Probabilités et Statistiques

Philippe Duchon & Bernard Perrot

Septembre 2005

Table des matières

Première partie

Probabilités

Chapitre 1

Théorie des probabilités

Sommaire

1.1 Présentation informelle

La théorie des probabilités est la théorie mathématique qui a pour but de définir un "degré de vraisemblance" pour le résultat d'expériences soumise à une part d'incertitude et d'aléatoire. L'origine, dans le phénomène modélisé, de cette part d'aléatoire n'est pas importante : elle peut provenir d'une incapacité assumée à calculer avec une précision suffisante un comportement déterministe (comme par exemple le mouvement d'un cube de matière qui rebondit sur une surface dure jusqu'à rester au repos ; le résultat d'un jet de dé n'est "aléatoire" que dans la mesure où la face qui termine sa course au-dessus dépend de manière suffisamment complexe et chaotique des conditions initiales du lancer), ou d'un phénomène que l'on considère comme "intrinsèquement probabiliste".

On va donc définir, pour un certain ensemble de conditions que peut vérifier le résultat (*événements*), ce degré de vraisemblance, sous la forme d'un réel positif appelé *probabilité de l'événement* ; la convention est que cette probabilité est proportionnelle au degré de certitude que l'on a que cet événement se produise lors de l'expérience.

Il existe un certain nombre de règles assez naturelles que doivent vérifier ces probabilités. Citons-en quelques unes :

- si un événement ne peut pas se produire, sa probabilité sera de 0 ;
- si un événement A se produit toujours lorsqu'un autre événement B se produit (la condition qui définit A est toujours vérifiée lorsque celle qui définit B l'est ; par exemple, dans le cas où l'événement A correspond à la condition "le nombre obtenu est impair" et l'événement B, à la condition "le nombre obtenu est un entier premier différent de 2"), la probabilité de A sera au moins égale à la probabilité de B.
- si deux événements A et B sont *incompatibles*, c'est-à-dire qu'il est impossible que les deux se produisent simultanément (par exemple, "le résultat est pair" et "le résultat est 3 ou 5"), alors la probabilité que l'un ou l'autre se produise ("le résultat est pair, ou 3, ou 5") est égal à la *somme* de leurs probabilités.

Ces conditions, et d'autres qui sont également raisonnables, correspondent à ce que fournit la théorie mathématique de la *mesure*, utilisée pour définir l'*intégrale de Lebesgue*.

On en rajoute une, sous la forme d'un facteur de normalisation : la probabilité d'un événement *certain* (dont la condition est toujours remplie) est de 1.

Le choix adopté dans ce cours est de présenter séparément (chapitres 2 et 3), d'une part le cas *discret*, où les calculs se font au moyen de sommes et de séries ; et le cas *continu*, où les calculs se font au moyen d'intégrales. Une présentation générale, faisant appel à la théorie de la mesure et de l'intégrale de Lebesgue, unifierait les deux (les séries devenant des cas particuliers d'intégrales) et ferait disparaître la distinction artificielle entre espaces discrets et espaces continues ; elle sort toutefois du cadre de ce cours.

Néanmoins, le vocabulaire de base, ainsi que les définitions générales, sont données, pour information, dans ce premier chapitre.

1.2 Vocabulaire des probabilités

1.2.1 Univers et événements

Pour une expérience donnée, on considère un certain ensemble (fini ou infini ; dans la pratique, la différence importante n'est pas entre fini et infini, mais plutôt entre ensemble fini ou dénombrable d'une part, et infini non dénombrable d'autre part), appelé *univers*, et qui représente *l'ensemble de tous les résultats possibles de l'expérience*. Les éléments de l'univers sont appelés *événements élémentaires*.

Exemple 1.1 *Si l'expérience que l'on souhaite modéliser est le lancer d'un dé cubique, et que le résultat auquel on s'intéresse est le nombre indiqué par le dé à l'issue d'un lancer, l'univers sera typiquement* $\Omega = \{1, 2, 3, 4, 5, 6\}$.

Exemple 1.2 *Si l'expérience modélisée consiste à placer un compteur Geiger à proximité d'un materiau radioactif et à mesurer le temps avant un* clic*, on doit envisager que le temps soit un réel positif arbitrairement grand ; l'univers naturel est donc* $\mathbb{R}^+ = [0, +\infty[$.

La définition d'un *espace de probabilités* va permettre de déterminer un degré de vraisemblance, appelé *probabilité*, non seulement aux événements élémentaires, mais à certaines *parties* de l'univers, que l'on désigne sous le terme général d'*événements*.

Plus précisément, l'*ensemble* $\mathcal{F} \subset \mathcal{P}(\Omega)$ des événements considérés doit satisfaire un certain nombre d'axiomes :

1. $\Omega \in \mathcal{F}$, et $\emptyset \in \mathcal{F}$;
2. si $A \in \mathcal{F}$, alors $\overline{A} \in \mathcal{F}$ ($\overline{A}$ désigne le complémentaire de A dans Ω : $\overline{A} = \Omega - A$) ;
3. si $A \in \mathcal{F}$ et $B \in \mathcal{F}$, alors $A \cup B \in \mathcal{F}$; plus généralement, si $\{A_i\}_{i \in I}$ est une famille *finie ou dénombrable* d'événements (c'est-à-dire que, pour tout $i \in I$, $A_i \in \mathcal{F}$), alors on a

$$\bigcup_{i \in I} A_i \in \mathcal{F}$$

(c'est-à-dire que l'union d'une famille finie ou dénombrable d'événements, est aussi un événement)

Un tel ensemble d'événements est appelé une *tribu*[1].

On remarque que, quelque soit l'univers Ω, l'ensemble $\mathcal{P}(\Omega)$ de *toutes* les parties de Ω est toujours une tribu. C'est d'ailleurs très souvent la tribu que l'on utilisera lorsque Ω est fini ou dénombrable.

Exercice 1.1 *Montrer que l'on obtient une définition équivalente de la notion de tribu, si l'on remplace le dernier axiome par :* l'intersection *d'une famille finie ou dénombrable d'événements est encore un événement.*

Remarque 1.3 *Si l'on retire l'axiome exigeant que le complémentaire d'un événement soit également un événement, on obtient la définition d'une* topologie *sur* Ω.

La tribu des événements représente l'ensemble de tous les résultats que l'on considère comme "observables".

Il faut penser à un événement comme à une *condition* que vérifient certains (parfois aucun) des événements élémentaires, un événement étant formé de *tous les événements élémentaires qui vérifient la condition.* Ainsi, dans le cas du lancer de dé, l'événement "le résultat est pair" correspond à $A = \{2, 4, 6\}$.

Deux événements sont dits *incompatibles*, si *leur intersection est vide* ; cela correspond à deux conditions telles qu'aucun résultat possible ne vérifie les deux conditions.

Plus généralement, le vocabulaire logique sur les conditions se traduit exactement par des opérations sur les ensembles événements, comme indiqué dans le tableau suivant.

disjonction disjonction dénombrable	$A \cup B$ $\cup_n A_n$	A se produit, ou B se produit un au moins des A_n se produit
conjonction conjonction dénombrable	$A \cap B$ $\cap_n A_n$	A et B se produisent tous les deux tous les A_n se produisent
complémentation événement impossible événements incompatibles	$\overline{A} = \Omega - A$ $\emptyset$ $A \cap B = \emptyset$	A ne se produit pas ne peut pas se produire A et B ne peuvent pas se produire simultanément
inclusion	$A \subset B$	chaque fois que A se produit, B se produit
partition	$\Omega = \cup_n A_n$, $A_i \cap A_j = \emptyset$	exactement un des A_n se produit toujours

[1] On parle également parfois de *σ-algèbre* ; si l'axiome sur l'union dénombrable est remplacé par un axiome, plus faible, demandant que l'union d'une famille *finie* d'événements soit un événement, on parle alors d'*algèbre d'ensembles.*

Exercice 1.2 *On considère une suite infinie $(A_n)_{n\geq 1}$ d'événements. On pose*

$$B = \bigcup_{n\geq 1} \left(\bigcap_{m\geq n} A_n \right),$$

et

$$C = \bigcap_{n\geq 1} \left(\bigcup_{m\geq n} A_n \right).$$

Montrer que B et C sont des événements, et les décrire chacun d'une phrase par rapport aux A_n, à la manière du tableau précédent. Montrer que B et C ne dépendent pas de l'ordre dans lequel sont numérotés les A_n. (B est appelé limite inférieure des A_n, et noté $\liminf A_n$; C est appelé limite supérieure des A_n, et noté $\limsup A_n$.)

Tribu engendrée, tribu borélienne

Il est facile de vérifier que l'intersection de deux tribus, ou même d'un ensemble infini de tribus, est encore une tribu. Par conséquent, si $\mathcal{C}$ est un ensemble quelconque de parties de l'univers Ω, *l'intersection de toutes les tribus contenant $\mathcal{C}$* (il en existe toujours au moins une : la tribu de toutes les parties de Ω) est encore une tribu. Cette tribu, appelée *tribu engendrée par $\mathcal{C}$*, et notée $\sigma(\mathcal{C})$, est la plus petite tribu contenant $\mathcal{C}$: toute tribu contenant $\mathcal{C}$ contient $\sigma(\mathcal{C})$.

La notion de tribu engendrée permet de définir aisément des tribus, en ne spécifiant que le "minimum" d'événements : on se contente de décrire l'ensemble $\mathcal{C}$ des événements dont on a "besoin", et on choisit la tribu $\sigma(\mathcal{C})$.

C'est particulièrement utile si l'espace Ω possède déja, par exemple, une structure d'*espace métrique* : la *tribu borélienne* est la tribu engendrée par les *ensembles ouverts*.

On note $\mathcal{B}$ (respectivement, $\mathcal{B}(\mathbb{R}^k)$ la tribu borélienne sur $\mathbb{R}$ (respectivement, $\mathbb{R}^k$).

1.2.2 Mesure de probabilités, espace probabilisé

Une fois déterminés l'univers de tous les événements élémentaires, ainsi que la tribu des événements, il reste à choisir une *loi de probabilités*.

Définition 1.4 *Une loi (ou mesure) de probabilités sur une tribu $\mathcal{F}$ est une application de $\mathcal{F}$ dans $[0,1]$, typiquement notée $\mathbb{P}$, et qui vérifie les axiomes suivants :*

1. $\mathbb{P}(\Omega) = 1$;
2. *(σ-additivité) si $(A_i)_{i\in I}$ est une famille finie ou dénombrable d'événements* deux à deux incompatibles *(c'est-à-dire que l'on a $A_i \cap A_j = \emptyset$ dès que $i \neq j$), alors on a*

$$\mathbb{P}\left(\bigcup_{i\in I} A_i \right) = \sum_{i\in I} \mathbb{P}(A_i)$$

Exercice 1.3 *Montrer que la définition d'une mesure de probabilités implique les affirmations suivantes :*

- $\mathbb{P}(\emptyset) = 0$;

– *si A est un événement quelconque, alors*

$$\mathbb{P}(\overline{A}) = 1 - \mathbb{P}(A);$$

– *si A et B sont des événements incompatibles, alors on a*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B);$$

– *si A et B sont des événements quelconques, alors on a*

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

Définition 1.5 *Le couple $(\Omega, \mathcal{F})$ (où $\mathcal{F}$ est une tribu sur Ω) est appelé* espace mesurable *; le triplet $(\Omega, \mathcal{F}, \mathbb{P})$ (où $(\Omega, \mathcal{F})$ est un espace mesurable, et $\mathbb{P}$ une mesure de probabilités sur $\mathcal{F}$) est appelé* espace probabilisé *ou* espace de probabilités.

Exemple 1.6 *Reprenons l'exemple du lancer d'un dé. Avec l'univers $\Omega = \{1, 2, 3, 4, 5, 6\}$, on utilise comme tribu, l'ensemble $\mathcal{P}(\Omega)$ de toutes ses parties, et, dans le cas d'un dé non pipé, on définit la loi de probabilité $\mathbb{P}$ de telle sorte que chaque singleton ait la même probabilité. Il est aisé de vérifier que cela détermine entièrement $\mathbb{P}$, et que l'on a alors, pour tout ensemble $A \subset \Omega$,*

$$\mathbb{P}(A) = \frac{\#A}{6}.$$

Exemple 1.7 *Dans le cas du compteur Geiger, $\Omega = \mathbb{R}^+$. La tribu naturelle n'est pas l'ensemble de toutes les parties de Ω, mais la* tribu des boréliens, *qui peut être définie comme "la plus petite tribu contenant tous les intervalles" (il est facile de démontrer que cette description détermine bien une unique tribu ; il n'est en revanche pas facile[2] de démontrer que cette tribu n'est pas celle de toutes les parties de Ω).*

Reste à décrire la loi de probabilités. La physique prévoit que, pour chaque événement intervalle *(de la forme $A = [t_1, t_2]$), on ait*

$$\mathbb{P}(A) = \lambda \left(e^{-t_1/\lambda} - e^{-t_2/\lambda} \right),$$

où $\lambda > 0$ est un paramètre qui dépend du matériau considéré. Il se trouve que cette condition détermine entièrement la probabilité $\mathbb{P}$ sur toute la tribu des boréliens.

1.2.3 Événements indépendants

La notion d'*indépendance* entre événements, modélise la situation où le fait qu'un événement se produise ou non, n'a aucune influence sur le fait que l'autre se réalise ou non. Un exemple type est celui du lancer simultané de deux dés de différentes couleurs (afin qu'il soit clair que l'on puisse distinguer le "dé bleu" du "dé rouge") : si les deux dés ne s'influencent pas l'un l'autre (ce qui serait le cas, par exemple, s'ils étaient aimantés, les deux pôles Nord des aimants correspondant aux deux faces 1), on s'attend à ce que tout événement portant uniquement sur le premier dé ("le dé bleu donne un résultat pair") soit indépendant de tout événement ne portant que sur le second ("le dé rouge donne 1").

[2] Cela nécessite d'ailleurs l'axiome du choix...

Définition 1.8 *Deux événements A et B sont indépendants, si l'on a*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B).$$

Plus généralement, les éléments d'une famille $(A_i)_{i\in I}$ finie ou infinie sont indépendants, si l'on a, pour toute partie finie $J \subset I$,

$$\mathbb{P}\left(\bigcap_{i\in J} A_i\right) = \prod_{i\in J} \mathbb{P}(A_i).$$

Ainsi, l'indépendance d'événements est une notion qui ne concerne pas seulement la tribu, mais la mesure de probabilité.

Exercice 1.4 *L'espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ est a priori quelconque.*

1. *Soient A et B deux événements. Montrer que A et B sont indépendants si et seulement si A et $\overline{B}$ le sont, si et seulement si $\overline{A}$ et $\overline{B}$ le sont.*
2. *Généralisation à une famille finie quelconque : soit $(A_i)_{1\leq i\leq k}$ une famille de k événements indépendants, et soit $k' \leq k$. On pose, pour tout $i \leq k$,*
   – $A'_i = A_i$ si $i \leq k'$,
   – $A'_i = \overline{A_i}$ si $i > k'$.
   Montrer que les événements de la famille $(A'_i)_{1\leq i\leq k}$ sont indépendants.
3. *Généralisation à une famille quelconque : soit $(A_i)_{i\in I}$ une famille quelconque d'événements indépendants, et soit $(A'_i)_{i\in I}$ une famille d'événements telle que, pour tout i, on ait $A'_i = A_i$ ou $A'_i = \overline{A_i}$. Montrer que $(A_i)_{i\in I}$ est une famille d'événements indépendants si et seulement si $(A'_i)_{i\in I}$ en est une.*

Il est important de retenir que l'indépendance n'est pas une relation transitive : il se peut parfaitement que A soit indépendant de B et B indépendant de C, mais que A ne soit pas indépendant de C (le cas le plus simple étant $A = C$: si $\mathbb{P}(A) \neq 0$ et $\mathbb{P}(A) \neq 1$, A n'est pas indépendant de lui-même).

Plus subtilement, l'indépendance n'est pas la même chose que l'*indépendance k à k.*

Définition 1.9 *Les éléments d'une famille $(A_i)_{i\in I}$, finie ou infinie, sont* indépendants k à k *($k \geq 2$), si, pour tout $2 \leq k' \leq k$ et toute sous-famille $(A_i)_{i\in J}$ de taille k' (c'est-à-dire $\#J = k'$), les événements de la sous-famille sont indépendants.*

Exercice 1.5 *Décrire un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$, et trois événements A, B, et C dans cet espace, qui soient indépendants 2 à 2, mais pas indépendants.*

1.2.4 Probabilités conditionnelles

La notion de probabilité conditionnelle sert, entre autres, à modéliser des situations où une expérience bien déterminée doit avoir lieu (l'espace de probabilités est "connu"), mais, sans avoir le résultat exact (la donnée d'un événement élémentaire dans l'univers), on a une information partielle (on sait qu'un événement donné se produit, *i.e.*, l'événement élémentaire est contraint à appartenir à un certain événement. La question est alors de savoir en quoi cette information nous force à modifier notre estimation du degré de vraisemblance d'autres événements (la loi de probabilités).

Les enfants du voisin de M. Martin

Commençons par une petite histoire, afin de fixer les idées.

M. Martin a un voisin, avec lequel il discute de temps en temps, sans être forcément très familier avec la vie privée de celui-ci. Il sait déjà que ce voisin a exactement deux enfants, mais il n'en sait pas plus. En mathématicien amateur mais éclairé, M. Martin estime que la probabilité que l'ainé soit un garçon est de $1/2$, qu'il en est de même pour le cadet, et que les événements en question sont indépendants. Il en déduit aisément que la probabilité que l'un au moins des enfants de son voisin soit une fille, est de $3/4$.

Lors d'une conversation entre voisin, M. Martin apprend ("Mon fils Bob est malade") que l'un au moins des enfants de son voisin est un garçon. Il révise rapidement son estimation : la probabilité que son voisin ait au moins une fille est maintenant de...[3]

Quelques jours plus tard, M. Martin en apprend un peu plus sur la famille de ses voisins : il s'avère que Bob est l'ainé des enfants. De nouveau, il réévalue son estimation : la probabilité que l'un au moins des enfants du voisin est maintenant de...[4]

Dans cette situation, il faut bien être conscient de ce que les calculs de probabilités ne font que refléter l'ignorance de M. Martin : tout au long de l'histoire, les sexes des enfants du voisin sont parfaitement déterminés, et donc nullement aléatoires, et, à tout moment, la probabilité "réelle" que le voisin ait une fille parmi ses enfants est soit de 0 (s'il n'en a pas), soit de 1 (s'il en a), mais inconnue. Les estimations successives ne prétendent que refléter, respectivement, "la probabilité qu'un couple ait au moins une fille, sachant qu'il a exactement deux enfants", "la probabilité qu'un couple ait au moins une fille, sachant qu'il a exactement deux enfants dont au moins un est un garçon", et "la probabilité qu'un couple ait au moins une fille, sachant qu'il a exactement deux enfants et que l'ainé est un garçon". (Il y a en fait deux autres hypothèses cachées, et qui sont légèrement fausses en réalité : d'une part, que les naissances de garçons et de filles sont équiprobables ; d'autre part, que les sexes d'enfants successifs d'un même couple sont indépendants.)

Probabilité conditionnelle

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilités, et soit $A \in \mathcal{F}$ un événement dont on suppose seulement $\mathbb{P}(A) > 0$.

Définition 1.10 *Soit $B \in \mathcal{F}$ un événement quelconque.*
On appelle probabilité de B sachant A, *et on note $\mathbb{P}(B|A)$, le réel*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

La raison pour laquelle on parle encore de probabilité est contenue dans la propriété suivante.

Proposition 1.11 *L'application*

$$\begin{array}{rccc} \mathbb{P}' : & \mathcal{F} & \to & [0,1] \\ & F & \mapsto & \mathbb{P}(F|A) \end{array}$$

est une loi de pobabilité sur $\mathcal{F}$, appelée loi de probabilité sachant A.

[3] $2/3$ (à vérifier !)
[4] $1/2$

Preuve: Le fait que les valeurs prises par $\mathbb{P}'$ soient comprises entre 0 et 1 est trivial :

$$\begin{aligned}\mathbb{P}'(B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \\ &\leq \frac{\mathbb{P}(A)}{\mathbb{P}(A)} = 1\end{aligned}$$

puisque $A \cap B \subset A$, ce qui implique $\mathbb{P}(A \cap B) \leq \mathbb{P}(A)$.

De même, il est immédiat de vérifier que l'on a bien $\mathbb{P}'(\Omega) = 1$. Reste donc à vérifier la σ-additivité.

Soit $(B_i)_{i \in I}$ une famille d'événements deux à deux disjoints, et $B = \cup_{i \in I} B_i$.

Posons $C_i = B_i \cap A$. D'après la définition d'une tribu, les C_i sont des événements, et puisque les B_i sont deux à deux disjoints, les C_i le sont également. Par conséquent, on a bien

$$\mathbb{P}(B \cap A) = \mathbb{P}\left(\bigcup_{i \in I} C_i\right) = \sum_{i \in I} \mathbb{P}(C_i) = \sum_{i \in I} \mathbb{P}(B_i \cap A).$$

En divisant par $\mathbb{P}(A)$, on obtient

$$\mathbb{P}(B|A) = \sum_{i \in I} \mathbb{P}(B_i|A),$$

ce qui démontre la σ-additivité et termine la preuve. □

Remarque 1.12 *Lorsque l'événement A est de probabilité strictement positive, l'indépendance de A et B est équivalente à $\mathbb{P}(B|A) = \mathbb{P}(B)$; on retrouve l'idée que "savoir que A se produit, ne modifie pas la probabilité de B".*

1.2.5 Formule des probabilités totales

Dans de nombreuses situations, il est plus facile d'évaluer la probabilité d'un événement B sachant qu'un événement A se produit, plutôt que d'évaluer directement la probabilité de B. Lorsque l'on est capable de faire cela pour un ensemble d'événements A_i qui forment une partition de l'univers, la *formule des probabilités totales* permet[5] d'exprimer la probabilité de l'événement en dehors de tout conditionnement.

Proposition 1.13 (Formule des probabilités totales) *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilités, et soit $(A_i)_{i \in I}$ une famille, finie ou dénombrable, d'événements deux à deux disjoints et qui forme une* partition *de Ω (i.e., $\cup_i A_i = \Omega$).*

Alors on a, pour tout événement B,

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(A_i)\mathbb{P}(B|A_i)$$

(si certains des A_i sont de probabilité nulle, alors $\mathbb{P}(B|A_i)$ n'est pas défini, mais le terme correspondant est considéré comme nul)

[5] Ce n'est pas forcément sa seule utilité !

Preuve: Les A_i formant une partition de Ω, les $A_i \cap B$ forment une partition de B. Par définition d'une loi de probabilités, on a donc

$$\mathbb{P}(B) = \sum_{i \in I} \mathbb{P}(B \cap A_i).$$

Comme on a, pour tout i (d'après la définition de la probabilité conditionnelle, si A_i est de probabilité strictement positive ; sinon, par la convention énoncée) $\mathbb{P}(B \cap A_i) = \mathbb{P}(A_i)\mathbb{P}(B|A_i)$, on obtient immédiatement la formule des probabilités totales. □

1.2.6 Conditionnements successifs

Considérons une suite décroissante d'événements : $(A_n)_{n \geq 1}$, avec la condition $A_{n+1} \subset A_n$. Cela correspond au cas où la condition qui définit A_{n+1} est toujours plus stricte que celle qui définit A_n.

Proposition 1.14 (Formule des conditionnements successifs) *Dans les conditions précédentes, on a, pour tout entier N,*

$$\begin{aligned} \mathbb{P}(A_N) &= \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_2)\dots\mathbb{P}(A_N|A_{N-1} \\ &= \mathbb{P}(A_1)\prod_{k=1}^{N-1}\mathbb{P}(A_{k+1}|A_k). \end{aligned}$$

Preuve: La preuve se fait par récurrence sur N (c'est un bon exercice que de la rédiger proprement). □

1.2.7 Formules de Bayes

Les deux versions proposées ici de la formule de Bayes s'utilisent lorsque l'on veut "inverser" un conditionnement : dans les deux cas, on exprime une probabilité *sachant B*, en partant de *la probabilité de B sachant un autre événement.*

Proposition 1.15 (Formule de Bayes simple) *Soient A et B deux événements, tous deux de probabilité strictement positive. Alors on a*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A)\mathbb{P}(B|A)}{\mathbb{P}(B)}.$$

Proposition 1.16 (Formule de Bayes composée) *Soit B un événement de probabilité strictement positive, et $(A_i)_{i \in I}$ une* partition *(finie ou dénombrable) de l'univers en événements de probabilités toutes strictement positives. Alors on a, pour tout $j \in I$,*

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j)\mathbb{P}(B|A_j)}{\sum_{i \in I}\mathbb{P}(A_i)\mathbb{P}(B|A_i)}.$$

Preuve: (des deux propositions) Dans les deux cas, le numérateur est égal à $\mathbb{P}(A \cap B)$, et le dénominateur, à $\mathbb{P}(B)$. □

1.3 Variables aléatoires

La notion, fondamentale, de "variable aléatoire" représente toute grandeur, le plus souvent numérique, que l'on peut être amené à évaluer sur le *résultat* d'une expérience aléatoire. C'est donc, tout naturellement, une *fonction* définie sur l'ensemble de l'univers Ω, et à valeurs dans $\mathbb{R}$.

Il y a là, d'une certaine façon, une contradiction dans les termes : une "variable aléatoire" n'est pas une variable, c'est plutôt une fonction ; et elle n'a rien en elle-même d'aléatoire, c'est l'espace sur lequel elle agit qui est probabiliste.

L'image à avoir en tête est la suivante : un ensemble est déterminé (l'univers de tous les événements élémentaires), sur lequel sont définies une ou plusieurs fonctions jugées intéressantes. Le "doigt invisible du destin" choisit "au hasard" (selon la distribution de probabilité $\mathbb{P}$ – c'est ici que l'analogie se mord la queue, la théorie des probabilités ne parlant *in fine* que de la théorie des probabilités) **un** élément de l'univers (l'événement élémentaire qui se réalise) ; et chaque variable aléatoire fournit alors une valeur, qui n'est aléatoire que parce que l'élément de l'univers où elle a été évaluée est lui-même aléatoire.

Il existe toutefois une condition, qui peut paraître technique, qui fait que *toute* fonction réelle définie sur l'univers n'est pas forcément une variable aléatoire : *l'image réciproque par la fonction de tout ensemble ouvert de $\mathbb{R}$, doit être un élément de la tribu $\mathcal{F}$.* Cette condition, qui porte le nom de *mesurabilité* (on parle de *fonction mesurable*), ne pose généralement aucun problème lorsque l'univers est fini ou dénombrable, et, dans la pratique, il sera généralement admis, même dans des univers non dénombrables, qu'elle est remplie pour toutes les fonctions que l'on considèrera.

Définition 1.17 *Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé. On appelle* variable aléatoire réelle, *toute fonction mesurable définie que Ω, à valeurs dans $\mathbb{R}$.*

1.3.1 Quelques exemples

Variable indicatrice d'un événement

Soit A un événement (une partie de Ω) quelconque. On note $\mathbf{1}_A$, la fonction définie sur Ω par

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{si } \omega \in A \\ 0 & \text{si } \omega \notin A \end{cases}$$

Cette variable aléatoire est appelée *indicatrice de l'événement A*. On a bien évidemment $\mathbb{P}(\mathbf{1}_A = 1) = \mathbb{P}(A)$, et par conséquent, $\mathbb{P}(\mathbf{1}_A = 0) = 1 - \mathbb{P}(A)$.

Lancer de deux dés

Construisons un espace de probabilités permettant de modéliser le lancer de deux dés cubiques non pipés. Les deux dés sont des objets physiques distinguables, on prendra donc comme univers l'ensemble des *couples* d'entiers de 1 à 6 :

$$\Omega = \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\}.$$

L'univers étant fini, nous allons nous simplifier la vie et prendre comme tribu, la tribu de toutes les parties de Ω (ce qui, Ω étant de cardinal 36, nous donne donc un gentil total de 2^{36} événements, soit une bonne soixantaine de milliards).

Définissons rapidement la loi de probabilité à mettre sur notre espace : ce sera la *loi uniforme*, qui accorde à chacun des 36 événements élémentaires un poids identique, et est donc définie par

$$\mathbb{P}(A) = \frac{\#A}{36}$$

pour toute partie A de Ω.

La *variable aléatoire* "résultat du premier dé", que nous noterons X_1 (il est de tradition d'utiliser des lettres majuscules pour les variables aléatoires), sera donc définie par

$$\begin{array}{rccc} X_1 : & \Omega & \to & \mathbb{R} \\ & (i,j) & \mapsto & i \end{array}$$

De même, la variable aléatoire "résultat du deuxième dé" sera X_2 :

$$\begin{array}{rccc} X_2 : & \Omega & \to & \mathbb{R} \\ & (i,j) & \mapsto & j \end{array}$$

Exercice 1.6 *Définir formellement les variables aléatoires "somme des résultats des deux dés" et "plus petit des résultats des deux dés".*

Exercice 1.7 *Vérifier que, dans l'espace de probabilités défini, les événements "le premier dé donne un résultat pair" et "le deuxième dé donne 2" sont indépendants.*

Les événements "le premier dé donne 3" et "la somme des deux dés est 7" sont-ils indépendants ?

1.3.2 Loi d'une variable aléatoire

Une remarque sur les notations

Considérons une variable aléatoire réelle X, définie sur un espace $(\Omega, \mathcal{F}, \mathbb{P})$.

D'après la définition d'une variable aléatoire, dès lors que A est un ensemble borélien de $\mathbb{R}$ (par exemple, un intervalle, ou une réunion finie ou dénombrable d'intervalles), l'*ensemble des événements élémentaires dont l'image par X se trouve dans A* doit être un élément de $\mathcal{F}$. Cet événement, qui est formellement $X^{-1}(A)$, est noté $\{X \in A\}$.

$$\{X \in A\} = X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

Plus généralement, on utilisera des notations comme

$$\{a \leq X < b\} = X^{-1}([a,b[) = \{\omega \in \Omega : a \leq X(\omega) < b\}.$$

Loi d'une variable aléatoire, fonction de répartition

Pour tout ensemble $A \in \mathcal{B}$, l'ensemble $\{X \in A\}$ est un événement de $\mathcal{F}$; il a donc une *probabilité*, $\mathbb{P}(\{X \in A\})$, que l'on note $\mathbb{P}(X \in A)$ pour alléger les notations.

Il est immédiat de vérifier que l'application

$$\begin{array}{ccc} \mathcal{B} & \to & \mathbb{R} \\ A & \mapsto & \mathbb{P}(X \in A) \end{array}$$

est elle-même une mesure de probabilité sur l'espace $(\mathbb{R}, \mathcal{B})$; on parle de *mesure image* de $\mathbb{P}$ par X, ou, plus souvent, de *loi de la variable aléatoire* X.

La loi d'une variable aléatoire est entièrement déterminée par la donnée des *probabilités d'intervalles* : la donnée de $\mathbb{P}(X \in I)$ pour tous les *intervalles* I. On peut même aller plus loin : *la loi d'une variable aléatoire est entièrement déterminée par la donnée de* $\mathbb{P}(X \leq x)$ *pour* $x \in \mathbb{R}$.

