, via le choix d'une statistique de test pertinente : confer § 12.4 *infra*). ♣

Remarque (AN'). La notion de niveau du test est fondamentalement *fréquentiste* : et, de manière générale, toute la théorie des tests d'hypothèse nulle se place entièrement dans le paradigme fréquentiste ! On peut d'ailleurs comparer la notion de niveau d'un test avec le "niveau bayésien" dans le cadre bayésien (définition (RZ)) :

— Le niveau *bayésien* était le supremum, sur les valeurs possibles de l'*observation* correspondant à un test *positif*, de la probabilité que le test se trompe (autrement dit, qu'on soit en réalité sous l'hypothèse *nulle*), probabilité qui est considérée dans le contexte *sachant la valeur correspondante de l'observation*. (Il s'agit donc d'un contexte de nature bayésienne : le contexte probabiliste $\mathbb{P}(\bullet \mid X = x)$ étant, en quelque sorte, "la probabilité à postériori dans le cas où l'observation vaut x").

— Le niveau *fréquentiste* est aussi un supremum de la probabilité que le test se trompe en répondant « positif » dans un cas où c'est l'hypothèse nulle qui est vraie : néanmoins, cette fois-ci on considère le supremum sur les valeurs possibles du *paramètre caché* correspondant à l'hypothèse *nulle*, et la probabilité considérée est celle que le terst réponde *positif*, probabilité qui est considérée dans le contexte *sachant la valeur correspondante du paramètre caché*. (Il s'agit donc d'un contexte de nature *fréquentiste* : c'est le contexte $\mathbb{P}_\theta$, qui correspond, en quelque sorte, à "la véritable probabilité dans le cas où le paramètre caché vaut θ").

Ainsi, en passant du paradigme bayésien au paradigme fréquentiste, on a été obligés de considérer des contextes probabilistes différents (dans la mesure où on ne peut plus parler du contexte à postériori), ce qui nous a amenés à inverser les rôles joués par le paramètre caché et l'observation. ♣

Remarque (AO'). Ici j'ai simplement qualifié α de « niveau » ; mais il faudrait en toute rigueur parler de « niveau *de risque* » : et on utilise, à l'inverse, l'expression « niveau *de confiance* » pour désigner $1 - \alpha$. Néanmoins, en pratique, on choisit toujours des niveaux de risque inférieurs ou égaux à 1/2, de sorte qu'il n'y a aucun

risque d'ambiguïté : si on parle d'un test « de niveau 92 % », cela se réfèrera forcément à un niveau de confiance (auquel cas $\alpha = 8$ %) ; alors que, lorsqu'on parle d'un test « de niveau 20 % », il s'agira à l'inverse d'un niveau *de risque*.

À noter, au passage, que la notation traditionnelle 'α' sous-entend qu'on parle d'un niveau *de risque* : les niveaux de confiance, eux, sont traditionnellement notés $1 - \alpha$. ♧

Remarque (AP'). Pour déterminer le niveau d'un test, il est essentiel d'être capable de définir l'hypothèse nulle $\mathscr{H}_0$ sous la forme $\{\theta \in \Theta_0\}$ (ce qui est toujours possible, par définition de ce qu'est une hypothèse), puisque le sous-espace Θ_0 intervient dans le supremum définissant le niveau. Et notez bien que ce supremum portera sur le paramètre caché *lui-même*, par sur une quantité d'intérêt explicative : par exemple, si nous prenons le modèle du pédagogue avec l'espace originel $\Theta^{\text{orig}} \underset{\text{déf}}{=} \mathbb{R} \times \mathbb{R}_+^*$ du paramètre caché, et que nous considérons l'hypothèse nulle $\{\sigma \geqslant 0{,}7\,\sigma_{\text{réf}}\}$, alors le supremum pour la définition du niveau de sera pas un supremum sur la seule variable σ, mais bien sur le *couple* (μ, σ) constituant le paramètre caché, quand bien même μ n'intervient pas explicitement dans la définition de l'hypothèse nulle ! — Plus précisément, on aura alors que

$$\alpha \underset{\text{déf}}{=} \sup\{\mathbb{P}_{\mu,\sigma}(\mathit{Test}) \mid \mu \in \mathbb{R}, \sigma \geqslant 0{,}7\,\sigma_{\text{réf}}\}. \tag{AQ'}$$

♧

Remarque (AR'). Dans la suite du cours, il sera fréquent que j'introduise seulement les notations $\mathscr{H}_0$ et $\mathscr{H}_1$, puis que je me mette à utiliser les notations 'Θ_0' et 'Θ_1' sans les avoir définies formellement : il faudra alors comprendre que Θ_i est le sous-ensemble de Θ caractérisé par le fait que $\mathscr{H}_i = \{\theta \in \Theta_i\}$. Parfois, l'association implicite se fera en sens inverse : je n'introduirai formellement que les ensembles Θ_0 et Θ_1, puis utiliserai les notations '$\mathscr{H}_0$' et '$\mathscr{H}_1$' sans prendre le temps de préciser que $\mathscr{H}_i$ est défini par $\{\theta \in \Theta_i\}$.

En fait, j'assimilerai très même souvent l'hypothèse en tant que telle et le sous-ensemble de Θ correspondant : il pourra ainsi m'arriver d'écrire des choses comme « $\mathscr{H}_0$ est topologiquement fermée », auquel cas je voudrai en réalité que c'est l'ensemble Θ_0 qui est un sous-ensemble fermé de Θ. ♧

Remarque (AS'). Notez que, de même que le concept d'estimateur (ou de prédicteur), le concept de « test » est très creux au niveau de sa définition "nue" : « être un test », en tant que tel, cela ne présente quasiment aucun intérêt... Ce qui est intéressant, c'est d'être un test *pertinent* ! De ce point de vue, « être de niveau α » est un tel critère de pertinence, de même que « être convergent » était un critère de pertinence pour un estimateur. Nous verrons néanmoins que le fait d'atteindre un niveau donné n'est pas en soi un critère de qualité suffisant pour un test : ce qui compte aussi, c'est qu'il ait une bonne *puissance*... Ce concept de « puissance » sera introduit un peu plus loin (définition (AU')). ♧

Puissance d'un test

Nous avons expliqué que, lorsqu'on procède à des tests fréquentistes, les hypothèses nulle et alternative sont traitées de façon dissymétrique : le « niveau » du test, en particulier, se réfère uniquement au cas où on se trompe en répondant VRAI alors qu'on est en réalité dans l'hypothèse nulle. Néanmoins, si on ne considérait que ce critère, la théorie des tests deviendrait complètement triviale : il suffirait de

ne considérer que les tests répondant systématiquement FAUX, et ce serait imbattable...! ☿

Bien entendu, en pratique, on veut *non seulement* que notre test ne se trompe pas trop souvent lorsqu'on est dans l'hypothèse nulle, mais *aussi* qu'il parvienne à identifier correctement l'hypothèse alternative dans certains cas... Ce sont les notions d'*erreur de seconde espèce* et de *puissance*, que nous allons à présent présenter, qui permettent de quantitfier cela.

Définition (AT') (Erreurs de première et de seconde espèce)**.** On dit qu'un test commet une *erreur de première espèce* lorsqu'il tranche en faveur de l'hypothèse alternative alors que c'est l'hypothèse nulle qui était correcte ; et qu'il commet une *erreur de seconde espèce* lorsqu'il tranche en faveur de l'hypothèse nulle alors que c'est l'hypothèse alternative qui était correcte. ♡

! **Définition (AU').** Lorsqu'on utilise un test booléen *Test* pour tester une hypothèse nulle $\{\theta \in \Theta_0\}$ contre l'hypothèse alternative $\{\theta \in \Theta_1\}$ (avec $\Theta_1 := \Theta \setminus \Theta_0$) :

— La *fonction de risque de première espèce* est l'application, définie sur Θ_0, qui associe à une valeur de θ le risque de première espèce correspondant : $\Theta_0 \ni \theta \mapsto \mathbb{P}_\theta(Test)$.

— En pratique, on s'intéresse uniquement au *supremum* de la fonction de risque de seconde espèce, qu'on appelle *niveau* du test, et qu'on note traditionnellement α (on retombe donc sur la définition (AG')) :

$$\alpha := \sup\{\mathbb{P}_\theta(Test) \mid \theta \in \Theta_0\}. \qquad \text{(AV')}$$

— La *fonction de risque de seconde espèce* est l'application, définie sur Θ_1, qui associe à une valeur de θ le risque de seconde espèce correspondant. On utilise généralement la lettre β pour la désigner :

$$\Theta_1 \ni \theta \mapsto \beta(\theta) := \mathbb{P}_\theta(\neg Test). \qquad \text{(AW')}$$

On considère également fréquemment le complément à 1 de la fonction de risque de seconde espèce, qu'on appelle *fonction de puissance* du test :

$$\Theta_1 \ni \theta \mapsto 1 - \beta(\theta) = \mathbb{P}_\theta(Test). \qquad \text{(AX')}$$

— En pratique, contrairement à ce qu'on fait pour le risque de première espèce, le supremum du risque de seconde espèce (ou, de manière équivalente, l'infimum de la fonction de puissance) n'est jamais considéré.

♡

Remarque (AY'). Vous vous demandez certainement pourquoi on considère le supremum en ce qui concerne le risque de première espèce, mais qu'on ne le fait pas en ce qui concerne le risque de seconde espèce. La raison en est que, dans les modèles qu'on est amené à considérer en pratique, il est possible de rendre un de ces deux risques uniformément petit, mais pas les deux simultanément... L'explication à ce phénomène sera abordée un peu plus loin dans ce polycopié : néanmoins, à ce stade, vous avez déjà suffisamment de nouveautés à intégrer dans ce chapitre pour qu'il soit préférable de simplement accepter l'affirmation ci-dessus telle quelle dans un premier temps ! ☺

En tout cas, cela justifie ce traitement dissymétrique des hypothèses complémentaires : puisqu'il est impossible de contrôle *uniformément* les deux risques à la

fois, on décide qu'il y aura une hypothèse pour laquelle le risque sera contrôlé uniformément, et l'autre pour laquelle on se devra se contenter d'une fonction de risque selon la valeur du paramètre caché : et c'est celle qu'on choisit pour le contrôle uniforme qu'on qualifier d'« hypothèse nulle », et l'autre qu'on qualifie d'« hypothèse alternative » ! ♧

12.3 Construction d'un test à partir d'une statistique

Notion de statistique de test

La définition (AG') nous définit ce qu'est, de manière générale, un test. Pour *construire* un test, la procédure de base (dont nous verrons un extension un peu plus loin [Procédure (CQ')]) est la suivante :

Définition (AZ') (Statistique de test unilatérale)**.** Lorsqu'on cherche à tester l'hypothèse nulle $\mathscr{H}_0$ contre l'hypothèse alternative $\mathscr{H}_1$, une *statistique de test à droite* est une statistique $t(X) =: T$ à valeurs réelles dont on "espère" qu'elle a "tendance" à prendre des valeurs plus grandes lorsque $\mathscr{H}_1$ est vérifiée que lorsque $\mathscr{H}_0$ est vérifiée. Dans ce cas, on dit aussi qu'on « soupçonne les grandes valeurs de T » (en partant toujours du principe que « être suspect », cela signifie « être de nature à nous inciter à *rejeter l'hypothèse nulle* »). !

À l'inverse, une *statistique de test à gauche* est une statistique $t(X) =: T$ à valeurs réelles dont on "espère" qu'elle a "tendance" à prendre des valeurs plus petites (au sens de « plus proches de $-\infty$ », quand bien même la valeur absolue deviendrait plus grande) lorsque $\mathscr{H}_1$ est vérifiée que lorsque $\mathscr{H}_0$ est vérifiée. Dans ce cas, on dit aussi qu'on « soupçonne les petites valeurs de T ». (On peut aussi dire « les valeurs les plus à gauche » pour souligner que « petit » signifie ici « petit en valeur *signée* », pas en valeur absolue).

Collectivement, les statistiques de test à gauche et à droite sont qualifiées de *statistiques de test unilatérales*. Une statistique de test unilatérale étant donnée, on parlera du *critère de suspicion* associé à cette statistique pour dire si on en soupçonne les grandes valeurs ou au contraire les petites valeurs (autrement dit, pour dire s'il s'agit d'une statistique à droite ou à gauche) : un « critère de suspicion » ne peut donc, à ce stade, que prendre deux modalités : « on soupçonne les valeurs les plus à droite » ou « on soupçonne les valeurs les plus à gauche ». (Nous verrons un peu plus tard une troisième modalité, confer définition (CO')).

Une statistique de test unilatérale étant donnée, on dira qu'un test est associé à cette statistique de test lorsque :

— Dans le cas d'une statistique de test à droite : le test est de la forme $\{T > t_{\text{seuil}}\}$ pour un certain $t_{\text{seuil}} \in [-\infty, +\infty]$;
— Dans le cas d'une statistique de test à gauche : le test est de la forme $\{T < t_{\text{seuil}}\}$ pour un certain $t_{\text{seuil}} \in [-\infty, +\infty]$.

♡

Remarque (BA'). Bien noter que, de même que la définition d'« être un estimateur » et « être un test », la définition d'« être une statistique de test » est, en tant que telle, "creuse" : quand on dit qu'on « espère » qu'une statistique de test à droite prendra de plus grandes valeurs sous $\mathscr{H}_1$ que sous $\mathscr{H}_0$, non seulement cela ne veut, formellement parlant, rien dire ; mais en plus, quand bien même la statistique prendrait des valeurs comparables sous $\mathscr{H}_1$ et sous $\mathscr{H}_0$, voire aurait tendance à prendre

de plus *petites* valeurs sous $\mathscr{H}_1$ que sous $\mathscr{H}_0$, cela n'empêcherait pas, *techniquement*, de la traiter comme une statistique de test à droite : le cas échéant, cette statistique ne serait *pas intelligente* (au sens où on n'arriverait pas à construire un test *pertinent* à partir d'elle) ; mais du point de vue de la définition formelle il n'y aurait pas de problème à dire quand même que c'est une statistique de test à droite ! Prendre une statistique (réelle) donnée comme statistique de test, puis décider d'en soupçonner les grandes ou les petites valeurs, est donc un *choix* du statisticien : tous les choix sont techniquement possibles, et aucun ne sera formellement faux : en revanche, certains seront plus pertinents que d'autres... ! ♣

Remarque (BB'). Notez que j'ai décidé de toujours utiliser des inégalités *strictes* pour définir les test : lorsque la statistique de test est exactement égale à la valeur-seuil, on penche systématiquement pour l'hypothèse nulle. Techniquement parlant, ce choix n'est pas obligatoire ; néanmoins cette convention s'avèrera très pratique dans la suite (en particulier pour définir un test à partir d'une statistique, d'un critère de suspicion et d'un niveau, cf. procédure (BD') *infra*) : je vous conseille donc de toujours la respecter. (Car la violer n'apporterait quasiment rien et risquerait de rendre certaines des constructions ci-dessous fausses...). ♣

Remarque (BC'). En pratique, on construit quasiment toujours les tests en passant par une statistique de test (qui peut néanmoins être, dans certains cas, une statistique de test *bilatérale*, confer § 12.5). (Et d'ailleurs, même dans les rares cas où aucune statistique de test n'est formellement introduite, on pourrait toujours considérer que le test *Test* est en fait associé à la statistique $\mathbf{1}_{Test}$ avec critère de suspicion à droite... ! ☺). ♣

Construction d'un test de niveau imposé

Dans la définition (AZ') ci-dessus, ainsi que dans l'exemple introductif, nous avons utilisé la statistique de test en fixant *d'abord* un seuil pour la statistique, *puis* en étudiant le niveau associé à ce seuil. En pratique, lorsqu'on cherche à construire un test booléen, on procède généralement dans l'autre sens : on *fixe* d'abord le niveau qu'on désire pour notre test ; puis on *cherche* le seuil permettant d'atteindre ce niveau (mais sans pousser le seuil trop loin pour garder une chance de conclure en faveur de l'hypothèse alternative lorsqu'elle est vraie). La procédure pour trouver le niveau adéquat est toujours la même, et vous devez la connaitre :

! **Procédure (BD').** Une partition de l'espace du paramètre caché en hypothèse nulle et hypothèse alternative étant donnée, un niveau $\alpha \in]0, 1/2]$ étant fixé, et une statistique de test unilatérale (avec son critère de suspicion) étant choisie, on construit canoniquement, à partir de ces éléments, « le » test de niveau α de la façon suivante :

— Dans le cas d'une suspicion à gauche, on fixe le seuil à

$$t_{\text{seuil}} = \inf\{\text{Qtile}(\text{Loi}_\theta(T); \alpha) \mid \theta \in \Theta_0\} \qquad \text{(BE')}$$

(puis on considèrera le test $\{T < t_{\text{seuil}}\}$) ;

— Dans le cas d'une suspicion à droite, on fixe le seuil à

$$t_{\text{seuil}} = \sup\{\text{Qtile}(\text{Loi}_\theta(T); 1-\alpha) \mid \theta \in \Theta_0\} \qquad \text{(BF')}$$

(puis on considèrera le test $\{T > t_{\text{seuil}}\}$). ♡

Pour comprendre d'où vient cette procédure (et pour la retenir), voyons comment on démontre que cela donne bien un test de niveau α. Nous ne traiterons ici que le cas des tests à droite, mais celui des tests à gauche serait similaire.

Démonstration de la validité de la procédure. Plaçons-nous donc dans le cas d'un test à droite. Soit t_* un seuil envisagé, et demandons-nous à quelle condition est-ce que le test $Test := \{T > t_*\}$ de l'hypothèse nulle $\{\theta \in \Theta_0\}$ sera de niveau α.

Pour que le test soit de niveau α, par définition du niveau, il faut et il suffit qu'on ait, pour tout $\theta \in \Theta_0$, que $\mathbb{P}_\theta(Test) \leqslant \alpha$, autrement dit, que $\mathbb{P}_\theta(T > t_*) \leqslant \alpha$. Cela peut encore se réécrire, par passage au complémentaire, en disant que la fonction de répartition de $\mathrm{Loi}_\theta(T)$, évaluée en t_*+, doit valoir au moins $1-\alpha$:

$$\mathrm{Répart}(\mathrm{Loi}_\theta(T); t_*+) \geqslant 1-\alpha. \qquad (\texttt{BG'})$$

Mais dire qu'il y a au moins une fraction $(1-\alpha)$ de la masse de $\mathrm{Loi}_\theta(T)$ à gauche de t_*, c'est équivalent à dire que le seuil auquel on sépare une fraction $(1-\alpha)$ à gauche d'une fraction α à droite est situé *avant* t_*, autrement dit, que le quantile de niveau $(1-\alpha)$ de $\mathrm{Loi}_\theta(T)$ est plus petit que t_* ! La condition que nous avons écrite sur la fonction de répartition est ainsi équivalente [††] à :

$$\mathrm{Qtile}\big(\mathrm{Loi}_\theta(T); 1-\alpha\big) \leqslant t_*. \qquad (\texttt{BH'})$$

Ainsi, pour qu'on ait bien un test de niveau α, le seuil t_* doit être (et il suffit qu'il le soit) supérieur ou égal à $\mathrm{Qtile}(\mathrm{Loi}_\theta(T); 1-\alpha)$; et ce, pour tous les $\theta \in \Theta_0$: ce qui revient bien à dire que t_* doit être au moins égal au supremum de ces quantiles lorsque θ parcourt Θ_0. Parmi l'ensemble de ces choix possible pour fixer le seuil, il est alors naturel de prendre le plus petit. En effet, comme nous le verrons plus loin, la condition sur le niveau contrôle simplement le risque de répondre « positif » alors qu'on est sous l'hypothèse nulle ; mais, autant que possible, on voudrait *aussi* limiter le risque de répondre « négatif » à tort : or, plus t_* est grand, plus le test $\{T > t_*\}$ répondra facilement « négatif » ; et dès lors, pour contrôler le second type de risque, il est logique de chercher à prendre t_* aussi petit que possible parmi les valeurs compatibles avec le niveau requis ! Ce qui nous conduit à la définition du t_{seuil} de la procédure. ♤

Remarque (BI'). La preuve ci-dessus est écrite en termes assez formels ; mais une fois que vous l'avez comprise, son idée doit vraiment être simple à retenir (de sorte que vous pourrez immédiatement écrire la formule pour t_{seuil}) : pour qu'on soit de niveau α, il faut que, sous les différents contextes $\mathbb{P}_\theta(\bullet)$ correspondant à l'hypothèse nulle, la probabilité de dépasser le seuil n'excède pas α ; autrement dit, que le seuil soit au-delà du quantile de niveau $(1-\alpha)$ de $\mathrm{Loi}_\theta(T)$: et puisqu'on veut que cela soit le cas pour tous les θ de Θ_0, on va prendre le supremum de ces quantiles lorsque θ décrit l'hypothèse nulle ! ◡ ♧

Remarque (BJ'). Lorsqu'on détermine le seuil comme un supremum ou infimum de quantiles, il est fréquent qu'on doive procéder à une approximation à un moment donné, ne serait-ce que pour arrondir la valeur qui est, en toute généralité,

[††]. En toute rigueur, pour que l'équivalence soit parfaite, il faudrait préciser qu'on prend dans (??) la version *continue à droite* de la fonction de quantile, càd. que le quantile de niveau "$1-\alpha$" est, en toute rigueur, le quantile de niveau "$(1-\alpha)-$"... Néanmoins, conformément à notre habitude, nous négligerons ces subtilités en ce qui concerne les fonctions de quantile ♋

un nombre réel non décimal. (Il peut aussi arriver qu'on ne sache pas déterminer *exactement* les quantiles, mais seulement en donner un encadrement). Dans ce cas, il convient de toujours faire l'approximation dans le sens qui assure que le quantile reste bien de niveau α : ainsi, dans le cas d'un test à droite, on prendre une approximation du seuil dont on sache qu'elle est située un peu *au-dessus* du seuil critique (et pas en-dessous, auquel cas on augmenterait la probabilité que le test soit positif, ce qui risquerait d'enfreindre la condition de niveau α) ; tandis que dans le cas d'un test à gauche, on doit prendre une approximation située un peu *en-dessous* de la vraie valeur. Typiquement, lorsqu'il s'agit de faire des arrondis, on fera l'arrondi du seuil par excès pour un test à gauche, resp. par défaut pour un test à droite.Pour souligner le fait que les approximations opérées garantissent qu'on conserve l'exigence d'« être de niveau α », on parle d'approximations *conservatives*. Cette idée de conservativité est essentielle dans la science de la maitrise des risques (industriels, sanitaires, financiers), où nos modèles sont entachés d'une certaine incertitude, mais où on veut néanmoins avoir une garantie solide sur le niveau de risque résiduel !

Dans un sens, c'est aussi l'idée de conservativité qui explique pourquoi, quand la statistique de test est pile au niveau du seuil, on décide que le test sera négatif : pour respecter le contrat « être de niveau α », on peut en effet se permettre de répondre un peu moins souvent « positif » que le maximum requis, mais en aucun cas de répondre « positif » *trop* souvent ! Donc, dans le cas-limite, mieux vaut répondre « négatif » [‡‡] ♧

Exemples de constructions de tests

Nous allons conclure cette section en montrant des exemples de constructions de tests. Mais avant de nous lancer dans ces exemples, il convient de souligner un point qui intervient fréquemment dans ce genre de constructions :

! *Remarque* (BK'). Le supremum qui intervient dans la définition du niveau d'un test peut fréquemment être restreint à un certain sous-ensemble de Θ_0 : moralement, l'idée est que, dans le cas (mettons) d'un test à droite, pour étudier le supremum des valeurs $\mathbb{P}_\theta(T > t_{\text{seuil}})$, il y a certaines valeurs de θ qu'on peut se permettre d'omettre, parce qu'il y a d'autres valeurs θ', situées également dans Θ_0, pour lesquelles T prend des valeurs "encore plus grandes" sous $\mathbb{P}_{\theta'}$ que sous $\mathbb{P}_\theta$, de sorte que ces θ-là ne contribueront pas au supremum. En général, pour étudier le supremum, on peut en fait se limiter à un sous-ensemble de l'hypothèse nulle correspondant plus ou moins au "bord" de Θ_0 ♧

Remarque (BL'). En continuation de la remarque précédente, il est également fréquent que, au moins sur le sous-ensemble de Θ_0 considéré, la loi la statistique de test sous $\mathbb{P}_\theta$ soit en fait la même *pour tous les θ à considérer* [∗] : dès lors, le

[‡‡]. Pour être plus précis, il faut distinguer deux cas : lorsque les lois $\text{Loi}_\theta(T)$ sont diffuses, en fait, la probabilité que la statistique de test vaille *exactement* t_{seuil} est nulle (sous la vraie loi), de sorte que définir le test par une inégalité large ou stricte ne change absolument rien en pratique. En revanche, si les lois $\text{Loi}_\theta(T)$ sont discrètes, il devient très important de bien définir le test par une inégalité stricte ! En effet dans ce cas, il se passe (génériquement) la chose suivante : le test défini, conformément aux explications ci-dessus, par $\{T > t_{\text{seuil}}\}$ aura un niveau strictement inférieur à α (et restera donc « de niveau α ») ; par contre, le test défini par $\{T \geqslant t_{\text{seuil}}\}$ (avec inégalité *large*) aura un niveau *strictement supérieur* à α, et ne constituera donc pas un test de niveau α... !

[∗]. Attention : Je ne suis pas en train de dire que la statistique T aurait la même loi sous absolument tous les $\mathbb{P}_\theta(\bullet)$: une telle statistique de test ne présenterait en effet aucun intérêt, puisqu'elle aurait le même comportement sous l'hypothèse nulle et sous l'hypothèse alternative !

calcul du supremum peut en fait se faire en considérant *une seule* loi... Ce qui est évidemment bien plus simple ! (En fait, un certain nombre de test classiques sont justement construits de façon à faire en sorte que la statistique de test ait toujours la même loi pour $\theta_{\checkmark}$ dans Θ_0 : confer principe (CN') *infra*).

En outre, cette idée que le supremum définissant le niveau peut être restreint à un sous-ensemble de valeurs plus simples, voire être ramené à un calcul sur une seule loi, marchera exactement de la même façon lorsqu'il s'agira de déterminer le seuil définissant le test à partir d'un certain niveau. (Confer les exemples ci-dessous). ♧

Remarque (BM'). En fait, la remarque précédente avait déjà été plus ou moins utilisée dans l'exemple de la § 12.1, lorsqu'on avait convenu que, pour évaluer le risque que le test tranche en faveur des collègues alors que c'est l'enseignant qui avait raison, il fallait considérer un cas où l'enseignant avait raison "de justesse" (et, en l'occurrence, vu que l'hypothèse nulle revenait à affirmer que $\{\mu \geqslant \mu_{\text{réf}}\}$, nous avions choisi de nous placer dans la sitation où $\mu_{\checkmark} = \mu_{\text{réf}} + 0{,}1$). En fait, ce qu'on aurait pu montrer en étant plus rigoureux, c'est que, quel que soit le seuil t_{seuil} choisi pour définir le test, le supremum des valeurs $\mathbb{P}_{\mu,\sigma_{\text{réf}}}(M < t_{\text{seuil}})$ pour $\mu \geqslant \mu_{\text{réf}}$ serait atteint précisément pour $\mu = \mu_{\text{réf}}$ (ce qui correspond bien au cas situé "au bord" de l'hypothèse nulle) : le véritable niveau du test se calculant donc, en l'occurrence, en évaluant $\mathbb{P}_{\mu_{\text{réf}},\sigma_{\text{réf}}}(M < 95)$. (Et cette probabilité est très proche, en pratique, de la probabilité $\mathbb{P}_{\mu_{\text{réf}}+0{,}1,\sigma_{\text{réf}}}(M < 95)$ que nous avions calculée dans la § 12.1 : elles valent respectivement 5,9 % et 5,6 %). ♧

Exemple (BN'). Reprenons, pour commencer, l'exemple de notre sous-section introductive : on considère donc le modèle du pédagogue où l'espace du paramètre caché est restreint à $\Theta^{\text{hsc}} := \mathbb{R} \times \{\sigma_{\text{réf}}\}$, et on cherche à tester l'hypothèse nulle $\{\mu \geqslant \mu_{\text{réf}}\} =: \mathscr{H}_0$ contre l'hypothèse alternative $\{\mu < \mu_{\text{réf}}\} =: \mathscr{H}_1$. On choisit d'utiliser comme statistique de test la moyenne de la première promotion testée, autrement dit, $M := \left(\sum_{i=0}^{n-1} X_i\right) / n$; avec suspicion pour les petites valeurs. La seule différence que nous allons prendre par rapport à la § 12.1 est sur la façon de choisir le seuil du test : contrairement à ce que nous avions fait précédemment, nous allons ici imposer qu'on souhaite avoir un test de niveau $\alpha := 5$ %, et regarder comment on en déduit le seuil approprié : cela nous permettra ainsi de mettre en pratique la procédure (BD') et la remarque (BK') ci-dessus.

En vertu de la procédure (BD'), notre test devra être de la forme $\{M < m_{\text{seuil}}\}$, avec

$$m_{\text{seuil}} = \inf_{\mu \geqslant \mu_{\text{réf}}} \text{Qtile}(\text{Loi}_{\mu,\sigma_{\text{réf}}}(M); \alpha) \,[\dagger] = \inf_{\mu \geqslant \mu_{\text{réf}}} \text{Qtile}\big(\text{Normale}(\mu, \sigma_{\text{réf}}^2 / n); \alpha\big). \quad \text{(BO')}$$

Nous allons maintenant voir comment le calcul de cet infimum peut être grandement simplifié. D'après les propriétés des lois normales, $\text{Normale}(\mu, \sigma_{\text{réf}}^2 / n) = \mu + \text{Normale}(0, \sigma_{\text{réf}}^2 / n)$. Or, puisque l'application $x \mapsto \mu + x$ est croissante, pour

Ce que je dis, c'est seulement que notre statistique aura fréquemment la même loi sous tous les $\mathbb{P}_\theta(\bullet)$ *pour* θ *dans* Θ_0 — ou du moins, pour θ dans l'ensemble "critique", marquant le "bord" de Θ_0, auquel nous avons fait allusion à l'instant.

[†]. Notez ici que, même s'il *semble* qu'on ait pris notre infimum sur μ plutôt que sur θ — alors que j'ai pourtant bien expliqué [remarque (AP') *supra*] qu'il fallait toujours prendre l'infimum sur θ *lui-même* ! —, la forme spécifique de l'espace du paramètre caché *dans cet exemple* fait que μ est en bijection avec $\theta = (\mu, \sigma_{\text{réf}})$

toute loi $\boldsymbol{P}$, pour tout $p \in]0,1[$, on a $\mathrm{Qtile}(\mu + \boldsymbol{P}; p) = \mu + \mathrm{Qtile}(\boldsymbol{P}; p)$: en l'occurrence, on a donc

$$\mathrm{Qtile}\big(\mathrm{Normale}(\mu, \sigma^2_{\text{réf}} / n); \alpha\big) = \mathrm{Qtile}\big(\mathrm{Normale}(0, \sigma^2_{\text{réf}} / n); \alpha\big) + \mu, \qquad (\texttt{BP'})$$

d'où

$$\inf_{\mu \geqslant \mu_{\text{réf}}} \mathrm{Qtile}\big(\mathrm{Normale}(\mu, \sigma^2_{\text{réf}} / n); \alpha\big) = \mathrm{Qtile}\big(\mathrm{Normale}(0, \sigma^2_{\text{réf}} / n); \alpha\big) + \mu_{\text{réf}}$$
$$= \mathrm{Qtile}\big(\mathrm{Normale}(\mu_{\text{réf}}, \sigma^2_{\text{réf}} / n); \alpha\big) : \qquad (\texttt{BQ'})$$

on voit ainsi qu'en fait, l'infimum des quantiles pour toutes les valeurs de θ dans Θ_0 correspond au quantile pour θ égal à la valeur "critique" $(\mu_{\text{réf}}, \sigma_{\text{réf}})$ située au bord de Θ_0.

Il ne reste plus qu'à utiliser un logiciel numérique pour calculer la fonction de table que nous avons fait apparaitre :

```
> qnorm(5 / 100, muref, sqrt(sigma ^ 2 / n))
[1] 94.73974
```

cela nous suggère, en fin de compte, de prendre pour test $\{\boldsymbol{M} < 94{,}7\}$, en prenant garde d'arrondir la valeur numérique trouvée *par dessous* [‡] pour être sûr que, même après arrondi, le test reste bien de niveau α. ♧

Exemple (BR'). Pour bien illustrer la remarque (BK'), reprenons l'exemple précédent, et regardons quel est l'impact de l'arrondi final sur le véritable niveau du test (dans la suite, nous utiliserons la notation $94{,}7 =: \tilde{\mu}_{\text{seuil}}$, et nous noterons $\tilde{\alpha}$ le niveau du test correspondant). Par définition, on a que

$$\tilde{\alpha} \underset{\text{déf}}{=} \sup_{\mu \geqslant \mu_{\text{réf}}} \mathbb{P}_{\mu, \sigma_{\text{réf}}}(\boldsymbol{M} < \tilde{\mu}_{\text{seuil}}) = \sup_{\mu \geqslant \mu_{\text{réf}}} \mathbb{P}\big(\mathrm{Normale}(\mu, \sigma_{\text{réf}}) < \tilde{\mu}_{\text{seuil}}\big). \qquad (\texttt{BS'})$$

La remarque (BK') nous suggère alors que le supremum ci-dessus sera atteint pour $\mu = \mu_{\text{réf}}$: et de fait, ce sera bien le cas. Mais comment le démontrer ? Ici, plutôt que de chercher à expliciter la façon dont $\mathbb{P}\big(\mathrm{Normale}(\mu, \sigma_{\text{réf}}) < \tilde{\mu}_{\text{seuil}}\big)$ dépend de μ, nous allons juste *comparer* les valeurs entre $\mu_{\text{réf}}$ et les autres valeurs de μ. Soit donc $\mu \geqslant \mu_{\text{réf}}$; notre but va être de montrer que

$$\mathbb{P}\big(\mathrm{Normale}(\mu, \sigma_{\text{réf}}) < \tilde{\mu}_{\text{seuil}}\big) \leqslant \mathbb{P}\big(\mathrm{Normale}(\mu_{\text{réf}}, \sigma_{\text{réf}}) < \tilde{\mu}_{\text{seuil}}\big). \qquad (\texttt{BT'})$$

Pour ce faire, on observe que, d'après les propriétés des lois normales, on a

$$\mathrm{Normale}(\mu, \sigma_{\text{réf}}) = \mathrm{Normale}(\mu, \sigma_{\text{réf}}) + \mu - \mu_{\text{réf}} : \qquad (\texttt{BU'})$$

en particulier, si $\boldsymbol{M}$ est une v.a. de loi $\mathrm{Normale}(\mu_{\text{réf}}, \sigma^2_{\text{réf}} / n)$, alors la v.a. $\boldsymbol{M'} := \boldsymbol{M} + \mu - \mu_{\text{réf}}$ sera de loi $\mathrm{Normale}(\mu, \sigma^2_{\text{réf}} / n)$; et en outre, puisque $\mu \geqslant \mu_{\text{réf}}$, on aura presque-surement que $\boldsymbol{M'} \geqslant \boldsymbol{M}$ [§]. Mais, au vu de cette comparaison, l'évènement $\{\boldsymbol{M'} < \tilde{\mu}_{\text{seuil}}\}$ implique l'évènement $\{\boldsymbol{M} < \tilde{\mu}_{\text{seuil}}\}$, de sorte que

$$\mathbb{P}(\boldsymbol{M'} < \tilde{\mu}_{\text{seuil}}) \geqslant \mathbb{P}(\boldsymbol{M} < \tilde{\mu}_{\text{seuil}}) : \qquad (\texttt{BV'})$$

[‡]. En l'occurrence, l'arrondi par-dessous correspond à l'arrondi au plus proche ; mais si nous avions plutôt voulu arrondir point près, ou au centième de près, cette fois-ci notre règle aurait conduit à prendre des arrondis à resp. 94 et 94,73, qui n'auraient pas coïncidé avec l'arrondi au plus proche !

[§]. Cette idée de trouver des variables aléatoires instanciant les lois $\mathrm{Normale}(\mu_{\text{réf}}, \sigma^2_{\text{réf}} / n)$ et $\mathrm{Normale}(\mu, \sigma^2_{\text{réf}} / n)$ sur un même espace probabilisé, d'une façon qu'on ait une comparaison pertinente entre les deux v.a. qui soit de type *presque-sure* (ou, plus généralement, qui soit une comparaison en probabilité), s'appelle « réaliser un *couplage* entre les deux lois ».

ce qui, quand on se rappelle quelles lois suivent resp. $\boldsymbol{M}'$ et $\boldsymbol{M}$, signifie bien que, conformément à ce qu'on souhaitait prouver, on a que $\mathbb{P}\big(\text{Normale}(\mu, \sigma^2_{\text{réf}} / n) < \tilde{\mu}_{\text{seuil}}\big) \geqslant \mathbb{P}\big(\text{Normale}(\mu_{\text{réf}}, \sigma^2_{\text{réf}} / n) < \tilde{\mu}_{\text{seuil}}\big)$!

Ainsi, on a que

$$\tilde{\alpha} = \texttt{répartNormale}(\mu_{\text{réf}}, \sigma^2_{\text{réf}} / n; \tilde{\mu}_{\text{seuil}}-), \qquad (\texttt{BW'})$$

ce qu'il ne reste plus qu'à calculer numériquement [¶] :

```
> pnorm(94.7, muref, sqrt(sigma ^ 2 / n))
[1] 0.04873142
```

On trouve ainsi que le niveau véritable de notre test est de 4,88 % [‖] (en arrondissant ici *par-dessus*, afin qu'il soit bien légitime de dire que « c'est un test de niveau 4,88 % »). ♧

12.4 Recherche de statistiques de test pertinentes

Pour appliquer la théorie des tests de façon utile, il va falloir construire des tests qui auront à la fois un bon niveau (typiquement, $\alpha \leqslant 10$ %) et une fonction de puissance aussi élevée que possible. Comment construire de tels tests ? Comme expliqué ci-dessus, en pratique on construit les tests à partir de statistiques de test (associées à un critère de suspicion) : la question devient donc de savoir comment on peut trouver une bonne statistique de test !

En tant qu'ingénieurs généralistes, en fait, il devrait être exceptionnel que vous soyez amené(e) à créer vous-mêmes des statistiques de test : le point essentiel sera surtout de s'assurer qu'une statistique de test donnée est pertinente ! Et sur ce point, en général, on procédèra plus souvent par des raisonnements informels que par des preuves rigoureuses, l'enjeu étant juste de s'assurer que le choix de la statistique de test est "raisonnable"...

Le principe fondamental pour vérifier qu'une statistique de test est pertinente est toujours le même. Ci-dessous nous allons légèrement anticiper sur la section suivante du chapitre, en parlant d'un critère de suspicion que vous n'avez pas encore rencontré : le « critère de suspicion bilatéral ». Ce qu'on entend par là, et la façon de construire les tests correspondants, sera développé un peu plus tard.

Principe (BX'). *Considérons un modèle statistique, où l'espace du paramètre caché est partitionné en deux hypothèses complémentaires, avec les notations standard. Une statistique de test $t(\boldsymbol{X})$ est appropriée dans les cas suivants :* !

(i) Lorsque $\text{Loi}_\theta(t(\boldsymbol{X}))$ *a "tendance à prendre des valeurs plus grandes* [**] *lorsque* $\theta \in \boldsymbol{\Theta}_1$ *que lorsque* $\theta \in \boldsymbol{\Theta}_0$ [††] *" : dans ce cas, le critère de suspicion approprié est un critère de suspicion à* droite.

(ii) Symétriquement, lorsque $\text{Loi}_\theta(t(\boldsymbol{X}))$ *a tendance à prendre des valeurs plus petites lorsque* $\theta \in \boldsymbol{\Theta}_1$ *que lorsque* $\theta \in \boldsymbol{\Theta}_0$ *: dans ce cas, il convient de soupçonner les valeurs les plus à* gauche.

[¶]. Notez que, comme les lois normales sont diffuses, il n'est pas important de savoir si on considère la fonction de répartition en $\tilde{\mu}_{\text{seuil}}-$ ou en $\tilde{\mu}_{\text{seuil}}+$.

[‖]. Cette valeur est très proche de 5 %, ce qui montre que notre arrondi n'était effectivement pas bien méchant ! ☺

(iii) *Lorsque* $\mathrm{Loi}_\theta(t(X))$ *n'a jamais tendance à prendre des valeurs trop grandes aussi longtemps que* $\theta \in \Theta_0$*, qu'il existe des valeurs* $\theta \in \Theta_1$ *pour lesquelles* $\mathrm{Loi}_\theta(t(X))$ *prend des valeurs (nettement) plus grandes que dans* n'importe quel cas *associé à* Θ_0*, et qu'il n'existe aucune valeur* $\theta \in \Theta_1$ *pour lesquelles* $\mathrm{Loi}_\theta(t(X))$ *prenne des valeurs (nettement) plus petites que* tous *les cas associés à* Θ_0 *: alors le critère de suspicion approprié est une suspicion à* droite.

(iv) *Symétriquement, lorsque* $\mathrm{Loi}_\theta(t(X))$ *n'a jamais tendance à prendre des valeurs trop grandes aussi longtemps que* $\theta \in \Theta_0$*, qu'il existe des valeurs* $\theta \in \Theta_1$ *pour lesquelles* $\mathrm{Loi}_\theta(t(X))$ *prend des valeurs (nettement) plus petites que dans n'importe quel cas associé à* Θ_0*, et qu'il n'existe aucune valeur* $\theta \in \Theta_1$ *pour lesquelles* $\mathrm{Loi}_\theta(t(X))$ *prenne des valeurs (nettement) plus grandes que tous les cas associés à* Θ_0 *: alors le critère de suspicion approprié est une suspicion à* gauche.

