Probabilités élémentaires

Marc Diener

23 octobre 2006

Table des matières

Avant Propos

Ce texte est écrit à l'occasion d'un cours de Probabilité donné en deuxième année d'études de Mathématiques et Physique à l'Université de Nice Sophia-Antipolis.

Le projet, que je soupçonne original et que je vais évoquer ci-dessous, a été rendu possible grâce à l'allongement du premier cycle des études, à l'occasion de la mise en accord des enseignements supérieurs dans l'Union Européenne. En effet cette réforme permet de concevoir la formation académique au calcul des probabilités sur une période plus longue tout en imposant, à mon avis, que des étudiants qui quitteraient l'enseignement supérieur sans avoir achevé ce cycle d'étude le feraient en détenant une réelle formation.

Le calcul des probabilités a connu, en France au moins, un gain d'intérêt hors du monde académique dont peut de branches des Mathématiques a pu bénéficier, la théorie des nombres appliquée au codage étant un autre exemple d'un tel intérêt nouveau.

Cet attrait est certe lié a l'augmentation de l'utilisation des outils statisques anciens et nouveaux, rendue possible biensûr par la baisse impressionnante du coût du calcul statistique, mais aussi et surtout, l'accumulation dans les entreprises et les administrations, d'entrepos de données se prètant à la fouille de données au prix quasi exclusif de la formations à ces outils des personnels auquels sont confiés ces entrepos.

Mais il me semble qu'au moins une autre circonstance explique cet intérêt extra-académique. Elle est apparue au début des année 70 du siècle écoulé, dans la question de la gestion des risques financiers par les contrats d'option. J'ai personnelement découvert cette application du calcul des probabilité à la lecture des notes que notre collègues Imme van den Berg a préparé pour les étudiants niçois au printemps 1996. J'y ai observé qu'une autre utilisation du langage des Probabilités pouvait exister que celle du paradigme statistique auquel le mot "probabilité" est dû, je pense. Je pense que cette utilisation n'a été rendue possible que grâce à la présentation moderne introduite au début des années 30 par Kolmogorov qui, au delà d'une remarquable extension des méthodes de Lebesgue et outils de Borel, a permis un véritable dépassement de la théorie de la mesure, en rendant accessible au calcul une formalisation de la multiplicité des futurs possibles et la prise en compte dynamique de ce qui "ne sera plus aléatoire", au fur et à mesure que se "révele l'information". En comprenant comment les financiers mathématiciens utilisent des probabilités j'ai eu la conviction qu'un nouveau paradigme pour leur utilisation est né à coté de l'ancien paradigme incarné par la Statistique. Voilà pourquoi j'aimerais appelé ce cours *Paradigmes nouveau et ancien pour les Probabilités*.

J'ai dit l'importance que je vois au travaux de Kolmogorov ; on comprendra donc aisément ma consternation devant le fait que les retombées de ces travaux soient mal connues de la communauté des mathématiciens au point que ce n'est que récemment qu'on a jugé utile d'introduire cette théorie au programme de l'Agrégation en Mathématiques. Bien entendu elle n'est pas "facile", mais tel est souvent le cas de l'exposition d'une nouvelle théorie, et ce n'est pas ce niveau de difficultés qui repousse généralement les mathématiciens, bien au contraire même !

Je ne suis pas loin de penser que c'est l'approche axée sur l'ancien paradigme qui explique les réticences qui ont conduit à ce que le nouveau ne bénéficie pas du minimum de curiosité que ne manque pas de sussiter une nouvelle théorie. Il me semble que l'on peut faire remonter à Jacques Bernoulli (1654-1705) l'origine de l'ancien paradigme, avec sa "Loi des Grands Nombres" qui établi le lien entre la fréquence observée des succès à une expérience (obtenir un double-six aux dés, par exemple) dont le résultat est incertain, mais pouvant être répétée *ad libitum* de manière "indépendante", et la "probabilité" de ce succès. Or, alors que ce mathématiciens était sous la protection d'un des plus grands de son époque, Leibnitz, il fallu attendre 1713 pour que soit publié (par son frère, Nicolas) son *Ars conjectandi*, c'est-à-dire plusieurs années après sa mort ! Il semble bien que ce soit essentiellement la responsabilité de Leibnitz si ce travail n'a pas été publié plus tôt, non par malveillance, mais par réelle infirmité de ce co-fondateur de l'Analyse, qui s'est révélé sourd aux raisonnements probabilistes de son élève.

Je cite cette anecdote, non pour accabler Leibnitz (dont, en non-standardiste, je me sens un peu le

descendant), mais au contraire pour lui rendre l'hommage d'être le meilleur mathématicien qui n'aura rien compris aux probabilités, ... et il n'était pas le seul, hélas. Je vois dans cet incident de l'histoire des Mathématiques la première occurence d'une suite d'incompréhensions qui font qu'à une date encore très récente on pouvait sans ruiner sa réputation de mathématicien cultivé affirmer que les Probabilités ne sont pas des Mathématiques, affirmation qui n'est plus émise par quiconque a un peu appris la théorie (moderne) des Probabilités.

Mais faut-il avoir compris toutes les subtilités de la méthode de Kolmogorov pour se prémunir de ce ridicule ? Je ne le pense pas, et suis convaincu que sa manière d'introduire les probabilités via un espace probabilisé abstrait permet de dissiper l'essentiel des ombres qui obscurcissait pour bien de mathématiciens le discours probabiliste, dès le contexte élémentaire de variables aléatoires finies, ce qui evite complètement l'aspect technique dans l'approche moderne[1]. Et ceci nous ramène à ce projet de premier cours de probabilités, en deuxième année de Licence.

(à suivre...)

[1] De fait, je ne fais là que reprendre l'idée d'Edward Nelson dans *Radically Elementary Probability Theory*, PUP (1987)

Chapitre 1

Algèbres d'évènements et variables aléatoires

1.1 Indexer toutes les valeurs possibles

Définition : Soit Ω un ensemble. On appelle *algèbre* (ou algèbre de Boole) sur Ω toute partie $\mathcal{A}$ de $\mathcal{P}(\Omega)$ ayant les propriétés suivantes :

1. $\emptyset \in \mathcal{A}$
2. si $A \in \mathcal{A}$, alors $A^c \in \mathcal{A}$
3. si $A \in \mathcal{A}$ et $B \in \mathcal{A}$, $A \cap B \in \mathcal{A}$

Exemples : $\mathcal{A}_0 = \{\emptyset, \Omega\}$ et $\mathcal{A}_\infty = \mathcal{P}(\Omega)$ sont deux exemples extrêmes l'algèbres sur Ω ; $\mathcal{A}_0$ est la "plus petite" et $\mathcal{A}_\infty$ est la "plus grande", au sens où, pour toute algèbre $\mathcal{A}$ sur Ω, on a $\mathcal{A}_0 \subseteq \mathcal{A} \subseteq \mathcal{A}_\infty$. Notez que si $(\mathcal{A}_i)_{i\in I}$ est une famille d'algèbres sur Ω, alors $\mathcal{A}^* := \bigcap_{i\in I} \mathcal{A}_i$, l'intersection des A_i, est aussi une algèbre sur Ω ; bien entendu, par définition de $\bigcap_{i\in I} \mathcal{A}_i$ et pour $A \subseteq \Omega$ on a $A \in \mathcal{A}^*$ si et seulement si $A \in \mathcal{A}_i$ pour tout $i \in I$.

Proposition 1.1 *Soient $\mathcal{A}$ une algèbre sur Ω, $N := \{1, \ldots, n\}$, et $(A_k)_{k\in N}$ une famille de sous-ensembles $A_k \in \mathcal{A}$; alors*

$$\bigcap_{k\in N} A_k \in \mathcal{A} \text{ et } \bigcup_{k\in N} A_k \in \mathcal{A} \tag{1.1}$$

On dit encore qu'une algèbre est stable par intersection finie *et* réunion finie.

Preuve : Montrons d'abord, par récurrence sur le *cardinal* (ou nombre d'éléments) n de N, que $\bigcap_{k\in N} A_k \in \mathcal{A}$. Pour $n = 1$ c'est une hypothèse ; supposons la relation vraie lorsque N est remplacé par $N' := \{1, \ldots, n-1\}$; donc $A' := \bigcap_{k\in N'} A_k \in \mathcal{A}$; comme $\mathcal{A}$ est une algèbre, on a donc $\bigcap_{k\in N} A_k = A' \cap A_k \in \mathcal{A}$ puisque $A_k \in \mathcal{A}$.

Par ailleurs $A := \bigcup_{k\in N} A_k = \bigcup_{k\in N} (A_k^c)^c = \left(\bigcap_{k\in N} A_k^c\right)^c \in \mathcal{A}$ puisque, comme nous venons le voir, $B := \bigcap_{k\in N} A_k^c \in \mathcal{A}$ du fait que $A_k^c \in \mathcal{A}$ et que $A = B^c \in \mathcal{A}$. □

Définition : On dit que l'algèbre $\mathcal{A}$ est une σ-algèbre (ou une *tribu*) si (1.1) est encore vrai pour $N = \mathbb{N}$. En d'autres termes, une σ-algèbre est une algèbre *stable par intersection dénombrable* et *réunion dénombrable.*

Définition : Soit Ω un ensemble et $\mathcal{B}$ une algèbre sur Ω. On appelle *variable aléatoire* (en abrégé v.a.) sur $(\Omega, \mathcal{B})$ toute fonction $X : \Omega \longrightarrow \mathbb{R}$ telle qu'on ait $\{X \leq x\} \in \mathcal{B}$ pour tout $x \in \mathbb{R}$. Si $\mathcal{A}$ est une algèbre sur Ω, on dit que la v.a. X est $\mathcal{A}$-mesurable si et seulement si $\{X \leq x\} \in \mathcal{A}$ pour tout $x \in \mathbb{R}$.

Une v.a. sur $(\Omega, \mathcal{B})$ est donc une fonction $\mathcal{B}$-mesurable.

Remarques :

1. $\{X \leq x\}$ dénote l'ensemble des $\omega \in \Omega$ tels que $X(\omega) \leq x$. Si, pour $G \subseteq \mathbb{R}$, on dénote par $X^{-1}(G)$ l'ensemble des $\omega \in \Omega$ tels que $X(\omega) \in G$, on a alors $\{X \leq x\} = X^{-1}(]-\infty, x])$. Observez que X est une fonction de Ω dans $\mathbb{R}$, alors que X^{-1} ainsi définie est une fonction de $\mathcal{P}(\mathbb{R})$ dans $\mathcal{P}(\Omega)$; elle est toujours définie, alors que la *fonction réciproque* de X n'est définie que si X est bijective (une situation guère intéressante dès qu'on veut pouvoir considérer plusieurs v.a. simultanément).
2. Une v.a. sur $(\Omega, \mathcal{P}(\Omega))$ n'est donc ni plus ni moins qu'une fonction définie sur Ω. Ce cas nous suffira amplement dans le cas où Ω est fini ; mais bien-entendu dans ce cas la v.a. ne pourra prendre qu'un nombre fini de valeurs, ce qui nous donnera un cadre pour parler d'un jeu de pile-ou-face (choisir par exemple $\Omega = \{0, 1\}$ et $X(\omega) = \omega$, mais $\Omega = \{1, \ldots, 6\}$ et $X(\omega) = \mathrm{mod}\,(\omega, 2)$ convient aussi), d'un jeu de dés, ou de toutes les valeurs inférieurs à 10^9 ([1]) d'une action cotée en $\frac{1}{64}$-èmes d'USD (prendre $\Omega = \{1, \ldots, 64 \cdot 10^9\}$ et $X(\omega) = \omega/64$)... Les vecteurs aléatoires (voir ci-dessous) sur un Ω fini nous donnerons aussi une manière de traiter d'une succession de d tirages d'un dé (prendre $\Omega = \{1, \ldots, 6\}^d$). En revanche, si nous voulons que notre v.a. puisse prendre toutes les valeurs réelles, ou si nous voulons considérer une suite infinie de tirages à pile-ou-face, un Ω infini sera indispensable. Dans ce cas on choisit généralement pour $\mathcal{B}$ la σ-algèbre (ou "tribu") des *boréliens* (voir cours de L3).
3. Une v.a. *vectorielle* est une fonction $X : \Omega \longrightarrow \mathbb{R}^d$, $X(\omega) = (X_1(\omega), \ldots, X_d(\omega))$, telle que pour tout $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, on a $\{X_i \leq x_i\} \in \mathcal{B}$ pour tout indice $i = 1..d$. Elle est $\mathcal{A}$-mesurable si et seulement si toutes ses "fonctions coordonnées" X_i le sont.
4. Pour une v.a. vectorielle X à valeurs dans $\mathbb{R}^d$ et $x \in \mathbb{R}^d$, par $X(\omega) \leq x$ nous entendons $X_i(\omega) \leq x_i$ pour tout $i = 1..d$, et nous dénoterons par $\{X \leq x\}$ de sous-ensemble $\bigcap_{i=1..d}\{X_i \leq x_i\}$.

Commentaire : Aussi curieux que cela puisse paraître, nous découvrirons progressivement que la manière dont une v.a. est définie (c'est-à-dire comment Ω et $\omega \mapsto X(\omega)$ sont explicitement choisis) est rarement le problème qui retient l'attention ; tout au plus s'assure-t-on qu'il est possible de les choisir ayant une propriété donnée. Cette observation ne peut être que sibyllines à ce stade ; elle n'est là que pour expliquer pourquoi nous ne donnerons guère d'exemples explicites. La notion de v.a. indépendantes qui sera introduite au chapitre suivant nous donnera l'occasion d'illustrer cette question.

Exercice : $X^{-1}(G)$ s'appelle l'*image réciproque* de G par la fonction X. Montrez qu'on a les relations suivantes : $X^{-1}(\emptyset) = \emptyset$, $X^{-1}(\mathbb{R}) = \Omega$, $X^{-1}(G^c) = (X^{-1}(G))^c$, $X^{-1}(G \cap H) = X^{-1}(G) \cap X^{-1}(H)$, $X^{-1}(G \cup H) = X^{-1}(G) \cup X^{-1}(H)$.

1.2 Engendrer des algèbres

Définition : Soit $\mathcal{X} \subseteq \mathcal{P}(\Omega)$ une famille de parties de Ω. On appelle *algèbre engendrée par* $\mathcal{X}$ la plus petite algèbre contenant $\mathcal{X}$; on la note $< \mathcal{X} >$; elle est égale à l'intersection de toutes les algèbres contenant $\mathcal{X}$.

Exemple : Si $\mathcal{X} = \{A, B\}$, pour $A \subseteq \Omega$ et $B \subseteq \Omega$, on a

$$< \mathcal{X} >= \{\emptyset, A, A^c, B, B^c, A \cap B, A^c \cup B^c, A \cap B^c, A^c \cup B, A^c \cap B, A \cup B^c, A^c \cap B^c, A \cup B, \Omega\}$$

On définit de même la *σ-algèbre engendrée par* $\mathcal{X}$ comme la plus petite σ-algèbre contenant $\mathcal{X}$; on la note $< \mathcal{X} >_\sigma$ et elle est égale à l'intersection de toutes les σ-algèbres contenant $\mathcal{X}$; la notion de σ-algèbre sera étudiée en L3 ; ici nous ne donnerons aucune preuve relatives aux σ-algèbres et admettrons systématiquement tout résultat faisant appel à cette notion, essentielle pour les v.a. prenant une infinité de valeurs.

Définition : Soit X une v.a. sur $(\Omega, \mathcal{B})$; on note $\alpha(X)$ l'algèbre engendrée par les $\{X \leq x\}$, pour tout $x \in \mathbb{R}$:

$$\alpha(X) :=< \{\{X \leq x\} \ , \ x \in \mathbb{R}\} > .$$

L'algèbre $\alpha(X)$ est la plus petite algèbre $\mathcal{A}$ telle que X soit $\mathcal{A}$-mesurable. On définit de manière analogue $\sigma(X)$, la σ-algèbre engendrée par les $\{X \leq x\}$, pour tout $x \in \mathbb{R}$; ainsi $\sigma(X) =< \{\{X \leq x\} \ , \ x \in \mathbb{R}\} >_\sigma$.