La fonction $F_X : x \mapsto \mathbb{P}(X \leq x)$ est appelée *fonction de répartition de la variable aléatoire* X ; la fonction de répartition caractérise la loi de la variable aléatoire.

Proposition 1.18 *La fonction de répartition F d'une variable aléatoire vérifie :*

1. *F est croissante sur $\mathbb{R}$;*
2. *$\lim_{x \to +\infty} F(x) = 1$;*
3. *$\lim_{x \to -\infty} F(x) = 0$;*
4. *F est* continue à droite *en tout point x :*

$$F(x) = \lim_{y \to x^+} F(y);$$

La croissance implique de plus que

$$F(x^-) = \lim_{y \to x^-} F(y)$$

est bien définie pour tout réel x, et que l'on a $F(x^-) \leq F(x)$.

De plus, on a, pour tout x, $\mathbb{P}(X < x) = F(x^-)$ et $\mathbb{P}(X = x) = F(x) - F(x^-)$.

Enfin, toute fonction F qui vérifie les assertions 1 à 4 est la fonction de répartition d'une variable aléatoire définie sur un certain espace de probabilités.

Dans la pratique, il n'est pas rare que l'on ne décrive une variable aléatoire que par sa loi, sans se préoccuper de décrire explicitement l'espace de probabilités sur lequel elle est définie.

1.3.3 Variables aléatoires indépendantes

Deux variables aléatoires X et Y, définies sur le même espace, sont dites *indépendantes*, si, pour tous les choix d'intervalles I et J, les événements $\{X \in I\}$ et $\{Y \in J\}$ sont indépendants.

Cela implique que cette indépendance d'événements reste vraie si I et J ne sont pas des intervalles, mais des ensembles boréliens quelconques (donc, en particulier, I et J peuvent être n'importe quelle réunion dénombrable d'intervalles).

La définition s'étend immédiatement au cas de n variables aléatoires ; pour une famille infinie de variables aléatoires, on recourt à la même "astuce" que pour les familles infinies d'événements : une infinité de variables aléatoires sont indépendantes, si toutes les sous-familles finies sont indépendantes.

Comme pour les événements, la notion d'indépendance des variables aléatoires correspond à la situation où chacune des deux variables est le fruit d'une expérience distincte, les expériences en question ne pouvant pas s'influencer l'une l'autre (par exemple, deux lancers successifs d'un même dé). On a toutefois besoin, pour pouvoir raisonner sur les deux variables simultanément (par exemple, calculer la probabilité qu'un au moins des deux dés donne 6), de supposer que les deux sont définies sur le *même* espace de probabilités, qui représente alors l'ensemble de tous les résultats possibles des *deux* expériences, prises comme une seule "grosse" expérience.

De plus, il arrive que deux variables aléatoires se trouvent être indépendantes, alors qu'elles dépendent conjointement de deux mêmes autres variables (voir, par exemple, la construction du paragraphe 5.2.4, page 59).

1.3.4 Loi d'un couple de variables, lois marginales

Soient X et Y deux variables aléatoires, définies sur le même espace. On peut voir le couple (X, Y) (couple de fonctions) comme une sorte de variable aléatoire à valeurs dans $\mathbb{R}^2$.

Loi conjointe

La *loi du couple* (X, Y), ou *loi conjointe* des deux variables aléatoires, est la loi de probabilités $\mathbb{P}_{X,Y}$ sur $(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$, définie par

$$\mathbb{P}_{X,Y}(A) = \mathbb{P}((X, Y) \in A) = \mathbb{P}(\{\omega \in \Omega : (X(\omega), Y(\omega)) \in A\}).$$

De même que la loi d'une variable aléatoire est caractérisée par les probabilités des intervalles, la loi d'un couple de variables aléatoires est caractérisée par les $\mathbb{P}_{X,Y}(U \times V)$, où U et V sont des intervalles quelconques.

Lois marginales

La loi du couple (X, Y) caractérise les lois de X et de Y, appelées *lois marginales*.

Exercice 1.8 *Exprimer explicitement la loi marginale de X : étant donné un borélien quelconque A, donner l'expression de $\mathbb{P}_X(A)$ au moyen de $\mathbb{P}_{X,Y}$.*

En revanche, la réciproque est fausse : la connaissance des lois marginales de X et de Y ne permet absolument pas, en général, de déterminer la loi du couple (X, Y). La seule exception notable est lorsque les variables X et Y sont indépendantes : *si deux variables aléatoires sont indépendantes, leurs lois marginales caractérisent la loi du couple.*

Chapitre 2

Lois de probabilités discrètes

Sommaire

L'ensemble de toutes les lois de probabilités réelles que l'on peut potentiellement décrire est extrêmement vaste, et contient des exemples "pathologiques" dont l'intérêt pratique est limité[1]. Dans ce cours, nous nous limiterons à deux sous-classes beaucoup plus simples, mais qui à elles deux recouvrent une grande partie des cas pratiquement intéressants : les probabilités discrètes, qui font l'objet de ce chapitre, et les probabilités diffuses à densité continue, qui seront étudiées au chapitre suivant.

[1] On sait qu'il y a équivalence entre loi de probabilités et fonction de répartition ; on se fera donc une idée de "à quel point une loi de probabilités peut être tordue" en réfléchissant "à quel point une fonction croissante, continue à droite, peut être tordue" ; la réponse est "assez tordue".

2.1 Espaces de probabilités discrets

On appelle "discret" un espace de probabilités dont l'univers Ω est fini ou dénombrable. Dans l'immense majorité des cas, la tribu est alors la tribu $\mathcal{P}(\Omega)$ de toutes les parties de Ω. Enfin, pour préciser la loi de probabilités, il suffit de préciser la probabilité de chaque singleton.

Définition 2.1 *Un espace de probabilités discret est défini par la donnée d'un ensemble fini ou dénombrable Ω, et d'une application*

$$\mathbb{P} : \Omega \to \mathbb{R}$$

qui vérifie

1. $\mathbb{P}(\omega) \geq 0$ *pour tout* $\omega \in \Omega$;
2. $\sum_{\omega\in\Omega} \mathbb{P}(\omega) = 1$.

La tribu est $\mathcal{F} = \mathcal{P}(\Omega)$, et la probabilité est étendue à $\mathcal{F}$ par la formule de somme :

$$\mathbb{P}(A) = \sum_{\omega\in A} \mathbb{P}(\omega).$$

2.2 Variables aléatoires discrètes

Une variable aléatoire définie sur un espace discret, est appelée *variable aléatoire discrète*. En tant que fonction définie sur un espace fini ou dénombrable, il est clair qu'une telle variable aléatoire ne peut prendre qu'un ensemble fini ou dénombrable de valeurs[2].

2.3 Espérance d'une variable aléatoire discrète

Soit X une variable aléatoire discrète, qui prend ses valeurs dans un ensemble $E = \{x_n, n \in \mathbb{N}\} \subset \mathbb{R}$.

L'*espérance* (ou *espérance mathématique*) de X, notée $\mathbb{E}(X)$, est définie par la formule

$$\mathbb{E}(X) = \sum_n x_n \mathbb{P}(X = x_n),$$

sous réserve que cette série soit absolument convergente[3] :

$$\sum_n |x_n| \mathbb{P}(X = x_n) < +\infty.$$

La justification du terme d'*espérance* tient dans les propriétés asymptotiques comme la "loi des grands nombres" qui sera vue au Chapitre 4 ; c'est, en quelque mots, la valeur moyenne que l'on peut s'attendre à voir prendre par la variable aléatoire X (ce qui ne veut absolument pas dire que l'espérance fait partie des valeurs prises par X).

Terminologie : Lorsqu'une variable aléatoire admet une espérance, on parle de variable aléatoire *intégrable*. Cela provient du fait que la théorie sous-jacente est essentiellement un cas particulier de la théorie de l'*intégrale de Lebesgue*.

[2] Techniquement, c'est même une meilleure définition d'une variable aléatoire discrète : une variable aléatoire X pour laquelle il existe un ensemble fini ou dénombrable A tel que $\mathbb{P}(X \in A) = 1$. Il est alors possible d'avoir des variables aléatoires discrètes dans un espace de probabilités qui ne l'est pas

[3] Ceci assure que la valeur de $\mathbb{E}(X)$ ne dépend pas de la *numérotation* des valeurs prises par X, en particulier.

2.3.1 Formule de transfert

Soit f une fonction quelconque, définie sur l'ensemble des valeurs prises par la variable aléatoire X, et à valeurs réelles. Alors $f \circ X$ est une nouvelle variable aléatoire, typiquement notée $f(X)$, et on a, sous réserve de convergence absolue des séries, on a

$$\mathbb{E}(f(X)) = \sum_n f(x_n)\mathbb{P}(X = x_n).$$

Preuve: Posons $Y = f(X)$, et soit $E' = \{y_m, m \in \mathbb{N}\}$ l'ensemble des valeurs prises par Y. La définition de l'espérance donne

$$\mathbb{E}(Y) = \sum_m y_m \mathbb{P}(Y = y_m).$$

Revenons aux sources : l'événement $\{Y = y_m\}$ se réécrit en

$$\begin{aligned} \{Y = y_m\} &= \{\omega \in \Omega : f(X(\omega)) = y_m\} \\ &= \bigcup_{x \in U_m} \{\omega \in \Omega : X(\omega) = x\}, \end{aligned}$$

où $U_m = f^{-1}(y_m) = \{x \in E : f(x) = y_m\}$; la réunion dans la formule précédente est une *réunion disjointe*, ce qui permet de passer aux probabilités :

$$\mathbb{P}(Y = y_m) = \sum_{x \in U_m} \mathbb{P}(X = x).$$

En insérant cette expression dans la formule pour l'espérance, il vient

$$\begin{aligned} \mathbb{E}(Y) &= \sum_m y_m \sum_{x \in U_m} \mathbb{P}(X = x) \\ &= \sum_m \sum_{x \in U_m} f(x)\mathbb{P}(X = x) \\ &= \sum_{x \in E} f(x)\mathbb{P}(X = x). \end{aligned}$$

(la dernière égalité provenant du fait que E n'est autre que la réunion disjointe des U_m) □

Remarque 2.2 *La définition que nous avons donnée de l'espérance, n'est autre que le résultat de l'application de la formule de transfert à $f = X$, si l'on prend comme définition alternative (et équivalente, sur les espaces discrets) de l'espérance,*

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega).$$

2.3.2 Propriétés de l'espérance

Variable aléatoire constante

Pour toute constante réelle λ, la fonction constante sur Ω, partout égale à λ, est bien évidemment une variable aléatoire dont l'espérance est λ.

Positivité

Si X est une variable aléatoire discrète *à valeurs positives ou nulles*, alors $\mathbb{E}(X) \geq 0$, l'égalité n'étant possible que si l'on a $\mathbb{P}(X = 0) = 1$.

Preuve: En reprenant l'une ou l'autre des formules pour l'espérance, on a une série à termes positifs ou nuls, d'où le résultat. De plus, tous les termes sont nuls si et seulement si on a $\mathbb{P}(X = x) = 0$ pour chaque $x > 0$, ce qui implique que l'on ait $\mathbb{P}(X = 0) = 1$. □

Linéarité

Si X et Y sont deux variables aléatoires discrètes, définies sur le même espace, et ayant chacune une espérance, et si λ et μ sont deux réels quelconques, alors $Z = \lambda X + \mu Y$, comme variable aléatoire encore définie sur le même espace, a une espérance, et l'on a

$$\mathbb{E}(Z) = \lambda\mathbb{E}(X) + \mu\mathbb{E}(Y).$$

Preuve: La formule de transfert (qui s'applique, *mutatis mutandis*, aux fonctions de plusieurs variables) donne :

$$\begin{aligned}\mathbb{E}(Z) &= \sum_{m,n}(\lambda x_n + \mu y_m)\mathbb{P}(X = x_n, Y = y_m) \\ &= \lambda\sum_n x_n \sum_m \mathbb{P}(X = x_n, Y = y_m) + \mu\sum_m y_m \sum_n \mathbb{P}(X = x_n, Y = y_m).\end{aligned}$$

Par ailleurs, on a, pour tout m, la réunion disjointe

$$\{X = x_n\} = \bigcup_m \{X = x_n, Y = y_m\},$$

d'où l'égalité sur les probabilités,

$$\mathbb{P}(X = x_n) = \sum_m \mathbb{P}(X = x_n, Y = y_m),$$

et de même,

$$\mathbb{P}(Y = y_m) = \sum_n \mathbb{P}(X = x_n, Y = y_m).$$

En reportant ces égalités dans la formule pour $\mathbb{E}(Z)$, on obtient exactement

$$\mathbb{E}(Z) = \lambda\mathbb{E}(X) + \mu\mathbb{E}(Y).$$

□

En conséquence, les plus férus d'algèbre linéaire pourront dire que, sur un espace de probabilités discret, *l'ensemble des variables aléatoires ayant une espérance est un espace vectoriel, et l'espérance est une forme linéaire*. Il s'agit en fait d'un *espace vectoriel normé* si l'on prend comme norme de X, $\mathbb{E}(|X|)$.

Produit de variables aléatoires indépendantes

Si X et Y sont deux variables aléatoires discrètes, *indépendantes*, alors la variable aléatoire $Z = XY$ est intégrable si X et Y le sont, et $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. Réciproquement, si XY est intégrable, alors soit l'une des deux variables X et Y est nulle avec probabilité 1, soit X et Y sont toutes deux intégrables.

Preuve:

Commençons par prouver le sens direct ; supposons simplement que X et Y sont intégrables.

Une fois de plus, on applique la formule de transfert :

$$\begin{aligned}\mathbb{E}(Z) &= \sum_{m,n} x_n y_m \mathbb{P}(X = x_n, Y = y_m) \\ &= \sum_{m,n} x_n y_m \mathbb{P}(X = x_n)\mathbb{P}(Y = y_m) \\ &= \sum_{n} x_n \mathbb{P}(X = x_n) \sum_{m} y_m \mathbb{P}(Y = y_m) \\ &= \mathbb{E}(X)\mathbb{E}(Y).\end{aligned}$$

Ceci prouve la formule, sous réserve que les séries soient absolument convergentes. Pour le vérifier, il suffit de refaire le même calcul avec les variables $|X$ et $|Y|$; les séries sont alors à termes positifs.

Reste à montrer que, si XY est intégrable, X et Y le sont (en supposant que ni X ni Y n'est nulle avec probabilité 1) ; nous nous contentons, par symétrie, de montrer que X est intégrable.

Puisque Y n'est pas nulle avec probabilité 1, il existe forcément $y_0 \neq 0$ tel que $\mathbb{P}(Y = y_0) > 0$. Formons les deux variables aléatoires

$$\begin{aligned}Z' &= |Z| = |X||Y| \\ Z_0 &\qquad Z'\mathbf{1}_{\{Y=y_0\}}\end{aligned}$$

(en rappelant que $\mathbf{1}_A$ désigne la variable indicatrice de l'événement A, qui vaut 1 sur A et 0 sur son complémentaire)

La variable Z' étant la valeur absolue de Z, l'intégrabilité de l'une entraîne celle de l'autre. De plus, on a $0 \leq Z_0 \leq Z'$ sur tout Ω, et donc l'intégrabilité de Z' entraîne également celle de Z_0.

Or il est facile de voir (en appliquant, une fois de plus, la formule de transfert) que l'on a

$$\mathbb{E}(Z_0) = \mathbb{P}(Y = y_0) \sum_{n} |x_n| \mathbb{P}(X = x_n),$$

et donc cette série converge ; ce qui prouve bien que X est intégrable. □

2.4 Variance et covariance

2.4.1 Variance

On dit qu'une variable aléatoire X est *de carré intégrable* si son carré X^2 est intégrable.

Exercice 2.1 *Montrer que si X est de carré intégrable, alors elle est intégrable. (**Indication :** séparer la somme selon les valeurs de X : majorer séparément la part dûe aux valeurs plus grandes ou plus petite que 1 en valeur absolue)*

Donner un contre-exemple à la réciproque, c'est-à-dire, décrire une variable aléatoire qui soit intégrable sans être de carré intégrable.

Lorsque X est de carré intégrable, on appelle *variance* de X, et on note $\mathbf{Var}(X)$,

$$\mathbf{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 .$$

Proposition 2.3 *Soit X une variable aléatoire de carré intégrable.*

1. $\mathbf{Var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right)$;
2. $\mathbf{Var}(X) \geq 0$;
3. $\mathbf{Var}(X) = 0$ *si et seulement si il existe* $\lambda \in \mathbb{R}$ *tel que* $\mathbb{P}(X = \lambda) = 1$.

Preuve: Nous prouvons la première assertion ; les deux autres en sont des conséquences triviales. Posons $\mu = \mathbb{E}(X)$, et $Z = (X - \mu)^2$; calculons $\mathbb{E}(Z)$ (dont l'intégrabilité est une conséquence des calculs) : par linéarité, il vient

$$\begin{aligned} \mathbb{E}(Z) &= \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 \\ &= \mathbb{E}(X^2) - 2\mu^2 + \mu^2 \\ &= \mathbb{E}(X^2) - \mu^2 = \mathbf{Var}(X). \end{aligned}$$

□

De la linéarité de l'espérance, on déduit aisément $\mathbf{Var}(\alpha X) = \alpha^2 \mathbf{Var}(X)$ pour tout réel α. On définit également l'*écart-type* $\sigma(X) = \sqrt{\mathbf{Var}(X)}$, pour lequel on a naturellement

$$\sigma(\alpha X) = |\alpha|\sigma(X)$$

pour tout réel α.

2.4.2 Covariance

Considérons deux variables aléatoires X et Y, toutes deux de carré intégrable, et vérifions que XY est intégrable.

En effet, $\{|X| \leq |Y|\}$ est bien un événement (le vérifier !), et on peut écrire

$$\begin{aligned} |XY| &= |XY|\mathbf{1}_{\{|X|\leq|Y|\}} + |XY|\mathbf{1}_{\{|X|>|Y|\}} \\ &\leq Y^2\mathbf{1}_{\{|X|\leq|Y|\}} + X^2\mathbf{1}_{\{|X|>|Y|\}} \\ &\leq Y^2 + X^2. \end{aligned}$$

Comme X^2 et Y^2 sont toutes deux intégrables, il en résulte (par positivité de l'espérance) que XY l'est aussi. (Au passage, puisque $(X+Y)^2 = X^2 + Y^2 + 2XY$, le même calcul permet de montrer que, si X et Y sont de carrés intégrables, leur somme l'est également.)

On définit alors la *covariance* de X et Y par

$$\mathbf{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y),$$

et la linéarité de l'espérance permet encore de démontrer (de la même manière que pour la variance) que l'on a

Proposition 2.4

$$\mathbf{Cov}(X,Y) = \mathbb{E}\left((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right).$$

On remarque évidemment que la variance n'est qu'un cas particulier de covariance : $\mathbf{Var}(X) = \mathbf{Cov}(X,X)$.

La définition de la covariance, et la propriété sur l'espérance de variables aléatoires indépendantes, donnent immédiatement

Proposition 2.5 *Si X et Y sont deux variables aléatoires indépendantes et de carré intégrable, alors*

$$\mathbf{Cov}(X,Y) = 0.$$

Attention : la réciproque est complètement fausse !

Exercice 2.2 *Donner un exemple de deux variables aléatoires qui soient de covariance nulle, mais non indépendantes.*

2.4.3 Coefficient de corrélation

Variance et covariance présentent l'inconvénient de dépendre d'un changement d'échelle (changement des unités dans lesquelles sont exprimées les valeurs des variables aléatoires). Le *coefficient de corrélation*, défini par

$$\rho_{X,Y} = \frac{\mathbf{Cov}(X,Y)}{\sigma(X)\sigma(Y)},$$

n'a pas cet inconvénient :

Exercice 2.3 *Soient X et Y deux variables aléatoires de carré intégrable, et chacune de variance non nulle, et α et β deux réels strictement positifs.*

Montrer que l'on a $\rho_{\alpha X,\beta Y} = \rho(X,Y)$.

Qu'en est-il si $\alpha > 0$ et $\beta < 0$?

Exercice 2.4 *Montrer que l'on a, sous les même hypothèse que précédemment,*

$$-1 \leq \rho_{X,Y} \leq 1,$$

et dire dans quel(s) cas on peut avoir égalité.

2.4.4 Variance d'une somme

En développant $\mathbf{Var}(X+Y)$, et en appliquant la linéarité de l'espérance, il vient immédiatement

$$\mathbf{Var}(X+Y) = \mathbf{Var}(X) + \mathbf{Var}(Y) + 2\mathbf{Cov}(X,Y).$$

Exercice 2.5 *Généraliser cette formule au cas d'une somme de n variables aléatoires : $S = X_1 + X_2 + \cdots + X_n$ (toutes les X_i sont de carré intégrable).*

En supposant de plus que les X_i sont toutes indépendantes *et* de même loi, *montrer que, si l'on pose*

$$Z = \frac{1}{n}\left(X_1 + \cdots + X_n\right),$$

on a

$$\mathbf{Var}(Z) = \frac{1}{n}\mathbf{Var}(X_1).$$

2.5 Variables aléatoires à valeurs entières

Un cas particulier important de variables aléatoires discrètes, est celui des variables à valeurs *entières positives*. Il s'agit d'une classe importante, non seulement parce que beaucoup d'exemples classiques de lois de probabilités sont de ce type, mais aussi parce que chaque fois qu'une variable aléatoire est censée mesurer un *comptage* de quelque chose, on se trouve naturellement en face de telles variables aléatoires entières. Il se trouve que l'on dispose d'outils supplémentaires pour travailler sur de telles variables aléatoires.

2.5.1 Une nouvelle formule pour l'espérance

Soit X une variable aléatoire à valeurs dans $\mathbb{N}$ (c'est-à-dire que l'on a $\mathbb{P}(X \in \mathbb{N}) = 1$). Alors, sous réserve de convergence,

$$\mathbb{E}(X) = \sum_{n\geq 0} \mathbb{P}(X > n) = \sum_{n\geq 1} \mathbb{P}(X \geq n).$$

Preuve: Nous montrons la deuxième forme. La probabilité $\mathbb{P}(X \geq n)$ s'exprime sous la forme

$$\mathbb{P}(X \geq n) = \sum_{k=n}^{+\infty} \mathbb{P}(X = k).$$

La formule proposée est donc

$$E = \sum_{n=1}^{+\infty} \left(\sum_{k=n}^{+\infty} \mathbb{P}(X = k) \right).$$

Changeons simplement l'ordre de sommation : d'abord selon k, puis selon n, avec simplement la condition $1 \leq n \leq k$:

$$\begin{aligned} E &= \sum_{k=1}^{+\infty} \left(\sum_{n=1}^{k} \mathbb{P}(X = k) \right) \\ &= \sum_{k=1}^{+\infty} k\mathbb{P}(X = k). \end{aligned}$$

Il suffit de rajouter le terme $0\mathbb{P}(X = 0)$, soit 0, pour obtenir la définition de l'espérance de X ; on a donc bien $E = \mathbb{E}(X)$, comme annoncé. □

2.5.2 Série génératrice de probabilités

Lorsque X est une variable aléatoire à valeurs dans $\mathbb{N}$, la *loi* de X est entièrement décrite par la *suite* $(\mathbb{P}(X = n))_{n\geq 0}$. Un moyen bien commode pour manipuler de telles suites est celui des *séries entières*.

Définition 2.6 *Soit X une variable aléatoire à valeurs entières positives ou nulles. On appelle* série génératrice de probabilités de X, *la série entière*

$$G_X(z) = \sum_{n\geq 0} \mathbb{P}(X = n) z^n.$$

Remarque 2.7 *Il résulte de la formule de transfert que, si z est une valeur pour laquelle la série entière converge absolument (i.e., à l'intérieur du rayon de convergence de la série, ou sur le cercle de convergence), alors $G_X(z)$ n'est autre que $\mathbb{E}(z^X)$.*

Du fait que l'on ait $\mathbb{P}(X \in \mathbb{N}) = 1$, on déduit immédiatement que la série converge pour $z = 1$, avec pour somme $G_X(1) = 1$. La série entière étant à coefficient réels positifs ou nuls, cela entraine que la convergence est absolue, et donc que *le rayon de convergence est au moins égal à* 1. En particulier, il s'agit d'une fonction analytique, donc indéfiniment dérivable, sur $]-1, 1[$.

Il est évident que la série génératrice de probabilités est entièrement déterminée par la loi de X. La réciproque est également vraie : *la loi de X est entièrement caractérisée par sa série génératice de probabilités.* En effet, on a immédiatement, par développement,

$$\mathbb{P}(X = n) = \frac{1}{n!} G_X^{(n)}(0).$$

Calculs d'espérances par séries génératrices

En dérivant formellement la série génératrice, puis en prenant la valeur pour $z = 1$, il vient :

$$\begin{aligned}
G'_X(z) &= \sum_{n\geq 1} n\mathbb{P}(X = n) z^{n-1} \\
G'_X(1) &= \sum_{n\geq 1} n\mathbb{P}(X = n) = \mathbb{E}(X) \\
G''_X(z) &= \sum_{n\geq 2} n(n-1)\mathbb{P}(X = n) z^{n-1} \\
G''_X(1) &= \sum_{n\geq 2} n(n-1)\mathbb{P}(X = n) = \mathbb{E}(X(X-1))
\end{aligned}$$

Par conséquent, on peut calculer l'espérance par

$$\mathbb{E}(X) = G'_X(1),$$

mais aussi la variance par

$$\mathbf{Var}(X) = G''_X(1) + G'_X(1) - \left(G'_X(1)\right)^2.$$

À quelles conditions ces expressions sont-elles valables ? Il faut pour cela qu'il soit correct de dériver les séries entières, puis de les évaluer en 1. C'est le cas si le rayon de convergence est *strictement* supérieur à 1, ce qui est équivalent à dire que l'on ait

$$\mathbb{P}(X = n) = O(\rho^n)$$

pour une constante $\rho < 1$. Mais c'est aussi le cas si le rayon de convergence est *égal* à 1, si les fonctions dérivées $G'_X(z)$ et $G''_X(z)$ ont des limites finies en $z = 1$. En effet, les séries sont des séries à termes positifs, on peut donc intervertir sommation et limites sur l'axe réel positif.

Dans la pratique, si l'on arrive à exprimer, pour $|x| < 1$, la série génératrice de probabilités comme une fonction dont la dérivée (respectivement, la dérivée seconde) admet une limite finie en 1, alors la formule précédente pour l'espérance (respectivement, la variance) est valable.

Séries génératrices multivariées

Si X et Y sont deux variables aléatoires à valeurs entières, on peut définir la série génératrice à deux variables pour le *couple* (X, Y) :

$$G_{X,Y}(s,t) = \sum_{m,n} \mathbb{P}(X = m, Y = n) s^m t^n.$$

C'est une série entière à deux variables, qui est au moins convergente sur $]-1,1[\times]-1,1[$. La définition peut bien entendu s'étendre à plus de deux variables aléatoires (la série aura une variable formelle par variable aléatoire).

Cette série multivariée devient surtout intéressante si les variables aléatoires sont *indépendantes* ; en effet, on a alors

$$\begin{aligned} G_{X,Y}(s,t) &= \sum_{m,n} \mathbb{P}(X = m, Y = n) s^m t^n \\ &= \sum_{m,n} \mathbb{P}(X = m)\mathbb{P}(Y = n) s^m t^n \\ &= \left(\sum_m \mathbb{P}(X = m) s^m \right) \left(\sum_n \mathbb{P}(Y = n) t^n \right) \\ &= G_X(s) G_Y(t). \end{aligned}$$

Séries génératrice d'une somme de variables indépendantes

En substituant la même variable formelle dans les deux de la série génératrice à deux variables, on obtient

$$G_{X,Y}(z,z) = \sum_{m,n} \mathbb{P}(X = m, Y = n) z^{m+n} = G_{X+Y}(z).$$

Cette relation est surtout utile dans le cas où X et Y sont indépendantes : on a alors la

Proposition 2.8 *Soient X et Y deux variables aléatoires indépendantes, à valeurs entières. Alors, pour $Z = X + Y$, on a*

$$G_Z(z) = G_X(z) G_Y(z).$$

Ce résultat s'étent naturellement au cas de plus de 2 variables aléatoires indépendantes.

2.6 Exemples de lois discrètes

Nous passons en revue quelques exemples classiques de lois de probabilités discrètes, qui apparaissent fréquemment dans la modélisation de phénomènes variés.

2.6.1 Loi de Bernoulli

Une variable aléatoire B qui ne peut prendre que les deux valeurs 0 et 1 (c'est-à-dire telle que $\mathbb{P}(B \in \{0,1\}) = 1$, est appelée *variable de Bernoulli*. Si l'on note $p = \mathbb{P}(B = 1)$, on a donc naturellement $\mathbb{P}(B = 0) = 1 - p$. Le paramètre p peut prendre n'importe quelle valeur $0 \leq p \leq 1$; les cas $p = 0$ et $p = 1$ sont bien entendu dégénérés.

La loi de Bernoulli apparaît chaque fois que l'on modélise le résultat du lancer d'une pièce de monnaie (jeu de pile ou face) qui peut être biaisée ; une pièce équilibrée correspond à $p = 1/2$.

L'espérance, la variance, et la série génératrice de probabilités, pour une Bernoulli B_p de paramètre p, sont données par

$$\begin{aligned} \mathbb{E}(B_p) &= p \\ \mathbf{Var}(B_p) &= p(1-p) \\ G_{B_p}(z) &= (1-p) + pz. \end{aligned}$$

2.6.2 Loi binomiale

La loi binomiale admet deux paramètres : un entier $N \geq 0$, et un réel $0 \leq p \leq 1$. C'est la loi de la somme de N variables de Bernoulli indépendantes, toutes de même paramètre p. Autrement dit, on obtient une variable binomiale en effectuant N lancers indépendants d'une pièce de monnaie (biaisée ou non, peu importe ; toutefois, la probabilité d'obtenir Pile doit être la même pour tous les lancers), le résultat étant le nombre de fois que l'on obtient Pile.

Par linéarité de l'espérance, et de la variance dans le cas de sommes de variables indépendantes, on obtient immédiatement l'espérance et la variance d'une binomiale $B_{N,p}$ de paramètres (N, p) :

$$\begin{aligned} \mathbb{E}(B_{N,p}) &= Np \\ \mathbf{Var}(B_{N,p}) &= Np(1-p). \end{aligned}$$

La série génératrice de probabilités s'obtient en utilisant le résultat sur les séries génératrices de sommes de variables indépendantes :

$$G_{B_{N,p}}(z) = \big(G_{B_p}(z)\big)^N = (1 - p + pz)^N .$$

La formule du binôme, appliquée à cette dernière série génératrice, donne les probabilités de la loi binomiale[4] :

$$G_{B_{N,p}}(z) = \sum_{k=0}^{N} \binom{N}{k} (1-p)^{N-k} p^k z^k,$$

d'où l'on déduit

$$\mathbb{P}(B_{N,p} = k) = \binom{N}{k} p^k (1-p)^k.$$

[4] La notation $\binom{N}{k}$ représente le coefficient binomial $N!/(k!(N-k)!)$, parfois aussi noté C_N^k.