(v) *Lorsque, aussi longtemps que* θ *est dans* Θ_0*,* $\mathrm{Loi}_\theta(t(X))$ *n'a jamais tendance à prendre des valeurs ni trop petites, ni trop grandes ; qu'il existe en revanche des valeurs* θ *dans* Θ_1 *pour lesquelles* $\mathrm{Loi}_\theta(t(X))$ *prend des valeurs nettement plus grandes que dans les cas associés à* Θ_0*, et qu'il existe aussi des valeurs* θ *dans* Θ_1 *pour lesquelles* $\mathrm{Loi}_\theta(t(X))$ *prend des valeurs nettement plus petites que dans les cas associés à* Θ_0 *: alors le critère de suspicion approprié est une suspicion* bilatérale. ◇

Remarque (BY'). En fait, à partir d'une statistique de test et d'un critère de suspicion, on peut *toujours* définir un test de niveau α : prendre le "mauvais" critère de suspicion ne signifie donc pas qu'on n'arrivera pas à construire de test, mais que ce test sera dénué d'intérêt, parce que sa puissance sera beaucoup trop faible ! (typiquement, quand le critère de suspicion est mal choisi, ce qui risque de se passer est que la puissance du test se retrouve uniformément majorée par α (qui, rappelons-le, est typiquement une valeur très petite) : ainsi, lorsque l'hypothèse alternative est vraie, on ne conclut alors quasiment jamais en sa faveur, et ce, peu importe la valeur $\theta_\checkmark$ dans $\mathscr{H}_1$! ♧

! *Remarque* (BZ'). On notera que, lorsqu'on vérifie qu'une statistique de test est pertinente (et qu'on détermine du critère de suspicion approprié), c'est *à cet endroit*, et *seulement* à cet endroit, qu'on est amené à regarder *aussi* quelle est l'hypothèse alternative $\mathscr{H}_1$! (Alors que, pour déterminer le seuil correspondant à un niveau donné, ou pour déterminer le niveau correspondant à un seuil donné, seule l'hypothèse nulle intervient...).

C'est tout à fait logique, puisque l'idée qu'une statistique de test soit pertinente va se traduire par la *puissance* qu'elle permet d'obtenir pour un niveau de risque donné. Or pour avoir une bonne puissance, il faut que les valeurs prises de la statistique de test sous l'hypothèse alternative soient (en général) suffisamment

[††]. Dans le cadre de ce cours, nous ne chercherons pas à définir précisément ce qu'on entend par là. De manière générale, l'idée serait de dire qu'il y a *domination stochastique* (au moins de manière approximative) des lois $\mathrm{Loi}_\theta(t(X))$ pour $\theta \in \Theta_0$ par les lois $\mathrm{Loi}_\theta(t(X))$ pour $\theta \in \Theta_1$.

[‡‡]. Ici, il faut comprendre qu'on exige que, pour *n'importe quelles v valeurs* $\theta_0 \in \Theta_0$ et $\theta_1 \in \Theta_1$, $\mathrm{Loi}_{\theta_1}(t(X))$ a tendance à prendre des valeurs plus grandes (au moins approximativement) que $\mathrm{Loi}_{\theta_0}(t(X))$. De manière générale, juste dire, par exemple « pour tout $\theta_0 \in \Theta_0$, on peut trouver un $\theta_1 \in \Theta_1$ tel que $\mathrm{Loi}_{\theta_1}(t(X))$ ait tendance à prendre des valeurs plus grandes que $\mathrm{Loi}_{\theta_0}(t(X))$; et pour pour tout $\theta_1 \in \Theta_1$, on peut trouver un $\theta_0 \in \Theta_0$ tel que $\mathrm{Loi}_{\theta_0}(t(X))$ ait tendance à prendre des valeurs plus petites que $\mathrm{Loi}_{\theta_1}(t(X))$ » ne serait pas suffisant ! Néanmoins, s'il s'agit juste de trouver le critère de suspicion approprié pour une statistique de test donnée, on pourra se contenter d'un tel argument.

différentes des valeurs pouvant être typiquement prises sous l'hytpothèse nulle pour qu'on puisse se rendre compte que l'hypothèse nulle n'est pas plausible : on est donc amené à *comparer* le comportement de la statistique de test entre les hypothèses nulle et alternative pour comprendre (au moins de façon grossière) si la puissance va bien se comporter. ♧

Nous allons maintenant voir un exemple où on peut vérifier la pertinence d'une statistique de test à l'aide de nos critères qualitatifs. En fait, nous allons reprendre l'exemple trouvé ci-dessus, et la statistique de test de Neyman-Pearson que nous avions calculée : il s'agira juste de vérifier que cela conduit bien à quelque chose satisfaisant les critères qualitatifs du principe présenté ci-dessus ! ☺

Exemple (CA'). Nous sommes donc dans le modèle du pédagogue, en testant l'« hypothèse du vieux grincheux [*] » $\{\boldsymbol{\mu} = \mu_{\text{réf}} \text{ et } \boldsymbol{\sigma} = \sigma_{\text{réf}}\}$ contre l'hypothèse alternative complémentaire (dans $\boldsymbol{\Theta}^{\text{orig}}$). Et nous voulons utiliser pour ce faire la statistique de test [†]

$$\boldsymbol{T} \coloneqq (\boldsymbol{M} - \mu_{\text{réf}})^2 + \boldsymbol{S}^2 - 2\sigma_{\text{réf}}^2 \ln \boldsymbol{S}, \tag{CB'}$$

dont nous allons feindre de ne pas connaitre le critère de suspicion approprié.

Commençons par observer qu'ici, puisque l'hypothèse nulle est simple, il y a une seule loi $\text{Loi}_\theta(\boldsymbol{T})$ possible lorsque $\theta \in \boldsymbol{\Theta}_0$: dès lors, on peut dire automatiquement que cette loi ne pourra jamais avoir tendance à prendres des valeurs arbitrairement grandes, ni arbitrairement petites, dans un tel cas.

Maintenant, que se passe-t-il sous l'hypothèse alternative ? Il y a quatre façons (non exclusives) dont le vieux grincheux peut avoir tort, selon ce que vaut le vrai paramètre $\theta_{\checkmark} = (\mu_{\checkmark}, \sigma_{\checkmark})$:

- Soit parce que $\mu_{\checkmark} < \mu_{\text{réf}}$, c.-à-d. que la nouvelle méthode fait en réalité globalement régresser le niveau des élèves : dans ce cas, sous réserve que l'écart entre $\mu_{\text{réf}}$ et $\mu_{\checkmark}$ soit suffisamment grand, on s'attend à ce que les notes soient très mauvaises, autrement dit que $\boldsymbol{M}$ soit (sauf coup de chance) nettement plus petit que $\mu_{\text{réf}}$, ce qui va rendre la statistique de test très élevée à cause de son premier terme.
- Soit, à l'inverse, parce que $\mu_{\checkmark} > \mu_{\text{réf}}$, c.-à-d. que la nouvelle méthode fait globalement progresser les élèves : auquel cas $\boldsymbol{M}$ sera nettement plus grand que $\mu_{\text{réf}}$, ce qui fera à nouveau exploser le premier terme de la statistique de test.
- Soit parce que $\sigma_{\checkmark} < \sigma_{\text{réf}}$, c.-à-d. que la nouvelle méthode donne des résultats plus homogènes que l'ancienne : dans ce cas, on s'attend à avoir un $\boldsymbol{S}$ très proche de 0, auquel cas la statistique de test explosera à cause de son troisième terme.
- Soit parce que $\sigma_{\checkmark} > \sigma_{\text{réf}}$, c.-à-d. que la nouvelle méthode crée plus d'hétérogénéité entre les élèves que l'ancienne : dans ce cas, sous réserve que le ratio entre $\sigma_{\text{réf}}$ et $\sigma_{\checkmark}$ soit suffisamment grand, on s'attend à ce que les notes soient très dispersées, autrement dit que $\boldsymbol{S}$ soit grand, ce qui fera exploser la statistique de test à cause de son second terme (certes, dans ce cas le troisième

[*]. Concernant cette appellation, confer exemple (CF') un peu plus loin.

[†]. Comme nous le verrons dans l'exemple (CF'), cette statistique de test est en fait celle du rapport des vraisemblances, ce qui implique en particulier que son critère de suspicion sera à droite : mais nous allons feindre d'ignorer ce dernier point... Au demeurant, le raisonnement que nous allons présenter ci-dessous n'a rien de spécifique au rapport des vraisemblances, et pourrait fonctionner de même avec plein d'autres statistiques de test ! ☺

terme sera négatif, mais la décroissance logarithmique ne compense pas la croissance quadratique ![‡]).

On voit donc que, quelle que soit la façon dont le vieux grincheux a tort, la statistique de test tendra à prendre des valeurs plus grandes que sous l'hypothèse nulle : ainsi la statistique de test est effectivement pertinente, et le critère de suspicion approprié consiste effectivement à suspecter ses grandes valeurs ! ◡̈ ♣

Rapport de vraisemblances

Au-delà des principes généraux énoncé dans les sous-sections précédente et suivante, il existe une procédure systématique pour construire une statistique de test pertinente : c'est de considérer le *rapport des vraisemblances* entre l'hypothèse alternative et l'hypothèse nulle (avec suspicion du côté où l'hypothèse alternative prend de plus grandes valeurs relativement à l'hypothèse nulle). Je n'ai pas marqué ce principe comme à retenir, dans le mesure où on l'utilise assez rarement en pratique ; néanmoins, c'est une "ligne directrice" qu'il est utile de conserver à l'esprit :

Principe (CC'). *De manière générale, une statistique de test pertinente pour tester une hypothèse alyternative $\mathscr{H}_1$ contre une hypothèse nulle $\mathscr{H}_0$ consiste en la v.a. dont la réalisation est le rapport de vraisemblances* [§]

$$\frac{\mathscr{L}(\mathscr{H}_1 \mid X = x_{\checkmark})}{\mathscr{L}(\mathscr{H}_0 \mid X = x_{\checkmark})}, \tag{CD'}$$

le critère de suspicion consistant alors à trouver suspectes les grandes valeurs de ce rapport. ♢

> *Remarque* (CE'). Il existe un théorème, appelé « théorème de Neyman-Pearson » (voir § **??**), qui dit que cette statistique de test est *rigoureusement* optimale (dans un sens bien précis) dans le cas où Θ_0 et Θ_1 sont toutes les deux des singletons. Lorsque ce n'est pas le cas, on n'a malheureusement plus de tel résultat d'optimalité ; néanmoins la stratégie de Neyman-Pearson reste pertinente malgré tout. ♣

Exemple (CF'). Voyons un exemple pour lequel le statistique du rapport des vraisemblances peut être calculée. On se place toujours dans le modèle du pédagogue, avec pour espace du paramètre caché la version originelle $\Theta^{\text{orig}} = \mathbb{R} \times \mathbb{R}_+^*$; mais cette fois-ci, plutôt que de chercher à savoir si la nouvelle méthode va faire progresser les élèves comme prévu ou non, on s'intéresse à l'affirmation d'un vieux professeur grincheux qui affirme que « de toutes façons, le choix de la méthode pédagogique n'a aucune influence sur les résultats des élèves ! ». Ainsi, le vieux professeur défend l'hypothèse que $\{\theta = \theta_{\text{réf}}\}$, avec $\theta_{\text{réf}} \coloneqq (\mu_{\text{réf}}, \sigma_{\text{réf}})$: nous prendrons cette hypothèse comme l'hypothèse nulle. Notre hypothèse alternative, ici, sera que « le professeur grincheux a tort », à savoir, que la nouvelle méthode pédagogique donne des résultats différents de l'ancienne, que ce soit parce que la moyenne change et/ou parce que la variance change : autrement dit, on prend $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ tout

[‡]. Plus précisément, on vérifie sans difficulté que la fonction $s \mapsto s^2 - 2\sigma_{\text{réf}}^2 \ln s$ a les variations suivantes sur $\mathbb{R}_+^*$: tendant vers l'infini près de 0, elle décroit jusqu'à l'abscisse $s = \sigma_{\text{réf}}$, avant de croitre à nouveau jusque vers l'infini lorsque $s \to \infty$.

[§]. Rappelons que la notion de vraisemblance d'une hypothèse composite a été définie dans la définition (UJ).

entier et $\Theta_1 \coloneqq \Theta \setminus \{\theta_{\text{réf}}\}$. Calculons le rapport des vraisemblances dans ce cas. Nous avons vu (cf. § 8.4) que

$$\ln \mathscr{L}((\boldsymbol{\mu}, \boldsymbol{\sigma}) = (\mu, \sigma) \mid X = x_{\checkmark}) = -n\big(\ln \sigma - \tfrac{1}{2}\sigma^{-2}\big((\mu - m_{\checkmark})^2 + s_{\checkmark}^2\big)\big). \qquad (\texttt{CG'})$$

Pour calculer la vraisemblance de l'hypothèse alternative, il faut passer au supremum (en observant que le supremum sur Θ_1 est égal au supremum sur Θ tout entier par continuité) : vu que, à σ fixé, le supremum est atteint pour $\mu = m_{\checkmark}$, il ne reste plus qu'à optimiser en σ sous la contrainte $\{\mu = m_{\checkmark}\}$:

$$\begin{aligned}\ln \mathscr{L}(\boldsymbol{\theta} \in \Theta_1 \mid X = x_{\checkmark}) &= \sup_{\sigma \in \mathbb{R}_+^*} \big(-n \ln \sigma - \tfrac{1}{2} n s_{\checkmark}^2 \sigma^{-2}\big) \\ &= -n \ln s_{\checkmark} - \tfrac{1}{2} n s_{\checkmark}^2 s_{\checkmark}^{-2} = -n(\ln s_{\checkmark} + \tfrac{1}{2}),\end{aligned}$$

où l'avant-dernière égalité vient de ce que le supremum est atteint pour $\sigma = s_{\checkmark}$, comme nous l'avions vu lorsque nous avions calculé l'estimateur du maximum de vraisemblance. On obtient donc le rapport de vraisemblances suivant :

$$\ln \frac{\mathscr{L}(\boldsymbol{\theta} \neq \theta_{\text{réf}} \mid X = x_{\checkmark})}{\mathscr{L}(\boldsymbol{\theta} = \theta_{\text{réf}} \mid X = x_{\checkmark})} = -n\big(\ln(s_{\checkmark}/\sigma_{\text{réf}}) + \tfrac{1}{2}\sigma_{\text{réf}}^{-2}(m_{\checkmark} - \mu_{\text{réf}})^2 + \tfrac{1}{2}(\sigma_{\text{réf}}^{-2} s_{\checkmark}^2 - 1)\big), \quad (\texttt{CH'})$$

qui peut être mis en bijection croissante avec la statistique de test suivante (dont on suspectera donc les grandes valeurs), plus simple :

$$(M - \mu_{\text{réf}})^2 + S^2 - 2\sigma_{\text{réf}}^2 \ln S. \qquad (\texttt{CI'})$$

♧

Remarque (CJ'). Comme les calculs ci-dessus le montrent, même dans le cas particulièrement simple du modèle du pédagogue avec hypothèse nulle simple, il peut être assez laborieux de calculer le rapport des vraisemblances : et calculer le seuil approprié pour le test correspondant peut s'avérer encore plus compliqué ! C'est pourquoi, malgré son caractère extrêmement général, la statistique de Neyman-Pearson n'est utilisée que dans une minorité de cas : le plus souvent, il s'avèrera bien plus commode de bricoler une statistique de test ad hoc, facile à calculer et pour laquelle on sache déterminer efficacement des seuils, en s'appuyant simplement sur les principes qualitatifs présentés plus haut pour vérifier la pertinence de cette statistique de test. Et dans les cas où on souhaitera utiliser la stratégie de Neyman-Pearson, il faudra en général procéder à un certain nombre d'approximations pour se ramener à des choses calculables... ♧

Statistique à loi uniforme

Un autre principe utile à connaitre pour trouver des statistiques de test pertinentes est le suivant, qui s'appuie directement sur le principe (BX') :

Principe (CK'). *Si on arrive à trouver une statistique $t(X)$ ayant la même loi*[¶] *sous tous les $\mathbb{P}_\theta(\bullet)$ pour $\theta \in \Theta_0$, mais une loi (en général) différente pour $\theta \in \Theta_1$, alors c'est très probablement une statistique de test pertinente. En fonction de la façon dont les lois pour $\theta \in \Theta_1$ sont susceptibles de différer de la loi (commune) pour $\theta \in \Theta_0$, le critère de suspicion approprié pourra être à gauche, à droite, ou bilatéral.* ◇ !

Remarque (CL'). En fait, il n'y a pas besoin que la statistique $t(X)$ ait *exactement* la même loi sous tous les $\mathbb{P}_\theta(\bullet)$ de $\theta \in \Theta_0$ pour que ce principe reste valable : si la loi de $t(X)$ sous tous ces $\mathbb{P}_\theta(\bullet)$ est toujours *quasiment* la même, au sens où on est capable de déterminer un contrôle uniforme sur la ou les queue(s) appropriées des $\mathrm{Loi}_\theta(t(X))$ (par exemple, si toutes ces lois ont la même moyenne et la même variance, ce qui permet alors de les contrôler via l'inégalité de Bienaymé-Tchebychev), alors cela marche aussi ! ☺ Cependant, le cas qu'on rencontre le plus souvent est où les $\mathrm{Loi}_\theta(t(X))$ sont rigoureusement identiques : cas qui présente en outre l'avantage d'être plus facile à comprendre et à traiter ! ☺ ♧

Remarque (CM'). Le principe (CK') sera très utilisé dans le chapitre **??** pour construire des *intervalles de confiance.* ♧

Il y a également une variante du principe ci-dessus où l'idée de « avoir toujours la même loi » ne vaut que sur le "bord" de l'hypothèse nulle :

! **Principe (CN').** *Lorsque l'hypothèse nulle Θ_0 est définie par une inégalité (disons, $\gamma(\theta) \leqslant \gamma_0$), si on arrive à trouver une statistique $t(X)$ qui :*

(i) A toujours la même loi (ou quasiment la même loi) dans le cas d'égalité *(autrement dit, lorsque θ vérifie $\gamma(\theta) = \gamma_0$, ce qui correspond donc au "bord" de l'hypothèse nulle) ;*

(ii) A tendance à prendre des valeurs plus petites que dans le cas d'égalité quand on est à l'intérieur de l'hypothèse nulle (autrement dit, lorsqu'on considère des θ pour lesquels $\gamma(\theta) < \gamma_0$) ;

(iii) A tendance, à l'inverse, à prendre des valeurs plus grandes que le cas d'égalité quand on est dans l'hypothèse alternative (i.e. que $\gamma(\theta) > \gamma_0$),

alors il s'agit (normalement) d'une statistique de test pertinente, avec critère de suspicion à droite.

On a bien entendu un résultat similaire (avec critère de suspicion à gauche) lorsque les valeurs de $t(X)$ deviennent plus grandes à l'intérieur de Θ_0, mais plus petites dans Θ_1. ♢

12.5 Tests bilatéraux

J'ai évoqué un peu plus haut l'existence d'un critère de suspicion « bilatéral » : c'est à ce point, et aux développements afférents, que nous allons nous intéresser dans cette section.

À titre l'introduction, reprenons l'exemple du vieux professeur grincheux évoqué ci-dessus (cf. exemple (CF')), à ceci près que cette fois-ci nous considérons que tout le monde s'accorde à dire que les différentes méthodes pédagogiques ne changent rien à l'hétérogénéité des résultats des élèves ; autrement dit, nous nous restreignons à $\Theta = \mathbb{R} \times \{\sigma_{\text{réf}}\}$. Dans ce cadre, le vieux professeur défend l'hypothèse nulle $\{\theta = \theta_{\text{réf}}\}$, avec $\theta_{\text{réf}} := (\mu_{\text{réf}}, \sigma_{\text{réf}})$. On voit qu'il y a qualitativement *deux* façons dont l'hypothèse nulle peut être fausse : soit l'hypothèse nulle est fausse parce que la vraie valeur de μ est $\mu_\checkmark < \mu_{\text{réf}}$, soit l'hypothèse nulle est fausse parce qu'on aurait $\mu_\checkmark > \mu_{\text{réf}}$.

Si nous reprenons ce que nous avons vu dans les sections précédentes, nous voyons que :

[¶]. Ou « quasiment la même loi », confer remarque (CL').

— S'il s'agissait de tester l'hypothèse nulle $\{\boldsymbol{\mu} = \mu_{\text{réf}}\}$ contre l'hypothèse alternative $\{\boldsymbol{\mu} > \mu_{\text{réf}}\}$, une statistique de test appropriée serait la moyenne $\boldsymbol{M}$ obtenue par les élèves, avec critère de suspicion à droite ;
— Si, en revanche, il s'agissait de tester l'hypothèse nulle $\{\boldsymbol{\mu} = \mu_{\text{réf}}\}$ contre l'hypothèse alternative $\{\boldsymbol{\mu} < \mu_{\text{réf}}\}$, on trouverait *aussi* que $\boldsymbol{M}$ est une statistique appropriée, mais cette fois-ci avec critère de suspicion à *gauche* !

On a donc toujours la même statistique de test $\boldsymbol{M}$; mais cette fois-ci il y a deux types de valeurs extrêmes qu'on a envie de rejeter : d'un côté, les valeurs trop grandes de $\boldsymbol{M}$ nous semblent suspectes du point de vue de l'hypothèse nulle parce qu'elles nous inciteraient plutôt à penser que $\boldsymbol{\mu} > \mu_{\text{réf}}$; et de l'autre, les valeurs trop petites de $\boldsymbol{M}$ sont *aussi* suspectes, cette fois-ci parce qu'elles nous inciteraient plutôt à penser que $\boldsymbol{\mu} < \mu_{\text{réf}}$! En résumé, on a envie de dire que plus les valeurs de $\boldsymbol{M}$ sont "extrêmes" (que ce soit vers la droite ou vers la gauche), plus elles sont suspectes... Eh bien, c'est précisément cela qu'on appellera « critère de suspicion bilatéral » :

Définition (CO'). Lorsqu'on associe un *critère de suspicion bilatéral* à une statistique de test $\boldsymbol{T}$, cela signifie qu'on considère que les valeurs les plus extrêmes sont les plus suspectes :
— D'une part, asymptotiquement (au-delà d'une certaine valeur, je veux dire), on considère que plus la réalisation $t_{\checkmark}$ de la statistique de test est grande, plus cela nous incite à pencher vers l'hypothèse alternative au détriment de l'hypothèse nulle ;
— D'autre part, toujours asymptotiquement (mais cette fois-ci *en deçà* d'une certaine valeur), on considère que plus la réalisation $t_{\checkmark}$ de la statistique de test est petite, plus cela nous incite à pencher vers l'hypothèse alternative au détriment de l'hypothèse nulle !

♡

Remarque (CP'). Attention ! Il pourrait être tentant de croire que le critère ci-dessus équivaudrait à quelque chose comme « plus la valeur absolue $|T|$ est grande, plus c'est suspect ». Sauf que cela sous-entendrait qu'il devrait être *exactement aussi suspect* d'observer, mettons, $t_{\checkmark} = -177$ que $t_{\checkmark} = +177$... Or, cela n'a aucune raison d'être le cas ! (et ce ne le sera pas en général, confer plus loin). Tout ce que dit la définition ci-dessus, c'est que le niveau de suspicion associé à l'observation de la valeur t est maximal pour $t \to -\infty$, puis décroit lorsqu'on se rapproche des valeurs "centrales", avant d'augmenter à nouveau pour devenir, à nouveau, maximal lorsque $t \to +\infty$. ♧

Dans le cas d'un critère de suspicion bilatéral, puisque nous devons rejeter à la fois les valeurs de $\boldsymbol{T}$ trop petites et les valeurs trop grandes, il ne va pas y avoir un, mais deux seuils. Mais attention : c'est le risque *total* d'avoir un test positif qui doit demeurer $\leqslant \boldsymbol{\alpha}$, par juste le risque associé à chaque seuil ! Cela nous amène donc à définir notre test à partir des intervalles de fluctuation :

Procédure (CQ'). Une partition de l'espace du paramètre caché en hypothèse nulle et hypothèse alternative étant donnée, un niveau $\boldsymbol{\alpha} \in]0, 1/2]$ étant fixé, et une statistique de test *bilatérale* $\boldsymbol{t}(\boldsymbol{X})$ étant choisie, on construit canoniquement, à partir de ces éléments, « le » test de niveau $\boldsymbol{\alpha}$ de la façon suivante. On construit l'*intervalle d'acceptation* de notre test comme l'union des intervalles de fluctuation !

des $\mathrm{Loi}_\theta(t(X))$ pour θ dans l'hypothèse nulle :

$$I_{\mathrm{acc}} \coloneqq \mathrm{Adh}\Big(\bigcup_{\theta\in\Theta_0} \big[\mathrm{Qtile}\big(\mathrm{Loi}_\theta(t(X)); \alpha/2\big), \mathrm{Qtile}\big(\mathrm{Loi}_\theta(t(X)); 1-\alpha/2\big)\big] \Big). \qquad (\texttt{CR'})$$

Quitte à remplacer 8_{acc} par son adhérence topologique (ce qu'on peut faire sans scrupule, car cela ne change rien du point de vue des applications pratiques), cela s'écrit aussi

$$I_{\mathrm{acc}} = \big[\inf_{\theta\in\Theta_0} \mathrm{Qtile}\big(\mathrm{Loi}_\theta(t(X)); \alpha/2\big), \sup_{\theta\in\Theta_0} \mathrm{Qtile}\big(\mathrm{Loi}_\theta(t(X)); 1-\alpha/2\big)\big]. \qquad (\texttt{CS'})$$

Ensuite, on définit notre test comme $\{t(X) \notin I_{\mathrm{acc}}\}$, autrement dit, on accepte d'hypothèse nulle si et seulement si la statistique de test tombe dans l'intervalle d'acceptation. ♡

12.6 Régime asymptotique des tests

Intérêt de l'étude asymptotique des tests

Le problème avec les mathématiques, c'est que bien souvent la beauté des constructions théoriques se heurte à la dure réalité de calculs inextricables... Les tests n'échappent malheureusement pas à cette règle : l'information cruciale sur un test, son niveau (dans le cas booléen) ou sa p-valeur, est bien souvent trop compliquée à calculer exactement. En revanche, on peut parfois obtenir des approximations intéressantes... C'est ce qui conduit à l'objet de cette section : les tests *asymptotiques*. Dans le cadre asymptotique, on a souvent la possibilité de prouver des résultats mathématiques intéressants qui nous rassurent quant à la pertinence de tel ou tel test : autrement dit, on va mettre au point des méthodes statistiques qui sont censées bien marcher sous réserve que le nombre d'observations soit suffisamment grand.

Les tests asymptotiques, comme leur nom l'indique, appartiennent au domaine de la statistique... asymptotique : autrement dit, on va se placer dans un modèle statistique avec un (ou plusieurs) paramètres du modèle pour lequel on considère un certain régime asymptotique (cf. § 7.4), et pour toute valeur λ du paramètre du modèle, on considèrera un test booléen $T^{(\lambda)} : \mathcal{X}^{(\lambda)} \to \{\texttt{VRAI}, \texttt{FAUX}\}$; nous chercherons alors à dire comment se comporte le test $T^{(\lambda)}$ pour un certain régime asymptotique de λ : typiquement, on aura $\lambda \in \mathbb{N}$ et on étudiera la limite $\lambda \to \infty$.

Remarque (CT'). Attention !! Les mathématiciens adorent faire de la statistique asymptotique, parce que cela leur permet d'avoir de jolis théorèmes. Mais les ingénieurs, eux, doivent s'en méfier comme de la peste ! En effet, il est bien beau de savoir que le résultat est valable si le nombre d'observations est « suffisamment grand », mais comment savoir si le nombre d'observations dont on dispose *en pratique* est *réellement* suffisamment grand ?… C'est pourquoi un ingénieur raisonnable ne devrait appliquer une méthode asymptotique que s'il dispose au moins d'une méthode (idéalement rigoureuse, à défaut heuristique) pour savoir si l'approximation asymptotique est raisonnable ou non ! ♧

Tests asymptotiques

! **Définition** (CU') (Niveau asymptotique d'un test)**.** Dans un modèle ayant un paramètre pour lequel on considère une asymptotique, soit $Test^{(\lambda)}$ un test booléen

de l'hypothèse $\{\theta \in \Theta_0\}$. On dit que le test est *asymptotiquement de niveau* α lorsque :

$$\forall \theta \in \Theta_0 \quad \overline{\lim_{\lambda \to \infty}} \, \mathbb{P}(Test^{(\lambda)} \mid \theta = \theta) \leqslant \alpha. \qquad (\texttt{CV'})$$

♡

Remarque (CW'). Notez que la définition pour un test asymptotique requiert seulement qu'il y ait convergence pour chaque θ, mais qu'il est tout à fait possible que chaque θ converge à son propre rythme, et qu'on ne puisse trouver aucune borne de convergence valable pour tous les θ à la fois ! ♧

Proposition (CX'). *Si un test $Test^{(\lambda)}$ est de niveau α (pour tout λ s'entend), alors il est asymptotiquement de niveau α(Par contre, la réciproque est fausse).* ♢

Démonstration. Immédiat. ♤

Remarque (CY'). Il est tout-à-fait possible d'avoir un niveau asymptotique pour un test égal à… 0 (même si en pratique on préfère fixer un niveau α et s'y tenir, ou mieux, raisonner en termes de p-valeurs, confer chapitre suivant) : cela n'a en fait rien de bien extraordinaire, pas plus que la notion de consistance que nous allons voir ci-après. ♧

Consistance

Toujours dans le cadre du régime asymptotique, un critère de qualité très important, lorsqu'on parle de tests, est la notion de *consistance* :

Définition (CZ'). Un test $Test^{(\lambda)}$ de l'hypothèse $\{\theta \in \Theta_0\}$ est dit *consistant* lorsque, pour tout $\theta \notin \Theta_0$, $\mathbb{P}(Test^{(\lambda)} \mid \theta = \theta) \to 1$ lorsque $\lambda \to \infty$. ♡ !

Remarque (DA'). On a l'impression avec la définition ci-dessus que la philosophie du test est inversée : asymptotiquement, on n'a plus aucun risque d'erreur dans le cas où $\theta_{\checkmark} \in \Theta_1$, alors qu'il reste toujours un risque α de faux positif si $\theta_{\checkmark} \in \Theta_0$... Mais, comme dit dans la remarque (CT'), la « vraie » statistique est en fait la statistique *non* asymptotique (ou au moins *uniformément* asymptotique) : les résultats asymptotiques sont juste là pour nous rassurer mathématiquement sur ce qu'on fait ! ◡̈ ♧

Remarque (DB'). Tous les tests que nous serons amenés à voir dans ce cours, lorsqu'il y a un régime asymptotique, seront en fait consistants. On peut en fait montrer que, au moins en ce qui concerne les modèles d'échantillonnage, sous des hypothèses générales assez souples sur le modèle probabiliste et la structure de l'hypothèse nulle, il est toujours possible de trouver un test consistant pour l'hypothèse nulle. ♧

Exemple (DC') (Test du χ^2). Le test du χ^2 [‖] est un test extrêmement classique pour tester l'indépendance de deux variables qualitatives. Soient $\mathcal{X}$ et $\mathcal{Y}$ des ensembles finis non vides et $\mu_{\checkmark}$ une loi portée par $\mathcal{X} \times \mathcal{Y}$. Le paramètre caché de notre modèle probabiliste est $\mu_{\checkmark}$ elle-même, et l'expérience statistique consiste à observer un nombre n librement fixé de réalisations i. i. d. de la loi $\mu_{\checkmark}$. Ces réalisations

[‖]. 'χ' est la lettre grecque « chi » — attention, on prononce /ki/ —; à l'oral, on lit « test du chi-deux » — prononcé /kidø/.

sont notées resp. $(x_{1\checkmark}, y_{1\checkmark}), \dots, (x_{n\checkmark}, y_{n\checkmark})$, vues comme les réalisations des variables aléatoires $(X_1, Y_1), \dots, (X_n, Y_n)$.

Les test du χ^2 se propose de tester l'hypothèse nulle que, sous la véritable loi $\mu_\checkmark$, les variables X et Y sont indépendantes. Autrement dit, notant $\mu(x, y)$ la probabilité sous μ de tirer le couple (x, y), resp. $\mu(x) \coloneqq \sum_{y' \in \mathcal{Y}} \mu(x, y')$ la probabilité que la première variable soit x, resp. $\mu(y) \coloneqq \sum_{x' \in \mathcal{X}} \mu(x', y)$ la probabilité que la seconde variable soit y, on veut tester l'hypothèse que

$$\forall (x, y) \in \mathcal{X} \times \mathcal{Y} \quad \mu_\checkmark(x, y) = \mu_\checkmark(x)\mu_\checkmark(y). \tag{DD'}$$

Pour ce faire, on introduit une statistique de test appelée *statistique du* χ^2, définie de la façon suivante. Notons $N(x, y)$ la variable aléatoire comptant le nombre d'observations i pour lesquelles on a $\{(X_i, Y_i) = (x, y)\}$, resp. $N(x) = \sum_{y' \in \mathcal{Y}} N(x, y')$ le nombre de fois qu'on a $\{X_i = x\}$, resp. $N(y) = \sum_{x' \in \mathcal{X}} N(x', y)$ le nombre de fois qu'on a $\{Y_i = y\}$. Alors la statistique du χ^2 est

$$T \coloneqq n \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{N(x, y)^2}{N(x)N(y)} - n. \tag{DE'}$$

(En prenant la convention $0/0 = 0$ si nécessaire). (On peut montrer que T est toujours positive).

Le résultat fondamental sur lequel s'appuie le test du χ^2 est le suivant :

Proposition (DF'). *Si $\mu_\checkmark$ vérifie l'hypothèse d'indépendance, alors quand $n \to \infty$, on a*

$$\mathrm{Loi}_\checkmark(T^{(n)}) \to \mathrm{ChiDeux}\big((|\mathcal{X}| - 1) \times (|\mathcal{Y}| - 1)\big) \tag{DG'}$$

lorsque $\mu_\checkmark$ associe une masse non nulle à tous les éléments de $\mathcal{X}$ et $\mathcal{Y}$, et plus généralement

$$\mathrm{Loi}(T^{(n)}) \to \mathrm{ChiDeux}\big((|\mathrm{supp}_X \mu_\checkmark| - 1) \times (|\mathrm{supp}_Y \mu_\checkmark| - 1)\big), \tag{DH'}$$

où $|\mathrm{supp}_X \mu_\checkmark|$ est le nombre de $x \in \mathcal{X}$ pour lesquels $\mu_\checkmark(x) > 0$, resp. $|\mathrm{supp}_Y \mu_\checkmark|$ est le nombre de $y \in \mathcal{Y}$ pour lesquels $\mu_\checkmark(y) > 0$. ♢

On en déduit en particulier que pour un niveau t fixé, lorsque n tend vers l'infini,

$$\overline{\lim}\, \mathbb{P}(T^{(n)} > t) \leqslant \mathbb{P}\big(\mathrm{ChiDeux}\big((|\mathcal{X}| - 1) \times (|\mathcal{Y}| - 1)\big) > t\big) \tag{DI'}$$

(où l'on a admis le fait que $\mathbb{P}\big(\mathrm{ChiDeux}(d) > t\big)$ est croissant lorsque d augmente — ce qui est très facile à vérifier dès qu'on connait la définition de la loi du ChiDeux).

Par conséquent, cela suggère le protocole de test suivant : étant fixé un niveau α, chercher le plus grand t tel que $\mathbb{P}\big(\mathrm{ChiDeux}\big((|\mathcal{X}| - 1) \times (|\mathcal{Y}| - 1)\big) \geqslant t\big) \geqslant \alpha$, puis rejeter l'hypothèse nulle lorsque $T > t$.

La proposition (DF') nous dit alors précisément que ce test sera asymptotiquement de niveau α. Par ailleurs, il est aisé de voir que le test proposé ci-dessus est consistant. ♧

Exemple (DJ') (Test du χ^2, suite). Dans le test du χ^2 proposé ci-dessus, on peut cependant démontrer que le niveau $\alpha^{(n)}$ de ce test pour une valeur donnée de n ne converge *pas*, quand n tend vers l'infini, vers α (il n'y a donc pas « uniforme consistance » du test), mais vers une valeur un peu plus élevée.

Néanmoins, dans le cas de ce modèle, on peut en fait trouver une règle qui permet d'obtenir une estimation quantitative sur la convergence du test vers son asymptotique. Pour ce faire, un certain seuil $N_{\min}$ étant introduit (on prend généralement $N_{\min} = 5$), on considère la statistique modifiée

$$\check{T} \coloneqq \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} n\frac{\check{N}(x,y)^2}{N(x)N(y)} - n, \tag{DK'}$$

où l'on pose

$$\check{N}(x,y) = \begin{cases} N(x,y) & \text{pour } N(x,y) \geqslant N_{\min} \\ 0 & \text{pour } N(x,y) < N_{\min} \end{cases}; \tag{DL'}$$

cette statistique a la même asymptotique que T sous l'hypothèse nulle, et conduit donc elle aussi à un test de niveau α. En outre, quoique le test fondé sur $\check{T}$ soit un peu moins puissant que celui fondé sur T, il est lui aussi consistant. Et surtout, cette fois-ci notre test a un niveau borné *uniformément* sur Θ_0 de façon *non* asymptotique, uniformément en n, le niveau en question pouvant être fixé arbitrairement proche de α pour peu qu'on ait choisi $N_{\min}$ suffisamment grand.

En pratique, pour $N_{\min} = 5$ et pour des valeurs de n de quelques dizaines ou quelques centaines, le niveau réel ainsi obtenu pour le test modifié est raisonnablement proche de α (à condition qu'on n'ait pas choisi α trop petit). ♧

Chapitre 13

La p-valeur

13.1 p-valeur associée à une statistique de test

Dans le chapitre précédent, vous venez de voir comment fonctionne un test booléen. Maintenant "oubliez tout ce que vous venez d'apprendre [*]" : en pratique, nous allons utiliser une autre façon de formuler les tests, plus précise que la notion "binaire" de test booléen. À savoir : la *p-valeur* !

Pour comprendre la raison derrière l'introduction du concept de p-valeur, revenons sur la procédure que nous avions suivie lorsqu'il y a une statistique de test T avec critère de suspicion unilatéral. (Pour fixer les idées, disons ici qu'on suspecte les *petites* valeurs de T). Notre procédure fonctionnait ainsi : nous étant fixé un niveau de tolérance α pour le risque de première espèce, nous avions choisi un seuil $t_{\text{seuil}}(\alpha)$ tel que

$$\forall \theta \in \Theta_0 \quad \mathbb{P}_\theta(T < t_{\text{seuil}}(\alpha)) \leqslant \alpha\,; \tag{DM'}$$

et, tant qu'à faire, nous avons choisi ce seuil aussi grand que possible : autrement dit, nous avons pris

$$t_{\text{seuil}}(\alpha) = \inf\{\text{Qtile}(\text{Loi}_\theta(T);\alpha) \mid \theta \in \Theta_0\}, \tag{DN'}$$

où $\text{Qtile}(\bullet;\alpha)$ désigne le quantile de niveau α (dans le cas d'un critère de suspicion à droite, nous aurions pris cette fois-ci le supremum des quantiles de niveau $1-\alpha$). La réponse de notre test (booléen) est alors `VRAI` si $t_{\checkmark} < t_{\text{seuil}}(\alpha)$, resp. `FAUX` si $t_{\checkmark} \geqslant t_{\text{seuil}}(\alpha)$.

Par conséquent, on voit que la réponse de notre test booléen résulte de la conjonction de deux facteurs :

— La valeur effectivement observée $t_{\checkmark}$ de la statistique de test ;

— Le niveau de risque α (via la détermination de $t_{\text{seuil}}(\alpha)$).

[*]. Façon de parler...! ☋ Vous devez *quand même* maitriser la notion de test booléen : d'une part, parce qu'il arrive parfois qu'on le rencontre en pratique malgré tout ; et d'autre part, parce que comprendre les tests booléens est une étape extrêmement utile pour comprendre les p-valeurs (l'objet du présent chapitre) ainsi que les intervalles de confiance (chap. 14). Ce que je veux dire par « oubliez les tests booléens », c'est que, aussi loin qu'il s'agit d'étudier la plausibilité d'une hypothèse nulle, l'approche par p-valeur est à la fois plus rapide, plus précise et plus riche : c'est pourquoi, dans la pratique industrielle, cette pratique méritera d'être systématiquement préférée dans la mesure du possible ! Au demeurant, c'est bien ce que font les logiciels de statistique : il expriment toujours les résultats des tests fréquentistes d'hypothèse en termes de p-valeur, jamais en termes booléens ! ☊

Il se trouve que la façon de conjuguer ces deux facteurs peut être faite sans jamais avoir à expliciter $t_{\text{seuil}}(\alpha)$. En effet, pour tout $t \in \mathbb{R}$, introduisons la quantité

$$p(t) \coloneqq \sup\{\mathbb{P}_\theta(T \leqslant t) \mid \theta \in \Theta_0\}. \tag{DO'}$$

On a alors le résultat suivant, pas très difficile à démontrer :

Lemme (DP').

$$t_{\text{seuil}}(\alpha) = \inf\{t \in \mathbb{R} \mid p(t) > \alpha\}. \tag{DQ'}$$

◇

Comme, en outre, la fonction $t \mapsto p(t)$ est évidemment croissante, il découle du lemme (DP') que, sauf cas particulier [†] (que nous négligerons dans la suite), la condition $\{t_{\checkmark} < t_{\text{seuil}}(\alpha)\}$ est équivalente à la condition $\{p(t_{\checkmark}) \leqslant \alpha\}$, laquelle ne demande cette fois-ci que l'évaluation de $p(t_{\checkmark})$, sans avoir à calculer explicitement $t_{\text{seuil}}(\alpha)$!