[1] 10^9 = un milliard = one billion en anglais US = one thousand million en anglais EU

Proposition 1.2 *Si X ne prend qu'un nombre fini n de valeurs $x_1, \ldots, x_n$, $\alpha(X)$ est aussi l'algèbre engendrée par les $A'_{x_i} := \{X = x_i\}$, $i = 1..n$. On a $\sigma(X) = \alpha(X)$.*

Preuve : On peut supposer les x_i numérotés par ordre strictement croissant : $x_1 < x_2 < \ldots < x_n$; posons $A_i := \{X \leq x_i\}$. Par définition $\alpha(X) =< \{A_1, A_2, \ldots, A_n\} >$; posons $\alpha'(X) :=< \{A'_{x_1}, A'_{x_2}, \ldots, A'_{x_n}\} >$. On voit facilement que

$$A'_{x_1} = A_1 \ , \ A'_{x_i} = A_i - A_{i-1} := A_i \cap A_{i-1}^c \in \alpha(X) \ , \text{ et } A_j = \bigcup_{i \leq j} A'^c_{x_i} \in \alpha'(X).$$

Donc, comme $A'_{x_i} \in \alpha(X)$ pour tout i, on a donc $\alpha'(X) \subseteq \alpha(X)$ puisque $\alpha'(X)$ est la plus petite algèbre ayant cette propriété. De même, comme $A_j \in \alpha'(X)$ pour tout j, on a donc $\alpha(X) \subseteq \alpha'(X)$ puisque $\alpha(X)$ est la plus petite algèbre ayant cette propriété. D'où $\alpha(X) = \alpha'(X)$ □

Observons que les A_{x_i} sont les *atomes* de l'algèbre $\alpha(X)$ au sens suivant :

Définition : On dit que $A \in \mathcal{A}$ est un *atome* de l'algèbre A si et seulement si pour toute parties non vides $C \in \alpha(X)$ telle que $C \subseteq A$ on a nécessairement $C = A$. Nous dirons que l'algèbre $\mathcal{A}$ est *atomisée* si et seulement si pour tout $D \in \mathcal{A}$ il existe une famille finie (ou dénombrable) $(A_i)_{i \in I(D)}$ d'atomes telle que $D = \bigcup_{i \in I(D)} A_i$.

Il est facile de voir que, sous les hypothèses de la proposition 1.2, $\alpha(X)$ est une algèbre atomisée ; les $(A_{x_i})_{i=1..n}$ constituent une *partition* de Ω (c'est-à-dire que $\Omega = \bigcup_{i=1..n} A_{x_i}$ et $A_{x_{i'}} \cap A_{x_{i''}} = \emptyset$ si $i' \times i''$.)

Exercice : Soit X une v.a. prenant deux valeurs $x_1 < x_2$. Donner la liste des éléments de $\alpha(X)$. Même question lorsque X prend trois valeurs $x_1 < x_2 < x_3$. On suppose que X prend quatre valeurs $x_1 < x_2 < x_3 < x_4$; quel est le cardinal[2] $Card(\alpha(X))$ de $\alpha(X)$?

Théorème 1.3 *Soient X et Y deux v.a. sur $(\Omega, \mathcal{B})$. La v.a. Y est $\alpha(X)$-mesurable si et seulement si il existe une fonction Borel-mesurable[3] $f : X(\Omega) \longrightarrow Y(\Omega)$ telle que $Y = f(X)$, c'est à dire que $Y(\omega) = f(X(\omega))$ pour tout $\omega \in \Omega$.*

Preuve : Nous ne prouvons ce théorème que dans le cas élémentaire où X et Y satisfont les hypothèses de la proposition 1.2. Soient $x_1, \ldots, x_n$ les valeurs de X et $y_1, \ldots, y_m$ les valeurs de Y ; notons $A_{x_i} : \{X = x_i\}$ et $B_{y_j} := \{Y = y_j\}$. Les $(A_{x_i})_{i=1..n}$ forment une partion de Ω de même que les $(B_{y_j})_{j=1..m}$; de plus les $(A_{x_i})_{i=1..n}$ sont les atomes de $\alpha(X)$ qu'ils engendrent.

Supposons tout d'abord qu'il existe f telle que $Y = f(X)$ (le fait que f est $\mathcal{B}$-mesurable n'est pas utile dans ce cadre élémentaire). Soit y quelconque et soient $\{y_{j_1}, \ldots, y_{j_k}\} := \{y_j \in Y(\Omega), \ \ y_j \leq y\}$; donc $\{Y \leq y\} = \{Y = y_{j_1}\} \cup \ldots \cup \{Y = y_{j_k}\}$.

Par ailleurs, pour tout $y_j \in Y(\Omega)\}$, on

$$\begin{aligned} \{Y = y_j\} &= \{\omega \in \Omega, \ \ Y(\omega) = y_j\} = \{\omega \in \Omega, \ \ f(X(\omega)) = y_j\} \\ &= \{\omega \in \Omega, \ \ X(\omega) = x_i \text{ et } f(x_i) = y_j\} \\ &= \bigcup_{x_i \in X(\Omega), f(x_i) = y_j} A_{x_i}, \end{aligned}$$

cette réunion comportant un nombre fini de termes puisque $X(\Omega)$ est fini. Ainsi, nous voyons que $\{Y \leq y\}$ est une réunion finie d'atomes de $\alpha(X)$; c'est donc bien un élément de $\alpha(X)$: la v.a. Y est donc bien $\alpha(X)$-mesurable.

Réciproquement, supposons que Y est $\alpha(X)$-mesurable ; on a donc $B_{y_j} \in \alpha(X)$ qui est une algèbre atomisée : donc pour tout $j = 1..m$, il existe des $x_1^j, \ldots, x_{n_j}^j$ tels que $B_{y_j} = A_{x_1^j} \cup \ldots \cup A_{x_{n_j}^j}$. Il suffit donc, pour $x \in \{x_1^j, \ldots, x_{n_j}^j\}$, de poser $f(x) = y_j$. Comme les $(B_j)_{j=1..n}$ forment une partition de Ω, ceci définit $f(x_i)$ (de manière univoque) pour tout $x_i \in X(\Omega) = \{x_1, \ldots, x_n\}$. En dehors de $X(\omega)$ on donne à f une valeur constante arbitraire ; par exemple $f(x) = 0$ si $x \in X(\Omega)^c$. A présent, nous voyons que pour tout $\omega \in \Omega$, il existe j tel que $Y(\omega) = y_j$, donc $\omega \in B_{y_j} = A_{x_1^j} \cup \ldots \cup A_{x_{n_j}^j}$, d'où $X(\omega) \in \{x_1^j, \ldots, x_{n_j}^j\}$, et donc $f(X(\omega)) = y_j = Y(\omega)$. □

[2]Rappelons qu'on appelle *cardinal* d'un ensemble fini le nombre de ses éléments. On dit que D est de cardinal *dénombrable* (ou simplement que D est dénombrable) s'il existe une bijection entre D et l'ensemble des entiers positifs $\mathbb{N}$

[3]cette hypothèse n'est importante que dans le cas où $X(\omega)$ n'est pas fini. La Borel-mesurabilité sera étudiée en L3.

1.3 Des évènements ni vrais ni faux

Définition : Les éléments de $\mathcal{B}$ sont appelés les *évènements* de $(\Omega, \mathcal{B})$, ou simplement des évènements, si Ω et $\mathcal{B}$ sont sous-entendu ou supposés fixés une fois pour toute.

Le fait qu'on a supposé que $\mathcal{B}$ est une algèbre implique que l'intersection $(A, B) \mapsto A \cap B$ et la réunion $(A, B) \mapsto A \cup B$ sont des lois de composition interne sur les éléments de $\mathcal{B}$ et la complémentation $c : A \mapsto A^c$ est une involution (c'est-à-dire que $c^2 = \mathrm{Id}$) dans $\mathcal{B}$. Rappelons que sur les parties de tout ensemble Ω on a les relations suivantes :

Proposition 1.4 *Si A, B, et C sont des parties de Ω, alors*

1. $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$,
2. $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$,
3. $(A \cup B)^c = A^c \cap B^c$,
4. $(A \cap B)^c = A^c \cup B^c$.

En associant à tout évènement A la v.a. $\mathbb{I}_A$, on peut établir une correspondance biunivoque entre évènements et v.a. à valeurs dans $\{0, 1\}$. Pour x et y dans $\mathbb{R}$, posons

$$x \wedge y = \mathrm{Min}\,\{x, y\} \text{ et } x \vee y = \mathrm{Max}\,\{x, y\},$$

ce qui définit les opérations $X \wedge Y$ et $X \vee Y$ sur les v.a. de la façon usuelle : $X \wedge Y(\omega) = X(\omega) \wedge Y(\omega)$, et $X \vee Y(\omega) = X(\omega) \vee Y(\omega)$. Nous avons alors les relations

$$1 - \mathbb{I}_A = \mathbb{I}_{A^c}, \tag{1.2}$$
$$\mathbb{I}_A \wedge \mathbb{I}_B = \mathbb{I}_{A \cap B} = \mathbb{I}_A \cdot \mathbb{I}_B, \tag{1.3}$$
$$\mathbb{I}_A \vee \mathbb{I}_B = \mathbb{I}_{A \cup B} = 1 - (1 - \mathbb{I}_A)(1 - \mathbb{I}_B). \tag{1.4}$$

Exercice : Démontrer ces relations.

Il est d'usage de considérer "le jet du dé amène un *as*", ou "le jet de la pièce amène un *pile*" comme étant un "évènement". Le fait qu'il soit "aléatoire" signifie précisément qu'il peut se produise ou non. Ceci se code très facilement par une v.a. $\mathbb{I}_A$; au fait que le dé amène un as correspond au fait que $\mathbb{I}_A(\omega) = 1$, l'autre issue correspondant au fait que $\mathbb{I}_A(\omega) = 0$: cela dépend de "l'état du monde" ω et si $\omega \in A$ ou $\omega \in A^c$. Nous voyons qu'ainsi le vrai et le faux se codent par 1 et 0 respectivement.

Au prochain chapitre nous allons montrer comment, en affectant une "probabilité" à chaque évenement on pourra quantifier "l'espérance" d'un v.a. et nous verrons plus tard dans quel sens cette espérance est la meilleure approximation déterministe (c'est-à-dire indépendante de l'état du monde) d'une v.a.

1.4 Questions

1. Que vaut $< \{A\} >$, où $A \subseteq \Omega$?
2. Soit $\Omega = \{1, 2, 3, 4, 5, 6\}$, $A = \{1, 2, 3\}$, et $B = \{2, 4, 6\}$. Que vaut $< \{A, B\} >$?
3. Soit $\Omega = \mathbb{N}$, $A = 2\mathbb{N} := \{2n, n \in \mathbb{N}\}$. Que vaut $< \{A\} >$?
4. Pourquoi toute fonction $X : \Omega \longrightarrow \mathbb{R}$ est-elle une v.a. sur $(\Omega, \mathcal{P}(\Omega))$?
5. On pose $\Omega = \{1, \ldots, 6\}$, $X(\omega) = \mathrm{mod}\,(\omega, 2)$, $Y(\omega) = \mathrm{mod}\,(\omega, 3)$, $Z(\omega) = \mathrm{mod}\,(\omega, 4)$. Y est-elle $\alpha(X)$ mesurable ? X est-elle $\alpha(Y)$ mesurable ? Z est-elle $\alpha(X)$ mesurable ? X est-elle $\alpha(Z)$ mesurable ?

Chapitre 2

Probabilité et espérance

Dans ce qui suit, Ω désigne l'ensemble des "états du monde" ω et $\mathcal{B}$ l'algèbre des évènements considérés sur Ω. Si Ω est fini, on pourra supposer que $\mathcal{B} = \mathcal{P}(\Omega)$; si Ω est infini, il faudra, pour certaines propriétés énoncées, supposer que $\mathcal{B}$ soit une σ-algèbre, par exemple la σ-algèbre des boréliens. Nous ne nous intéressons pas ici à expliquer en détails pourquoi ces précautions sont nécessaires et renvoyons au cours de probabilités de L3 sur cet aspect.

2.1 Probabilité et espérance des v.a. élémentaires

Définition : On appelle *probabilité sur* $(\Omega, \mathcal{B})$ toute fonction $\mathbb{P} : \mathcal{B} \longrightarrow [0,1]$ telle que $\mathbb{P}(\Omega) = 1$ et $\mathbb{P}(A \dot\cup B) = \mathbb{P}(A) + \mathbb{P}(B)$, où $A \dot\cup B$ désigne[1] simplement l'union $A \cup B$ tout en exprimant qu'*on suppose* que $A \cap B = \emptyset$.

Une manière d'interpréter $\mathbb{P}$ est que $\mathbb{P}(A)$ "mesure l'espérance[2] qu'on peut avoir que l'évènement A se réalisera". Cette mesure n'a nullement besoin d'exprimer les "chances" qu'a A de se réaliser (même si, comme nous le verrons avec la *loi des grands-nombres*, cela peut servir à cela dans un sens à préciser). Nous verrons ci-dessous des exemple d'autres mesures utiles. En revanche, nous attendons que cette espérance soit linéaire dans le sens suivant : nous avons vu qu'à tout évènement A nous pouvons faire correspondre la v.a. $\mathbb{I}_A$, où $\mathbb{I}_A(\omega)$ vaut 1 ou 0 selon que "A est vrai ou non" c'est-à-dire selon que "l'état du monde" ω appartient à A ou non. Si l'on désigne par $\mathbb{E}$ l'espérance (mathématique), la linéarité évoquée ici signifie simplement que pour tous évènements A et B de $\mathcal{B}$ et tous réels a et b, on a $\mathbb{E}(a\mathbb{I}_A + b\mathbb{I}_B) = a\mathbb{E}(\mathbb{I}_A) + b\mathbb{E}(\mathbb{I}_B) = a\mathbb{P}(A) + b\mathbb{P}(B)$, et en particulier $\mathbb{E}(\mathbb{I}_A) = \mathbb{P}(A)$. Nous allons devoir expliquer, étant donné une probabilité $\mathbb{P}$ sur $\mathcal{B}$, comment l'étendre à une espérance $\mathbb{E}(X)$ définie pour toute v.a. $\mathcal{B}$-mesurable. Nous ne donnerons la construction que dans le cas élémentaire et indiquerons, sans preuve, à quelle formules on abouti lorsqu'on étend la notion d'espérance au moyen des constructions classiques de la théorie de la mesure.

Voici ce cas élémentaire : considérons une v.a. X sur $(\Omega, \mathcal{B})$ et notons $\mathcal{X} := X(\Omega)$. Supposons que cet ensemble des valeurs de X soit fini : $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$. Nous dirons qu'une telle v.a. est *élémentaire*. Pour $i = 1..n$, soit $A_i := \{X = x_i\}$, l'ensemble des $\omega \in \Omega$ tels que $X(\omega) = x_i$. Il est facile de voir que $\Omega = \dot\bigcup_{i=1..n} A_i$ et que les A_i sont bien dans $\mathcal{B}$ puisque la v.a. X est par hypothèse $\mathcal{B}$ mesurables (expliquez cela !) ; donc $\mathbb{P}(\{X = x_i\}) = \mathbb{P}(A_i)$ est bien défini pour tout $x_i \in \mathcal{X}$. Cette petite construction nous permet d'écrire $X = x_1\mathbb{I}_{A_1} + x_2\mathbb{I}_{A_2} + \ldots + x_n\mathbb{I}_{A_n}$ (vérifiez cela, en envisageant la valeurs de chacun des deux membres de l'égalité lorsque $\omega \in A_i$). Par linéarité de l'espérance postulée ci-dessus, nous devons donc poser $\mathbb{E}(X) = \mathbb{E}(x_1\mathbb{I}_{A_1} + x_2\mathbb{I}_{A_2} + \ldots + x_n\mathbb{I}_{A_n}) = x_1\mathbb{E}(\mathbb{I}_{A_1}) + x_2\mathbb{E}(\mathbb{I}_{A_2}) + \ldots + x_n\mathbb{E}(\mathbb{I}_{A_n})$, et donc

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_i \text{ où } p_i := \mathbb{P}(A_i) \ \ (= \mathbb{E}(\mathbb{I}_{A_i})). \tag{2.1}$$

[1] De façon analogue, on notera $\dot\bigcup_{i\in I} A_i$ la réunion $\bigcup_{i\in I} A_i$ tout en exprimant qu'*on suppose* que $A_i \neq A_j$ si $i \neq j$

[2] le sens de la locution "mesurer l'espérance" n'est pas précisé ici et nous laissons au lecteur la liberté de choisir une intuition qui lui convienne. Avec l'exemple de paris sur une course de chevaux, nous lui proposons un peu plus loin un exemple où cette mesure de l'espérance est peut-être différente de ce qui lui vient à l'esprit. Nous figerons bientôt le sens du mot espérance dans un sens mathématique, d'ailleurs déduit de celui de probabilité, le mot mesure admettant lui aussi un sens mathématique, qui sera, lui, précisé en L3

Exemple : Voici un exemple de construction d'une probabilité associée à une situation à l'issue inconnue, et où la probabilité considérée ne prétend pas refléter les chances des diverses issues possibles. Considérons un cafetier qui souhaite organiser un pari sur une course de chevaux, pour le seul agrément de ses clients, et sans avoir l'intention de s'enrichir mais avec l'exigence de ne pas s'appauvrir non plus. La course comporte n chevaux numérotés $i = 1..n$. Voici une formulation probabiliste de cette situation :

Soit Ω l'ensemble des états du monde, et notons A_i l'évènement "le cheval i est le gagnant" (nous n'envisagerons pas ici le cas de gagnants ex aequo). Notons X_i le gain par euro misé sur le cheval i ; on a donc $X_i = x_i \mathbb{I}_{A_i}$, où x_i est un nombre à préciser. Notons s_i la somme des mises sur le cheval i, et $s := \sum_{i=1}^n s_i$ la mise totale. La somme totale à payer aux parieurs par le cafetier s'écrit donc

$$S := \sum_{i=1}^{n} s_i x_i \mathbb{I}_{A_i}.$$

Comme de cafetier ne veut pas s'appauvrir, il faut donc que $S \leq s$, c'est-à-dire que $S(\omega) \leq s$ pour tout état du monde $\omega \in \Omega$. En choisissant $\omega \in A_{i_0}$ on voit immédiatement que le contrainte du cafetier implique que $x_{i_0} \leq \frac{s}{s_{i_0}}$ pour n'importe quel i_0 (faites le calcul en supposant que $\omega \in A_{i_0}$).