2.6.3 Loi géométrique

La loi géométrique est paramétrée par un réel $0 < p < 1$ (les cas dégénérés $p = 0$ et $p = 1$ n'ayant que peu d'intérêt). On obtient une variable aléatoire géométrique dans la situation suivante : on effectue une même expérience, qui a une probabilité p de "réussir", autant de fois, de manière indépendante, qu'il est nécessaire pour obtenir le premier succès. Ainsi, si l'on a une suite infinie de variables de Bernoulli $(B_n)_{n\geq 1}$, indépendantes, toutes de même paramètre p, le *plus petite indice* d'une des B_k qui valle 1, est une variable aléatoire géométrique (de paramètre p) :

$$X_p = \inf\{k : B_k = 1\}.$$

(La notation G_p serait plus naturelle pour indiquer une variable géométrique, mais nous utilisons ici la lettre X pour éviter la confusion avec la notation des séries génératrices de probabilités.)

Cette relation permet de déterminer les probablités de la loi géométrique : la probabilité que X_p soit supérieure (strictement) à k, est exactement la probabilité que les k premières Bernoulli soient toutes nulles, soit $(1-p)^k$. On a donc

$$\begin{aligned}
\mathbb{P}(X_p > k) &= (1-p)^k \\
\mathbb{P}(X_p = k) &= \mathbb{P}(X_p > k-1) - \mathbb{P}(X_p > k) \\
&= (1-p)^{k-1} - (1-p)^k \\
&= (1-p)^{k-1}p.
\end{aligned}$$

Cette expression permet d'écrire la série génératrice de probabilités :

$$\begin{aligned}
G_{X_p}(z) &= \sum_{k\geq 1} p(1-p)^{k-1}z^k \\
&= pz\sum_{k\geq 0}(1-p)^k z^k \\
&= \frac{pz}{1-(1-p)z}.
\end{aligned}$$

On obtient alors, simplement par dérivation, espérance et variance :

$$\begin{aligned}
\mathbb{E}(X_p) &= \frac{1}{p} \\
\mathbf{Var}(X_p) &= \frac{1-p}{p^2}.
\end{aligned}$$

2.6.4 Loi de Poisson

La loi de Poisson de paramètre λ ($\lambda > 0$ étant un réel positif quelconque) est définie par ses probabilités : une variable aléatoire Z est de Poisson si l'on a, pour chaque entier $k \geq 0$,

$$\mathbb{P}(Z = k) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

Cette définition peut paraître quelque peu arbitraire, mais la loi de Poisson apparaît en fait assez naturellement dans différentes situations. La première est comme "cas limite" de binomiales : si l'on considère des binomiales avec un paramètre N qui devient grand, mais d'espérance constante (*i.e.* $p = a/N$ pour une constante a), la loi binomiale ressemble alors a une Poisson, plus facile à manipuler :

Proposition 2.9 *Pour $a > 0$ fixé, on a, pour tout k,*

$$\lim_{N\to+\infty} \mathbb{P}(B_{N,a/N} = k) = e^{-a}\frac{a^k}{k!}.$$

Preuve:

$$\begin{aligned}
\mathbb{P}(B_{N,a/N} = k) &= \frac{N!}{k!(N-k)!}\left(\frac{a}{N}\right)^k\left(1-\frac{a}{N}\right)^{N-k} \\
&= \frac{a^k}{k!}\left(1-\frac{a}{N}\right)^{N-k}\prod_{i=1}^{k-1}\frac{N-i}{N} \\
&= \frac{a^k}{k!}\left(1-\frac{a}{N}\right)^{N}\left(1-\frac{a}{N}\right)^{-k}\prod_{i=1}^{k-1}\left(1-\frac{i}{N}\right).
\end{aligned}$$

Dans cette dernière expression, le premier facteur est constant, le deuxième tend vers e^{-a}, et le reste est un produit d'un nombre constant (car k est fixe) de facteurs dont chacun tend vers 1. La limite est donc bien $e^{-a}a^k/k! = \mathbb{P}(Z = k)$. □

Un exemple d'utilisation de ce genre de propriétés est le suivant : on a un ensemble de 80 personnes dont on suppose que les dates d'anniversaires sont indépendantes, chacun des 365 jours de l'année ayant la même probabilité 1/365 d'être le jour anniversaire de chaque personne. Alors la loi du nombre de personnes nées le premier janvier est, en réalité, la loi binomiale de paramètres $N = 80$ et $p = 1/365$, mais l'approximation poissonnienne suggère qu'elle est "presque" une loi de Poisson de paramètre $80/365 = 0.22 \pm 0.001$.

La série génératrice de probabilités de la loi de Poisson se calcule fort bien :

$$\begin{aligned}
S_P(z) &= \sum_{k\geq 0} e^{-\lambda}\frac{\lambda^k z^k}{k!} \\
&= e^{-\lambda}\sum_{k\geq 0}\frac{(\lambda z)}{k!} \\
&= e^{-\lambda}e^{\lambda z} = e^{\lambda(z-1)}.
\end{aligned}$$

Par dérivation, on obtient alors facilement $\mathbb{E}(P) = \lambda$ et $\mathbf{Var}(P) = \lambda$.

Par ailleurs, la série génératrice de probabilités permet de démontrer facilement le théorème suivant :

Théorème 2.10 *Soient X et Y deux variables aléatoires indépendantes, X suivant la loi de Poisson de paramètre λ et Y la loi de Poisson de paramètre μ. Alors $X + Y$ suit la loi de Poisson de paramètre $\lambda + \mu$.*

Preuve: On utilise simplement la propriété sur la série génératrice d'une somme de variables indépendantes :

$$\begin{aligned}
G_{X+Y}(z) &= G_X(z)G_Y(z) \\
&= e^{\lambda(z-1)}e^{\mu(z-1)} = e^{(\lambda+\mu)(z-1)}.
\end{aligned}$$

On reconnaît dans G_{X+Y} la série génératrice de probabilités d'une loi de Poisson de paramètre $\lambda + \mu$. □

Cette propriété sur les sommes de variables de Poisson a un pendant que l'on appelle parfois le *fractionnement* des variables de Poisson :

Théorème 2.11 *Soit $\lambda > 0$ et $0 < p < 1$. On considère un ensemble de Z "individus", où Z suit une loi de Poisson de paramètre λ. Chaque individu est alors coloré en bleu avec probabilité p, et en rouge avec probabilité $1 - p$, indépendamment les uns des autres, et indépendamment de la valeur de Z.*

Soit Z_b le nombre d'individus bleus, et Z_r le nombre d'individus rouges.

Alors, Z_b suit la loi de Poisson de paramètre $p\lambda$, Z_r suit la loi de Poisson de paramètre $(1-p)\lambda$, et Z_b et *Z_r* sont indépendantes.

Chapitre 3

Lois de probabilités à densité

Sommaire

Lorsque l'on a à modéliser des grandeurs aléatoires qui prennent des valeurs réelles générales, on est amené à considérer des lois de probabilités qui ne sont plus discrètes. Le plus souvent, on a affaire à des variables aléatoires faisant partie de la classe des *lois à densité*.

3.1 Variables aléatoires diffuses

3.1.1 Notion de densité

Définition 3.1 *Une variable aléatoire réelle X a une densité s'il existe une fonction f, définie et continue par morceaux*[1] *sur $\mathbb{R}$, telle que l'on ait, pour tout intervalle $I = [a, b]$,*

$$\mathbb{P}(X \in I) = \int_a^b f(x)dx.$$

[1] La condition de continuité n'est pas *sricto sensu* nécessaire, mais elle permet de traiter tous les cas dont nous avons besoin dans le cadre de ce cours.

La fonction f est alors appelée la densité *de X. Une variable aléatoire (ou une loi de probabilités sur $\mathbb{R}$) qui admet une densité est dite* diffuse.

Remarque 3.2
- *La fonction f ne peut prendre que des valeurs positives ou nulles, sinon il serait facile de trouver un intervalle, éventuellement très petit, sur lequel f est strictement négative, ce qui se traduirait par une probabilité strictement négative que X prenne sa valeur dans cet intervalle.*
- *La fonction f est* unique, *aux valeurs aux points de discontinuité près. On peut donc légitimement parler de* la *densité d'une variable aléatoire.*
- *La variable aléatoire prend forcément sa valeur dans $\mathbb{R}$, ce qui impose à la densité une condition supplémentaire :*
$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$
- *La formule donnant la probabilité que X prenne sa valeur dans un ensemble s'étend au cas où l'ensemble n'est pas nécessairement un intervalle, mais un borélien quelconque (donc en particulier une réunion finie ou dénombrables d'intervalles ouverts ou fermés) :*
$$\mathbb{P}(X \in A) = \int_A f(x)dx.$$
- *Aussi dérangeant que cela puisse paraître au premier abord,* la probabilité qu'une variable à densité prenne une valeur donnée, est toujours nulle *: $\mathbb{P}(X = x) = 0$ pour tout $x \in \mathbb{R}$. En effet, la formule donne pour l'intervalle réduit à un point $\{x\} = [x, x]$:*
$$\mathbb{P}(X = x) = \int_x^x f(t)dt = 0.$$

Il convient de noter que l'on peut définir des lois de probabilités "hybrides", sortes de mélanges de lois discretes et à densité. Le plus simple consiste à choisir une suite $(x_n)_{n\in\mathbb{N}}$ de valeurs, une suite de probabilités $(p(x_n))$ avec $\sum_n p(x_n) < 1$, et une fonction positive f telle que
$$\sum_n p(x_n) + \int_{\mathbb{R}} f(t)dt = 1.$$
On définit alors la loi d'une variable X par
$$\mathbb{P}(X \in A) = \sum_{x_n \in A} p(x_n) + \int_A f(x)dx.$$

3.1.2 Fonction de répartition

Lorsqu'une variable aléatoire admet une densité f, la fonction de répartition
$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t)dt$$
est continue (ce n'est pas le cas lorsque la loi contient une part discrète, comme dans la remarque précédente), et est en fait une *primitive* de la densité f. La fonction de répartition constitue souvent le meilleur moyen de calculer la loi d'une variable aléatoire.

3.1.3 Couples de variables diffuses

On dira qu'un couple de variables aléatoires (X, Y) a pour densité une fonction de deux variables f, si l'on a, pour tous intervalles $I = [a, b]$ et $J = [c, d]$,

$$\mathbb{P}(X \in I, Y \in J) = \int_{I \times J} f(x, y)dx\ dy = \int_a^b \left(\int_c^d f(x, y)dy \right) dx.$$

Cette formule par une intégrale double s'étend à n'importe quel domaine mesurable D de $\mathbb{R}^2$:

$$\mathbb{P}((X, Y) \in D) = \int_D f(x, y)dx\ dy$$

En particulier, la densité permet de retrouver les lois marginales de X et de Y :

$$\mathbb{P}(X \in A) = \mathbb{P}(X \in A, Y \in \mathbb{R}) = \int_A \left(\int_{\mathbb{R}} f(x, y)dy \right) dx,$$

et cela suggère que la densité de X devrait être donnée par

$$x \mapsto \int_{\mathbb{R}} f(x, y)dy.$$

(Il se peut toutefois que cette intégrale ne converge pas pour tout x, et que la loi de X n'admette pas de densité.)

Ces considérations s'étendent facilement au cas de vecteurs de plus de deux variables aléatoires, mais donnent lieu à des intégrales multiples.

Attention : il se peut que deux variables aléatoires X et Y aient chacune une densité, mais que leur loi conjointe n'en ait pas. L'exemple le plus simple est donné par un couple formé d'une première variable diffuse, et d'une seconde qui dépend de manière déterministe de la première : ainsi, si on a $Y = |X|$, le couple (X, Y) n'aura pas de densité, même si X en a une.

3.1.4 Couples de variables indépendantes

Lorsque les variables X et Y sont indépendantes et toutes deux diffuses, le couple (X, Y) admet automatiquement une densité qui n'est autre que le produit des densités de X et de Y : si X a pour densité la fonction f, et Y la fonction g, alors la densité de (X, Y) est la fonction h définie par

$$h(x, y) = f(x)g(y).$$

3.1.5 Densité d'une somme

On suppose que le couple (X, Y) admet une densité h. Dans ces conditions, la densité f de la variable aléatoire $X + Y$ est (sous réserve de convergence des intégrales),

$$f(x) = \int_{-\infty}^{+\infty} h(t, x - t)dt.$$

Dans le cas particulier où X et Y sont indépendantes, et de densités respectives f et g, la densité h de $X + Y$ devient

$$h(x) = \int_{-\infty}^{+\infty} f(t)g(x - t)dt$$

3.1.6 Densité image

Il arrive souvent que, disposant d'une variable aléatoire diffuse X, on ait besoin de déterminer la densité d'une variable aléatoire $Y = \varphi(X)$, pour une certaine fonction φ.

Proposition 3.3 (Densité d'une variable aléatoire image) *Soit X une variable aléatoire, à valeurs dans un espace $\Omega = \mathbb{R}^d$, de densité f.*

- *Soit φ une fonction définie sur Ω, injective, à valeurs dans $\mathbb{R}^d$, de fonction réciproque Ψ (i.e., $\Psi \circ \varphi(x) = x$ pour tout $x \in \Omega$) ; alors, si Ψ est suffisamment régulière, $Y = \varphi(X)$ a également une densité g, et on a*

$$g(y) = f(\Psi(y))|\det(J_\Psi(y))|,$$

  où $J_\Psi(y)$ désigne la matrice jacobienne de Ψ au point y.
- *Si φ est une fonction définie sur Ω, à valeurs dans $\mathbb{R}^d$,* injective par morceaux *(c'est-à-dire que l'on peut partitionner Ω en domaines $D_1, \ldots, D_k$, de telle sorte que la restriction φ_k de φ à D_k soit injective, de réciproque Ψ_k), alors $Y = \varphi(X)$ a également une densité g, et on a*

$$g(y) = \sum_k f(\Psi_k(y))|\det(J_{\Psi_k}(y)|,$$

  (la sommation s'étend, pour chaque y, à l'ensemble des k tels que y soit dans le domaine de définition de Ψ_k ; autrement dit, à l'ensemble des k tels que y soit dans l'image de φ_k)

Preuve: Soit $B \subset \mathbb{R}^d$ un domaine "raisonnable" (par exemple, ouvert borné) ; on a l'union disjointe

$$\begin{aligned}\{Y \in B\} &= \bigcup_k \{X \in D_k, \varphi_k(X) \in B\} \\ &= \bigcup_k \{X \in D_k \cap \Psi_k(B)\},\end{aligned}$$

d'où l'expression pour la probabilité :

$$\begin{aligned}\mathbb{P}(Y \in B) &= \sum_k \mathbb{P}(X \in D_k \cap \Psi_k(B)) \\ &= \sum_k \int_{D_k \cap \Psi_k(B)} f(x)dx.\end{aligned}$$

En effectuant, dans chaque intégrale, le changement de variable $x = \Psi_k(y)$, on obtient

$$\mathbb{P}(Y \in B) = \sum_k \int_{\varphi_k(D_k) \cap B} f(\Psi_k(y))|J_{\Psi_k(y)}|dy.$$

La condition $y \in \varphi_k(D_k)$ revient à dire que Ψ_k soit définie en y, ce qui donne la formule de l'énoncé, en tenant compte de la restriction sur la sommation. □

La formulation de la Proposition 3.3 peut paraître intimidante, mais elle est dans la pratique simple d'utilisation. Ainsi, dans le cas $\varphi(x) = x^2$, on a naturellement deux branches

inverses $\Psi_1(y) = \sqrt{y}$ (définie sur $[0, +\infty[$, à valeurs dans $[0, +\infty[$) et $\Psi_2(y) = -\sqrt{y}$ (définie sur $]0, +\infty[$, à valeurs dans $]-\infty, 0[$), ce qui donne pour la densité de $Y = X^2$,

$$g(y) = \frac{f(\sqrt{y}) + f(-\sqrt{y})}{2\sqrt{y}} \mathbf{1}_{y>0}.$$

La formule de la densité image permet également de prouver la formule donnant la densité d'une somme de deux variables aléatoires : en utilisant la transformation $\varphi : (x_1, x_2) \mapsto (x_1, x_1+x_2)$, de réciproque $\Psi : (y_1, y_2) \mapsto (y_1, y_2-y_1)$ (le déterminant de la matrice jacobienne étant 1 partout), on obtient, pour le couple $(Y_1, Y_2) = \varphi(X_1, X_2)$, la densité $h(y_1, y_2-y_1)$; de là, on applique la formule de la densité d'une marginale pour obtenir la densité de $Y_2 = X_1 + X_2$:

$$g(y) = \int_{\mathbb{R}} h(t, y-t)dt.$$

Densité et transformations affines

Il existe un cas particulier, mais extrêmement utile, d'application de la formule de la densité image : c'est lorsque la transformation φ est affine.

Proposition 3.4 (Cas des variables aléatoires réelles) *Soit X une variable aléatoire réelle, de densité f, et soient $\alpha \neq 0$ et β deux réels.*

Alors la variable aléatoire $Y = \alpha X + \beta$, a pour densité la fonction g définie par

$$g(y) = \frac{1}{|\alpha|} f\left(\frac{y-\beta}{\alpha}\right).$$

Proposition 3.5 (Cas multidimensionnel) *Soit $\mathbf{X} = (X_1, \ldots, X_d)$ une variable aléatoire à valeurs dans $\mathbb{R}^d$, de densité f (f est donc une fonction de d variables réelles), et soient A une matrice carrée de taille d, inversible, et $\mathbf{b} = (b_1, \ldots, b_d)$ un vecteur (les vecteurs sont considérés comme des vecteurs colonne). Posons $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$; alors la variable aléatoire $\mathbf{Y}$ a pour densité la fonction g, définie par*

$$g(\mathbf{y}) = |\det(A)|^{-1} f(A^{-1}(\mathbf{y} - \mathbf{b})).$$

3.2 Moments d'une variable aléatoire diffuse

3.2.1 Espérance

L'*espérance* d'une variable aléatoire X, de densité f, est définie par

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x)dx,$$

sous réserve que cette intégrale soit absolument convergente (c'est-à-dire, que l'intégrale sur $\mathbb{R}$ de la valeur absolue de $xf(x)$ soit finie).

Propriétés de l'espérance

Les propriétés énoncées pour l'espérance des variables aléatoires discrètes (positivité, linéarité, espérance d'un produit de variables aléatoires indépendantes) restent vraies dans le cas de variables diffuses.

Formule de transfert

La formule de transfert s'adapte au cas diffus : si X a pour densité f, et si φ est une fonction continue, alors la variable $Y = \varphi(X)$ a pour espérance

$$\mathbb{E}(Y) = \int_{-\infty}^{+\infty} \varphi(x) f(x) dx.$$

3.2.2 Variance et covariance

Les définitions de la variance, de la covariance, et du coefficient de corrélation qui ont été données au chapitre 2, sont également valables pour les variables diffuses – tout est défini à partir de l'espérance. Les mêmes propriétés sont également vraies ; en particulier, deux variables aléatoires indépendantes ont une covariance et un coefficient de corrélation nuls, mais la réciproque est fausse en général.

3.3 Exemples de lois diffuses

Dans cette section, après avoir introduit la notion, fort utile pour l'étude des variables aléatoires diffuses, de fonction caractéristique, nous passons en revue quelques exemples classiques de lois diffuses qui interviennent souvent.

3.3.1 Fonction caractéristique

La *fonction caractéristique* d'une variable aléatoire diffuse, est en quelque sorte le pendant continu de la série génératrice de probabilités pour une variable à valeurs entières.

Définition 3.6 *Soit X une variable aléatoire. On note Φ_X, et on appelle* fonction caractéristique de X, *la fonction définie par*

$$\Phi_X(t) = \mathbb{E}\left(e^{itX}\right)$$

pour les valeurs de t pour lesquelles l'espérance est définie.

La formule de transfert donne, si la densité de X est f,

$$\begin{aligned} \Phi_X(t) &= \int_{-\infty}^{+\infty} e^{itx} f(x) dx \\ &= \int_{-\infty}^{+\infty} \cos(tx) f(x) dx + i \int_{-\infty}^{+\infty} \sin(tx) f(x) dx. \end{aligned}$$

En d'autres termes, la fonction caractéristique est, à un changement de variable près, la transformée de Fourier de la densité.

Remarque 3.7 *Si les variables aléatoires considérées sont à valeurs* réelles *(plutôt que complexes), ce qui est toujours le cas dans le cadre de ce cours, la variable aléatoire e^{itX} est toujours de module* 1, *donc intégrable et de carré intégrable, et ce, pour tout t. Donc, en particulier, la fonction caractéristique d'une variable aléatoire réelle est définie (comme variable aléatoire complexe) sur tout $\mathbb{R}$.*

De même que pour la série génératrice de probabilités, la fonction caractéristique détermine complètement la loi de la variable aléatoire. Elle permet, de manière analytique, de retrouver les moments (espérance, variance, ...) sous la forme de dérivées :

$$\begin{aligned} \mathbb{E}(X) &= -i\Phi'_X(0) \\ \mathbf{Var}(X) &= -\Phi''_X(0) + \left(\Phi'_X(0)\right)^2 \end{aligned}$$

De plus, dans le cas de variables aléatoires indépendantes, on peut exprimer facilement la fonction caractéristique de leur somme :

Proposition 3.8 *Soient X et Y deux variables aléatoires indépendantes, et soit $Z = X + Y$. Alors pour tout t,*

$$\Phi_Z(t) = \Phi_X(t)\Phi_Y(t).$$

Preuve: Tout simplement, si X et Y sont indépendantes, alors e^{itX} et e^{itY} le sont aussi. On a donc

$$\mathbb{E}(e^{itZ}) = \mathbb{E}(e^{itX}e^{itY}) = \mathbb{E}(e^{itX})\mathbb{E}(e^{itY}).$$

□

Cette propriété se généralise évidemment à un nombre quelconque de variables aléatoires indépendantes.

3.3.2 Loi uniforme sur un intervalle borné

Soit $I = [a, b]$ un intervalle borné. La loi uniforme sur I est définie par sa fonction de répartition :

$$F(t) = \begin{cases} 0 & (t \leq a) \\ \frac{t-a}{b-a} & (a \leq t \leq b) \\ 1 & (t > b) \end{cases}$$

En conséquence, si $J = [c, d]$ est un sous-intervalle ($a \leq c$, $d \leq b$), la probabilité qu'un variable aléatoire uniforme sur I prenne sa valeur dans J est tout simplement $F(d) - F(c) = (d - c)/(b - a)$: cette probabilité ne dépend que de la *longueur* du sous-intervalle, et non de sa position à l'intérieur de I. C'est ce qui correspond le mieux à la notion intuitive de "réel aléatoire entre a et b".

La densité de la loi uniforme sur I est définie par

$$f(t) = \begin{cases} 0 & (t \notin J) \\ \frac{1}{b-a} & (t \in J) \end{cases}$$

L'espérance est $(a + b)/2$, et la variance, $(b - a)^2/12$.

Les lois uniformes sont conservées par transformations affines :

Proposition 3.9 *Si U est une variable aléatoire uniforme sur $[a, b]$, et $\alpha \neq 0$ et β deux réels, alors $V = \alpha U + \beta$ est uniforme sur $[\alpha a + \beta, \alpha b + \beta]$ (si $\alpha > 0$) ou $[\alpha b + \beta, \alpha a + \beta]$ (si $\alpha < 0$).*

En particulier, si U est uniforme sur $[0, 1]$, $1 - U$ est également uniforme sur $[0, 1]$.

3.3.3 Loi exponentielle

Définition et paramètres

La loi exponentielle est paramétrée par un réel $\lambda > 0$, et a pour densité

$$f_\lambda(t) = \lambda e^{-\lambda t}.$$

En conséquence, sa fonction de répartition est

$$F(t) = \mathbb{P}(X < t) = \int_0^t \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}.$$

La fonction caractéristique se calcule explicitement :

$$\begin{aligned} \Phi_X(t) = \mathbb{E}(e^{itX}) & \\ &= \lambda \int_0^{+\infty} e^{itx} e^{-\lambda x} dx \\ &= \lambda \int_0^{+\infty} e^{(it-\lambda)x} dx \\ &= \frac{\lambda}{\lambda - it} = \frac{1}{1 - it/\lambda}. \end{aligned}$$

On en déduit immédiatement, par dérivations,

$$\begin{aligned} \mathbb{E}(X) &= \frac{1}{\lambda} \\ \mathbf{Var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

Propriétés de conditionnement

La loi exponentielle a surtout des propriétés quand on calcule des probabilités conditionnelles : Pour tous $s, t > 0$,

$$\mathbb{P}(X \in [t, t+s] | X > t) = \mathbb{P}(X \in [0, s])$$

Et, plus généralement, pour tous $r, s, t > 0$ avec $r < s$,

$$\mathbb{P}(X \in [t+r, t+s] | X > t) = \mathbb{P}(X \in [r, s]).$$

Ces deux propriétés s'interprètent en visualisant X comme l'instant où quelque chose se produit. Conditionner par l'événement $\{X > t\}$ correspond exactement à se placer à l'instant t, où l'on observe que le "quelque chose" ne s'est pas encore produit. Les propriétés ci-dessus (qui sont caractéristiques des lois exponentielles : une variable aléatoire qui vérifie la seconde est forcément une variable exponentielle) disent simplement que le temps depuis lequel on attend n'a pas d'influence sur la loi du temps que l'on aura encore a attendre.

La loi exponentielle est une loi qui intervient lorsque l'on modélise le fait, pour un appareil, de tomber en panne "de manière aléatoire". On fait l'hypothèse que l'appareil ne vieillit pas : la probabilité qu'il tombe en panne pour la première fois dans la prochaine seconde (ou la prochaine minute, ou le prochain jour) ne dépend pas de l'âge de l'appareil (c'est-à-dire du temps pendant lequel l'appareil n'est jamais tombé en panne). Si l'on note X la variable aléatoire qui décrit le moment où l'appareil (mis en service au temps 0) tombe en panne pour la première fois, X suit une loi exponentielle.

Propriétés de changement d'échelle

On vérifiera sans peine, par exemple en calculant des fonctions de répartition, les propriétés suivantes :

- Si X est une variable exponentielle de paramètre λ, alors μX est une variable exponentielle de paramètre λ/μ.
- Si X et Y sont des variables exponentielles indépendantes, de paramètres respectifs λ et μ, alors $\min(X, Y)$ est une variable exponentielle de paramètre $\lambda + \mu$.

3.3.4 Lois gaussiennes

Les lois gaussiennes, ou lois normales, dont fait partie la loi dite "normale réduite", sont particulièrement importantes dans les applications statistiques.

Une variable aléatoire suit la loi normale réduite $\mathcal{N}(0, 1)$, si elle a pour densité

$$f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Si X est normale réduite, $Y = m + \sigma X$ est appelée *variable gaussienne (ou normale) de paramètres m et σ* pour tous $m \in \mathbb{R}$ et $\sigma > 0$. La densité d'une telle variable aléatoire est, d'après la Proposition 3.4,

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

La loi gaussienne de paramètres m et σ est notée $\mathcal{N}(m, \sigma)$.

Il n'y a pas de formule analytique pour la fonction de répartition. En revanche, la fonction caractéristique se calcule fort bien : pour une normale réduite X, on a

$$\begin{aligned}
\Phi_X(t) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{itx} e^{-\frac{x^2}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2 - 2itx}{2}} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x-it)^2 + t^2}{2}} dx \\
&= e^{-\frac{t^2}{2}}.
\end{aligned}$$

Pour une gaussienne Y de loi $\mathcal{N}(m, \sigma)$, $Y = m + \sigma X$ où X est normale, et on a donc, par simple changement de variable,

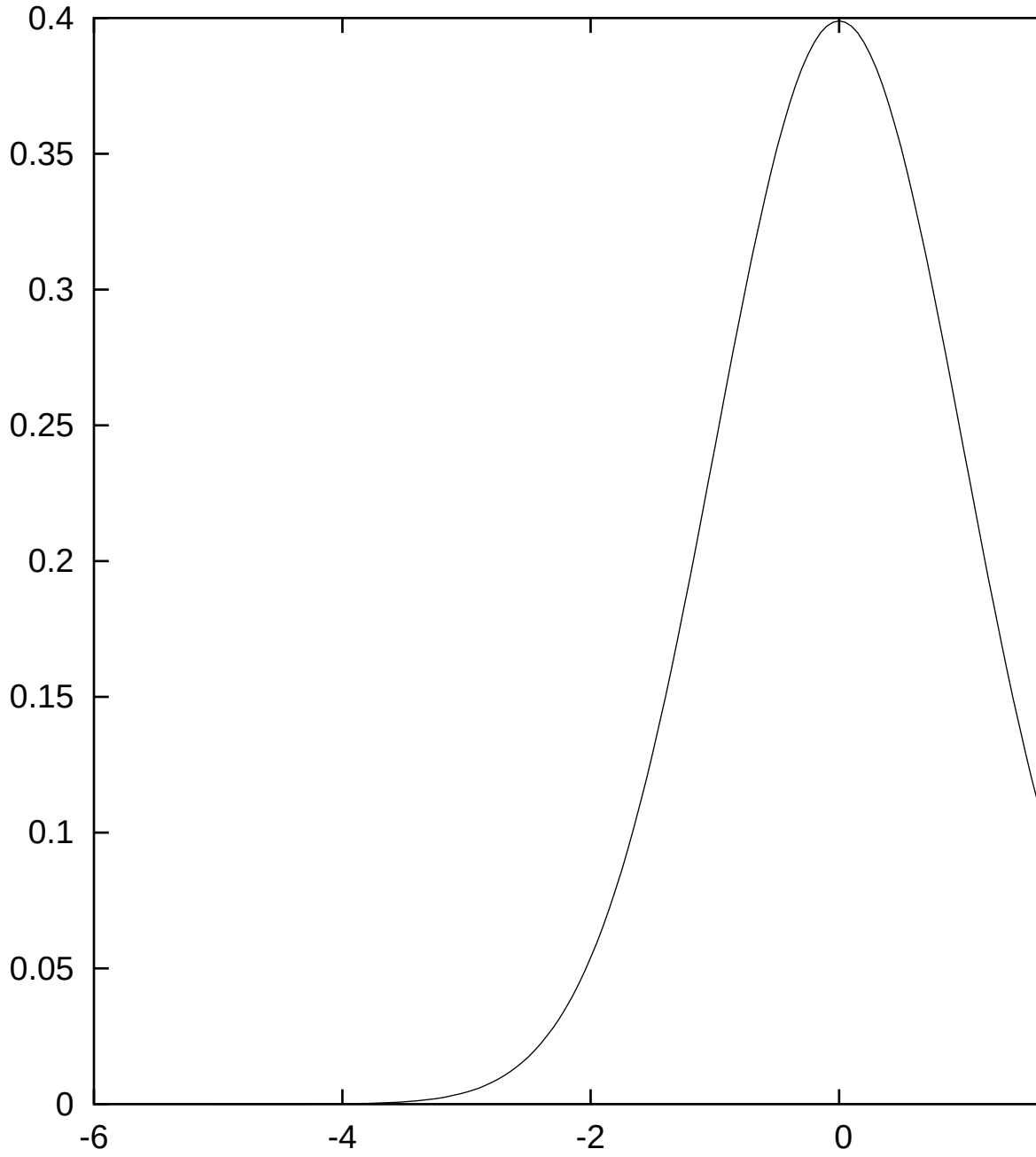
$$\begin{aligned}
\Phi_Y(t) &= \mathbb{E}(e^{it(m+\sigma X)}) && (3.1) \\
&= e^{itm}\mathbb{E}(e^{it\sigma X}) && (3.2) \\
&= e^{itm}\Phi_X(\sigma t) = e^{-\frac{\sigma^2 t^2}{2} + itm} && (3.3)
\end{aligned}$$

Espérance et variance s'obtiennent alors simplement en dérivant la fonction caractéristique :

$$\begin{aligned}
\mathbb{E}(Y) &= m \\
\mathbf{Var}(Y) &= \sigma^2
\end{aligned}$$

(les paramètres sont m et σ sont donc respectivement l'espérance et l'écart-type : on parle parfois de *gaussienne d'espérance m et d'écart-type σ*.