La valeur $p(t_{\checkmark})$, que dorénavant nous noterons $p_{\checkmark}$, permet donc de savoir si le test booléen de niveau α associé à notre statistique de test (pour le critère de suspicion considéré) est positif ou négatif, ce qui montre clairement son intérêt ! C'est pourquoi on lui donne un nom : on l'appelle la *p-valeur* de notre test. Comme la définition ci-dessous l'explique, on peut comprendre la *p*-valeur comme « la probabilité (supremale), lorsque l'hypothèse nulle est vraie, d'obtenir un résultat au moins aussi suspect que celui qui a été effectivement observé » :

! **Définition (DR')** (*p*-valeur pour une statistique de test, cas unilatéral)**.** Soit T une statistique destinée à servir de statistique de test pour l'hypothèse nulle $\{\theta \in \Theta_0\}$, et soit $t_{\checkmark}$ la réalisation de cette statistique. Alors la *p-valeur* [‡] de notre test est la (réalisation de) statistique $p_{\checkmark}$ définie ainsi : c'est le supremum, sur toutes les valeurs θ appartenant à l'hypothèse nulle, de la probabilité que, sous $\mathbb{P}_\theta$, la statistique de test T prenne une valeur au moins aussi suspecte que la valeur $t_{\checkmark}$ effectivment observée. Autrement dit, c'est la quantité suivante :

— Si ce sont les petites valeurs de T qui sont suspectes (« test à gauche ») :

$$p_{\checkmark} \coloneqq \sup_{\theta \in \Theta_0} \mathbb{P}(T \leqslant t_{\checkmark} \mid \theta = \theta). \tag{DS'}$$

[†]. Le « cas particulier » en question concerne la situation où on a *exactement* $p(t_{\checkmark}) = \alpha$. Néanmoins, on peut considérer comme assuré que cela n'arrivera jamais en pratique, car les valeurs α susceptibles d'intervenir dans vos analyses seront toujours des seuils de références fixés indépendamment du modèle (p. ex. 5 %, 1 ‰, ...), seuils auxquels la *p*-valeur, qui est en général un nombre réel tout ce qu'il y a de plus quelconque, n'aura aucune raison d'être rigoureusement égale... ! D'où ma volonté de ne pas s'ennuyer avec ce cas d'égalité. En fait, il serait *possible* de traiter rigoureusement l'équivalence entre théorie des tests booléens et approche par *p*-valeurs d'une façon qui ne présente aucune exception au niveau des cas d'égalité ; mais cela obligerait alors à revenir sur la théorie des tests booléens, et même sur celle des intervalles de flucutation, pour savoir dans quel cadre on doit ouvrir ou pas les bornes de tel ou tel intervalle : ce qui représenterait beaucoup de complexité supplémentaire pour un apport dérisoire... ☹ Cet apport serait en outre d'autant plus dérisoire que nous savons très bien (merci George Box ☺) que les nombres réels intervenant dans les modèles statistiques et les observations sont eux-mêmes inévitablement entachés d'une certaine imprécision : il serait donc passablement stupide de compliquer le cours pour la seule satisfaction de savoir traiter rigoureusement le cas d'une égalité parfaite qui, non seulement n'a aucune raison de se produire, mais n'a de toutes façons aucune réalité concrète ! ☺

[‡]. La *p*-valeur est traditionnellement notée par la lettre 'p', d'où son nom. Dans le cadre de ce cours, nous utiliserons plutôt la notation '$p_{\checkmark}$', reflétant ainsi le fait que la *p*-valeur est en fait la réalisation d'une certaine statistique — statistique que nous pourrons alors noter 'P' lorsqu'on la verra en tant que variable aléatoire).

— Si ce sont les grandes valeurs de T qui sont suspectes (« test à droite ») :

$$p_{\checkmark} \coloneqq \sup_{\theta\in\Theta_0} \mathbb{P}(T \geqslant t_{\checkmark} \mid \theta = \theta). \qquad (\texttt{DT'})$$

♡

Remarque (DU'). Attention ! Lorsqu'on définit la p valeur, on considère (le supremum de) la probabilité d'être *au moins aussi* suspect que l'observation ; alors que lorsqu'on définissait le test booléen, on considérait le fait d'être *strictement plus suspect* que (l'infimum de) le quantile de la loi de T ! On a donc une inégalité *large* dans le cadre de la définition (DR'), contrairement à la procédure (BD') qui requérait une inégalité *stricte*... [§] ♧

On peut adapter la définition de la p-valeur aux tests bilatéraux :

Définition (DV') (p-valeur pour les tests bilatéraux)**.** Dans le cas d'une statistique de test T avec critère de suspicion bilatéral, la p-valeur déduite de la statistique de test T est définie de la façon suivante, où $t_{\checkmark}$ est la réalisation de T (en pratique, on utilisera plutôt de la majoration donnée par le membre de droite [¶]) : !

$$p_{\checkmark} \coloneqq 2 \sup_{\theta\in\Theta_0} \big(\mathbb{P}_\theta(T \leqslant t_{\checkmark}) \wedge \mathbb{P}_\theta(T \geqslant t_{\checkmark})\big) \leqslant 2\big(\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(T \leqslant t_{\checkmark}) \wedge \sup_{\theta\in\Theta_0} \mathbb{P}_\theta(T \geqslant t_{\checkmark})\big). \qquad (\texttt{DW'})$$

♡

Remarque (DX'). La formule ci-dessus peut avoir l'air passablement obscure ; mais en réalité, elle s'interprète de façon tout à fait logique, comme je vais vous l'expliquer à présent.

De manière générale, l'idée de la p-valeur, c'est qu'on considère la probabilité d'être *au moins aussi suspect* que $t_{\checkmark}$. Dans le cas d'un critère de suspicion unilatéral, disons à droite, c'était facile à interpréter : cela signifiant juste « probabilité d'être au moins aussi *grand* que $t_{\checkmark}$ ». Maintenant, si on est *bilatéral*, comment faut-il traduire les choses ?...

Prenons un exemple pour comprendre. Mettons ici que la probabilité d'être strictement inférieur à $t_{\checkmark}$ vaut 11 %, que celle d'être exactement égal à $t_{\checkmark}$ vaut 2 %, et que celle d'être strictement supérieur à $t_{\checkmark}$ vaut 87 %. En l'occurrence, puisqu'il y a moins de probabilité à gauche de $t_{\checkmark}$ qu'à droite de $t_{\checkmark}$, ce qui rend ce $t_{\checkmark}$-là suspect, c'est le fait qu'il soit particulièrement *petit* ! Ainsi, la probabilité d'être au moins aussi suspect que $t_{\checkmark}$ *en étant suspect dans le même sens que lui* vaut ici 13 %. Cette

[§]. En fait, ce n'est pas si paradoxal qu'il n'y parait à première vue. Certes, pour les tests booléens, on avait à considérer des évènements du type $\{T < t_{\text{seuil}}\}$, alors que pour la p valeur, on considère des évènements du type $\{T \leqslant t_{\checkmark}\}$. Mais, dans le premier cas, la variable aléatoire T était destinée à être remplacée *in fine* par $t_{\checkmark}$; alors que dans le second, 'T' fait en réalité référence à un "autre" tirage de la statistique de test, si on *relançait* l'expérience, tirage qu'on pourrait appeler 'T''... Dès lors, si on se place au niveau des réalisations, on s'aperçoit qu'on considère dans le premier cas l'affirmation « $t_{\checkmark} < t_{\text{seuil}}$ », et dans le second cas, l'affrimation « $t'_{\checkmark} \leqslant t_{\checkmark}$ ». Le membre commun entre ces deux affirmations étant $t_{\checkmark}$, si on veut réellement rendre les deux affirmations comparables, on va placer "$t_{\checkmark}$" du côté gauche à chaque fois : et on se rend compte alors que les expressions deviennent resp. « $t_{\checkmark} < t_{\text{seuil}}$ » et « $t_{\checkmark} \not< t'_{\checkmark}$ » : ainsi, en fait, on s'intéressait bien, dans les deux cas, à des questions sur le fait que $t_{\checkmark}$ soit strictement inférieur, ou pas, à une valeur de comparaison !

[¶]. Dans les situations rencontrées en pratique, en effet, la comparaison entre les valeurs $\mathbb{P}_\theta(T \leqslant t_{\checkmark})$ et $\mathbb{P}_\theta(T \geqslant t_{\checkmark})$ a lieu dans le même sens pour toutes les valeurs de $\theta \in \Theta_0$, de sorte que la majoration ne perd alors aucunement en précision.

valeur correspond au minimum « $\mathbb{P}(T \leqslant t_{\checkmark}) \wedge \mathbb{P}(T \geqslant t_{\checkmark})$ » qui apparait dans notre formule : selon que $t_{\checkmark}$ est situé du côté gauche de la distribution de T ou du côté droit de la probabilité, la « probabilité d'être au moins suspect que lui en penchant du même côté » sera soit $\mathbb{P}(T \leqslant t_{\checkmark})$, soit $\mathbb{P}(T \geqslant t_{\checkmark})$: et dans tous les cas, ce sera toujours *la plus petite* des deux probabilités en question : étant donné que dire que $t_{\checkmark}$ est suspect *du fait de sa petitesse* (plutôt que du fait de sa grandeur) devient précisément à dire que « être plus petit que $t_{\checkmark}$ » est quelque chose de plus rare que « être plus grand que $t_{\checkmark}$ » !

Soit ; nous avons donc compris ce fameux symbole de minimum. Maintenant, d'où vient le facteur 2 ? Eh bien, si nous reprenons l'exemple ci-dessus, nous avons dit que la probabilité d'être au moins aussi suspect que $t_{\checkmark}$ *en étant suspect du fait qu'on est petit* valait 13 % : d'accord ; mais on pourrait tout aussi bien être au moins aussi suspect que $t_{\checkmark}$ *en étant suspect du fait qu'on est grand* ! (Puisqu'un valeur "très très" grande sera bel et bien plus suspect qu'une valeur juste "très" petite). Mais que vaut cette seconde probabilité ? En fait, comme nous avons décidé de répartir le risque de façon symétrique entre les deux queues de la distribution, « être plus suspect qu'un niveau donné du fait qu'on est grand » doit avoir la même probabilité que « être plus suspect que ce même niveau donné du fait qu'on est petit », d'où la multiplication par 2 [‖].

Donc en fait, on devrait plutôt écrire le membre du milieu dans la formule comme

$$\sup_{\theta \in \Theta_0} \big(2 \times \big(\mathbb{P}_\theta(T \leqslant t_{\checkmark}) \wedge \mathbb{P}_\theta(T \geqslant t_{\checkmark})\big)\big), \qquad (\texttt{DY'})$$

pour bien rendre visible le fait qu'on est toujours en train de considérer la probabilité d'être au moins aussi suspect : mais cela rend la formule difficilement lisible du fait du grand nombre de parenthèses... ⁀

Un mot, enfin, en ce qui concerne le membre de droite de la formule. Une fois qu'on a compris l'idée que « la probabilité d'être au moins aussi suspect, dans le cas bilatéral, c'est le double de celle des deux valeurs qui est la plus faible entre, d'une part, la probabilité d'être au moins aussi petit, et d'autre part, la probabilité d'être au moins aussi grand », on interprète le membre de droite en disant : « la (majoration de la) p-valeur dans le cas bilatéral peut se définir ainsi : on regarde la p-valeur qu'on aurait eue dans le cas d'un critère de suspicion à gauche, d'une part ; puis la p-valeur qu'on aurait eue dans le cadre d'un critère de suspicion à droite, d'autre part ; on prend *la plus petite* de ces deux p-valeurs (puisque c'est celle qui nous dit quel est le critère de suspiction qui rendrait notre observation effective la plus suspecte) ; et on multiplie par 2 pour tenir compte du fait qu'il y a *deux* façons possibles, considérées de façon symétrique, d'être suspect ». Ouf ! ◡ ♧

Remarque (DZ'). On remarquera que la formule donnée ci-dessus pour déterminer la p-valeur (ainsi que sa majoration) dans le cas d'un test bilatéral peut produire, dans certains cas, des valeurs *strictement supérieures* à 1 : ce qui n'est cohérent, ni avec l'interprétation de la p valeur comme un supremum de probabilités, ni avec le reste du chapitre (où il est toujours entendu que les p-valeurs sont $\leqslant 1$), ni avec

[‖]. Si on veut vraiment rentrer dans le détail, il s'avère malheureusement que la phrase que je viens d'écrire requiert, pour être parfaitement correcte, la prise en compte d'un certain nombre de subtilités lorsqu'on considère des distributions de probabilité discrètes... Je ne rentrerai pas dans ces subtilités ici, vous demandant simplement de me croire sur parole quand je dis que le raisonnement ci-dessus justifiant la multiplication par 2 est, malgré son allure un peu bancale, bel et bien solide dans le cadre de la façon dont ce cours traite les statistiques de test ! ☺

l'usage fait dans les logiciels (qui ne renvoient jamais de p-valeur supérieure à 1). Pour autant, la formule n'est pas *techniquement* fausse pour autant, car rien n'empêche de construire la théorie des p-valeurs en autorisant les valeurs > 1 ; et tous les théorèmes que nous avons énoncés restent alors encore valables... En fait, dans le paradigme où on autorise les p-valeurs à dépasser 1, on s'aperçoit que, si P est une p-valeur, alors $P \wedge 1$ est *aussi* une p-valeur, ce qui explique pourquoi il est en fait inutile de s'encombre de valeurs > 1. Un corolaire de ce point est que, dans la formule ci-dessus définissant la p-valeur, on pourrait rajouter « $\wedge 1$ » sans rien changer à la validité de ce qui est écrit. (Du coup, si à un moment vous calculez par cette formule une p-valeur de, mettons, 1,04, vous pouvez tout à fait légitimement préciser « que l'on peut tronquer à 1 », et considérer ensuite que la *véritable* p-valeur est plutôt 1). ♧

Comme annoncé, à partir de la p-valeur, on peut avoir la réponse à n'importe quel test booléen :

Théorème (EA') (Lien entre p-valeur et résultats des tests booléens)**.** *Une statistique de test (et un critère de suspicion (unilatéral) associé) étant fixée, le test de niveau α défini à partir de cette statistique sera positif si, et seulement si, la p-valeur associée à cette statistique est inférieure ou égale à α* [**] [††] ◇

!

Remarque (EB')*.* Ainsi, la p-valeur nous permet d'un seul coup de donner la réponse des tests booléens pour *n'importe quelle* valeur de α : soit α est plus petit que la p-valeur et alors on accepte l'hypothèse nulle ; soit α est plus grand, et alors on la rejette ! Du coup, plutôt que de fixer arbitrairement un α dès le départ, il est beaucoup plus malin de calculer la p-valeur, et de considérer que celle-ci donne une réponse *graduée* à notre test. ♧

!

13.2 Notion intrinsèque de p-valeur

Cette courte section est entièrement "culturelle", et peut donc être sautée sans encombre ☺

Lorsque nous avions défini les tests booléens, nous avions d'abord dit rigoureusement ce que signifiait « être un test booléen de niveau α » *dans l'absolu*, avant de voir, *en pratique*, comment on construisait de tels tests booléens à partir de statistiques de test. Dans le traitement ci-dessus que nous avons fait des p-valeurs, par contre, nous nous sommes appuyés de façon *fondamentale* sur l'idée de statistique de test : notre définition n'a en effet pas de sens en l'absence de statistique de test ! Est-ce à dire qu'il n'y a pas de définition « absolue » de la notion de p-valeur ?...

[**]. En fait, de la façon dont ont été construits les tests booléens dans ce cours (à savoir, en incluant systématiquement les valeurs seuil dans la zone de négativité), ce théorème n'est pas cent pour cent rigoureux... Un théorème exact serait le suivant : d'une part, si le test booléen de niveau α est positif, alors la p-valeur sera inférieure ou égale à α ; d'autre part, le test {la p-valeur est $\leqslant \alpha$} est lui-même un test booléen de niveau α. (Mais il est possible que, lorsque la statistique de test est précisément égale à la valeur seuil, le test construit à partir de la p-valeur soit positif, alors que par convention le test construit à partir de la statistique de test est négatif au niveau du seuil). Néanmoins, les cas où l'équivalence écrite dans l'énoncé ne tient pas sont extrêmement particuliers, donc il vaut mieux retenir cet énoncé-là, qui de toutes façons est "quasiment vrai", seuls les cas où la p-valeur est rigoureusement égale à α pouvant poser problème.

[††]. On remarquera que, pour le coup, le test consistant à regarder si la p-valeur est $\leqslant \alpha$ est un test booléen qui sera *positif* au niveau de la valeur seuil, contrairement à ce qui a été fait dans le chapitre précédent. Il y a des raisons techniques profondes à ce que j'aie préféré écrire les choses ainsi, mais sur lesquelles je préfère ne pas m'apesantir... En pratique, soyez rassurés : si jamais vous rencontrez une p-valeur égale à *exactement* 6 %, et que vous écrivez « puisqu'on est pile au niveau du seuil, le test est négatif, donc on accepte l'hypothèse nulle au risque 6 % », je considèrerai cela comme correct aussi ! ☺ Et de toutes façons, ce genre d'égalité exactes n'arrive normalement *jamais* dans la vie réelle... ☹

En fait, si : on peut bel et bien définir la notion de p-valeur sans passer par des statistiques de test ; mais comme c'est un usage assez rare en pratique, il m'a semblé dispensable de préciser cette définition dans le cadre de ce cours. En fait, il y a même *deux* définitions possibles de la notion de p-valeur : la première, la plus usuelle dans le ouvrages de statistique, où on généralise légèrement l'idée de « test reposant sur une statistique donnée » en celle de « famille de tests imbriqués » ; et la seconde, très générale, où la notion de « statistique de p-valeur » est définie de façon complètement directe, qui est ma préférée ☃ Voici ces deux définitions :

Définition (EC') (p-valeur associée à une famille de tests imbriqués)**.** Soit un modèle statistique avec les notations standards, et une division de l'espace du paramètre caché en des parties nulle et alternatives Θ_0 et Θ_1. On dit qu'un ensemble $(Test_\alpha)_{\alpha\in]0,1[}$ est une *famille de tests imbriqués* pour l'hypothèse nulle $\{\theta \in \Theta_0\}$ lorsque :

— D'une part, pour tout $\alpha \in]0,1[$, $Test$ est un test de notre hypothèse nulle au niveau de risque [‡‡] α ;

— D'autre part, pour tous $\alpha < \alpha'$, $Test_\alpha \implies Test_{\alpha'}$: autrement dit, si une observation x donne lieu à une réponse positive pour le test $Test_\alpha$, elle devra aussi donner une réponse positive pour tous les $Test_{\alpha'}$ avec $\alpha' > \alpha$.

Dans ces conditions, la p-valeur associée à l'observation $x_\checkmark$ pour notre famille de tests est définie comme :

$$p_\checkmark := \inf\{\alpha \in]0,1[\mid test_\alpha(x_\checkmark)\} ; \tag{ED'}$$

autrement dit, c'est la "première" valeur de α pour laquelle, lorsqu'on relève progressivement le niveau de risque du test, on se met à avoir un test positif pour l'observation $x_\checkmark$. ♡

Définition (EE') (Notion générale de statistique de p-valeur)**.** De manière générale, un modèle statistique et une hypothèse nulle étant donnés (avec les notations standard), une *statistique de p-valeur* est une statistique P à valeurs dans $[0,1]$ telle que

$$\forall \alpha \in [0,1] \quad \forall \theta \in \Theta_0 \quad \mathbb{P}_\theta(P < \alpha) \leqslant \alpha \text{ [*]}. \tag{EF'}$$

♡

Proposition (EG') (Test par p-valeur et test booléen, version sans statistique de test)**.** *Une statistique de p-valeur P étant donnée pour un test de l'hypothèse nulle $\mathscr{H}_0$, pour $\alpha \in]0,1[$ un niveau de risque, le test booléen $\{P \leqslant \alpha\}$ est un test de niveau α de l'hypothèse $\mathscr{H}_0$.* ♢

13.3 Interprétation d'une p-valeur

!! **Principe** (EH')**.** *La p-valeur étant la probabilité d'obtenir un résultat au moins aussi suspect dans le cas où l'hypothèse nulle est vraie, plus la p-valeur est petite et plus on sera enclin à rejeter l'hypothèse nulle. Voir aussi le paragraphe « Interprétation d'une p-valeur » un peu plus loin.* ♢

Remarque (EI')*.* Lorsqu'on arrondit une p-valeur, il est préférable de le faire toujours par *excès*, afin d'être conservatif concernant le niveau du test. (Confer remarque (BJ')). ♧

! [‡‡]. À noter que, exceptionnellement, on est amené ici à considérer des niveau de risque plus grands que 1/2.

[*]. Dans le jargon technique, deux lois Q_0 et Q_1 sur $\mathbb{R}$ étant données, on dit que Q_1 *domine stochastiquement* Q_0 lorsque la fonction de répartition de Q_0 est partout supérieure où égale à la fonction de répartition de Q_1 (ce qui signifie, intuitivement, que Q_1 prend ses valeurs "plus à droite" que Q_0). Avec ce vocabulaire, la condition de la formule peut se reformuler en disant que, pour tout θ participant de l'hypothèse nulle, la loi de la P sous $\mathbb{P}_\theta$ doit être stochastiquement dominée par $\mathrm{Unif}^{\mathrm{me}}(0,1)$.

Définition (EJ'). En analyse statistique, lorsqu'on souhaite exprimer l'idée qu'une observation semble suffisamment incompatible avec une hypothèse nulle (autrement dit, que la p-valeur du test soit suffisamment petite) pour qu'on puisse conclure au rejet de cette hypothèse nulle, on parle de résultat « significatif » : par exemple, une phrase comme « les résultats de la promotion cobaye sont significativement meilleurs que ce à quoi on aurait pu s'attendre avec l'ancienne méthode » signifie, en termes plus techniques : « le test de l'hypothèse alternative selon laquelle la nouvelle méthode pédagogique donne des résultats strictement meilleurs que l'ancienne, lorsqu'on l'applique aux données de la promotion cobaye, est positif pour un niveau de risque très faible (ou, de façon équivalente, a une très petite p-valeur) ».

Attention ! Dans le vocabulaire courant, « significativement meilleur » n'exprime en général pas le *niveau de confiance* qu'on peut avoir sur le fait qu'on est effectivement meilleur, mais *à quel point* on est meilleur (cela correspond au concept de « taille d'effet », cf. § 17.5) : l'usage statistique de ce mot est donc bien particulier, et on doit être très prudent avant d'utiliser le mot « significatif » devant des non-statisticiens ! (et, à l'inverse, on veillera à ce que les statisticiens, eux, pourront être induits en erreur si on a voulu employer l'adjectif « significatif » au sens du langage courant...). ♡

Remarque (EK'). La p-valeur étant une grandeur quantitative, son interprétation qualitative en termes de « résultat plus ou moins significatif » ne saurait être définie de façon canonique. Cependant, l'usage a consacré certains seuils et appellations qu'il est préférable d'avoir vu au moins une fois afin de ne pas trop s'en éloigner : voir la table 13.1. Vous n'êtes pas obligés de les suivre à la lettre (de sorte que je ne vous demande pas de les apprendre par cœur), mais cela vous donnera au moins une idée de ce qui peut être considéré comme un usage correct : par exemple, dire qu'une p-valeur à 6,2 % est « significative » ne peut guère être considéré comme fautif ; par contre, dire qu'elle est « très significative » est assez clairement inacceptable... ! ♧

Remarque (EL'). Vous serez peut-être surpris, dans la table 13.1, de voir qu'on considère qu'il n'y a *aucune* significativité pour une p-valeur supérieure à 20 %, alors qu'une p-valeur peut potentiellement prendre n'importe quelle valeur entre 0 et 1 : dès lors, ne semblerait-il pas logique de commencer à trouver un peu de significativité dès lors que la p-valeur est inférieure à 1/2 ?...

Eh bien non, ce serait là une conclusion parfaitement inepte ! Parler de significativité pour une p-valeur supérieure à 20 %, ce serait considérer que, lorsqu'on voit un évènement qui avait 21 % de chances de se produire sous l'hypothèse nulle, ce serait suffisant pour dire « ouh là, 21 % de chances, c'est vraiment improbable : le fait que l'évènement se soit produit est donc un indice substantiel que l'hypothèse nulle est en fait fausse » ! Pour prendre un exemple concret, imaginons que vous vous demandiez si un dé donné est équilibré ou s'il a tendance à privilégier les grandes valeurs ; et que pour trancher, vous lanciez le dé une fois, lancer qui vous donne un '6' (ce qui correspond à une p-valeur de 1/6 pour l'hypothèse nulle que le dé n'est pas truqué, si on considère que comme statistique de test le résultat du dé, avec suspicion des grandes valeurs). Vous semblerait-il raisonnable d'y voir là un début de preuve que le dé est truqué ? C'est tout au plus un indice *très, très faible* ! Et c'est bien ce que dit notre tableau, qui qualifie une p-valeur de 17 % de « très faiblement significative ». ♧

p-valeur	Interprétation
$20\ \% \leqslant p_{\checkmark}$	Non significatif
$10\ \% \leqslant p_{\checkmark} < 20\ \%$	Très faiblement significatif
$5\ \% \leqslant p_{\checkmark} < 10\ \%$	Faiblement significatif
$2\ \% \leqslant p_{\checkmark} < 5\ \%$	Significatif
$1\ \% \leqslant p_{\checkmark} < 2\ \%$	Assez fortement significatif
$1\ ‰ \leqslant p_{\checkmark} < 1\ \%$	Fortement significatif
$10^{-6} \leqslant p_{\checkmark} < 1\ ‰$	Très fortement significatif
$p_{\checkmark} < 10^{-6}$	Preuve considérable comme “définitive”

TABLE 13.1 – Interprétations conventionnelles de la p-valeur en termes de significativité

13.4 Un exemple de test par p-valeur

Dans cette section, nous allons développer un exemple montrant comment un test par p-valeur se construit et se calcule en pratique. Nous allons nous placer, à nouveau, dans le cas du pédagogue ; mais contrairement au chapitre précédent, nous prenons cette fois-ci $\Theta = \mathbb{R} \times \mathbb{R}_+^*$ tout entier : autrement dit, on envisage ici que la variance des notes des élèves puisse éventuellement changer suite à l'introduction de la nouvelle méthode. L'hypothèse nulle que l'on souhaite tester est l'affirmation de l'enseignant que la nouvelle méthode va faire progresser ses élèves d'*au moins* un point en moyenne, autrement dit c'est l'hypothèse $\{\mu \geqslant \mu_{\min}\}$, avec $\mu_{\min} := 76$.

Pour une telle situation, en consultant la littérature [†] [‡], on trouve que la statistique de test appropriée est la suivante, avec test à gauche (autrement dit, plus on ira vers $-\infty$ et plus l'hypothèse nulle sera suspecte) :

$$T := \frac{M - \mu_{\min}}{S}, \tag{EM'}$$

où nous rappelons que nous avions convenu de noter resp. par M et S les moyenne et écart-type empiriques des observations :

$$M := n^{-1} \sum_{i=0}^{n-1} X_i \, ; \qquad S := \left(n^{-1} \sum_{i=0}^{n-1} (X_i - M)^2 \right)^{1/2}.$$

Ayant pris connaissance de la réalisation $t_{\checkmark}$ de notre statistique de test, nous souhaitons déterminer la p-valeur associée à celle-ci. Puisque notre critère de suspicion est à gauche, et au vu de l'hypothèse nulle que nous souhaitons tester, ladite p-valeur est définie comme

$$\sup_{\substack{\mu \geqslant \mu_{\min} \\ \sigma \in \mathbb{R}_+^*}} \mathbb{P}\left(\frac{M - \mu_{\min}}{S} \leqslant t_{\checkmark} \;\middle|\; \theta = (\mu, \sigma) \right). \tag{EN'}$$

Pour calculer cela, il nous faudrait connaitre, pour $(\mu, \sigma) \in [\mu_{\min}[\times \mathbb{R}_+^*$, la *loi* de la variable aléatoire $(M - \mu_{\min}) / S$ sous le contexte $\mathbb{P}_{\mu,\sigma}$. De manière générale, cette loi est susceptible dépendre de la valeur de (μ, σ) au sein de l'hypothèse nulle ; néanmoins, nous espérons être dans un cas où un des principes (CK') ou (CN')

[†]. Concernant la façon d'exploiter la littérature, confer la § 13.5 ci-après.

[‡]. En l'occurrence, il s'avère d'ailleurs que cette statistique des test est aussi celle suggérée par le rapport des vraisemblances ! ☺

s'applique. En l'occurrence, vu que notre hypothèse nulle fait intervenir une inégalité, nous serions le cas échéant dans le cadre du principe (CN') ; et le "bord" de notre hypothèse nulle correspondrait aux cas où $\mu = \mu_{\min}$.

Est-il donc vrai que, sur ce bord, la loi de $(M - \mu_{\min})/S$ ne dépend pas de σ ? De fait, oui ! C'est même assez facile à démontrer : en effet, lorsque μ vaut $\mu_{\min}$, alors sous le contexte $\mathbb{P}_{\mu_{\min},\sigma}$, en introduisant les variables "réduites" $\check{X}_i \coloneqq (X_i - \mu_{\min})/\sigma$, on a que

$$\frac{M - \mu_{\min}}{S} \underset{\text{déf}}{=} \frac{n^{-1}\sum_{i=1}^n (\mu_{\min} + \sigma\check{X}_i) - \mu_{\min}}{\mathrm{var}_{\mathrm{emp}}^{1/2}\big((\mu_{\min} + \sigma\check{X}_i)_i\big)} = \frac{n^{-1}\sigma\sum_{i=1}^n \check{X}_i}{\sigma\,\mathrm{var}_{\mathrm{emp}}^{1/2}((\check{X}_i)_i)} = \frac{\sum_{i=1}^n \check{X}_i}{n\,\mathrm{var}_{\mathrm{emp}}^{1/2}((\check{X}_i)_i)}, \tag{EO'}$$

dont la loi ne dépend donc que de la loi jointe des $\check{X}_i$; or, peu importe la valeur de σ, les variables réduites sont, sous $\mathbb{P}_{\mu_{\min},\sigma}(\bullet)$, indépendantes [en tant que fonctions de v.a. indépendantes] et chacune de loi normale standard [d'après les propriétés d'invariance des lois normales], de sorte que leur loi jointe est toujours la même.

En fait, dans les situations pratiques, le fait que la statistique de test ait toujours la même loi dès lors que μ vaut $\mu_{\min}$ nous serait donné par la littérature ☺ D'autre part, la littérature nous dirait aussi — ce qui est très important ! — *quelle* est cette loi que suit la statistique de test. En l'occurrence, il s'agit du résultat suivant :

Théorème (EP'). *Dans le modèle du pédagogue*[§]*, lorsque le paramètre caché est de la forme* $(\mu_{\min}, \sigma)$*, la loi de la statistique de test introduite ci-dessus vaut*

$$n^{-1/2} \times \mathrm{T}_{\mathrm{St}}(n-1), \tag{EQ'}$$

où "$\mathrm{T}_{\mathrm{St}}(n-1)$*" est ce qu'on appelle la* loi t de Student à $(n-1)$ degrés de liberté*, dont les tables sont implémentées dans tous les bons logiciels utilisables pour traiter des données statistiques.* ◇

Grâce à ce théorème, nous avons que, lorsque (μ, σ) est au bord de notre hypothèse nulle, la probabilité que la statistique de test tombe en-dessous de $t_{\checkmark}$ est toujours la même, et vaut

$$\mathbb{P}\big(n^{-1/2} \times \mathrm{T}_{\mathrm{St}}(n-1) \leqslant t_{\checkmark}\big). \tag{ER'}$$

Cependant, dans le supremum (EN') définissant notre p-valeur, il faut considérer non seulement les cas où $\mu = \mu_{\min}$, mais *aussi* ceux où $\mu > \mu_{\min}$... ! Pour ces cas-là, en vertu du principe (CN'), notre espoir est que la statistique de test ait tendance à prendre des valeurs *plus à droite* que dans le cas du "bord", de sorte que sa probabilité d'être $\leqslant t_{\checkmark}$ sera plus petit que pour les cas du bord, et n'interviendra donc pas dans la détermination du supremum. Cet espoir semble parfaitement raisonnable, puisqu'on comprendre bien que T, qui a tendance à être d'autant plus grande que les X_i sont élevés (à cause du terme en M au numérateur), devrait donc tendre à prendre de plus grandes valeurs sous $\mathbb{P}_{\mu,\sigma}(\bullet)$ pour $\mu > \mu_{\min}$ que sous $\mathbb{P}_{\mu_{\min},\sigma}$.

Est-il possible de rendre rigoureux cet argument intuitif ? Oui ; par exemple en considérant à nouveau les variables aléatoires "réduites" de façon à suivre la loi Normale$(0,1)$. En l'occurrence, notre paramètre (μ, σ) étant fixé avec $\mu > \mu_{\min}$,

[§]. Le « modèle du pédagogue » est bien connu de la littérature — quoique pas sous ce nom, bien entendu ! ☿ — : il s'agit en effet du modèle d'échantillonnage associé à des lois normales de paramètres inconnus, qui apparait naturellement dans nombre de situations pratiques ☊

nous définissons les $\check{X}_i$ comme $(X_i - \mu) / \sigma$: on a alors, comme précdémment, que les $\check{X}_i$ i.i.d. normaux standards ; tandis que, en ce qui concerne la façon de ré-écrire la statistique de test à partir des variables réduites, on voit apparaitre un terme supplémentaire par rapport à (EO')

$$T = \frac{\sum_{i=1}^{n} \check{X}_i}{n \operatorname{var}_{\mathrm{emp}}^{1/2}((\check{X}_i)_i)} + \frac{\mu - \mu_{\min}}{\sigma \operatorname{var}_{\mathrm{emp}}^{1/2}((\check{X}_i)_i)}. \tag{ES'}$$

Or, la variable aléatoire que constitue ce terme supplémentaire — et que nous noterons 'Ξ' dans les lignes qui suivent — ne prend que des valeurs positives ! On a donc donné un sens rigoureux à l'idée que la loi de T, sous $\mathbb{P}_{\mu,\sigma}(\bullet)$, "prend des valeurs plus grandes que" $\left(\sum_{i=1}^{n} \check{X}_i\right) / n \operatorname{var}_{\mathrm{emp}}^{1/2}((\check{X}_i)_i)$; et en particulier, on est maintenant en mesure de démontrer qu'elle a moins de risque de tomber en-dessous de $t_{✓}$, par le raisonnement suivant :

$$\begin{aligned}\mathbb{P}_{\mu,\sigma}(T \leqslant t_{✓}) &= \mathbb{P}_{\mu,\sigma}\left(\frac{\sum_{i=1}^{n} \check{X}_i}{n \operatorname{var}_{\mathrm{emp}}^{1/2}((\check{X}_i)_i)} + \leqslant \underbrace{t_{✓} - \Xi}_{\leqslant t_{✓} \text{ p.-s.}}\right) \\ &\leqslant \mathbb{P}_{\mu,\sigma}\left(\underbrace{\frac{\sum_{i=1}^{n} \check{X}_i}{n \operatorname{var}_{\mathrm{emp}}^{1/2}((\check{X}_i)_i)}}_{\text{de loi } n^{-1/2} \times \mathrm{T}_{\mathrm{St}}(n-1)} \leqslant t_{✓}\right) = \mathbb{P}\left(n^{-1/2} \times \mathrm{T}_{\mathrm{St}}(n-1) \leqslant t_{✓}\right). \end{aligned} \tag{ET'}$$

Par conséquent, dans le supremum défini par (EN'), les cas correspondant au bord de l'hypothèse nulle valent tous $\texttt{répartT}_{\mathtt{St}}(n-1; n^{1/2} t_{✓} +)$, tandis que les autres cas ne sont jamais plus grand que cette valeur : *in fine*, on a donc trouvé que (la réalisation de) notre p-valeur vaut

$$\texttt{répartT}_{\mathtt{St}}(n-1; n^{1/2} t_{✓} +) \quad \ddot{\smile} \tag{EU'}$$

Remarque (EV'). Dans le calcul ci-dessus, nous avons *démontré* rigoureusement que seuls les cas au bord de l'hypothèse nulle nécessitaient d'être pris en compte pour le calcul du supremum. Néanmoins, comme nous l'avons signalé, c'était en réalité très intuitif... Dans un contexte d'exercice ou d'examen, vous auriez vous contenter [¶], dans un tel cas, d'une justification intuitive, du type : « Il est évident, au vu de sa forme, que la statistique de test prend (à σ fixé) des valeurs d'autant plus grandes que μ est grand : en effet, quand μ augmente, le numérateur tend à augmenter, tandis que les valeurs prises par le dénominateur restent les mêmes. Dès lors, pour déterminer le supremum définissante la p-valeur, il suffit de considérer les cas où μ est minimal (à σ fixé), autrement dit, ceux où $\mu = \mu_{\min}$ ». ♧

Application numérique. La formule pour la p-valeur que nous avons donnée s'implémente ainsi sous R :

```
> mumin = 76
> notes = c(91.7, 69.6, 81.5, 45.0, 69.7, 77.3, 42.1, 70.7, 66.9, 93.6,
+ 75.7, 93.7, 59.8, 75.2, 49.8, 88.2, 63.7, 90.0, 54.2, 57.2, 72.8, 58.3)
> (n = length(notes))
[1] 22
> (m = mean(notes)) # Moyenne empirique
```

[¶]. À moins, bien sûr, qu'il ne fût explicitement demandé une démonstration formelle ! ☹

```
[1] 70.30455
> (s = sqrt(mean(notes ^ 2) - m ^ 2)) # Écart-type empirique
[1] 15.20698
> pt(sqrt(n) * (m - mumin) / s, n - 1)
[1] 0.04677142
```

Une telle p-valeur de 4,7 % est généralement considérée comme significative (quoique pas *très* significative non plus) : c'est un indice fort pour rejeter l'hypothèse nulle, même si gros coup de manchance reste vaguement envisageable sans faire preuve de trop de mauvaise foi. Mais si l'enseignant est raisonnable, il devrait reconnaitre qu'il s'est trompé : sa méthode, manifestement, échoue à faire progresser les élèves autant qu'il ne l'aurait espérer ! (voire, les fait carrément régresser).

Il peut être intéressant de regarder ce qu'aurait donné le résultat de notre test pour d'autres valeurs de l'observation. Par exemple, si toutes les notes avaient été deux points plus élevées (avec donc une moyenne empirique de 72,30, l'écart-type empirique restant le même), la p-valeur serait montée à 13,4 %, ce qui pour le coup n'est que très peu significatif : cela montre donc qu'il est tout à fait possible qu'on ait une moyenne de promotion plus de deux points et demi *en dessous* du seuil de référence tout en jugeant largement envisageable l'hypothèse que, *en réalité*, la méthode passe, quand on moyenne l'effet du hasard, *progresser* les élèves d'au moins un point !

Imaginons aussi le cas où les notes des élèves auraient eu la même moyenne que dans la réalité (soit 70,30), mais avec une dispersion *plus grande* : mettons, par exemple, qu'on aurait dans ce scénario $s_{\checkmark} = 21$. Dans ce cas aussi, la p-valeur serait sensiblement plus élevée (à 10,9 %) : des notes plus hétérogènes pour la promotion tendant en effet à indiquer que la nouvelle méthode présente manifestement une valeur assez élevée de σ, ce qui crée un fort "bruit" sur les notes faisant de la moyenne de la première promotion une moins bonne approximation de la véritable valeur $\mu_{\checkmark}$... À l'inverse, une dispersion plus petite, disons $s_{\checkmark} = 11$, (toujours avec la même moyenne) ferait plonger la p-valeur à 1,3 % : de tels résultats permettraient alors de réfuter l'hypothèse de l'enseignant avec une confiance très élevée !

13.5 Utilisation pratique d'un test par l'ingénieur

Ci-dessus, le traitement des statistiques de test a été mené assez en détail. Nous avons notamment vu comment utiliser les rapports de vraisemblances pour déterminer des statistiques de test pertinentes, comment on pouvait introduire des variables réduites pour ramener la loi de la statistique de test (sous l'hypothèse nulle) à un objet mathématique bien spécifié, et comment démontrer, si nécessaire, que seuls certains cas particuliers de l'hypothèse nulle sont à prendre en compte. Voir comment tous ces éléments marchent est en effet important, me semble-t-il, pour bien comprendre la démarche du test.

Dans la pratique néanmoins, l'ingénieur généraliste (non statisticien) n'aura pas à ré-inventer la roue ! Concrètement, que fait-il face à un problème donné ? Eh bien, il ouvre son ouvrage de statistique favori pour voir s'il y a un test bien adapté à cette situation ! On tombe alors sur une description dans le style suivant :

Tester la moyenne d'une population normale de variance inconnue

Lorsque $X_1, \dots, X_n$ sont des v.a. normales i.i.d. Normale($\mu_✓, \sigma_✓$) avec $\mu_✓$ et $\sigma_✓$ inconnus, pour tester l'hypothèse $\{\mu_✓ = \mu_{\text{réf}}\}$ (ou $\{\mu_✓ \leqslant \mu_{\text{réf}}\}$, ou $\{\mu_✓ \geqslant \mu_{\text{réf}}\}$), il faut utiliser la statistique de test

$$T := n^{1/2} \frac{\text{moy}(X_1, \dots, X_n) - \mu_{\text{réf}}}{\text{var}_{\text{emp}}^{1/2}(X_1, \dots, X_n)}, \qquad (\texttt{EW'})$$

sachant que, lorsque $\mu_✓ = \mu_{\text{réf}}$, cette variable aléatoire suit la loi $\text{T}_{\text{St}}(n-1)$.

Comment faut-il utiliser une description ? C'est ce que je détaille ci-dessous, la procédure étant ici illustrée pour le cas « tester la moyenne d'une population normale de variance inconnue », mais étant parfaitement générale :

! **Procédure (EX').** Déjà, le texte de l'ouvrage de statistique nous fournit déjà la statistique de test et la loi de référence à laquelle il faut se ramener !

Il vous appartient néanmoins de vous répondre à la question « dois-je faire un test à gauche, à droite ou bilatéral ? ». Sur ce point, on pourrait bien sûr détailler le raisonnement mathématique ; mais le simple bon sens montre que, pour tester l'hypothèse $\{\mu \leqslant \mu_{\text{réf}}\}$, c'est un test à droite qu'il faut faire : en effet, quand l'hypothèse nulle est (largement) fausse, cela signifie que les X_i tendent à être (beaucoup) plus grands que $\mu_{\text{réf}}$, mais alors leur moyenne aura aussi tendance à être (beaucoup) plus grande, et donc la statistique de test aura tendance à être (très) positive : ce sont donc les plus petites grandes (au sens de « plus positives ») de la statistique de test qui seront les plus suspectes, alors qu'à l'inverse, de petites valeurs semblent indiquer que $\mu_✓$ est plus *petite* que $\mu_{\text{réf}}$, et ne doivent donc pas conduire au rejet de l'hypothèse nulle ! De même, si on avait eu à tester l'hypothèse nulle $\{\mu \geqslant \mu_{\text{réf}}\}$, il aurait convenu de faire un test à gauche, et pour tester l'hypothèse nulle $\{\mu = \mu_{\text{réf}}\}$, un test bilatéral.

Le point suivant consiste à voir comment seuls certains cas méritant d'être pris en compte. Ainsi, si l'hypothèse nulle que nous souhaitons tester est $\{\mu \leqslant \mu_{\text{réf}}\}$ (on aurait un argument similaire pour $\{\mu \geqslant \mu_{\text{réf}}\}$), vous allez argüer (ou démontrer si vous avez un doute) que, pour $\theta \in \Theta_0$, le cas où la valeur de $\mathbb{P}(T \geqslant t_✓ \mid \theta = \theta)$ est le plus élevé est forcément lorsque $\mu = \mu_{\text{réf}}$, de sorte que vous pouvez vous limiter à ce cas.