Notons que jusqu'ici nous n'avons pas fait appel à la notion de probabilité. Postulons maintenant qu'une même probabilité $\mathbb{P}^*$ permette de prendre en compte les mises de tous les joueurs ; ceci signifie que chaque euro misé sur le cheval i se fait avec l'espoir de gagner au moins un euro, où le sens donné au mot "espoir" s'accorde tout-à-fait avec le fait que dans certains états du monde on perde sa mise, bien entendu. Nous pouvons à présent revenir à un discours exclusivement mathématique.

La contrainte des parieurs sur l'évènement A_i s'écrit donc

$1 \leq \mathbb{E}^*(X_i) = \mathbb{E}^*(x_i \mathbb{I}_{A_i}) = x_i \mathbb{E}^*(\mathbb{I}_{A_i}) = x_i \mathbb{P}^*(A_i)$, où $\mathbb{E}^*$ désigne l'espérance au sens de la probabilité $\mathbb{P}^*$.

Donc $\mathbb{P}^*(A_i) \geq \frac{1}{x_i} \geq \frac{s_i}{s}$ pour tout $i = 1..n$. A présent, comme $\Omega = \dot{\bigcup}_{i=1..n} A_i$, nous obtenons

$$1 = \mathbb{P}^*(\Omega) = \mathbb{P}^*\left(\dot{\bigcup}_{i=1..n} A_i\right) = \sum_{i=1..n} \mathbb{P}^*(A_i) \geq \sum_{i=1..n} \frac{s_i}{s} = 1,$$

qui n'est possible que si $\sum_{i=1..n} \mathbb{P}^*(A_i) = \sum_{i=1..n} \frac{s_i}{s}$, qui n'est, à son tour, possible que si

$$\mathbb{P}^*(A_i) = \frac{s_i}{s}.$$

Nous voyons donc qu'ici la mesure de l'espérance que le cheval i gagne qui se dégage de ce pari est égale à la proportion des mises qui se sont portées sur ce cheval : c'est une "mesure de marché", le marché des clients prêts à parier un euro dans le seul "espoir" d'en gagner autant mais avec la possiblité, selon l'état de monde, d'en gagner plus.

A noter qu'il est à présent possible de vérifier que toutes les inégalités que nous avons dégagées pour la modélisation doivent donc être des égalité : en particulier $S = s$, qui exprime que le cafetier ne gagne rien à cela (sauf si certains parieurs heureux décident de payer une tournée générale, mais ça, c'est une autre affaire !). Nous voyons aussi que pour $\mathbb{P}^*$, l'espérance de gain par euro misé sur le cheval i_0 est égale à $\mathbb{E}^*(X_{i_0}) = \mathbb{E}^*(x_{i_0} \mathbb{I}_{A_{i_0}}) = x_{i_0} \mathbb{P}^*(\mathbb{I}_{A_{i_0}}) = \frac{s}{s_{i_0}} \times \frac{s_{i_0}}{s} = 1$. En d'autres termes, l'espérance est égale à la mise : c'est l'idée que l'on se fait d'une loterie équitable.

Notons que généralement les cafetiers n'ont pas l'esprit frondeur et ne se permettent pas d'organiser des jeux d'argent, qui ne sont, rappelelons-le, autorisés que dans un cadre très strict : ils préfèrent passer un accord avec le PMU ou la Française des Jeux (FdJ), qui partagent avec l'état (40%) et les cafetiers, une part importante des mises, et ne redistribue que le solde aux parieurs. Ce faisant, on remplace la contraite $S \leq s$ par $S < 40\% s$. On en déduit alors que les client de la FdJ acceptent allègrement que "l'espoir de gain" pour la probabilité de marché $\mathbb{P}^*$ soit strictement inférieur à la mise ; cela peut avoir un sens si l'on pense que pour un cheval i_0 on a $\mathbb{P}(A_{i_0}) > \mathbb{P}^*(A_{i_0})$ pour une autre probabilité $\mathbb{P}$ qui leur semble plus pertinente que la probabilité de marché $\mathbb{P}^*$.

2.2 Extension du domaine de l'espérance

Une fois définie l'espérance $\mathbb{E}$ par (2.1) pour les v.a. élémentaires, on l'étend, "par approximation", aux v.a. quelconques en posant $\mathbb{E}(X) = \lim_k \mathbb{E}(X_k)$ où $(X_k)_{k=1,2,...}$ est une suite de v.a. élémentaires "tendant" vers X. Esquissons comment cela sera fait en L3. Par exemple, si X ne prend qu'un ensemble

dénombrable de valeurs $(x_i)_{i\in\mathbb{N}}$, si $\sum_{i=0}^{+\infty}|x_i|p_i<+\infty$, $p_i:=\mathbb{P}(\{X=x_i\})$, il suffit de poser $X_k(\omega)=X(\omega)$ si $X(\omega)=x_i$ pour $i\leq k$, et $X_k(\omega)=0$ sinon ; nous voyons que le choix des X_k dépend de l'ordre dans lequel on a numéroté les valeurs de X, mais on peut montrer que l'hypothèse que $\sum_{i=0}^{+\infty}|x_i|p_i<+\infty$ assure que ceci est sans conséquence sur la valeur de

$$\mathbb{E}(X):=\lim_{k\to+\infty}\mathbb{E}(X_k)=\lim_{k\to+\infty}\sum_{i=0}^{k}x_i\mathbb{P}(\{X=x_i\})=\sum_{i=0}^{+\infty}x_i\mathbb{P}(\{X=x_i\})=\sum_{i=0}^{+\infty}x_ip_i \text{ où } p_i:=\mathbb{P}(\{X=x_i\}).$$

Si l'ensemble des valeurs de X n'est pas dénombrable, il convient de "bricoler" un peu plus : on commence par scinder X en différence de deux v.a. positives $X=X_+-X_-$, avec $X_+(\omega)=\mathrm{Max}\{X(\omega),0\}$ et $X_-(\omega)=\mathrm{Max}\{-X(\omega),0\}$, puis, pour chaque v.a. positive Y, on l'approche (de façon monotone) par des v.a. Y_k, avec $Y_k(\omega)=y_i^{(k)}$ si $0=y_0^{(k)}\leq y_i^{(k)}\leq Y(\omega)<y_{i+1}^{(k)}\leq k$, et $Y_k(\omega)=0$ si $Y(\omega)>k$, les $y_i^{(k)}$ constituant une discrétisation $[0..k]_{\frac{1}{2^k}}$ de l'intervalle $[0,k]$ par des point espacés de $\frac{1}{2^k}$. Chaque v.a. aléatoire $X_+^{(k)}$ et $X_-^{(k)}$ ainsi confectionnée est alors élémentaire. La théorie de la mesure montre alors que si $\lim_k\mathbb{E}(X_+^{(k)})<+\infty$ et $\lim_k\mathbb{E}(X_-^{(k)})<+\infty$, alors $\lim_{k\to+\infty}X_+^{(k)}-X_-^{(k)}=X$ et on peut poser $\mathbb{E}(X):=\lim_k\mathbb{E}(X_+^{(k)})-\lim_k\mathbb{E}(X_-^{(k)})$, toute autre manière d'approcher X_- et X_+ de façon monotone par des v.a. élémentaires conduisant nécessairement au même résultat (la définition précise et la preuve élégante de toutes ces affirmations est précisément l'objet de la théorie de Lebesgue).

On définit ainsi $\mathbb{E}(X)$ pour une large classe de v.a., dites "intégrables" (voir ci-dessous un exemple de v.a. non intégrable). Nous allons indiquer ici, sans démonstration, comment se calcule $\mathbb{E}(X)$ dans un certain nombre de cas de v.a. non élémentaires, en particulier lorsque la *loi* de X admet une *densité*.

Définition : On appelle *fonction de répartition* de la v.a. X sur $\mathbb{R}^d$ ($d\geq 1$) la fonction $F_X:\mathbb{R}^d\longrightarrow[0,1]$ définie par

$$F_X(x):=\mathbb{P}(\{X\leq x\}) \text{ (où } X(\omega)\leq x \text{ signifie } X_i(\omega)\leq x_i \text{ pour tout i=1..d)};$$

on l'appelle aussi la *loi* de X. On dit que la loi de X admet une *densité* f_X si la fonction $f_X:\mathbb{R}^d\longrightarrow\mathbb{R}^+$ est telle que pour tout $x_0\in\mathbb{R}^d$ on a $F_X(x_0)=\int_{x\leq x_0}f_X(x)dx$.

Remarques : La fonction de répartition est croissante puisque si $x'<x''$ on a $\{X\leq x'\}\subseteq\{X\leq x''\}$ et donc

$$F_X(x')=\mathbb{P}(\{X\leq x'\})\leq\mathbb{P}(\{X\leq x'\})+\mathbb{P}(\{x'<X\leq x"\})=\mathbb{P}(\{X\leq x''\})=F_X(x'').$$

Par monotonie, comme F_X est bornée (par 0 et 1), les limites $\lim_{x\to-\infty}F_X(x)$ et $\lim_{x\to+\infty}F_X(x)$ existent, et on montre que nécessairement

$$\lim_{x\to-\infty}F_X(x)=0 \text{ et } \lim_{x\to+\infty}F_X(x)=1$$

Si F_X est dérivable de dérivée F_X' continue, la dérivée $f_X:=F_X'$ est nécessairement positive puisque F_X est croissante, et X admet $f_X:=F_X'$ pour densité. Observez que l'hypothèse de mesurabilité des v.a. assure préciséement que $\{X\leq x\}\in\mathcal{B}$ pour tout $x\in\mathbb{R}$; les nombres $\mathbb{P}(\{X\leq x\})$ sont donc tous bien définis.

Exercice : Montrer que $\lim_{x\to-\infty}F_X(x)=0$ et $\lim_{x\to+\infty}F_X(x)=1$ dans le cas où la v.a. X ne prend qu'un nombre fini de valeurs $x_1<x_2<\ldots<x_n$.

Exemples : On dit que la v.a. X suit une *loi uniforme* sur $[a,b]$, et on note $X\rightsquigarrow\mathcal{U}([a,b])$ si et seulement si $a<b$, et si $x\in[a,b]$ alors $\mathbb{P}(\{X\leq x\})=\frac{x-a}{b-a}$; si $x\leq a$ alors $\mathbb{P}(\{X\leq x\})=0$; si $x\geq b$ alors $\mathbb{P}(\{X\leq x\})=1$.

On dit que la v.a. suit une *loi normale centrée réduite* et on note $X\rightsquigarrow\mathcal{N}(0,1)$ si elle admet pour densité la fonction $f_X(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, dite *fonction gaussienne*[3] ; le théorème central-limit nous expliquera pourquoi cette loi est capitale, notamment en statistiques.

Les observations faites en début de section conduisent au théorème suivant dont la preuve sera donnée en cours de Probabilités de L3 :

[3]en l'honneur de Karl Friedrich Gauss

Théorème 2.1 *Soit X une v.a. à valeurs dans $\mathbb{R}^d$. Si X ne prends qu'un nombre fini de valeurs $x_1, \ldots, x_n$ alors*

$$\mathbb{E}(X) = \sum_{i=1}^{n} x_i p_i \ , \text{ où } p_i := \mathbb{P}(\{X = x_i\}). \tag{2.2}$$

Si X ne prend qu'une suite dénombrable de valeurs $(x_i)_{i\in\mathbb{N}}$ et si $\sum_{i=0}^{+\infty} |x_i| p_i < +\infty$, alors

$$\mathbb{E}(X) = \sum_{i=1}^{+\infty} x_i p_i \ , \text{ où } p_i := \mathbb{P}(\{X = x_i\}). \tag{2.3}$$

Soit[4] *$X : \Omega \longrightarrow \mathbb{R}$ (donc $d = 1$). Si la loi de X admet une densité f_X et si $\int |x| f_X(x)dx < +\infty$, alors*

$$\mathbb{E}(X) = \int_{x\in\mathbb{R}} x f_X(x)dx = \int_{-\infty}^{+\infty} x f_X(x)dx. \tag{2.4}$$

Si la loi de X admet une densité f_X, si la v.a. Y est définie par $Y = g(X)$, et si $\int |g(x)| f_X(x)dx < +\infty$, alors

$$\mathbb{E}(Y) = \int g(x) f_X(x)dx. \tag{2.5}$$

Notons qu'il convient encore de s'assurer que les fonctions f_X et g mentionnées dans ce paragraphe sont "Borel-mesurables" c'est-à-dire suffisamment régulières pour que les intégrales mentionnées aient bien un sens, une limitation guère contraignante dans un bonne théorie de l'intégration.

Exercice : Montrer que si $X \rightsquigarrow \mathcal{U}([a, b])$, alors $\mathbb{E}(X) = \frac{a+b}{2}$; déterminer la loi de $Y = \alpha X + \beta$. Montrer que si $X \rightsquigarrow \mathcal{N}(0, 1)$, alors $\mathbb{E}(X) = 0$; déterminer la loi de $Y = \alpha X + \beta$.

Exercice : On dit que X suit une *loi de Cauchy* si et seulement si la loi de X admet pour densité la fonction $f_X(x) = \frac{C}{1+x^2}$. Determiner la valeur de la constante C. Montrer que $\int |x| f_X(x)dx \left(:= \int_{-\infty}^{+\infty} |x| f_X(x)dx\right) = +\infty$: on dit que la v.a. X *n'est pas intégrable*. En particulier l'espérance $\mathbb{E}(x)$ n'est pas définie. **Rep :** $C = \pi$.

2.3 Quelques propriétés des probabilités et de leur espérance

Proposition 2.2 *Soit $\mathbb{P}$ une probabilité sur $(\Omega, \mathcal{B})$. Soient A, B, $(A_i)_{i=1..n}$ des évènement. On a les relations suivantes :*

1. $\mathbb{P}(\emptyset) = 0$
2. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
4. *Si les A_i sont deux-à-deux disjoints, alors $\mathbb{P}\left(\dot{\bigcup}_{i=1..n} A_i\right) = \sum_{i=1..n} \mathbb{P}(A_i)$, et en particulier, si $A = \{\omega_1, \ldots, \omega_n\}$ est fini et $\{\omega_i\} \in \mathcal{B}$, alors $\mathbb{P}(A) = \sum_{i=1..n} \mathbb{P}(\{\omega_i\})$.*
5. *Si $\Omega = \dot{\bigcup}_{i=1..n} A_i$, alors $\mathbb{P}(B) = \sum_{i=1..n} \mathbb{P}(B \cap A_i)$.*
6. $\mathbb{P}(A) = \mathbb{E}(\mathbb{I}_A)$.

Théorème 2.3 *Pour toutes v.a. intégrables X et Y, et tous réels a et b on a*

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y) \ ;$$

si $X \leq Y$ (au sens où $X(\omega) \leq Y(\omega)$ pour tout $\omega \in \Omega$), alors $\mathbb{E}(X) \leq \mathbb{E}(Y)$

[4] Si d est quelconque, on a $X = (X_1, \ldots, X_d)$ et $\mathbb{E}(X) := \big(\mathbb{E}(X_1), \ldots, \mathbb{E}(X_d)\big)$, avec $\mathbb{E}(X_i) = \int_{x\in\mathbb{R}^d} x_i f_X(x)dx$.