FIG. 3.1 – Densité de la loi normale réduite



Remarque 3.10 *Lorsqu'une variable aléatoire a une fonction caractéristique de la forme $e^{-Q(t)}$, où $Q(t)$ est une fonction polynôme du second degré (dont le coefficient dominant ne peut qu'être positif, et le coefficient constant, nul), on peut immédiatement en conclure qu'il s'agit d'une gaussienne ; les paramètres exacts peuvent être alors identifiés en la mettant sous la forme (3.3).*

Théorème 3.11 *Soient X et Y deux variables aléatoires indépendantes, gaussiennes, de paramètres respectifs (m, σ) et (m', σ'). Alors $X + Y$ est gaussienne, d'espérance $m + m'$ et de variance $\sigma^2 + \sigma'^2$.*

Preuve: Le plus simple est de calculer la fonction caractéristique de $X + Y$, qui, par indépendance, est le produit des fonctions caractéristiques. On peut soit faire tout le calcul explicitement pour vérifier que l'on obtient bien la fonction caractéristique de la loi annoncée, soit se contenter de vérifier que l'on obtient la forme de la fonction caractéristique d'une gaussienne, et calculer séparément les paramètres par linéarité de l'espérance et de la variance dans le cas indépendant. □

Ce théorème se généralise de lui-même à plus de deux variables aléatoires : une somme d'un nombre fini quelconque de variables gaussiennes indépendantes, est encore gaussienne.

Vecteurs gaussiens

Définition 3.12 *Soit $\mathbf{X} = (X_1, \ldots, X_N)$ un vecteur (considéré comme vecteur colonne), composé de N variables aléatoires indépendantes, chacune de loi normale réduite $\mathcal{N}(0, 1)$.*

Soient également A une matrice carrée de taille $M \times N$, et $\mathbf{m} = (m_1, \ldots, m_M)$ un vecteur (colonne) à coefficients réels.

Alors, le vecteur $\mathbf{Y} = (Y_1, \ldots, Y_M)$, défini par

$$\mathbf{Y} = A\mathbf{X} + \mathbf{m},$$

est appelé vecteur gaussien.

Remarque 3.13
- *On a vu qu'une combinaison affine de deux variables aléatoires gaussiennes indépendantes, est encore une variable aléatoire gaussienne (à condition d'accepter de considérer les constantes comme des variables gaussiennes de variance nulle). Il en découle qu'une combinaison affine d'un nombre fini quelconque de variables gaussiennes, est encore gaussienne. En particulier, les coordonnées d'un vecteur gaussien sont forcément gaussiennes.*
- *De plus, une combinaison linéaire quelconque des coordonnées d'un vecteur gaussien, étant encore une combinaison affine des gaussiennes indépendantes de départ, est également gaussienne.*
- *On peut démontrer que cette dernière propriété* caractérise *en fait les vecteurs gaussiens : un vecteur aléatoire est gaussien si et seulement si n'importe quelle combinaison affine de ses coordonnées est une variable gaussienne.*

La définition que nous avons donnée d'un vecteur gaussien, permet de calculer facilement espérance, variance, et covariance des coordonnées d'un vecteur gaussien : l'espérance est

naturellement $\mathbb{E}(Y_i) = m_i$; pour la covariance, si les coefficients de A sont notés $a_{i,j}$, on a

$$\begin{aligned}\mathbb{E}(Y_iY_k) &= \mathbb{E}\left((m_i + \sum_{j=1}^{N} a_{i,j}X_j)(m_k + \sum_{\ell=1}^{N} a_{k,\ell}X_\ell)\right) \\ &= m_im_k + \sum_{j,\ell} a_{i,j}a_{k,\ell}\mathbb{E}(X_jX_\ell).\end{aligned}$$

Par le fait que les X_j sont des normales réduites indépendantes, on a $\mathbb{E}(X_jX_\ell) = 0$ si $j \neq \ell$, et $\mathbb{E}(X_j^2) = 1$; on obtient donc

$$\mathbb{E}(Y_iY_k) = m_im_k + \sum_{j=1}^{N} a_{i,k}a_{j,k},$$

d'où la covariance

$$\mathbf{Cov}(Y_i, Y_k) = \sum_{j=1}^{N} a_{i,k}a_{j,k} = ({}^tA.A)_{i,j}.$$

On a donc montré que *les coefficients de la matrice ${}^tA.A$ donnent les covariances des coordonnées du vecteur* $\mathbf{Y}$.

De plus, dans le cas où $M = N$ et où la matrice A est *inversible*, on peut exprimer $\mathbf{X}$ en fonction de $\mathbf{Y}$ sous la forme :

$$\mathbf{X} = A^{-1}(\mathbf{Y} - \mathbf{m});$$

cela permet, par l'intermédiaire de la formule de la densité image, d'obtenir la densité du vecteur $\mathbf{Y}$: la fonction Ψ est affine, de matrice A^{-1} ; son jacobien est donc $\det(A^{-1}) = (\det(A))^{-1}$. En appliquant la formule, on obtient

$$f(\mathbf{y}) = \frac{1}{|\det(A)|(2\pi)^{N/2}} e^{-\frac{1}{2}||A^{-1}(\mathbf{y}-\mathbf{m})||^2}.$$

Cette formule est plus souvent exprimée en fonction, non de la matrice A, mais de la matrice des covariances $C = {}^tA.A$: on a $\det(C) = \det(A)^2$, d'où la formule pour la densité

$$f(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^N \det(C)}} e^{-\frac{1}{2}({}^t\mathbf{y}-{}^t\mathbf{m})C^{-1}(\mathbf{y}-\mathbf{m})}. \tag{3.4}$$

Inversement, si la densité d'un vecteur $(Y_1, \ldots, Y_N)$ de N variables aléatoires est de la forme (3.4) avec une matrice symétrique C, alors C est la matrice des covariances, et si l'on factorise C en ${}^tA.A$ où A est une matrice carrée (ce qui est toujours possible), alors on peut vérifier (en appliquant la formule de la densité image dans l'autre sens) que $\mathbf{X} = A^{-1}(\mathbf{Y} - \mathbf{m})$ est un vecteur composé de N variables aléatoires indépendantes, normales réduites. En particulier, on en déduit que $\mathbf{Y}$ est bien un vecteur gaussien.

Une autre propriété importante des coordonnées d'un vecteur gaussien est la suivante : *deux coordonnées Y_1 et Y_2 d'un même vecteur gaussien sont indépendantes, si et seulement si leur covariance est nulle.*

3.3.5 Lois du χ^2 et de Student

Les lois du χ^2 et de Student sont les lois de variables aléatoires qui sont facilement définies à partir d'un vecteur gaussien réduit (c'est-à-dire, dont les composantes sont des variables aléatoires gaussiennes réduites, indépendantes).

Lois du χ^2 (khi-deux)

Définition 3.14 *On appelle* loi du χ^2 à d degrés de liberté*, la loi de la variable aléatoire*

$$U_d = \sum_{k=1}^{d} X_k^2,$$

où les variables $X_1, \ldots, X_d$ sont d variables aléatoires indépendantes, de loi normale réduite.

En d'autres termes, la loi du χ^2 à d degrés de liberté est la loi du carré de la norme (euclidienne) d'un vecteur gaussien de dimension d, dont la matrice des covariances est la matrice identité, et le vecteur des espérances, le vecteur nul.

Lorsque d est assez grand ($d > 30$, mettons), on peut approximativement (c'est une conséquence du Théorème Central Limite, que l'on verra au chapitre 4, et du calcul de $\mathbb{E}(X^4)$ lorsque X est normale réduite) remplacer la loi du χ^2 à d degrés de liberté, par la loi $\mathcal{N}(d, \sqrt{2d})$.

Lois de Student

Définition 3.15 *On appelle* loi de Student à d degrés de liberté*, la loi de la variable aléatoire*

$$S_d = \frac{X_0}{\sqrt{\frac{1}{d}\sum_{k=1}^{d} X_k^2}},$$

où les variables $X_0, \ldots, X_d$ sont $d+1$ variablea aléatoires indépendantes, de loi normale réduite.

On reconnaît, sous la racine carrée du dénominateur, une variable du χ^2 à d degrés de liberté (divisée par d).

Densités

Les lois du χ^2 et de Student sont toutes les deux des lois diffuses. Pour obtenir leurs densités, une méthode consiste à partir de la densité d'un vecteur gaussien réduit (de dimension $N = d$ pour la loi du χ^2, $N = d + 1$ pour Student) :

$$f_N(\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi}}\right)^N e^{-(x_1^2+\cdots+x_N^2)/2},$$

et d'effectuer le changement de variables en coordonnées sphériques[2]

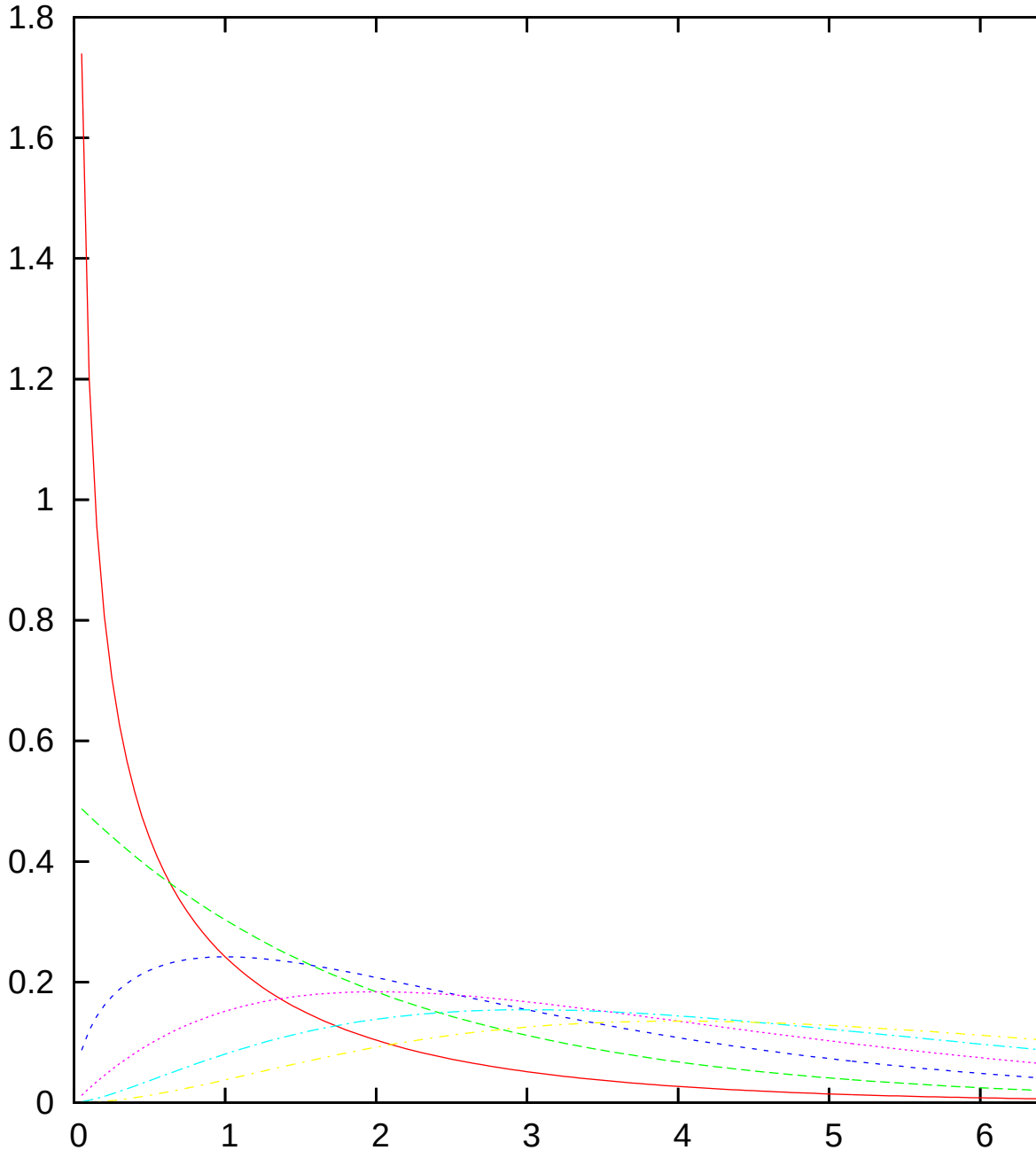
$$\begin{aligned}
x_1 &= r\cos(\theta_1)\ldots\cos(\theta_{N-2})\cos(\theta_{N-1}) \\
x_2 &= r\cos(\theta_1)\ldots\cos(\theta_{N-2})\sin(\theta_{N-1}) \\
x_i &= r\cos(\theta_1)\ldots\cos(\theta_{N-i})\sin(\theta_{N-i+1} \\
x_N &= r\sin(\theta_1)
\end{aligned}$$

avec $r \in]0, +\infty[$, $\theta_1 \in [0, 2\pi[$, $\theta_i \in]-\pi/2, \pi/2[$. Le jacobien vaut en valeur absolue (après calculs),

$$|J| = r^{N-1}\left|\cos^{N-2}(\theta_1)\right|\cos^{N-3}(\theta_2)\ldots\cos(\theta_{N-2}).$$

[2] En dimension N, tout de même.

FIG. 3.2 – Densités de lois du χ^2



On remarque que ce jacobien s'exprime comme un produit de fonctions des différentes variables ; en conséquence, les calculs d'intégrales multiples se transforment en produits d'intégrales simples, pour peu que les domaines d'intégration s'expriment comme des produits cartésiens en ces variables.

Pour la loi du χ^2, on obtient la fonction de répartition

$$F_N(x) = \mathbb{P}(U_N \leq x),$$

puis la densité par différentiation, en intégrant la densité sur le domaine $x_1^2 + \cdots + x_N^2 \leq x$, soit exactement $r \leq \sqrt{x}$. Le résultat, tous calculs faits, est le suivant :

Proposition 3.16 *La loi du χ^2 à d degrés de liberté a pour densité*

$$\frac{1}{\Gamma(d/2)2^{d/2}} x^{\frac{d}{2}-1} e^{-x/2} \mathbf{1}_{x>0}.$$

Ici, Γ désigne la "fonction Gamma", définie par

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt,$$

dont les propriétés classiques

$$\begin{aligned} \Gamma(n+1) &= n! \\ x\Gamma(x) &= \Gamma(x+1) \\ \Gamma(1/2) &= \sqrt{\pi} \end{aligned}$$

permettent de calculer les valeurs pour x entier ou demi-entier.

Pour la loi de Student, on forme une variable de Student

$$S = X_N / \sqrt{1/(N-1)(X_1^2 + \cdots + X_{N-1}^2)};$$

la condition $S \leq x$ correspond exactement, en coordonnées sphériques, à $\tan(\theta_{n-1}) \leq x/(N-1)$. En intégrant pour obtenir la fonction de répartition, puis en dérivant pour la densité, on obtient :

Proposition 3.17 *La loi de Student à d degrés de liberté a pour densité, sur $\mathbb{R}$:*

$$\frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{d\pi}\,\Gamma\left(\frac{d}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{d}\right)^{\frac{d+1}{2}}}.$$

Chapitre 4

Théorèmes asymptotiques

Sommaire

Il existe de nombreux résultats sur ce qui se passe lorsque l'on observe un grand nombre de variables aléatoires indépendantes. Les plus classiques portent les noms de *loi des grands nombres* et de *théorème central limite*.

4.1 Quelques inégalités utiles

Nous commençons par donner deux inégalités qui permettent toutes les deux de *majorer* la probabilité qu'une variable aléatoire soit "loin" de son espérance.

4.1.1 Inégalité de Markov

Proposition 4.1 (Inégalité de Markov) *Soit X une variable aléatoire à valeurs* positives ou nulles*, et telle que $\mathbb{E}(X)$ existe.*

Alors on a, pour tout $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Preuve: Posons $X_a = a.\mathbf{1}_{\{X \geq a\}}$ (de telle sorte que X_a vaut 0 si $X < a$, et a si $X \geq a$). X_a est donc une variable aléatoire à valeurs positives, et l'on a $\mathbb{E}(X_a) = a\mathbb{P}(X \geq a)$. Mais on a également $X_a \leq X$ (toujours), et par conséquent, $\mathbb{E}(X_a) \leq \mathbb{E}(X)$, soit $a\mathbb{P}(X \geq a) \leq \mathbb{E}(X)$, ce qui est l'inégalité recherchée. □

L'inégalité de Markov peut également être écrite sous la forme suivante, qui est équivalente (en posant $a = \lambda \mathbb{E}(X)$) :

$$\mathbb{P}(X \geq \lambda \mathbb{E}(X)) \leq \frac{1}{\lambda}.$$

Remarquons que, bien évidemment, cette inégalité ne fournit aucune information si $a \leq \mathbb{E}(X)$ (la majoration dit que la probabilité est inférieure à un nombre qui est lui-même plus grand que 1). En revanche, elle permet d'affirmer qu'une variable aléatoire (positive) ne peut pas être plus grande que deux fois son espérance avec probabilité plus grande que 1/2.

De plus, l'inégalité de Markov est "ce qui se fait de mieux" si l'on ne connaît que l'espérance, comme le montre l'exercice suivant.

Exercice 4.1 *Soit $a > 0$. Donner un exemple de variable aléatoire pour laquelle l'inégalité de Markov est une égalité.*

4.1.2 Inégalité de Tchebycheff

Lorsque l'on connaît à la fois l'espérance et la variance d'une variable aléatoire, l'*inégalité de Tchebycheff*[1] est souvent meilleure (plus forte) que celle de Markov.

Proposition 4.2 (Inégalité de Tchebycheff) *Soit X une variable aléatoire ayant une espérance μ et une variance σ^2.*

Alors, pour tout $\lambda > 0$, on a

$$\mathbb{P}(|X - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}.$$

Preuve: Ce n'est rien d'autre que l'inégalité de Markov, appliquée à la variable aléatoire $Y = (X - \mu)^2$ dont l'espérance est σ^2 ; en effet,

$$\{|X - \mu| \geq \lambda\sigma\} = \{(X - \mu)^2 \geq \lambda^2\sigma^2\} = \{Y \geq \lambda^2 \mathbb{E}(Y)\}.$$

□

Une autre façon, équivalente, d'écrire l'inégalité de Tchebycheff est la suivante :

$$\mathbb{P}(|X - \mu| \geq \lambda) \leq \frac{\sigma^2}{\lambda^2}.$$

4.2 Différentes notions de convergence

Il existe de nombreuses façons classiques de définir une notion de limite pour une suite de variables aléatoires ou de lois de probabilités. La plupart sortent du cadre de ce cours, mais il est nécessaire d'en présenter quelques unes.

[1] La graphie internationale est *Chebyshev*.

4.2.1 Convergence en loi

La *convergence en loi* ne concerne pas les variables aléatoires en tant que telles, mais plutôt leurs *lois de probabilités* ; c'est une notion de limite dans l'espace des lois de probabilités sur $\mathbb{R}$. On dira tout de même, par abus de langage, qu'une suite de variables aléatoires converge en loi vers une variable X si leurs lois de probabilités convergent vers celle de X ; en particulier, rien n'oblige à ce que les variables aléatoires soient définies sur le même espace de probabilités.

Définition 4.3 *Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires réelles, de fonctions de répartition respectives $(F_n)_{n\geq 0}$, et soit X une variable aléatoire de fonction de répartition F.*

On dira que la suite $(X_n)_{n\geq 0}$ converge en loi *(ou* en distribution *vers X, ce que l'on note*

$$X_n \xrightarrow{\mathcal{L}} X,$$

si l'on a

$$\lim_{n\to+\infty} F_n(x) = F(x)$$

pour tout x où F est continue. *En particulier, si X est une variable diffuse (donc à fonction de répartition partout continue), la convergence en loi vers X revient exactement à la convergence simple des fonctions de répartition vers celle de X.*

La convergence en loi de (X_n) vers la loi de X, se note parfois également

$$X_n \xrightarrow{(d)} X.$$

Une façon classique de démontrer la convergence en loi est le théorème suivant, dû à Paul Lévy, et que nous admettrons :

Théorème 4.4 ("Théorème de Paul Lévy") *1. Si une suite (X_n) de variables aléatoires converge en loi vers une variable X, alors la suite (φ_n) de leurs fonctions caractéristiques converge simplement vers la fonction caractéristique φ de X. De plus, cette convergence est uniforme sur tout intervalle borné.*

2. *Si une suite (X_n) de variables aléatoires est telle que la suite (φ_n) des fonctions caractéristiques correspondantes converge simplement vers une fonction φ, et que* la partie réelle de φ est continue en 0, *alors φ est la fonction caractéristique d'une (unique) loi de probabilité, et la suite (X_n) converge en loi vers cette loi de probabilités.*

Cas de la convergence en loi discrète

Il est possible d'avoir une suite de variables discrètes qui convergent en loi vers une loi diffuse (c'est d'ailleurs le cas lorsqu'on applique le Théorème Central Limite à des variables discrètes), ou une suite de variables diffuses qui convergent en loi vers une loi discrète. Toutefois, il existe un cas particulier de convergence en loi de variables discrètes vers une loi discrète qui peut se caractériser de manière beaucoup plus simple que le cas général.

Proposition 4.5 *Soit E un ensemble fini de réels.*

Alors, une suite $(X_n)_{n\geq 0}$ de variables aléatoires à valeurs dans E, converge en loi vers la loi d'une variable X, si et seulement si on a, pour tout $x \in E$,

$$\lim_{n\to+\infty} \mathbb{P}(X_n = x) = \mathbb{P}(X = x).$$

En d'autres termes, dans l'espace des lois de variables aléatoires à valeurs dans un espace fini, la convergence en loi est la même chose que la convergence classique dans un espace vectoriel de dimension finie.

4.2.2 Convergence en probabilités

Définition 4.6 *Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires, et X une variable aléatoire, toutes définies sur le même espace de probabilités.*

On dit que la suite (X_n) converge en probabilités vers 0 *si, pour tout $\epsilon > 0$,*

$$\lim_{n\to+\infty} \mathbb{P}(|X_n| > \epsilon) = 0.$$

On dit que la suite (X_n) converge en probabilités vers X *si la suite $(X_n - X)$ converge en probabilités vers* 0.

La convergence en probabilités de la suite $(X_n)_{n\geq 0}$ vers X, se note traditionnellement

$$X_n \overset{(p)}{\to} X.$$

4.2.3 Convergence presque sûre

Définition 4.7 *Soit $(X_n)_{n\geq 0}$ une suite de variables aléatoires, et X une variable aléatoire, toutes définies sur le même espace de probabilités Ω.*

*On définit l'*ensemble de convergence de la suite (X_n) vers X *comme étant*

$$\mathcal{C} = \left\{\omega \in \Omega : \lim_n X_n(\omega) = X(\omega)\right\},$$

et on dit que (X_n) converge presque sûrement vers X *si*

$$\mathbb{P}(\mathcal{C}) = 1.$$

Exercice 4.2 *Montrer que l'ensemble $\mathcal{C}$ est bien un événement.*

Contrairement à la convergence en loi, la convergence presque sûre concerne bien les variables aléatoires – pour envisager qu'une suite de variables aléatoires converge presque sûrement, il est *indispensable* qu'elles soient définies sur le même espace de probabilités. De plus, la limite presque sûre, lorsqu'elle existe, est "presque" unique :

Proposition 4.8 *Si une même suite (X_n) converge presque sûrement à la fois vers X et vers Y, alors $\mathbb{P}(X = Y) = 1$: à part éventuellement sur un ensemble de probabilité nulle, X et Y sont identiques.*

La convergence presque sûre de $(X_n)_{n\geq 0}$ vers X, se note

$$X_n \to X \text{ (p.s.)}$$

Des trois notions de convergence présentées ici, la convergence presque sûre est la plus forte, et la convergence en loi est la plus faible : **la convergence presque sûre implique la convergence en probabilités, qui elle-même implique la convergence en loi**.

4.3 Lois des grands nombres

Les multiples théorèmes connus collectivement sous le nom de *lois des grands nombres* peuvent être considérés comme une justification *a posteriori* de la notion de probabilité.

4.3.1 Une loi des grands nombres

Théorème 4.9 (Loi des grands nombres) *Soit $(X_n)_{n\geq 1}$, une suite infinie de variables aléatoires indépendantes, toutes de même loi que la variable $X = X_1$. On suppose de plus que X admet une espérance et une variance σ^2.*

On définit la variable $\overline{X}_N$, comme la moyenne *des N premières X_i :*

$$\overline{X}_N = \frac{1}{N}\left(X_1 + \cdots + X_N\right).$$

Alors la suite $(\overline{X}_n)_{n\geq 1}$ converge en probabilité vers la constante $a = \mathbb{E}(X)$.

Preuve: Les variables X_i étant indépendantes, la somme de N d'entre elles a pour variance $N\sigma^2$. En divisant par N, on obtient

$$\mathbf{Var}(\overline{X}_N) = \frac{\sigma^2}{N}.$$

Par ailleurs, la linéarité de l'espérance donne immédiatement $\mathbb{E}(\overline{X}_N) = \mathbb{E}(X)$.

Soit $\epsilon > 0$; appliquons l'inégalité de Tchebychev à $\overline{X}_N$:

$$\mathbb{P}(|\overline{X}_N - \mathbb{E}(X)| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2 N}.$$

À ϵ fixé, cette probabilité tend donc bien vers 0 lorsque N tend vers $+\infty$. □

Remarque 4.10 *La preuve précédente n'utilise pas réellement, ni le fait que les variables sont indépendantes, ni celui qu'elles ont la même loi ; il suffirait qu'elles soient deux à deux indépendantes, ou même simplement non corrélées (toutes covariances nulles), et qu'elles aient toutes même espérance et même variance, pour que le calcul de la variance de $\overline{X}_N$ soit correct.*

Remarque 4.11 *La loi des grands nombres présentée ici porte le nom de loi* faible *des grands nombres ; la loi* forte *demande la convergence presque sûre au lieu de la convergence en probabilités. Les hypothèse du théorème 4.9 sont en fait suffisantes pour obtenir la convergence presque sûre, et donc une loi forte, mais la démonstration en est plus compliquée.*

4.3.2 Loi des grands nombres et notion intuitive de probabilité

Historiquement, la notion naïve de probabilité a évolué à partir de la constatation empirique que, si la même pièce de monnaie (équilibrée) est lancée un grand nombre de fois, la *proportion* de lancers qui donnent un résultat Pile se stabilise toujours vers 1/2 ; si la pièce de monnaie n'est pas parfaitement équilibrée, la proportion de lancers Pile ne se stabilise pas forcément vers 1/2, mais elle se stabilise encore. Des constatations similaires ont été faites avec, par exemple, des dés : si un dé non pipé est lancé un grand nombre de fois, la proportion de lancers qui donnent 1 se stabilise à 1/6, et la moyenne des lancers se stabilise à 3.5.

La première constatation a donné naissance à la notion intuitive de probabilité (c'est, dans un grand nombre d'essais[2], *toujours* la proportion d'essais qui vérifient la condition), et la dernière, à celle d'espérance (c'est, dans un grand nombre d'expériences[3], la valeur moyenne que l'on observe *toujours*.

Ces constatations ne font pas partie des *axiomes* de la théorie mathématique des probabilités, mais elles en sont une des *conséquences*, par l'intermédiaire de la loi des grands nombres.

Considérons, en effet, une expérience qui a une probabilité p ($0 < p < 1$) de "réussir". Cela se modélise naturellement par un espace de probabilités dans lequel on dispose d'un événement A, de probabilité p ; la variable aléatoire caractéristique de A, $\mathbf{1}_A$, est donc une variable de Bernoulli de paramètre p.

Pour modéliser une suite potentiellement infinie de répétitions de l'expérience, nous considérons donc une suite infinie de variables aléatoires (X_n), indépendantes, chacune étant une variable de Bernoulli de paramètre p (la valeur de la n-ème variable indiquant si la n-ème expérience réussit ou échoue). La *proportion* de succès parmi les N premières expériences est donc $(X_1 + \cdots + X_N)/N = \overline{X}_N$.

Dans cette situation, la loi faible des grands nombres prédit que la probabilité que cette proportion ne soit pas comprise entre $p - \epsilon$ et $p + \epsilon$, tend vers 0 lorsque N tend vers l'infini ; la loi forte, elle, prédit que, avec probabilité 1, la proportion de succès tend vers p lorsque N tend vers l'infini.

4.4 Théorème central limite

Les différentes versions de la loi des grands nombres nous renseignent sur ce qui se passe lorsque l'on somme un grand nombre de variables aléatoires indépendantes, centrées, de même loi, et que l'on moyenne en divisant par leur nombre : le résultat tend en gros vers 0, c'est-à-dire vers quelque chose qui n'est plus aléatoire, mais déterministe. Cela ne veut pourtant pas dire que la somme tend vers 0 ; on a eu besoin de diviser par N.

Si maintenant, au lieu de diviser la somme par le nombre de variables aléatoires (N), on divise par sa racine carrée ($\sqrt{N}$), ce qui est évidemment plus faible, il se trouve que le résultat du passage à la limite *reste* aléatoire, mais que *la loi de ce résultat aléatoire ne dépend quasiment pas de la loi des variables aléatoires de départ* : c'est toujours une loi gaussienne. C'est essentiellement ce qu'énonce le "théorème central limite"[4], et la raison pour laquelle la loi normale est aussi importante.

Théorème 4.12 (*Central Limit Theorem*) *Soit $(X_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes, toutes de même loi, d'espérance nulle*[5] *et de variance finie σ^2.*

Posons $S_N = \sum_{i=1}^{N} X_i$, et $Y_N = S_N/(\sigma\sqrt{N})$. Alors la suite $(Y_n)_{n\geq 1}$ converge en loi vers la loi normale $N(0,1)$.

[2] indépendants

[3] indépendantes toujours

[4] Cette vilaine expression à la grammaire douteuse est un anglicisme provenant de *central limit theorem*, elle-même étant une traduction de l'Allemand *zentraler Grentzwertsatz* proposé par le...Hongrois(!) Pólya ; une expression telle que "théorème de la limite centrale" serait peut-être plus propre, mais l'usage a consacré le théorème sous ce nom.

[5] Si ce n'est pas le cas, on remplace X_n par $X'_n = X_n - \mathbb{E}(X_1)$.

Preuve: Nous donnons seulement la preuve du théorème dans le cas où les X_n ont une densité f et sont de plus de cube intégrable ($\mathbb{E}(X_1^3)$ existe), bien que le théorème soit valable sans ces conditions ; on trouvera une preuve plus complète dans [2]. La preuve utilise le théorème 4.4. Nous allons donc nous attacher à démontrer la convergence des fonctions caractéristiques des Y_N vers celle de la loi normale.