Mais alors, dans le cas en question, T suit la loi $\text{T}_{\text{St}}(n-1)$, dont vous disposez de la fonction de répartition, ce qui vous permet de calculer la p-valeur : celle-ci correspond en effet à $\mathbb{P}(\text{T}_{\text{St}}(n-1) \geqslant t_✓)$, autrement dit à $1 - \texttt{répartT}_{\texttt{St}}(n-1; t_✓)$ (ici peu importe si on considère la fonction de répartition en $t_✓-$ ou en $t_✓+$, car les lois de Student sont diffuses), où la fonction $\texttt{répartT}_{\texttt{St}}$ sera implémentée dans votre logiciel de statistique préféré et $t_✓$ sera calculé à partir de la formule pour T en prenant les valeurs des observations effectives.

Dans le cas d'un test de l'hypothèse nulle $\{\mu \geqslant \mu_{\text{réf}}\}$, où le critère de suspicion est à gauche, la formule pour la p-valeur aurait de même été donnée par $\texttt{répartT}_{\texttt{St}}(n-1; t_✓)$; et dans le cas d'un test de l'hypothèse nulle $\{\mu = \mu_{\text{réf}}\}$, où le critère de suspicion est bilatéral, la formule aurait été $2 \times \big(\big(1 - \texttt{répartT}_{\texttt{St}}(n-1; t_✓)\big) \wedge \texttt{répartT}_{\texttt{St}}(n-1; t_✓)\big)$. ♡

Chapitre 14

Intervalles de confiance et de prédiction fréquentistes

14.1 Intervalles de confiance

Définition et interprétation

Les tests d'hypothèses nous permettent d'obtenir de l'information sur une quantité d'intérêt $\varphi(\theta) =: \varphi$, dans la mesure où ils peuvent nous permettre de rejeter avec une certitude assez grande des hypothèses nulles sur la valeur de φ qui ne seraient pas réalistes. Cependant, comme on sait pas à l'avance quelles valeurs de φ sont réalistes ou non, cela implique que, avant d'avoir une idée honnête de ce que peut valoir $\varphi(\theta)$, il va falloir tester un grand nombre d'hypothèses nulles distinctes, ce qui n'est guère pratique... Nous, ce qu'on voudrait plutôt, c'est tout simplement connaitre directement quelle est la plage des valeurs plausibles pour $\varphi(\theta)$!

C'est à ce problème que répondent les intervalles de confiance. Il s'agit en fait, du point de vue mathématique, de mener *simultanément* des tests pour toutes les valeurs envisageables de $\varphi(\theta)$; mais nous allons voir que les résultats se présentent sous une apparence fort différente.

Définition (EY') (Intervalle de confiance)**.** Soit $\Theta \ni \theta \mapsto \text{Loi}(X \mid \theta = \theta))$ un modèle statistique, $\varphi(\theta)$ notre quantité d'intérêt ; et soit $I(X)$ une statistique à valeurs dans les intervalles de $\mathbb{R}$ [*]. On dit que $I(X)$ est un *intervalle de confiance au niveau* α pour $\varphi(\theta)$ lorsque, pour tout $\theta \in \Theta$, $\mathbb{P}(I(X) \not\ni \varphi(\theta) \mid \theta = \theta) \leqslant \alpha$; autrement dit, pour tout $\theta \in \Theta$, « $I(X) \not\ni \varphi(\theta)$ » est un test de niveau α de l'hypothèse nulle $\{\theta = \theta\}$. ♡ !

Définition (EZ') (Intervalles de confiance asymptotiques)**.** Dans le cas d'un modèle statistique ayant un paramètre du modèle pour lequel on considère une asymptotique, on définit de même le concepts d'intervalle de confiance *asymptotiquement de niveau* α lorsque pour tout $\theta \in \Theta$, $\overline{\lim}_{n\to\infty} \mathbb{P}_\theta(I(X) \not\ni \varphi(\theta)) \leqslant \alpha$. ♡ !

Remarque (FA')*.* En fait, rien dans les définitions ci-dessus n'oblige conceptuellement $I(X)$ à être un intervalle : cela pourrait être n'importe quelle partie de

[*]. N'oubliez pas qu'une « statistique » peut être à valeurs dans *n'importe quel type* d'objets mathématiques : ce n'est pas forcément un nombre réel ! (Nous avions d'ailleurs vu, dans le chapitre 11, le cas des lois empiriques, qui sont des statistique à valeurs « distribution de probabilité »...). En l'occurrence, $I(X)$ est à valeurs « intervalles » : c'est donc un *intervalle* défini à partir de la variable aléatoire X.

$\mathbb{R}$ (et il faudrait alors plutôt parler de « zone de confiance »). D'ailleurs, de manière générale, $\boldsymbol{\varphi}(\boldsymbol{\theta})$ pourrait être à valeurs dans n'importe quel ensemble $\mathscr{G}$, et alors $\boldsymbol{I}(\boldsymbol{X})$ serait une partie de $\mathscr{G}$. Cependant, essentiellement toutes les situations pratiques concernent des intervalles de $\mathbb{R}$; d'autant que, comme nous allons le voir ci-dessous, on dispose alors d'une procédure (partiellement) standardisée pour construire des intervalles de confiance. C'est pourquoi nous ne nous intéresserons qu'à ce cas ici. ♣

Remarque (FB'). Contrairement à l'intervalle de confiance bayésien (alias « intervalle de croyance »), qui correspond à un intervalle de fluctuation pour la loi à postériori et peut donc être défini "x par x" (puisque la loi à postériori a bien un sens x par x), un intervalle de confiance fréquentiste ne peut être défini que comme une *application* de $\mathscr{X}$ dans l'ensemble *Ivl* des intervalles de $\mathbb{R}$. ♣

> *Remarque* (FC'). En fait, on peut montrer que sous des hypothèses très peu restrictives, pour tout $x \in \mathscr{X}$, on peut trouver un intervalle de confiance $ICF : \mathscr{X} \to Ivl$ tel quel $ICF(x) = \varnothing$! Alors que dans le cadre bayésien (une fois, s'entend, qu'on a choisi notre priore), cela n'est pas possible, puisque la réalisation $ICB(x_{\checkmark})$ d'un intervalle de croyance $ICB : \mathscr{X} \to Ivl$ doit nécessairement être un intervalle de fluctuation pour la loi à postériori de $\varphi(\boldsymbol{\theta})$.
>
> Au passage, le fait qu'on puisse toujours trouver un intervalle de croyance fréquentiste ICF pour lequel $ICF(x) = \varnothing$ souligne un point particulièrement important en statistique fréquentiste (dont nous reparlerons dans la chapitre 17) : à savoir, qu'il faut décider de la procédure d'analyse statistique *avant* de connaitre (ou du moins de regarder) les données : car en "bricolant" notre technique d'analyse au vu de nos données, on pourrait arriver à "conclure" à n'importe quel résultat qui nous arrange, aussi absurde soit-il... ! ♣

! *Remarque* (FD'). À cause de ce que nous venons de dire, il ne serait pas correct de conclure une analyse statistique par intervalle de confiance fréquentiste par une phrase comme « il y a une probabilité $(1-\alpha)$ que $\boldsymbol{\varphi} \in ICF(x_{\checkmark})$ » : en fait, cette formulation correspondrait à l'interprétation d'un intervalle de confiance *bayésien*...! Tout ce que nous pouvons dire lors du calcul d'un intervalle de confiance fréquentiste, c'est qu'*avant* de lancer l'analyse statistique, donc *avant* de tirer $x_{\checkmark}$, on savait qu'il y aurait une probabilité $(1-\alpha)$ que notre intervalle de confiance aille contenir la véritable valeur de $\varphi_{\checkmark}$. Mais peut-être qu'il y a certains x pour lesquels notre intervalle de confiance ne contiendra pas aussi souvent (voire jamais, dans le cas d'un intervalle vide !) la véritable valeur de $\boldsymbol{\varphi}$, et d'autres pour lesquels au contraire $ICF(x)$ contiendra encore plus souvent (voire toujours, dans le cas d'un intervalle total !) la véritable valeur... Seule l'analyse bayésienne pourrait nous en dire plus long sur ce que vaut la probabilité de contenir la vraie valeur *une fois qu'on a observé* $x_{\checkmark}$!

Pour cette raison, je vous demande de toujours suivre une des deux formulations suivantes (avec une préférence pour la première, plus rigoureuse) :

- « Notre procédure conduit à un intervalle de confiance au risque α pour $\boldsymbol{\varphi}$ dont la réalisation (pour nos données effectives) est $ICF(x_{\checkmark})$ » ;
- « Au risque α, on conclut que $\varphi_{\checkmark} \in ICF(x_{\checkmark})$ ». (La locution « au risque α » ne veut à vrai dire pas dire grand-chose ; mais elle a le mérite d'être simple et de distinguer la phrase de conclusion pour un intervalle de confiance fréquentiste de celle pour un intervalle de croyance ! ☺).

♣

Cette définition étant posée, il y a deux méthodes principales pour fabriquer des intervalles de confiance, que nous allons maintenant présenter.

Première méthode : Test avec paramètre

L'idée de cette méthode est de calculer le résultat du test de l'hypothèse nulle $\{\boldsymbol{\varphi} = \varphi\}$ pour chaque valeur possible de $\boldsymbol{\varphi}$, et à renvoyer comme intervalle de confiance l'ensemble des valeurs compatibles. Cela donne la procédure suivante, dans laquelle l'adjectif « asymptotique » correspond au cas d'un intervalle de confiance asymptotique :

Procédure (FE'). !

1° Pour tout $\varphi \in \mathbb{R}$, on considère un test (asymptotique) de niveau α $Test_\varphi$ de l'hypothèse $\{\boldsymbol{\varphi} = \varphi\}$, ce qui nous donne donc une famille de tests $(Test_\varphi)_{\varphi\in\mathbb{R}}$: pour chacun de ces tests, on regarde quelle est la zone d'acceptation pour l'observation. On s'arrangera évidemment pour que tous ces tests s'obtiennent par le même calcul, où seule la valeur de φ changera. Il peut également être souhaitable que tous ces tests utilisent la même statistique de test $\boldsymbol{T}$ (auquel cas la loi de la statistique de test, et donc la région d'acceptation, dépendra bien entendu de la valeur de φ).

2° Pour une réalisation $x_\checkmark$ de l'observation (ou pour une réalisation $t_\checkmark$ de la statistique de test), la réalisation de l'intervalle de confiance au niveau α pour $\varphi_\checkmark$ sera alors donnée par l'ensemble des $\varphi \in \mathbb{R}$ tels que le test $Test_\varphi$ soit négatif au vu de cette observation (autrement dit, tel que $x_\checkmark$ (ou $t_\checkmark$) soit dans la région d'acceptation de $Test_\varphi$). ♡

Remarque (FF'). Une variante de cette méthode consiste à tester toutes les hypothèses élémentaires $\{\boldsymbol{\theta} = \theta\}$ pour $\theta \in \boldsymbol{\Theta}$, à regarder la "zone de confiance" correspondante dans $\boldsymbol{\Theta}$, puis à prendre l'image par $\boldsymbol{\varphi}$ de cette zone de confiance. Par exemple, si on note $\boldsymbol{T}^{(\theta)}$ la statistique utilisée pour tester l'hypothèse $\{\boldsymbol{\theta} = \theta\}$, et que tous nos tests reposent sur un critère de suspicion bilatéral, on pose

$$\boldsymbol{\Theta}_{\mathrm{conf}\checkmark} \coloneqq \{\theta \in \boldsymbol{\Theta} \mid t^{(\theta)}_\checkmark \text{ est dans l'intervalle de fluctuation au risque } \alpha \text{ de } \mathrm{Loi}_\theta(\boldsymbol{T}^{(\theta)})\}, \tag{FG'}$$

qui est l'ensemble des valeurs de $\boldsymbol{\theta}$ jugées crédibles (au risque α) au vu de notre observation ; et la réalisation de notre intervalle de confiance est alors donnée par $\boldsymbol{\varphi}(\boldsymbol{\Theta}_{\mathrm{conf}\checkmark})$. ♧

Je n'ai pas encore écrit d'exemple pour la procédure (FE') ; mais la mise en pratique est assez directe... Cette méthode a l'avantage d'être (théoriquement) applicable à essentiellement tout modèle, surtout une fois prise en compte la remarque (FF') ; néanmoins, les calculs peuvent se révéler particulièrement lourds, voire impraticables... D'où l'intérêt de la seconde méthode que nous allons à présent exposer, qui ne s'applique que dans certains cas, mais qui est beaucoup plus rapide.

Seconde méthode : Combiner intelligemment la quantité d'intérêt et l'observation

Nous allons d'abord expliquer le raisonnement conduisant à cette seconde méthode, avant d'expliquer, dans la procédure (FH'), comment on applique cette méthode en pratique.

L'idée est de partir d'une variable aléatoire $t(\boldsymbol{\varphi}, X)$ [†] dépendant de l'observation et de la quantité d'intérêt $\boldsymbol{\varphi}$, dont la loi est la même sous tous les contextes probabilistes $\mathbb{P}_\theta(\bullet)$ [‡] (ou, dans le cas d'un test asymptotique, dont la loi sous $\mathbb{P}_\theta^{(n)}$ tend, lorsque $n \to \infty$, vers une limite indépendante de θ). (Attention : rien ne garantit que trouver une telle v.a. sera facile, ni même possible... C'est pourquoi cette méthode est moins générale que la précédente. Néanmoins ici nous supposerons disposer d'une telle variable, pertinente pour notre objectif [§]).

Ensuite, en remplaçant $\boldsymbol{\varphi}$ par φ dans l'expression de $t(\boldsymbol{\varphi}, X)$, on obtient une statistique de test pour l'hypothèse $\{\boldsymbol{\varphi} = \varphi\}$. Il faut en toute rigueur vérifier que cette statistique est intelligente, et en particulier que si $\varphi_{\checkmark}$ n'est pas réellement égal à φ, on s'attend à ce que la statistique de test en question prenne des valeurs bien différentes de ce que serait sa loi sous l'hypothèse nulle. Normalement, la situation doit être la suivante : notre statistique de test (qui est à valeurs dans $\mathbb{R}$ ou un des ses sous-ensembles) aura tendance à s'écarter dans une direction si $\varphi_{\checkmark} < \varphi$, et dans l'autre direction si $\varphi_{\checkmark} > \varphi$; de sorte que le critère de suspicion consistera à soupçonner les valeurs les plus extrêmes, les valeurs les moins suspectes étant celles les plus centrales. Par conséquent, l'intervalle d'acceptation pour la statistique de test sera toujours $[q_{\alpha/2}, q_{1-\alpha/2}]$, où $q_{\alpha/2}$ et $q_{1-\alpha/2}$ sont les quantiles aux niveaux respectifs $\alpha/2$ et $1-\alpha/2$ de la loi censée être suivie par la statistique de test sous l'hypothèse nulle, autrement dit, de la loi $\mathrm{Loi}\big(t(\varphi, X) \;\big|\; \boldsymbol{\varphi} = \varphi\big)$, laquelle loi est connue [¶] : c'est en effet la loi $\mathrm{Loi}_{\checkmark}\big(t(\boldsymbol{\varphi}, X)\big)$, vu que celle-ci est indépendante de la valeur de $\theta_{\checkmark}$.

On regarde alors, pour une observation donnée, quel est l'ensemble des valeurs φ pour lesquelles la statistique construite ci-dessus aurait donné un test négatif : cet ensemble de valeurs est la réalisation de notre intervalle de confiance pour la quantité d'intérêt $\boldsymbol{\varphi}$.

La procédure que nous venons de décrire, où nous faisons intervenir explicitement des tests bilatéraux, est en fait inutilement compliquée : nous ne l'avons donnée qu'à titre d'éclaircissement pédagogique. En réalité, le résultat de cette procédure est rigoureusement identique à celui de la procédure suivante, bien plus simple :

! **Procédure (FH').**

1° On part d'une variable aléatoire $t(\boldsymbol{\varphi}, X)$ bien choisie [‖], dépendant de l'observation et de la quantité d'intérêt $\boldsymbol{\varphi}$, dont la loi (sous $\mathbb{P}_{\checkmark}$) est connue (ou, dans le

[†]. Attention : Même si cette variable aléatoire joue un peu le même rôle qu'une statistique de test, on ne peut pas la qualifier de « statistique », puisqu'elle dépend de $\boldsymbol{\varphi}$, qui n'est pas une fonction de l'observation... !

[‡]. En pratique, plutôt que de dire qu'on a la même loi sous tous les contextes probabilistes $\mathbb{P}_\theta(\bullet)$, je dirai fréquemment « la loi de $t(\boldsymbol{\varphi}, X)$ est connue, et ce, indépendamment de la valeur $\theta_{\checkmark}$ ».

[§]. Dans l'absolu, trouver une variable aléatoire vérifiant la propriété requise pourrait toujours être fait en prenant, par exemple, $t(\boldsymbol{\varphi}, X) \equiv 0$: cependant, ce ne serait pas un choix *pertinent*, parce que la fonction $t(\varphi, x)$ ne dépendrait pas *réellement* de φ, de sorte que la connaissance de sa valeur n'apporterait aucune information sur $\boldsymbol{\varphi}$ — sans compter qu'ici la loi de $t(\boldsymbol{\varphi}, X)$ serait une loi de Dirac, pour laquelle on ne peut pas partitionner le support en des zones de confiance et de rejet non triviales...

[¶]. Comme nous sommes dans le cadre fréquentiste, il n'est *prima facie* pas clair qu'on sache déterminer la loi conditionnée par $\{\boldsymbol{\varphi} = \varphi\}$: c'est uniquement à cause des propriétés particulières de la v.a. $t(\boldsymbol{\varphi}, X)$ qu'un tel conditionnement peut être calculé ici.

[‖]. Dans les exercices et problème qui vous seront soumis, cette variable aléatoire sera donnée directement par l'énoncé, ou du moins fortement suggérée.

cas d'un test asymptotique, tend vers une loi connue lorsque $n \to \infty$), indépendamment de la valeur de $\theta_{\checkmark}$.

2° On détermine l'intervalle de fluctuation IF pour la loi de la v.a. $t(\varphi, X)$. Ici la notion d'« intervalle de fluctuation » a bien du sens, même dans le cadre fréquentiste, car nous avons supposé avoir construit notre v.a. de sorte que $\mathrm{Loi}_\theta\big(t(\varphi, X)\big)$ ne dépende pas de θ. Cet intervalle de flucutation correspond, informellement, aux valeurs "crédibles" pour $t(\varphi, X)$.

3° La réalisation de notre intervalle de confiance correspond alors aux valeurs de φ pour lesquelles $t(\varphi, x_{\checkmark}) \in IF$.

♡

Remarque (FI'). Ci-dessus, j'ai supposé que la loi de $t(\varphi, X)$ devait être *exactement* la même sous toutes les lois $\mathbb{P}_\theta$. En réalité, ce n'est pas vraiment nécessaire, et tout ce qui compte est que cette loi ne varie "pas trop", au sens où on peut trouver un intervalle IF qui, pour tout $\theta \in \Theta$, soit un intervalle de fluctuation au risque α de $\mathrm{Loi}_\theta\big(t(\varphi(\theta), X)\big)$: concrètement, il suffit pour ce faire de considérer tous les intervalles de fluctuation de toutes les $\mathrm{Loi}_\theta\big(t(\varphi(\theta), X)\big)$, puis de prendre IF comme étant leur réunion. Cependant, je ne connais pas vraiment de cas où cette construction donne lieu à quelque chose d'intéressant bien que les lois $\mathrm{Loi}_\theta\big(t(\varphi(\theta), X)\big)$ soient distinctes : c'est pourquoi j'ai seulement mis en valeur le cas où on a une identité parfaite entre toutes ces lois. ♧

Exemples de construction d'intervalles de confiance

Nous allons maintenant donner deux exemples de construction d'intervalles de confiance. Tous les deux s'appuient sur la seconde méthode présentée ci-dessus, mais dans les deux cas nous soulignons le lien avec l'approche par test avec paramètre plutôt que de recourir directement à la procédure (FH'). Le premier exemple est non asymptotique, tandis que le second est asymptotique.

Exemple (FJ'). Considérons le modèle du pédagogue (p. **??**), pour lequel nous cherchons un intervalle de confiance pour μ. Il se trouve que dans ce modèle, on peut montrer que, quelle que soit la valeur de $\theta_{\checkmark}$, la variable aléatoire

$$\frac{n^{-1}\sum_{i=1}^n X_i - \mu_{\checkmark}}{\mathrm{var}_{\mathrm{B}}^{1/2}(X_1, \dots, X_n) \,/\, n^{1/2}} \tag{FK'}$$

suit (sous $\mathbb{P}_{\checkmark}$) une loi qu'on appelle « loi de Student à $(n-1)$ degrés de liberté », notée $\mathrm{T}_{\mathrm{St}}(n-1)$. Nous ne chercherons pas à démontrer ce résultat ici : ce que nous allons faire, en revanche, est de montrer comment on peut l'utiliser pour obtenir un intervalle de confiance sur la valeur de μ. (Plus généralement, dans ce cours, le but de vos exercices sera de savoir utiliser des résultats mathématiques comme le précédent pour construire des intervalles de confiance, pas de démontrer les résultats en question).

Pour commencer, que cherchons-nous ? Réponse : à savoir quelles valeurs de μ sont plausibles ou pas. En quoi la propriété ci-dessus nous y aide-t-elle ? Réponse : parce qu'elle spécifie la loi que doit suivre une quantité qui ne dépend que de μ et de l'observation. Par conséquent, pour $\mu \in \mathbb{R}$, si la quantité

$$\frac{n^{-1}\sum_{i=1}^n X_i - \mu}{\mathrm{var}_{\mathrm{B}}^{1/2}(X_1, \dots, X_n) \,/\, n^{1/2}} \tag{FL'}$$

a une valeur particulièrement peu plausible pour une loi $\mathrm{T}_{\mathrm{St}}(n-1)$, ce sera vraisemblablement que μ n'était pas la vraie valeur de $\mu_{\checkmark}$! Maintenant, il s'agit de dire quelles sont les valeurs pour la quantité ci-dessus qui nous laissent le plus soupçonner que le véritable $\mu_{\checkmark}$ n'est pas égal à μ. Si par exemple μ est plus petit que la véritable valeur $\mu_{\checkmark}$, on s'attend à ce que les valeurs trouvées pour notre statistique de test dévient vers $+\infty$; tandis que si μ est plus grand que $\mu_{\checkmark}$, on s'attend à ce qu'elles dévient vers $-\infty$. Par conséquent, nous allons considérer que les valeurs les plus suspectes pour notre statistique de test sont celles qui sont les plus "extrêmes", et donc que les valeurs les moins suspectes sont celles les plus "centrales". (Comme expliqué plus haut, normalement lorsque notre but est de construire un intervalle de confiance, on doit toujours se retrouver dans ce cas).

Sous l'hypothèse que $\mu_{\checkmark}$ est effectivement égal à μ, la statistique de test doit suivre la loi $\mathrm{T}_{\mathrm{St}}(n-1)$, en vertu de ce que nous avons dit ci-dessus. Nous voulons une procédure qui rejette alors l'hypothèse que $\{\boldsymbol{\mu} = \mu\}$ dans une proportion au plus α des cas, et par ailleurs nous avons dit qu'il nous paraissait intelligent de rejeter les valeurs trop petites ou trop grandes : on va par conséquent se placer dans une situation de test *bilatéral*. Soient donc $q_{\alpha/2}^{(n-1)}$ et $q_{1-\alpha/2}^{(n-1)}$ les quantiles aux niveaux resp. $\alpha/2$ et $(1-\alpha/2)$ de la loi $\mathrm{T}_{\mathrm{St}}(n-1)$: nous décidons de rejeter l'hypothèse nulle lorsque la statistique de test est $< q_{\alpha/2}$ ou $> q_{1-\alpha/2}$. Maintenant, la condition sous laquelle nous acceptons l'hypothèse nulle, à savoir

$$q_{\alpha/2}^{(n-1)} \leqslant \frac{n^{-1}\sum_{i=1}^{n} X_i - \mu}{\mathrm{var}_{\mathrm{B}}^{1/2}(X_1, \dots, X_n)/n^{1/2}} \leqslant q_{1-\alpha/2}^{(n-1)} \qquad (\texttt{FM'})$$

peut se réécrire

$$\mu \in \Big[n^{-1}\sum_{i=1}^{n} X_i - q_{1-\alpha/2}^{(n-1)}\,\mathrm{var}_{\mathrm{B}}^{1/2}(X_1, \dots, X_n)/n^{1/2},\; n^{-1}\sum_{i=1}^{n} X_i - q_{\alpha/2}^{(n-1)}\,\mathrm{var}_{\mathrm{B}}^{1/2}(X_1, \dots, X_n)/n^{1/2} \Big] :$$

ça y est, nous l'avons, notre intervalle de confiance (au niveau α), c'est-à-dire l'ensemble des valeurs de μ pour lesquelles nous acceptons l'hypothèse que $\{\boldsymbol{\mu} = \mu\}$ (au risque α) ! ♧

Exemple (FN'). Voyons maintenant un exemple plus compliqué. Dans cet exemple, on considère le modèle suivant de sondage électoral (idéalisé). Il y a un très grand nombre d'électeurs, tellement grand qu'on peut le considérer comme infini, dont une proportion $\pi_{A\checkmark}$ souhaite voter pour la candidate A, une proportion $\pi_{B\checkmark}$ pour le candidat B, et une proportion $\pi_{C\checkmark}$ pour la candidate C (donc dans ce modèle, l'espace du paramètre caché est donc $\Theta = \{(\pi_A, \pi_B, \pi_C) \mid \pi_A + \pi_B + \pi_C = 1\}$; et on note $(\boldsymbol{\pi}_A, \boldsymbol{\pi}_B, \boldsymbol{\pi}_C) =: \boldsymbol{\theta}$). On sonde n électeurs, qu'on suppose tirés uniformément et indépendamment dans la population ; parmi ceux-là, on obtient X_A électeurs qui souhaitent voter pour A, X_B électeurs qui souhaitent voter pour B et X_C électeurs qui souhaitent voter pour C (l'espace de l'observation est donc $\mathcal{X}^{(n)} = \{(x_A, x_B, x_C) \in \mathbb{N}^3 \mid x_A + x_B + x_C = n\}$ — notez que notre modèle possède n comme paramètre du modèle).

Supposons que nous nous intéressions à la quantité $\varphi(\boldsymbol{\theta}) := \boldsymbol{\pi}_A - \boldsymbol{\pi}_B$, pour laquelle nous cherchons à établir un intervalle de confiance. Notons $\hat{\boldsymbol{\pi}}_A := X_A/n$, resp. $\hat{\boldsymbol{\pi}}_B := X_B/n$, la proportion parmi les sondés d'électeurs souhaitant voter pour A,

resp. pour B. Le théorème-limite central nous assure alors que, lorsque n tend vers l'infini,

$$\mathrm{Loi}_{\checkmark}\left(\frac{(\hat{\pi}_A - \hat{\pi}_B) - (\pi_{A\checkmark} - \pi_{B\checkmark})}{n^{1/2}}\right) \to \mathrm{Normale}\big(0, \pi_{A\checkmark}(1-\pi_{A\checkmark}) + \pi_{B\checkmark}(1-\pi_{B\checkmark}) + 2\pi_{A\checkmark}\pi_{B\checkmark}\big).$$

La variable aléatoire de notre membre de gauche a bien l'allure que nous souhaitons, car elle ne dépend que de l'observation et de la quantité d'intérêt $\varphi(\theta_{\checkmark})$. Par contre, le membre de droite ne va pas du tout, car il dépend du paramètre caché $\theta_{\checkmark}$! Pour pallier cela, on peut essayer de diviser par l'écart-type de la variable normale :

$$\mathrm{Loi}_{\checkmark}\left(\frac{(\hat{\pi}_A - \hat{\pi}_B) - (\pi_{A\checkmark} - \pi_{B\checkmark})}{\big(\pi_{A\checkmark}(1-\pi_{A\checkmark}) + \pi_{B\checkmark}(1-\pi_{B\checkmark}) + 2\pi_{A\checkmark}\pi_{B\checkmark}\big)^{1/2} n^{1/2}}\right) \overset{n\to\infty}{\to} \mathrm{Normale}(0,1). \tag{FO'}$$

Cela règle le problème du membre de droite, mais alors nous avons à nouveau un souci avec le membre de gauche, car maintenant il ne dépend pas uniquement (outre les observations) de $(\pi_A - \pi_B)$ mais aussi de $(\pi_A(1-\pi_A)+\pi_B(1-\pi_B)+2\pi_A\pi_B)^{1/2}$, qui n'est pas une fonction de $\varphi(\theta)$! L'idée est alors de remplacer le facteur $(\pi_A(1-\pi_A)+\pi_B(1-\pi_B)+2\pi_A\pi_B)^{-1/2}$ par quelque chose qui en est très proche, mais qui ne dépend que des observations : typiquement, en remplaçant π_A et π_B par des *estimateurs* de ces quantités (cf. chap. **??**) : le choix le plus naturel est, en l'occurrence, de remplacer π_A par $\hat{\pi}_A$ et π_B par $\hat{\pi}_B$. Pour notre modèle, on peut alors montrer que, dans ce cas, on a effectivement

$$\mathrm{Loi}_{\checkmark}\left(\frac{(\hat{\pi}_A - \hat{\pi}_B) - (\pi_{A\checkmark} - \pi_{B\checkmark})}{\big(\hat{\pi}_A(1-\hat{\pi}_A) + \hat{\pi}_B(1-\hat{\pi}_B) + 2\hat{\pi}_A\hat{\pi}_B\big)^{1/2} n^{1/2}}\right) \overset{n\to\infty}{\to} \mathrm{Normale}(0,1). \tag{FP'}$$

Maintenant, nous pouvons reprendre le raisonnement de l'exemple précédent. Considérons, pour $\delta \in \mathbb{R}$, la statistique de test

$$\frac{(\hat{\pi}_A - \hat{\pi}_B) - \delta}{\big(\hat{\pi}_A(1-\hat{\pi}_A) + \hat{\pi}_B(1-\hat{\pi}_B) + 2\hat{\pi}_A\hat{\pi}_B\big)^{1/2} n^{1/2}} \tag{FQ'}$$

pour tester l'hypothèse $\{\pi_A - \pi_B = \delta\}$. Si l'hypothèse est vraie, cette statistique de test suivra asymptotiquement la loi Normale$(0,1)$. Si maintenant l'hypothèse est fausse, dans le cas où $\pi_{A\checkmark} - \pi_{B\checkmark} < \delta$, la statistique de test aura tendance à s'écarter vers $-\infty$, tandis que dans le cas où $\pi_{A\checkmark} - \pi_{B\checkmark} > \delta$, elle aura tendance à s'écarter vers $+\infty$. Nous allons donc considérer comme suspectes les valeurs les extrêmes prises par la loi Normale$(0,1)$ et comme plausibles ses valeurs les plus centrales. (Comme dit précédemment, ce critère de suspicion était en fait attendu dans le contexte de la construction d'un intervalle de confiance). Soit α le niveau (asymptotique) de l'intervalle de confiance que nous souhaitons mettre en place, et soient $q_{\alpha/2}$ et $q_{1-\alpha/2}$ les quantiles aux niveaux $\alpha/2$ et $(1-\alpha/2)$ de la loi Normale$(0,1)$. Regarder si

$$\frac{(\hat{\pi}_A - \hat{\pi}_B) - \delta}{\big(\hat{\pi}_A(1-\hat{\pi}_A) + \hat{\pi}_B(1-\hat{\pi}_B) + 2\hat{\pi}_A\hat{\pi}_B\big)^{1/2} / n^{1/2}} \notin [q_{\alpha/2}, q_{1-\alpha/2}] \tag{FR'}$$

sera alors bien un test asymptotiquement de niveau α de l'hypothèse nulle $\{\pi_A - \pi_B = \delta\}$. Et l'ensemble des valeurs pour lesquelles l'hypothèse nulle est acceptée est le

suivant : (où nous avons utilisé, pour simplifier la formule, qu'on a $q_{\alpha/2} = -q_{1-\alpha/2}$ par symétrie de la loi normale standard),

$$[\hat{\boldsymbol{\pi}}_A - \hat{\boldsymbol{\pi}}_B \pm q_{1-\alpha/2}(\hat{\boldsymbol{\pi}}_A(1-\hat{\boldsymbol{\pi}}_A) + \hat{\boldsymbol{\pi}}_B(1-\hat{\boldsymbol{\pi}}_B) + 2\hat{\boldsymbol{\pi}}_A\hat{\boldsymbol{\pi}}_B)^{1/2} / n^{1/2}] : \qquad (\mathsf{FS}')$$

cela est donc un intervalle de confiance asymptotiquement de niveau $\boldsymbol{\alpha}$ pour la quantité $\boldsymbol{\pi}_A - \boldsymbol{\pi}_B$. ♧

14.2 Intervalles de prédiction

Définition

! **Définition** (FT') (Intervalle de prédiction (fréquentiste))**.** Soit un modèle statistique de prédiction ayant pour espace du paramètre caché $\boldsymbol{\Theta}$ et pour espaces des observations resp. passée et future $\mathcal{X}$ et $\mathcal{Y}$, la loi d'une observation pour la valeur θ du paramètre étant $\mathrm{Loi}_\theta(X, Y)$. Soit $g \coloneqq \mathcal{Y} \to \mathbb{R}$ une application définissant une quantité d'intérêt réelle $g(Y)$. Notant *Ivl* l'ensemble des intervalles de $\mathbb{R}$, on dit alors qu'une application $IPF : \mathcal{X} \to Ivl$ définit un *intervalle de prédiction (fréquentiste) au risque* $\boldsymbol{\alpha}$ *pour* $g(Y)$ lorsque, pour tout $\theta \in \boldsymbol{\Theta}$,

$$\mathbb{P}_\theta(IPF(X) \not\ni g(Y)) \leqslant \alpha. \qquad (\mathsf{FU}')$$

♡

Remarque (FV')*.* Bien que l'intervalle de prédiction fréquentiste porte sur une fonction de l'observation future (et non du paramètre caché), la condition sur le fait que l'intervalle contienne la quantité d'intérêt est quantifiée, elle, par les valeurs possibles du paramètre caché ! C'est parfaitement naturel, quand on y réfléchit bien : en statistique, tout le problème est qu'on ignore la valeur du paramètre caché, mais qu'on sait en revanche ce qui se passe à valeur donnée du paramètre caché ; par conséquent, pour faire des prédictions qui fonctionnent bien, il faut avoir quelque chose qui fonctionne bien quelle que soit la valeur du paramètre caché : et cela n'a rien à voir avec le fait que notre prédiction elle-même concerne le paramètre caché ou l'observation future ! ♧

Remarque (FW')*.* Toutes les remarques sur la différence entre l'interprétation bayésienne et l'interprétation fréquentiste que nous avions faites pour les intervalles de confiance s'appliquent aussi, *mutatis mutandis*, aux intervalles de prédiction. ♧

Construction

Lorsqu'il s'agit de construire un intervalle de prédiction fréquentiste, l'idée de faire une famille de tests paramétrée par la valeur testée pour la quantité d'intérêt ne fonctionne plus : en effet, par nature, un test d'hypothèse ne peut porter que sur le paramètre caché, et pas sur l'observation future... En revanche, la technique exposée en § 14.1, elle, peut être adaptée sans difficulté. Cela donne la procédure suivante :

! **Principe** (FX')**.**

1º On détermine une variable aléatoire

$$t(X, .g(Y)) \qquad (\mathsf{FY}')$$

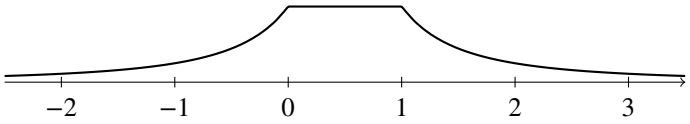

FIGURE 14.1 – Densité de la loi de $(Y_3 - \min(X_1, X_2)) / |X_2 - X_1|$ lorsque X_1, X_2, Y_3 sont tirées indépendamment selon une loi uniforme.

telle que $\mathrm{Loi}_\theta\big(t(X, .g(Y))\big)$ *ne dépende pas de* θ [**] *Dans les exercices et problèmes qui vous seront soumis, cette variable aléatoire sera donnée directement par l'énoncé, ou du moins fortement suggérée.*

2° *On détermine ensuite l'intervalle de fluctuation* J *pour* $\mathrm{Loi}_{\checkmark}\big(t(X, .g(Y))\big)$. *(Rappelons que cette loi est connue indépendamment de* $\theta_{\checkmark}$*). L'intervalle* J *est donc tel que*

$$\forall \theta \in \Theta \qquad \mathbb{P}_\theta\big(t(X, .g(Y)) \notin J\big) \leqslant \alpha. \tag{FZ'}$$

(De même que dans la remarque (FI')*, on n'était en fait pas obligé que* $\theta \mapsto \mathrm{Loi}_\theta\big(t(X, .g(Y))\big)$ *fût constante : tout ce qui compte est d'avoir la propriété ci-dessus, qu'on peut obtenir par exemple en prenant pour* J *la réunion de tous les intervalles de fluctuation des* $\mathrm{Loi}_\theta\big(t(X, .g(Y))\big)$. *Mais en pratique, je ne connais pas de cas intéressant où la loi de* $t(X, .g(Y))$ *soit variable...).*

3° *Un intervalle de prédiction pour* $.g(Y)$ *est alors donné par*

$$x \mapsto \{g \in \mathbb{R} \mid t(x, g) \in J\}. \tag{GA'}$$

◇

[††].

Un exemple

Exemple (GB'). On considère un modèle où l'observation passée est un couple d'observations $X_1, X_2 \in \mathbb{R}$ et l'observation future est $Y_3 \in \mathbb{R}$, sachant que X_1, X_2, Y_3 sont i.i.d. selon une loi $\mathrm{Unif}^{\mathrm{me}}(\alpha_{\checkmark}, \beta_{\checkmark})$ avec $\alpha_{\checkmark}$ et $\beta_{\checkmark}$ inconnus ($\alpha_{\checkmark} < \beta_{\checkmark}$). On peut montrer que, quels que soient $\alpha_{\checkmark}$ et $\beta_{\checkmark}$, la loi (sous $\mathbb{P}_{\checkmark}$) de $(Y_3 - \min(X_1, X_2))/|X_2 - X_1|$ est toujours la même : et plus précisément, on a (voir la figure 14.1)

$$\mathbb{P}_{\checkmark}\left(\frac{Y_3 - \min(X_1, X_2)}{|X_2 - X_1|} \in \mathrm{d}t\right) = \begin{cases} \frac{1}{3}(1-t)^{-2}\,\mathrm{vol}(\mathrm{d}t) & \text{pour } t \leqslant 0\,; \\ 1/3 \times \mathrm{vol}(\mathrm{d}t) & \text{pour } t \in]0, 1]\,; \\ \frac{1}{3}t^{-2}\,\mathrm{vol}(\mathrm{d}t) & \text{pour } t > 1. \end{cases} \tag{GC'}$$

On calcule que l'intervalle de fluctuation à 90 % de cette loi est $[-3{,}48, 4{,}48]$: il y donc 90 % de chances que $(Y_3 - \min(X_1, X_2))/|X_2 - X_1| \in [-3{,}48, 4{,}48]$, autrement dit que Y_3 tombe entre $\min(X_1, X_2) - 3{,}48 \times |X_2 - X_1|$ et $\min(X_1, X_2) + 4{,}48 \times |X_2 - X_1|$, qui sont donc les bornes de notre intervalle de prédiction. L'application numérique donne, pour $x_{1\checkmark} = 64$ et $x_{2\checkmark} = 61$, qu'on aura

$$Y_3 \in [50{,}58, 74{,}42] \tag{GD'}$$

au risque 10 %. ♧

[††]. Attention : Ici il s'agit bien d'une dépendance en θ ! Confer remarque (FV').

Quatrième partie

Du bon usage de la statistique

Chapitre 15

Choix de la priore

15.1 Problématique

Principe général pour choisir la priore

Le choix de la priore est absolument crucial en analyse bayésienne : en partant de deux priores différentes, on arrive à deux postériores différentes, et donc potentiellement à deux conclusions différentes pour notre analyse... Mais alors, comment faut-il choisir la priore ?! La réponse est paradoxale : il n'y a pas forcément *une* façon bien définie de choisir la priore... mais on ne peut pas la choisir arbitrairement pour autant !

Principe (GE'). *La priore a un sens bien précis : elle exprime la probabilité à priori que le paramètre caché vaille telle ou telle valeur. Mais comment donner un sens à la « probabilité » que, par exemple, la nouvelle méthode de l'enseignant donne un meilleur résultat moyen ?... Nous avons en effet l'habitude de dire « la probabilité d'un évènement, c'est la proportion de fois où cet évènement se produirait si on répétait un très grand nombre de fois l'expérience à l'identique » : c'est ce qu'on appelle parfois la définition* fréquentiste *de la notion de probabilité. Or ici, il n'y a pas d'expérience répétable à l'identique : l'enseignant a eu l'idée d'*une *nouvelle méthode bien particulière, qui est peut-être efficace et peut-être pas, mais cela ne dépend en rien du* hasard *: simplement du fait que l'enseignant n'arrive pas à* deviner *à l'avance les valeurs de* $\boldsymbol{\mu}$ *et* $\boldsymbol{\sigma}$*... On entre alors dans la vision* épistémique *de la notion de probabilité. Si, par exemple, l'enseignant dit « Il y a 30 % de probabilité que* $\boldsymbol{\sigma}$ *soit plus petit que* $\sigma_{\text{réf}}/2$ *», cela veut en fait dire qu'il serait exactement aussi* surpris *que s'apercevoir que* $\{\boldsymbol{\sigma} < \sigma_{\text{réf}}/2\}$ *qu'il le serait de voir qu'une variable aléatoire uniformément tirée entre* 0 *et* 1 *soit inférieure à* 0,30 *: par exemple, si on lui proposait un pari rapportant* 5 € *(moins déduction de la mise) dans le cas où* $\{\boldsymbol{\sigma} < \sigma_{\text{réf}}/2\}$ *(après expérimentation de la nouvelle méthode, s'entend, sur un nombre d'élèves tellement grand qu'on n'aurait plus de doute sur la vraie valeur de* $\boldsymbol{\sigma}$*), il serait prêt à entrer dans un tel pari pour une mise de* 1 €*, mais pas pour une mise de* 2 €[∗]. ◇ !!