Preuve : Comme d'habitude, nous ne prouvons ce théorème que dans le cas de v.a. élémentaires. Soient X et Y deux v.a. élémentaires, $\mathcal{X} := X(\Omega)$ et $\mathcal{Y} := Y(\Omega)$ les ensembles (fini) de leurs valeurs. Montrons tout d'abord que $\mathbb{E}(aX) = a\mathbb{E}(X)$; notons $a\mathcal{X} := \{ax, x \in \mathcal{X}\} = ax(\Omega)$. Comme $\{aX = ax\} = \{X = x\}$, on a

$$\mathbb{E}(aX) = \sum_{\xi \in a\mathcal{X}} \xi \mathbb{P}(\{aX = \xi\}) = \sum_{x \in \mathcal{X}} ax\mathbb{P}(\{aX = ax\}) = a \sum_{x \in \mathcal{X}} x\mathbb{P}(\{X = x\}) = a\mathbb{E}(X).$$

Il reste à montrer que $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$; à cette fin, posons $Z := X + Y$ et $\mathcal{Z} := Z(\Omega)$. notons $\mathcal{X} + \mathcal{Y} := \{x + y | x \in \mathcal{X}, y \in \mathcal{Y}\}$. On a bien-entendu $\mathcal{Z} \subseteq \mathcal{X} + \mathcal{Y}$. On a de plus $\{X = x\} = \dot{\bigcup}_{y \in \mathcal{Y}} \{X = x\} \cap \{Y = y\} =: \dot{\bigcup}_{y \in \mathcal{Y}} \{X = x, Y = y\}$, et donc

$$\begin{aligned}
\mathbb{E}(X) &= \sum_{x \in \mathcal{X}} x\mathbb{P}(\{X = x\}) = \sum_{x \in \mathcal{X}} x\mathbb{P}\left(\dot{\bigcup}_{y \in \mathcal{Y}} \{X = x, Y = y\}\right) \\
&= \sum_{x \in \mathcal{X}} x \sum_{y \in \mathcal{Y}} \mathbb{P}(\{X = x, Y = y\}) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} x\mathbb{P}(\{X = x, Y = y\}) \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} x\mathbb{P}(\{X = x, Y = y\}) \qquad \text{; de même a-t-on} \\
\mathbb{E}(Y) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} y\mathbb{P}(\{X = x, Y = y\}) \qquad \text{et donc} \\
\mathbb{E}(X) + \mathbb{E}(Y) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} x\mathbb{P}(\{X = x, Y = y\}) + \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} y\mathbb{P}(\{X = x, Y = y\}) \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} (x + y)\mathbb{P}(\{X = x, Y = y\}) \text{ , puis en regroupant par même valeur } z \text{ de } x + y \\
&= \sum_{z \in \mathcal{X} + \mathcal{Y}} \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y},\ x+y=z} z\mathbb{P}(\{X = x, Y = y\}) \\
&= \sum_{z \in \mathcal{Z}} \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y},\ x+y=z} z\mathbb{P}(\{X = x, Y = y\}) \\
&\qquad \text{car, pour } (x, y) \in \mathcal{X} \times \mathcal{Y} \text{ et } (z =)x + y \in \mathcal{Z}^c, \text{ on a } \{X = x, Y = y\} = \emptyset \\
&= \sum_{z \in \mathcal{Z}} z\mathbb{P}\left(\dot{\bigcup}_{(x,y) \in \mathcal{X} \times \mathcal{Y},\ x+y=z} \{X = x, Y = y\}\right) \\
&= \sum_{z \in \mathcal{Z}} z\mathbb{P}(\{Z = z\}) = \mathbb{E}(Z) = \mathbb{E}(X + Y).
\end{aligned}$$

Finalement, supposons que $X \leq Y$, et montrons que $\mathbb{E}(X) \leq \mathbb{E}(Y)$, ou encore, que $0 \leq \mathbb{E}(Y) - \mathbb{E}(X) = \mathbb{E}(Y - X) =: \mathbb{E}(U)$, où $U := Y - X$. Soit $\mathcal{U} := U(\Omega)$; pour tout $u \in \mathcal{U}$ on a donc $u = U(\omega) = Y(\omega) - X(\omega) \geq 0$; comme $\mathbb{P}(\{U = u\}) \geq 0$, on a bien $\mathbb{E}(U) = \sum_{u \in \mathcal{U}} u\mathbb{P}(\{U = u\}) \geq 0$.

□

Kolmogorov[5] :   : Gauss[6]

[5] Andrey Nikolaevich Kolmogorov (1903-1987)
[6] Karl Friedrich Gauss (1777-1855)

.

Chapitre 3

Probabilité et espérance conditionnelles

3.1 Probabilité conditionnelle

Rappelons que nous comprenons la probabilité d'un évenement $A \subseteq \Omega$ comme une mesure de l'espérance que nous avons que c'est l'évènement A qui se réalisera et non son contraire A^c, c'est-à-dire une mesure de l'espérance que nous avons que l'état du monde ω^* où l'on se trouve est tel que $\omega^* \in A$. Imaginons à présent que nous souhaitons déterminer comment réévaluer cette espérance si nous considérons comme acquis que $\omega^* \in B$ (soit qu'on a une information qui nous assure de cela, soit qu'on souhaite simplement traiter séparément les deux cas $\omega^* \in B$ et $\omega^* \notin B$). On appelle cette nouvelle probabilité la *probabilité* de A *sachant* B, et on la note $\mathbb{P}_B(A)$ (ou $\mathbb{P}(A|B)$). Soulignons que $\mathbb{P}_B$ est bien une probabilité sur (tout) Ω, simplement veut-on au-moins que $\mathbb{P}_B(A) = 0$ si $A \subseteq B^c$ (puisque si $\omega^* \in B$, on est certain que $\omega^* \notin A$) ; on veut aussi que $\mathbb{P}_B(B) = 1$ (puisqu'on envisage ici que le cas où $\omega^* \in B$). L'idée est alors de poser $\mathbb{P}_B(A) = c\mathbb{P}(A \cap B)$ ce qui assurera facilement que $\mathbb{P}_B$ est une probabilité, et comme $\mathbb{P}_B(B) = 1$, nous voyons que $c = 1/\mathbb{P}(B)$; bien-entendu ceci impose que $\mathbb{P}(B) \neq 0$. Le fait qu'un évènement B soit de probabilité nulle ou non étant souvent capitale, on dit qu'un évènement est *négligeable* si et seulement si $\mathbb{P}(B) = 0$.

D'où finalement la définition

Définition : Soit B un évenement non-négligeable de $(\Omega, \mathcal{B})$. On appelle *probabilité conditionnelle sachant* B la fonction $\mathbb{P}_B : \mathcal{B} \longrightarrow [0,1]$ définie, pour tout évènement $A \in \mathcal{B}$, par $\boxed{\mathbb{P}_B(A) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}}$.

Proposition 3.1 *Si $\mathbb{P}$ est une probabilité sur $(\Omega, \mathcal{B})$, et si $B \in \mathcal{B}$ n'est pas négligeable, alors $\mathbb{P}_B$ est également une probabilité sur $(\Omega, \mathcal{B})$.*

Exercice : Montrer la proposition 3.1 et vérifier que $\mathbb{P}_B(A) = \mathbb{E}(\mathbb{I}_A \mathbb{I}_B)/\mathbb{E}(\mathbb{I}_B)$.

On dit qu'une famille d'évènements $(A_i)_{i=1..n}$ forme une système complet d'évènements (s.c.é) si et seulement si [1] $\Omega = \dot{\bigcup}_{i=1..n} A_i$, c'est-à-dire que chaque $\omega \in \Omega$ appartient à un A_i et un seul. Dans ce cas, pour tout évènement $B \in \mathcal{B}$ on a $B = \dot{\bigcup}_{i=1..n} (B \cap A_i)$ et donc $\mathbb{P}(B) = \sum_{i=1}^n \mathbb{P}(B \cap A_i) = \sum_{i=1}^n \frac{\mathbb{P}(B \cap A_i)}{\mathbb{P}(A_i)} \mathbb{P}(A_i)$, d'où la *formule de la probabilité totale*

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i). \tag{3.1}$$

Imaginons connues les probabilité $\mathbb{P}(B|A_i)$ et $\mathbb{P}(A_i)$ pour tout $i = 1..n$. La *formule de Bayes*[2] suivante permet d'en déduire les $\mathbb{P}(A_i|B)$.

[1] Certains auteurs se contentent de demander que les A_i soient deux-à-deux disjoints et $\mathbb{P}\left(\bigcup_{i=1..n} A_i\right) = 1$; voyez-vous la différence ? Que peut-on dire de $N := \left(\bigcup_{i=1..n} A_i\right)^c$?

[2] Thomas Bayes (1702-1761), publication (postume) de *An Essay Toward Solving a Problem of Chances*, en 1754.

Théorème 3.2 (formule de Bayes) *Soit $(A_i)_{i=1..n}$ un s.c.é. de $(\Omega, \mathcal{B})$. Alors, pour tout évènement B non-négligeable, et tout $j = 1..n$, on a*

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\mathbb{P}(B)} \left(= \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \right).$$

Preuve : Elle se réduit à l'application de la définition de $\mathbb{P}(A_j|B)$ et de $\mathbb{P}(B|A_j)$:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(A_j \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \cap A_j)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\mathbb{P}(B)},$$

et on termine par la formule (3.1). □

Exercice : Deux usines se partagent la production mondiale de vistemboirs[3]. Un vistemboir choisi au hasard[4] a la probabilité 0.5 de provenir de l'une comme de l'autre usine. Les vistemboirs de la première usine sont défectueux avec une probabilité de 2% et ceux de la seconde usine sont défectueux avec une probabilité de 6%. On choisit un vistemboir au hasard : qu'elle est la probabilité qu'il provienne de la première usine (Indication : poser $A_i :=$"le vistemboir est produit dans l'usine i", et $B :=$"le vistemboir est défectueux". Cet exemple explique pourquoi cette formule est parfois appelée "formule des probabilités des causes".)

3.2 Indépendance

Cette définition assez anodine de probabilité conditionnelle nous conduit à une notion capitale en probabilité : celle d'indépendance. Au chapitre 1 nous avons vu ce qu'est une v.a. Y qui est X-mesurable : c'est une simplement fonction déterministe de X dans le sens où $Y = g(X)$ pour une fonction déterministe $x \mapsto g(x)$; donc $Y(\omega)$ est entièrement connu dès que $X(\omega)$ est connu. A l'inverse, nous souhaitons une notion d'indépendance telle la connaissance de $X(\omega)$ ne nous apprenne rien sur $Y(\omega)$. En fait nous allons définir ce que sont deux évènements A et B indépendants et demanderons que les évènements $\{X \leq x\}$ et $\{Y \leq y\}$ soient (tous) indépendants. L'idée que deux évènements A et B sont indépendants est que le fait de supposer, par exemple, que B a lieu ne change pas la probabilité de A, c'est-à-dire que $\mathbb{P}_B(A) = \mathbb{P}(A)$, ce qu'il n'a de sens que si $\mathbb{P}(B) \neq 0$. On veux aussi s'affranchir de l'apparente disymétrie et que $\mathbb{P}_A(B) = \mathbb{P}(B)$ (sous reserve cette fois que $\mathbb{P}(A) \neq 0$). Explicitons ces deux identités ; nous obtenons

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

En "chassant les dénominateurs", nous voyons que les deux relations se réduisent à l'unique relation $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$ qui a en outre l'avantage d'être définie même si $\mathbb{P}(A) = 0$ ou $\mathbb{P}(B) = 0$. Ceci nous conduit donc tout naturellement à la définition

Définition : Soit $\mathbb{P}$ une probabilité sur $(\Omega, \mathcal{B})$, et soient $A \in \mathcal{B}$ et $B \in \mathcal{B}$ deux évènements. On dit que les évènement A et B sont *indépendants pour la probabilité* $\mathbb{P}$ si et seulement si $\boxed{\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)}$.

Soient X et Y deux v.a. sur $(\Omega, \mathcal{B})$. On dit que les v.a. X et Y sont *indépendantes pour la probabilité* $\mathbb{P}$ si et seulement si pour tous x et y les évènements $\{X \leq x\}$ et $\{Y \leq y\}$ sont indépendants.

Notations : Il est commode d'écrire $A \perp\!\!\!\perp B$ pour noter que les évènements A et B sont indépendants, et d'écrire $X \perp\!\!\!\perp Y$ pour noter que les v.a. X et Y sont indépendantes.

Exemple : Pièces de monnaie indépendantes : Considérons l'ensemble des états du monde Ω suivant :

$$\Omega = \{0,1\}^2 = \{0,1\} \times \{0,1\} = \{\omega = (\omega_1, \omega_2), \omega_i = 0..1, i = 1..2\}.$$

Posons $X_1(\omega_1, \omega_2) = \omega_1$ et $X_2(\omega_1, \omega_2) = \omega_2$. Considérons deux pièces de monnaie, et modélisons le fait que la première tombe sur Pile par l'évènement $\{X_1 = 1\}$, l'évènement $\{X_2 = 1\}$ modélisant quant à lui que la seconde pièce tombe sur Pile. Soient $p := \mathbb{P}(\{X_1 = 1\})$ et $q := \mathbb{P}(\{X_2 = 1\})$. Supposons les deux pièces indépendantes. Notons $A := \{X_1 = 1\}$ et $B := \{X_2 = 1\}$; l'indépendance des deux pièces implique celle de A et B, donc $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) = pq$, et donc $\mathbb{P}(\{(X_1, X_2) = (1,1)\}) = pq$. En procédant de même pour les autres combinaisons de valeurs de X_1 et X_2, nous voyons que les deux pièces sont indépendantes si et seulement si on a le tableau des probabilités suivant :

[3] "Vistemboir" : nouvelle de Jacques Perret, La Machin, NRF, Gallimard, 1953.

[4] de "azzahr", le jeu de dés, en arabe ; ne me demandez pas comment on fait pour "choisir au hasard"...

$\downarrow X_1 = *, X_2 = * \longrightarrow$	1	0	$\mathbb{P}(\{X_1 = *\}) \downarrow$
1	pq	$p(1-q)$	p
0	$(1-p)q$	$(1-p)(1-q)$	$1-p$
$\mathbb{P}(\{X_2 = *\} \longrightarrow$	q	$1-q$	

Exercice : Pièces indiscernables : On reprend les notation de l'exemple précèdent. Montrer que $X_1 \perp\!\!\!\perp X_2 \quad \Rightarrow \quad \mathbb{P}(X_1 \neq X_2) = p + q - 2pq$. On dit que les deux pièces sont *indiscernables* si $\mathbb{P}(X_1 \neq X_2) = \mathbb{P}(X_1 = X_2 = 1) = \mathbb{P}(X_1 = X_2 = 0) = \frac{1}{3}$; montrer que deux pièces indiscernables ne peuvent être indépendantes.

Définition : Soit $(A_i)_{i\in I}$ une famille d'évènements (l'ensemble I peut être infini). On dit que ces évènements sont indépendants si et seulement si $I_0\mathbb{P}(\bigcap_{i\in I_0} A_i) = \Pi_{i\in I_0}\mathbb{P}(A_i)$ pour tout sous-ensemble *fini*. Il est commode de dénoter par $\coprod_{i\in I} A_i$ la propriété que ces évènements sont indépendants.

On dit que les v.a. d'une famille $(X_i)_{i\in I}$ sont indépendantes si et seulement si pour tous $x_i \in \mathcal{X}_i := X_i(\Omega)$ les évènements $(\{X_i \leq x_i\})_{i\in I}$ sont indépendants. Il est commode de dénoter par $\coprod_{i\in I} X_i$ la propriété que ces v.a. sont indépendantes.

Exercice : Evènements (mutuellement) indépendants : On reprend les v.a. indépendantes de l'exemple, et on choisit $p = \frac{1}{2} = q$; on pose $X = X_1$, $Y = X_2$, et $Z = |X - Y|$. Quelles sont les valeurs possibles pour la v.a. Z ? Les v.a. X et Y sont indépendantes par construction ; montrer que les v.a. X et Z sont indépendantes, et que les v.a. Y et Z sont indépendantes. On pose $A := \{X \leq 0\}$, $B := \{Y \leq 0\}$, et $C := \{Z \leq 0\}$. Calculer $\mathbb{P}(A)$, $\mathbb{P}(B)$, $\mathbb{P}(C)$, et $\mathbb{P}(A \cap B \cap C)$. Montrer que les v.a. X, Y, et Z *ne* sont *pas* indépendantes. Cet exemple montre que trois v.a. peuvent être indépendantes deux-à-deux sans être indépendantes. C'est pourquoi, pour des v.a. qui sont indépendantes au sens de la définition que nous avons données, on dit parfois qu'elles sont *mutuellement* indépendantes.

Proposition 3.3 *Si les v.a. $(X_i)_{i\in I}$ ne prennent chacune qu'un ensemble fini ou dénombrable de valeurs $\mathcal{X}_i := X_i(\Omega) = \{x_j^i, j = 1, 2, \ldots\}$, les v.a. $(X_i)_{i\in I}$ sont indépendantes si et seulement si pour tous $x^i \in \mathcal{X}_i$ les évènements $(\{X_i = x^i\})_{i\in I}$ sont indépendants.*

Soient $(X_i)_{i\in I}$ une famille de v.a. et $(g_i)_{i\in I}$ une famille de fonctions $g_i : X_i(\Omega) \longrightarrow \mathbb{R}$. Posons $Y_i := g_i(X_i)$. Si les v.a. $(X_i)_{i\in I}$ sont indépendantes, alors les v.a $(Y_i)_{i\in I}$ sont également indépendantes.

Théorème 3.4 *Si les v.a. X et Y sont indépendantes, on a $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$.*

Preuve : Voir cours ; c'est facile. □

Remarque : nous avions déjà vu que $\mathbb{E}$ est linéaire, et qu'on a donc $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$. Nous voyons ici que si les v.a. X et Y sont indépendantes, on a de plus $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(X)$. Ceci est évidemment très commode pour les calculs d'espérance, mais soulignons bien que si la linéarité est toujours assurée, la relation $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(X)$ est généralement fausse lorsque l'hypothèse d'indépendance n'est pas assurée.