Soit φ la fonction caractéristique des variables aléatoires X_n :

$$\varphi(t) = \int e^{itx} f(x)dx.$$

La formule de Taylor avec reste intégral donne, à l'ordre 3, pour la fonction exponentielle,

$$e^u = 1 + u + \frac{u^2}{2} + \frac{u^3}{2}\int_0^1 (1-v)^2 e^{uv} dv.$$

Appliquée en $u = itx$, elle devient

$$e^{itx} = 1 + itx - \frac{t^2x^2}{2} - i\frac{t^3x^3}{2}\int_0^1 (1-v)^2 e^{itxv} dv.$$

L'intérêt de cette formulation est que la partie intégrale est immédiatement majorée (en module) par 1. En reportant cette décomposition dans l'expression de $\varphi(t)$, il vient

$$\varphi(t) = \int f(x)dx + it\int xf(x)dx - \frac{t^2}{2}\int x^2 f(x)dx - i\frac{t^3}{2}\int x^3 f(x)H(t,x)dx.$$

La première intégrale vaut 1 (car f est une densité), la deuxième est nulle (c'est la définition de l'espérance de X_1), la troisième vaut $\mathbb{E}(X^2) = \sigma^2$, et la troisième est majorée en module par $\mathbb{E}(|X^3|)$. Nous avons donc

$$\varphi(t) = 1 - \frac{t^2\sigma^2}{2} + o(t^2).$$

La fonction caractéristique de $Y_N = (X_1 + \cdots + X_N)/(\sigma\sqrt{N})$ vaut

$$\begin{aligned} \varphi_N(t) &= \left(\varphi\left(\frac{t}{\sigma\sqrt{N}}\right)\right)^N \\ &= \left(1 - \frac{t^2}{2N} + o\left(\frac{t^2}{N}\right)\right)^N \end{aligned}$$

En appliquant la relation bien connue

$$\lim_{N\to+\infty}\left(1 + \frac{u}{N}\right)^N = e^u,$$

nous obtenons donc (t est un réel quelconque mais fixé, c'est N qui tend vers $+\infty$) :

$$\lim_{N\to+\infty} \varphi_N(t) = e^{-t^2/2}.$$

Nous avons donc montré que la suite des fonctions caractéristiques des Y_N converge simplement vers la fonction caractéristique de la loi normale ; le théorème de Paul Levy permet d'en conclure que les Y_N convergent en loi vers la loi normale. □

Ce théorème a une importance capitale en statistiques : chaque fois que l'on dispose d'un grand nombre N d'échantillons indépendants d'une même loi d'espérance μ et de variance σ^2, on peut faire l'approximation que la somme, centrée et renormalisée

$$Y = \frac{1}{\sqrt{N}} \left(\left(\sum_{i=1}^{N} X_i \right) - N\mu \right)$$

suit la loi gaussienne $\mathcal{N}(0, \sigma)$.

Chapitre 5

Simulation numérique de lois de probabilités

Sommaire

5.1 Principes généraux

5.1.1 Définition d'une simulation numérique

Simuler numériquement une loi de probabilité donnée, consiste à donner, par des moyens numériques, les valeurs $x_1, \ldots, x_N$ prises par N variables aléatoires *indépendantes* $X_1, \ldots, X_N$ dont chacune suit la loi de probabilité spécifiée.

Nous sommes donc confrontés à un semi paradoxe : munis d'un ordinateur, objet dont on a pris grand soin d'assurer que le comportement est aussi déterministe que possible, nous souhaitons lui faire produire un résultat aléatoire.

Une solution possible à ce problème en apparence insurmontable consiste à utiliser un dispositif extérieur à l'ordinateur, tel qu'un capteur de température ou un dispositif de mesure du temps, et à supposer que, à un grand niveau de précision, la valeur observée peut être considérée comme aléatoire (si l'on mesure, par exemple, le temps entre deux frappes sur la touche d'un clavier, et qu'on ne garde que les millièmes de seconde, on peut raisonnablement considérer que chaque résultat est aléatoire, et que les résultats successifs sont indépendants). De tels dispositifs ont toutefois, le plus souvent, un problème de débit (dans le cas de la mesure du temps entre deux frappes sur un clavier, on est dépendant de l'activité de l'utilisateur, et le débit ne peut qu'être très limité ; dans les cas où l'on mesure une grandeur physique à

intervalles réguliers, augmenter le débit revient à rapprocher les mesures, ce qui introduit des corrélations entre les valeurs obtenues, que l'on ne peut plus considérer comme indépendantes).

Une solution alternative, bien qu'imparfaite, consiste à créer des *générateurs pseudo-aléatoires*, qui sont des dispositifs (typiquement purement informatiques) qui fournissent des valeurs qui ne sont pas réellement aléatoires, mais qui y ressemblent suffisamment (typiquement, on s'assure empiriquement de la "qualité" d'un tel générateur en vérifiant qu'il passe avec succès des tests statistiques variés).

5.1.2 Simulation de la loi uniforme sur $[0,1]$

La plupart des méthodes classiques de simulation présupposent que l'on dispose préalablement d'un moyen de simuler la loi uniforme sur l'intervalle $[0,1]$. La plupart des langages de programmation proposent d'ailleurs une fonction permettant de simuler la loi uniforme.

La grande majorité des méthodes numériques reposent sur le calcul d'une suite par récurrence. On peut par exemple définir une suite $(x_n)_{n\geq 0}$ par la donnée d'une valeur initiale x_0, et par une relation de récurrence de la forme

$$x_{n+1} = ax_n + b \mod M,$$

où a, b et M sont des entiers choisis avec soin (voir plus bas). On considère alors que $r_n = x_n/M$ représente une bonne approximation d'une variable aléatoire uniforme sur $[0,1]$, et que termes successifs de la suite peuvent être vus comme des échantillons indépendants (c'est cette dernière supposition qui est la plus sujette à caution).

Les contraintes sur les paramètres sont généraleent les suivantes : M doit être le plus grand possible, typiquement $M = 2^q$ sur une machine où les entiers sont représentés sur q bits ; a doit être premier avec M (c'est-à-dire impair si $M = 2^q$), et il est préférable que b soit impair. Ceci assure que deux termes successifs de la suite ne peuvent pas prendre la même valeur, ce qui serait désastreux puisque la suite serait alors stationnaire, tous les termes suivants étant égaux.

De tels générateurs pseudo-aléatoires passent avec succès certains tests statistiques (test du χ^2 par exemple ; voir le chapitre 9), et peuvent être considérés comme satisfaisants. Ils sont toutefois toujours limités. On trouvera dans [4] et dans [5] des commentaires intéressants sur les générateurs pseudo-aléatoires.

5.1.3 Méthode de la transformation inverse

La méthode dite de la transformation inverse, permet en théorie de simuler n'importe quelle loi de probabilités, en supposant que l'on sait simuler la loi uniforme sur $[0,1]$.

Considérons la fonction de répartition F de la loi à simuler. F est, par définition, croissante sur $\mathbb{R}$, et continue à droite en tout point. On en déduit que, pour tout $u \in]0,1[$, l'ensemble

$$F^{-1}([u,1]) = \{x : F(x) \geq u\}$$

est un intervalle non vide, non borné à droite, fermé à gauche, dont nous notons $G(u)$ la borne inférieure :

$$\{x : F(x) \geq u\} = [G(u), +\infty[$$

La fonction G ainsi définie sur $]0,1[$, est une sorte de fonction réciproque généralisée de F : si F est une bijection de $\mathbb{R}$ sur $]0,1[$, G n'est rien d'autre que sa fonction réciproque, et

l'on a $F(G(x)) = x$ pour tout $x \in]0,1[$. De manière générale, on a ici, pour tous a, b tels que $a < b$ et tout $u \in]0,1[$,

$$F(a) < u \leq F(b) \iff a < G(u) \leq b$$

Cette fonction réciproque, quand on est en mesure de la calculer, permet de simuler la loi de probabilité en question :

Théorème 5.1 *Soit F la fonction de répartition d'une loi de probabilité μ, et soit G sa fonction réciproque généralisée.*

Alors, si U est une variable aléatoire uniforme sur $[0,1]$, $V = G(U)$ est une variable aléatoire de loi μ (c'est-à-dire de fonction de répartition F).

Preuve: Vérifions que la fonction de répartition est bien F :

$$\begin{aligned} \mathbb{P}(V \leq x) &= \mathbb{P}(G(U) \leq x) \\ &= \mathbb{P}(U \leq F(x)) \\ &= F(x). \end{aligned}$$

La deuxième égalité provient de la propriété précédente sur la fonction réciproque généralisée ; la dernière, du fait que la variable U, étant uniforme sur $[0,1]$, a pour fonction de répartition la fonction identité sur $[0,1]$. □

5.2 Exemples de simulations

Dans chacun des exemples qui suivent, on indique comment obtenir un échantillon de la loi recherchée en utilisant un ou plusieurs échantillons indépendants de la loi uniforme. Si la formule est de la forme $X = h(U)$, cela signifie donc que, à partir d'une simulation $u_1, \ldots, u_N$ de la loi uniforme, on obtiendra une simulation $x_1, \ldots, x_N$ de la loi cherchée au moyen de la transformation

$$x_i = h(u_i);$$

Lorsque la formule est de la forme $X = h(U, V)$, on utilise deux simulations uniformes $u_1, \ldots, u_N$ et $v_1, \ldots, v_N$ en posant

$$x_i = h(u_i, v_i),$$

ou une seule, deux fois plus longue, en posant

$$x_i = h(u_{2i-1}, u_{2i})$$

(il est important, pour préserver l'indépendance, de ne pas "réutiliser" le même échantillon pour plus d'un échantillon de X).

5.2.1 Loi uniforme sur un intervalle borné

On a vu, parmi les propriétés des lois uniformes, que, si U est uniforme sur $[0,1]$, $\alpha U + \beta$ est uniforme sur $[\beta, \alpha + \beta]$ (cas $\alpha > 0$). Par conséquent, on peut simuler la loi uniforme sur $[a, b]$ en posant

$$X = a + (b - a)U.$$

5.2.2 Loi exponentielle

La loi exponentielle de paramètre λ, a pour fonction de répartition, $F(x) = 1 - e^{-ax}$. Un calcul élémentaire donne la fonction réciproque

$$G(x) = -\frac{1}{a}\ln(1-x) = \frac{\ln\left(\frac{1}{1-x}\right)}{a}.$$

La méthode de la transformation inverse donne donc la formule :

$$X = \frac{1}{a}\ln\left(\frac{1}{1-U}\right).$$

Remarquons que, $1-U$ étant elle-même uniforme sur $[0,1]$ lorsque U l'est, on peut aussi appliquer la formule

$$X = \frac{1}{a}\ln(1/U).$$

En particulier, si U est uniforme sur $[0,1]$, $-\ln(U)$ est exponentielle de paramètre 1.

5.2.3 Loi de Poisson

Plutôt que d'appliquer la méthode de la transformation inverse, il est préférable de combiner la simulation de la loi exponentielle décrite ci-dessus, avec le résultat suivant :

Théorème 5.2 *Soit $(Y_n)_{n\geq 1}$ une suite de variables aléatoires indépendantes, exponentielles de paramètre 1. On pose*

$$X = \sup\{n : Y_1 + \cdots + Y_n \leq \lambda\}.$$

Alors X suit la loi de Poisson de paramètre λ.

Preuve: Il suffit de calculer $\mathbb{P}(X = k)$, pour un entier k quelconque. On a

$$\mathbb{P}(X = k) = \mathbb{P}\left(\sum_{i=1}^{k} Y_i \leq \lambda, \sum_{i=1}^{k+1} Y_i > \lambda\right).$$

Le $(k+1)$-uplet $(Y_1, \ldots, Y_{k+1})$ a pour densité $e^{-(y_1+\cdots+y_{k+1})}$, on a donc

$$\mathbb{P}(X = k) = \int_A e^{-(y_1+\cdots+y_{k+1})} dy_1 \ldots dy_{k+1},$$

l'ensemble d'intégration A étant défini par

$$A = \{(y_1, \ldots, y_{k+1}) : y_i > 0, \sum_{i=1}^{k} y_i < \lambda, \sum_{i=1}^{k+1} y_i > \lambda\}.$$

En intégrant d'abord par rapport à la variable y_{k+1}, il vient

$$\mathbb{P}(X = k) = \int_{\sum_{i=1}^{k} y_i < \lambda} \left(\int_{\lambda - \sum_{i=1}^{k} y_i}^{+\infty} e^{-y_{k+1}} dy_{k+1}\right) e^{-(y_1+\cdots+y_k)} dy_1 \ldots dy_k.$$

L'intégrale intérieure se calcule aisément :

$$\int_{\lambda-\sum_{i=1}^{k} y_i}^{+\infty} e^{-y} dy = e^{-\lambda+\sum_{i=1}^{k} y_i}.$$

La probabilité devient donc

$$\mathbb{P}(X = k) = e^{-\lambda} \int_{\sum_{i=1}^{k} y_i < \lambda} dy_1 \dots dy_k.$$

Le changement de variable $y_i = \lambda x_i$ donne

$$\mathbb{P}(X = k) = e^{-\lambda} \lambda^k \int_{\sum_{i=1}^{k} x_i < 1} dx_1 \dots dx_k.$$

Pour calculer cette dernière intégrale, on pose

$$I_k(x) = \int_{\sum_{i=1}^{k} x_i < \mathbf{x}} dx_1 \dots dx_k$$

(nous cherchons à calculer $I_k(1)$), et on calcule par récurrence : $I_1(x) = x$, et

$$\begin{aligned} I_k(x) &= \int_0^x \left(\int_{\sum_{i=1}^{k-1} x_i < x - x_k} dx_1 \dots dx_{k-1} \right) dx_k \\ &= \int_0^x I_{k-1}(x-t) dt = \int_0^x I_{k-1}(t) dt, \end{aligned}$$

avec comme solution $I_k(x) = x^k/k!$ et $I_k(1) = 1/k!$.

On obtient donc $\mathbb{P}(X = k) = e^{-\lambda}\lambda^k/k!$, ce qui est bien la loi de Poisson de paramètre λ. □

L'algorithme de simulation de la loi de Poisson de paramètre λ, basé sur ce théorème, se contentera de simuler des variables exponentielles jusqu'à ce que leur somme dépasse λ, le résultat de la simulation étant le nombre de variables simulées (moins 1).

Exercice 5.1 *Reformuler l'algorithme pour qu'il travaille directement avec des variables uniformes plutôt que des variables exponentielles.*

5.2.4 Loi normale

La fonction de répartition de la loi normale réduite est

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt,$$

et ni elle ni sa fonction réciproque ne peuvent être exprimées analytiquement. Ceci rend peu attrayante la méthode de la transformation inverse pour la simulation des lois gaussiennes.

Une solution est fournie par le théorème suivant :

Théorème 5.3 *Soient ρ et θ deux variables aléatoires indépendantes, ρ suivant la loi exponentielle de paramètre 1, et θ la loi uniforme sur $[0, 2\pi]$.*

On pose $X = \sqrt{2\rho}\cos(\theta)$ et $Y = \sqrt{2\rho}\sin(\theta)$. Alors X et Y sont indépendantes, et suivent toutes deux la loi normale $\mathcal{N}(0, 1)$.

Preuve: Calculons la densité de (X, Y), au moyen de la formule de la densité image. Le couple (ρ, θ) a pour densité

$$f(\rho, \theta) = \frac{1}{2\pi} e^{-\rho}$$

(le produit des densités des lois marginales, puisque les variables sont indépendantes). Par ailleurs, la transformation

$$\Psi : (\rho, \theta) \mapsto (X, Y)$$

établit une bijection entre $]0, +\infty[\times[0, 2\pi[$ et $\mathbb{R}^2 \backslash \{(0, 0)\}$[1], la transformation inverse Ψ^{-1} étant donnée par

$$\begin{aligned} \rho &= \frac{x^2 + y^2}{2} \\ \theta &= \begin{cases} \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) & (y \geq 0) \\ 2\pi - \arccos\left(\frac{x}{\sqrt{x^2+y^2}}\right) & (y < 0) \end{cases} \end{aligned}$$

La matrice jacobienne de Ψ est

$$J_\Psi(\rho, \theta) = \begin{pmatrix} \frac{1}{\sqrt{2\rho}}\cos(\theta) & -\sqrt{2\rho}\sin(\theta) \\ \frac{1}{\sqrt{2\rho}}\sin(\theta) & \sqrt{2\rho}\cos(\theta) \end{pmatrix},$$

et le jacobien (son déterminant) est donc de 1.

Le couple (X, Y) a donc pour densité, au point (x, y),

$$g(x, y) = f(\Psi^{-1}(x, y)) = e^{-(x^2+y^2)/2} = e^{-x^2/2} e^{-y^2/2}.$$

Cette densité s'exprime comme le produit d'une densité fonction de x, et d'une densité fonction de y : les variables X et Y sont donc indépendantes ; de plus, les deux densités en question sont celle de la loi normale $\mathcal{N}(0, 1)$, ce qui termine la preuve. □

Ce théorème nous donne une solution pour simuler la loi normale, et même pour obtenir 2 échantillons indépendants X et Y à partir de 2 échantillons indépendants U et V de la loi uniforme sur $[0, 1]$:

$$\begin{aligned} X &= \sqrt{-2\ln(U)}\cos(2\pi V) \\ Y &= \sqrt{-2\ln(U)}\sin(2\pi V) \end{aligned}$$

La simulation d'une gaussienne non réduitee $\mathcal{N}(m, \sigma)$ s'en déduit par la transformation affine décrite en 3.3.4.

Enfin, la définition de la loi du χ^2 nous donne immédiatement une méthode pour simuler cette loi : il suffit de simuler N gaussiennes réduites indépendantes, et de prendre la somme de leurs carrés.

[1] Le fait qu'il "manque" un point d'un côté, et un segment de l'autre, n'est pas important : dans les deux cas, la probabilité des points manquants est nulle.

Chapitre 6

Processus markoviens discrets

Sommaire

Il arrive fréquemment que l'on ait besoin de décrire l'évolution aléatoire d'une quantité au cours du temps. Dans ce cas, on utilise une famille de variables aléatoires (X_t), indexée par le "temps". On parle alors de *processus stochastique*, ou plus simplement de *processus*.

On distingue classiquement les processus selon que le temps est représenté de manière continue (t prend ses valeurs dans $\mathbb{R}$, ou dans $\mathbb{R}^+$; on parle alors de processus *à temps continu*), ou de manière discrète (t prend ses valeurs dans $\mathbb{N}$ ou $\mathbb{Z}$; le processus est alors simplement une *suite*, éventuellement doublement infinie, de variables aléatoires, et l'on parle alors de processus *à temps discret*). Une seconde distinction peut être faite selon que les *valeurs* prises par les variables aléatoires sont dans un espace discret ou continu. Dans ce cours, nous nous intéressons exclusivement à des processus à temps discret, à valeurs dans un ensemble discret, et appartenant à une classe particulière, appelés *processus markoviens* ou *chaînes de Markov*.

6.1 Généralités

6.1.1 Définition

Définition 6.1 *Une suite de variables aléatoires $(X_n)_{n\geq 0}$, à valeurs dans un ensemble* fini ou dénombrable *E, est un* processus markovien discret *si, pour tout n et toute suite $e_0, \ldots, e_n, e_{n+1}$ d'éléments de E telle que*

$$\mathbb{P}(X_0 = e_0, X_1 = e_1, \ldots, X_n = e_n) > 0,$$

on a

$$\mathbb{P}(X_{n+1} = e_{n+1} | X_0 = e_0, \ldots, X_n = e_n) = \mathbb{P}(X_{n+1} = e_{n+1} | X_n = e_n).$$

Cette propriété porte le nom de propriété de Markov.

Cette définition doit être comprise en considérant que n représente l'instant présent. La suite $(e_0, \ldots, e_n)$ représente donc l'ensemble du *passé* (connu) du procesus, et la variable aléatoire X_{n+1}, son futur proche.

La propriété de Markov qui définit un processus markovien peut donc s'exprimer ainsi : **le futur (en fait, sa loi de probabilités) ne dépend *que* du présent, et pas du reste du passé.** On parle aussi de processus *sans mémoire*.

6.1.2 Quelques exemples

Marche aléatoire sur un échiquier

On considère les 64 cases d'un échiquier classique, numérotées arbitrairement de 1 à 64. On place initialement un cavalier dans la case 1, et on applique la règle suivante : au k-ème coup, on dénombre les mouvements possibles pour le cavalier (selon sa position, il y en aura 2, 3, 4, 6 ou 8) ; on choisit alors l'un de ces coups au hasard, uniformément, et indépendamment des coups précédents.

Si l'on note X_i, le numéro de la case occupée par le cavalier après le i-ème mouvement, la suite $(X_n)_{n\geq 0}$ est un exemple typique de processus markovien discret.

Un jeu de hasard

Un joueur, initialement muni d'une somme M, joue à un jeu de hasard selon la règle suivante : au n-ème coup, il mise n, sauf s'il n'a plus les fonds nécessaires, auquel cas il s'arrête définitivement de miser. Le jeu est équitable : le joueur a probabilité $1/2$ de gagner une somme égale à sa mise, et probabilité $1/2$ de perdre sa mise. Les parties sont indépendantes (par exemple, chaque partie est résolue par un lancer d'une pièce équilibrée).

Si X_n désigne la richesse du joueur après n parties (qu'il ait ou non misé à chacune d'elles), alors (X_n) est un nouvel exemple de processus markovien discret. (Cet exemple présente une différence importante avec le précédent : le temps intervient dans la "règle" qui est appliquée.)

6.1.3 Spécification d'un processus markovien

On appelle *probabilités de transitions*, les probabilités conditionnelles

$$p_{k,e,e'} = \mathbb{P}(X_{k+1} = e' | X_k - e)$$

qui sont *a priori* définies pour tout triplet $(k, e, e') \in \mathbb{N} \times E \times E$ tel que $\mathbb{P}(X_k = e)$.

Les valeurs prises par un processus markovien (les éléments de E) sont souvent appelées *états* du processus ; les *suites* de valeurs que peut prendre le processus au cours du temps sont, elles, appelées *trajectoires* du processus.

Proposition 6.2 *La loi d'un processus markovien discret (X_n) ne dépend que de la loi de X_0 et des probabilités de transition.*

Preuve: La loi du processus est entièrement donnée par les

$$\mathbb{P}(X_0 = e_0, X_1 = e_1, \ldots, X_n = e_n)$$

pour tout n et toute suite d'états $(e_0, \ldots, e_n)$.

Or, on a (formule de Bayes)

$$\mathbb{P}(X_0 = e_0, \ldots, X_n = e_n) = \mathbb{P}(X_0 = e_0)\mathbb{P}(X_1 = e_1|X_0 = e_0)\ldots\mathbb{P}(X_n = e_n|X_0 = e_0, \ldots, X_{n-1} = e_{n-1})$$

et chaque probabilité conditionnelle est, d'après la propriété de Markov, donnée par la probabilité de transition :

$$\mathbb{P}(X_0 = e_0, \ldots, X_n = e_n) = \mathbb{P}(X_0 = e_0)p_{0,e_0,e_1}p_{1,e_1,e_2}\ldots p_{n-1,e_{n-1},e_n}.$$

On a bien exprimé la probabilité de la trajectoire uniquement en fonction de la loi de X_0 et des probabilités de transition. □

Un processus de Markov est donc déterminé par la donnée d'une loi de probabilités initiale, et, pour chaque instant k, de l'ensemble des probabilités de transition $p_{k,e,e'}$.

6.2 Chaînes de Markov

6.2.1 Définition et théorème fondamental

Un processus markovien discret est dit *homogène* (on parle plus souvent de *chaîne de Markov*[1], si ses probabilités de transition ne dépendent pas du temps : pour tous états e et e', et tous n et k tels que les probabilités conditionnelles sont définies,

$$\mathbb{P}(X_{k+1} = e'|X_k = e) = \mathbb{P}(X_{n+1} = e'|X_n = e).$$

En d'autres termes, les probabilités de transition $p_{k,e,e'}$ ne dépendent pas de k, mais seulement de e et e' :

$$p_{k,e,e'} = p_{e,e'}.$$

Un jeu de hasard, dans lequel on joue une succession de coups sans changer les règles du jeu, fournit typiquement un exemple de chaîne de Markov. L'exemple précédent du cavalier, est une chaîne de Markov ; le jeu de hasard où la mise du joueurs évolue en fonction du temps, n'en est pas une.

Matrice de transition d'une chaîne de Markov

Dans le cas où l'ensemble des états E est de cardinal fini N, on peut définir la *matrice de transition* de la chaîne comme étant la matrice $N \times N$, P, dont lignes et colonnes sont indexées par les éléments de E, et dont les coefficients sont définis par

$$P_{e,e'} = \mathbb{P}(X_{k+1} = e'|X_k = e)$$

(la chaîne étant supposée homogène, cette définition ne dépend pas de k).

[1] La terminologie n'est pas complètement figée : ce que nous appelons ici processus de Markov discret est parfois appelé chaîne de Markov ; ce que nous définissons ici sous le terme de chaîne de Markov est alors appelé chaîne de Markov homogène.

Dans le cas où E est infini dénombrable, la matrice de transition est définie de la même manière, à ceci près qu'il s'agit alors d'une matrice infinie. Nous ne traiterons pas en détail ce cas.

Une remarque importante est la suivante : dans la matrice de transition d'une chaîne de Markov, la somme des coefficients de chaque ligne vaut 1 :

$$\sum_{e' \in E} P_{e,e'} = 1 \quad (\forall e \in E).$$

En effet, cette somme se réécrit en

$$\sum_{e \in E} \mathbb{P}(X_{k+1} = e' | X_k = e) = \mathbb{P}(X_{k+1} \in E | X_k = e) = 1.$$

Une matrice à coefficients réels positifs, dont la somme des coefficients de chaque ligne vaut 1, est appelée *matrice stochastique*.

Loi de l'état au temps n

On a vu que la loi du processus est entièrement déterminée par la loi de l'état initial et par la matrice de transition. Le théorème suivant fournit un moyen analytique de déterminer la loi de l'etat en un temps futur, connaissant l'état présent.

Proposition 6.3 *Soit, pour tout k, $P^{(k)}$ la matrice dont les coefficients sont définis par*

$$P^{(k)}_{e,e'} = \mathbb{P}(X_{n+k} = e' | X_n = e)$$

(cette définition ne dépend pas de n, par homogénéité)

Alors on a

$$P^{(k)} = P^k.$$

De plus, si, pour $k \geq 0$, V_k est le vecteur ligne (aux coefficients indexés par les éléments de E) défini par

$$(V_k)_e = \mathbb{P}(X_k = e),$$

alors on a

$$V_k = V_0.P^k.$$

Preuve: Nous démontrons la première partie par récurrence sur k. Pour $k = 1$, elle est vraie – c'est la définition de la matrice de transition d'une chaîne de Markov.

Supposons vraie la propriété pour $k \geq 1$, et montrons qu'elle est alors encore vraie pour $k + 1$. Soient e et e' deux états de la chaîne, et n un temps tel que $\mathbb{P}(X_n = e) > 0$.

On a clairement

$$\{X_n = e, X_{n+k+1} = e'\} = \bigcup_{e'' \in E} \{X_n = e, X_{n+k} = e'', X_{n+k+1} = e'\},$$

l'union étant une union disjointe ; cette égalité se traduit donc en une somme sur les probabilités, et, par division par $\mathbb{P}(X_n = e)$, sur les probabilités conditionnelles :

$$\mathbb{P}(X_{n+k+1} = e' | X_n = e) = \sum_{e'' \in E} \mathbb{P}(X_{n+k} = e'', X_{n+k+1} = e' | X_n = e).$$

On peut réécrire les probabilités conditionnelles du membre droit, en utilisant la propriété de Markov et l'hypothèse de récurrence :

$$\begin{aligned}\mathbb{P}(X_{n+k}=e'', X_{n+k+1}=e'|X_n=e) &= \mathbb{P}(X_{n+k}=e''|X_n=e)\mathbb{P}(X_{n+k+1}=e'|X_{n+k}=e'')\\ &= P^{(k)}_{e,e''}P_{e'',e'}.\end{aligned}$$

En reportant cette expression dans celle pour $\mathbb{P}(X_{n+k+1}=e'|X_n=e)=P^{(k+1)}_{e,e'}$, on obtient la formule

$$P^{(k+1)}=\sum_{e''\in E}P^{(k)}_{e,e''}P_{e'',e'},$$

qui est la formule du produit de matrices. On a donc la relation $P^{(k+1)}=P^{(k)}P=P^{k+1}$.

La deuxième partie de la proposition découle directement de la première, car la formule des probabilités totales donne

$$\mathbb{P}(X_k=e')=\sum_{e\in E}\mathbb{P}(X_0=e)\mathbb{P}(X_k=e'|X_0=e),$$

ce qui peut être immédiatement réécrit sous la forme du produit

$$V_k=V_0.P^{(k)}.$$

□

6.2.2 Exemple d'une marche aléatoire sur une ligne

On considère un "marcheur ivre" qui, à chaque seconde, soit se déplace d'un pas, soit vers la droite (avec probabilité p), soit vers la gauche (avec probabilité q), soit, avec probabilité $1-p-q$, reste sur place – le tout, en oubliant tout ce qu'il a fait jusqu'à présent (c'est-à-dire que le pas qu'il effectue à la n-ème seconde est indépendant des pas qu'il a fait auparavant). Notre marcheur est cependant confiné dans un couloir dont la longueur est de N pas.

On peut envisager différents types de "conditions aux bords", qui permettent de décider ce qui se passe si le marcheur est à un bout du couloir : conditions "absorbantes" (le marcheur s'arrête définitivement), conditions "réfléchissantes" (s'il est à un bout du couloir, le marcheur effectue avec probabilité $p+q$ le pas qui l'en éloigne), conditions "butoir" (s'il est au fond du couloir à droite, le marcheur reste sur place avec probabilité $1-q$, et effectue un pas vers la gauche avec probabilité q, et symétriquement au fond à gauche) ; tous les processus stochastiques obtenus (où la variable X_n représente la position du marcheur au bout de n secondes) ont en commun d'être des chaînes de Markov que l'on appelle traditionnellement des *marches aléatoires*.

Dans le cas $N=5$, $p=q=1/2$ (si $p=q$, on parle de marche aléatoire *symétrique*), avec bords absorbants, la matrice de transition est

$$P=\begin{pmatrix}1 & 0 & 0 & 0 & 0\\ 1/2 & 0 & 1/2 & 0 & 0\\ 0 & 1/2 & 0 & 1/2 & 0\\ 0 & 0 & 1/2 & 0 & 1/2\\ 0 & 0 & 0 & 0 & 1\end{pmatrix}$$

On peut alors calculer les puissances de cette matrice : si k est pair,

$$P^k = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 3/4 & 0 & 0 & 0 & 1/4 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/4 & 0 & 0 & 0 & 3/4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} + \left(\frac{1}{2}\right)^{1+k/2} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 1 & -1 \\ -1 & 0 & 2 & 0 & -1 \\ -1 & 1 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix};$$

si k est impair,

$$P^k = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 3/4 & 0 & 0 & 0 & 1/4 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/4 & 0 & 0 & 0 & 3/4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} + \left(\frac{1}{2}\right)^{3/2+k/2} \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 2 & 0 & -1 \\ -2 & 2 & 0 & 2 & -2 \\ -1 & 0 & 2 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Dans tous les cas, on voit que, lorsque k tend vers $+\infty$, la matrice P^k tend vers

$$Q = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 3/4 & 0 & 0 & 0 & 1/4 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/4 & 0 & 0 & 0 & 3/4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Exercice 6.1 *Dire comment on peut interpréter la i-ème ligne de la matrice P^k, ainsi que la i-ème ligne de la matrice Q.*

6.3 Propriétés asymptotiques des chaînes de Markov

Dans cette section, on suppose que les chaînes de Markov considérées sont à ensemble d'états E *fini*.