[∗]. En supposant ici que, pour des montants inférieurs à 5 € en valeur absolue, l'utilité de « recevoir un gain net de x € » (une perte étant comptée comme un gain négatif) est pratiquement une fonction affine de x : cette approximation est très raisonnable lorsque les montants en jeu sont petits, car dans ce cas il n'y a guère d'aversion (ni de propension) au risque, confer Remarque (HQ').

Probabilités épistémiques

La notion de probabilité épistémique suggère deux difficultés conceptuelles, que nous discutons dans cette sous-section.

Point (GF') (Subjectivité des probabilités épistémiques). Une première difficulté avec la notion de probabilités épistémiques est que ces quantités sont fondamentalement *subjective* : dans le cas du modèle du pédagogue, par exemple, une enseignante expérimentée arrivera certainement à deviner avec beaucoup plus de précision l'effet de la nouvelle méthode qu'un enseignant novice, et attribuera donc une loi différente à $\boldsymbol{\theta}$ pour la *même* situation... Mais est-ce à dire que l'enseignante expérimentée a *raison* pour sa loi et que le novice a tort ? Non : le point est qu'une probabilité épistémique n'est définie que *par rapport à la personne qui pose cette probabilité*, et est donc correcte dès lors que la personne a bien pris en compte, de façon appropriée, toute l'information dont elle disposait ! ♧

Remarque (GG'). Il convient de noter une subtilité dans le point précédent : « prendre en compte de façon appropriée toute l'information dont on dispose » inclut, le cas échéant, l'information concernant le fait que d'autres personnes (ayant accès à des informations que nous n'avons pas) attribuent telle ou telle probabilité épistémique au phénomène concerné ! Ainsi, dans le cas de l'enseignant novice, cela signifie qu'il ne devra pas donner la même loi à priori à $\boldsymbol{\theta}$ selon qu'il a eu, ou non, l'occasion de discuter avec sa collègue expérimentée : en cas de discussion, et pour peu qu'il accorde un minimum de foi à l'expertise de sa collègue, il devra modifier sa croyance à priori pour tenir compte de l'avis de celle-ci... [†] ♧

Point (GH'). La seconde difficulté est que, bien souvent, nous avons tendance à nous tromper sur notre évaluation des probabilités épistémiques ! Lequel (ou laquelle) d'entre vous n'a-t-il jamais dit à des amis, par exemple, « oui, je viendrai à votre soirée ; c'est sûr à 99,99 % », et pourtant a eu un contretemps de dernière

[†]. En mathématiques et en philosophie de la connaissance, la problématique de la prise en compte des probabilités épistémiques émises par des tiers donne lieu à des développements que je trouve absolument fascinants... Imaginons ainsi deux individus supposés parfaitement rationnels, qui ont recueilli chacun de son côté diverses informations concernant une affirmation donnée. S'ils n'ont pas communiqué entre eux, les probabilités épistémiques qu'ils auront concernant cette affirmation seront évidemment différentes : rien d'étonnant jusque-là. À l'inverse, si chacun va transmettre à l'autre la totalité des informations dont il dispose, les deux individus se retrouvent alors avec exactement les mêmes informations, et (puisqu'ils sont parfaitement rationnels) vont en déduire la même probabilité épistémique. Mais bien entendu, communiquer la totalité des informations dont on dispose est en pratique impossible, tant le détail de ces informations est encodé de façon subtile dans notre cerveau... Mais c'est alors qu'un théorème fascinant intervient : supposons que le premier individu (appelons-le $\boldsymbol{A}$) envoie au second ($\boldsymbol{B}$) un message indiquant simplement la probabilité épistémique qu'il attribue au phénomène. $\boldsymbol{B}$ répond ensuite en informant $\boldsymbol{A}$ de la probabilité épistémique qu'il attribue au phénomène *après prise en compte du premier message de* $\boldsymbol{A}$. Puis $\boldsymbol{A}$ réplique à son tour en indiquant sa probabilité épistémique *après prise en compte du message de* $\boldsymbol{B}$, et ainsi de suite... Dans ce cas, on peut montrer que les probabilités épistémiques de $\boldsymbol{A}$ et $\boldsymbol{B}$ doivent converger (en les supposant parfaitement rationnels) vers une valeur commune au fil des échanges, et assez rapidement qui plus est ! Or, on observe en pratique que ce n'est pas ce qui se passe : des gens très intelligents et très rationnels peuvent camper sur des évaluations épistémiques fortement différentes concernant le même phénomène, même en ayant connaissance de leur discordance... En philosophie, on appelle cela la question du *désaccord entre pairs épistémiques*. Les philosophes essayent de comprendre si l'omniprésence de tels désaccords est l'indication que les humains sont irrémédiablement profondément irrationnels, ou s'il n'existe pas des mécanismes subtils qui font que le théorème mathématique sur la convergence des probabilités épistémiques ne s'applique pas dans la vraie vie... ?

minute... ?! Pourtant, lorsque nous annonçons un tel niveau de certitude, l'annulation ne devrait survenir qu'une fois sur 10 000 : beaucoup plus que le nombre de soirées auxquelles vous êtes jamais allés... De manière générale, notre intuition est très mauvaise pour évaluer les probabilités extrêmement petites (ou, par complémentaire, extrêmement proches de 1) : par exemple, nous avons beaucoup de mal à admettre que la probabilité que deux personnes prises au hasard dans le monde soient nées le même jour à moins de 5 minutes d'intervalle est beaucoup plus grande que la probabilité qu'une grille de loto donnée remporte le jackpot ! (ce qui consiste à trouver 5 numéros corrects parmi 49, plus 1 parmi 10). Et que cette probabilité elle-même est beaucoup plus grande que la probabilité qu'une météorite qui tombe sur Terre se trouve atterrir à moins de 2 km de chez vous...

Or, autant la probabilité épistémique que nous pouvons attribuer à un évènement dépend de l'*information* dont nous disposons, autant on ne peut pas lui donner *n'importe quelle* valeur selon notre convenance ! En effet, il faut dans tous les cas que, par exemple, les évènements dont nous prédisons à priori que la probabilité qu'ils arrivent vaut 10 % se produisent effectivement environ 10 fois pour cent prédictions [‡]) (même s'il ne s'agit pas là à proprement parler de la répétition d'une expérience, puisque ce sont des évènements complètement différents à chaque fois : c'est ce qu'on appelle, dans le langage technique, la *calibration* correcte des probabilités épistémiques [§]. Il faut donc une méthode permettant de quantifier de façon raisonnablement précise la façon de convertir nos informations et nos croyances en probabilités... ♧

Nous allons présenter ci-après les façons appropriées pour choisir une priore. Comme il faut s'y attendre, la méthode à suivre dépendra du type d'informations dont nous disposons (ou pas) sur le paramètre caché. Nous allons commencer par présenter les cas où la loi du paramètre caché peut être déterminée sans ambigüité, en allant progressivement vers des situations où l'information sur le paramètre caché sera décrite de façon de plus en plus vague.

Remarque (GI'). Dans le cadre de ce cours, on ne vous demandera normalement pas de choisir vous-mêmes la priore d'un modèle statistique ; cependant, il est important que vous connaissiez les règles pour le choix d'une priore, et que vous soyez

[‡]. Ici j'ai supposé implicitement que les différentes prédictions concernées portaient sur des évènements pouvant être considérés comme indépendants les uns des autres. En effet, il va de soi que, si je m'intéresse aux cent évènements « il pleuvra sur la place Stanislas le 22 septembre 2099 à 17 h 00 », « il pleuvra sur la place Stanislas le 22 septembre 2099 à 17 h 01 », ..., « il pleuvra sur la place Stanislas le 22 septembre 2099 à 18 h 39 », que j'attribue à chacun de ces cent évènements une probabilité de 10 %, et que pourtant aucun d'entre eux ne se réalise, ce n'est pas le signe d'une mauvaise calibration de ma part : simplement que, pour des évènements aussi corrélés entre eux, il n'y a aucune raison de s'attendre à observer un phénomène de type « loi des grands nombres » !...

[§]. Nous avons vu que deux individus rationnels peuvent très bien attribuer des probabilités épistémiques différentes au même évènement, pour peu qu'ils s'appuient sur des informations différentes (et qu'ils n'aient pas pu en échanger entre eux) : ainsi, le fait de savoir qu'un individu donné attribue une probabilité épistémique de tant ou tant à tel ou tel phénomène ne peut jamais constituer une preuve que cet individu est irrationnel (à moins, bien sûr, de connaitre précisément toutes les informations dont il dispose). En revanche, les probabilités épistémiques émises par un individu rationnel *doivent* être correctement calibrées : ainsi, si les pronostics de votre belle-sœur concernant les matchs de football de Ligue 1 sont tels que, quand elle dit qu'une équipe a au moins 90 % de chances de gagner, on constate en pratique, sur le long terme, que l'équipe désignée comme favorite ne gagne effectivement que dans 70 % des situations ainsi pronostiquées par votre belle-sœur, le problème n'est pas que votre belle-sœur manque d'information : cela prouve que, forcément, les évaluations qu'elle fait des probabilités épistémiques ne traitent (en général) pas les informations dont elle dispose selon un raisonnement correct !

capables de justifier le choix d'une priore donnée par l'énoncé, voire de proposer une priore en vertu de certains critères spécifiés. ♧

15.2 Cas où la priore est non ambigüe

Il s'agit là du cas le plus facile : c'est lorsque la situation confrontée correspond *réellement* à tirer au sort la valeur $\theta_{\checkmark}$ de $\boldsymbol{\theta}$ (quoique sans la révéler), puis à tirer $x_{\checkmark}$ [¶] selon la loi $\mathrm{Loi}(X \mid \boldsymbol{\theta} = \theta_{\checkmark})$!

Exemple (GJ'). Plaçons-nous dans la situation du chasseur, mais avec une petite nuance dans l'interprétation du problème mathématique : ici, on suppose que les qualités du chasseur sont parfaitement connues ; mais par contre, qu'on ignore la qualité de l'*arme* avec laquelle il va tirer : notre chasseur possède en effet trois fusils indistinguables à l'œil nu, mais dont l'un a un léger défaut de fabrication qui fait chuter sa précision : si le chasseur tire avec un bon fusil, il aura un taux de réussite de 30 % ; tandis que s'il tire avec le fusil défectueux, il n'aura un taux de réussite que de 15 %... La question correspond ici à identifier l'état du fusil avec lequel le chasseur tire : comme c'est équivalent à déterminer son taux de réussite, mathématiquement parlant, nous sommes bien dans la situation de base.

Quelle est la priore correspondante ? La réponse ne souffre d'aucune ambigüité : puisque les fusils sont indistinguables à l'œil nu, l'arme qu'utilise notre chasseur est prise au hasard uniformément parmi ses trois fusils ; et il y a donc deux chances sur trois d'avoir un bon fusil (auquel cas $\theta_{\checkmark}$ vaut 30 %), et une chance sur trois d'avoir le fusil défectueux (auquel cas $\theta_{\checkmark}$ vaut 15 %). Ainsi, il *faut* prendre

$$\mathrm{Loi}_{\mathrm{pr}}(\boldsymbol{\theta}) = \tfrac{2}{3}\delta_{30\ \%} + \tfrac{1}{3}\delta_{15\ \%}. \qquad (\mathsf{GK}')$$

♧

Ce genre de situations où la priore est déterminée sans ambigüité sont celles que vous avez dû voir dans vos exercices de lycée. Mais toute la force de la statistique bayésienne est, justement, que sa démarche peut aussi s'appliquer dans des cas où la notion de « probabilité à priori » est définie de façon beaucoup plus floue ! C'est ce que nous verrons dans les § 15.4 et 15.5. Mais auparavant, il importe de préciser une "méta-règle" qui doit s'appliquer *dans tous les cas* pour déterminer une priore, quelle que soit la partie de ce chapitre à laquelle on se réfère : c'est ce que nous allons voir dans la section suivante !

15.3 Postériore d'hier, priore d'aujourd'hui

!! **Principe** (GL'). *Si on a déjà des informations sur $\boldsymbol{\theta}$ issues d'une expérience précédente, alors il faut inclure ces informations dans la loi à priori qu'on attribue à $\boldsymbol{\theta}$!*

En particulier, dans le cas où l'« expérience précédente » en question est formalisée en termes de statistique bayésienne, la loi à priori *pour $\boldsymbol{\theta}$ qu'il convient de prendre pour l'expérience* présente *est celle que l'expérimentateur de l'expérience* précédente *(peu importe qu'il s'agisse de nous-même ou quelqu'un d'autre) aurait trouvée comme loi* à postériori. ♢

[¶]. Ou éventuellement $(x_{\checkmark}, y_{\checkmark})$ si on est dans une situation de prévision, mais, le cas échéant, en ne révélant que $x_{\checkmark}$ dans un premier temps.

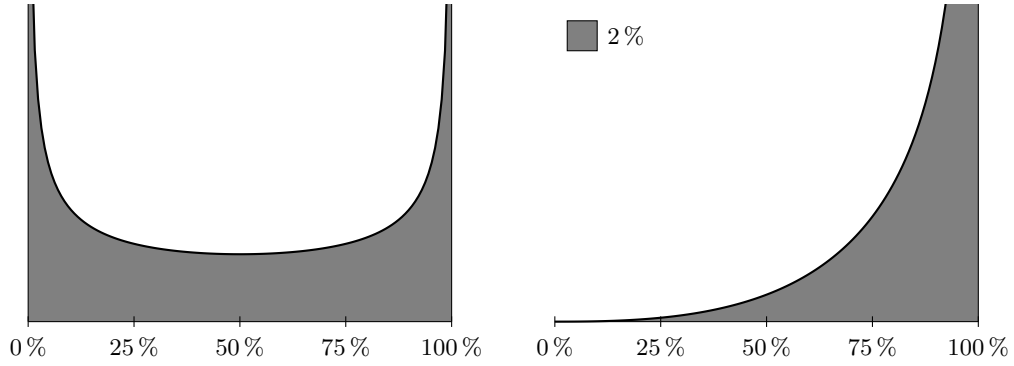

FIGURE 15.1 – Tracés, à la même échelle, de la loi arcsinus (à gauche) et de la loi Bêta($3\frac{1}{2}$, 1/2) (à droite).

Exemple (GM'). Prenons à nouveau le cas du chasseur. Normalement, quand un nouveau candidat arrive, les statisticiens du club du Bouchonnois choisissent comme priore pour $\boldsymbol{\theta}$ la loi arcsinus. Cependant, ici le cas est un peu particulier : car le candidat qui est examiné aujourd'hui a procédé à trois tirs d'essai hier (les conditions de ces tirs d'essais étaient les mêmes que pour le test), qu'il a réussis tous les trois ! Quelle serait la postériore attribuée à $\boldsymbol{\theta}$ pour un chasseur ayant réussi trois tirs ? Ce serait dans ce cas la loi Bêta($3\frac{1}{2}$, 1/2), qui est très nettement différente de la loi arcsinus ! (voir figure 15.1). Et cela est parfaitement logique : puisqu'ici, au vu de ses trois tirs d'essais, avant même que notre candidat commence le test officiel, nous sommes déjà fortement convaincus qu'il a une bonne fiabilité...! ♧

Remarque (GN'). Dans l'exemple précédent, le résultat qu'on obtiendra, en partant de la priore Bêta($3\frac{1}{2}$, 1/2), après analyse bayésienne des 25 tirs de test du chasseur, sera exactement le même que celui qu'on aurait obtenu en partant de la priore arcsinus, *mais en intégrant les trois premiers tirs à nos données !* (et donc en considérant que l'expérience a consisté à tirer sur 28 plateaux en tout). En effet, dans les deux cas, il s'agit du traitement bayésien des *mêmes* informations (à ceci près que dans le premier cas elles ont été traitées en deux étapes, alors que dans le second cas on a procédé directement à un traitement global) : il est donc intuitivement évident que les deux calculs doivent conduire au même résultat !

Il s'agit là, bien entendu, d'un phénomène parfaitement général : dire « on prend la postériore du début comme priore pour la fin » ou dire « on considère que l'expérience englobe le début et la fin », ce ne sont que deux façons différentes, et mathématiquement parfaitement équivalentes, de prendre en compte toute l'information disponible ! ☺ ♧

Remarque (GO'). En sciences expérimentales, l'idée de choisir la priore au vu des observations précédentes est particulièrement importante. En effet, il est fréquent que plusieurs chercheurs réalisent des expériences différentes sur le même phénomène (par exemple, l'effet du tabagisme sur le cancer du poumon), correspondant donc au même paramètre caché (en l'occurrence, la façon dont fumer augmente — ou pas — le risque de cancer). Dès lors, il serait absurde et même malhonnête de faire table rase des expériences passées lorsqu'on mène une nouvelle étude : au contraire, notre croyance à priori sur le paramètre caché *doit* être la croyance à postériori consécutive aux expériences prédécentes ! (sauf, bien sûr, si on a des raisons de penser que ces expériences précédentes ont été truquées ou mal conçues statistiquement). Imaginez ainsi qu'un élève vienne me voir et me montre des données épidémiologiques sur sa famille, menées à partir d'une priore non informative (cf. § 15.5 *infra*), dont il déduise,

au prix d'une analyse statistique impeccable, qu'il n'y a que 40 % de chances qu'il existe un lien entre tabagisme et cancer. Vais-je *réellement* croire qu'il y a 60 % de chances que fumer soit sans aucun risque ? Certainement pas ! Car, au vu des nombreuses études scientifiques déjà menées préalablement sur le sujet, j'avais du fait des résultats de ces études de très fortes raisons de croire à priori (c.-à-d. avant de voir les résultats sur la famille de mon élève) que fumer augmente le risque de cancer : le choix d'une priore non informative par mon élève n'était donc pas pertinent ! La façon correcte de traiter les résultats d'une telle étude serait plutôt celle d'une *révision de croyances*, du type suivant : avant d'avoir vu l'enquête de mon élève, j'attribuais une probabilité de 99,995 % à l'existence d'un lien entre tabagisme et cancer ; après avoir vu son enquête, je n'y attribue plus "qu'"une probabilité de 99,994 %.

Curieusement, la démarche que je viens d'exposer n'est *pas* celle de la plupart des travaux de sciences expérimentales : les chercheurs dans ces matières ont au contraire tendance à utiliser l'approche *fréquentiste*, et à analyser les résultats des différentes expériences *indépendamment*. Peut-être est-ce parce que l'approche bayésienne est trop subtile à manipuler de façon convaincante sur de tels cas concrets, à moins que ce ne soit simplement parce que cette approche n'est pas celle qui a été enseignée aux chercheurs dans leur jeunesse... ¨˙ Seules les *méta-études* utilisent parfois cette approche bayésienne pour, justement, mettre à bout les résultats des différentes études individuelles : elles arrivent alors à montrer de façon quantifiable comment beaucoup de petites preuves peuvent conduire à une certitude très grande, ou quelle est la conclusion générale qu'il convient de tirer de résultats individuels en apparence contradictoires ; et ce, grâce à la méthode bayésienne en particulier ˘ ♧

15.4 Traduction mathématique d'une expertise préalable

Je parle d'« expertise » dans le cas où on n'a pas d'information de l'énoncé qui nous donne une connaissance *parfaite* de la priore, mais où on a cependant une idée *honnête* de ce que vaut cette priore au vu de notre expertise. En fait, « utiliser notre expertise pour choisir la priore » s'apparente à l'idée de « utiliser les expériences précédentes pour faire de la postériore de ces expériences la priore de notre modèle », mais dans un cadre où on n'est pas forcément à même de poser un modèle précis, comme nous allons le voir sur les exemples suivants.

Exemple (GP'). Dans les exemples précédents traitant du problème du chasseur, nous avons proposé une priore de type neutre (voir § 15.5 *infra*) pour définir la loi à priori de $\boldsymbol{\theta}$ sur $]0,1[$, en l'occurrence la loi arcsinus. Mais en fait, pour peu que le club du Bouchonnois ait un peu d'expérience dans le recrutement des candidats, il peut faire bien mieux...! Par exemple, à force de voir des candidats défiler, puis d'avoir pu juger de leurs performances sur le plus long terme pour ceux qui ont été recrutés [‖], le club s'est certainement fait une idée honnête de la proportion de bons chasseurs (disons ici qu'un « bon chasseur » est un chasseur capable de toucher au moins un plateau sur quatre, autrement dit pour lequel $\{\boldsymbol{\theta} \geqslant 1/4\}$) et de mauvais chasseurs. Ainsi, si sur les 20 candidats précédents ils s'est avéré y avoir 8 bons chasseurs (soit 40 % des candidats passés) et 12 mauvais chasseurs (soit 60 %), plutôt que de poser une priore arcsinus pour $\boldsymbol{\theta}$ ($\boldsymbol{\theta}$ représentant le niveau

[‖]. Ici nous ferons l'hypothèse que le club ne rejette un candidat que s'il n'y a aucun doute raisonnable sur le fait que ce soit un mauvais chasseur. Dès lors, le club connait pour chaque candidat son statut de bon ou de mauvais chasseur : soit que son test n'ait laissé aucun doute raisonnable, soit qu'il a été recruté et qu'on a pu jauger finement de son niveau réel par la suite.

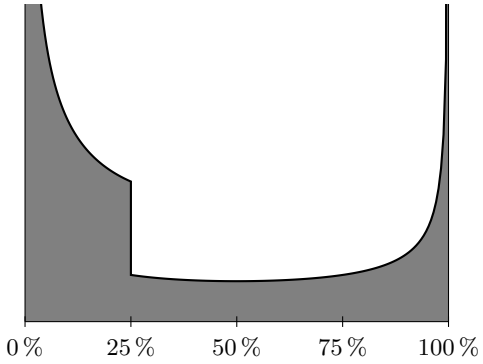

FIGURE 15.2 – Une priore donnant une masse de 60 % à l'hypothèse que le candidat ait une fiabilité inférieure à 1/4, resp. une masse de 40 % à l'hypothèse que sa fiabilité soit supérieure.

d'un nouveau candidat), il sera plus judicieux de poser

$$\mathbb{P}_{\mathrm{pr}}(\boldsymbol{\theta} \in \mathrm{d}\theta) = \begin{cases} \alpha_{\mathrm{bon}} \times \mathbb{P}\big(\mathrm{LoiArcsin} \in \mathrm{d}\theta\big) & \text{pour } \theta \geqslant 1/4\,; \\ \alpha_{\mathrm{mal}} \times \mathbb{P}\big(\mathrm{LoiArcsin} \in \mathrm{d}\theta\big) & \text{pour } \theta < 1/4, \end{cases} \qquad (\texttt{GQ'})$$

avec

$$\alpha_{\mathrm{bon}} \coloneqq \frac{40\ \%}{\mathbb{P}(\mathrm{LoiArcsin} \geqslant 1/4)}\,; \qquad \alpha_{\mathrm{mal}} \coloneqq \frac{60\ \%}{\mathbb{P}(\mathrm{LoiArcsin} < 1/4)}\,; \qquad (\texttt{GR'})$$

autrement dit, on *pondère* les différentes possibilités par rapport à la priore arcsinus de façon à avoir, après pondération, une probabilité à priori de resp. 40 % et 60 % pour être resp. un bon ou un mauvais chasseur.

D'ailleurs, même si le club ne dispose pas de statistiques précises sur ses candidats passés, il peut toujours se servir de son expérience pour proposer une estimation "au doigt mouillé" de la distribution des niveaux parmi les candidats potentiels : ainsi, si l'expérience semble montrer qu'il y a un peu plus de "mauvais" chasseurs que de bons, et que parmi ces bons chasseurs, rares sont les chasseurs qui soient carrément "très bons", le club peut chercher, en guise de priore, une distribution de probabilité sur $]0,1[$ qui donne une masse de 60 % à l'intervalle $]0,1/4[$ (les mauvais chasseurs), 30 % à l'intervalle $[1/4,3/4[$ (les chasseurs assez bons), et 10 % à l'intervalle $[3/4,1[$ (les très bons chasseurs). En outre, pour éviter de créer des seuils trop nets au niveau des valeurs-seuils 1/4 et 3/4, il peut être judicieux de souhaiter aussi que la densité à priori soit C^1, ce qui revient à dire qu'on veut que la fonction de répartition de la priore *interpole* de façon régulière les valeurs 0 % en 0, 60 % en 1/4, 90 % en 3/4 et 100 % en 1 : on peut résoudre un tel problème d'interpolation à l'aide de techniques d'analyse numérique[**] , ce que j'ai fait pour obtenir une

[**]. En analyse numérique, vous avez vu qu'une manière de résoudre ce genre de problème d'interpolation est de chercher, parmi les densités de probabilité respectant les contraintes sur la masse totale des intervalles respectifs $]0,1/4[$, $[1/4,3/4[$ et $[3/4,1[$, celle qui minimise l'"énergie" $\int_0^1 f'(x)^2\mathrm{d}x$ [††], et que dans ce cas la fonction de répartition d'obtient à l'aide de « splines cubiques ». La méthode que j'ai suivie pour obtenir la figure 15.3 est apparentée à l'idée de minimisation de l'énergie, mais avec quelques fioritures supplémentaires pour tenir comptes des spécificités du problème.

[††]. Notez que c'est le carré de la dérivée *première* de la densité qu'on considère ici, puisque l'interpolation porte sur la fonction de répartition, de sorte que la dérivée seconde de la fonction de répartition est bien la densité première de la densité.

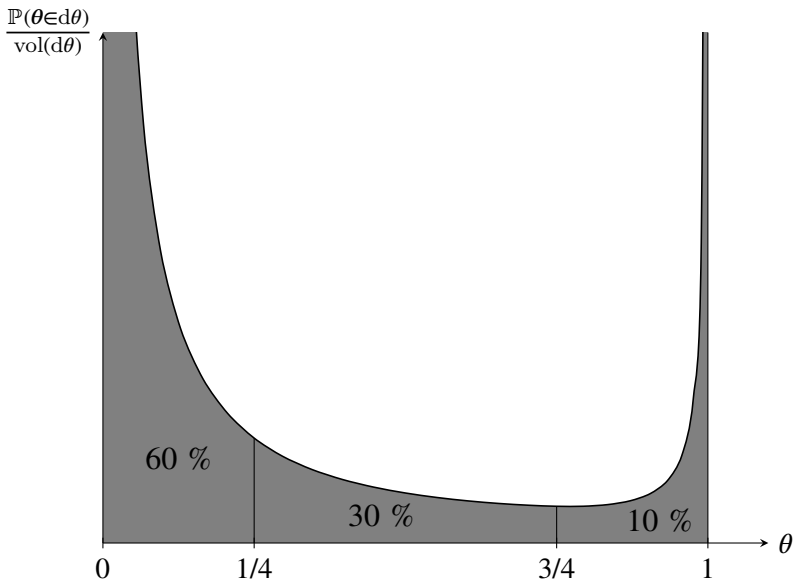

FIGURE 15.3 – La loi représentée ci-dessus propose une priore sur $\boldsymbol{\theta}$ traduisant l'expertise informelle du club de chasse : il y a plus de mauvais chasseurs que de bons ; et parmi les bons, rares sont les très bons. Concrètement, on a fait en sorte que 60 % de la masse soit comprise entre 0 et 1/4, 30 % de la masse entre 1/4 et 3/4, et 10 % de la masse entre 3/4 et 1. On a en outre imposé que le comportement qualitatif de la priore près des valeurs extrêmes 0 et 1 soit le même que celui de la loi arcsinus. (Et, sous l'ensemble des contraintes ci-dessus, on a choisi une priore dont la densité soit "aussi régulière que possible").

distribution de probabilité "aussi raisonnable que possible" vérifiant les contraintes requises, dont vous pouvez voir le tracé en figure 15.3.

La priore de la figure 15.3 est donc une traduction en termes numériques de l'expertise de notre club, que nous avions initialement formulée en termes assez vagues :

> *« Il y a un peu plus de mauvais chasseurs que de bons ; et parmi ces bons chasseurs, rares sont les chasseurs qui soient carrément très bons. »*

Il importe néanmoins de noter que le fait de fixer les seuils à 1/4 et 3/4, de prendre des proportions de 60 %, 30 % et 10 %, et le choix de la méthode d'interpolation, tout cela a été fait "au doigt mouillé" [‡‡] : la façon dont le club a traduit ici son expertise sous forme d'une distribution de probabilité est parfaitement honnête ; néanmoins, son expertise n'étant pas quantifiable en des termes parfaitement précis, cette traduction garde une part d'arbitraire... !

♧

!! *Remarque* (GS'). Lorsqu'on estime une priore par expertise, on sait que notre choix

[‡‡]. Concernant le seuil de 1/4 et la proprition de 60 %, vous aurez noté que, dans l'exemple précédant celui-ci, ils n'avaient pas été définis "au doigt mouillé", mais respectivement en tant que *définition* de la notion de « bon chasseur » et que proportion *observée* historiquement. Néanmoins, même si j'ai repris les mêmes valeurs pour l'exemple actuel, on n'est pas obligé de considérer cette fois-ci que ces valeurs correspondent resp. à une définition et à une observation historique exactes : en fait, l'idée ici est plutôt que, sans jamais rien avoir formalisé rigoureusement ni compté précisément, le club considère qu'il considère les candidats comme mauvais lorsque leur fiabilité est inférieure à *environ* 1/4, et que le « un peu plus » de mauvais chasseurs que de bons correspond à une proportion d'*environ* 60 %.

risque de ne refléter qu'imparfaitement la réalité... Dans ce cas, entre deux maux possibles, il est essentiel de choisir le moindre, à savoir : il vaut mieux risquer d'avoir une priore qui attribue une probabilité de 10^{-3} à quelque chose qui se produit en fait avec probabilité 10^{-6}, que l'inverse ! Dit en d'autres termes, il vaut mieux que notre priore soit trop bien répartie entre les différentes possibilités (en restant donc assez prudent sur la façon de traduire en priore les indices dont on dispose) que l'inverse. Ou encore : il vaut mieux garder un soupçon d'approche "neutre" (voir § 15.5) dans la composition de notre priore experte. Ou encore encore : il vaut mieux rester plutôt ouvert à la probabilité d'un phénomène "exceptionnel"... ♧

Remarque (GT'). La justification de la remarque ci-dessus provient de concepts de théorie de l'information : pour mesurer "de combien se trompe" une probabilité approchée $\boldsymbol{Q}$ par rapport à une probabilité véritable $\boldsymbol{P}$ en termes de manque d'information, l'outil adéquat est la *divergence de Kullback–Leibler* $\mathscr{D}_{\mathrm{KL}}(\boldsymbol{P} \parallel \boldsymbol{Q})$. On peut montrer que dans la situation évoquée ci-dessus, si on estime que la probabilité d'un évènement rare est η au lieu de ε, cela donne une divergence de Kullback–Leibler proportionnelle à $\eta - \varepsilon - \ln(\eta/\varepsilon)\varepsilon$. L'application numérique indique alors que le fait de croire qu'un évènement rare a une probabilité 10^{-6} au lieu de 10^{-3} en réalité a un cout, en termes d'information, environ 6 fois plus grand que de croire qu'un évènement rare a une probabilité de 10^{-3} au lieu de 10^{-6} en réalité. ♧

Exemple (GU'). L'exemple ci-dessous montre pourquoi il ne faut pas suivre aveuglément les indices dont on dispose. Dans le cadre de l'exemple (GP'), nous avons déterminé la priore à partir du fait que, dans le passé, 16 candidats s'étaient révélés mauvais contre seulement 4 bons ; et nous en avions déduit que la probabilité à priori qu'un nouveau candidat soit un bon chasseur devait être de l'ordre de 20 %. Imaginons maintenant que *tous* les 20 candidats précédents recontrés par le club aient été mauvais. Aurait-il fallu en déduire qu'il était (à priori) *certain* qu'un nouveau candidat soit un mauvais chasseur ? Évidemment non : après tout, si la proportion de bons chasseurs parmi les candidats potentiels vaut en réalité, disons, 2 %, il n'y a même rien de bien étonnant à n'avoir trouvé aucun bon chasseur parmi les candidats précédents ! (Puisque sur 20 candidats, on doit alors trouver seulement 0,4 bons chasseurs en moyenne). Il est donc plus raisonnable de supposer que la probabilité à priori que le nouveau candidat soit un bon chasseur est certes faible, mais pas nulle ; typiquement, de la prendre de l'ordre de 2,5 % (soit la moitié d'un vingtième), puisque c'est le seuil à partir duquel il n'y a plus rien d'étonnant à n'avoir trouvé aucun bon chasseur parmi les 20 candidats précédents. ♧

Remarque (GV'). En fait, une façon mathématiquement plus solide de traiter le cas de l'exemple ci-dessus (où tous les candidats passés étaient de mauvais chasseurs) aurait consisté à utiliser un argument « hyper-bayésien » : dans ce cas, on considère que la priore elle-même a été tirée aléatoirement selon une « hyper-priore », autrement dit une distribution de probabilité *sur les distributions de probabilité sur* $\boldsymbol{\Theta}$, qu'on aurait choisie en fonction d'un argument neutre (voir le paragraphe suivant), puis ajustée au vu des résultats des 20 candidats précédents en vertu de ce que nous avons vu à la § 8.2. (Il faut ensuite prendre l'espérance pour transformer la postériore sur la priore en *une* priore effective). On trouverait alors, dans le cas où tous les chasseurs précédents étaient mauvais, une probabilité *à priori* pour que le chasseur suivant soit bon de 2,4 %. De manière intéressante, si on applique le même argument dans le cas où 4 candidats sur 20 se sont avérés bons dans le passé, on arrive par l'argument hyper-bayésien à trouver qu'une priore raisonnable doit donner à notre nouveau candidat une probabilité de 21,4 % : si cela ne change pas grand-chose au schmilblick (et montre donc que notre proposition de priore de l'exemple (GP') était pertinente), on voit cependant que nous avons effectivement *équilibré* la priore (au sens où nous avons considéré que le probabilité d'être un bon chasseur était un peu moins faible que ce suggéraient les données) par rapport à ce qu'aurait suggéré une confiance trop aveugle

en nos données passées : ainsi, l'argument hyper-bayésien rejoint les conclusions de la remarque (GS') !

Néanmoins, les approches hyper-bayésiennes dépassent le cadre de ce cours, et nous ne les aborderons donc pas davantage ici. ♣

! *Remarque* (GW'). En fait, l'exemple (GU') ci-dessus illustre une idée qui est un peu moins générale que celle exprimée par la remarque (GS'), mais qui a l'avantage d'être plus facile à appréhender et à vérifier :

> La priore doit donner une masse non nulle à *toutes* les zones de l'espace du paramètre caché ! (À moins, s'entend, qu'il n'existe un argument pour éliminer certaines zones avec une certitude *absolue*).

En termes plus formels, cela revient à dire (en utilisant les notations génériques) que, lorsque $\boldsymbol{\Theta}$ est muni d'une topologie naturelle, notre priore doit donner une masse strictement positive à n'importe quel ouvert non vide de $\boldsymbol{\Theta}$. ♣

15.5 Priores non informatives

Voyons enfin le cas où on n'a absolument *aucune* information sur ce que peut bien pouvoir valoir $\boldsymbol{\theta}$ (pas même par expertise) ; auquel cas le choix le plus raisonnable est manifestement de choisir une priore qui donne « la même chance à chaque possibilité ». Typiquement, si $\boldsymbol{\Theta}$ est un espace fini, on choisira comme loi à priori pour $\boldsymbol{\theta}$ la loi uniforme sur $\boldsymbol{\Theta}$.

Exemple (GX'). Dans le cas du chasseur, où $\boldsymbol{\Theta} =]0,1[$, un choix raisonnable de priore non informative serait ainsi de prendre $\mathrm{Loi}(\boldsymbol{\theta}) = \mathrm{Unif}^{\mathrm{me}}(0,1)$. ♣

Exemple (GY'). Dans le modèle du pédagogue, demandons-nous quelle loi à priori on souhaite mettre sur le paramètre de dispersion $\boldsymbol{\sigma}$, à valeurs dans $\mathbb{R}_+^*$. Il pourrait être tentant de prendre la priore (impropre) proportionnelle à la mesure de Lebesgue sur $\mathbb{R}_+^*$; néanmoins, quand on y réfléchit bien, ce n'est pas si satisfaisant que cela. En effet, une priore uniforme sur $\mathbb{R}_+^*$ pour $\boldsymbol{\sigma}$ reviendrait à dire, par exemple, qu'on considère qu'il est à priori deux fois plus probable d'avoir $\boldsymbol{\sigma}$ entre 100 et 110 que d'avoir $\boldsymbol{\sigma}$ inférieur à 5 ; or, on a plutôt envie de dire que c'est le second évènement qui devrait être plus probable à priori : car il y a des façons très différentes les unes par rapport aux autres d'avor une dispersion inférieure à 5 (une dispersion de 4 n'a rien à voir avec une dispersion de 0,4 !), alors que le premier évènement nuos donne sur $\boldsymbol{\sigma}$ une information qu'on perçoit intuitivement comme beaucoup plus spécifique... En fait, quand on y réfléchit mieux, on réalise qu'il est intuitivement plus satisfaisant de vouloir que ce soit l'*ordre de grandeur* de $\boldsymbol{\sigma}$ qui soit uniforme : autrement dit, de donner une densité à priori uniforme à $\log \boldsymbol{\sigma}$, plutôt qu'à $\boldsymbol{\sigma}$ lui-même. Si on fait cela, on dira en substance qu'on considère comme à priori aussi probable que $\boldsymbol{\sigma}$ soit entre 2 et 4, qu'ente 10 et 20, qu'entre 50 et 100 : et cette fois-ci, cela colle beaucoup mieux avec notre diée intuitive de « n'avoir aucune information sur ce que vaut le paramètre de dispersion $\boldsymbol{\sigma}$ » ! ◡̈

Dès lors, si on prend pour loi à priori sur $\ln \boldsymbol{\sigma}$ la mesure (impropre) proportionnelle à la mesure de Lebesgue sur $\mathbb{R}$, on en déduit la loi à priori sur $\boldsymbol{\sigma}$ lui-même en prenant (toujours à constante multiplicative impropre près) la mesure-image de ladite mesure de Lebesgue par la fonction exponentielle : et les calculs nous donnent alors, pour $\boldsymbol{\sigma}$, une priore (impropre) caractérisée par

$$\mathbb{P}(\boldsymbol{\sigma} \in \mathrm{d}\sigma) \propto \sigma^{-1} \, \mathrm{vol}_1(\mathrm{d}\sigma). \qquad (\mathtt{GZ'})$$

♧

Remarque (HA'). Comme le montre l'exemple ci-dessus, le problème de « prendre une priore uniforme » n'est pas si simple qu'il n'en a l'air, car lorsqu'on a affaire à des distributions de probabilité non discrètes, la notion d'« uniformité » n'est pas stable par changement de variables... Voyons une autre situation où c'est encore plus flagrant (et avec des priores propres). Imaginons que, dans le club de chasse des Canardeurs (concurrent de celui du Bouchonnois), on utilise une échelle pour mesurer la qualité d'un chasseur : la « qualité » (au sens des Canardeurs) d'un tireur, noté cette fois-ci $\boldsymbol{\tau}$, étant sa probabilité de réussite *après deux essais*. Il est alors évident que la « qualité » $\boldsymbol{\tau}$ et la « fiabilité » $\boldsymbol{\theta}$ sont reliées par le relation bijective $\boldsymbol{\tau} \coloneqq 2\boldsymbol{\theta} - \boldsymbol{\theta}^2$; de sorte que, du point de vue mathématique, décrire le niveau du tireur par $\boldsymbol{\theta}$ ou par $\boldsymbol{\tau}$ est rigoureusement équivalent (on parle de *reparamétrage*). Cependant, mettre une loi uniforme sur $\boldsymbol{\tau}$ n'est *pas* la même chose que mettre une loi uniforme sur $\boldsymbol{\theta}$... ♧

Comme le montre la remarque ci-dessus, la simple idée d'« uniformité » n'est pas toujours suffisante pour déterminer une priore non informative de façon canonique. Alors, peut-on faire mieux... ? Nous n'entrerons pas dans le détail de cette question dans le cadre de ce cours ; néanmoins je présente ci-dessous en digression, à titre de complément, deux grandes stratégies qu'on peut suivre pour tenter d'être neutre indépendamment du paramétrage.

Définition (HB') (Priore de Haar)**.** Une stratégie pertinente pour déterminer une priore non informative consiste à demander que la distribution à priori sur $\boldsymbol{\theta}$ soit invariante par un certain nombre de transformations "naturelles" sur l'espace $\boldsymbol{\Theta}$: par exemple, dans le cas où $\boldsymbol{\Theta}$ est le cercle unité (si par exemple $\boldsymbol{\theta}$ représente la phase d'un phénomène sinusoïdal), la loi uniforme sur le cercle unité est la seule mesure qui soit invariante par toutes les rotations. Une distribution de probabilité caractérisée par de telles propriétés d'invariance est qualifié de *mesure de Haar*.

Dans le cadre de ce cours, je ne vous demanderai pas de déterminer des priores en tant que mesures de Haar ; en revanche, il se peut que je vous demande de vérifier qu'une priore donnée par une certaine formule vérifie certaines propriétés d'invariance, et que le cas échéant, vous sachiez y voir un signe que cette priore vérifie une certaine forme de neutralité, puisqu'elle est indépendante du choix qu'on fait entre différentes façons naturelles de paramétrer $\boldsymbol{\Theta}$ (la façon de passer d'un paramétrage à l'autre correspondant, justement, aux transformations « naturelles » évoquées ci-dessus). ♡

Exemple (HC'). En fait, la priore sur $\boldsymbol{\sigma}$ proposée dans l'exemple (GY') *supra* correspond à une priore de Haar. En effet, un groupe de transformations agissant naturellement sur $\mathbb{R}_+^*$ est celui des changements d'échelle : $\sigma \mapsto k\sigma$ pour $k \in \mathbb{R}_+^*$. Or, les mesures proportionnelles à $\sigma^{-1}\,\mathrm{vol}(\mathrm{d}\sigma)$ sont les seules qui soient invariantes par ces transformations, de sorte que la mesure de probabilité (impropre) qui s'en déduit est un choix naturel de priore non informative pour $\boldsymbol{\sigma}$. ♧

Définition (HD') (Priore de Jeffreys)**.** La *priore de Jeffreys* a une définition plus subtile, liée à la théorie de l'information, et qu'il serait au-delà de vos connaissance actuelles de définir rigoureusement : il s'agit de dire, en substance, que, pour tout $\theta \in \boldsymbol{\Theta}$, la probabilité que $\boldsymbol{\theta}$ soit difficile à distinguer de θ doit être identique.