A noter que la réciproque "$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(X) \Rightarrow X \perp\!\!\!\perp Y$" est fausse également : soit X une v.a. (élémentaire) symétrique, c'est-à-dire $\mathbb{P}(\{X \geq +x\} = \mathbb{P}(\{X \leq -x\}))$ pour tout x, alors $Z := X^3$ est également symétrique, et donc $\mathbb{E}(X) = 0 = \mathbb{E}(Z)$. Posons à présent $Y := X^2$; Y n'est bien-entendu pas indépendant de X (sauf si $X = 0$), et pourtant $\mathbb{E}(XY) = \mathbb{E}(X \cdot X^2) = \mathbb{E}(Z) = 0 = 0 \cdot \mathbb{E}(X^2) = \mathbb{E}(X) \cdot \mathbb{E}(X^2) = \mathbb{E}(Y)$.

Exercice : Démontrer ce qui vient d'être affirmé ci-dessus : donner un exemple de v.a. élémentaire symétrique X (choisir $X(\Omega) = \{-1, +1\}$. Montrer que $\mathbb{P}(X = +1) = \mathbb{P}(X = -1) = \frac{1}{2}$; en déduire que $\mathbb{E}(X) = 0 = \mathbb{E}(X^3)$. Montrer que X et $Y := X^2$ ne sont pas indépendantes.

3.3 Espérance conditionnelle

Soit tout d'abord Y une v.a. sur $(\Omega, \mathcal{B})$, et $A \in \mathcal{B}$ un évènement non-négligeable. Nous pouvons donc considérer la probabilité conditionnelle $\mathbb{P}_A$ sur $(\Omega, \mathcal{B})$. L'*espérance* de Y sachant A, notée $\mathbb{E}(Y|A)$, est l'espérance de Y pour cette probabilité $\mathbb{P}_A$, c'est-à-dire le *nombre*

$$\mathbb{E}(Y|A) := \mathbb{E}^{\mathbb{P}_A}(Y) := \frac{1}{\mathbb{P}(A)} \sum_{y\in Y(\Omega)} y\mathbb{P}(\{Y = y\} \cap A\}.$$

C'est donc la moyenne des valeurs de Y pondérée par la probabilité qu'ont ces valeurs lorsque l'évènement A est supposé assuré.

Si l'on considère le cas où l'évènement A est la survenue d'une valeur (non négligeable) d'une v.a. X, l'espérance conditionnelle permet d'associer à la v.a. Y une nouvelle v.a. $\overline{Y}_X$ ne dépendant que des valeurs de X. Voici comment :

Nous supposons que la v.a. X est élémentaire et notons $\mathcal{X} := X(\Omega) = \{x_1, \ldots, x_n\}$ l'ensemble de ses valeurs prises (avec probabilité non nulle, par définition d'une v.a. élémentaire). Ce qui suit n'est pas restreint à ce cas particulier mais l'exposition du cas général implique des constructions masquant trop facilement le but poursuivi.

Soit $x \in \mathcal{X}$; l'évènement $B_x := \{X = x\}$ est de probabilité non nulle et on peut considérer la probabilité $\mathbb{P}_{B_x}$ conditionnellement à cet évènement. Soit à présent Y une v.a. ; notons $\mathbb{E}_x(Y)$ l'espérance de Y pour la probabilité $\mathbb{P}_{B_x}$. Ce nombre ne prend donc en compte que les valeurs $Y(\omega)$ que prend Y lorsque $X(\omega) = x$; en effet

$$\mathbb{E}_x(Y) = \sum_{y \in Y(\Omega)} y\mathbb{P}_{B_x}(\{Y = y\}) = \frac{1}{\mathbb{P}(\{X = x\})} \sum_{y \in Y(\Omega)} y\mathbb{P}(\{Y = y\} \cap \{X = x\}),$$

et $\mathbb{P}(\{Y = y\} \cap \{X = x\}) = 0$ si $y \notin Y(\{X = x\})$, puisque $\{Y = y\} \cap \{X = x\} = \emptyset$ dans ce cas. Cette formule montre que le nombre $\mathbb{E}_x(Y)$ est la moyenne des valeurs de Y pondérée par la probabilité de $\{Y = y\} \cap \{X = x\}$. Si l'on effectue cette opération pour chaque $x \in X(\Omega)$ on définit une fonction $x \mapsto g(x) := \mathbb{E}_x(Y)$; nous pouvons alors considérer la nouvelle v.a. $\overline{Y}_X := g(X)$; par construction elle est constante sur les parties de la partition de Ω en les $\{X = x\}$, $x \in \mathcal{X}$, et cette valeur constante est une moyenne de Y sur $\{X = x\}$; nous pouvons donc comprendre $\overline{Y}_X$ comme la v.a. X-mesurable "ressemblant le plus à Y" ; en particulier, si Y est (déjà) X-mesurable, on a $\overline{Y}_X = Y$. En revanche, si Y et X sont indépendantes, ce traitement dégrade complétement la v.a. Y en une constante (la constante qui ressemble le plus à Y, à savoir $\mathbb{E}(X)$). Encore un mot : $\overline{Y}_X$ se note $\mathbb{E}(Y|X)$, notation commode dans les calculs, mais peut-être pas très évocatrice.

A noter que la v.a. X n'a été utilisée dans cette construction que pour définir la partition de Ω en les atomes de $\alpha(X)$ que sont les évènement $\{X = x\}$, pour $x \in \mathcal{X}$. Soit $\mathcal{A}$ est une algèbre quelconque et supposons que ses atomes A_i soient de probabilité non nulle. La même construction conduit à la v.a. $\mathbb{E}(Y|\mathcal{A})$, appelée "espérance conditionnelle de Y relativement à l'algèbre $\mathcal{A}$", qui est constante sur les atomes A_i de $\mathcal{A}$, définie par

$$\mathbb{E}(Y|\mathcal{A})(\omega) = \mathbb{E}(Y|A_i) \text{ pour tout } \omega \in A_i.$$

Exercice : Soit X une v.a. élémentaire. Montrer que $\mathbb{E}(Y|X) = \mathbb{E}(Y|\alpha(X)\}$; en déduire que $\mathbb{E}(Y|X) = \mathbb{E}(Y|g(X))$ pour toute application injective g. Ceci montre le rôle réduit des valeurs de la v.a. X dans $\mathbb{E}(Y|X)$: seule l'*information révélée par* X (c'est-à-dire $\alpha(X)$) est essentielle.

Théorème 3.5 *Soient $\mathcal{A}^+ \supseteq \mathcal{A}^-$ deux algèbres, alors $\mathbb{E}(\mathbb{E}(X|\mathcal{A}^+)|\mathcal{A}^-) = \mathbb{E}(X|\mathcal{A}^-)$.*

Si Z est $\mathcal{A}$-mesurable, alors $\mathbb{E}(ZX|\mathcal{A}) = Z\mathbb{E}(X|\mathcal{A})$.

Soit Y une v.a. $\mathcal{A}$-mesurable. Alors $Y = \mathbb{E}(X|\mathcal{A})$ si et seulement si $\mathbb{E}(Y\mathbb{I}_A) = \mathbb{E}(X\mathbb{I}_A)$ pour tout $A \in \mathcal{A}$

Exercice : Montrer ce théorème 3.5 dans le cas où les atomes de $\mathcal{A}$ et $\mathcal{A}^+$ (et donc ceux de $\mathcal{A}^-$) ne sont pas négligeables et Z est une v.a. élémentaire.

Thomas Bayes (1702-1761) :

Chapitre 4

Mesure de la variabilité

Le propre d'une v.a. c'est précisément de pouvoir prendre plus qu'une valeur, l'importance de ces valeurs étant contrôlée par la probabilité choisie. Nous considérons ici le cas des v.a. unidimensionnelles. Cette étude est également utile pour des v.a. à d dimensions, dans la mesure où elle s'applique à chacune des composantes. Toutefois, le cas multidimensionnel ouvre la question du comportement des diverses composantes l'une vis-à-vis de l'autre qui fera l'objet du chapitre suivant. Dans tout ce chapitre on supposera systématiquement que X *est de carré intégrable* (on note cela $X \in L^2(\Omega)$), c'est-à-dire que $\mathbb{E}(X^2) \leq \infty$, ce qui est toujours vrai lorsque X est une v.a. élémentaire.

4.1 Variance

Soit X une v.a. (unidimensionelle) et $\lambda \in \mathbb{R}$ un nombre. Considérons la quantité $\mathrm{Var}_\lambda(X)$ définie par

$$\mathrm{Var}_\lambda(X) := \mathbb{E}((X-\lambda)^2).$$

Il s'agit de l'espérance de la v.a. $(X-\lambda)^2$ qui est positive, donc d'espérance positive, et nous voyons que si $\mathrm{Var}_\lambda(X) = 0$ l'évènement $\{X \neq \lambda\}$ est de probabilité nulle. Voilà pourquoi on peut comprendre $\mathrm{Var}_\lambda(X)$ comme une mesure[1] de la variabilité de la v.a. X vis-à-vis de la valeur fixée λ. Une question naturelle maintenant est de déterminer (si possible) une valeur de λ vis-à-vis de laquelle la variablité de X est minimale. Le problème est simple ; en voici la solution :

Proposition 4.1 *Parmi toutes les valeurs de $\lambda \in \mathbb{R}$, celle pour pour laquelle $Var_\lambda(X)$ est la plus petite est $\lambda^* := \mathbb{E}(X)$.*

Preuve : Soit $\varphi(\lambda) := \mathrm{Var}_\lambda(X) = \mathbb{E}((X-\lambda)^2) = \lambda^2 - 2\lambda\mathbb{E}(X) + \mathbb{E}(X^2)$; le minimum λ^* de cette fonction polynôme de degré 2 de λ est la solution de $\varphi'(\lambda^*) = 0$. Or $\varphi'(\lambda) = 2(\lambda - \mathbb{E}(X))$, ce qui montre que $\lambda^* = \mathbb{E}(X)$ comme annoncé. □

Il est donc naturel de retenir $\mathrm{Var}_{\lambda_*}(X)$ comme mesure intrinsèque de la variabilité de la v.a. X ; c'est ce que l'on fait en posant la définition suivante :

Définition : On appelle *variance* de la v.a. unidimensionnelle X et on note $\mathrm{Var}\,(X)$ le nombre défini par

$$\boxed{\mathrm{Var}\,(X) := \mathbb{E}((X-\mathbb{E}(X))^2)}$$

On appelle é*cart-type* de la v.a. réelle X et on note $\sigma(X)$ le nombre défini par

$$\boxed{\sigma(X) := \sqrt{\mathrm{Var}\,(X)} = \sqrt{\mathbb{E}((X-\mathbb{E}(X))^2)}}$$

Proposition 4.2 *Pour tout $a \in \mathbb{R}$, on a* $\boxed{Var\,(aX) = a^2\,Var\,(X)}$ *et* $\boxed{\sigma(aX) = |a|\sigma(X)}$.

[1] Notons que si X et λ ont une "dimension" au sens de la Physique, c'est-à-dire représentent par exemple un nombre de mètres (m) ou d'euros (EUR), alors $\mathrm{Var}_\lambda(X)$ s'exprime en m^2 ou EUR^2, et il serait plus naturel de considérer la racine de $\mathrm{Var}_\lambda(X)$; les calculs en seraient en revanche compliqués et il est aisé de vérifier conceptuellement qu'ils aboutiraient aux mêmes conclusions.

Exercice : On suppose que $\sigma(X) = 0$; que peut-on dire de $\mathbb{P}(\{X \neq \mathbb{E}(X)\})$? (on supposera que la v.a. X est élémentaire)

Proposition 4.3 (formule de Huygens) $\boxed{Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2}$

Preuve : . Notons $x := \mathbb{E}(X)$; on a Var $(X) := \mathbb{E}((X-x)^2) = \mathbb{E}(X^2 - 2xX + x^2) = \mathbb{E}(X^2) - 2x\mathbb{E}(X) + x^2 = \mathbb{E}(X^2) - 2\mathbb{E}(X)^2 + (\mathbb{E}(X))^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. □

Remarque : La formule de Huygens peut encore s'écrire

$$\mathbb{E}(X^2) = (\mathbb{E}(X))^2 + \text{Var}\,(X). \tag{4.1}$$

Il s'agit en fait d'une formule de Pythagore : nous verrons au chapitre suivant en quoi la v.a. $\mathbb{E}(X)$ (en fait, non aléatoire) est toujours orthogonale à la v.a. $X - \mathbb{E}(X)$. Si l'on comprend que l'application $Y \mapsto \mathbb{E}(Y^2)$ comme le carré d'une "norme" $\|Y\|_{L^2}$, nous voyons que la formule de Huygens-Pythagore (4.1) n'est autre que $\|X\|_{L^2}^2 = \|\mathbb{E}(X)\|_{L^2}^2 + \|X - \mathbb{E}(X)\|_{L^2}^2$. En fait $Y \mapsto \sqrt{\mathbb{E}(X^2)}$ n'est "pas tout-à-fait" une norme, puisque $\|Y\|_{L^2} = 0$ n'entraine pas que $Y = 0$, mais seulement que $\mathbb{P}(\{Y \neq 0\}) = 0$, une petite subtilité à laquelle il faut s'habituer : on dit que $Y = 0$ *presque sûrement* (p.s.).

4.2 Variance de quelques lois

4.2.1 Loi de Bernoulli

Définition : On dit que la v.a. X suit une *loi de Bernoulli* et on note $X \rightsquigarrow \mathcal{B}(1,p)$ où $p \in]0,1[$ si et seulement si $X(\Omega) = \{0,1\}$ et $\mathbb{P}(\{X = 1\}) = p$.

Proposition 4.4 $\boxed{\textit{Si } X \rightsquigarrow \mathcal{B}(1,p), \textit{ alors } \mathbb{E}(X) = p \textit{ et } Var(X) = p(1-p).}$

4.2.2 Loi binomiale

La *loi binomiale* $\mathcal{B}(n,p)$ est la loi suivie par toute somme de n v.a. de Bernoulli $\mathcal{B}(1,p)$ indépendantes. La proposition suivante en précise quelques caractéristiques :

Proposition 4.5 *Soient $X_1, \ldots, X_n$ n v.a. de Bernoulli indépendantes, $X_i \in \mathcal{B}(1,p)$ pour tout $i = 1..n$. Soit $X = X_1 + \ldots + X_n \rightsquigarrow \mathcal{B}(n,p)$; alors,* $\boxed{\textit{pour tout } k = 0..n,\ \mathbb{P}(X = k) = C_n^k = \frac{n\,!}{k\,!(n-k)\,!}.}$

De plus $\boxed{\mathbb{E}(X) = np}$ *et* $\boxed{Var(X) = np(1-p)}$.

Preuve : Posons $\mathcal{C}_n^k := \{x = (x_1, \ldots, x_n) \in \{0,1\}^n \mid x_1 + \ldots + x_n = k\}$. On sait que Card $(\mathcal{C}_n^k) = C_n^k = \frac{n\,!}{k\,!(n-k)\,!}$. En effet, on a $\mathcal{C}_n^k = \mathcal{Y}_{n-1}^k \dot\cup \mathcal{Z}_{n-1}^{k-1}$ avec $\mathcal{Y}_{n-1}^k = \{x = (y_1, \ldots, y_{n-1}, 0)$ avec $y = (y_1, \ldots, y_{n-1}) \in \mathcal{C}_{n-1}^k\}$ et $\mathcal{Z}_{n-1}^{k-1} = \{x = (z_1, \ldots, z_{n-1}, 1)$, avec $z = (z_1, \ldots, z_{n-1}) \in \mathcal{C}_{n-1}^{k-1}\}$. Donc Card $(\mathcal{C}_n^k)$ = Card $(\mathcal{Y}_{n-1}^k)$ + Card $(\mathcal{Z}_{n-1}^{k-1})$. On en déduit facilement, par récurrence, que Card $(\mathcal{C}_n^k) = C_n^k := \frac{n\,!}{k\,!(n-k)\,!}$. A présent

$$\begin{aligned}
\mathbb{P}(\{X = k\}) &= \mathbb{P}(\{X_1 = x_1, \ldots, X_n = x_n \text{ , avec } x = (x_1, \ldots, x_n) \in \mathcal{C}_n^k\}) \\
&= \mathbb{P}\left(\dot\bigcup_{x \in \mathcal{C}_n^k} \{X_1 = x_1, \ldots, X_n = x_n\}\right) = \sum_{x \in \mathcal{C}_n^k} \mathbb{P}(\{X_1 = x_1, \ldots, X_n = x_n\}) \\
&= \sum_{x \in \mathcal{C}_n^k} \mathbb{P}(\{X_1 = x_1\} \cap \ldots \cap \{X_n = x_n\}) \\
&= \sum_{x \in \mathcal{C}_n^k} \mathbb{P}(\{X_1 = x_1\}) \cdots \mathbb{P}(X_n = x_n\}) \text{ car les } X_i \text{ sont indépendantes} \\
&= \sum_{x \in \mathcal{C}_n^k} p^k (1-p)^{n-k} \text{ puisque } x_1 + \ldots + x_n = k \\
&= p^k (1-p)^{n-k} \sum_{x \in \mathcal{C}_n^k} 1 = p^k (1-p)^{n-k} \text{Card}\,(\mathcal{C}_n^k) = C_n^k p^k (1-p)^{n-k}.
\end{aligned}$$

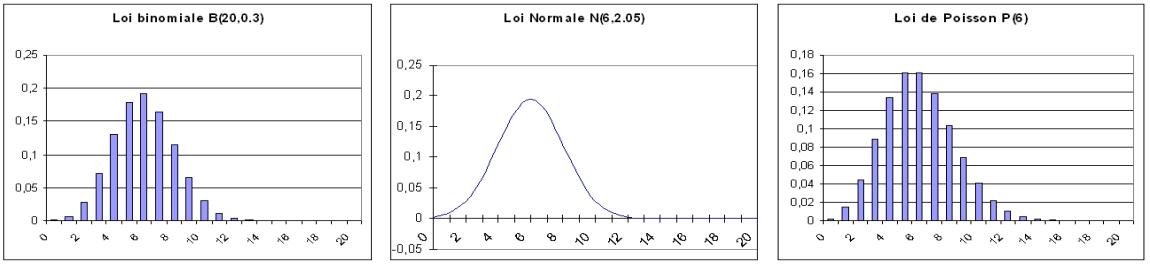


FIG. 4.1 – Histogramme et graphe de trois lois ayant même espérance 6 : la loi Binomiale $\mathcal{B}(20, 0.3)$, la loi normale $\mathcal{N}(6, \sqrt{4.2})$, et la loi de Poisson $\mathcal{P}(6)$. Les deux première ont la même variance 4.2 ; la valeur de la variance (6) de la loi de Poisson est imposée par le choix de son espérance. Nous verrons dans quelle mesure une loi binômiale $\mathcal{B}(n, p)$ peut être approchée, pour n grand, par une loi normale (théorème central-limit) de même espérance et même variance, ou par une loi de Poisson $\mathcal{P}(\lambda)$, pour $\lambda = \lim_n np_n$.