On a vu précédemment que le vecteur V_0P^n décrit la loi de l'état au temps n, X_n, de la chaîne de Markov. On peut naturellement se demander si les coefficients de ce vecteur ont ou non une limite lorsque n tend vers l'infini. Cela amène naturellement à s'interroger sur les valeurs propres de la matrice de transition.

6.3.1 Propriétés spectrales des matrices stochastiques

Théorème 6.4 *Soit M une matrice réelle carrée de taille N, à coefficients positifs ou nuls, et dont la somme des coefficients de chaque ligne est* 1. *(Une telle matrice est appelée* matrice stochastique.*)*

Alors 1 *est valeur propre de M, et toute valeur propre λ de M (réelle ou non) vérifie* $|\lambda| \leq 1$.

Preuve: Il est facile de vérifier que le vecteur (colonne) dont tous les coefficients sont égaux à 1, est vecteur propre (à droite) de la matrice M.

Soit maintenant λ une valeur propre, et U un vecteur propre (à droite) de M pour la valeur propre λ. Soit i un indice tel que $|U_i|$ soit maximal. On a $\lambda U_i = (MU)_i$, donc (inégalité triangulaire)

$$|\lambda||U_i| \leq \sum_{j=1}^{N} M_{ij}|U_j|;$$

en majorant chaque $|U_j$ par $|U_i|$, on en déduit

$$|\lambda||U_i| \leq \left(\sum_{j=1}^{N} M_{ij}\right)|U_i| = |U_i|$$

puisque la matrice est stochastique. On a donc bien $|\lambda| \leq 1$. □

Dans le cadre des chaînes de Markov, on est amené à considérer non les vecteurs propres (colonnes) à droite, mais les vecteurs propre à gauche de la matrice de transition, c'est-à-dire les vecteurs transposés des vecteurs propres à droite de sa matrice transposée. Toutefois, le spectre d'une matrice réelle étant le même que celui de sa matrice transposée, le théorème précédent reste pertinent.

Par ailleurs, le produit de deux matrices stochastiques est encore une matrice stochastique. Cela a une conséquence importante :

Proposition 6.5 *Soit M une matrice stochastique de taille N. Si la suite $(M^n)_{n\geq 0}$ converge vers une matrice M', alors M' est encore une matrice stochastique.*

Preuve: La convergence est ici une convergence simple, coefficient par coefficient :

$$M'_{i,j} = \lim_{n\to+\infty} (M^n)_{i,j};$$

c'est la convergence dans l'espace vectoriel $\mathbb{R}^(N^2)$, de dimension finie, donc complet. Or l'ensemble des matrices stochastiques est fermé dans cet espace, donc la limite d'une suite de matrices stochastiques, si elle existe, est aussi une matrice stochastique. □

Note : l'extension naïve de cette proposition au cas de matrices stochastiques infinies serait fausse ; en effet, s'il est vrai (en définissant proprement les matrices infinies stochastiques) que le produit de matrices stochastiques infinies reste une matrice stochastique, il n'est plus vrai que la limite simple fait de l'ensemble des matrices stochastiques un espace complet. C'est essentiellement pour éviter ce genre de difficultés que nous nous limitons aux chaînes d'espace d'états fini.

Le théorème suivant est fondamental pour décrire le comportement asymptotique d'une chaîne de Markov, mais sa démonstration complète est assez technique. Nous ne donnons ici qu'une preuve dans un cas particulier ; la preuve générale est donnée en appendice.

Théorème 6.6 *Soit P une matrice carrée stochastique, de taille N.*

- *Si 1 est valeur propre simple de P, et si toutes les valeurs propres de P autres que 1 sont de module strictement inférieur à 1, alors la suite de matrice $(P^n)_{n\geq 0}$ converge vers une matrice stochastique Π. De plus, toutes les lignes de la matrice Π sont identiques.*

- *Si* 1 *est valeur propre multiple de* P*, et que toutes les autres valeurs propres sont de module strictement inférieur à* 1*, la suite* $(P^n)_{n\geq 0}$ *converge vers une matrice stochastique* Π.
- *Si* P *a au moins une valeur propre autre que* 1 *dont le module est égal à* 1*, alors la suite* $(P^n)_{n\geq 0}$ *ne converge pas. Toutefois, la suite* $(Q_n)_{n\geq 1}$*, définie par*

$$Q_n = \frac{1}{n}\left(I + P + \cdots + P^{n-1}\right),$$

converge vers une matrice stochastique Π *(ou, dit autrement, la suite* (P^n) *converge* au sens de Cesaro *vers une matrice stochastique* Π*).*

Preuve partielle : Nous démontrons les deux premières parties de ce théorème dans le cas particulier où la matrice P est *diagonalisable* ; pour une preuve plus complète, utilisant la réduction de Jordan au lieu de la diagonalisation, on pourra consulter [3]. Dans ce cas, il existe une matrice inversible A (matrice de changement de base) telle que $P = A^{-1}DA$, où D est une matrice diagonale dont les coefficients diagonaux sont exactement les valeurs propres de P (avec leur ordre de multiplicité). On a alors, en élevant à la puissance n, $P^n = A^{-1}D^nA$; la matrice D^n est toujours diagonale, avec comme coefficients diagonaux, les puissances n-èmes des valeurs propres. Pour toutes les valeurs propres λ avec $|\lambda| < 1$, on a $\lambda^n \to 0$. Par conséquent, le comportement de la matrice D^n est relativement simple : si 1 est la seule valeur propre de module 1, D^n converge vers une matrice diagonale D' dont les coefficients diagonaux sont tous 0 ou 1 ; sinon, D^n ne converge pas.

Le premier cas du théorème en découle : la suite P^n converge alors vers la matrice $\Pi = A^{-1}D'A$, et la matrice D' a un seul coefficient diagonal égal à 1 (disons, le k-ème). Si l'on exprime les cofficients de Π par la formule du produit, on obtient

$$\Pi_{i,j} = (A^{-1})i, k(A)_{k,j},$$

ce qui nous permet d'affirmer que chaque ligne de Π est proportionnelle à la k-ème ligne de la matrice A. Mais Π, comme limite d'une suite de matrices stochastiques, est elle-même stochastique. Par conséquent, toutes les lignes de Π sont identiques.

Le deuxième cas est du même acabit. Les matrices D^n convergent encore vers une matrice D', et P^n vers $A^{-1}D'A$ qui, comme limite d'une suite de matrices stochastiques, est encore une matrice stochastique. Il reste à vérifier que ses lignes ne sont pas toutes égales, ce qui revient à démontrer qu'elle n'est pas de rang 1. Il se trouve que le rang de Π est exactement la multiplicité de 1 comme valeur propre ; en effet, on peut reconnaître dans la matrice $\Pi = A^{-1}D'A$, la matrice de la projection sur le sous-espace engendré par les vecteurs colonnes de A^{-1} correspondant aux coefficients non nuls de D'.

Dans le troisième cas, il est clair que les matrices D^n ne convergent pas (les coefficients diagonaux de la forme λ^n, avec $|\lambda| = 1$ mais $\lambda \neq 1$, n'ont pas de limite), mais cela n'implique pas immédiatemant que P^n ne converge pas (il pourrait y avoir des compensations lors du produit $A^{-1}D^nA$). □

6.3.2 Distribution limite et ergodicité

Considérons différentes chaînes de Markov de même matrice de transition P, mais qui diffèrent par la loi de l'état initial X_0. En fait, pour chaque état $e \in E$, nous pouvons définir

une chaîne $(X_n^{(e)})_{n\geq 0}$, la *chaîne partant de l'état* e, en donnant comme loi à $X_0^{(e)}$,

$$\mathbb{P}(X_0^{(e)} = e) = 1.$$

Alors, la matrice P^n peut être interprétée de la façon suivante : *la ligne e de la matrice P^n donne la loi de l'état $X_n^{(e)}$*. Par conséquent, si la suite de matrices (P^n) a une limite lorsque n tend vers l'infini, alors pour tout état initial e, *la loi de $X_n^{(e)}$ converge vers une loi de probabilité qui est donnée par la ligne e de la matrice limite*. C'est cette situation qui a été observée dans l'exemple de la marche aléatoire symétrique du paragraphe 6.2.2. Notons que, dans ce cas, les lignes de la matrice limite ne sont pas toutes les mêmes : la loi limite de X_n dépend de l'état initial.

Le vecteur $V^n = VP^n$, qui décrit la loi de l'état au temps n, est appelé *vecteur d'état*.

Définition 6.7 *Une chaîne de Markov est dite* ergodique *si le vecteur d'état au temps n, $V.P^n$, tend vers une limite V^∞, et si cette limite ne dépend pas du vecteur initial V.*

Note : *stricto sensu*, la propriété d'*ergodicité* n'est pas une propriété de la chaîne de Markov, mais de sa matrice de transition ; notre définition d'une chaîne de Markov incluant le choix de la loi de l'état initial X_0, la propriété d'ergodicité est une propriété collective de toutes les chaînes de Markov qui partagent la même matrice de transition. Cet abus de langage est toutefois consacré par l'usage.

Lemme 6.8 *Une chaîne de Markov de matrice de transition P est ergodique si et seulement si la suite de matrices P^n a une limite Π lorsque n tend vers $+\infty$.*

Preuve: Supposons la chaîne ergodique. Alors, il existe une loi de probabilités $V^{(\infty)}$ (vue comme vecteur ligne) telle que, pour tout vecteur stochastique V, on ait

$$V^{(\infty)} = \lim_{n\to+\infty} VP^n.$$

En prenant comme vecteur V, le vecteur V_i dont tous les coefficients sont nuls, sauf le i-ème qui vaut 1, V_iP^n est la i-ème ligne de la matrice P^n ; l'ergodicité de la chaîne implique donc que toutes les lignes de P^n convergent vers une même ligne $V^{(\infty)}$: P^n converge vers une matrice stochastique dont chaque ligne est égale à $V^{(\infty)}$.

Réciproquement, supposons que chaque ligne de P^n converge vers une même ligne $V^{(\infty)}$ (hypothèse de convergence de P^n vers une matrice stochastique de rang 1). Alors, si V est un vecteur ligne stochastique, on vérifie immédiatement que VP^n converge vers la même ligne. □

En combinant le Lemme 6.8 et le Théorème 6.6, on obtient le théorème suivant, qui caractérise les matrices de transition des chaînes ergodiques :

Théorème 6.9 *Une chaîne de Markov de matrice de transition P est ergodique si et seulement si la valeur propre 1 est valeur propre simple, et est l'unique valeur propre de module 1. La loi limite de la chaîne est alors donnée par l'unique vecteur propre à gauche (ligne) de la matrice P (c'est-à-dire, les probabilités limites sont les coefficients de l'unique tel vecteur propre dont la somme des coefficients soit 1).*

6.4 Graphe d'une chaîne de Markov

Le cours sur les graphes n'a lieu qu'en deuxième semestre de première année, et cette section met donc quelque peu la charrue avant les boeufs. Toutefois, il semble dommage de parler de chaînes de Markov sans mentionner leur représentation sous la forme de graphes orientés pondérés, d'autant plus que la caractérisation des chaînes de Markov ergodiques est beaucoup plus intuitive dans ce cas.

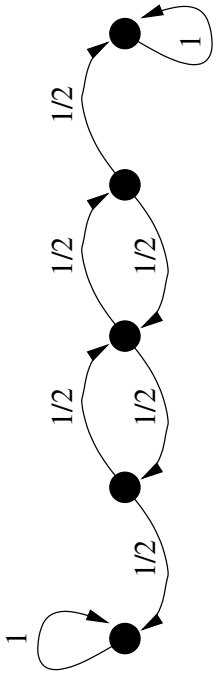
Considérons une chaîne de Markov d'espace d'états fini E, et de matrice de transition P. Nous pouvons former un "graphe orienté" (un ensemble de points, appelés *sommets*, et de flèches allant chacune d'un des points à un autre (ou au même), appelées *arcs*), de la manière suivante :

- à chaque état $e \in E$, est associé un sommet, lui aussi noté e ;
- pour chaque couple $(e, e') \in E^2$, tel que $p_{e,e'} > 0$, on a une flèche (arc) allant de e à e', sur laquelle on indique la valeur de $p_{e,e'}$.

La disposition exacte des sommets n'est pas importante, et il se peut que les arcs se croisent.

La Figure 6.4 montre le graphe de la "marche aléatoire" étudiée au paragraphe 6.2.2.

FIG. 6.1 – Une marche aléatoire avec bords absorbants



On peut alors visualiser le fonctionnement de la chaîne de Markov de la manière suivante : si, à un temps n quelconque, la chaîne se trouve dans l'état e (ce qui signifie $X_n = e$), l'état au temps $n + 1$ (la valeur de X_{n+1}) est déterminé en choisissant aléatoirement l'un des arcs sortant de e (chaque arc (e, e') ayant probabilité $p_{e,e'}$ d'être celui qui est choisi), et on a alors $X_{n+1} = e'$. En quelque sorte, on effectue une suite de déplacements aléatoires le long des arcs du graphe, les poids $p_{e,e'}$ des différents arcs indiquant à chaque fois la probabilité d'emprunter un arc plutôt qu'un autre.

Dans un tel graphe, on dira qu'un sommet (un état) e' est *accessible* à partir d'un autre état e, s'il existe une succession d'arcs permettant de passer de e à e'.

On peut alors convenir d'appeler *composante puits*, tout ensemble d'états C non vide, qui vérifie les propriétés suivantes :

– si $e \in C$ et $e' \in C$, alors e' est accessible à partir de e ;

– si $e \in C$ et $e' \in C$, alors il n'existe pas d'arc de e à e' (c'est-à-dire que $p_{e,e'} = 0$).

La vision intuitive d'une composante puits est la suivante : si "un jour" on se trouve dans une composante puits, on n'en resortira plus jamais (puisqu'aucun arc ne permet d'en sortir) ; en revanche, depuis n'importe quel sommet d'une telle composante, il est possible d'aller à n'importe quel autre sommet de la composante.

Il est facile de se convaincre que le graphe possède forcément au moins une composante puits (qui peut être réduite à un seul sommet, ou au contraire contenir tous les états). Dans l'exemple de la Figure 6.4, il existe deux composantes puits, chacune étant composée de l'un des sommets extrêmes de la chaîne.

Une chaîne de Markov qui ne possède qu'une seule composante puits, est dite *irréductible*.

Avant de pouvoir caractériser l'ergodicité, il nous faut encore définir la *périodicité* d'une composante puits. Nous dirons qu'il existe un *cycle de longueur d* passant par un état e, s'il existe une suite de d arcs, formant un chemin (c'est-à-dire que la destination d'un arc est l'origine du suivant), et telle que e soit l'origine du premier arc et la destination du dernier. Alors, la *périodicité* de la composante puits et le PGCD (plus grand diviseur commun) de toutes les longueurs de cycles passant par les sommets de la composante. Dans l'exemple de la Figure 6.4, les deux composantes puits sont de périodicité 1, car les deux sommets extrêmes sont porteurs de "boucles" (des arcs allant d'eux-mêmes à eux-mêmes), ce qui donne immédiatement des cycles de longueur 1.

Nous admettrons le théorème suivant :

Théorème 6.10 *Une chaîne de Markov (d'espace d'états fini) est ergodique, si et seulement si elle est* irréductible *(une seule composante puits) et si l'unique composante puits est* de périodicité 1.

De plus, si c'est le cas, la distribution limite V vérifie $V_e = 0$ pour tout état qui ne se trouve pas dans la composante puits, et $V_e > 0$ pour tout état de la composante puits.

Remarquons que, d'après ce théorème, l'ergodicité d'une chaîne de Markov ne dépend pas réellement des *valeurs* des probabilités de transition, mais seulement de celles qui sont strictement positives (puisque le théorème précédent s'exprime entièrement en fonction des arcs qui sont présents dans le graphe, et n'utilise pas les probabilités de transitions $p_{e,e'}$).

On pourrait se demander quel est le lien entre les Théorèmes 6.9 et 6.10. Essentiellement, la condition de périodicité correspond au fait que 1 soit la seule valeur propre de module 1, et la condition de connexité, à la condition sur la multiplicité de la valeur propre 1.

Enfin, notons que ces théorèmes ne seraient pas vrais, tels quels, dans le cas de chaînes de Markov à espace d'états *infinis dénombrables*.

Deuxième partie

Statistiques

Chapitre 7

Généralités sur les statistiques

Sommaire

7.1 Terminologie et notations

Ce cours porte sur les *statistiques inductives*, dont le but est, à partir d'un *échantillon* d'une population, d'induire les propriétés de certains *caractères* de cette population.

7.1.1 Petit lexique Probabilités-Statistiques

Le terme de *population* recouvre ce que l'on appelle généralement *univers* en Probabilités : c'est un ensemble d'individus dont on désire induire des propriétés ; les individus jouent donc le rôle des *événements élémentaires* de la théorie des Probabilités. Implicitement, la loi de probabilités sur une population finie est toujours la loi uniforme (s'il y a N individus, chacun a une probabilité $1/N$).

Dans la terminologie standard en Statistiques, le mot de "caractère" correspond assez précisément à ce qui a précédemment été appelé "variable aléatoire" : la valeur de chaque caractère est définie individuellement pour chaque individu ; sur l'ensemble de la population, le caractère se présente donc comme une *fonction*, à valeurs dans un ensemble qui peut être $\mathbb{R}$ (par exemple, la taille des individus, exprimée en centimètres, si les individus sont des personnes) ou un ensemble fini abstrait ($\{\text{Masculin}, \text{Féminin}\}$ pour le sexe), défini sur la même population.

Lorsque X est un caractère, un *échantillon de taille n* de modèle X est une suite $X_1, \ldots, X_n$ de n variables aléatoires indépendantes, chacune ayant la loi de X. Cela correspond à choisir aléatoirement (uniformément) et indépendamment n individus parmi la population, et à noter,

pour chacun d'entre eux, la valeur de leur caractère X (Dans une population finie, il est alors possible que le même individu soit "interrogé" plus d'une fois ; c'est assez probable si n est sensiblement plus grand que $\sqrt{N}$, et tres improbable si n est sensiblement plus petit que $\sqrt{N}$). La suite $x_1, \ldots, x_n$ de valeurs (numériques ou non) ainsi obtenues est alors appelée *série statistique* ; en termes de théorie des Probabilités, ce n'est rien d'autre qu'une réalisation de la loi de l'échantillon.

Lorsque le "caractère" X est à valeurs dans un espace produit $E \times F$, on considère en fait qu'il s'agit d'un couple (Y, Z) (Y étant à valeurs dans E, et Z à valeurs dans F) et l'on parle de *série statistique double*. En cas de triplets (respectivement, de n-uplets), on parle de *série statistique triple* (respectivement, *n-uple*).

7.1.2 Notations usuelles

La Figure 7.1 présente quelques notations usuelles, et les noms qui leurs sont associés. Ici, les notations X et Y désignent deux caractères numériques d'une même population, $(X_1, Y_1), \ldots, (X_N, Y_N)$ un échantillon de taille N, et $(x_1, y_1), \ldots, (x_N, y_N)$ la série statistique double associée.

FIG. 7.1 – Notations usuelles en Statistiques

Notation	**Formule**	**Nom**
$\overline{X}$	$\frac{1}{N}\sum_k X_k$	estimateur (standard) de moyenne
$\overline{x}$	$\frac{1}{N}\sum_k x_k$	moyenne de la série statistique
$C^*_{X,X}$	$\frac{1}{N}\sum_k (X_k - \overline{X})^2$	estimateur (standard) de variance
$c_{x,x}$	$\frac{1}{N}\sum_k (x_k - \overline{x})^2$	variance de la série statistique
$C^*_{X,Y}$	$\frac{1}{N}\sum_k (X_k - \overline{X})(Y_k - \overline{Y})$	estimateur (standard) de covariance
$c_{x,y}$	$\frac{1}{N}\sum_k (x_k - \overline{x})(y_k - \overline{y})$	covariance de la série statistique
$R^*_{X,Y}$	$C^*_{X,Y}/\sqrt{C^*_{X,X}C^*_{Y,Y}}$	estimateur de coefficient de corrélation
$r_{x,y}$	$c_{x,y}/\sqrt{c_{x,x}c_{y,y}}$	coefficient de corrélation de la série statistique

Le sens de la notion d'*estimateur* sera précisé au chapitre 8.

On notera l'usage mnémotechnique de *majuscules* pour les fonctions, et de *minuscules* pour leurs valeurs numériques.

Les formules suivantes (dans lesquelles l'opérateur est appliqué à des fonctions qui ne sont pas des caractères d'origine) sont toutes des conséquences de la linéarité de cet opérateur, et sont le pendant statistiques des formules vues sur l'espérance, la variance et la covariance :

$$\begin{aligned}
C^*_{X,X} &= \overline{X^2} - \left(\overline{X}\right)^2 = \overline{(X - \overline{X})^2} \\
c_{x,x} &= \overline{x^2} - (\overline{x})^2 = \overline{(x - \overline{x})^2} \\
C^*_{X,Y} &= \overline{(XY)} - \overline{X}.\overline{Y} = \overline{(X - \overline{X})(Y - \overline{Y})} \\
c_{x,y} &= \overline{(xy)} - \overline{x}.\overline{y} = \overline{(x - \overline{x})(y - \overline{y})}
\end{aligned}$$

7.2 Représentations graphiques

Il est fréquent de représenter graphiquement, sous forme condensée, des séries statistiques. Nous passons en revue quelques représentations classiques, en détaillant les situations dans lesquelles elles sont appropriées.

7.2.1 Diagramme "en bâtons"

Dans les cas où le caractère étudié X suit une loi discrète, et où le nombre de valeurs distinctes prises n'est pas trop important, le diagramme dit "en bâtons" est assez approprié.

Si les valeurs prises par la série statistique $x_1, \dots, x_N$ sont $a_1, \dots, a_m$, on appelle *fréquence de la valeur a_i dans la série statistique*, le *nombre* d'occurrences de la valeur a_i, divisé par N ; ainsi, si on note p_i cette fréquence,

$$p_i = \frac{1}{N} \#\{j \leq N : x_j = a_i\}.$$

On remarque que l'on a automatiquement

$$\sum_{i=1}^{m} p_i = 1.$$

Pour former le diagramme en bâtons, on place, à chaque abscisse a_i (en supposant que les valeurs sont numériques ; si ce n'est pas le cas, on ordonne les valeurs dans un ordre arbitraire, et on utilise l'abscisse i à la place de a_i), un trait vertical de hauteur proportionnelle à p_i.

Justification heuristique du diagramme en bâtons

On peut considérer que la fréquence p_i est la valeur prise par une variable aléatoire P_i, qui serait la proportion, dans un échantillon de taille N, des individus dont le caractère X vaut exactement a_i. Si l'on note

$$p'_i = \mathbb{P}(X = a_i),$$

la variable aléatoire $\mathbf{1}_{a_i} = \mathbf{1}_{\{X=a_i\}}$ (définie dans l'espace de probabilités où est définie X, et que l'on peut donc considérer comme un caractère de notre population) a pour espérance p'_i, et P_i **n'est autre que l'estimateur standard de moyenne de ce nouveau caractère**.

Par ailleurs, la loi des grands nombres vue au Chapitre 4 permet d'affirmer que, pour tout i ($1 \leq i \leq m$), *la variable aléatoire P_i (qui est en fait une variable $P_i^{(N)}$, dépendant de la taille N de l'échantillon) converge en probabilités vers la constante p'_i lorsque N tend vers l'infini.* Par conséquent, on peut considérer que, si N est "assez grand", le diagramme en bâtons donnera une représentation fidèle de la *loi du caractère X* – ce qui est l'objectif.

7.2.2 Histogramme

Considérons un caractère numérique X, et une série statistique $x_1, \dots, x_N$ associée à un échantillon $X_1, \dots, X_N$ de taille N.

Si la loi que suit X est une loi diffuse, la probabilité de chaque valeur possible est 0, et, en conséquence, la probabilité que deux valeurs x_i et x_j soient égales, est nulle également. On utilise le plus souvent, pour représenter graphiquement la série statistique, un *histogramme*.

On définit une *partition* de l'ensemble des valeurs en une succession d'*intervalles* $(A_i)_{1\leq i\leq m}$: $A_i = [a_{i-1}, a_i[$, de telle sorte que toutes les valeurs prises par la série statistique soient comprises entre a_0 et a_m.

Chaque intervalle est renommé "classe", et se voit attribuer une fréquence de la même manière qu'au paragraphe précédent :

$$p_i = \frac{1}{N}\#\{j \leq N : a_{i-1} \leq X_j < a_i\}.$$

On a donc toujours

$$\sum_{i=1}^{m} p_i = 1.$$

La différence avec le diagramme en bâtons est que chaque classe A_i est représentée graphiquement par un rectangle situé entre les abscisses a_{i-1} et a_i, et dont *l'aire* (et non la hauteur) est proportionnelle à p_i.

La Figure 7.2 représente, sur un même graphique, la densité de la loi normale réduite, et l'histogramme d'un échantillon de taille 8000 de la loi normale réduite (échantillon obtenu par la méthode de simulation décrite en 5.2.4).

Justification heuristique de l'histogramme

De même que pour le diagramme en bâtons, les fréquences P_i convergent en probabilités (lorsque la taille de l'échantillon tend vers l'infini) vers les probabilités

$$p'_i = \mathbb{P}(X \in A_i);$$

pour peu que les classes soient suffisamment petites pour que la *densité* de X soit à peu près constante sur chaque classe, on s'attend donc à ce que, sur un échantillon assez grand, *l'histogramme donne une représentation assez fidèle de la densité du caractère étudié.*

Remarques sur le choix des classes

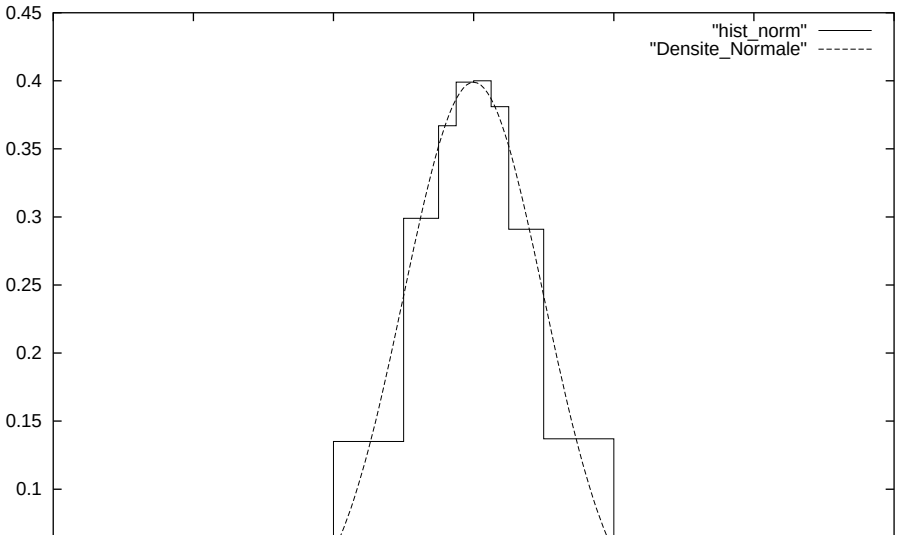
Pour que l'histogramme soit significatif, il est nécessaire de respecter certaines règles dans le choix des classes A_i. D'une part, il faut que chaque classe soit assez petite pour que la fonction densité ne subisse pas des variations trop importantes à l'intérieur de la classe ; mais d'autre part, il est bon que chaque classe soit assez grande pour contenir plusieurs termes de la série statistique (un ordre de grandeur raisonnable est d'au moins 5, sauf éventuellement pour les classes extrêmes A_1 et A_m).

Ces deux contraintes, prises ensemble, vont typiquement imposer que la taille N de la série statistique soit assez grande.

7.3 Régression

Lorsque l'on dispose d'une série statistique double $(x_1, y_1), \ldots, (x_N, y_N)$, on peut bien entendu construire une représentation graphique séparée pour les deux séries $x_1, \ldots, x_N$ et $y_1, \ldots, y_N$, mais cela ne permettra jamais de mettre en évidence d'éventuels liens entre les caractères X et Y.

FIG. 7.2 – Histogramme d'un échantillon de la loi normale réduite



Pour cela, on considère le *nuage de points* du plan $M_1, \ldots, M_N$, le point M_i étant défini par ses coordonnées (x_i, y_i). On cherche alors une courbe "modèle" qui ait la propriété que les points du nuage soient "à peu près situés" sur la courbe.

Cette démarche peut être appliquée à des séries statistiques triples (on a alors un nuage de points dans l'espace, et on cherche une surface qui approche bien le nuage), ou même, bien que cela devienne plus difficile à se représenter mentalemetn, à des séries n-uples (dans un espace à n dimensions). On parle alors de *régression multivariée.*

7.3.1 Droite de régression

Le cas le plus courant de régression est la *régression linéaire* où la courbe que l'on cherche à "coller" à la série statistique double, est en fait une droite. En d'autres termes, on cherche une relation de la forme $Y = aX + b$ qui soit "presque" vérifiée par les caractères X et Y.

Remarque 7.1 (Régression logarithmique) *Parfois, on peut être amené à rechercher une régression linéaire non pas directemetn entre X et Y, mais entre X et $\log(Y)$ ou entre $\log(X)$ et Y.*

Pour chercher une droite qui approche au mieux un ensemble de N points du plan, il faut préciser le sens de "au mieux". Le sens consacré par l'usage, et qui a l'avantage de se prêter à des calculs raisonnables, est celui des *moindres carrés*, dont le sens mathématique est le suivant : on cherche, parmi toutes les droites du plan, celle qui *minimise* la somme des *carrés* des écarts entre les ordonnées des points du nuage et celles des points de mêmes abscisses de la droite. Autrement dit, on cherche a et b qui minimisent la somme

$$\sum_{k=1}^{N} \left(y_k - (ax_k + b)\right)^2 .$$

On cherche donc à minimiser la norme (euclidienne) du vecteur (de dimension N) $(ax_k + b - y_k)_{1 \leq k \leq N}$.

Les méthodes générales de résolution des problèmes de moindres carrés seront vues dans l'UV "Algorithmique Numérique". La solution générale est caractérisée par des *équations normales* qui, dans le cas qui nous occupe, se ramènent à

$$\begin{aligned} a\overline{x^2} + b\overline{x} &= \overline{xy} \\ a\overline{x} + b &= \overline{y} \end{aligned}$$

Le déterminant de ce système est $\overline{x^2} - (\overline{x})^2 = c_{x,x}$, qui est strictement positif sauf dans le cas trivial où *tous les x_k sont égaux* (cas trivial, car alors tous les points du nuage sont situés sur la droite d'équation $X = x_k$, qui est d'emblée la meilleure droite d'approximation).

Dans tous les cas non triviaux, le système a donc une solution unique donnée par

$$\begin{aligned} a &= \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{c_{x,y}}{c_{x,x}} \\ b &= \overline{y} - a\overline{x} \end{aligned}$$

Au passage, la dernière équation nous apprend que *le centre de gravité du nuage de points se trouve sur la droite de régression linéaire.*

Interprétation du coefficient de corrélation

On vient de voir que la covariance $c_{x,y}$ intervient dans le calcul de la droite de régression linéaire ; il se trouve que le coefficient de corrélation peut également être interprété par rapport à cette même droite.