La priore de Jeffreys présente l'avantage de pouvoir toujours être définie, et ce, de façon dénuée d'ambiguïté, dès qu'on est face à un modèle de statistique paramétrique. ♡

Exemple (HE'). Dans le modèle du chasseur, le calcul de la priore de Jeffreys montre qu'il s'agit de la loi de l'arcsinus ! C'est notamment pour cette raison que j'en ai fait mon choix de priore par défaut dans le cadre de ce polycopié. (Et aussi parce que ce choix présentait un certain nombre d'avantages pédagogiques ☺). ♧

Remarque (HF'). C'est principalement lorsqu'on recourt à des priores [*] non informatives qu'on se retrouve confronté à des priores impropres. (Peu importe la méthode non informative retenue, d'ailleurs : aussi bien la stratégie uniforme que celles de Haar et de Jeffreys produisent facilement des mesures de masse totale infinie !). Ce qui n'est pas un cas si rare que cela : car comme nous l'avons fait remarquer plus haut, il vaut mieux éviter de se fier trop aveuglément à son avis sur la valeur de $\boldsymbol{\theta}$, et donc utiliser (au moins partiellement) une forme d'approche non informative dans son raisonnement... ♧

Avant de clore ce chapitre, mentionnons un concept très intéressant émergeant naturellement en théorie de l'information, et qui permet de poser une priore "universelle" sur absolument tout type de données. [Nota : ci-dessous je ne définis pas certains des termes intervenant (comme « machine de Turing universelle »), dans la mesure où mon but est simplement de vous donner un aperçu des idées exposées, pas de vous faire un cours de théorie de l'information]. Voici ce que j'aimerais dire ici à ce sujet :

Définition (HG'). La *distribution de probabilité de Solomonoff* est la distribution de probabilité sur $\{0,1\}^\omega$ (autrement dit l'ensemble des suites infinies de bits) qu'on obtient à partir d'une machine de Turing universelle qui n'efface jamais sa sortie, lorsque la bande d'entrée et la bande de sortie sont initialisées à une suite de bits i.u.d. Informellement, cela correspond à la loi de la sortie d'un programme informatique aléatoire (écrit dans un langage suffisamment général).

"La" distribution de Solomonoff est en fait dépendante du choix de machine de Turing universelle opéré ; mais on peut montrer que tous les choix possibles amènent à des probabilités ayant des densités bornées les unes par rapport aux autres, de sorte qu'on peut considérer qu'on a bien défini « une » distribution de probabilité "à densité relative bornée près". ♡

Définition (HH'). L'*inférence de Solomonoff* consiste à encoder l'ensemble des données passées et futures sous la forme d'une suite de bits (les données passées étant autodélimitées et précédant les données futures), et à faire de l'inférence prédictive sur des données futures en considérant que notre jeu de données a été généré à priori selon la distribution de probabilité de Solomonoff. ♡

Point (HI'). Il existe des arguments philosophiques forts pour considérer l'inférence de Solomonoff comme la "bonne" façon de faire de la statistique bayésienne en l'absence d'information spécifique à priori. On peut montrer notamment montrer que, dans le cadre des modèles d'échantillonnage paramétriques, l'approche de Solomonoff re-donne les mêmes conclusions qu'on aurait obtenues à partir de technques d'analyse statistique "classique". En outre, cela permet d'obtenir des réponses solidement justifiées à des questions comme « si, depuis $n \gg 1$ jours que l'humanité existe, le Soleil s'est levé tous les jours, quelle est la probabilité qu'il se lève demain ? » (en supposant ici qu'on n'ait aucune notion d'astronomie par ailleurs), où la théorie de Solomonoff nous dit en l'occurrence que la réponse correcte est de l'ordre de $1/n$. De même, dans le cas du modèle du chasseur, la théorie de Solomonoff fournit un argument en faveur de la priore impropre Bêta$(0,0)$ par rapport à d'autres choix possibles. Dans les problématiques de *sélection de modèle* que vous verrez notamment en analyse de données et en intelligence artifcielle, c'est encore les raisonnements à la Solomonoff qui sont cachés derrière le critère de sélection BIC (pour « *Bayesian information criterion* »). ♧

Remarque (HJ'). On peut nénamoins démontrer que la probabilité de Solomonoff est rigoureusement impossible à calculer (et il s'agit là d'une impossibilité *théorique*, pas seulement pratique : *aucun* algorithme, si compliqué soit-il, ne permet de calculer la probabilité de Solomonoff ! — C'est lié aux problématiques d'incomplétude en logique mathématique) : il s'agit donc d'un outil *philosophique* plus que pratique, qui permettra tout au plus d'avoir des heuristiques intelligentes (comme le critère BIC) en fonction de certaines options plutôt que d'autres. L'idée est néanmoins remarquable sur le plan conceptuel, car elle indique que, même en l'absence totale d'information, il existe malgré tout une priore sur nos données qui aboutira à des

[*]. Ou éventuellement à des hyper-priores, confer remarque (GV').

conclusions pertinentes : grâce à la priore de Solomonoff, on peut donc utiliser l'analyse bayséienne comme une théorie *complètement générale* de l'idée de connaissance, nous permettant (sur le papier !) de répondre à *toute* question du type : « quelle croyance devons-nous tirer de telles données au vu de telles informations dont nous disposions précédemment [†] ? » ! ♧

[†]. Les « informations dont nous disposions précédemment étant vues », dans le cadre de la théorie de Solomonoff, comme des données passées encore plus anciennes que ce que le modèle appelle « observation passée » : en effet, grâce à son encodage binaire, la théorie de Solomonoff n'a aucune difficulté à agréger des informations de typologies à première vue différentes.

Chapitre 16

Fonctions d'utilité

16.1 Illustration du concept

Dans les parties II et III du cours, nous avons expliqué que, lorsqu'on cherche à évaluer quantitativement la pertinence d'une procédure d'estimation, de prévision ou plus généralement de décision, on le faisait en passant par une *fonction de perte*, qui nous dit à quel point une décision s'avère mauvaise dans une situation donnée (la description de la situation en question étant reflétée, typiquement, par la valeur d'une certaines quantité d'intérêt), et que le but était alors de minimiser l'espérance de valeur prise par ladite perte.

Dans ce chapitre, je voudrais développer plus avant cette notion de perte, dans la mesure où il me semble qu'il s'agit là d'une idée dont la compréhension est absolument essentielle *pour un ingénieur*. Du strict point de vue de la *théorie statistique*, cependant, comprendre précisément le sens des fonctions d'utilité et la façon de les employer en contexte industriel n'est pas central : c'est pourquoi, afin de ne pas trop se disperser dans les enjeux de ce cours, je n'ai indiqué aucun élément de cette annexe comme étant à retenir. Pour autant, je vous conseille vivement d'essayer de lire et de comprendre ce qui suit, parce que ce sera intéressant et important pour vous *de manière générale* ; mais en revanche, lors de vos révisions pré-examen, vous pourrez vous contenter de regarder le passage qui suit en diagonale ! ⌣

Pour commencer, rappelons la définition de la notion d'utilité (ou de perte) :

Définition (HK'). Lorsqu'on est amené à prendre une décision, représentée mathématiquement comme un élément d'un certain espace $\mathscr{D}$, et que l'impact de cette décision dépend de la valeur prise par une certaine quantité d'intérêt, représentée mathématiquement comme un élément d'un certain espace $\mathscr{G}$, pour comparer deux prises de décisions en tenant compte de l'incertitude sur la quantité d'intérêt, l'outil mathématique le plus classique consiste à introduire une *fonction d'utilité* $u : \mathscr{G} \times \mathscr{D} \to \mathbb{R}$, dont l'interprétation est la suivante :

(i) La valeur $u(g, \hat{d})$ nous dit à quel point est-ce que la décision $\hat{d}$ amènera des conséquences bénéfiques si la quantité d'intérêt s'avère valoir g : plus cette valeur est grande, plus les conséquences sont bénéfiques à nos yeux ;

(ii) En outre, entre deux situations incertaines où l'utilité que nous obtiendrons effectivement est aléatoire, notre préférence va à celle où *l'espérance* de l'utilité est la meilleure (et nous sommes indifférents entre les deux situations lorsque l'espérance est identique).

L'opposé d'une fonction d'utilité est qualifié de *fonction de perte* (traditionnellement notée 'ℓ') : c'est exactement le même principe, sauf que cette fois-ci nous avons une préférence pour les valeurs les plus *petites* ![*] ♡

Commentaire (HL'). Détaillons un peu le point (ii) de la définition ci-dessus. Imaginons que je vous donne le choix entre les deux options suivantes :

Option A Je vous offre un cadeau aléatoire, déterminé par le tirage d'une pièce de monnaie non truquée : ce cadeau sera soit un vélo électrique si la pièce tombe sur pile, soit un jeu de quilles finlandaises si la pièce tombe.

Option B Je vous offre un lave-vaisselle, quoi qu'il arrive. (Dans cette option je lance *quand même* la pièce, mais on n'exploitera pas son résultat [†]).

Selon le choix que vous aurez fait et le résultat de la pièce, vous gagnerez soit un jeu de quilles, soit un lave-vaisselle, soit un vélo : si nous appelons resp. u_0, u_1, u_2 les bénéfices que le gain de chacun de ces trois cadeaux représente à vos yeux, la fonction d'utilité correspondant à cette situation est alors définie sur $\{\text{pile}, \text{face}\}\times\{\mathsf{A}, \mathsf{B}\}$ par

$$\begin{cases} u(\text{pile}, \mathsf{A}) = u_2 \,; \\ u(\text{face}, \mathsf{A}) = u_0 \,; \\ u(\text{pile}, \mathsf{B}) = u_1 \,; \\ u(\text{face}, \mathsf{B}) = u_1. \end{cases} \qquad (\mathsf{HM}')$$

Notez bien ici que ces valeurs u_0, u_1, u_2 reflètent vos préférences *personnelles*, avec la subjectivité que cela comporte : ainsi, s'il y a bien du sens à dire qu'un certain triplet (u_0, u_1, u_2) modélise correctement vos préférences (au sens où il conduit effectivement aux décisions que vous prendriez dans une telle expérience), ce n'est pas pour autant qu'un triplet (u_0, u_1, u_2) qui serait correct *pour vous* serait pour autant correct pour votre voisine, et vice-versa !

Essayons maintenant de comprendre plus précisément ce que les valeurs u_0, u_1, u_2 disent sur vos préférences. Déjà, si (comme on peut raisonnablement le supposer) vous préférez le vélo au lave-vaisselle, et le lave-vaisselle aux quilles, cela devra se refléter par le fait que $u_2 > u_1 > u_0$. Mais à part cela, l'écart entre u_1 et u_2 doit-il être plus petit, plus grand, ou égal à celui entre u_0 et u_1 ? Sur ce point, votre réponse devra refléter celle des deux options A ou B ayant votre préférence, la première ayant pour espérance u_1 tandis que la seconde a pour espérance $(u_0 + u_2)/2$! Ainsi, si vous préférez l'option A (autrement dit si, à vos yeux, avoir une chance sur deux de voir votre lave-vaisselle se transformer en vélo électrique suffit à compenser le fait d'avoir un risque sur deux de le voir se transformer en quilles), cela doit se refléter par le fait que $(u_0 + u_2)/2 > u_1$, autrement dit que $u_2 - u_1 > u_1 - u_0$: ce qui traduit le fait que, à vos yeux, l'écart entre le bénéfice du vélo et celui du lave-vaisselle est supérieur à l'écart entre le bénéfice du lave-vaisselle et celui du jeu de quilles. Si à l'inverse vous préférez l'option B, cela signifie que vous attribuez (dans votre

[*]. En fait, dans ce cours, comme dans la plupart des ouvrages de statistique, nous utiliserons le formalisme de le fonction de perte plutôt que celui de la fonction d'utilité : mais j'ai trouvé que la définition centrale était plus aisée à saisir lorsqu'on l'écrivait avec le convention de l'utilité.

[†]. Ce protocole étrange sert simplement à unifier la présentation mathématique entre les deux options. En pratique bien sûr, si vous choisissez l'option B, je ne lancerai pas *réellement* la pièce : le résultat de la pièce sera alors simplement *virtuel* (ce qui n'a aucune incidence, puisqu'il n'intervient pas dans le cadeau que vous obtenez !), et ne servira qu'à simplifier les calculs de comparaison entre les deux options.

système de préférences personnel) moins d'écart entre le vélo et le lave-vaisselle qu'entre le lave-vaisselle et les quilles, ce qui doit donc se traduire par le fait que $u_2 - u_1 < u_1 - u_0$: et cela correspond bien, au niveau des espérances, à la propriété que $(u_0 + u_2)/2 < u_1$!

Nous voyons ainsi que votre préférence entre les options A et B reflète lequel des deux écarts $u_2 - u_1$ et $u_1 - u_0$ doit être le plus grand (lorsque u_0, u_1, u_2 correspondent aux utilités pour votre cas personnel, s'entend). Mais on peut même aller plus loin, en interprétant le sens à donner à la valeur *précise* du ratio $(u_2 - u_1)/(u_1 - u_0)$! Pour ce faire, supposons que ma pièce de monnaie soit déséquilibrée et ait en fait une certaine probabilité $p \in]0, 1[$ de tomber sur pile, probabilité que vous comme moi connaissons. (Ce n'est donc pas de la triche, juste que le tirage au sort ne se fait pas de façon uniforme). Il est clair alors que votre préférence entre les options A et B va dépendre de la valeur de p : pour p tendant vers de 0, l'option A devient essentiellement équivalente à gagner un jeu de quilles à coup sûr, donc vous préférez l'option B (rappelons que nous avons supposé, pour les besoins de cet exemple, que vous préfériez le lave-vaisselle au jeu de quilles), tandis que pour p tendant vers 1, l'option A tend vers le gain du vélo électrique à coup sûr, auquel cas vous la préférez à l'option B. En outre, plus p est élevé, plus l'option A devient intéressante à vos yeux : il y aura donc un certain seuil "critique" p_c en-deçà duquel vous préfèrerez B et au-delà duquel vous préfèrerez A, et tel que vous serez indifférent·e si la probabilité de pile vaut exactement p_c. Cette indifférence au seuil p_c soit se traduire par le fait que, pour cette valeur de p, les deux utilités doivent devenir identiques :

$$p_c u_2 + (1 - p_c) u_0 = u_1, \tag{HN'}$$

ce qui est équivalent à

$$\frac{u_2 - u_1}{u_1 - u_0} = \frac{1}{p_c} - 1. \tag{HO'}$$

Ainsi, le ratio entre $u_2 - u_1$ et $u_1 - u_0$ traduit le seuil auquel vous basculez entre l'option B et l'option A : lorsque ce seuil est petit ($p_c \ll 1$), même une faible probabilité de gagner le vélo suffit à vos yeux à compenser le fait de risquer de perdre le lave-vaisselle pour les quilles, et cela se voit au fait que $u_2 - u_1 \gg u_1 - u_0$ (vous mettez beaucoup plus d'écart entre le lave-vaisselle et le vélo qu'entre les quilles et le lave-vaisselle), tandis que lorsque ce seuil est grand $1 - p_c \ll 1$, même une faible probabilité de perdre le lave-vaisselle vous décourage de tenter de gagner le vélo, signifiant alors que vous placez beaucoup plus d'écart entre les quilles et le lave-vaisselle qu'entre le lave-vaisselle et le vélo. ♧

16.2 Considérations sur la notion d'utilité

Remarque (HP'). Comme le commentaire (HL') a dû le rendre clair, notez bien que la notion d'utilité, du point de vue industriel, est un concept fondamentalement *subjectif* : elle formalise *votre* niveau de satisfaction relatif concernant telle situation par rapport à telle autre. Ainsi, si une entreprise décide d'intégrer les notions de développement durable parmi ses objectifs, la façon dont cela se traduira mathématiquement sera au niveau de la fonction d'utilité que l'entreprise cherche à maximiser (ou plus exactement, dont elle cherche à maximiser l'espérance) : cette fonction se mettant alors à refléter une notion de succès *globale* incluant le résultat financier et les aspects environnementaux, sociaux et éthiques !... ♧

Remarque (HQ'). Entre autres notions subjectives, la notion d'utilité permet de modéliser l'*aversion au risque* (ou parfois la propension au risque). Imaginez par exemple que vous dirigiez une entreprise dont la capital est d'environ 10 M€. Une ingénieure vient vous proposer une idée risquée : si vous la mettez en pratique, il y a une chance sur deux que cela fasse gagner 200 k€ à l'entreprise, mais une chance sur deux que cela lui fasse perdre 100 k€... Que décidez-vous ? Normalement, vous allez considérer que, *en moyenne*, l'idée en question est gagnante, et que cela mérite donc qu'on la mette en œuvre ! Mais maintenant, imaginez que cette même ingénieure vienne vous proposer une idée qui a une chance sur deux de faire gagner 18 M€ à l'entreprise, et une change sur deux de lui faire perdre 9 M€... L'espérance de gain est toujours positive, et même 90 fois plus grande que dans la situation précédente ! Pourtant, cette fois-ci vous allez estimer que le risque ne mérite pas d'être couru : en effet, vu le capital de l'entreprise, si l'idée échoue, on ne va pas "juste" perdre de l'argent, mais se retrouver carrément en situation de faillite ou de quasi-faillite, ce qui aura des conséquences catastrophiques dont il sera très difficile de sortir, même à long terme ! Alors qu'en cas de succès, certes, on triple la valeur de l'entreprise, mais d'un point de vue qualitatif, ce n'est pas un changement si radical que cela...

La notion d'utilité permet de fournir un modèle justifiant cette décision mathématiquement : en effet, ce que vous voulez maximiser, ce n'est pas l'espérance du résultat financier *en tant que tel*, mais l'espérance de *l'utilité* de ce résultat ! Et ici le fonction qui, à un résultat fiancier donné, associera son utilité sera typiquement *concave*, traduisant une *aversion au risque* : entre deux solutions ayant la même espérance financière "brute", votre préférence va à celle qui implique le moins d'aléa ! Pour l'exemple que j'ai donné ci-dessus, une fonction d'utilité qui modélise bien la décision prise consiste à attribuer à un gain financier g une utilité de

$$\ln\left(1+\frac{g}{10\ \text{M€}}\right). \tag{HR'}$$

On voit alors que ne rien faire a une utilité nulle, gagner 200 k€ a une utilité de $+0{,}019\,8$, perdre 100 k€ a une utilité de $-0{,}010\,1$, gagner 18 M€ a une utilité de $+1{,}03$, et perdre 9 M€ a une utilité de $-2{,}30$: on vérifie bien que, en termes d'espérance, appliquer l'idée de l'ingénieure dans la première situation (gain d'utilité moyen : $+0{,}004\,8$) par rapport à ne rien faire, mais que dans la seconde situation, le gain d'utilité moyen ($-0{,}64$) est en fait une lourde perte !! ♧

Remarque (HS'). À l'inverse, notez qu'il peut exister des situations de propension au risque : imaginez ainsi un élève de Mines Nancy en grande difficulté sur ses résultats, qui risque très fortement de se faire exclure de l'École, à moins qu'il n'obtienne au moins 15/20 à l'examen de statistique. Cet élève va évidemment faire son maximum pour réussir, mais se pose la question de savoir s'il doit réviser "normalement" en travaillant tous les chapitres du cours, ou faire des "impasses" et privilégier certains sujets très précis en espérant qu'ils tombent à l'examen... Dans le premier cas, il considère qu'il devrait obtenir typiquement environ 12/20 (à plus ou moins un point près) [‡] ; alors que dans le second cas, il a une chance sur deux d'obtenir environ 16/20 si le sujet qu'il a travaillé en priorité tombe à l'examen, et une chance sur deux d'obtenir environ 6/20 si ce sujet ne tombe pas ! Ici, faire l'impasse est une situation *à la fois* plus défavorable en moyenne "brute" et plus risquée ; pourtant, il est clair que c'est dans l'intérêt de l'élève de suivre cette stratégie : car, foutu pour foutu, qu'est-ce que ça changera pour lui d'avoir eu 12 ou 6 s'il doit être exclu dans les deux cas... ?! Donc, dans une telle situation (rare en pratique, néanmoins ☡), la fonction d'utilité reflèterait une propension au risque. ♧

[‡]. Ah ben oui, hein, on a dit que c'était un élève en difficulté : donc, même en travaillant dur, il va avoir du mal à obtenir un résultat excellent !...

Remarque (HT'). Lorsqu'on introduit une fonction d'utilité, la valeur zéro pour l'utilité n'a aucune signification particulière, et il n'y a pas non plus d'unité de mesure pour quantifier les utilités : ainsi, du point de vue "physique", transformer les valeurs d'une fonction d'utilité selon une bijection affine croissante conduit exactement à la *même* modélisation — et, en particulier, les conclusions qu'on tirera quant aux décisions qu'il convient de prendre seront *exactement* les mêmes ! En termes d'utilité, tout ce qui a du sens, c'est :

— Le sens de comparaison entre deux situations : telle situation est-elle préférable à telle autre ?

— Entre trois situations $\mathsf{A}, \mathsf{B}, \mathsf{C}$ pour lesquelles on préfère A à B et B à C, le *ratio* entre l'ampleur de notre préférence pour A sur B et l'ampleur de notre préférence pour B sur C [§].

♧

Remarque (HU'). Dans le contexte statistique, on peut aussi observer que cela ne change rien, du point de vue des décisions à prendre, de remplacer une fonction d'utilité $u(g, \hat{d})$ par une variante $u(g, \hat{d}) + c(g)$ qu'on a translatée selon un terme ne dépendant que de g. Cela expliquera pourquoi, lorsqu'on s'intéressera aux fonctions de perte concernant l'estimation, on pourra toujours se permettre de supposer qu'une reconstitution exacte correspond à une perte de zéro. ♧

Remarque (HV'). Cette modélisation de la prise de décision en contexte incertain par le fait qu'on souhaite maximiser l'espérance d'une utilité est clairement intéressante ; mais on peut se demander néanmoins s'il n'existerait pas des situations de prise de décision qu'on ne pourrait pas modéliser ainsi... Le *théorème de von Neumann-Morgenstern* énonce que non : sous certaines hypothèses très raisonnables exprimant des contraintes de rationalité minimales sur un processus de prise de décision, on montre que nos décisions correspondront nécessairement à optimiser l'espérance d'une certaine fonction d'utilité. C'est pourquoi, en mathématiques et en économie, on utilise quasiment toujours ce formalisme pour modéliser les questions de décision optimale. ♧

[§]. Du point de vue philosophique, cela entraine des considérations particulièrement intéressantes. À cause de la propriété d'invariance que nous venons d'expliquer, si, demain, vous vous mettiez à tout ressentir de façon deux fois plus intense, aussi bien la joie que la tristesse, en toute chose, alors votre façon de vous comporter resterait *exactement* la même. Pourtant, du point de vue d'une personne souhaitant suivre une morale « utilitariste » (c-à-d. une morale dont le but est de maximiser le bonheur total de l'humanité), l'exacerbation de vos ressentis devrait justifier que vos besoins soient satisfaits en priorité par rapport à ceux des personnes dont la fonction d'utilité varie sur des plages moins larges... Mais comment une telle personne pourra-t-elle savoir que vos ressentis sont exacerbés, puisque cette exacerbation est *intrinsèquement* impossible à mesurer ? Et d'ailleurs, est-il juste de fonder une philosophie morale sur un concept (à savoir, le bonheur total de l'humanité) qu'il est ainsi fondamentalement impossible de mesurer ?! C'est là une objection philosophique très importante à l'utilitarisme, que Robert Nozick a popularisée avec son expérience de pensée du « monstre utilitariste », une personne dont les utilités seraient si énormément exacerbées que cela justifierait que tout le reste de l'humanité se sacrifie pour le moindre de ses désirs... Il est tentant, de ce fait, de vouloir refuser de comparer l'utilité entre deux personnes différentes. Mais pourtant, lorsqu'on doit prendre une décision favorisant une personne sur une autre (exemple : « si je fais des crêpes, ma fille sera contente et mon garçon déçu ; mais si je fais des gaufres, ce sera l'inverse ; or je n'ai pas le temps de faire les deux à la fois : que dois-je donc choisir ? »), notre souhait de justice va bel et bien nous amener à comparer les utilités entre des personnes distinctes ! (Ainsi, si mon garçon n'a qu'une très légère préférence pour les gaufres sur les crêpes, alors que ma fille adore les crêpes et déteste les gaufres, il semble logique que je prépare des crêpes, dans l'intérêt collectif !). Est-ce que cela est possible ? Est-ce que cela a seulement du sens ? Est-ce que cela ne risque pas de nous faire préférer injustement ceux qui ressentent les émotions plus violemment ?... La notion mathématique d'utilité, sous son apparence anodine, amène de grandes questions philosophiques ! ♨

Remarque (HW'). Dans le cadre de ce cours, il ne vous sera jamais demandé de proposer vous-mêmes des fonctions d'utilité ; tout au plus de justifier le choix d'une fonction d'utilité qui vous aura été donnée par l'énoncé. ♧

16.3 Un exemple en contexte industriel

Pour finir ce chapitre, présentons un exemple de contexte industriel dans lequel il y a une fonction de perte qui s'impose assez clairement (ce qui est souvent le cas en pratique) :

Exemple (HX'). Considérons un gisement contenons une quantité $\boldsymbol{\varphi}$ de minerai. Après avoir entrepris une campagne de prospection qui lui a donné une idée (entachée néanmoins d'incertitude) sur la valeur de $\boldsymbol{\varphi}$, une entreprise doit prendre la décision, notée $\boldsymbol{\delta}$, d'exploiter ou non le gisement (la valeur VRAI signifiant qu'on exploite le gisement, et la valeur FAUX qu'on ne l'exploite pas). On suppose ici qu'exploiter le gisement engendre des frais f fixes, et rapporte a par unité de minerai extrait, les valeurs de f et a étant connues ; on suppose également que, une fois l'exploitation du gisement entamée, même si le gisement se révèle décevant, il sera toujours préférable poursuivre l'exploitation jusqu'au bout, dans la mesure où c'est amont de l'exploitation qu'on engage l'essentiel des frais. La question est : dans ce contexte, quelle fonction d'utilité (ou de perte) prendre par rapport à la quantité d'intérêt $\boldsymbol{\varphi}$ et à la décision $\boldsymbol{\delta}$?

Ici, il semble assez raisonnable d'associer la notion d'utilité au bénéfice financier réalisé par l'entreprise. Si on n'exploite pas le gisement, il n'y a ni frais ni perte ; et si on l'exploite, le bénéfice final vaura $a\varphi_{\checkmark} - f$ (où $\varphi_{\checkmark}$ désigne la quantité véritable de minerai contenue dans le gisement) : on prendra donc

$$u(\varphi, \delta) = \mathbf{1}_{\delta}(a\varphi - f). \tag{HY'}$$

Voyons maintenant comment reformuler cette fonction d'utilité dans le formalisme de la fonction de perte normalisée. De manière générale, la perte est l'opposée de l'utilité, soit $\mathbf{1}_{\delta}(f - a\varphi) =: \ell_{\text{brut}}(\varphi, \delta)$ en l'occurrence. Maintenant, normaliser la fonction de perte signifie qu'on va retrancher de celle-ci la valeur minimale qu'elle prend pour une quantité d'intérêt donnée (ce qui ne change pas la signification "physique" de la perte), de sorte que, à quantité d'intérêt donnée, le minimum de la fonction de perte renormalisée soit toujours 0.

Il est évident que, à valeur fixée de φ, le minimum de la perte est atteint pour $\delta = \text{FAUX}$ lorsque $\varphi < f/a$ (il n'est pas intéressant d'exploiter le gisement lorsque la quantité de minerai s'avère trop faible pour couvrir les frais), resp. pour $\delta = \text{VRAI}$ lorsque $\varphi \geqslant f/a$: le minimum valant alors respectivement 0 et $f - a\varphi$. On obtient donc la fonction de perte normalisée suivante :

$$\ell_{\text{norm}}(\varphi, \delta) = \ell_{\text{brut}}(\varphi, \delta) - \mathbf{1}_{\varphi \geqslant f/a}(f - a\varphi) = (\mathbf{1}_{\varphi \geqslant f/a} - \mathbf{1}_{\delta}) \cdot (a\varphi - f). \tag{HZ'}$$

Ce résultat peut se comprendre ainsi : lorsque $\delta = \{\varphi \geqslant f/a\}$, cela signifie que l'entreprise a pris la "bonne" décision [¶], et dans ce cas sa perte est nulle, ce qui

[¶]. Ici je veux dire que la décision apparait *rétrospectivement* comme la bonne, au sens où cette décision coïncide avec celle qu'on aurait dû prendre si on avait connu *à l'avance* la quantité de minerai ! Bien entendu, en pratique, il peut arriver qu'on prenne une décision *correctes* (au sens où il est logique de la prendre au vu des informations dont on dispose) qui s'avèrera rétrospectivement

signifie qu'on a fait aussi bien que possible. Lorsque $\varphi < f\,/\,a$ mais que $\delta = \mathtt{VRAI}$ (autrement dit, l'entreprise a pris la décision d'exploiter un gisement qui ne contenait pas assez de minerai pour voucrir les frais), la perte vaut $f - a\varphi$ et représente l'argent globalement perdu par l'entreprise (en prenant en compte ce que le minerai a tout de même rapporté) dans cette exploitation qu'il aurait mieux valu ne pas faire. Et lorsque $\varphi \geqslant f\,/\,a$ mais que $\delta = \mathtt{VRAI}$ (autrement dit, l'entreprise a renoncé à exploiter un gisement qui était en fait rentable), la perte vaut $a\varphi - f$: autrement dit, elle représente le *manque à gagner* par rapport à la situation où on aurait exploité le gisement !

Pour finir, on peut aussi définir une fonction de perte liée à l'*estimation* de la quantité de minerai contenue dans le jugement. Ci-dessus la fonction de perte que nous considérions s'entendait par rapport à une quantité d'intérêt et à une *décision*. Maintenant, imaginons qu'on cherche plutôt à estimer à quel point c'est grave de croire qu'il y a une quantité de minerai $\hat{\varphi}$ lorsque la quantité réelle est φ. (On cherche donc une fonction de perte où on compare deux quantités φ et $\hat{\varphi}$ qui vivent dans le même espace). En fait, cela se ramène à la situation précédente : on va juste dire que, si $\varphi < f\,/\,a$, cela signifie qu'on croit qu'il y a trop peu de minerai dans le gisement pour qu'il vaille la peine d'exploiter celui-ci : dans ce cas, l'entreprise prendra la décision de ne pas exploiter et la perte (normalisée) sera $\mathbf{1}_{\varphi \geqslant f\,/\,a}(a\varphi - f)$; et que si $\varphi \geqslant f\,/\,a$, à l'inverse, notre estimation conduira l'entreprise à exploiter le gisement, d'où une perte de $\mathbf{1}_{\varphi < f\,/\,a}(f - a\varphi)$. En résumé, si on note $\delta(\hat{\varphi}) \coloneqq \{\hat{\varphi} \geqslant f\,/\,a\}$ la décision d'exploitation qui découle logiquement de l'estimation $\hat{\varphi}$, notre perte vaut

$$\ell(\varphi, \hat{\varphi}) = (\mathbf{1}_{\delta(\hat{\varphi})} - \mathbf{1}_{\varphi \geqslant f\,/\,a}) \cdot (f - a\varphi) = \begin{cases} \mathbf{1}_{\varphi \geqslant f\,/\,a}(a\varphi - f) & \text{si } \hat{\varphi} < f\,/\,a\,; \\ \mathbf{1}_{\varphi < f\,/\,a}(f - a\varphi) & \text{si } \hat{\varphi} \geqslant f\,/\,a. \end{cases} \tag{IA'}$$

En l'occurrence, cela peut se ré-écrire plus élégamment à l'aide des fonctions « partie positive » $\bullet^+$ et « partie négative » $\bullet^-$:

$$\ell(\varphi, \hat{\varphi}) = \begin{cases} (a\varphi - f)^+ & \text{si } \hat{\varphi} < f\,/\,a\,; \\ (a\varphi - f)^- & \text{si } \hat{\varphi} > f\,/\,a. \end{cases} \tag{IB'}$$

♧

Remarque (IC'). Ici j'ai choisi d'assimiler la fonction de perte à la perte financière de l'enteprise en tant que telle : autrement dit, je suppose que passer d'un bénéfice de 10 M€ à un bénéfice de 12 M€ est exactement aussi intéressant que passer d'un déficit de 1 M€ à un bénéfice de 1 M€... En pratique, dans bien des cas l'intérêt véritable qu'on a à gagner de l'argent diminue à mesure qu'on en a gagné déjà beaucoup, alors qu'à l'inverse être ruiné est une catastrophe qu'il convient d'éviter absolument... Autrement dit, la véritable fonction d'utilité est plutôt une fonction concave du gain financier : en termes d'économie, on parle dans une telle situation d'*aversion au risque*[‖]. Cependant, lorsqu'on ne parle que de petits montants

"mauvaises" (au sens où, après coup, on se rend compte qu'une autre décision aurait conduit à un meilleur résultat, à cause d'éléments dont on ne disposait pas encore au moment où on a pris notre décision) : dans ce cas, ce n'est pas parce qu'on n'a pas pris la "bonne" décision qu'on en serait blâmable pour autant ! ☹

[‖]. À noter qu'il existe aussi, quoique ce soit plus rare, des situations de *propension au risque* ; confer p.ex. la remarque **??** *supra*.

(par exemple, si la mine que vous hésitez à exploiter n'est qu'une des centaines de concessions que possède votre entreprise multinationale), le phénomène de concavité est si peu marqué pour les montants mis en jeu par une seule mine qu'on peut alors le négliger et assimilier la fonction de perte à (l'opposé de) le bilan financier de l'exploitation, comme je l'ai fait ici. ♧

Remarque (ID'). Dans l'exemple (HX'), nous avons pu déterminer la fonction de perte pertinente pour notre problème sans vraiment d'ambigüité. Cependant, autant la *définition* de la fonction de perte est bien claire dans un contexte industriel (c'est ce qu'on « gagne » si on sait que $\boldsymbol{\varphi}$ vaut φ et qu'on agit en conséquence, moins ce qu'on « gagne » si on croit que $\boldsymbol{\varphi}$ vaut $\hat{\varphi}$), autant sa *quantification* sous forme numérique n'est pas évidente : car évaluer numériquement « ce qu'on gagne » est assez subjectif... Déjà, il y a l'aversion au risque dont j'ai parlé dans la remarque précédente. Mais il y a aussi, notamment, tous les enjeux autour de la *responsabilité sociétale de l'entreprise*, qui sont difficiles à quantifier lorsqu'on doit les comparer aux enjeux financiers ! Imaginons ainsi que vous ayez reçu une offre d'un fournisseur en Amérique du Sud, offre intéressante financièrement et dont la légalité ne pose pas question, mais que vous ayez de bonnes raisons de soupçonner que ce fournisseur soit en fait lié au narcotrafic local, et que lui donner des débouchés aiderait indirectement les gangs à prospérer. Sous réserve que vous ayez un minimum d'éthique, vous devrez alors, dans la construction de votre fonction de perte, tenir compte, non seulement du bilan financier (clairement positif) qu'un tel changement de fournisseur impliquerait pour votre entreprise, mais aussi de l'impact sociétal (clairement négatif) de celui-ci en Amérique du Sud ! Mais comment pondérer les deux ?... Une réponse naïve consisterait à introduire un poids *infini* à l'aspect éthique, en disant « j'ai une éthique, donc je ne veux faire que des choses parfaitement propres ». Mais on se rend vite compte que le choix de l'infini ne marche pas en pratique, parce que les choses sont plus compliquées que cela : par exemple, peut-être que le lien de votre fournisseur avec le narcotrafic est en fait très ténu, et que passer un accord avec lui ne dégradera pas grand-chose à la situation des gens sur place, alors que d'un autre côté, les bénéfices de votre entreprise seront si considérables que cela lui permettra, grâce à ses produits bon marché, d'aider la vie de millions de clients, d'offrir de meilleurs salaires à des milliers d'employés, de faire des dons humanitaires... Donc, dans tous les cas, il y a bien une question de *pondération*. Mais la déterminer comportera nécessairement une part de subjectivité... ♧

Chapitre 17

Pièges et paradoxes des tests

17.1 Que signifie « accepter une hypothèse nulle » ?

Déjà, soulignons que, dans la mesure du possible, il est méthodologiquement préférable de conclure l'analyse d'un test booléen en parlant du rejet (ou pas) de l'hypothèse nulle, plutôt que de la validation (ou pas) de l'hypothèse alternative :

Point (IE'). Lorsqu'on écrit la phrase de conclusion d'un test positif, il est généralement recommandé, par souci de clarté méthodologique, de préférer la formulation « on est enclin à rejetter $\mathscr{H}_0$ » à la formulation « on valide l'hypothèse $\mathscr{H}_1$ » ; et de même, pour la phrase de conclusion d'un test négatif, il est recommandé de préférer la formulation « l'observation est compatible avec $\mathscr{H}_0$ » à la formulation « on n'est pas en mesure de prouver l'hypothèse $\mathscr{H}_1$ ».

En effet, c'est là la philosophie même du principe de test statistique : la conclusion vaut par rapport à l'hypothèse nulle, *et uniquement par rapport à celle-ci* ; l'hypothèse « alternative », comme son nom l'indique, est simplement la réunion de toutes les hypothèses autres que l'hypothèse nulle... Certes, c'est généralement le fait de valider l'hypothèse *alternative* qui nous intéresse *réellement* ; néanmoins, écrire la conclusion en fonction de l'hypothèse nulle limitera les risques de mécompréhension du sens de votre analyse par ceux qui la liront ! ♧

Remarque (IF'). Le point méthodologique ci-dessus est même encore plus flagrant lorsqu'il s'applique à un test par p-valeur, dans la mesure où la probabilité à laquelle cette p-valeur se réfère concerne la loi de la statistique de test *sous l'hypothèse nulle !*. ♧

Par ailleurs, profitons-en pour remettre une couche sur la façon dont il convient d'interpréter resp. la positivité et (surtout !) la négativité d'un test. Nous avons certes déjà évoqué cela dans le chapitre 12 ; mais il s'agit d'une erreur d'interprétation si classique que remettre une couche dessus ne peut pas faire de mal :

Point (IG'). Lorsqu'on dit « on rejette l'hypothèse nulle au risque α », cela signifie « on a observé un phénomène qui a une chance moindre que α de se produire sous $\mathscr{H}_0$; et donc il est peu vraisemblable que l'hypothèse $\mathscr{H}_0$ soit effectivement vérifiée ». Cela est une conclusion d'autant plus forte que α sera petit. ♧ **!**

Point (IH'). À l'inverse, « accepter l'hypothèse nulle au risque α » ne signifie rien d'autre que « ne pas être en mesure de rejeter l'hypothèse nulle au risque α ». En particulier, ce n'est absolument pas synonyme de « rejeter l'hypothèse alternative au risque α » ! ♧ **!!**

Remarque (II'). Notons au passage, que, pas contraposition du point (IG'), la conclusion « accepter l'hypothèse nulle au risque α » est celle fois-ci une conclusion d'autant plus *faible* que α est petit ! Néanmoins, gardez bien à l'esprit qu'une telle conclusion d'acceptation n'a qu'une portée très faible dans tous les cas, peui importe la valeur de α... ♧

Remarque (IJ'). On notera ainsi qu'il est tout-à-fait possible et pas choquant du tout d'accepter (disons, au risque 10 %) plusieurs hypothèses contradictoires entre elles : cela signifie juste qu'aucune de ces hypothèses ne peut être exclue par le résultat du test ! ♧

Exemple (IK'). Pour illustrer la remarque ci-dessus, imaginons qu'un détective enquête sur le meurtre qui vient d'avoir lieu au cours d'une soirée dans un manoir. L'unique témoin n'a pu voir qu'une vague silhouette de l'assassin : celui-ci avait, selon son témoignage, la stature d'un homme. (On suppose ici que tous les convives masculins on grosso-modo la même stature, plus grande que celle de toutes les convives féminines). Bien sûr, la témoin a pu se tromper ; mais on admettra qu'il y a moins d'une chance sur dix pour que, voyant la silhouette d'une femme, elle la prenne pour celle d'un homme.

Du point de vue statistique, la quantité d'intérêt (explicative) à laquelle s'intéresse notre détective est alors l'identité de l'assassin ; et l'observation sur laquelle il s'appuie pour inférer cette quantité d'intérêt est le récit de la témoin. Dans ces conditions, lorsque le détective se demande si Madame Pervenche est l'assassin, il peut rejeter l'hypothèse nulle {M$^{\text{me}}$ Pervenche est l'assassin} au risque 10 % : puisque, fût-elle réellement coupable, il y avait moins de 10 % de risque que sa silhouette fût jugée masculine. Ainsi, le test innocente cette suspecte ! (au risque 10 %).