Finalement $\mathbb{E}(X_1+\ldots+X_n) = \mathbb{E}(X_1)+\ldots+\mathbb{E}(X_n) = np$; le fait que $\mathrm{Var}\,(X_1+\ldots+X_n) = \mathrm{Var}\,(X_1)+\ldots+\mathrm{Var}\,(X_n) = np(1-p)$ résulte du fait que les v.a. sont indépendantes et de la proposition 4.9 ci-dessous. □

4.2.3 Loi de Poisson

Définition : Soit $\lambda > 0$; on dit que la v.a. X suit une *loi de Poisson* de paramètre λ et on note $X \rightsquigarrow \mathcal{P}(\lambda)$ si et seulement si $X(\Omega) = \mathbb{N}$ et $\mathbb{P}(X = k) = c\frac{\lambda^k}{k\,!}$, avec $c = e^{-\lambda}$.

Proposition 4.6 $\boxed{\textit{Si } X \rightsquigarrow \mathcal{P}(\lambda),\ \textit{alors } \mathbb{E}(X) = \lambda \textit{ et } Var(X) = \lambda.}$

Preuve : laissée en exercice. Indication : pour le calcul de $\mathrm{Var}\,(X)$ commencer par calculer $\mathbb{E}(X(X-1))$ puis en déduire $\mathbb{E}(X^2)$ et finalement $\mathrm{Var}\,(X)$ en appliquant la formule de Huygens. □

4.2.4 Loi Exponentielle

Définition : Soit $\lambda > 0$; on dit que la v.a. X suit une *loi exponentielle* de paramètre λ et on note $X \rightsquigarrow \mathcal{E}(\lambda)$ si et seulement si $X(\Omega) = \mathbb{R}^+$ et X admet pour densité la fonction $f_X(x) = ce^{-\lambda x}\mathbb{I}_{\mathbb{R}^+}(x)$, avec $c = \lambda$.

Proposition 4.7 $\boxed{\textit{Si } X \rightsquigarrow \mathcal{E}(\lambda),\ \textit{alors } \mathbb{E}(X) = \frac{1}{\lambda} \textit{ et } Var(X) = \frac{1}{\lambda^2}.}$

Preuve : laissée en exercice. □

4.2.5 Loi normale

Définition : Soient $\mu \in \mathbb{R}$ et $\sigma > 0$; on dit que la v.a. X suit une *loi normale* d'espérance μ et écart-type σ et on note $X \rightsquigarrow \mathcal{N}(\mu, \sigma)$ si et seulement si $X(\Omega) = \mathbb{R}$ et X admet pour densité la fonction $f_X(x) = ce^{-\frac{(x-\mu)^2}{2\sigma^2}}$, avec $c = \frac{1}{\sigma\sqrt{2\pi}}$.

Proposition 4.8 $\boxed{\textit{Si } X \rightsquigarrow \mathcal{N}(\mu, \sigma),\ \textit{alors } \mathbb{E}(X) = \mu \textit{ et } Var(X) = \sigma^2.}$

Preuve : laissée en exercice. Indication : considérer d'abord le cas de $Y = \frac{1}{\sigma}(X - \mu) \rightsquigarrow \mathcal{N}(0, 1)$. □

Exercice : Que pouvez-vous dire de $\mathbb{P}(\{X = 21\})$ et $\mathbb{P}(\{X \leq 21\})$ pour X suivant chacune des trois lois de la figure 4.1.

4.3 Variance et indépendance

Proposition 4.9 *Soient $X_1, \ldots, X_n$ n v.a.. Alors*

$$\boxed{Var(X_1 + \ldots + X_n) = Var(X_1) + \ldots + Var(X_n) \quad \textit{si les v.a. } X_1, \ldots, X_n \textit{ sont indépendantes.}}$$

Preuve : Posons $Y_i := X_i - \mathbb{E}(X_i)$ de telle sorte que $\mathbb{E}(Y_i) = \mathbb{E}(X_i - \mathbb{E}(X_i)) = \mathbb{E}(X_i) - \mathbb{E}(\mathbb{E}(X_i)) = \mathbb{E}(X_i) - \mathbb{E}(X_i) = 0$. A présent, nous voyons que

$$\begin{aligned}\mathrm{Var}\,(X_1 + \ldots + X_n) &= \mathbb{E}((X_1 + \ldots + X_n - \mathbb{E}(X_1 + \ldots + X_n))^2) = \mathbb{E}((X_1 - \mathbb{E}(X_1) + \ldots + X_n - \mathbb{E}(X_n))^2) \\ &= \mathbb{E}((Y_1 + \ldots + Y_n)^2) = \mathbb{E}\left(\sum_{i,j=1}^{n} Y_i Y_j\right) = \sum_{i,j=1}^{n} \mathbb{E}(Y_i Y_j).\end{aligned}$$

C'est à présent que nous utilisons l'hypothèse que pour $i \neq j$ les v.a. Y_i et Y_j sont indépendantes (puisque fonction chacune d'une des v.a. indépendantes X_i et X_j) et donc $\mathbb{E}(Y_i Y_j) = \mathbb{E}(Y_i)\mathbb{E}(Y_j) = 0 \cdot 0 = 0$, puisque $\mathbb{E}(Y_i) = 0 = \mathbb{E}(Y_j)$. Donc, dans la sommes $\sum_{i,j=1}^{n} \mathbb{E}(Y_i Y_j)$, ne subsistent que les termes tels que $i = j$, et on obtient

$$\mathrm{Var}\,(X_1 + \ldots + X_n) = \sum_{i=1}^{n} \mathbb{E}(Y_i Y_i) = \sum_{i=1}^{n} \mathbb{E}(Y_i^2) = \sum_{i=1}^{n} \mathbb{E}((X_i - \mathbb{E}(X_i))^2) = \sum_{i=1}^{n} \mathrm{Var}\,(X_i).$$

□

Remarque : l'hypothèse d'indépendance dans la proposition 4.9 est essentielle ; en effet, soit par exemple X une v.a. de variance non nulle ; posons $X_1 := +X$ et $X_2 := -X$. Alors

$$\begin{aligned}\mathrm{Var}\,(X_1 + X_2) &= \mathrm{Var}\,(X - X) = \mathrm{Var}\,(0) = 0 \\ &\neq 2\mathrm{Var}\,(X) = \mathrm{Var}\,(X) + \mathrm{Var}\,(X) = \mathrm{Var}\,(X) + \mathrm{Var}\,(-X) = \mathrm{Var}\,(X_1) + \mathrm{Var}\,(X_2).\end{aligned}$$

Le corollaire suivant est une version probabiliste de l'adage que veut qu'il vaille mieux mettre ses oeufs dans deux paniers :

Corollaire 4.10 *Soient X_1, X_2, ..., X_n n v.a. indépendantes ; soit $M := \frac{1}{n}(X_1 + \ldots + X_n)$ leur moyenne. Alors $Var(M) = \frac{1}{\sqrt{n}} Var(X_1)$.*

Exercice : Montrer le corollaire. Donner un modèle probabiliste étayant l'adage ; **indication :** postuler que la seule manière de briser un oeuf est de faire tomber le panier qui le contient et dans ce cas tous les oeufs du panier sont brisés. Discuter ce postulat et l'adage en termes de variance.

Exercice : Soient X et Y deux v.a. de même loi. Lorsqu'on découvre les probabilités et la notion (essentielle) de loi d'une v.a. il vient immanquablement le moment où l'on se pose la question de savoir si pour deux v.a. X et Y avoir la même loi est synonyme d'être égales... Qu'en pensez-vous ? Montrer que si X et Y sont indépendantes, $\mathrm{Var}\,(X-Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y)$. Voyez-vous le lien avec l'interrogation évoquée juste avant ? Si l'on renonce à l'hypothèse d'indépendance, peut-on avoir $\mathrm{Var}\,(X-Y) > \mathrm{Var}\,(X) + \mathrm{Var}\,(Y)$? (N'hésitez pas à supposer que $\mathbb{E}(X) = 0 = \mathbb{E}(Y)$). Il serait naturel que vous vous demandiez alors si la variance $\mathrm{Var}\,(X-Y)$ peut être arbitrairement grande ; la réponse est non : au chapitre suivant, l'inégalité de Cauchy-Schwarz nous permettra de montrer que $\mathrm{Var}\,(X - Y) \leq (\sigma(X) + \sigma(Y))^2$.

Chapitre 5

Expression et mesure de l'interdépendance

Nous considérons ici le cas d'un vecteur àléatoire (vct.a.) à deux dimensions $Z = (X, Y) \in \mathbb{R}^2$, pour éviter les lourdeurs du type $Z = (Z_1, \ldots, Z_d) \in \mathbb{R}^d$, (qu'on appelle aussi v.a. de $\mathbb{R}^d$). Ce que nous développerons pour la v.a. X s'adapte généralement facilement à la v.a. Y ; dans ce cas, nous laisserons à la sagacité de la lectrice (ou du lecteur) le soin d'assurer cette adaptation.

5.1 Composantes d'un vecteur aléatoire

Soit $Z = (X, Y)$ un vecteur aléatoire (vct.a.) sur Ω, et $z_0 = (x_0, y_0) \in \mathbb{R}^2$. Rappelons que $\{Z \leq z_0\}$ désigne l'évènement $\{X \leq x_0, Y \leq y_0\} := \{\omega \in \Omega | X \leq x_0, Y \leq y_0\}$; la fonction de répartition (ou loi) de Z est la fonction $F_Z : \mathbb{R}^2 \longrightarrow [0, 1]$ définie par $F_Z(z_0) := \mathbb{P}(\{Z \leq z_0\})$. Les composantes X et Y sont également des v.a., à valeurs dans $\mathbb{R}$. Elles ont donc chacune également une loi $F_X(x_0) := \mathbb{P}(\{X \leq x_0\})$ et $F_Y(y_0) := \mathbb{P}(\{Y \leq y_0\})$.

Définition : Les fonctions de répartition F_X et F_Y s'appellent les *lois marginales* du vecteur aléatoire $Z = (X, Y)$. La fonction de répartition $F_Z = F_{(X,Y)}$ s'appelle la *loi jointe* des v.a. X et Y. Les lois marginales se déduisent facilement de la loi jointe :

Proposition 5.1 $\boxed{F_X(x_0) = F_{(X,Y)}(x_0, +\infty)} := \lim_{y_0 \to +\infty} F_Z(x_0, y_0)$ *et* $\boxed{F_Y(y_0) = F_{(X,Y)}(+\infty, y_0)}$.

Preuve : $F_X(x_0) = \mathbb{P}(\{X \leq x_0\}) = \mathbb{P}(\{X \leq x_0, Y < +\infty\}) =^1 \lim_{y_0 \to +\infty} \mathbb{P}(\{X \leq x_0, Y \leq y_0\}) = \lim_{y_0 \to +\infty} F_Z(x_0, y_0)$. On procède de manière similaire pour $F_Y(y_0)$. □

En revanche, sans hypothèse complémentaire, il n'est pas possible de résoudre le problème de retrouver la loi jointe à partir des lois marginales. L'indépendance des v.a. est une hypothèse qui donne une solution à ce problème.

Proposition 5.2 *Supposons que les v.a. X et Y soient* indépendantes. *Alors on retrouve la loi jointe à partir des lois marginales par la formule :* $\boxed{F_{(X,Y)}(x_0, y_0) = F_X(x_0) F_Y(y_0)}$. *En particulier*

- *Si (X, Y) est élémentaire, alors* $p_{z_0} := \boxed{\mathbb{P}(\{X = x_0, Y = y_0\}) = \mathbb{P}(\{X = x_0\})\mathbb{P}(\{Y = y_0\})} =: p_{x_0} p_{y_0}$
- *Si X et Y admettent des densités f_X et f_Y, alors* $\boxed{f_{(X,Y)}(x, y) = f_X(x) f_Y(y)}$.

Preuve : Comme X et Y sont indépendantes, les évènements $\{X \leq x_0\}$ et $\{Y \leq y_0\}$ sont indépendants. On en déduit que

$$F_{(X,Y)}(x_0, y_0) := \mathbb{P}(\{X \leq x_0, Y \leq y_0\}) = \mathbb{P}(\{X \leq x_0\} \cap \{Y \leq y_0\}) = \mathbb{P}(\{X \leq x_0\})\mathbb{P}(\{Y \leq y_0\}) =: F_X(x_0) F_Y(y_0).$$

Si les v.a. sont élémentaires et indépendantes l'une de l'autre les évènement $\{X = x_0\}$ et $\{Y = y_0\}$ sont indépendants. Donc

$$p_{z_0} := \mathbb{P}(\{X = x_0, Y = y_0\}) = \mathbb{P}(\{X = x_0\} \cap \{Y = y_0\} = \mathbb{P}(\{X = x_0\})\mathbb{P}(\{Y = y_0\}) =: p_{x_0} p_{y_0}.$$

[1] ici nous supposons implicitement que $\mathcal{B}$ est une tribu et que la probabilité $\mathbb{P}$ est σ-additive : voir cours de L3.

Si les v.a. X et Y admettent des densités, on voit facilement que $f_{(X,Y)}(x_0,y_0) = \frac{\partial^2 F_{(X,Y)}}{\partial x \partial y}(x_0,y_0)$, $f_X(x_0) = \frac{dF_X}{dx}(x_0)$, et $f_Y(y_0) = \frac{dF_Y}{dy}(y_0)$; la dernière relation s'en déduit à partir de la première. □

5.2 Copules

Définition : Soit $C : [0,1] \times [0,1] \longrightarrow [0,1]$ telle que

$$C(0,t) = 0 = C(t,0) \text{ et } C(1,t) = t = C(t,1) \text{ pour tout } t \in [0,1] \tag{5.1}$$

et

$$C(u_+,v_+) - C(u_+,v_-) - C(u_-,v_+) + C(u_-,v_-) \geq 0 \tag{5.2}$$

pour tous $0 \leq u_- \leq u_+ \leq 1$ et $0 \leq v_- \leq v_+ \leq 1$. On dit que les v.a. X et Y *admettent C pour copule* si et seulement si pour tout $(x_0,y_0) \in \mathbb{R}^2$ on a

$$F_{(X,Y)}(x_0,y_0) = C(F_X(x_0), F_Y(y_0)). \tag{5.3}$$

Théorème 5.3 (Abe Sklar, 1959) *Tout couple (X,Y) de v.a. admet une copule C (au moins).*

Preuve : Nous donnons la preuve dans le cas où les v.a. X et Y admettent des fonctions de répartition continues strictement croissantes, $F_X : [x_-,x_+] \longrightarrow [0,1]$ et $F_Y : [y_-,y_+] \longrightarrow [0,1]$, avec $x_- := \text{Inf}\{x_0 | F_X(x_0) \in]0,1[\}$ et $x_+ := \text{Sup}\{x_0 | F_X(x_0) \in]0,1[\}$ et similaire pour y_- et y_+ (ces nombres étant possiblement égaux à $\pm\infty$). Dans ce cas la relation $F_{(X,Y)}(x_0,y_0) = C(F_X(x_0), F_Y(y_0))$ pour $u = F_X(x_0)$ et $v = F_Y(y_0)$ montre qu'il faut prendre $\boxed{C(u,v) := F_{(X,Y)}(F_X^{-1}(u), F_Y^{-1}(v))}$. On montre alors que C ainsi défini a également les autres propriétés annoncées. □

Exemple : Si les v.a. X et Y sont indépendantes, on voit immédiatement qu'elles admettent la copule $C^\perp(u,v) := uv$. Les fonctions $C^-(u,v) := \max\{u+v-1, 0\}$ et $C^+(u,v) := \min\{u,v\}$, appelées *copules extrèmes de Fréchet*, ont la propriété que pour toute copule C, on a $C^-(u,v) \leq C(u,v) \leq C^+(u,v)$ pour tout $(u,v) \in [0,1]^2$.