Considérons la quantité suivante : le *carré de la norme euclidienne* du vecteur des écarts défini préalablement. Cela donne

$$\begin{aligned}\sum_{k=1}^{N}(ax_k+b-y_k)^2 &= \sum_{k=1}^{N}(ax_k+(\overline{y}-a\overline{x})-y_k)^2\\ &= \sum_{k=1}^{N}(a(x_k-\overline{x})-(y_k-\overline{y}))^2\\ &= \sum_{k=1}^{N}\left(a^2(x_k-\overline{x})^2-2a(x_k-\overline{x})(y_k-\overline{y})+(y_k-\overline{y})^2\right)\\ &= N\left(a^2c_{x,x}-2ac_{x,y}+c_{y,y}\right).\end{aligned}$$

Si maintenant on se souvient de la formule pour $a=c_{x,y}/c_{x,x}$, cette expression devient

$$\begin{aligned}N\left(\frac{(c_{x,y})^2}{c_{x,x}}-2\frac{(c_{x,y})^2}{c_{x,x}}+c_{y,y}\right) &= Nc_{y,y}\left(1-\frac{(c_{x,y})^2}{c_{x,x}c_{y,y}}\right)\\ &= Nc_{y,y}\left(1-r_{x,y}^2\right).\end{aligned}$$

Par conséquent, si l'on divise ce carré de norme euclidienne par N (ce qui correspond à se ramener à la "contribution individuelle" de chacun des N points du nuage), on obtient $c_{y,y}(1-r_{x,y}^2)$. Au passage, on vient de démontrer (puisque le carré d'une norme ne saurait être négatif, pas plus que la variance de la série Y) que $|r_{x,y}|\leq 1$; on a surtout obtenu une interprétation de $r_{x,y}$, ou tout du moins de sa valeur absolue : *plus le coefficient de corrélation est faible (à variance $c_{y,y}$ fixée), mieux la droite de régression approche (au sens des moindres carrés) le nuage de points.*

Chapitre 8

Estimation

Sommaire

8.1 Estimation ponctuelle

8.1.1 Buts de l'estimation

Lorsqu'on est confronté à une série statistique $x_1, \ldots, x_n$ dont on suppose qu'elle provient d'un échantillon $X_1, \ldots, X_N$ d'un modèle X (ce qui, une fois de plus, revient à dire que l'on *suppose* que les valeurs $x_1, \ldots, x_N$ ont été obtenues par un procédé qui est mathématiquement équivalent à tirer N variables aléatoires indépendantes, toutes de même loi que X), on s'intéresse souvent au problème de *estimer* (au sens : proposer une valeur que l'on juge plausible) une grandeur a, qui est un paramètre dépendant de la loi du caractère X.

Des exemples classiques de tels "paramètres" sont l'espérance, ou la variance, de la loi ; mais on peut aussi s'intéresser à la valeur médiane (un x défini par $F(x) = 1/2$, si F est la fonction de répartition – en supposant qu'un tel x existe, ce qui est assuré si F est continue), ou par exemple à $a = \mathbb{P}(|X| > 1)$.

Une variable aléatoire A, définie comme une *fonction* f_N de N et de l'échantillon $X_1, \ldots, X_N$, sera appelée *estimateur du paramètre* a si l'on a des raisons de penser que la valeur de A, à savoir $f_N(x_1, \ldots, x_N)$, a tendance à être proche de a.

En d'autres termes, n'importe quelle variable aléatoire peut être considérée comme un estimateur de a, mais la plupart de ces "estimateurs" n'auront aucune des qualités particulières que l'on est en droit d'en attendre.

On parle d'*estimation ponctuelle* quand on cherche seulement à proposer une *valeur* comme estimation pour le paramètre évalué. L'*estimation par intervalle de confiance* sera étudiée en 8.3.

8.1.2 Qualités souhaitables d'un estimateur

Estimateur sans biais

Un estimateur A est un *estimateur sans biais* d'un paramètre a, si l'on a

$$\mathbb{E}(A) = a.$$

Exemple 8.1 (Estimateur sans biais pour l'espérance) *$\overline{X}$ est un estimateur sans biais de $\mathbb{E}(X)$ (ce qui justifie a posteriori le nom qui lui a été donné).*

Exemple 8.2 (Estimateur avec biais) *Posons $Z = f_N(X_1, \ldots, X_N) = \frac{1}{N+1}\sum_{k=1}^{N} X_k$; la linéarité de l'espérance donne immédiatement*

$$\mathbb{E}(Z) = \frac{N}{N+1}\mathbb{E}(X),$$

ce qui montre que Z n'est pas un estimateur sans biais de $\mathbb{E}(X)$ (sauf dans le cas particulier où $\mathbb{E}(X) = 0$, mais construire un estimateur d'espérance qui soit sans biais sur l'ensemble des lois d'espérance nulle n'est certainement pas un exploit). En revanche, $\mathbb{E}(Z)$ tend vers *$\mathbb{E}(X)$ lorsque la taille de l'échantillon tend vers l'infini : on dit qu'un tel estimateur est* asymptotiquement sans biais.

Estimateur convergent

Un estimateur A est un *estimateur convergent* d'un paramètre a, s'il *converge en probabilités* vers a lorsque N tend vers l'infini.

Exemple 8.3 (Estimateur convergent pour l'espérance) *La loi faible des grands nombres, telle qu'elle a été vue au chapitre 4, peut être reformulée en :* $\overline{X}$ est un estimateur convergent pour l'espérance $\mathbb{E}(X)$ *(au moins pour les lois admettant une variance).*

Proposition 8.4 *Un estimateur sans biais, et dont la variance tend vers 0 lorsque N tend vers l'infini, est convergent.*

Preuve: On applique l'inégalité de Tchebycheff. Soit σ_n^2 la variance de l'estimateur A sur un échantillon de taille N. Fixons un $\epsilon > 0$ quelconque ; l'inégalité de Tchebycheff donne

$$\mathbb{P}(|A_N - a| \geq \epsilon) \leq \left(\frac{\sigma_N}{\epsilon}\right)^2,$$

ce qui nous garantit que, pour tout $\epsilon > 0$, $\mathbb{P}(|A_N - a| > \epsilon)$ tend bien vers 0 lorsque N tend vers l'infini ; c'est exactement la définition de la convergence en probabilités vers la constante a. □

Un estimateur qui est à la fois *sans biais* et *convergent* est dit *correct*. Notre ami $\overline{X}$ est donc un estimateur correct de l'espérance de X.

Estimateur de variance minimum

Un estimateur A du paramètre a est dit *de variance minimum*, si l'on a

$$\mathbb{E}((A-a)^2) \leq \mathbb{E}((B-a)^2)$$

pour toute variable aléatoire B définie en termes de N et de l'échantillon.

Proposition 8.5 *Un estimateur de variance minimum est nécessairement sans biais.*

Preuve: Supposons que l'estimateur A a un biais, $\mathbb{E}(A) - a$. Alors, nous montrons que l'estimateur $B = A - (\mathbb{E}(A) - a)$ a une variance strictement plus faible (remarquons toutefois que l'estimateur B, qui diffère de A par une constante, n'est *a priori* pas accessible).

On a

$$(B-a)^2 = ((A-a) - (\mathbb{E}(A) - a))^2.$$

Par conséquent, en prenant les espérances (et en remarquant que $\mathbb{E}(A) - a$ est une constante), il vient

$$\mathbb{E}((B-a)^2) = \mathbb{E}((A-a)^2) - (\mathbb{E}(A) - a)^2,$$

qui est strictement inférieur à $\mathbb{E}((A-a)^2)$ dès lors que A a effectivement un biais non nul. □

8.1.3 Maximum de vraisemblance

La méthode dite du "maximum de vraisemblance" est la méthode la plus courante pour construire des estimateurs.

On définit la *vraisemblance* d'une série statistique $x_1, \ldots, x_N$, par

$$V(x_1, \ldots, x_N) = \mathbb{P}(X_1 = x_1, \ldots, X_N = x_N) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_N = x_n)$$

si X suit une loi discrète, et

$$V(x_1, \ldots, x_N) = f(x_1) \ldots f(x_N)$$

si X est une loi diffuse de densité f.

Cette fonction de vraisemblance dépend aussi du paramètre a ; on cherche alors, en fonction de $x_1, \ldots, x_N$, la valeur de a qui rend maximale cette vraisemblance. Si on parvient à l'exprimer sous la forme $\varphi(x_1, \ldots, x_N)$, on choisit comme estimateur (appelé *estimateur du maximum de vraisemblance*),

$$A = \varphi(X_1, \ldots, X_N).$$

Un exemple d'estimateur du maximum de vraisemblance

Nous nous plaçons dans la situation où notre loi de X est une loi de Poisson, de paramètre a, et où nous cherchons précisément un estimateur pour a.

La loi de Poisson (de paramètre a) étant une loi discrète avec $\mathbb{P}(X = k) = e^{-a}a^k/k!$, on exprime la vraisemblance d'une série statistique sous la forme

$$\begin{aligned} V(x_1, \ldots, x_N) &= \prod_{i=1}^{N} \mathbb{P}(X_k = x_k) \\ &= \prod_{i=1}^{N} e^{-a} \frac{a^{x_k}}{x_k!} \\ &= e^{-Na} \frac{a^{\sum_k x_k}}{\prod_k x_k!} \end{aligned}$$

Pour déterminer le maximum de vraisemblance, on cherche pour quelle valeur de a (comme fonction des entiers x_k) cette fonction prend une valeur maximale. On dérive donc par rapport à la variable a :

$$\frac{\partial V}{\partial a} = \left(-N + \frac{\sum_k x_k}{a} \right) e^{-Na} a^{\sum_k x_k} \prod_k \frac{1}{x_k!},$$

et on découvre que cette expression ne s'annule que si l'on a

$$N = \frac{\sum_k x_k}{a}.$$

On vérifie qu'il s'agit bien d'un maximum (c'est forcément le cas, pour des raisons de convergence de la série des probabilités), et on obtient donc l'expression de l'estimateur du maximum de vraisemblance :

$$A = \frac{1}{N} \sum_{k=1}^{N} X_k.$$

On est tout de même un peu déçu, même si l'on n'est pas trop surpris : on cherchait un estimateur du paramètre a, dont on sait par ailleurs (pour les lois de Poisson) que c'est l'espérance de la loi ; et l'estimateur du maximum de vraisemblance que l'on obtient, n'est autre que l'estimateur standard de l'esperance.

8.1.4 Estimation de la variance

Cas où l'espérance est connue

Si l'espérance μ de la loi considérée est connue, il est clair que l'estimateur

$$V^* = \frac{1}{N} \sum_{k=1}^{N} (X_n - \mu)^2$$

a pour espérance, la variance de X ; en d'autres termes, il s'agit d'un estimateur sans biais pour la variance.

Cas où l'espérance est inconnue

Si l'espérance est inconnue, on n'a naturellement pas accès à V^*. On peut remplacer, dans la formule de V^*, l'espérance inconnue μ par son estimateur standard $\overline{X}$; on obtient alors l'estimateur standard de variance $C^*_{X,X}$.

Cependant, le calcul de l'espérance de $C^*_{X,X}$ montre que cet estimateur n'est pas sans biais : on obtient

$$\mathbb{E}(C^*_{X,X}) = \frac{N-1}{N}\mathbf{Var}(X).$$

Le ratio $\frac{N-1}{N}$ tend vers 1 lorsque la taille de l'échantillon tend vers l'infini : $C^*_{X,X}$ est un estimateur de variance *asymptotiquement sans biais*. On peut toutefois obtenir un estimateur sans biais, en prenant comme nouvel estimateur,

$$V^*_{n-1} = \frac{N}{N-1}C^*_{X,X} = \frac{1}{N-1}\sum_{k=1}^{N}\left(X_k - \overline{X}\right)^2.$$

(L'indice $n-1$ dans la notation V^*_{n-1} est uniquement une notation mnémotechnique permettant de distinguer cet estimateur de l'estimateur V^*.)

Cas des lois normales

Lorsque l'on sait que la loi de X est une loi normale, on a des résultats supplémentaires sur les estimateurs présentés :

Proposition 8.6 *Si X suit une loi normale d'espérance connue μ, l'estimateur de variance V^* est l'estimateur du maximum de vraisemblance ; il est convergent.*
Si X suit une loi normale d'espérance inconnue, l'estimateur de variance V^*_{N-1} est convergent.

Preuve: Commençons par vérifier que le maximum de vraisemblance donne bien V^*.
La densité d'une variable aléatoire de loi $\mathcal{N}(\mu,\sigma)$ est

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}};$$

la densité du N-uplet en $(x_1,\ldots,x_N)$ est donc

$$F(x_1,\ldots,x_N) = \frac{1}{\sigma^N(2\pi)^{N/2}}e^{-\frac{\sum_k(x_k-\mu)^2}{2\sigma^2}}.$$

Pour déterminer le maximum de vraisemblance, nous cherchons pour quelle valeur de σ cette fonction est maximale (à x_i fixés) ; dériver par rapport à σ donne (tous calculs faits)

$$\left(\frac{\sum_k(x_k-\mu)^2}{\sigma^2} - N\right)\frac{1}{\sigma}F(x_1,\ldots,x_N).$$

Cette expression est nulle pour $\sigma^2 = (\sum_k(x_k-\mu)^2)/N$, et l'étude rapide du signe permet de vérifier qu'il s'agit bien d'un maximum. Le maximum de vraisemblance pour σ^2 est donc bien atteint pour V^*.

Pour démontrer la convergence, dans les deux cas, on se base sur la proposition 8.4. Les deux estimateurs V^* (cas où μ est connu) et V^*_{N-1} (cas où μ est inconnu) étant tous deux sans biais, il suffit de vérifier que leur variance tend vers 0. □

8.1.5 Efficacité d'un estimateur

Il existe une borne inférieure, que nous admettrons, à la *variance* d'un estimateur sans biais A pour une loi de X admettant une densité f :

Proposition 8.7 (Inégalité de Fréchet, Cramer et Rao) *Pour tout estimateur sans biais A d'un paramètre a, on a nécessairement*

$$\mathbf{Var}(A) \geq B_{X,a} = \frac{1}{N}\mathbb{E}\left(\left(\frac{\partial}{\partial a}\log f(X)\right)^2\right).$$

Cette inégalité conduit à définir le *coefficient d'efficacité* d'un estimateur sans biais :

Définition 8.8 *On appelle coefficient d'efficacité d'un estimateur sans biais A,*

$$e(A) = \frac{B_{X,a}}{\mathbf{Var}(A)}.$$

Un estimateur est dit efficace *si son coefficient d'efficacité est* 1 *(le maximum possible).*

Il arrive fréquemment que les estimateurs construits selon le maximum de vraisemblance soient efficaces.

8.2 Lois du χ^2 et de Student en Statistiques

La loi du χ^2 intervient très fréquemment lorsque l'on considère des variables gaussiennes ; à la lumière du Théorème Central Limite, il n'est pas surprenant que ces lois gaussiennes soient omniprésentes en statistiques.

8.2.1 Loi du χ^2 et estimateur de variance

Considérons un échantillon $X_1, \ldots, X_n$, de taille n, de modèle X suivant la loi $\mathcal{N}(m, \sigma)$. On pose, conformément à la notation usuelle,

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

On a alors la propriété suivante :

Théorème 8.9 *Posons*

$$Z = \sqrt{n}(\overline{X} - m)$$

et

$$U = \frac{1}{\sigma^2}\sum_{k=1}^{n}(X_k - \overline{X})^2.$$

Alors U et Z sont indépendantes, U suit la loi du χ^2 à $n-1$ degrés de liberté, et Z suit la loi normale réduite $\mathcal{N}(0, \sigma)$.

Preuve: Commençons par vérifier que Z, tel qu'indiqué, suit la loi $\mathcal{N}(0,\sigma)$. Le vecteur $\mathbf{X} = (X_1,\ldots,X_n)$, étant composé de n variables gaussiennes indépendantes, est un vecteur gaussien. Z s'exprime comme une combinaison affine des X_i :

$$Z = -\sqrt{n}m + \sum_{k=1}^{n} \frac{1}{\sqrt{n}} X_k;$$

par conséquent, Z suit une loi normale, dont on peut déterminer espérance et variance par linéarité ; on vérifie aisément que l'espérance est nulle, et la variance σ^2.

Posons maintenant, pour $1 \leq j \leq n-1$,

$$Y_j = -jX_{j+1} + \sum_{k=1}^{j} X_k = -j(X_{j+1} - m) + \sum_{k=1}^{j} (X_k - m).$$

Chaque Y_j est une variable gaussienne, dont on détermine également, par linéarité, espérance et variance :

$$\begin{aligned} \mathbb{E}(Y_j) &= 0 \\ \mathbf{Var}(Y_j) &= (j^2+j)\sigma^2. \end{aligned}$$

Par conséquent, si l'on pose $Z_j = \frac{1}{\sqrt{j+j^2}} Y_j$ pour $j \geq 1$ et $Z_0 = Z$, les n variables aléatoires $Z_0, Z_1, \ldots Z_{n-1}$ ont toutes la même loi $\mathcal{N}(0,\sigma)$, et le vecteur $\mathbf{Z} = (Z_0,\ldots,Z_{n-1})$ est un vecteur gaussien. Il n'est toutefois pas composé de variables indépendantes ; il nous faut calculer les covariances pour déterminer la loi de $\mathbf{Z}$.

Il est facile, à partir des expressions pour les Y_j, de calculer $\mathbb{E}(Y_jY_k)$:

$$\mathbb{E}(Y_jY_k) = \mathbb{E}((-jX_{j+1} + \sum_{i=1}^{j} X_i)(-kX_{k+1} + \sum_{\ell=1}^{k} X_\ell));$$

en développant, et en utilisant le fait que les X_i sont indépendants, il vient (en supposant $j < k$) :

$$\begin{aligned} \mathbb{E}(Y_jY_k) &= \sum_{i=1}^{j} \mathbb{E}(X_i^2) - j\mathbb{E}(X_j)^2 \\ &= j\sigma^2 - j\sigma^2 = 0. \end{aligned}$$

En d'autres termes, les Y_j sont indépendants (ce sont les coordonnées d'un vecteur gaussien, et toutes les covariances sont nulles). Effectuons le même calcul pour $\mathbb{E}(ZY_j)$:

$$\begin{aligned} \mathbb{E}(ZY_j) &= \mathbb{E}\left(\frac{1}{n}(-m + \sum_{i=1}^{n} X_i)(-jX_{j+1} + \sum_{k=1}^{j} X_j)\right) \\ &= \frac{1}{n}\left(-m\mathbb{E}(Y_j) + \sum_{i=1}^{j} \mathbb{E}(X_i^2) - j\mathbb{E}(X_{j+1}^2)\right) \\ &= 0. \end{aligned}$$

Par conséquent, les variables $(Z, Y_1, \ldots, Y_{n-1})$ sont indépendantes ; il en est donc de même de $(Z_0, Z_1, \ldots, Z_{n-1})$. Par conséquent, si l'on pose $V = \frac{1}{\sigma^2}\sum_{k=1}^{n-1} Z_k^2$, *$V$ suit la loi du χ^2 à $n-1$ degrés de liberté, et est indépendants de Z_0*. Il nous suffit donc, pour terminer la preuve, de vérifier que l'on a $U = V$.

Pour cela, nous évaluons séparément $\sigma^2 U + Z_0^2$ et $\sigma^2 V + Z_0^2$, et montrons qu'ils sont tous deux égaux à $\sum_{k=1}^{n}(X_k - m)^2$.

En effet, nous avons obtenu chaque Z_i comme combinaison linéaire des $(X_i - m)$, donc les vecteurs $\mathbf{Z} = (Z_0, \ldots, Z_{n-1})$ et $\mathbf{X}' = (X_1 - m, \ldots X_n - m)$ sont liés par une relation de la forme $\mathbf{Z} = Q\mathbf{X}'$, où Q est une matrice carrée dont les calculs de covariance et de variance ont montré qu'elle est *orthogonale*. Par conséquent, les vecteurs $\mathbf{Z}$ et $\mathbf{X}'$ ont même norme euclidienne.

Or, le carré de la norme de $\mathbf{Z}$ est

$$Z_0^2 + \sum_{k=1}^{n-1} Z_k^2 = Z_0^2 + \sigma^2 V;$$

et le carré de la norme de $\mathbf{X}'$ s'écrit

$$\begin{aligned}\sum_{k=1}^{n}(X_k - m)^2 &= \sum_{k=1}^{n}((X_k - \overline{X}) + (\overline{X} - m))^2 \\ &= \sum_{k]1}^{n}(X_k - \overline{X})^2 + n(\overline{X} - m)^2 + 2(\overline{X} - m)(-n\overline{X} + \sum_{k=1}^{n} X_k).\end{aligned}$$

Dans cette dernière expression, le premier terme est $\sigma^2 U$, le deuxième est Z_0^2, et le troisième est nul en raison de la définition de $\overline{X}$.

Par conséquent, nous avons

$$\sigma^2 U + Z_0^2 = ||\mathbf{X}'||^2 = ||\mathbf{Z}||^2 = \sigma^2 V + Z_0^2,$$

d'où $V = U$ comme annoncé. □

L'énoncé du théorème peut être réécrit en termes des estimateurs classiques de variance :

Théorème 8.10 *Soit $(X_1, \ldots, X_n)$ un échantillon de taille n de la loi $\mathcal{N}(m, \sigma)$.*
Alors $(n-1)V_{n-1}^*/\sigma^2$ suit la loi du χ^2 à $n-1$ degrés de liberté.

8.2.2 Loi de Student

Des calculs similaires à ceux effectués au paragraphe précédent, permettent d'obtenir le résultat suivant :

Proposition 8.11 *Soit $(X_1, \ldots, X_n)$ un échantillon de taille n de la loi normale $\mathcal{N}(m, \sigma)$. Posons*

$$\sigma_{n-1}^* = \sqrt{V_{n-1}^*}$$

(σ_{N-1}^* est un estimateur raisonnable de l'écart-type σ, basé sur l'estimateur sans biais de variance V_{n-1}^*.)

Alors la variable aléatoire

$$S = \sqrt{n}\frac{\overline{X} - m}{\sigma_{n-1}^*}$$

suit la loi de Student à $n-1$ degrés de liberté.

8.3 Estimation par intervalle de confiance

8.3.1 Principe d'un intervalle de confiance

Jusqu'à présent, nous nous sommes intéressés à l'*estimation ponctuelle*, dans laquelle, à partir d'un échantillon $X_1, \ldots, X_N$ d'une loi X, un estimateur fournit une valeur qui sert d'*estimation* de la valeur "exacte" d'un paramètre a de la loi de X.

Dans l'*estimation par intervalle de confiance*, le but est de fournir non pas une valeur unique, mais un *intervalle* (donc deux valeurs) dans lequel il est probable que la valeur a recherchée se trouve. Cela fournit donc une mesure de la *précision* d'un estimateur qui prend sa valeur dans l'intervalle en question.

Étant donné un échantillon $X_1, \ldots, X_N$, on cherche donc deux valeurs

$$\begin{aligned} \alpha &= \alpha(X_1, \ldots, X_N) \\ \beta &= \beta(X_1, \ldots, X_N) \end{aligned}$$

telles que

$$\mathbb{P}(a \in [\alpha(X_1, \ldots, X_N), \beta(X_1, \ldots, X_N)]) \geq 1 - \epsilon \tag{8.1}$$

Ici, $1 - \epsilon$ est appelé *niveau de confiance* ; des niveaux de confiance de 90%, 95% ou 99% sont fréquemment utilisés (le plus souvent, 95%, ce qui correspond à $\epsilon = 0.05$).

Remarque 8.12 *Dans l'équation (8.1), il est important de bien remarquer que a n'est pas aléatoire : c'est un paramètre qui dépend de la* loi *de X, qui est quelque chose d'inconnu mais déterministe. Ce sont les bornes de l'intervalle, $\alpha(X_1, \ldots, X_N)$ et $\beta(X_1, \ldots, X_N)$, qui sont aléatoires. L'équation pourrait être réécrite, de manière plus conforme à l'usage,*

$$\mathbb{P}(\alpha(X_1, \ldots, X_N) \leq a, \beta(X_1, \ldots, X_N) \geq a) \geq 1 - \epsilon.$$

En particulier, une fois que l'on a déterminé que deux formules pour α et β déterminent un intervalle de confiance à 95%, et si, par exemple, l'évaluation de α et β sur un échantillon donne $[3.13, 3.15]$, il est stricto sensu **incorrect** *de dire "il y a au moins 19 chances sur 20 pour que a se trouve entre 3.13 et 3.15". En effet, la valeur exacte de a (peut-être $a = \pi$) se trouve, ou ne se trouve pas, dans l'intervalle, sans qu'il y ait quoi que ce soit d'aléatoire là-dedans. Une formulation plus correcte serait, "nous venons d'observer un intervalle aléatoire, qui a au moins 19 chances sur 20 de contenir a ; cet intervalle est ici $[3.13, 3.15]$".*

Naturellement, lorsque l'on cherche une estimation par intervalle de confiance, on cherche à trouver un intervalle le plus petit possible (typiquement, de faible longueur $\beta - \alpha$), tout en maintenant le niveau de confiance requis.

8.3.2 Cas standard : intervalle centré sur un estimateur

Lorsque l'on dispose déjà d'un estimateur $A = A(X_1, \ldots, X_N)$, on cherche typiquement un intervalle *centré sur* A, c'est-à-dire un intervalle de la forme $[A - r, A + r]$, où r peut être exprimé en fonction de l'échantillon :

$$r = r(X_1, \ldots, X_N).$$

Il s'agit alors de trouver un r, le plus petit possible, tel que l'on ait

$$\mathbb{P}(|A(X_1, \ldots, X_N) - a| > r) \leq \epsilon.$$

Exemple 8.13 (Estimation d'une espérance, à variance connue) *Plaçons-nous dans le cas, il est vrai peu fréquent, où l'on cherche à estimer l'espérance $a = \mathbb{E}(X)$ de la loi, et où l'on connaît sa* variance σ^2 *– ou, de manière peut-être plus vraisemblable, dans le cas où l'on connaît une* majoration σ^2 *de la variance (c'est-à-dire que l'on sait que l'on a $\mathbf{Var}(X) \leq \sigma^2$).*

Dans ce cas, pour $S = X_1 + \cdots + X_N$, la variance de S est (au plus) $N\sigma^2$, et l'inégalité de Tchebycheff, nous permet d'affirmer que

$$\mathbb{P}(|S - Na| \geq \lambda\sigma\sqrt{N}) \leq \frac{1}{\lambda}.$$

En réexprimant cette inégalité en termes de l'estimateur standard d'espérance $\overline{X} = S/N$, on obtient

$$\mathbb{P}(|\overline{X} - a| \geq \lambda\sigma/\sqrt{N}) \leq \frac{1}{\lambda}.$$

En posant $\lambda = 20$, on obtient donc un intervalle de confiance à 95% de la forme

$$\left[\overline{X} - \frac{20\sigma}{\sqrt{N}}, \overline{X} + \frac{20\sigma}{\sqrt{N}}\right].$$

Exemple 8.14 (Estimation d'une espérance dans le cas gaussien) *Plaçons-nous maintenant dans le cas où la loi de X est la loi gaussienne $\mathcal{N}(m, \sigma)$, et où nous cherchons à estimer m. Notre estimateur naturel est $\overline{X}$, et nous chercherons un intervalle de confiance de la forme*

$$[\overline{X} - r\sigma, \overline{X} + r\sigma]$$

dans le cas où σ est connu (ce qui est le cas de l'exemple précédent, à ceci près que nous supposons ici que la loi est gaussienne), ou

$$[\overline{X} - r\sigma^*_{N-1}, \overline{X} + r\sigma^*_{N-1}]$$

dans le cas où, σ étant inconnu, nous le remplaçons par un (bon) estimateur.

La condition sur le seuil de confiance (qui va nous permettre de choisir r) est donc

$$\mathbb{P}(|\overline{X} - m| > r\sigma^*_{N-1}) \leq \epsilon,$$

ce qui peut être réécrit en

$$\mathbb{P}\left(\left|\frac{\sqrt{N}(\overline{X} - m)}{\sigma^*_{N-1}}\right| > \sqrt{N}r\right) \leq \epsilon.$$

Or nous connaissons la loi de $\sqrt{N}(\overline{X} - m)/\sigma^*_{N-1}$: c'est la loi de Student à $N - 1$ degrés de liberté *; ce qui nous permet, en fonction de la taille de l'échantillon, et en consultant la table appropriée pour la loi de Student, de trouver la valeur appropriée de r.*

8.3.3 Intervalle de confiance pour l'écart-type

Lorsque l'on cherche à estimer un écart-type, c'est le plus souvent que l'on cherche à *majorer* cet écart-type. À partir d'un estimateur Σ, on cherche donc plutôt un intervalle de confiance de la forme $[0, r\Sigma]$ (avec $r > 1$), c'est-à-dire une inégalité de la forme

$$\mathbb{P}(r\Sigma < \sigma) \leq \epsilon. \tag{8.2}$$

Supposons maintenant que la loi de X soit, une fois de plus, la loi gaussienne $\mathcal{N}(m, \sigma)$ (où σ au moins est inconnu). Typiquement, l'estimateur Σ sera σ^* (dans le cas où m est connu) ou σ^*_{N-1} (dans le cas, plus fréquent, où m est inconnu).

Dans le premier cas, (8.2) peut être réécrite en

$$\mathbb{P}\left(N\left(\frac{\sigma^*}{\sigma}\right)^2 < \frac{N}{r^2}\right) \leq \epsilon;$$

dans le second cas, en

$$\mathbb{P}\left((N-1)\left(\frac{\sigma^*_{N-1}}{\sigma}\right)^2 < \frac{N-1}{r^2}\right) \leq \epsilon.$$

Selon le cas, la variable aléatoire majorée est la loi du χ^2 à N ou à $N-1$ degrés de libertés ; la valeur minimale de r peut donc être déduite d'une table de la fonction de répartition (inverse) de la loi du χ^2.

Chapitre 9

Tests d'hypothèses

Sommaire

9.1 Principes d'un test d'hypothèse

Lors d'une *estimation*, nous cherchons à proposer une *valeur* pour un paramètre, en tentant d'assurer que la valeur proposée soit, de manière probable, proche de la valeur réelle du paramètre.

Lors d'un *test d'hypothèse*, l'objectif est de proposer un *verdict* à une question qui nous est posée.

Idéalement, *tester une hypothèse H contre une hypothèse K* (où H et K sont deux hypothèses incompatibles, dont on suppose qu'exactement une est vraie), consiste à répondre à la question "est-il vrai que H, plutôt que K ?".

Fréquemment, l'hypothèse K n'est pas spécifiée, ce qui signifie implicitement que c'est l'hypothèse contraire à H ("est-il vrai que H, plutôt que non-H ?").

Dans la pratique, tout ce dont on dispose pour émettre un verdict est un échantillon $X_1, \ldots, X_N$ de loi X (l'hypothèse H porte sur la loi de X). Il va donc nous falloir déterminer, dans l'ensemble des réalisations possibles de l'échantillon, une région R de *rejet* ; on rejettera l'hypothèse ("les données poussent à penser que l'hypothèse n'est pas vérifiée") si $\mathbf{X} \in R$ (l'échantillon tombe dans la région de rejet), et on l'acceptera ("les données ne sont pas suffisantes pour rejeter l'hypothèse") dans le cas contraire.