En revanche, si on applique le même test aux hypothèses nulles respectives {le Colonel Moutarde est l'assassin}, {le Révérend Olive est l'assassin} et {Le Professeur Violet est l'assassin}, cette fois-ci ces trois hypothèses seront toutes acceptées ! Est-ce que cela signifie qu'ils sont tous coupables ? Évidemment non, puisque le témoignage assure ici que l'assassin est unique ! Cela veut juste dire qu'on n'est, à ce stade, en mesure d'innocenter aucun des trois ☹ ♧

17.2 Comment choisir l'hypothèse nulle

Caractère fermé de l'hypothèse nulle

Lorsqu'on procède à un test statistique pour trancher entre deux hypothèses complémentaires, une question qui se pose naturellement est de savoir qui choisir comme hypothèse nulle et qui choisir comme hypothèse alternative. Bien souvent, il s'avère qu'on n'a pas vraiment le choix — ce qui peut d'ailleurs être fâcheux, car il y a souvent une hypothèse qu'on aimerait pouvoir démontrer, et pour cela on aurait besoin d'en faire notre hypothèse alternative... On a en effet la règle générale suivante (plus d'explications dans la suite) :

! **Principe** (IL'). *On choisira toujours*[*] *une hypothèse nulle correspondant à un sous-ensemble topologiquement* fermé *de l'espace du paramètre caché.*

En particulier, lorsque le test est défini par une ou plusieurs inégalités, il convient que ces inégalités soient larges *du côté de l'hypothèse nulle (et donc strictes du côté de l'hypothèse alternative).*

D'autre part, lorsque le test consiste à se demander si une certaine égalité (parfaite) est vérifiée ou cas, le cas « les deux termes sont égaux » devra toujours correspondre à l'hypothèse nulle*, et le cas « les deux membres sont distincts » à l'hypothèse* alternative*... !*

De manière plus générale, on peut retenir que « l'hypothèse nulle doit contenir le cas de référence », où par « cas de référence », je fais référence au cas d'égalité d'indépendance, etc. : soit que cette hypothèse corresponde au cas de référence lui-même*, soit qu'elle corresponde à une inégalité* large *consistant à être* au moins aussi *[quelque chose] que ce cas de référence.* ♢

> *Remarque* (IM'). En fait, quand on dit que l'hypothèse nulle est « fermée », il faudrait préciser par rapport à quelle topologie on entend cela... L'énoncé techniquement précis du principe ci-dessus serait que l'hypothèse nulle doit être fermée par rapport à la topologie de la distance $(\theta,\theta')\mapsto \|\mathrm{Loi}_{\theta'}(X)-\mathrm{Loi}_{\theta}(X)\|_{\mathrm{VT}}$, où $\|P'-P\|_{\mathrm{TV}}$ représente la « distance en variation totale » (dont nous ne donnerons pas la définition dans le cadre de ce cours) entre deux distributions de probabilité P et P' sur le même espace $\mathcal{X}$.
>
> En fait, dans tout ce cours, nous faisons implicitement l'hypothèse que l'application $\theta\mapsto \mathrm{Loi}_{\theta}(X)$ est continue depuis la topologie sur Θ (dont nous rappelons que, dans le cadre de ce cours, il s'agit toujours d'une partie d'un certain espace $\mathbb{R}^d$: cela définit donc la topologie sur Θ sans ambigüité) vers la topologie de la distance en variation totale sur $\mathscr{M}_1(\mathcal{X})$. (Et, de fait, tous les exemples présentés satisfont bien cette hypothèse). Dans ces conditions, « être une partie fermée de Θ au sens de la distance en variation totale entre les $\mathrm{Loi}_{\theta}(X)$ » implique « être une partie fermée de Θ vu comme un sous-espace de $\mathbb{R}^d$ » : donc, dire qu'il faut choisir l'hypothèse nulle fermée par rapport à la topologie de la variation totale implique bien qu'il faut choisir cette hypothèse nulle fermée pour la topologie "du modèle" sur Θ. (Même s'il est possible que certaines parties de Θ fermées au sens du modèle ne soient en réalité pas fermées pour la norme en variation totale : auquel cas, nonobstant leur caractère fermé, elles ne pourront pas convenir pour définir une hypothèse nulle intelligente). ♧

Remarque (IN'). Attention : ici, quand je dis que l'hypothèse nulle doit être fermée, c'est par rapport à la topologie sur Θ, pas sur l'espace $\mathbb{R}^d$ (dans lequel Θ est plongé) tout entier ! Par exemple, considérons le modèle du chasseur où on réduit Θ à l'ensemble $]0,0{,}5[\sqcup[0{,}8,0{,}9]$: cela correspondrait à la situation où le candidat qui se présente au club prétend être un certain tireur d'élite (dont on sait que, le cas échéant, sa fiabilité est comprise, au sens large, entre 80 % et 90 %), mais qu'on se demande s'il s'agit bien de lui où si ce ne serait pas plutôt son frère jumeau (dont on sait qu'il n'est pas capable d'atteindre 50 % de ses cibles, même si, là encore, on ignore sa fiabilité exacte). Eh bien, dans ce modèle-ci, l'hypothèse $]0,0{,}5[$ est, nonobstant les apparences, bel et bien *fermée*[†], dans la mesure où, *vu comme un sous-espace de* Θ, cet ensemble $]0,0{,}5[$ est en fait identique à l'ensemble $[0,0{,}5]$ (ou encore, par exemple, à l'ensemble défini par l'inégalité large $\{\theta\leqslant 0{,}6\}$). ♧

Remarque (IO'). Attention, l'expression « cas de référence » peut être un peu trompeuse. Mettons pas exemple qu'on cherche à rejeter l'hypothèse nulle selon laquelle les garçons seraient, en moyenne strictement plus forts en mathématiques que les filles (pour une certaine définition du concept de « force en mathématique », évaluée sous forme de nombre réel). Comme cette hypothèse n'est pas fermée, il procède du principe (IL') que cela ne constitue pas un bon choix d'hypothèse

[∗]. « Toujours » doit s'entendre ici comme « toujours, en *pratique* » : du point de vue *technique*, absolument rien n'empêche de choisir une hypothèse nulle non fermée ; néanmoins, comme nous l'expliquerons plus loin, cela ne présenterait aucun intérêt le cas échéant...

[†]. Elle est néanmoins ouverte *aussi*, ainsi qu'on pouvait le voir immédiatement d'après sa forme

nulle. Nous pourrions alors être tentés de dire qu'on n'a qu'à tester l'hypothèse nulle que « les garçons soient au moins aussi forts que les filles » ; sauf que cette hypothèse nulle ne nous intéresse pas non plus : car ce qu'on aimerait prouver, ce n'est pas que les filles soient strictement plus fortes que les garçons (on ne s'attend d'ailleurs pas à ce que cela soit vrai, pas plus qu'on ne s'attend à ce qu'il soit vrai que les garçons soient strictement plus forts que les filles) ; c'est que les niveaux moyens des filles et des garçons soient quasiment identiques... !

Dans ce cas, il s'avère que l'hypothèse nulle qu'il sera pertinent de tester est en fait « les garçons ont un avantage strict en mathématiques sur les filles, et pas complètement infime qui plus est [‡] » : autrement dit, si dans notre modèle, le niveau moyen des garçons est notée $\mu_{G\checkmark}$, et que le niveau moyen des filles est noté $\mu_{F\checkmark}$, notre hypothèse nulle correspondra à $\{\mu_G - \mu_F \geqslant \varepsilon\}$, où ε est une certaine valeur strictement positive, mais néanmoins extrêmement petite, de sorte qu'on peut considérer que les cas où $\{0 < \mu_G - \mu_F < \varepsilon\}$ correspondent *en pratique* à une égalité essentiellement parfaite entre garçons et filles !

Dans ce cas, vous pourriez m'objecter « hé ; cette nouvelle hypothèse nulle correspond certes à une inégalité large ; mais elle ne contient pas le cas de référence où garçons et filles ont le même niveau moyen ! ». Certes. Sauf qu'ici, ce n'est pas cela qu'il faut entendre par « cas de référence » : le « cas de référence », en fait, désigne la situation, pouvant être décrite par une égalité, *par rapport à laquelle notre hypothèse nulle se place* : et en l'occurrence, il s'agit donc du cas d'égalité $\{\mu_G = \mu_F + \varepsilon\}$ (qui signifie « les garçons ont le plus petit avantage imaginable sur les filles qui mériterait, le cas échéant, d'être signalé »), cas qui est bel et bien inclus dans notre hypothèse nulle ! ♧

Expliquons maintenant les « preuves » du fait qu'il convient de toujours prendre l'hypothèse nulle fermée :

Premier argument pour le principe (IL'). La raison pour laquelle il convient de placer le cas de référence au sein de l'hypothèse nulle provient de la philosophie même du test statistique (booléen), qui nous demande de tirer une conclusion qui soit valable au risque α *quelle que soit l'hypothèse ponctuelle* θ au sein de l'hypothèse nulle $\mathscr{H}_0$. (Et ici, notez que la dissymétrie entre les hypothèses nulle et alternative se manifeste clairement, puisqu'on ne pose, en revanche, aucune exigence qui soit uniforme sur Θ_1). Par conséquent, si l'hypothèse $\mathscr{H}_0$ est trop "grande", et qu'elle peut conduire à un ensemble de conséquences trop diverses, on ne parviendra pas construire de test statistique pertinent... Idéalement, il faut donc que l'hypothèse nulle soit réduite à un seul point, ou qu'on puisse plus ou moins se ramener à cette situation : ce point auquel on peut se réduire correspondant au fameux « cas de référence », il est naturel de l'inclure au sein de l'hypothèse nulle, puisque, de toutes façons, elle devra s'y ramener... ! ♤

Deuxième argument pour prendre l'hypothèse nulle fermée. Une seconde raison pour toujours prendre l'hypothèse nulle fermée est qu'on peut toujours fermer l'hypothèse nulle sans qu'on y perde rien ! Plus précisément, soit Θ_0 un sous-ensemble non nécessairement fermé de Θ et *Test* un test [§] de l'hypothèse $\{\theta \in \Theta_0\}$. Alors,

[‡]. Avec le vocabulaire qui sera introduit dans la § 17.5, cette idée que l'avantage soit « pas complètement infime » correspond à dire qu'on exige qu'il y ait une *taille d'effet* pas trop ridicule pour le phénomène de différence éventuelle entre filles et garçons.

[§]. J'écris ce point avec le cas d'un test booléen en tête ; mais en fait cela fonctionne exactement de la même façon dans le cas d'un test par p-valeur.

si nous notons $\bar{\Theta}_0$ l'adhérence de Θ_0, il se trouve que *Test* est *aussi* un test de l'hypothèse Θ_0 *sans* qu'on ait besoin d'en changer le niveau ou la p-valeur ! Et comme $\bar{\Theta}_0$ est un sur-ensemble de Θ_0, cela signifie qu'en considérant qu'on a en fait testé l'hypothèse $\{\theta \in \bar{\Theta}_0\}$, dans les cas où on rejette l'hypothèse nulle, la conclusion obtenue est plus forte, puisqu'on rejette plus de possibilités... ♤

L'argument qui précède peut même être mis en valeur en tant que théorème :

Théorème (IP'). *Si Test est un test binaire au niveau α (resp., si P est un test par p-valeur) de l'hypothèse nulle $\{\theta \in \Theta_0\}$, où Θ_0 est une partie non nécessairement fermée de l'espace Θ du paramètre caché, alors, notant $\bar{\Theta}_0$ la fermeture topologique de Θ_0 au sein de Θ, Test constitue automatiquement un test de niveau α (resp. P constitue un test par p-valeur) de l'hypothèse nulle $\{\theta \in \bar{\Theta}_0\}$* aussi. ♢

Troisième argument pour prendre l'hypothèse nulle fermée. Un dernier argument pour dire qu'une hypothèse nulle non fermée n'est pas pertinente et liée à l'idée de *consistance* du test, ou plus généralement à son risque de seconde espèce. En effet, il découle du théorème (IP') que, si $\theta_!$ est un point situé dans $\bar{\Theta}_0$ mais pas dans Θ_0, alors on aura $\mathbb{P}(\mathit{Test} \mid \theta = \theta_!) \leqslant \alpha$, de sorte que la fonction de risque de seconde espèce, en $\theta_!$, vaudra au moins $1-\alpha$: ce qui est catastrophique... Certes, le théorème ?? nous dit que, de manière générale, le *supremum* de cette fonction de risque de seconde espèce vaudra au moins $1-\alpha$; néanmoins, autant faire en sorte que, dans la mesure du possible, ce supremum ne soit pas *atteint* pour autant ! C'est d'autant plus que, à supposer qu'on s'intéresse à un régime asymptotique « $n \to \infty$ » pour notre modèle, on voudrait que, dans ce régime, notre test soit consistant, autrement dit, que le risque de deuxième espèce $\beta^{(n)}(\theta)$ tende, lorsque $n \to \infty$, vers 0, et ce pour tout $\theta \in \Theta_1$. Néanmoins, s'il existe un $\theta_!$ dans $\bar{\Theta}_0 \smallsetminus \Theta_0$, un tel $\theta_!$ sera bien dans Θ_1, mais vérifiera cependant $\beta^{(n)}(\theta) \geqslant 1-\alpha$ pour tout n : et dès lors, ce point de Θ_1 empêchera le test d'être consistant : même en prenant un paramètre n arbitrairement grand, lorsque θ est égal à $\theta_!$, il n'est pas possible de s'apercevoir qu'on n'est pas dans l'hypothèse nulle... !

Ainsi, non seulement fermer l'hypothèse nulle ne nuit pas au niveau de notre test (confer l'argument précédent), mais *ne pas* la fermer, par contre, nuirait à sa puissance ! Il est donc toujours préférable de fermer l'hypothèse nulle. ♤

> *Remarque* (IQ'). L'argument ci-dessus a été écrit dans le formalisme des tests booléens ; mais il se transpose également (au prix d'un peu de complexité technique) au cas des tests par p-valeur ☋ ♧

Remarque (IR'). L'argument précédent montre que le fait de prendre l'hypothèse nulle fermée n'est pas tant un *choix* du statisticien, qui ne serait affaire que de convention, mais bien une *contrainte*, due au principe même du test statistique, s'il souhaite pouvoir effectivement démontrer l'hypothèse alternative (quitte à devoir se placer en régime asymptotique) chaque fois qu'elle se présente...! !

Par exemple, dans le modèle du pédagogue, imaginons que notre enseignant se dise « d'accord ; je n'arrive peut-être pas à prouver que ma méthode fait strictement progresser les élèves, mais je pourrais au moins essayer de prouver qu'elle ne les fait pas régresser ! », en souhaite ainsi tester l'hypothèse nulle $\{\mu < \mu_{\text{réf}}\}$ en vue de la réfuter. La problème s'il choisit une telle hypothèse nulle, c'est que la méthodologie même du test impose qu'il y ait une probabilité au moins $(1-\alpha)$ d'accepter l'hypothèse nulle lorsqu'on a $\{\mu = \mu_{\text{réf}} - 10\}$, mais *aussi* lorsqu'on a $\{\mu = \mu_{\text{réf}} - 1\}$,

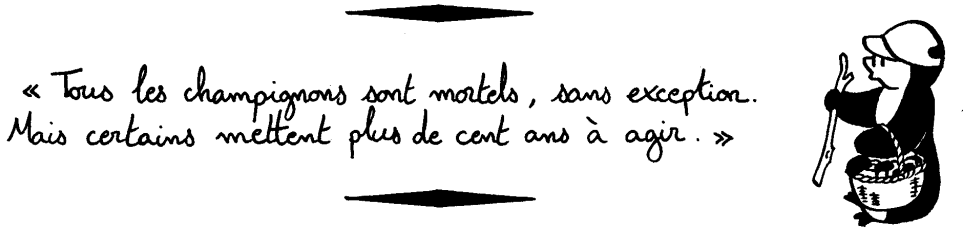

FIGURE 17.1 – Si nous appelons $\tau_{\checkmark} \in [0, \infty]$ le temps au bout duquel l'ingestion d'un champignon finit par tuer, l'affirmation « Tel champignon est mortel » correspond à l'hypothèse $\{\tau < \infty\}$, tandis que l'affirmation « Tel champignon est sans danger » correspond à l'hypothèse $\{\tau = \infty\}$. On voit donc que l'hypothèse de mortalité n'est pas fermée dans $[0, \infty]$; dès lors, elle ne saurait constituer un choix d'hypothèse nulle judicieux pour un test : en effet, si on admet que l'effet d'un champignon peut être arbitrairement long, il devient impossible de réfuter sa létalité... Bien entendu, Anicet est de mauvaise foi, car en pratique on peut mettre une borne au temps d'effet d'un champignon, de sorte que l'affirmation de sa mortalité devient alors une hypothèse parfaitement testable...! (source : *Anicet le Pingouin, « À en mourir de rire »*)

lorsqu'on a $\{\mu = \mu_{\text{réf}} - 0{,}1\}$, lorsqu'on a $\{\mu = \mu_{\text{réf}} - 0{,}01\}$, ... Et on pourrait alors montrer, par passage à la limite, quand dans ce cas, on aura nécessairement une probabilité au moins $(1 - \alpha)$ d'accepter l'hypothèse nulle lorsqu'on sera dans le cas $\{\mu = \mu_{\text{réf}}\}$. Or, *justement*, ce qui intéresserait notre enseignant, c'est d'être capable, même lorsque l'impact de sa méthode sur les élèves est nul, de *prouver* qu'elle n'a pas un impact négatif. Mais ça, si grand que soit le nombre d'élèves qu'il teste, ce n'est juste pas possible : puisque dans tous les cas, il aura une très grande probabilité de conclure que le résultat observé est compatible avec l'hypothèse d'une régression !

On peut le comprendre de façon plus intuitive en disant qu'un test de détection de la chute du niveau des élèves qui devrait réagir à la moindre chute, si petite soit-elle, sera tellement sensible qu'il sera obligé de réagir même quand il n'y a plus de chute du tout...

Dans le gag de la figure 17.1, on retrouve à nouveau cette idée qu'un test est incapable de détecter les points de l'hypothèse alternative qui seraient au bord de l'hypothèse nulle : même si, cette fois-ci, l'exploitation qui en est faite est de nature purement humoristique ! ☺ ♧

Intérêt pour l'hypothèse alternative

La sous-section ci-dessus concernait une caractéristique *technique* de l'hypothèse nulle : il faut la choisir fermée, pour des raisons purement *mathématiques*. Mais mettons que nous ne soyons pas soumis à cette contrainte, par exemple parce que nous devons trancher entre deux hypothèses simples (voir aussi l'importante remarque **??** *infra* sur ce point). Dans ce contexte, qui faut-il choisir comme hypothèse nulle, et qui faut-il choisir comme hypothèse alternative ?

!! **Principe (IS').** *Dans le cadre de la statistique fréquentiste, si on a le choix concernant les rôles à jouer par les deux hypothèses complémentaires qu'on considère, deux considérations doivent entre en jeu :*

(i) S'il y a une des deux hypothèses dont la preuve apporte une information plus intéressante que l'autre, par exemple, si la conclusion « on a prouvé $\mathscr{H}_{\mathrm{a}}$ » apporte, le cas échéant (car, bien sûr, $\mathscr{H}_a$ peut aussi être fausse, ou trop difficile à prouver), une information plus intéressante que la conclusion « on a prouvé $\neg\mathscr{H}_{\mathrm{a}}$ », alors cette hypothèse doit jouer le rôle d'hypothèse alternative *: en effet, si notre est négatif, la conclusion qu'il apportera ne sera de toutes façons que très peu intéressante : donc autant se concentrer sur ce qui serait le plus intéressant pour nous* dans le cas d'un test positif *!*

(ii) S'il y a une des deux hypothèses en faveur de laquelle pencher à tort serait plus fâcheux que l'autre (par exemple, si trancher en faveur de $\mathscr{H}_a$ alors qu'on a en réalité $\neg\mathscr{H}_{\mathrm{a}}$ a des conséquences plus fâcheuses que trancher en faveur de $\neg\mathscr{H}_{\mathrm{a}}$ alors qu'on a en réalité $\mathscr{H}_{\mathrm{a}}$, alors cette hypothèse doit jouer le rôle d'hypothèse nulle *: en effet, de par la nature du test, le seul risque qu'on soit effectivement en mesure de contrôler est le risque de première espèce, donc il convient d'appliquer ce contrôle au risque qu'on chercher le plus fortement à éviter !*

◇

Exemple (IT'). Le cas (i) est typique de la recherche pharmaceutique : le laboratoire qui développe une molécule donnée souhaite *prouver* que sa molécule apporte un bénéfice thérapeutique, afin d'obtenir son autorisation de mise sur le marché ! Dès lors, il procèdera à des tests où l'hypothèse nulle sera que la molécule est inefficace, voire nuisible. En effet, si l'hypothèse d'une efficacité trop faible était l'hypothèse alternative, on serait dans une situation stupide : soit le test serait négatif et alors on ne pourrait pas conclure de façon ferme, soit il serait positif et alors ça n'aurait de toutes façons quasiment aucun intérêt d'apprendre qu'on a travaillé pour rien... ! ♧

Exemple (IU'). Le cas (ii) est quant à lui typique de l'expertise judiciaire : notre philosophie contemporaine de la justice, en effet, est fondée sur le principe qu'il est plus grave de condamner un innocent que d'acquitter un criminel ! (D'où le principe « le doute doit profiter à l'accusé » [¶] — ou, en latin : « *In dubio pro reo* »). Par conséquent, lorsqu'une expertise judiciaire est appliquée (par exemple, pour analyser une trace de sang sur les lieux du crime), c'est l'hypothèse que l'accusé n'a rien à voir avec ce qui est reproché (donc, en l'occurrence, ce ce sang n'est pas le sien) qui doit être prise comme hypothèse nulle ! ♧

Remarque (IV'). Si on se trouve à la fois dans la situation du cas (i) et dans la situation du cas (ii), "normalement" ils doivent jouer dans le même sens. En effet, s'il plus grave de rejeter $\mathscr{H}_{\mathrm{a}}$ à tort que de la croire à tort, la prudence consiste à agir en partant du principe que $\mathscr{H}_{\mathrm{a}}$ est vraie : et dès lors, avoir la confirmation que $\mathscr{H}_{\mathrm{a}}$ est effectivement vraie ne serait pas si intéressant que ça, puisque de toutes façons nous aurions agi comme si elle était vraie... Ainsi, l'hypothèse qu'il est le plus grave de rejeter à tort est aussi celle qu'il est le moins intéressant de prouver : dans les deux cas, cela nous indique de prendre cette hypothèse comme hypothèse nulle ! ♧

[¶]. Confer p.ex. l'article 304 du code de procédure pénale, ou l'arrêt « Boutaffala v^{s} Belgique » rendu par la Cour Européenne des Droits de l'Homme le 28 juin 2022.

Remarque (IW'). On peut néanmoins s'amuser à chercher des cas où ces deux principes jouent en sens inverse. Prenons par exemple le problème auquel s'est retrouvée confrontée l'armée américaine au cours de la seconde guerre mondiale : y a-t-il, ou pas, la possibilité d'exploiter l'énergie nucléaire pour créer une arme révolutionnaire (à savoir, la bombe atomique) ? [‖] On a clairement l'impression qu'il est plus intéressant de savoir que la bombe atomique est bel et bien possible (il faut alors mettre le paquet dessus !) que d'apprendre qu'elle est en fait impossible (le cas échéant, *business as usual*, quoi !). Pourtant, croire à tort en l'impossibilité de la bombe atomique aurait des conséquences beaucoup plus grave (si l'Allemagne nazie, de son côté, se mettait à la développer, ce serait la défaite des Alliés) que de croire à tort à se possibilité (on y dilapiderait certes une partie de l'effort de guerre, mais rien de critique... !). ♧

Remarque (IX'). Cela dit, lorsqu'il est possible de faire jouer *soit* le rôle d'hypothèse nulle, *soit* le rôle d'hypothèse alternative, à chacun de nos deux hypothèses complémentaires, autant faire les deux tests correspondants : cela ne coute en effet qu'une augmentation de l'effort d'analyse statistique [**], pour un résultat qui a deux fois plus de chances d'être exploitable... ! ⌢ ♧

! *Remarque* (IY'). Il arrive fréquemment que le principe (IS') nous donne envie de tester une hypothèse nulle de la forme $\{\boldsymbol{\varphi} > \varphi_{\text{réf}}\}$ ou $\{\boldsymbol{\varphi} \neq \varphi_{\text{réf}}\}$: par exemple, si le pédagogue souhaite s'assurer que sa méthode n'est pas nuisible, ou si une entreprise veut s'assurer qu'elle ne discrimine pas ses employés en fonction de l'âge... Or, de telle hypothèses ne sont pas fermées, et semblent donc inaptes à jouer le rôle d'hypothèse nulle !

Néanmoins, dans un tel cas, nos hypothèses nulles sont équivalentes *en pratique* à des hypothèses nulles fermées, où l'on prend juste un tout petit peu de marche par rapport à l'égalité parfaite du cas de référence. Par exemple, au lieu de tester l'hypothèse nulle « la nouvelle méthode fait strictement régresser les élèves », l'enseignant peut choisir de plutôt l'hypothèse nulle « la nouvelle méthode fait régresser les élèves *d'au moins* 0,1 pt » : car, bien franchement, si elle fait régresser les élèves de 0,05 pt, *en pratique*, c'est comme si elle n'avait aucun effet ! De même, notre entreprise pourra tester l'hypothèse nulle « il existe un niveau discrimination avec l'âge suffisamment important pour avoir un imact réel sur la vie des employés » : car, à supposer qu'on arrive à démontrer que, toutes choses égales par ailleurs, un employé en fin de carrière touche 1 € par an en moins qu'une employée en début de carrière, on n'a pas envie de dire qu'un tel niveau de "discrimination" pose quelque souci que ce soit... ! Ce très léger déplacement du seuil de référence que pouvoir se ramener à des hypothèses nulles fermées est très lié à l'idée de *taille d'effet* que nous présenterons dans la § 17.5.

De ce fait, *en pratique*, il est toujours possible de définir n'importe quelle hy-

[‖]. De notre côté de l'histoire, la question parait stupide ! Mais en réalité, lorsque le projet *Manhattan* a été lancé, il n'était pas clair du tout qu'il fût effectivement possible de réaliser effectivement un bombe atomique, et encore moins, le cas échéant, qu'une telle arme apportât un avantage miliaire réel le cas échéant : ainsi, la fameuse lettre Einstein-Szilárd, enjoignant les États-Unis d'entreprendre des recherches dans cette direction, sous-estime le potentiel réel des bombes atomiques à un niveau frappant : ces physiciens, nonobstant leur exceptionnel génie à tous les deux, y écrivent en effet simplement qu'« il est concevable, quoique bien moins certain, que des bombes d'un nouveau type et extrêmement puissantes [puissent] être assemblées. Une seule bombe de ce type, transportée par bateau et explosant dans un port, pourrait très bien détruire l'ensemble du port ainsi qu'une partie de la zone aux alentours. Toutefois, de telles bombes pourraient très bien s'avérer trop lourdes pour un transport aérien »...

[**]. Sachant que, bien souvent, c'est plutôt la collecte des données qui est le facteur limitant...

pothèse comme nulle ou alternative tout en restant mathématiquement cohérent, quitte à utiliser la petite perturbation expliquée ci-dessus ☺ ♧

Voir aussi le théorème (IP') à ce sujet.

17.3 Trop de tests tuent le test !

Si vous procédez à deux tests différents pour une même hypothèse $\mathscr{H}_0$, disons un test T_1 au risque α_1 et un test T_2 au risque α_2, vous pouvez alors définir un test de synthèse T à partir de ces deux tests : ce test T sera défini comme étant positif si *au moins un des deux* tests T_1 et T_2 l'est — attendu qu'une seule conclusion positive suffit à rejeter définitivement $\mathscr{H}_0$, alors qu'à l'inverse une conclusion négative ne permet de l'accepter que jusqu'à preuve du contraire. Mais attention : quel est le niveau du test global T?... C'est $\alpha_1 + \alpha_2$ [††], puisque sous l'hypothèse nulle, il y a un risque α_1 d'échouer au premier test et un risque α_2 d'échouer au second...

Remarque (IZ'). Cela est évident, me direz-vous. Pourtant, on peut facilement se faire piéger en oubliant cet argument... Imaginons ainsi que je procède à 20 tests $T_0, \dots, T_{19}$ de la même hypothèse nulle, chaque test ayant un risque de 5 %. Le test global aura un risque de 100 %, c.-à-d. qu'il ne signifiera strictement rien : de la sorte, même si l'hypothèse nulle est effectivement vérifiée, il n'y aura rien d'étonnant à ce qu'au moins un test T_i s'avère positif — disons T_{13} pour fixer les idées. Mais si vous vous contentez de clamer « Le test T_{13} est positif, donc je peux rejeter l'hypothèse nulle au risque 5 % », votre conclusion sera malhonnête, car vous n'aurez pas mentionné que pour arriver à un tel résultat vous avez *aussi* procédé à 19 autres tests négatifs, et donc que la valeur de 5 % n'est pas le vrai niveau du test réellement opéré... C'est exactement ce qui se passe dans la bande dessinée de la figure 17.2. ♧

Principe (JA'). *Moralité : quand on fait plusieurs tests de la même hypothèse, il faut mentionner tous les tests opérés,* y compris ceux qui se sont révélés négatifs*, pour avoir le vrai niveau de confiance, lequel est la somme des niveaux de chaque test. Notez qu'il s'agit là d'un vrai souci en recherche scientifique, car on tend à omettre de publier les résultats non concluants (ceux pour lesquels les tests se sont relevés négatifs), ce qui conduit à considérer comme significatifs (c.-à-d. ayant un petit niveau de risque) des résultats qui ne le sont en fait pas...* ◇ !!

Principe (JB'). *Dans le même esprit, il est essentiel de définir votre protocole de test* avant *d'observer les données : sinon, en bricolant une région d'acceptation ad hoc conçue précisément pour accepter ou rejeter les résultats que vous avez obtenus, vous pourriez accepter ou rejeter n'importe quoi à votre guise...* ◇ !!

Remarque (JC'). Lorsqu'on est face à une situation comme celle de la figure 17.2 où l'hypothèse alternative de notre test apparait comme une union de plusieurs sous-hypothèses alternatives individuelles "élémentaires", il existe heureusement des méthodes pour pouvoir quand même regrouper les sous-tests individuels en un test global ayant le bon niveau de significativité ! La plus simple de ces méthodes est

[††]. $(\alpha_1 + \alpha_2)$ est la meilleure borne sur le niveau du test global dont on puisse être assuré dans le cadre présenté ici. On notera toutefois qu'en pratique, les tests T_1 et T_2 seront souvent indépendants voire positivement corrélés, auquel cas on pourra un peu (et parfois même beaucoup) améliorer cette borne...

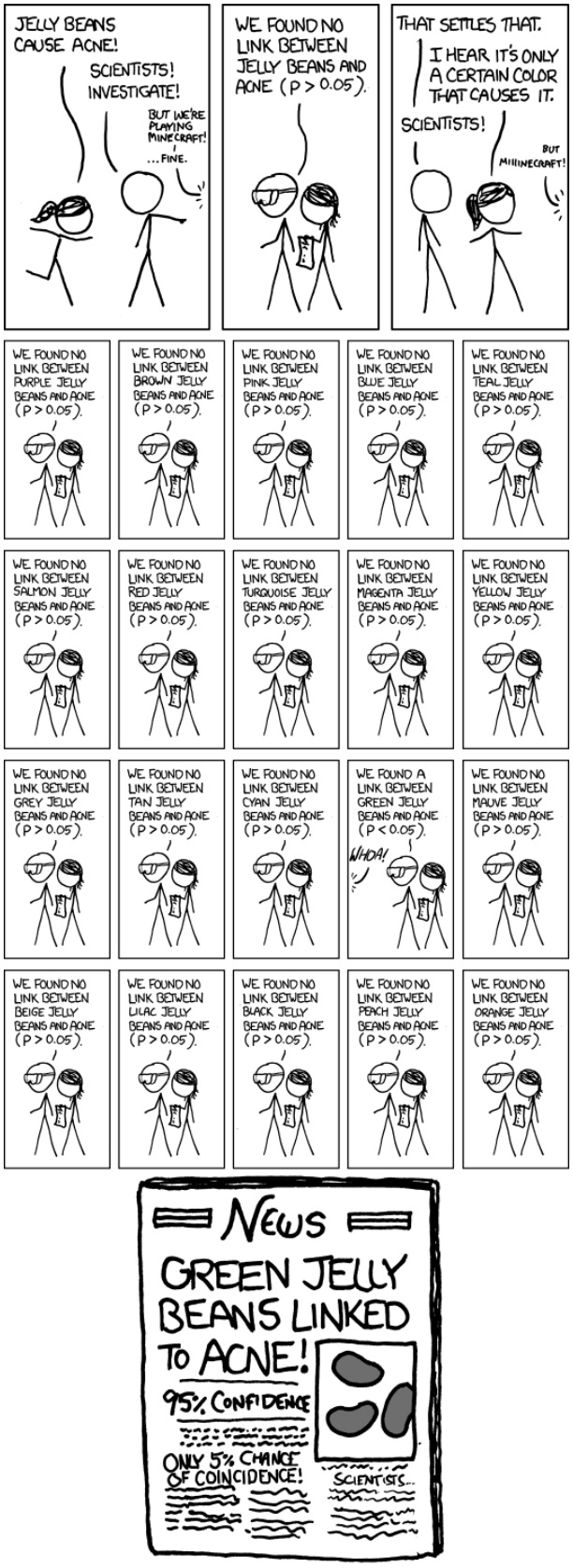

FIGURE 17.2 – Trop de tests tuent le test... (source : http://xkcd.com/882/)

.

la *correction de Bonferroni*, qui consiste, s'il y a m hypothèses alternatives individuelles ($m = 20$ dans l'exemple de la figure 17.2), à rejeter l'hypothèse nulle globale au niveau α lorsque l'une au moins des hypothèses nulles individuelles est rejetée au niveau α/m. (Ce qui revient à dire que, si les p-valeurs trouvées pour chaque sous-test individuel sont $p_1, \dots, p_m$, on renverra une p-valeur globale de $(\min_{1 \leqslant i \leqslant m} p_i) \times m$). Cela dit, on n'a rien sans rien : et lorsqu'on veut tester une hypothèse nulle conjonction de m sous-hypothèses élémentaires en utilisant cette correction de Bonferroni, la puissance du test devient souvent médiocre comparé à ce qu'elle était pour chaque sous-test individuel... ♧

Une autre façon de se faire piéger par l'adage « trop de tests tuent le test », plus insidieuse dans la mesure où fréquemment on ne s'en aperçoit pas, est d'oublier qu'IL FAUT DÉFINIR LE TEST AVANT DE REGARDER LES DONNÉES ! Ne pas le faire est une des erreurs les plus faciles à commettre. Vous observez un jeu de données, par exemple les décimales de π. Là, vous observez un phénomène apparemment remarquable : entre la position 762 et la position 767, il y a 6 décimales '9' successives [‡‡] ! Dès lors vous vous demandez : est-ce vraiment une coïncidence extraordinaire, ou est-ce que cela peut se produire facilement sous l'effet du seul hasard ? Pour cela, vous décidez de regarder la statistique de test « plus longue série de chiffres consécutifs identiques parmi les 767 premiers chiffres », sous l'hypothèse nulle que « les chiffres sont statistiquement indistinguables d'une série de chiffres i.u.d. » : on trouve une p-valeur de 6,84 ‰ [∗], ce qui est tout de même très faible. Faut-il en déduire qu'on a levé un grand mystère dans les décimales de π sur lequel la science doit se pencher dare-dare ?...

... En fait, sans doute pas. Certes, la probabilité d'observer 6 chiffres identiques à la suite dans une série de 767 chiffres aléatoires *est* très faible, rien à redire là-dessus. Mais aurions-nous choisi *spontanément* de regarder la longueur de la plus longue séquence de chiffres identiques parmi les 767 premiers chiffres si nous n'avions pas d'abord *vu* que les décimales de π présentaient cette série de six '9' ? Certainement pas avec la valeur 767 en tout cas : à la rigueur, nous aurions peut-être regardé la plus longue séquence parmi les 1 000 premiers chiffres (ce qui fait déjà légèrement augmenter la p-valeur calculée, à 8,92 ‰). Mais pourquoi 1 000, après tout ?... Nous aurions pu plutôt choisir de nous pencher sur les 100 premiers chiffres, ou les 10 000 premiers chiffres, ou... ! Et puis, nous aurions pu trouver tout aussi remarquable, par exemple, qu'un chiffre apparaisse 130 fois ou plus parmi les 1 000 premiers, ou qu'une séquence de 5 chiffres y apparaisse deux fois consécutivement (du genre, « ... 4124541245 ... »), ou qu'il y ait une série de 17 chiffres consécutifs tous compris entre 0 et 4, ou que la même séquence de 8 chiffres apparaisse à deux endroits différents (du genre, « ... 93779974 ... 99377974 ... »), etc., autant d'évènements dont la probabilité de survenue est de l'ordre de 1 % — et qui, eux, n'arrivent pas pour les vraies décimales de π [†]... Autrement dit, puisque nous avons fait notre test *en fonction* des données, c'est comme si nous avions fait tous les tests possibles selon ce qu'on aurait pu remarquer et eu envie de tester ! Et on se retrouve dans la situation précédente : trop de tests tuent le test...

Il faut donc vraiment insister sur ce point :

!!

[‡‡]. Authentique : Vous pouvez vérifier ! ☺

[∗]. Et non 7,62 ‰ comme on serait enclin à le penser, car le calcul de cette p-valeur est assez piégeux...

[†]. Oui, j'ai vérifié ☺

Principe (JD'). *Quand on procède à une étude statistique, il faut décider de ce qu'on va tester* avant *de connaitre les données, sinon on voit des phénomènes significatifs là où il n'y a rien !* ◇

Remarque (JE'). En pratique, comme il est difficile de définir complètement à l'avance le protocole d'analyse, ce qu'on fait souvent est qu'on regarde un premier jeu de données, en fonction de ce qu'il semble montrer, on se dit qu'on aurait aimé tester telle chose, puis on fait le test sur un second jeu de données indépendant du premier : là, il n'y a pas de biais. (Cette pratique est appelée *validation croisée* ; elle est très utilisée dans le domaines de l'apprentissage automatique que vous verrez en 2e année). Pour reprendre l'exemple de π, on pourrait par exemple essayer de confirmer l'hypothèse qu'il y a tendance à avoir de longues séries de chiffres identiques parmi les $\sim 1\,000$ premiers chiffres en regardant le développement de π en base 11 (lequel n'a, à priori, rien à voir avec celui en base 10) : et là, patatras, on ne trouve rien de mieux que 3 chiffres identiques consécutifs, soit une p-valeur supérieure à 0,5... ♧

Remarque (JF'). En réalité, lorsqu'on ne peut vraiment pas se permettre de procéder à de la validation croisée (par exemple, parce qu'on considère les données liées à un évènement passé qu'on ne peut pas reproduire à volonté), on peut quand même faire de l'analyse statistique, mais il faut alors être très prudent pour que les méthodes proposées ne tombent pas dans les pièges évoqués ci-dessus : on sera alors amené à utiliser des techniques spécifiques dans le style de la correction de Bonferroni... On qualifie de *méthodes d'analyse post hoc* les techniques destinées à faire de la statistique correcte sur des données collectées avant d'avoir défini les éléments que nous souhaitons regarder. ♧

17.4 Le paradoxe du test à droite

Plaçons-nous dans la situation suivante. On considère des données $x_{0\checkmark}, \dots, x_{15\checkmark}$ issues de la réalisation de variables aléatoires $X_0, \dots, X_{15}$ de même loi $\mathcal{N}(\mu_\checkmark, \sigma_\checkmark^2)$, données pour lesquelles on a mesuré une moyenne empirique $m_\checkmark = 0{,}5$ et un écart-type empirique $s_\checkmark = 1$. On voudrait regarder ce qui se passe quand on teste l'hypothèse nulle $\{\mu = 0\}$ et quand on teste l'hypothèse nulle $\{\mu \leqslant 0\}$ (sachant que $\sigma_\checkmark$ est inconnu). Dans les deux cas, on va utiliser que

$$\frac{M - \mu_\checkmark}{S/\sqrt{15}} \overset{\mathbb{P}_\checkmark}{\sim} \mathrm{T}_{\mathrm{St}}(15), \tag{JG'}$$

où $\mathrm{T}_{\mathrm{St}}(d)$ désigne la loi de Student de paramètre d.

Pour tester l'hypothèse nulle $\{\mu = 0\}$, on écrit que si $\mu_\checkmark = 0$, on doit avoir

$$\frac{M}{S/\sqrt{15}} \overset{\mathbb{P}_\checkmark}{\sim} \mathrm{T}_{\mathrm{St}}(15), \tag{JH'}$$

d'où, avec un risque de 5 %,

$$\frac{M}{S/\sqrt{15}} \in [-2{,}14, 2{,}14]. \tag{JI'}$$

Ici la réalisation du membre de gauche vaut 1,94, de sorte qu'on ne peut pas rejeter l'hypothèse $\{\mu = 0\}$.

Si maintenant on souhaite tester l'hypothèse nulle $\{\mu \leqslant 0\}$, on doit alors écrire que sous cette hypothèse, on a

$$\frac{M}{S/\sqrt{15}} \leqslant \frac{M - \mu_\checkmark}{S/\sqrt{15}} \overset{\mathbb{P}_\checkmark}{\sim} \mathrm{T}_{\mathrm{St}}(15). \tag{JJ'}$$

Dans ce cas-là, on n'a pas d'autre choix que de faire un test à droite [‡] : observant que pour $X \sim \mathrm{T}_{\mathrm{St}}(15)$, on a $X \leqslant 1{,}76$ au risque 5 %, on en déduit à fortiori que, sous l'hypothèse que $\{\mu \leqslant 0\}$, on doit avoir

$$\frac{M}{S/\sqrt{15}} \leqslant 1{,}76\,; \qquad (\mathsf{JK}')$$

et puisque la réalisation du membre de gauche toujours 1,94, cette fois-ci l'hypothèse est rejetée.

Voilà qui extrêmement choquant : on n'est pas en mesure de rejeter l'hypothèse $\{\mu = 0\}$; mais à partir de la même statistique de test, et pour le même risque, on peut rejeter l'hypothèse pourtant *plus large* $\{\mu \leqslant 0\}$...!! Comment expliquer ce paradoxe ?

La première remarque, c'est que le test consistant à regarder si $M/(S/\sqrt{15}) > 1{,}76$ est *aussi* un test de l'hypothèse nulle $\{\mu = 0\}$: la logique est donc sauve puisque, eussions-nous utilisé *exactement* le même test pour les deux hypothèses nulles, le rejet de l'hypothèse la plus large aurait bien entrainé celui de l'hypothèse la plus étroite...

Se pose alors naturellement la question : pourquoi ne pas avoir aussi utilisé un test à droite pour l'hypothèse $\{\mu = 0\}$? La réponse est à chercher dans le risque de seconde espèce. En effet, le défaut du test à droite est qu'il sera très peu effiace pour détecter des valeurs *strictement négatives* de μ : dans un tel cas en effet, aussi grand soit $|\mu_{\checkmark}|$, le test sera réussi avec au moins 95 % de probabilité, ce qui est très mauvais... Ce souci rend le test à droite, quoique *correct* formellement, *inadapté* en pratique au test de l'hypothèse $\{\mu = 0\}$, alors qu'il ne se pose pas pour tester l'hypothèse $\{\mu \leqslant 0\}$. [§]

« Oui, mais quand même... », me direz-vous. Le paradoxe subsiste : si vous me demandez sans plus de précision s'il est plausible que μ soit nul, la procédure adéquate m'amènera à vous répondre « oui » ; alors que si vous me demandez s'il est plausible que μ soit négatif ou nul, je répondrai « non »... Y a-t-il une explication "morale" à ce phénomène...?