Exercice : Montrer que les copules extrèmes de Fréchet sont bien des copules. Calculer la loi de v.a. uniformes sur $[0,1]$ dont la loi jointe admet cette copule. Calculer dans ce cas la covariance (voir ci-dessous) de ces v.a..

Exercice : Montrer que la *copule logistique de Gumbel* $C(u,v) := \frac{uv}{u+v-uv}$ est bien une copule.

Proposition 5.4 *Soient h et k deux fonctions strictement croissantes de $\mathbb{R}$ dans $\mathbb{R}$. Soient X et Y deux v.a., $X' := h(X)$ et $Y' := k(Y)$. Si le couple de v.a. (X,Y) admet la copule C, alors le couple (X',Y') admet également la copule C.*

Exercice : Montrer la proposition 5.4.

5.3 Covariance

Définition : Soient X et Y deux v.a. ; on appelle *covariance* de X et Y et on note $\text{Cov}(X,Y)$ le nombre $\boxed{\text{Cov}(X,Y) := \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))}$.

On appelle *corrélation* de X et Y et on note $\rho(X,Y)$ le nombre $\boxed{\rho(X,Y) := \dfrac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}}$.

Exemple : Soient X une v.a. et $\mathcal{E}$ une v.a. *centrée*, c'est-à-dire telle que $\mathbb{E}(\mathcal{E}) = 0$; on *suppose* que X et $\mathcal{E}$ sont *indépendantes*. Soient a et b deux réels ; posons $Y := aX + b$ et $Z := aX + b + \mathcal{E}$; nous voyons dans Y une fonction linéaire-affine de la grandeur aléatoire X et dans Z sa valeur perturbée par des erreurs de mesure $\mathcal{E}$ indépendantes de X. Alors $\boxed{\mathrm{Cov}\,(Z, X) = a\mathrm{Var}\,(X)} = \mathrm{Cov}\,(aX + b, X)$; en d'autres termes, dans ce cas $\mathrm{Cov}\,(X, Z)$ ne dépend pas de "l'erreur" $\mathcal{E}$ mais seulement de la pente a de la droite liant X et Y.

Exercice 5.1 Montrer que dans l'exemple on a bien $\mathrm{Cov}\,(Z, X) = a\mathrm{Var}\,(X)$. Soit $Z_n := aX + b + \mathcal{E}_n$, où $\mathcal{E}_n = \frac{1}{n}\mathcal{E}$. On suppose que $a\mathrm{Var}\,(X) \neq 0$; calculer $\rho(Z_n, X)$ et montrer que $\lim_{n\longrightarrow\infty} \rho(Z_n, X) = 1$.

Proposition 5.5 (inégalité de Cauchy-Schwarz) $\boxed{\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)}}$.

Preuve : Il suffit de remarquer que la fonction $\lambda \mapsto \mathbb{E}((\lambda X - Y)^2)$ est un polynôme du second degré qui n'est jamais négatif : son discriminant $\Delta = 4(\mathbb{E}(XY))^2 - 4\mathbb{E}(X^2)\mathbb{E}(Y^2)$ est donc négatif. L'inégalité de Cauchy-Schwarz en découle immédiatement. □

Proposition 5.6 *L'application $(X, Y) \mapsto Cov(X, Y)$ est bilinéaire et positive. La corrélation est à valeurs dans $[-1, 1]$. Si $\rho(X, Y) = \pm 1$ il existe a et b tel que $y = aX + b$ p.s. avec $sgn\,(a) = \rho(X, Y)$.*

Remarque : La preuve de ce résultat est sans difficulté ; la positivité tient au fait $\mathrm{Cov}\,(X, X) = \mathrm{Var}\,(X) \geq 0$. Le cas de $\rho(X, Y) = \pm 1$ fait l'objet de l'exercice 5.4

Il est intéressant de noter que Cov n'est pas définie-positive ; en effet, comme nous l'avons vu, le fait que $\mathrm{Var}\,(X) = 0$ n'entraine pas que $X = 0$, mais seulement que $X = \mathrm{Cste}(= \mathbb{E}(X))$ presque sûrement. En fait, il y a bien un produit scalaire derrière la notion de covariance : il s'agit de $< X, Y >:= \mathbb{E}(XY)$, à condition "d'assimiler[2]" deux variables X_1 et X_2 dès lors qu'elles sont égales presque-sûrement ($\mathbb{P}(\{X_1 = X_2\} = 1)$. Dans ce cas les v.a. d'espérance nulle forment un sous-espace de codimension 1, et $X \mapsto X' := X - \mathbb{E}(X)$ est la *projection orthogonale* sur ce sous-espace. L'écart-type $\sigma(X) := \sqrt{\mathrm{Var}\,(X)} =: \|X'\|$ est la *norme* de X' au sens de ce produit scalaire, et la covariance de X et Y n'est alors rien d'autre que le produit scalaire des projections X' et Y'. Dans cet esprit, la corrélation $\rho(X, Y)$ n'est autre que le cosinus de l'angle θ entre X' et Y' dans le sens $< X', Y' >= \|X'\|\|Y'\| \cos(\theta)$.

5.4 Exercices

Exercice 5.2 Cas de deux v.a. élémentaires Soient X et Y deux v.a. élémentaires dont les lois sont données par

x	1	2	3
$\mathbb{P}(\{X = x\})$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

, et

y	0	2	4
$\mathbb{P}(\{Y = y\})$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

On se propose de déterminer la loi jointe $F_Z(z)$ de $Z = (X, Y)$ et la copule couplant les lois de X et de Y dans deux situations distinctes. On pose $\mathcal{X} := X(\Omega)$ et $\mathcal{Y} = Y(\Omega)$.

1. On suppose que X et Y sont *indépendantes*. Former successivement les tableaux donnant les $p_z := \mathbb{P}(\{X = x, Y = y\})$, $F_Z(x, y) := \mathbb{P}(\{X \leq x, Y \leq y\})$, et $C(u, v)$ tels que $F_Z(x, y) = C(F_X(x), F_Y(y))$, pour tous $z := (x, y) \in \mathcal{X} \times \mathcal{Y}$, et tous $(u, v) \in F_X(\mathcal{X}) \times F_Y(\mathcal{Y})$.
   Que vaut $\mathrm{Cov}\,(X, Y)$?
2. On suppose à présent que les lois de X et Y sont couplées par la copule $C(u, v) := C^-(u, v) := \mathrm{Max}\,(u + v - 1, 0)$. Former successivement les tableaux donnant les $C(u, v)$, $F_Z(x, y)$, et p_z, pour tous $(u, v) \in C(F_X(\mathcal{X}), F_Y(\mathcal{Y}))$ et tout $z := (x, y) \in \mathcal{X} \times \mathcal{Y}$. Calculer la covariance $\mathrm{Cov}\,(X, Y)$ et la corrélation $\rho(X, Y)$ dans ce cas.

Exercice 5.3 Cas de deux v.a. continues Soient X et Y deux v.a. dont les lois sont données par $f_X(x) = \frac{1}{3}\mathbb{I}_{[0,3]}$ et $f_Y(y) = \frac{1}{2}\mathbb{I}_{[0,2]}$. On se propose de déterminer la loi jointe $F_Z(z)$ de $Z = (X, Y)$ et la copule couplant les lois de X et de Y dans deux situations distinctes. On pose $\mathcal{X} := X(\Omega)$ et $\mathcal{Y} = Y(\Omega)$.

[2] La relation $X_1 \sim X_2$ si et seulement si $\mathbb{P}(\{X_1 = X_2\}) = 1$ est une relation d'équivalence ; on considère le *quotient* de l'ensemble L^2 des v.a. telles que $\mathbb{E}(X^2)$ existe par cette relation d'équivalence.

1. On suppose que X et Y sont *indépendantes*. Calculer successivement la densité $f_Z(x,y)$, la fonction de répartition $F_Z(x,y) := \mathbb{P}(\{X \leq x, Y \leq y\})$, et la copule $C(u,v)$ (telle que $F_Z(x,y) = C(F_X(x), F_Y(y))$), pour tous $z := (x,y) \in \mathcal{X} \times \mathcal{Y}$, et tous $(u,v) \in [0,1] \times [0,1]$.
   Que vaut $\mathrm{Cov}\,(X,Y)$?
2. On suppose à présent que les lois de X et Y sont couplées par la copule $C(u,v) := C^+(u,v) := \mathrm{Min}\,(u,v)$. Montrer que[3] $F_Z(x,y) = ((0 \vee (\frac{x}{3} \wedge \frac{y}{2})) \wedge 1$.
   Représenter sur un schema du plan $\mathbb{R}^2_{x,y}$ sur quelles région la fonction F_Z est égale respectivement à 0, 1, $\frac{x}{3}$, et $\frac{y}{2}$, et montrer que si la loi de $Z := (X,Y)$ admettait une densité f_Z on aurait $f_Z(x,y) = 0$ "presque-partout".
   Montrer que $Z := (X,Y)$ a même loi que $Z' := (X, \frac{2}{3}X)$.
   En déduire la covariance $\mathrm{Cov}\,(X,Y)$ et la corrélation $\rho(X,Y)$ dans ce cas.

Exercice 5.4 Soient X_0 et Y_0 tels que $\mathbb{E}(X_0) = 0 = \mathbb{E}(Y_0)$ et $\mathrm{Var}\,(X_0) = 1 = \mathrm{Var}\,(Y_0)$.

1. Montrer que si $\rho(X_0, Y_0) = 1$, alors $\mathrm{Var}\,(X_0 - Y_0) = 0$.
2. Soit Z tel que $\mathbb{E}(Z) = 0 = \mathrm{Var}\,(Z)$; on suppose d'abord que $Z(\Omega) \subseteq \mathbb{N}$. Montrer que $\mathbb{P}(\{Z \neq 0\}) = 0$.
3. Soit toujours Z tel que $\mathbb{E}(Z) = 0 = \mathrm{Var}\,(Z)$. On suppose à présent que Z est absolument continue, c'est-à-dire qu'il existe une fonction $f_Z : \mathbb{R} \longrightarrow \mathbb{R}^+$, la densité de Z, telle que $\mathbb{P}(\{X \in [a,b]\}) = \int_a^b f_Z(z)dz$ pour tout $a \leq b$. Montrer que pour tout $a \leq b$ tels que $0 \notin [a,b]$ on a $\mathbb{P}(\{X \in [a,b]\}) = 0$; on dit que $Z = 0$ *presque-sûrement*, et on écrit "$Z = 0$ p.s.".
4. Soient X et Y tels que $\rho(X,Y) = \pm 1$. Trouver a et b tels que $Y = aX + b$; vérifier que $\mathrm{sgn}\,(a) = \rho(X,Y)$.

Maurice Fréchet (1878-1973) :   : Abe Sklar ()

[3] On note $a \wedge b := \mathrm{Min}\,(a,b)$ et $a \vee b := \mathrm{Max}\,(a,b)$.

Chapitre 6

Loi des grands nombres

Avec ce chapitre nous abordons le point essentiel du paradigme de l'application du calcul des probabilité qui a donné son nom à ce calcul : la relation qu'il y a entre l'observation du nombre des succès lors de "répétitions indépendantes" d'une expérience à l'issue incertaine, tel l'obtention d'une "face" dans un tirage à pile ou face, et la "chance de succès", dite "probabilité de réalisation de l'évènement". Nous commençons par deux inégalités qui nous serviront dans la preuve de ce point essentiel

6.1 Les inégalités de Markov et Bienaymé-Tchébicheff

Au chapitre 4 nous avons dit que la variance est une mesure de la dispertion de la loi d'un v.a. X loin de son espérance. L'inégalité de Bienaymé-Tchebicheff[1] précise ce point. La preuve de cette inégalité est "diaboliquement simple", l'idée se généralisant directement à une situation un peu plus générale dite "inégalité de Markov".

Proposition 6.1 (inégalité de Markov) *Soit Y une v.a. quelconque ; pour toute fonction $h : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$, $a \mapsto h(a)$, croissante strictement positive pour $a > 0$ telle que $\mathbb{E}(h(Y)) < +\infty$, et pour tout $a > 0$ on a*

$$\mathbb{P}(\{|Y| \geq a\}) \leq \frac{1}{h(a)}\mathbb{E}(h(|Y|)). \tag{6.1}$$

Preuve : L'idée tient dans le fait de choisir l'écriture linéaire de la probabilité, à savoir $\mathbb{P}(\{|Y| \geq a\}) = \mathbb{E}(\mathbb{I}_{\{|Y|\geq a\}})$ et de remarquer que, comme h est croissante positive, on a $h(a)\mathbb{I}_{\{|Y|\geq a\}}) \leq h(|Y|)\mathbb{I}_{\{|Y|\geq a\}})(\leq h(|Y|))$. La suite découle de la linéarité et de la positivité de l'espérance :

$$\mathbb{P}(\{|Y| \geq a\}) = \mathbb{E}(\mathbb{I}_{\{|Y|\geq a\}}) = \frac{1}{h(a)}\mathbb{E}(h(a)\mathbb{I}_{\{|Y|\geq a\}}) \leq \frac{1}{h(a)}\mathbb{E}(h(|Y|)\mathbb{I}_{\{|Y|\geq a\}}) \leq \frac{1}{h(a)}\mathbb{E}(h(|Y|))$$

□

En appliquant ce résultat à $Y := X - \mathbb{E}(X)$, en posant $h(x) := x^2$ et $a := \lambda^2$ nous obtenons l'inégalité de Bienaymé-Tchebicheff :

Proposition 6.2 *Pour toute v.a. $X \in L^2(\Omega)$ on a*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) \leq \frac{1}{\lambda^2} Var(X) \tag{6.2}$$

6.2 La Loi des Grands Nombres (LGN)

Théorème 6.3 *Soit $(X_i)_{i\geq 0}$ une suite de v.a. i.i.d., ayant une espérance μ et un écart-type. Pour chaque $n \geq 1$, soit $M_n := \frac{1}{n}(X_1 + \ldots + X_n)$ la moyenne des n premières. Alors la suite de ces moyennes M_n tend vers le nombre μ dans le sens suivant :*

$$\textit{Pour tout } \lambda > 0, \ \lim_{n\to+\infty} \mathbb{P}(\{|M_n - \mu| \geq \lambda\}) = 0. \tag{6.3}$$

[1] le français Bienaymé était ami du russe Tchebicheff (Chebyshev en transcription anglo-saxonne), tout comme du belge Quetelet- un des fondateurs de la "statistique" au sens étymologique : sciences de l'Etat (sociologie quantitative). Il est d'usage d'accoler le nom de Tchebicheff à l'inégalité de Bienaymé car c'est Tchebichev qui l'a utilisée le premier à la généralisation de la loi des Grands Nombres de Bernoulli

Preuve : Notons σ l'écart-type commun des X_i ; observons que l'espérance les M_n est également μ, et comme les v.a. X_i sont indépendantes, la variance des M_n est égale à $\frac{\sigma^2}{n}$; en effet

$$\mathbb{E}(M_n) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \text{ , et} \tag{6.4}$$

$$\mathrm{Var}\,(M_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\,(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}. \tag{6.5}$$

A présent, il suffit d'écrire l'inégalité de Tchébicheff pour M_n :

$$0 \leq \mathbb{P}(|M_n - \mu| \geq \lambda) = \mathbb{P}(|M_n - \mathbb{E}(M_n)| \geq \lambda) \leq \frac{1}{\lambda^2}\mathrm{Var}\,(M_n) \leq \frac{\sigma^2}{n\lambda^2}.$$

On conclut on observant que $\lim_{n\to+\infty}\frac{\sigma^2}{n\lambda^2} = 0$. □

Exemple : Considérons un phénomènes pouvant se produire ou non à chaque expérience, tel la sortie d'une face dans un tirage à pile ou face, ou la sortie d'un 6 dans le tirage d'un dé. On répète l'expérience de manière identique de manière à ce que le résultat d'une expérience n'influe pas sur les expériences suivantes. On compte le nombre $S(n)$ de fois où le phénomène s'est produit au cours des n premières expériences, et on forme le rapport $M(n) = S(n)/n$, égal au nombre moyen de "succès". On dira que le phénomène se produit de manière aléatoire avec la probabilité p dans les conditions de l'expérience choisies si l'on peut lui appliquer le modèle probabiliste suivant : l'apparition du phénomène lors de la i-ème expérience est une v.a. de Bernouilli $X_i \rightsquigarrow \mathcal{B}(1,p)$, les diverses v.a. X_i étant supposées indépendantes. Dans ce cas, le rapport $M(n) = S(n)/n$ observé est modélisé par la v.a. $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Que nous dit la Loi des Grands Nombres pour ce modèle ? On a $\mu = \mathbb{E}(X_i) = p\cdot 1 + (1-p)\cdot 0 = p$; donnons nous un $\lambda > 0$, par exemple $\lambda = 0.01$. Plus n est grand, plus il est improbable que $M_n \notin]p-\lambda, p+\lambda[$. Si le phénomène considéré est effectivement aléatoire (dans le conditions de l'expérience), les valeurs observées du rapport $M(n)$ doivent donc, au fur et à mesure que n augmente, "se grouper autour" d'une valeur $\hat{p}$, sans que cela n'exclue de "petite excursions" hors de $]\hat{p}-\lambda, \hat{p}+\lambda[$, celles-ci "devenant de plus en plus rares".