Il y a donc potentiellement deux "risques", qu'un test doit tenter de minimiser :

- le risque de rejeter H, alors que H est vraie (*risque de faux négatifs*, aussi appelé *risque de première espèce*) ;
- le risque d'accepter H, alors que H est fausse (*risque de faux positifs*, ou *risque de deuxième espèce*.

Traditionnellement, on tend à préférer, entre deux tests, celui qui a le plus petit risque de première espèce, le risque de deuxième espèce servant surtout à départager d'éventuels *ex æquo*. Cela revient à dire que, lors d'un test, on préfère accepter à tort une hypothèse fausse, plutôt que de rejeter à tort une hypothèse vraie ; on aura donc tendance à être parfois exagérément prudent avant de rejeter une hypothèse. Cette dissymétrie est importante ; il est possible que, avec le même échantillon, un test accepte l'hypothèse H contre l'hypothèse non-H, alors que le même test accepte l'hypothèse non-H contre l'hypothèse H.

Remarque 9.1 *Il arrive que le risque de première espèce soit noté en employant une notation de probabilité conditionnelle :*

$$\alpha = \mathbb{P}(\mathbf{X} \in R|H).$$

Cette notation est impropre, car elle porte à penser que l'hypothèse H correspond à un événement probabiliste. Or, il n'en est a priori rien : dans le modèle standard, l'hypothèse est, ou n'est pas, vérifiée, mais ne résulte pas d'une expérience aléatoire.

9.2 Test de Bayes

Si l'on se place dans un cadre dit *bayésien*, on suppose, contrairement à ce qui a été dit dans la remarque 9.1, que le modèle probabiliste de X (sa loi de probabilités) provient lui-même d'une expérience aléatoire dont on connaît *a priori* la loi.

Par exemple, si on s'intéresse à la proportion de pièces défectueuses dans la production d'une machine, X suivra une loi de Bernoulli de paramètre p (chaque pièce est défectueuse ($X = 1$) avec probabilité p, et correcte ($X = 0$) avec probabilité $1 - p$). Dans un modèle bayésien, p est elle-même une grandeur aléatoire, dont on spécifie la loi de probabilités.

Dans ce cadre, l'hypothèse H (qui est une affirmation qui porte sur la loi de X) devient elle-même probabiliste, et on peut donc parler de *la probabilité que H soit satisfaite* $\mathbb{P}(H)$, et de *la probabilité de rejeter H, sachant qu'elle est satisfaite* $\mathbb{P}(\mathbf{X} \in R|H)$.

Dans ce cas, si l'on est en mesure d'exprimer un "coût" aux deux types de risques :

– un coût C_H, associé au risque de première espèce (coût d'un rejet à tort) ;

– un coût C_K, associé au risque de deuxième espèce (coût d'une acceptation à tort) ;

alors on peut tenter de minimiser le *coût moyen du risque d'erreur*,

$$C = C_H \mathbb{P}(\mathbf{X} \in R|H) + C_K \mathbb{P}(\mathbf{X} \notin R|K).$$

(Ici, la *minimisation* se fait sur le choix de la *région de rejet*, qui définit le test.)

9.3 Test de Neymann-Pearson

Dans un test d'hypothèse classique, on n'a pas de modèle bayésien. On se donne un niveau acceptable α de risque de première espèce (probabilité de rejeter l'hypothèse alors qu'elle est vraie), typiquement $\alpha = 0.05$ (soit 5%), et on est donc amené à cherche une *région de rejet* R qui garantit que, pour un échantillon $\mathbf{X}$, on ait

$$\mathbb{P}(\mathbf{X} \in R) \leq \alpha$$

dès lors que la loi de X satisfait à l'hypothèse H.

Avant de considérer le test (la région de rejet) comme acceptable, on cherche également à vérifier le risque de deuxième espèce,

$$\mathbb{P}(\mathbf{X} \notin R),$$

est suffisamment petit dès lors que la loi de X ne satisfait pas à l'hypothèse H (ou, dès lors qu'elle satisfait à l'hypothèse K, lorsque K n'est pas exactement le contraire de H).

Les deux conditions sont, d'une certaine manière, des objectifs contradictoires : on peut *a priori* rendre α aussi petit que souhaité, en prenant R arbitrairement petite (avec $R = \emptyset$, on ne rejettera jamais à tort une hypothèse vraie, puisqu'on ne la rejettera jamais ; en contrepartie, on acceptera toujours une hypothèse fausse : le risque de deuxième espèce est de 1) ; et inversement, tenter de réduire le risque de deuxième espèce pousse à choisir une région R grande.

On distingue généralement entre les "hypothèses simple" et les "hypothèses composées" :

- Si l'hypothèse H détermine entièrement la loi de X, on parle d'*hypothèse simple*. Dans ce cas, H détermine entièrement le risque de première espèce $\mathbb{P}_H(\mathbf{X} \in R)$. Par exemple, si X est censé suivre une loi gaussienne $\mathcal{N}(m, 1)$, où m n'est pas connu, l'hypothèse "$m = 0$" est une hypothèse simple.
- Si, au contraire, l'hypothèse H ne détermine pas complètement la loi de X, on parle d'*hypothèse composée*. Le risque de rejeter à tort l'hypothèse n'est donc plus simplement défini (la probabilité de rejet à tort prend potentiellement une valeur différente pour chaque loi possible pour X, le risque de première espèce étant alors le maximum, ou le supremum, de ces valeurs). Dans le cas du modèle gaussien $\mathcal{N}(m, 1)$ pour X, des hypothèses comme "$m < 2$", "$m \in [-1, 1]$", ou même "$m \neq 0$" sont des hypothèses composée.

9.3.1 Test de comparaison des espérances

On se place dans la situation suivante : on dispose d'échantillons de deux caractères X et Y, d'espérances respectives m_X et m_Y, et on cherche à tester l'hypothèse $m_X = m_Y$.

Les échantillons $X_1, \ldots, X_N$ et $Y_1, \ldots, Y_M$ ne sont pas forcément de même taille, mais ils sont supposés indépendants.

Cas de grands échantillons

Dans le cas où les deux échantillons sont suffisamment grands (disons, M et N tous deux au moins de l'ordre de 30), on peut adopter l'approximation gaussienne suivante :

La loi de

$$\frac{(\overline{Y} - \overline{X}) - (m_Y - m_X)}{\sqrt{\frac{C^*_{X,X}}{N} + \frac{C^*_{Y,Y}}{M}}}$$

est approximativement la loi normale réduite $\mathcal{N}(0, 1)$.

Pour tester l'hypothèse "$m_X = m_Y$", on peut donc se permettre utiliser une condition de rejet de la forme

$$\frac{|\overline{Y} - \overline{X}|}{\sqrt{\frac{C^*_{X,X}}{N} + \frac{C^*_{Y,Y}}{M}}} > r,$$

où r est obtenu en consultant une table de la fonction de répartition de la loi normale réduite (par exemple, $r = 1.96$ pour $\alpha = 0.05$).

Cas de petits échantillons (loi gaussienne)

Si les échantillons sont de petite taille, le test à appliquer sera fortement dépendant de la nature des lois possibles pour X. Nous traitons ici le cas où X *et* Y *suivent toutes deux des lois gaussiennes, de même variance* σ^2. Une telle supposition devrait, dans un cas pratique, être justifiée.

Dans ce cas, on peut démontrer la propriété suivante :

Proposition 9.2 *La variable aléatoire*

$$Z = \sqrt{\frac{MN(M+N-2)}{M+N}} \frac{\overline{Y} - \overline{X} - (m_Y - m_X)}{\sqrt{NC^*_{X,X} + MC^*_{Y,Y}}}$$

suit la loi de Student à $M+N-2$ *degrés de liberté.*

Preuve: D'après le théorème 8.9, $\frac{1}{\sigma^2} NC^*_{X,X}$ suit la loi du χ^2 à $N-1$ degrés de libertés, $\frac{1}{\sigma^2} MC^*_{Y,Y}$ la loi du χ^2 à $M-1$ degrés de libertés, et ces deux quantités sont respectivement indépendantes de $\overline{X}$ et de $\overline{Y}$. De plus, par indépendance des échantillons $\mathbf{X}$ et $\mathbf{Y}$, toute fonction des X_i est indépendante de toute fonction des Y_j. Par conséquent, la quantité sous la racine dans le dénominateur de Z, une fois divisée par σ^2 :

$$Z_1 = \frac{1}{\sigma^2}(NC^*_{X,X} + MC^*_{Y,Y}),$$

comme somme de variables indépendantes suivant chacune une loi du χ^2, suit la loi du χ^2 à $M+N-2$ degrés de libertés, et est indépendante du numérateur.

Or, si l'on pose

$$Z_0 = \sqrt{\frac{MN}{M+N}}(\overline{Y} - \overline{X} - (m_Y - m_X)),$$

on voit que Z_0 suit une loi gaussienne, d'espérance nulle, et dont on calcule aisément la variance : $\overline{X}$ a variance σ^2/N, $\overline{Y}$ a variance σ^2/M, donc $\overline{Y} - \overline{X}$ a variance

$$\sigma^2 \left(\frac{1}{M} + \frac{1}{N} \right) = \sigma^2 \frac{M+N}{MN}.$$

Par conséquent, Z_0 a pour variance σ^2.

On a donc exprimé Z sous la forme

$$Z = \frac{Z_0}{\sqrt{(M+N-2)Z_1}},$$

où Z_0 et Z_1 sont indépendantes, Z_0 est de loi normale réduite $\mathcal{N}(0,1)$ et Z_1 est de loi du χ^2 à $M+N-2$ degrés de liberté ; ceci suffit pour montrer que Z suit la loi de Student à $M+N-2$ degrés de liberté. □

Maintenant, pour tester l'hypothèse H : "$m_X = m_Y$", nous écrivons la condition de rejet sous la forme

$$\sqrt{\frac{MN(M+N-2)}{M+N}} \frac{|\overline{Y} - \overline{X}|}{\sqrt{NC^*_{X,X} + MC^*_{Y,Y}}} > r,$$

et nous cherchons r dans la table de la fonction de répartition de la valeur absolue de la loi de Student à $M+N-2$ degrés de liberté ; par exemple, pour $M = 10$ et $N = 15$, on obtient $r = 2.069$ pour $\alpha = 0.05$.

9.3.2 Test d'ajustement du χ^2

Ce test, ou "test tu χ^2", est très souvent pratiqué, à tel point que, bien souvent, il est appliqué comme une "recette magique" par des gens qui ne connaissent rien d'autre aux statistiques.

On considère un échantillon $\mathbf{X} = (X_1, \ldots, X_N)$ de modèle X, et une loi de probabilités $\mathcal{P}$.

L'hypothèse H que l'on teste est "le caractère X suit la loi $\mathcal{P}$" ; on parle ici de *test non paramétrique*, car on ne s'intéresse pas seulement à un paramètre de la loi de X, mais à l'ensemble de sa loi.

Le principe est, comme dans la représentation d'un histogramme, de partitionner l'espace des valeurs du caractère en q intervalles, ou "classes" $A_1, \ldots, A_q$, et de considérer, pour chaque $1 \leq i \leq q$, le nombre de valeurs de l'échantillon qui sont dans la classe A_i, soit la variable aléatoire

$$N_i = \#\{j : 1 \leq j \leq N, x_j \in A_i\}.$$

Chaque N_i suit la loi binomiale $\mathcal{B}(N, p_i)$, avec $p_i = \mathbb{P}(X \in A_i)$ (les N_i ne sont toutefois pas indépendantes : on a forcément la relation $\sum_i N_i = N$).

Pour peu que N soit assez grand (typiquement, de l'ordre de 30 au moins, en tenant compte des remarques qui suivent), on peut appliquer le principe d'approximation suivant :

Sous l'hypothèse que X suit bien la loi $\mathcal{P}$, la variable aléatoire

$$U = \frac{1}{N} \sum_{i=1}^{q} \frac{(N_i - Np_i)^2}{p_i}$$

suit approximativement la loi du χ^2 à $q-1$ degrés de liberté.

Justification heuristique de l'approximation : Chaque N_i peut s'écrire sous la forme

$$N_i = \sum_{k=1}^{N} X_{i,k},$$

où $X_{i,k}$ est la variable indicatrice de l'événement "la k-ème mesure tombe dans A_i" ; les $X_{i,k}$ sont des variables de Bernoulli, de paramètre p_i, mais si $(X_{i,k})_{1 \leq k \leq N}$ sont indépendantes, $(X_{i,k})_{1 \leq i \leq q}$ ne le sont pas, puique l'on a la relation $\sum_i X_{i,k} = 1$.

Toutefois, on peut tout de même en déduire, par le Théorème Central Limite, que le vecteur $\mathbf{T} = (T_1, \ldots, T_q)$, avec

$$T_i = \frac{(\sum_{k=1}^{N} X_{i,k}) - Np_i}{\sqrt{Np_i}} = -\sqrt{Np_i} + \sum_{k=1}^{N} \frac{X_{i,k}}{\sqrt{Np_i}},$$

est approximativement gaussien, et centré (l'espérance de chaque composante est nulle, ce qui simplifie les calculs).

Pour mieux comprendre la loi de $\mathbf{T}$, nous examinons la matrice des covariances ; il nous faut donc calculer $\mathbb{E}(T_i T_j)$. En tenant compte des faits suivants :

- $X_{i,k} X_{j,k} = 0$ si $i \neq j$ (la k-ème mesure ne peut tomber à la fois dans A_i et dans A_j) ;
- $X_{i,k}^2 = X_{i,k}$ (ce sont des variables de Bernoulli) ;
- $X_{i,k}$ et $X_{j,\ell}$ sont indépendantes si $k \neq \ell$,

on arrive après un calcul peu compliqué (mais qui constitue un exercice intéressant du point de vue de la manipulation des sommes), à

$$\begin{aligned} \mathbf{Cov}(T_i, T_j) &= -\sqrt{p_i p_j}\ (i \neq j) \\ \mathbf{Var}(T_i) &= 1 - p_i \end{aligned}$$

Notons C la matrice de ces covariances ($C_{i,j} = \mathbb{E}(T_i T_j) = \mathbf{Cov}(T_i, T_j)$). Nous savons, parce que notre vecteur $\mathbf{T}$ est (approximativement) gaussien, et centré, que cette matrice des covariances détermine sa loi, et donc la loi du carré de sa norme, dont nous voulons précisément prouver qu'il s'agit d'une loi du χ^2, avec un degré de liberté de moins que la dimension. Pour cela, nous avons en fait besoin d'identifier la structure spectrale (valeurs propres et multiplicités) de cette matrice symétrique. (Dans la suite de la preuve, il faut considérer que le vecteur $\mathbf{T}$ a été remplacé par un "vrai" vecteur gaussien centré dont la matrice de covariances est C.)

Comme matrice réelle symétrique, C est diagonalisable dans une base orthonormale. Nous n'allons pas exactement exhiber une telle base, mais seulement donner assez d'informations pour identifier la structure spectrale.

Notons Q la matrice carrée, de dimension q, dont les coefficients sont donnés par

$$Q_{i,j} = \begin{cases} \sqrt{p_i} & (j = q) \\ -\sqrt{p_{j+1}} & (i = j < q) \\ \sqrt{p_j} & (j = i + 1 < q) \\ 0 & \text{sinon} \end{cases}$$

Ainsi, chaque vecteur colonne c_j de Q, sauf le dernier, a exactement deux coefficients non nuls : le coefficient diagonal et le suivant.

Il n'est pas difficile de vérifier les identités suivantes :

$$\begin{aligned} C.c_q &= 0 \\ C.c_j &= c_j\ (j < q) \\ (c_j, c_q) &= 0 \end{aligned}$$

En d'autres termes, les q vecteurs colonnes de la matrice Q sont tous des vecteurs propres : les $q-1$ premiers pour la valeur propre 1, et le dernier pour la valeur propre 0 ; et les $q-1$ premiers vecteurs propres sont tous orthogonaux au dernier (ce qui est en fait automatique, les sous-espaces propres d'une matrice symétrique étant toujours orthogonaux). En revanche, les vecteurs c_j ($j < q$) ne sont pas orthogonaux entre eux : même en divisant les colonnes par leurs normes, on n'obtiendra pas une matrice orthogonale.

Toutefois, il est également facile de vérifier que les $q-1$ vecteurs $(c_j)_{1 \leq j < q}$ sont *libres*, et forment donc une base d'un sous-espace de dimension $q-1$, qui est donc *le* sous-espace propre pour la valeur propre 1. On peut donc (sans calcul) remplacer ces $q-1$ vecteurs par une base orthonormée de ce sous-espace, et obtenir une base orthonormée de vecteurs propres de C (le vecteur c_q est déjà de norme 1), et donc, en les utilisant comme vecteurs colonne d'une matrice Q', on obtient maintenant une matrice *orthogonale* Q' qui diagonalise C :

$${}^t Q' C Q' = D,$$

où D est la matrice diagonale dont tous les coefficients diagonaux sont égaux à 1, sauf le dernier qui est nul.

D'après les résultats du paragraphe 3.3.4, le vecteur $\mathbf{Z} = Q'\mathbf{T}$ est également gaussien, centré, avec comme matrice de covariances la matrice D. Cette matrice étant diagonale, les composantes de $\mathbf{Z}$ sont indépendantes ; les $q-1$ ont comme variance 1, et sont donc normales réduites ; la dernière a une variance nulle, et est donc constante (et nulle, car centrée). Au total, le carré de la norme de $\mathbf{Z}$ est la somme des carrés de $q-1$ variables normales réduites indépendantes : sa loi est donc celle du χ^2 à $q-1$ degrés de liberté.

La matrice Q' étant orthogonale, la norme de $\mathbf{Z}$ est égale à celle de $\mathbf{T}$, ce qui prouve que le carré de la norme de $\mathbf{T}$ suit bien la loi du χ^2 à $q-1$ degrés de liberté. □

Pour pratiquer un test d'ajustement du χ^2, on procède donc comme suit : on choisit q et les classes A_i, et on calcule les probabilités p_i (qui sont données par les classes et la loi P que l'on veut tester). On forme donc le vecteur $\mathbf{T}$, et on rejette l'hypothèse si l'on a

$$U = \sum_{k=1}^{q} T_k^2 > r,$$

où r est obtenu dans la table du χ^2 en fonction du nombre de degrés de liberté et du seuil α choisi ; ainsi, pour 10 classes (9 degrés de liberté) et $\alpha = 0.05$, on aura $r = 16.92$.

Remarque 9.3 *Les tables du χ^2 contiennent également des colonnes pour des valeurs qui sont* très probablement *dépassées par le χ^2 (valeurs qui ont, par exemples, une probabilité* 0.95 *d'être inférieures à une variable suivant la loi du χ^2 ; pour* 9 *degrés de liberté, la valeur est de* 3.33*). Elles sont censées être utilisées de la manière suivante : si la valeur de U est* inférieure *à ce seuil, on considère typiquement que l'échantillon est* trop régulier*, ce qui indique généralement que les valeurs ne sont pas indépendantes ; on remet donc en cause, plutôt que l'hypothèse H, le processus de collecte des données.*

Remarque 9.4 (Choix de q et des classes) *Pour que le test d'ajustement soit valide, il faut prendre quelques précautions au niveau de la valeur de q et des classes A_i ; en effet, tout repose sur l'approximation gaussienne.*

On considère généralement valides les règles heuristiques suivantes :

- *La taille totale de l'échantillon doit être assez grande (au moins $N \geq 30$) ;*
- *Le nombre de classes, et les classes elles-mêmes, doivent être choisies de telle sorte que les effectifs espérés des classes, Np_i, soient tous, sauf au plus un, de l'ordre de* 5 *au moins. Cette contrainte limitera forcément le nombre de classes q en fonction de la taille N de l'échantillon.*
- *Si la loi P fait partie d'une famille de lois dépendant de k paramètres ($k = 2$ pour les gaussiennes ; $k = 1$ pour les lois exponentielles démarrant à 0, ...), il convient que q soit plusieurs fois supérieur à k ($q \geq 4k$, typiquement) pour que l'on puisse considérer le test comme discriminant.*

Dans tous les cas, le test d'ajustement du χ^2, puisqu'il ne fait pas la différence entre deux valeurs distinctes qui tombent dans la même classe, ne fera jamais la différence (à ensemble de classes fixé) entre deux lois qui donnent les mêmes probabilités p_i.

9.3.3 Test des longueurs

Le *test des longueurs* est un deuxième exemple de test non paramétrique. On considère *deux* séries statistiques, indépendantes, $x_1, \ldots, x_N$ et $y_1 \ldots, y_M$, correspondant à deux caractères X et Y, et on cherche à tester l'hypothèse H : "les caractères X et Y ont la même loi".

Un cas d'application d'un tel test serait le suivant : on désire savoir si l'application d'un processus particulier (un traitement médical, par exemple) a une influence sur un caractère (par exemple, le poids). On peut alors procéder ainsi : on définit deux populations ; l'une est soumise au processus, et on en tire une série statistique $x_1, \ldots, x_N$; l'autre n'y est pas soumise, et on en tire une seconde série statistique $y_1, \ldots, y_M$. L'hypothèse "le traitement n'a pas d'influence" est alors équivalente à l'hypothèse "les caractères X et Y suivent la même loi".

Le principe du test des longueurs est le suivant : les deux séries statistiques sont mises en commun, et les valeurs sont triées par ordre croissant ; on code le résultat par un mot de $M+N$ lettres sur l'alphabet $\{x, y\}$, dans lequel la k-ème lettre indique si la k-ème valeur provenait de l'échantillon $\mathbf{X}$ ou de l'échantillon $\mathbf{Y}$.

L'hypothèse "les caractères ont la même loi", est *équivalente* au fait que le mot obtenu suit la loi uniforme parmi tous les mots de $M+N$ lettres comportant N fois la lettre x et M fois la lettre y.

L'idée est alors de découper le mot en suites de x ou de y consécutifs (ces suites sont appelées "longueurs", ou "runs" en Anglais), et de compter le nombre L de "longueurs" obtenues. Si les lois de X et de Y ne sont pas les mêmes, on s'attend à ce que les x ou les y aient tendance à s'accumuler plus en certaines positions, ce qui aura tendance à faire diminuer L. On va donc chercher une condition de rejet de la forme $L < r$.

Pour déterminer r, il nous faut déterminer la *loi* de L, sous l'hypothèse H.

Sous cette hypothèse, le mot formé est un mot uniforme parmi les $\binom{M+N}{N}$ mots possibles. Par conséquent, si l'on note $C_{M,N,k}$ le nombre de mots comportant M fois la lettre x, N fois la lettre y, et ayant k longueurs, on a (sous hypothèse H),

$$\mathbb{P}(L = k) = \frac{C_{M,N,k}}{\binom{M+N}{N}}.$$

Proposition 9.5 *Si $k \leq 2\min(M,N)$, on a*

$$C_{M,N,k} = \binom{N-1}{k_1}\binom{M-1}{k_2} + \binom{N-1}{k_2}\binom{M-1}{k_2},$$

où

$$k_1 = \lceil \frac{k-2}{2} \rceil \text{ et } k_2 = \lfloor \frac{k-2}{2} \rfloor.$$

Pour $k = 2\min(M,N)+1$, on a, si par exemple $M < N$

$$C_{M,N,2M+1} = \binom{N-1}{M-1}.$$

Enfin, dans le cas particulier $M = N$ et $k = 2M$, on a $C_{N,N,2N} = 2$.

Preuve: La preuve que nous donnons est assez combinatoire.

Si un mot doit contenir k longueurs, et qu'il commence par exemple par la lettre x, alors il doit contenir soit $k/2$ longueurs de x et $k/2$ longueurs de y (dans le cas où k est pair), soit $(k+1)/2$ longueurs de x et $(k-1)/2$ longueurs de y (dans le cas où k est impair). Dans tous les cas, le nombre de longueurs de l'un des lettres est $k_1 + 1$, et le nombre de longueurs de l'autre lettre est $k_2 + 1$.

La remarque importante est la suivante : une fois choisie la première lettre du mot (x ou y), on peut *coder* de manière bijective en spécifiant, pour chaque lettre, quelles sont les positions, parmi les $N-1$ (pour x) ou $N-1$ (pour y) occurrences de cette lettre autre que la première, lesquelles sont des débuts de longueurs (la première occurrence d'une lettre est toujours le début d'une longueur).

Or, le nombre de façons de choisir, par exemple, k_1 positions parmi $N-1$, est $\binom{N-1}{k_1}$. Ainsi, le nombre de mots commençant par x, et ayant k_1+1 longueurs de x et k_2+1 longueurs de y, est

$$\binom{N-1}{k_1}\binom{M-1}{k_2}.$$

En sommant, on obtient la formule annoncée.

Dans le cas $M<N$ et $k=2M+1$, tous les mots doivent commencer par une longueur de x et se terminer par une longueur de x, et toutes les longueurs de y sont formées d'une seule lettre ; le même raisonnement donne la formule annoncée. □

Lorsque les entiers M et N sont tous les deux petits (par exemple, moins d'une quinzaine), on peut appliquer les formules exactes pour calculer la loi de k, et choisir r en conséquence : pour un seuil acceptable de risque α, on choisira un r tel que la fonction de répartition de la loi de L, en r, soit inférieure à 0.05 (ce qui signifie exactement que $\mathbb{P}(L\leq r)<0.05$: c'est la condition que nous voulons sur la probabilité de rejeter l'hypothèse si elle est vraie).

Lorsque M et N deviennent grands (il est préférable de maintenir M et N du même ordre de grandeur ; si M est de l'ordre de la centaine, il est déconseillé d'avoir N de l'ordre de la dizaine), on a typiquement recours à une loi approchée pour L.

En écrivant sous forme de quotients de produits de factorielles les formules à base de coefficients binomiaux :

$$\binom{a}{b}=\frac{a!}{b!(a-b)!},$$

et en appliquant soigneusement la formule d'approximation de Stirling :

$$a!=a^a e^{-a}\sqrt{2\pi a}(1+o(a)),$$

on obtient le résultat suivant, que nous admettrons :

Proposition 9.6 *Lorsque M et N sont suffisamment grands, la loi de L est proche de la loi normale $\mathcal{N}(m,\sigma)$, avec*

$$\begin{aligned} m &= 1+\frac{2MN}{M+N} \\ \sigma^2 &= \frac{2MN(M^2+N^2-M-N)}{(M+N)^2(M+N-1)} \end{aligned}$$

On peut alors déterminer r en fonction de α, en utilisant une table de la fonction de répartition de la loi normale réduite.

Annexe A

Tables

On trouvera dans ces pages, des tables donnant la fonction de répartition de lois importantes (loi normale réduite, lois du χ^2, lois de Student).

A.1 Loi normale réduite $\mathcal{N}(0,1)$

La fonction de répartition de la loi normale réduite est donnée par la formule

$$F(x) = \mathbb{P}(N \leq x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-x^2/2} dx.$$

Les valeurs de $F(x)$, pour des valeurs comprises entre 0 et 3, sont données dans la table A.1 ; on obtient x en ajoutant les valeurs en tête de la ligne et de la colonne, la table donnant alors $F(x)$. Pour des valeurs de x négatives, on utilise la parité de la densité, qui se traduit par

$$F(-x) = 1 - F(x)$$

A.2 Lois du χ^2

La table A.2 donne, pour différentes valeurs de α, et en fonction du nombre d de degrés de liberté, la valeur de r pour laquelle on a

$$\mathbb{P}(\chi^2_d > r) = \alpha.$$

Pour de grandes valeurs de d, on peut considérer que la loi du χ^2 à d degrés de liberté est égale à la loi normale $\mathcal{N}(d, 2d)$; c'est-à-dire que, si U_d suit la loi du χ^2 à d degrés de liberté, $(U_d - d)/\sqrt{2d}$ suit la loi normale réduite $\mathcal{N}(0,1)$, dont la fonction de répartition est donnée par la table A.1.

A.3 Lois de Student

La table A.3 donne, pour différents nombres de degrés de liberté et différentes valeurs de α, la valeur qui a probabilité α d'être dépassée *en valeur absolue* par une variable de Student à ce nombre de degrés de libertés : c'est le t tel que l'on ait

$$\mathbb{P}(|S_d| > t) = \alpha.$$

TAB. A.1 – Fonction de répartition de la loi normale réduite

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

TAB. A.2 – Table du χ^2

d	0.99	0.95	0.90	0.50	0.10	0.05	0.01
1	0.0002	0.004	0.016	0.455	2.706	3.842	6.635
2	0.020	0.103	0.211	1.386	4.606	5.994	9.227
3	0.115	0.352	0.584	2.366	6.252	7.816	11.346
4	0.297	0.711	1.064	3.357	7.780	9.489	13.279
5	0.554	1.145	1.610	4.351	9.237	11.071	15.089
6	0.872	1.635	2.204	5.348	10.645	12.592	16.813
7	1.239	2.167	2.833	6.346	12.017	14.067	18.476
8	1.646	2.733	3.490	7.344	13.362	15.508	20.091
9	2.088	3.325	4.168	8.343	14.684	16.919	21.666
10	2.558	3.940	4.865	9.342	15.987	18.307	23.209
11	3.053	4.575	5.578	10.341	17.275	19.675	24.725
12	3.571	5.226	6.304	11.340	18.550	21.026	26.217
13	4.107	5.892	7.042	12.340	19.812	22.362	27.688
14	4.660	6.571	7.789	13.339	21.064	23.685	29.141
15	5.229	7.261	8.547	14.339	22.307	24.996	30.578
16	5.812	7.962	9.312	15.339	23.542	26.296	32.000
17	6.408	8.672	10.085	16.338	24.769	27.587	33.408
18	7.015	9.390	10.865	17.338	25.990	28.869	34.805
19	7.633	10.117	11.651	18.338	27.203	30.143	36.191
20	8.260	10.851	12.443	19.338	28.412	31.410	37.566
21	8.897	11.591	13.240	20.337	29.615	32.671	38.932
22	9.543	12.338	14.041	21.337	30.813	33.925	40.289
23	10.196	13.091	14.848	22.337	32.007	35.172	41.638
24	10.856	13.848	15.659	23.337	33.196	36.415	42.980
25	11.524	14.611	16.473	24.337	34.382	37.652	44.314
30	14.95	18.49	20.60	29.34	40.26	43.77	50.89

Cela ne correspond pas directement à la fonction de répartition de la loi de Student : par symétrie (la densité est paire), on a

$$\mathbb{P}(S_d \leq t) = 1 - \frac{\alpha}{2}.$$

Pour un grand nombre de degrés de liberté, on pourra considérer que la loi de Student est gaussienne.

TAB. A.3 – Table de la loi de Student

d	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.355	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.100	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
80	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Bibliographie

[1] N. Bouleau, *Probabilités de l'ingénieur* – Hermann, 2002

[2] D. Foata, A. Fuchs, *Calcul des probabilités* – Dunod, 1998 (deuxième édition)

[3] D. Foata, A. Fuchs, *Processus stochastiques* – Dunod, 2002

[4] D.E. Knuth, *The Art of Computer Programming*, volume 2, *Seminumerical Algorithms* – Addison-Wesley, 1998 (troisième édition)

[5] W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes in C* – Cambridge University Press.