Eh bien, oui ! Car *poser une question, c'est déjà y répondre un peu...* [¶] Quand je vous demande « μ est-il nul ? », vous en déduisez intuitivement que cela sous-entend que μ a de bonnes chances d'être nul aussi bien que d'être non nul (disons avec une probabilité à priori de 1/2 pour chaque), sans que cela ne vous dise rien sur le fait que μ soit positive ou négative dans le cas non nul (disons qu'il y a une probabilité à priori de 1/4 pour chacune de ces deux possibilités). À l'inverse, quand je vous demande « μ est-il négatif ou nul ? », vous avez tendance à en déduire qu'il y a des chances équilibrées que μ soit négatif-ou-nul d'une part, ou strictement positif d'autre part (disons avec une probabilité à priori de 1/2 pour chaque), et dans le cas négatif ou nul, vous n'avez pas de raison de savoir si c'est strictement négatif ou pas, ce qui fait que vous pouvez attribuer, disons, une probabilité à priori de 1/4 à chacune de ces deux possibilités. Cela correspond aux probabilités à priori résumées par le tableau suivant :

Question	$\mathbb{P}_{\mathrm{pr}}(\mu < 0)$	$\mathbb{P}_{\mathrm{pr}}(\mu = 0)$	$\mathbb{P}_{\mathrm{pr}}(\mu > 0)$
$\{\mu = 0\}$?	25 %	50 %	25 %
$\{\mu \leqslant 0\}$?	25 %	25 %	50 %

[‡]. En effet, si on avait obtenu une conclusion plausible sur $\mathrm{T}_{\mathrm{St}}(15)$ qui aurait pris une autre forme qu'une majoration, on n'aurait été en mesure de rien en déduire sur $M/(S/\sqrt{15})$, vu que tout ce qu'on a est une majoration de cette variable aléatoire par une quantité suivant la loi $\mathrm{T}_{\mathrm{St}}(15)$.

[§]. On peut même être un peu plus précis en regardant la puissance du test en termes *asymptotiques*, c.-à-d. en faisant tendre n vers l'infini : asymptotiquement, le test à droite sera rejeté à coup sûr dès que $\mu_{\checkmark} > 0$, mais pas si $\mu_{\checkmark} < 0$. Ainsi, le risque asymptotique de seconde espèce est uniformément nul (i.e., le test est consistant) sur l'hypothèse alternative $\{\mu > 0\}$, mais pas sur l'hypothèse alternative $\{\mu \neq 0\}$, qui est trop large...

[¶]. Que la question posée soit révélatrice du résultat auquel on s'attend est l'évidence même dans la vie de tous les jours : ainsi, pour mesurer les connaissances d'un enfant, on lui posera à priori des questions plus faciles que face à un adulte, parce qu'on s'attend à des performances plus faibles...

Maintenant, faisons une modélisation grossière où nous disons que la probabilité d'obtenir une valeur de l'ordre de 1,94 vaut selon le cas :

Situation	Proba conditionnelle
$\mu_✓ < 0$	1 %
$\mu_✓ = 0$	4 %
$\mu_✓ > 0$	75 %

En appliquant le théorème de Bayes, on trouve alors des probabilités à postériori de :

Question	$\mathbb{P}_{\text{post}}(\mu < 0)$	$\mathbb{P}_{\text{post}}(\mu = 0)$	$\mathbb{P}_{\text{post}}(\mu > 0)$
« $\mu = 0$? »	1,2 %	9,5 %	89,3 %
« $\mu \leqslant 0$? »	0,6 %	2,6 %	96,8 %

Conclusion : les probabilités à postériori sont tellement différentes selon la façon dont est posée la question (à cause de la différence dans les probabilités à priori), que la probabilité à postériori de $\{\mu \leqslant 0\}$ n'est que de 3,2 % lorsqu'on a posé la seconde question, alors que lorsqu'on a posé la première question, la probabilité à postériori de la seule hypothèse $\{\mu = 0\}$ vaut 9,5 %... Il n'est donc pas surprenant que le second test échoue, mais pas le premier ☺

Cette explication n'est pas du tout aussi foireuse qu'elle n'y parait : en réalité, au prix de quelques modifications mineures, on peut réinterpréter une bonne partie de la théorie des tests en termes bayésiens...

17.5 *p*-valeur v^s^ taille d'effet

Dans beaucoup de logiciels appliquant les statistiques, pour un test d'indépendance par exemple, le résultat de l'analyse est donné uniquement sous la forme de *p*-valeur. On est alors enclin à penser que lorsque la *p*-valeur est minuscule (disons, inférieure à 1×10^{-6}), l'hypothèse nulle est "carrément fausse" ; et à l'inverse, si la *p*-valeur reste raisonnable, que l'hypothèse nulle est juste, ou tout au plus "légèrement fausse". Or ce serait là une grave erreur d'interprétation :

!! *Point* (JL'). Le niveau de certitude qu'on a sur la fausseté de l'hypothèse nulle n'est pas synonyme de "à quel point" l'hypothèse nulle est fausse ! ♧

Commençons par le second sophisme (une *p*-valeur raisonnable impliquerait que l'hypothèse nulle est "presque" vraie) : nous allons réfuter celui-ci à l'aide de l'exemple suivant.

Exemple (JM'). Imaginons que je sois chargé d'établir les risque d'inondations dans différentes zones : la législation prévoyant des mesures de sécurité particulières lorsqu'il y a en moyenne au moins une inondation tous les 70 ans, je me plonge dans les archives (qui remontent, dans notre exemple, jusqu'à il y a 231 ans) pour savoir combien il y a eu d'inondations historiquement. Mon modèle est donc le suivant : $\theta_✓$ est la véritale fréquence à laquelle interviennent les inondations, qui vit dans $\Theta := \mathbb{R}_+$; X est le nombre d'inondations au cours des t dernières années avec $t := 231$ (je prends l'année pour unité de temps), qui suit la loi Poisson($\theta_✓ t$) sous $\mathbb{P}_✓$. (Il semble en effet raisonnable de supposer que, à l'échelle de plusieurs années, les inondations sont des évènements ponctuels et sans mémoire). Mon hypothèse nulle est $\{\theta \geqslant \theta_{\text{réf}}\}$ avec $\theta_{\text{réf}} = 1/70$; et dans ce cas il est clair qu'il y a un test uniformément le plus puissant, consistant à prendre pour statistique de test la valeur X elle-même, et à considérer comme suspectes (c'est-à-dire inclinant à supposer qu'il y a *peu* d'inondations) les plus petites valeurs de X. Pour finir, parmi les différentes sous-hypothèses simples composant l'hypothèse nulle, c'est évidemment

toujours dans le cas-limite $\{\theta = \theta_{\text{réf}}\}$ que la probabilité d'observer X en-dessous d'une valeur donnée sera maximale.

Cela étant établi, on calcule que la p-valeur pour observer un nombre d'inondations $n_{\checkmark}$ sur la durée des archives est égale à $\mathbb{P}(\text{Poisson}(t\theta_{\text{réf}}) \leqslant n_{\checkmark}) \underset{\text{déf}}{=} \texttt{répartPoisson}(t\theta_{\text{réf}}; n_{\checkmark}+)$. Nous supposons ici qu'on a observé $n_{\checkmark} = 1$ observation sur la période des archives, ce qui donne le calcul numérique suivant pour la p-valeur :

```
> ppois(1, 231 / 70)
[1] 0.1585976.
```

Avec une p-valeur de 16 %, l'affaire semble entendue : si on a observé 1 inondation depuis le début des archives, c'est certes plutôt peu, mais absolument pas assez pour avoir une quelconque certitude sur le fait qu'il y ait moins d'une inondation tous les 70 ans... Mais est-ce à dire que, même si la période moyenne entre deux inondations est peut-être supérieure à 70 ans, elle n'est pas "trop" supérieure pour autant ? Ce n'est pas sûr du tout non plus ! Par exemple, si la fréquence réelle des inondations est d'une tous les 1 500 ans, cela donne déjà 14 % de chances d'avoir observé une inondation dans les 231 dernières années : autrement dit, quand le risque *réel* d'inondation est d'une tous les 1 500 ans, j'ai 14 % de chances d'avoir une p-valeur supérieure à 0,16 concernant l'hypothèse nulle qu'il y a (en moyenne) une inondation tous les 70 ans ou moins ! Autrement dit, il se peut très bien, sans avoir affaire à quelque coïncidence extraordinaire que ce soit, que l'hypothèse nulle soit *complètement* fausse et que pourtant on ne puisse pas la rejeter, ou seulement "mollement". ♧

Voyons maintenant le phénomène inverse. Je me demande si la saison de naissance impacte significativement l'intelligence. Pour ce faire, j'utilise les résultats des tests de l'armée (je suppose que j'ai eu le droit d'y accéder pour mes recherches), ce qui me permet d'avoir accès à un volume considérable de données : 4 millions de tests sur les 5 dernières années ! Les tests de l'armée divisent les résultats des tests en trois catégories : « intelligence faible », « intelligence moyenne » et « intelligence supérieure ». Je fais un test du χ^2 sur toutes mes données (on peut par exemple utiliser *R* pour ce faire, le test en que non plusstion y étant pré-implémenté), et je trouve une p-valeur de

```
        Pearson's Chi-squared test

data: M
X-squared = 4452.796, df = 6, p-value < 2.2e-16.
```

Damned ! La p-valeur est absolument minuscule : il semble donc que l'effet de la saison de naissance sur l'intelligence soit absolument dramatique !... Mais regardons maintenant les données de base :

Saison	Faible	Moyenne	Supérieure
Printemps	26 %	50 %	24 %
Été	24 %	52 %	24 %
Automne	24 %	50 %	26 %
Hiver	26 %	48 %	26 %

La différence entre la distribution de l'intelligence selon la saison de naissance est en fait très faible ! Simplement, comme nous avons un très grand nombre d'observations, nous pouvons voir avec une *très grande certitude* que cette différence

n'est pas *rigoureusement* nulle ! Mais cela nous fait une belle jambe, car en réalité nous n'avons pas envie de nous exciter sur une variation entre un taux de 24 % ou de 26 %... Ce dernier s'explique vraisemblablement par le fait que, les enfants faisant tous leur rentrée en septembre, ceux nés en automne commencent l'école un chouïa plus jeunes que ceux nés au printemps, et voient donc leur cerveau stimulé par l'école un peu plus tôt, ce qui leur confère un infime avantage... mais extrêmement inférieur aux fluctuations naturelles d'un individu à l'autre : si je prends un individu né en automne et un né au printemps, il y a 32,8 % de chances que celui d'automne soit dans une meilleure catégorie que celui né au printemps, contre 29,8 % de chances que ce soit l'inverse : autant dire que c'est kif-kif !

! **Définition (JN').** Dans le vocabulaire de la statistique, on parle de *taille d'effet* pour dire *à quel point* la paramètre caché s'écarte de la situation de référence définie par l'hypothèse nulle. ♡

Le choix d'un indicateur de taille d'effet dépend du modèle et est par ailleurs affaire de convention. Dans le cas des inondations, la taille d'effet (qui quantifie le niveau de sécurité de notre zone) pourrait être définie comme le nombre de périodes de 70 ans survenant, en moyenne, entre deux inondations, i.e. comme $\theta_{\text{réf}}/\boldsymbol{\theta}$: avec cette convention, l'indicateur de taille d'effet varie entre 0 et l'infini, notre critère étant que nous voulions que cet indicateur soit supérieur à 1. En l'espèce, notre conclusion est que, avec 1 inondation sur la période des archives, nous n'avons pas de preuve *significative* que la taille d'effet soit plus grande que 1, mais qu'il reste potentiellement parfaitement *plausible* que la taille d'effet soit *beaucoup* plus grande que 1, disons de l'ordre de 20 ! En d'autres termes, nous avons encore énormément d'incertitude sur la valeur de $\boldsymbol{\theta}$, ce qui est assez logique puisque nous avons très peu d'observations dans les archives (1 seule inondation, sachant que le nombre d'inondations est forcément entier...).

Pour l'exemple des tests d'intelligence, un indicateur de taille d'effet pertinent est de quantifier à quel point la connaissance de la valeur $\theta_{\checkmark}$ permet de deviner qui, de deux conscrits pris au hasard, et dont on ne connait que la saison de naissance, a la meilleure catégorie d'intelligence. J'explique un peu mieux l'idée : nous avons un « devin » qui doit désigner le plus intelligent de deux conscrits pris au hasard : il marquera 1 point s'il a vu juste, -1 point s'il a eu faux, et 0 point si les deux conscrits ont le même niveau. Notre devin, ayant connaissance de la vraie valeur de $\theta_{\checkmark}$, donnera sa réponse à chaque fois en fonction de ce qui lui donnera le meilleur résultat en espérance. Dans ces conditions, un bon indicateur de taille d'effet serait le score moyen (exprimé en pourcentage) obtenu par le devin : cet indicateur varie naturellement entre 0 % et 100 % [‖]. Comme les données collectées par l'armée sont très volumineuses, nous avons que la taille d'effet est non nulle avec une significativité extrêmement forte ; en revanche, si nous essayons d'estimer cette taille d'effet, nous trouvons qu'elle n'est que de 1,1 %, ce qui est sans intérêt...

[‖]. En fait, pour ce modèle, le score maximal de 100 % n'est pas mathématiquement possible — même si la corrélation entre saison et intelligence est parfaite — : car il y aura forcément des cas où les deux conscrits ont le même niveau d'intelligence, ou qu'ils sont nés à la même saison, et que le devin n'aura alors aucun indice pour le guider dans sa réponse... Plus précisément, vu les effectifs dans chaque catégorie d'intelligence, on peut montrer que la valeur maximale possible pour l'indicateur de taille d'effet vaut, en l'occurrence, 62,5 %. Mais en pratique, l'ordre de grandeur de ce maximum étant comparable à 100 %, cela ne changera essentiellement rien quand il s'agira de décider si une taille d'effet donnée doit être considérée comme digne d'intérêt ou pas ☺

Chapitre 18

Visualisation des données

18.1 Représentation des distributions de probabilité

Dans un contexte d'ingénierie et de communication, il est évidemment très important d'être capable de *représenter* graphiquement les distributions de probabilité qui nous intéressent, lorsque cela est possible, et de le faire de façon *pertinente* ! Il y a deux cas om les représentations sont telement classiques que j'irais presque jusqu'à dire qu'elles *doivent* être utilisées lorsque cela est possible :

Définition (JO') (Diagramme en bâtons)**.** On représente généralement les distributions de probabilités discrètes sur (un sous-ensemble de) $\mathbb{R}$ par des *diagrammes en bâtons* : en tout point ω doté d'une masse non nulle, on trace un segment vertical basé en ω sur l'axe des abscisses et de hauteur (proportionnelle à) la masse $P(\{\omega\})$. Notez que le légendage de l'axe des ordonnées n'est pas indispensable (quoique conseillé), dans la mesure où on sait que la somme des longueurs des segments, qui représente la masse de probabilité totale, doit nécessairement valoir 1. ♡ !

Exemple (JP'). Soit $\Omega = \mathbb{R}_+$. La *distribution binomiale négative* de paramètres 3 et $\frac{1}{2}$, notée NégBin(3, $\frac{1}{2}$), est la distribution de probabilité P portée par $\Omega' = \mathbb{N}$ telle que, pour tout $n \in \mathbb{N}$,

$$P(\{n\}) = (n+1)(n+2)2^{-(n+4)} \text{ :} \qquad \text{(JQ')}$$

a ainsi $P(\{0\}) = \frac{1}{8}$, $P(\{1\}) = \frac{3}{16}$, $P(\{2\}) = \frac{3}{16}$, $P(\{3\}) = \frac{5}{32}$, ... [∗].

Cette distribution de probabilité peut être représentée par le diagramme en bâtons de la distribution de la figure 18.1.

♧

Remarque (JR'). Sur la figure 18.1, j'ai *à la fois* gradué l'axe des ordonnées et le légendage "hors-sol" d'un fragment de longueur : c'est évidemment redondant... En pratique, choisissez l'option qui a votre préférence. ♧

Définition (JS') (Graphe de densité)**.** On peut représenter une distribution de probabilité à densité sur $\mathbb{R}$ simplement en traçant le graphe de la densité f. Alors la masse attribuée par cette probabilité à une partie $A \subseteq \mathbb{R}$ correspond à l'aire comprise entre la courbe et l'axe des abscisses située au-dessous de A. De même que dans le cadre discret, l'échelle n'est pas absolument nécessaire, puisqu'elle est automatiquement impliquée par le fait que l'aire sous la courbe vaille 1. ♡ !

[∗]. On peut montrer que la somme des $P(\{n\})$ vaut bien 1 : une méthode pour ce faire consiste à évaluer en $\frac{1}{2}$ le développement en série entière de la fonction $x \mapsto \frac{1}{8}(1-x)^{-3}$.

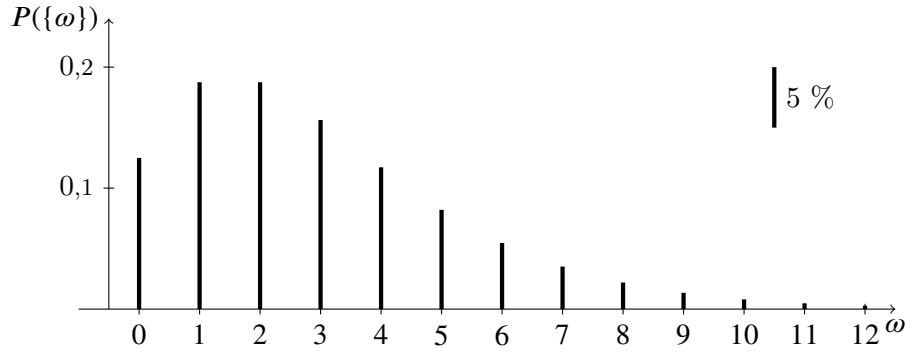

FIGURE 18.1 – Représentation par diagramme en bâtons de la distribution de probabilité NégBin(3, $\frac{1}{2}$).

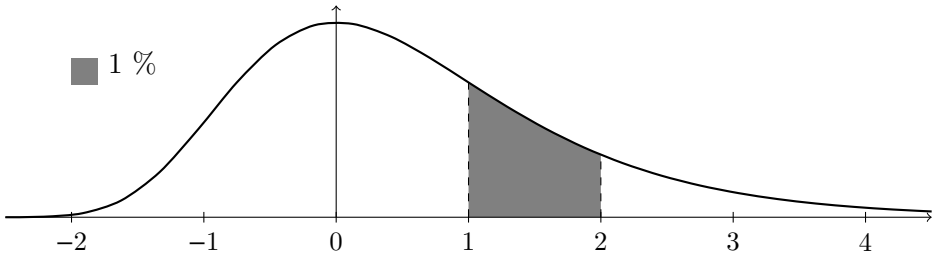

FIGURE 18.2 – Représentation graphique de la distribution de probabilité *Gumbel*(0, 1). L'aire grisée sous la courbe correspond à la masse donnée par cette distribution au segment [1, 2].

Exemple (JT'). La distribution de Gumbel standard, notée Gumbel(0, 1), est la distribution de probabilité sur $\mathbb{R}$ qui donne au voisinage infinitésimal dx de x la masse $\exp(-x - \mathrm{e}^{-x})\,\mathrm{vol}(\mathrm{d}x)$ [†]. On calcule alors que, par exemple, la masse donnée par cette distribution au segment [1, 2] est

$$\int_{x\in[1,2]} \exp(-x-\mathrm{e}^{-x})\,\mathrm{vol}(\mathrm{d}x) = \int_{x=1}^{2} \exp(-x-\mathrm{e}^{-x})\mathrm{d}x \text{ [‡]}$$
$$= \Big[\exp(-\mathrm{e}^{-x})\Big]_1^2 = \mathrm{e}^{-1/\mathrm{e}^2} - \mathrm{e}^{-1/\mathrm{e}} \approx 18{,}122\ \% ;$$

À l'inverse, la masse attribuée par cette distribution à un simple point comme {0}, ou à tout ensemble fini (ou même dénombrable) de points, serait nulle. La figure 18.2 donne la représentation par diagramme de densité d'une loi de Gumbel standard.

♧

Remarque (JU'). Vous noterez que je n'ai pas gradué l'axe des ordonnées en figure 18.2, mais que j'ai plutôt légendé "hors-sol" un petit bout de surface. En fait, je vous *déconseille* de graduer l'axe des abscisses dans un tel cas, car l'information sur la valeur de densité en elle-même est essentiellement sans intérêt, d'autant

[†]. On vérifie que cette densité s'intègre bien à 1 : $\int_{-\infty}^{\infty} \exp(-x-\mathrm{e}^{-x})\mathrm{d}x = \Big[\exp(-\mathrm{e}^{-x})\Big]_{-\infty}^{\infty} = 1-0 = 1$.

[‡]. Noter que la notation 'dx', qui se référait à une *zone* infinitésimale dans le premier membre, se réfère à un *accroissement* infinitésimal à partir du second membre.

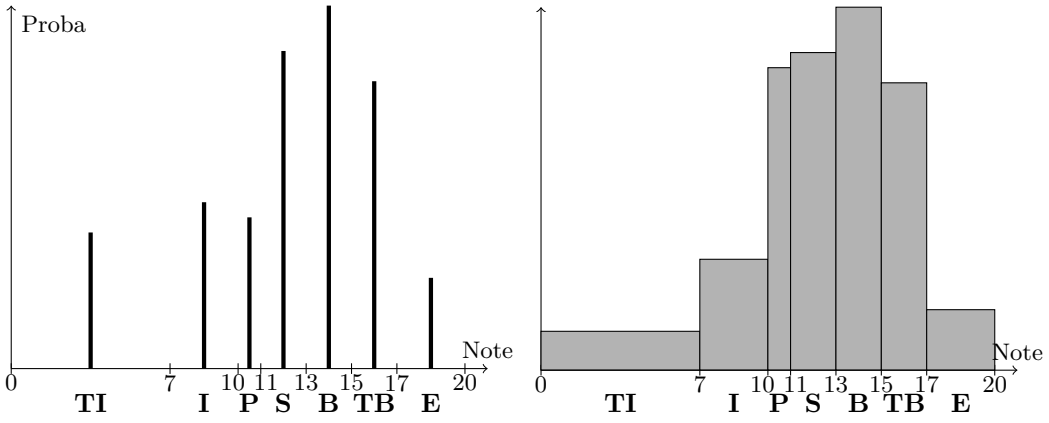

FIGURE 18.3 – Représentation de la distribution de notes de l'exemple `(JW')` selon qu'on opte pour une représentation en bâtons (à gauche) ou en histogramme (à droite).

qu'elle s'exprimera en général dans une unité bizarre et pas parlante du tout...! (Néanmoins en pratique, on ne rncontre que trop rarement le légendage par un petit bout de surface "hors-sol"). ♧

Remarque `(JV')`. Attention ; bien comprendre que les représentations des distributions de probabilité discrètes et à densité sont fondamentalement *différentes* : dans le premier cas en effet, la probabilité correspond à une *longueur*, alors que dans le second cas elle correspond à une *aire* ! C'est d'ailleurs une subtilité à prendre en compte lorsqu'on doit tracer une distribution de probabilité qu'on ne connait que sur des classes assez grossières : soit on utilise l'approche en bâtons et alors la *hauteur* des bâtons doit correspondre à la proportion de chaque classe ; soit un utilise un histogramme et alors c'est l'*aire* qui doit être proportionnelle à l'effectif — ce qui, évidemment, donne des résultats très différents dès que les classes n'ont pas toutes la même largeur ! Voir la figure **??** à titre d'illustration. ♧

Exemple `(JW')`. Supposons qu'à un examen, les mentions 'Excellent', 'Très Bien', 'Bien', 'Satisfaisant', 'Passable', 'Insuffisant' et 'Très Insuffisant' soient associées resp. aux intervalles de notes $[17, 20]$, $[15, 17[$, $[13, 15[$, $[11, 13[$, $[10, 11[$, $[7, 10[$ et $[0, 7[$, et que les proportions des notes dans les différentes mentions aient été de resp. 6 %, 19 %, 24 %, 21 %, 10 %, 11 % et 9 %. Observez comment la représentation de ces notes diffère selon qu'on les représente par un diagramme en bâtons ou par un histogramme...!

♧

Remarque `(JX')`. Lorsqu'on représente une distribution à densité sur $\mathbb{R}$, c'est toujours l'*aire* qui doit être proportionnelle à la probabilité. Il arrive dans certains cas qu'on souhaite utiliser une échelle *logarithmique* sur l'axe des abscisses : dans ce cas, conformément au principe ci-devant, la fonction tracée ne sera pas la densité de notre distribution de probabilité elle-même, mais celle de sa mesure-image par l'application logarithme... (voir plus loin). ♧

Cinquième partie

Modèles statistiques classiques

Annexe A

La régression linéaire

A.1 Régression linéaire simple

Définition (JY'). Le modèle de la régression linéaire simple s'appuie sur un paramètre caché traditionnellement noté (α, β, σ), où α, β et σ sont resp. à valeurs dans $\mathbb{R}$, $\mathbb{R}$ et $\mathbb{R}_+^*$. Il possède pour paramètres du modèle, dans sa version explicative, un entier n (que nous supposerons $\geqslant 3$) et des réels $x_0, \dots, x_{n-1}$. L'observation passée est notée $(Y_0, \dots, Y_{n-1})$; et suit la loi suivante pour la valeur (α, β, σ) du paramètre caché :

$$\vec{Y}_{[\![0,n[\![} \sim \bigotimes_{i=0}^{n-1} \text{Normale}(\alpha x_i + \beta, \sigma^2) \text{ :} \tag{JZ'}$$

autrement dit, les Y_i sont indépendants, chaque Y_i suivant une loi normale centrée sur $\alpha x_i + \beta$ et de variance σ^2 (avec la même variance pour tous les Y_i). ♡

Définition (KA'). Pour traiter le modèle de la régression linéaire simple, on a coutume de s'appuyer sur les paramètres et statistiques suivants :

- $\bar{x}$ est la moyenne empirique des x_i. (On peut donc le voir comme un paramètre du modèle, s'exprimant à partir des paramètres de base) ;
- s_x est l'écart-type empirique des x_i [*]. (Là aussi, on peut le voir comme un paramètre du modèle dérivé des paramètres de base) ;
- $\bar{Y}$ est la moyenne empirique des observations passées Y_i.
- $\boldsymbol{\alpha}$, qu'on peut appeler « pente de régression estimée », est la statistique définie par

$$\hat{\boldsymbol{\alpha}} := \frac{\text{cov}_{\text{emp}}\big((x_i, Y_i)\big)_{i \in [\![0,n[\![}}{s_x^2} \text{ ;} \tag{KB'}$$

- R est le coefficient de corrélation empirique entre les x_i et les Y_i :

$$R := \frac{\text{cov}_{\text{emp}}\big((x_i, Y_i)\big)_{i \in [\![0,n[\![}}{s_x \, \text{var}_{\text{emp}}^{1/2}(Y_i)_{i \in [\![0,n[\![}}. \tag{KC'}$$

- On appelle « droite de régression estimée » la droite (aléatoire, fonction de l'observation passée) de pente $\hat{\boldsymbol{\alpha}}$ et passant par le point de coordonnées $(\bar{x}, \bar{Y})$;
- On note $\hat{\boldsymbol{\beta}}$ l'ordonnée à l'origine de la droite de régression : $\boldsymbol{\beta} := \bar{Y} - \hat{\boldsymbol{\alpha}}\bar{x}$.

[*]. Dans de nombreux ouvrages de référence, aucune notation spécifique n'est introduite pour s_x : on écrit alors simplement « $\sum_i (x_i - \bar{x})^2$ » pour ce que nous noterons $n s_x^2$.

— Pour tout $i \in [\![0, n[\![$, on définit l'« erreur estimée » comme l'écart entre Y_i et l'ordonnée de la droite de régression à l'abscisse correspondante :

$$\hat{E}_i \coloneqq Y_i - \hat{\alpha} x_i - \hat{\beta}. \tag{KD'}$$

— On note

$$\hat{\sigma}_{\text{emp}} \coloneqq \Big(\sum_{i=0}^{n-1} \hat{E}_i^2 \,/\, n\Big)^{1/2};$$
$$\hat{\sigma}_{\text{B}} \coloneqq \Big(\sum_{i=0}^{n-1} \hat{E}_i^2 \,/\, (n-2)\Big)^{1/2}.$$

♡

Estimation de la pente de la droite de régression

Procédure (KE'). Dans ce contexte, l'estimateur standard pour α est $\hat{\alpha}$. Il correspond à la fois à l'estimateur des moindres carrés et à l'estimateur du maximum de vraisemblance. ♡

Remarque (KF'). Plus précisément, l'estimateur des moindres carrés pour le paramètre caché dans son ensemble est donné par le triplet $(\hat{\alpha}, \hat{\beta}, \bullet)$, où l'estimateur pour σ n'est pas précisé dans la mesure où tous les estimateurs possibles pour ce paramètre conduisent aux mêmes résultats en ce qui concerne les moindre carrés. En l'occurrence, appliquer l'estimation par moindre carrés consiste à trouver $\hat{\alpha}_{\checkmark}$ et $\hat{\beta}_{\checkmark}$ comme les valeurs $\hat{\alpha}$ et $\hat{\beta}$ minimisant

$$\sum_{i=0}^{n-1} \big(y_{i\checkmark} - (\hat{\alpha} x_i + \hat{\beta})\big)^2. \tag{KG'}$$

♧

Remarque (KH'). De même, l'estimateur du maximum de vraisemblance pour α provient (comme toujours) de l'estimateur du maximum pour le paramètre caché dans son ensemble : celui-ci est donné plus précisément par le triplet $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}_{\text{emp}})$. ♧

Théorème (KI'). *Dans le contexte de la régression linéaire, pour tester une hypothèse de la forme* $\{\alpha = \alpha_0\}$, $\{\alpha \leqslant \alpha_0\}$ *ou* $\{\alpha \geqslant \alpha_0\}$, *ou pour déterminer un intervalle de confiance pour* α, *on utilise la propriété suivante :*

$$\text{Loi}_{\checkmark}\left(\frac{\hat{\alpha} - \alpha_{\checkmark}}{\hat{\sigma}_{\text{emp}} \,/\, \sqrt{n-2}\, s_x}\right) = \text{T}_{\text{St}}(n-2). \tag{KJ'}$$

♢

Remarque (KK'). La statistique qui apparait au dénominateur de la variable aléatoire ci-dessus est appelée *erreur standard* du paramètre α. Ce nom provient de ce que cette quantité mesure l'ordre de grandeur typique de la différence entre $\hat{\alpha}$ et α, d'une façon que le quotient suive une certaine loi de Student. En fait, le carré de l'erreur standard correspond à un estimateur sans biais de la variance (sous la véritable loi) de l'estimateur $\hat{\alpha}$.

On notera que l'erreur standard peut aussi s'écrire sous la forme

$$\hat{\sigma}_{\text{B}} \,/\, \sqrt{n}\, s_x. \tag{KL'}$$

♧

Remarque (KM'). Le test de l'hypothèse nulle $\{\alpha = 0\}$ est particulièrement important, car le fait que $\alpha_{\checkmark}$ soit non nul, resp. nul, signifie que la valeur de l'abscisse x_i a un impact, resp. pas d'impact, sur la valeur de l'observation Y_i. En fait, $\alpha_{\checkmark}$ quantifie à quel point est-ce qu'une augmentation de x a tendance (sous la vraie loi) à entrainer une augmentation de Y. ♧

Estimation de l'intensité du bruit

Théorème (KN'). *On a la relation suivant, qui permet de simplifier le calcul de* $\hat{\sigma}$ *:*

$$n\hat{\sigma}^2 = n \operatorname{var}_{\mathrm{emp}}(y_i)_i - \hat{\alpha}^2 s_x^2. \tag{KO'}$$

◇

Procédure (KP'). Dans ce contexte, l'estimation standard pour σ est $\hat{\sigma}_{\mathrm{B}}$; parfois aussi on rencontre l'estimateur $\hat{\sigma}_{\mathrm{emp}}$. L'estimateur $\hat{\sigma}_{\mathrm{emp}}$ correspond au maximum de vraisemblance, tandis que $\hat{\sigma}_{\mathrm{B}}$ en est une variante qui présente la propriété que $\hat{\sigma}_{\mathrm{B}}^2$ est un estimateur sans biais de σ^2 (mais en tant qu'estimateur de σ lui-même, $\hat{\sigma}_{\mathrm{B}}$ est en revanche biaisé vers le haut). ♡

Théorème (KQ'). *Dans le contexte de la régression linéaire, pour tester une hypothèse de la forme* $\{\sigma = \sigma_0\}$*,* $\{\sigma \leqslant \sigma_0\}$ *ou* $\{\sigma \geqslant \sigma_0\}$*, ou pour déterminer un intervalle de confiance pour* σ*, on utilise la propriété suivante :*

$$\mathrm{Loi}_{\checkmark}\left(\frac{n\hat{\sigma}_{\mathrm{emp}}^2}{\sigma_{\checkmark}^2}\right) = \chi^2(n-2). \tag{KR'}$$

◇

Remarque (KS'). Le numérateur de la variable aléatoire ci-dessus peut aussi se ré-écrire comme $(n-2)\hat{\sigma}_{\mathrm{B}}^2$. ♧

Estimation de l'ordonnée de la droite de régression en une abscisse donnée La valeur $\beta_{\checkmark}$ correspond à l'ordonnée à l'origine de la véritable droite « $y = \alpha_{\checkmark} x + \beta$ » : pour cette raison, les logiciels anglophones l'appellent souvent « intercept » (car c'est la hauteur à laquelle la droite de régression *intercepte* l'axe des ordonnées). Cependant, en général dans les contextes industriels considérés, la valeur zéro de l'axe des abscisses ne pas de pertinence particulière : de manière générale, ce qu'on cherchera en général est valeur de la quantité d'intérêt $\alpha x_\star + \beta$ pour une certaine abscisse (connue) $x_\star$ qui pourra très bien ne pas être nulle. Il se trouve en fait que le problème n'est pas plus compliqué à résoudre pour une valeur générale de $x_\star$ que pour une valeur quelconque : c'est donc au cas général que nous nous intéresserons ici.

Procédure (KT'). Dans le contexte de la régression linéaire, l'estimateur standard pour $\alpha x_\star + \beta$ est $\hat{\alpha} x_\star + \hat{\beta}$: autrement dit, c'est l'estimateur qu'on obtient par substitution à partir des moindres carrés ou du maximum de vraisemblance. ♡

Théorème (KU'). *Dans le contexte de la régression linéaire, pour tester une hypothèse de la forme* $\{\alpha x_\star + \beta = y_{\text{réf}}\}$*,* $\{\alpha x_\star + \beta \leqslant y_{\text{réf}}\}$ *ou* $\{\alpha x_\star + \beta \geqslant y_{\text{réf}}\}$*, ou pour*

déterminer un intervalle de confiance pour $\boldsymbol{\alpha} x_\star + \boldsymbol{\beta} = y_{\text{réf}}$*, on utilise la propriété suivante :*

$$\text{Loi}_{\checkmark}\left(\frac{(\hat{\boldsymbol{\alpha}} x_\star + \hat{\boldsymbol{\beta}}) - (\alpha_{\checkmark} x_\star + \beta_{\checkmark})}{\hat{\boldsymbol{\sigma}}_{\text{emp}} / \sqrt{n-2} \times \left(1 + \left((x_\star - \bar{x}) / s_x\right)^2\right)^{1/2}} \right) = \text{T}_{\text{St}}(n-2). \qquad \text{(KV')}$$

♢

Remarque (KW'). On observe donc que l'erreur standard sur $\boldsymbol{\alpha} x_\star + \boldsymbol{\beta}$ dépend assez sensiblement de la valeur de $x_\star$: elle est minimal lorsque $x_\star$ est "au milieu" des valeurs des x_i, et commence à croitre sensiblement dès lors que $x_\star$ s'écarte de la plage des valeurs "typiques" de x_i. Informellement, on peut donc dire que la régression linéaire arrive à déterminer plus précisément l'emplacement de la droite de régression pour les abscisses correspondant au milieu du nuage de points que pour les abscisses correspondant à la périphérie du nuage de points. ♣

Remarque (KX'). Il est très important, lorsqu'on doit estimer un intervalle de confiance pour $\boldsymbol{\alpha} x_\star + \boldsymbol{\beta}$, de ne pas commencer par déterminer un intervalle de confiance sur $\boldsymbol{\beta}$ avant de « reporter » celui-ci dans la formule « $\alpha x_\star + \beta$ » : outre que cette stratégie de construction d'intervalle de confiance en deux temps est presque toujours une stratégie donnant des résultats imprécis de manière générale, ici cela peut être particulièrement catastrophique si les abscisses du nuage de point sont toutes éloignées de zéro, car dans ce cas on aura une très forte incertitude sur $\boldsymbol{\beta}$, bien plus forte que l'incertitude qu'on peut obtenir sur $\boldsymbol{\alpha} x_\star + \boldsymbol{\beta}$ en procédant convenablement... ♣

Intervalle de prédiction

Définition (KY'). On peut introduire une version prédictive du modèle de la régression linéaire : dans ce cas, on introduit un paramètre supplémentaire x_n dans le modèle ; et l'observation future, notée Y_n, suit la loi Normale$(\alpha x_n + \beta, \sigma^2)$, indépendamment des autres Y_i : il s'agit donc en substance du même modèle, mais avec une observation de plus qui n'est observée que dans le futur. ♡

Remarque (KZ'). Dans ce cas, les valeurs de $\bar{x}$, s_x et des statistiques $\bar{Y}$, &c. sont toujours déterminées à partir des indices i relatifs aux seules observations *passées.* ♣

Procédure (LA'). Le prédicteur standard pour Y_n est alors $\hat{\boldsymbol{\alpha}} x_n + \hat{\boldsymbol{\beta}}$. ♡

Théorème (LB'). *Dans ce contexte, pour déterminer un intervalle de confiance pour* Y_n*, on utilise la propriété suivante :*

$$\text{Loi}_{\checkmark}\left(\frac{(\hat{\boldsymbol{\alpha}} x_n + \hat{\boldsymbol{\beta}}) - Y_n}{\hat{\boldsymbol{\sigma}}_{\text{emp}} / \sqrt{n-2} \times \left(n + 1 + \left((x_n - \bar{x}) / s_x\right)^2\right)^{1/2}} \right) = \text{T}_{\text{St}}(n-2). \qquad \text{(LC')}$$

♢

Remarque (LD'). Pour peu que n ne soit pas trop petit, cette fois-ci l'erreur standard sur Y_n ne dépendra pas beaucoup de la valeur x_n, car la somme $n + 1 + \left((x_n - \bar{x}) / s_x\right)^2$ sera toujours proche de n en pratique. ♣

Notes bibliographiques

Jusqu'en 2016, le cours de statistique de Mines Nancy était dispensé par Thierry Verdel : son polycopié [6] (qui s'appuyait lui-même sur un document plus ancien de Claude Chambon) est toujours disponible en ligne, et constitue une approche tout à fait recommandable à la statistique pour l'ingénieur, en particulier pour découvrir les techniques statistiques particulières les plus utilisées (tests et intervalles de confiance de Student, de Fisher, du chi-deux ; régression linéaire ; analyse de la variance ; etc.), qui ne sont pas exposées dans le présent document. Cependant, l'approche est très différente entre mon polycopié (qui s'intéresse surtout aux *concepts* statistiques) et celui de M. Verdel (qui s'intéresse surtout aux *procédures*) ; c'est pourquoi je ne pense pas qu'il soit avisé de lire ces deux documents en parallèle : mais plutôt, de garder en tête l'existence du polycopié de M. Verdel pour le jour où vous aurez besoin d'en savoir davantage sur une méthode particulière !

Pour tout ce qui concerne la partie fréquentiste de ce cours, je me suis appuyé d'assez près sur deux excellents ouvrages que je vous recommande (et que vous trouverez à la bibliothèque de l'École) : celui de Cadre & Vial [1], particulièrement léger, idéal pour une introduction aux concepts de la statistique fréquentiste ; et celui de Stoltz & Rivoirard [4], plus fouillé, qui présente un grand nombre d'exemples d'applications détaillés, avec des implémentations informatiques dont les codes sont téléchargeables librement, parfait pour voir comment ces concepts se mettent en action ! Ces deux ouvrages doivent eux-mêmes beaucoup à l'école de statistique mathématique française, et notamment M^me^ Dominique Picard, dont l'auteur de ces lignes a lui-même eu la chance de suivre le cours [3] !

Pour la partie bayésienne, malheureusement, il ne semble pas exister d'ouvrage accessible (en tout cas pas en français) recoupant la deuxième partie de ce polycopié... Peut-être parce que ce qui y est dit est "trop facile" pour constituer l'objet d'un livre entier ? Peut-être à cause de la prééminence de la statistique fréquentiste au cours du XX^e^ siècle ?... Un ouvrage très célèbre est celui de Christian Robert [5] ; mais il part tout de suite d'un niveau avancé, et sera utile seulement pour celui souhaitant se spécialiser en statistique bayésienne. En fait, la statistique bayésienne étant souvent décriée comme une approche trop peu rigoureuse (mais ces critiques tendent à s'estomper, notamment à cause de l'augmentation de la puissance de calcul qui rend l'approche bayésienne particulièrement performante dans des cas où l'analyse purement mathématique rendrait les armes, en particulier en intelligence artificielle), on trouve plus facilement des ouvrages expliquant *pourquoi* la statistique bayésienne est une bonne idée, plutôt que pour détailler *comment* s'en servir ! Dans cette catégorie, outre le livre [5] déjà cité de Robert, je recommande à tous ceux qui sont intéressés par la découverte de l'extraordinaire puissance de la méthode bayésienne de lire le tout récent livre du célèbre vulgarisateur Lê Nguyên

Hoang [2], qui mêle mathématique, informatique et philosophie avec un enthousiasme passionnant !

Bibliographie

[1] Benoît Cadre et Céline Vial : *Statistique mathématique — Cours et exercices corrigés*. Ellipses, 2012. ISBN 978-2-7298-7323-3.

[2] Lê Nguyên Hoang : *La formule du savoir — Une philosophie unifiée du savoir fondée sur le théorème de Bayes*. Edp Sciences, 2018. ISBN 978-2-7598-2260-7.

[3] Dominique Picard : Statistique et modèles aléatoires. `https://www.lpsm.paris/pageperso/picard/pubps/Poly0304.pdf`, 2004.

[4] Vincent Rivoirard et Gilles Stoltz : *Statistique mathématique en action*. Vuibert, 2e édition, 2012. ISBN 978-2-311-00720-6.

[5] Christian P. Robert : *Le choix bayésien — Principes et pratique*. Springer, 2006. ISBN 978-2-287-25173-3.

[6] Thierry Verdel : Décision et prévision statistiques. `https://drive.google.com/open?id=1SAhHtaC8tvXhWO3v824kuR0ajhoP8HdT`, 2016.