Si tel est le cas, on considérera que "le phénomène se produit avec la probabilité $\hat{p}$".

L'exemple qui précède correspond à la Loi des Grands Nombre, telle qu'elle a été découverte par Jacques Bernouilli, pour le cas particulier de v.a. de Bernoulli, précisément, et qui constitue le *théorème de Bernoulli*. Ce cas à l'avantage de la simplicité, la loi commune des v.a. se réduisant au choix d'un unique paramètre p, que la Loi de Grands Nombre révèle par la *limite en probabilité* des $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. C'est cette application du théorème de Bernoulli qui fonde l'utilisation du calcul (abstrait) des probabilités comme cadre mathématique de la Statistique, littéralement "science des Etats", ou "Physique sociale" pour reprendre le nom d'un livre d'Adolphe Quételet (Bruxelles, 1869).

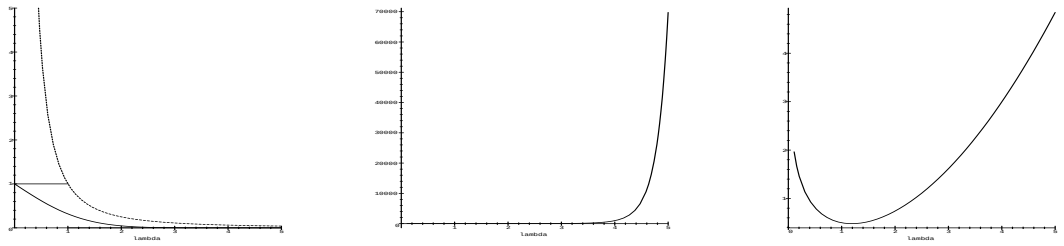
6.3 Portée réelle de ces résultats

6.3.1 Que nous apprennent vraiment l'inégalités de Bienaymé-Tchébicheff ?

Sur le premier graphique de la figure 6.1, nous avons représenté simultanément la fonction $\frac{1}{\lambda^2}\mathrm{Var}\,(X)$ pour une loi de variance 1, et la probabilité $\mathbb{P}\{|X - \mathbb{E}(X)| \geq \lambda\}$ pour deux v.a. centrées réduites : $X \rightsquigarrow \mathcal{N}(0,1)$ et $X = 2Y - 1$, avec $Y \rightsquigarrow \mathcal{B}(1, 0.5)$. Nous voyons que pour λ grand, la majoration est très grossière : pour $X = 2Y - 1$, nous voyons que, sauf en $\lambda = 1^-$ où loi et majorant se confondent, nous voyons que nous majorons 0 par $\frac{1}{\lambda^2}$. Sur le deuxième graphique, pour $X \rightsquigarrow \mathcal{N}(0,1)$, nous avons représenté le rapport du majorant divisé par la fonction qu'il est "chargé de majorer" ; nous λ grand, nous voyons que ce rapport devient considérable. Pour améliorer la lisibilité, nous avons représenté dans le troisième graphique le $\log_{10}$ du rapport précédent (les valeurs représentent donc le rapport comme un exposant de 10). Nous voyons que cette inégalité est à la fois médiocre, mais ne peut être améliorée en toute généralité à cause du cas de v.a. de Bernoulli pour lesquelles la majorations se révèle la pire.

6.3.2 Loi faible et loi forte

L'énoncé 6.3 de la LGN que nous avons donné est encore appelé "loi *faible* des grands nombres". La "loi *forte*" traite de l'ensemble $\Omega_0 = \{\omega \in \Omega \mid \lim_{n\to+\infty} M_n(\omega) = \mu\}$, c'est-à-dire de l'ensemble des



(a) **(b)** **(c)**

FIG. 6.1 – L'inégalité de Bienaymé-Tchebicheff : **(a)** en pointillé : la fonction $\frac{1}{\lambda^2}\mathrm{Var}\,(X)$ pour une loi de variance 1, en ligne continue la probabilité $\mathbb{P}\{|X - \mathbb{E}(X)| \geq \lambda\}$ pour deux v.a. centrées réduites : $X \rightsquigarrow \mathcal{N}(0,1)$ et $X = 2Y - 1$, avec $Y \rightsquigarrow \mathcal{B}(1, 0.5)$. Nous voyons que pour λ grand, la majoration est très grossière. **(b)** Pour $X \rightsquigarrow \mathcal{N}(0,1)$, quotient du majorant $\frac{1}{\lambda^2}\mathrm{Var}\,(X)$ par la valeur exacte $\mathbb{P}\{|X - \mathbb{E}(X)| \geq \lambda\}$. **(c)** $\log_{10}$ du quotient précédent.

états du monde pour lesquels la moyenne $M_n(\omega)$ tend effectivement vers l'espérance commune μ des v.a. X_i. Notons que pour avoir une suite infinie de v.a. indépendantes, Ω doit nécessairement être infini, et le plus simple pour avoir une simple suite de v.a. de Bernouilli, $\Omega = \{0,1\}^{\mathbb{N}^*}$, n'est pas dénombrable. Nous quittons donc le cadre élémentaire que nous nous sommes fixé pour ce cours. Indiquons néanmoins le résultat de la loi forte des grands nombres : le sous-ensemble Ω_0 ci-dessus est de probabilité égale à 1 ; en d'autres termes, les ω pour lesquels on n'a pas la convergence souhaitée forment un *ensemble négligeable*, c'est-à-dire de probabilité nulle. On dit aussi que la convergence $\lim_{n\to+\infty} M_n = \mu$ est *presque-sûre*, et on note $Y_n \xrightarrow{ps} \overline{Y}$ si la convergence $\lim_{n\to+\infty}(Y_n - \overline{Y}) = 0$ est presque-sûre.

6.3.3 Convergence en probabilité

L'énoncé 6.3 de la loi des grands nombres que nous avons donné s'exprime encore par la locution "la suite des M_n *tend en probabilité* vers le nombre μ". De façon générale, on dit que la suite de v.a. $(Y_n)_{n\geq 1}$ *tend en probabilité vers la v.a.* $\overline{Y}$, et on note $Y_n \xrightarrow{\mathcal{P}} \overline{Y}$ si et seulement si

$$\text{pour pour } \varepsilon > 0,\ \lim_{n\to+\infty} \mathbb{P}(\{|Y_n - \overline{Y}| \geq \varepsilon\}) = 0\ ;$$

en paraphrasant : quand n devient grand, il est de plus en plus improbable que $Y_n(\omega)$ s'écarte de $\overline{Y}(\omega)$ de plus de ε. On montre que la convergence presque-sûre implique toujours la convergence en probabilité.

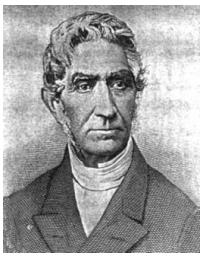
6.3.4 Et notre norme L^2 ?

Nous avons vu que $X \mapsto \|X\| := (\mathbb{E}(X^2))^{\frac{1}{2}}$ est une norme sur $L^2(\Omega)$. Nous avons donc la notion de convergence usuelle dans un espace normé, celle qui assure que $\lim_{n\to+\infty} \|Y_n - \overline{Y}\| = 0$: on dit dans ce cas que Y_n *tends vers* $\overline{Y}$ *dans* $L^2(\Omega)$, et on note $Y_n \xrightarrow{L^2} \overline{Y}$.

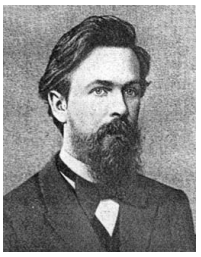
On montre que $Y_n \xrightarrow{L^2} \overline{Y}$ implique que $Y_n \xrightarrow{\mathcal{P}} \overline{Y}$. Par ailleurs, on montre que $Y_n \xrightarrow{ps} \overline{Y}$ implique que $Y_n \xrightarrow{L^2} \overline{Y}$. Nous voyons donc que la *convergence en probabilité* "$\xrightarrow{\mathcal{P}}$" est la plus faible de ces notions de convergence ; toutefois on montre aussi que si $Y_n \xrightarrow{\mathcal{P}} \overline{Y}$, alors il existe une sous-suite $k \mapsto n_k$ telle que $Y_{n_k} \xrightarrow{ps} \overline{Y}$.

Jacques Bernoulli (1654-1705) : 

Irénée-Jules Bienaymé (1796-1878) : 

Lambert Adolphe Jacques Quetelet (1796-1874) 

Pafnuty Lvovich Tchebicheff [Chebyshev] (1821-1894) : 

Andrei Andreyevich Markov (1856-1922) : 

Chapitre 7

Fonctions génératrices

7.1 Fonction génératrice des probabilités d'une v.a. entière

Soit $X \rightsquigarrow \mathcal{B}(n,p)$ une v.a. binômiale ; on a $p_X(k) := \mathbb{P}\{X=k\} = \binom{n}{k} p^k(1-p)^{n-k}$. On observe que les nombres $p_X(k)$ sont précisément les coefficients du polynôme (de la variable s) suivant :

$$G_X(s) := ((1-p)+ps)^n = \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k}s^k = \sum_{k=0}^{n} p_X(k)s^k = \mathbb{E}(s^X).$$

Cette dernière expression, $\mathbb{E}(s^X)$, se prête à la généralisation :

Définition : Soit X une v.a. entière. On appelle *fonction génératrice des probabilités* (fgp) de X, et on note G_X, la fonction

$$G_X(s) := \mathbb{E}(s^X) = \sum_{k=0}^{\infty} p_X(k)s^k, \text{ toujours avec } p_X(k) := \mathbb{P}(\{X=k\}).$$

Remarques :

- Comme $\sum_{k=0}^{\infty} p_X(k) = 1$, le rayon de convergence R_G de la série entière $\sum_{k=0}^{\infty} p_X(k)s^k$ est au moins égal à 1.
- Si X et Y ont même fgp, elles ont même loi. En effet

$$\mathbb{P}\{X=k\} = p_X(k) = \frac{G_X^{(k)}(0)}{k\ !} = \frac{G_Y^{(k)}(0)}{k\ !} = p_Y(k) = \mathbb{P}\{y=k\}.$$

Théorème 7.1 *Si $\mathbb{E}X^r$, le moment d'ordre r de la v.a. X, existe, la dérivé $G_X^{(r)}$ de la fgp de X vérifie*

$$G_X^{(r)}(1) = \mathbb{E}[X(X-1)\dots(X-r+1)], \tag{7.1}$$

où, lorsque $R_G = 1$, $G_X^{(r)}(1)$ désigne $\lim_{s\to 1^-} G^{(r)}(s) = \lim_{s\to 1^-} \sum_{k=0}^{\infty} \frac{d^r}{ds^r}(p(k)s^k)$.

Comme on sait seulement que $R_G \geq 1$, on est assuré de la dérivabilité de G_X que pour $|s|<1$. Comme $G_X(1)$ est toutefois bien défini, le résultat suivant s'applique ; il servira dans la preuve du théorème 7.1.

Lemme 7.2 (Abel) *Si la série entière de $f(x) := \sum_{n=0}^{\infty} a_n x^n$ converge pour tout $|x|<1$, et si $\sum_{n=0}^{\infty} a_n$ converge, alors $\lim_{x\to 1^-} f(x) = \sum_{n=0}^{\infty} a_n = f(1)$.*

Preuve : Comme les dérivées terme à terme d'une série entière ont le même rayon de convergence que cette série entière, le rayon de convergence de la série de $G_X^{(r)}(s)$ pour $|s|<R_G$ est lui aussi, au moins égal à 1. Comme $\mathbb{E}X^r$ existe, il en est de même de $\mathbb{E}X^s$ pour $0\leq s\leq r$, et donc $\mathbb{E}[X(X-1)\dots(X-r+1)]$ existe bien. Montrons le théorème pour $r=1$; le cas général se montrerait de même par récurrence. Pour $|s|<1$, on a, en dérivant terme à terme,

$$G'_X(s) = \sum_{k=0}^{\infty} \frac{d}{ds}(p(k)s^k) = \sum_{k=0}^{\infty} kp(k)s^{k-1},$$

d'où, par le lemme d'Abel 7.2 , $\lim_{s\to 1^-} G'_X(s) = \sum_{k=0}^{\infty} kp(k) = \mathbb{E}X$. □

Exemples :

– *Fonction génératrice des probablités d'une v.a. de Poisson* $X \rightsquigarrow \mathcal{P}(\lambda)$:
On a $\mathbb{P}\{X=k\} = e^{-\lambda}\frac{\lambda^k}{k\,!}$, donc

$$G_X(s) = \mathbb{E}s^X = \sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k\,!}s^k = e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda s)^k}{k\,!} = e^{-\lambda}e^{\lambda s} = e^{\lambda(s-1)}.$$

D'où $G'_X(s) = \lambda e^{-\lambda(s-1)}$, et donc $\mathbb{E}X = G'_X(1) = \lambda$. Plus généralement, $G^{(r)}_X(s) = \lambda^r e^{-\lambda(s-1)}$, et donc $\mathbb{E}[X(X-1)\dots(X-r+1)] = G^{(r)}_X(1) = \lambda^r$. En particulier $\lambda^2 = \mathbb{E}[X(X-1)] = \mathbb{E}X^2 - \mathbb{E}X$, donc $\mathbb{E}X^2 = \lambda^2 + \lambda$, d'où $\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

– *Fonction génératrice des probablités d'une v.a. de loi géométrique* $X \rightsquigarrow \mathcal{G}(\alpha)$:
On a $\mathbb{P}\{X=k\} = (1-\alpha)\alpha^{k-1}$, $k = 1, 2, \dots$, donc, comme $\sum_{k=1}^{\infty} q^k = \frac{q}{1-q}$,

$$G_X(s) = \mathbb{E}s^X = \sum_{k=1}^{\infty}(1-\alpha)\alpha^{k-1}s^k = \frac{1-\alpha}{\alpha}\sum_{k=1}^{\infty}(\alpha s)^k = (1-\alpha)\frac{s}{1-\alpha s},$$

donc $G'_X(s) = \frac{1-\alpha}{(1-\alpha s)^2}$, et $\mathbb{E}X = G'_X(1) = \frac{1-\alpha}{(1-\alpha)^2} = \frac{1}{1-\alpha}$; par ailleurs

$$\mathbb{E}X^2 - \mathbb{E}X = G''_X(s) = (1-\alpha)\frac{+2\alpha}{(1-\alpha)^3} = \frac{2\alpha}{(1-\alpha)^2},$$

d'où $\mathbb{E}X^2 = \frac{2\alpha}{(1-\alpha)^2} + \frac{1}{1-\alpha} = \frac{1+\alpha}{(1-\alpha)^2}$, et

$$\operatorname{Var} X = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{1+\alpha}{(1-\alpha)^2} - \frac{1}{(1-\alpha)^2} = \frac{\alpha}{(1-\alpha)^2}.$$

Proposition 7.3 *Si X et Y sont deux v.a. indépendantes, $G_{X+Y}(s) = G_X(s)G_Y(s)$.*

Preuve : Comme X et Y sont indépendantes, il en est de même pour les v.a. s^X et s^Y, d'où

$$G_{X+Y}(s) = \mathbb{E}[s^{X+Y}] = \mathbb{E}[s^X s^Y] = \mathbb{E}[s^X]\mathbb{E}[s^Y] = G_X(s)G_Y(s).$$

□

Exemples :

– *V.a. de Bernoulli :* on sait que la loi binômiale est la loi de la somme de n v.a de Bernoulli indépendantes : ceci explique que la fgp d'une v.a. binômiale, $G(s) = ((1-p)+ps)^n$, est la puissance n-ième de la fgp $(1-p)+ps$ d'une v.a. de Bernoulli.

– *Somme aléatoire de v.a. i.i.d. :* Soit N, X_1, …, X_k, …des v.a. entières indépendantes, les X_k étant identiquement distribuées. Notons respectivement G_N et G_X les fgp de N et des X_k. Soit $T := \sum_{k=1}^{N} X_k$ la somme aléatoire des X_k (un modèle possible pour le coût annuel des sinistres pour une compagnie d'assurance : N désigne le nombre de sinistres par an, et X_k représente le coût du k-ième). Alors $G_T = G_N \circ G_X$. En effet, $G_T(s) = \mathbb{E}(\mathbb{E}(s^T|N)) = \mathbb{E}g(N)$, avec

$$g(n) := \mathbb{E}(s^T|N=n) = \mathbb{E}\left(s^{\sum_{k=1}^n X_k}|N=n\right) \overset{\sharp}{=} \mathbb{E}\left(s^{\sum_{k=1}^n X_k}\right) \overset{\flat}{=} \prod_{k=1}^{n}\mathbb{E}(s^{X_k}) = (G_X(s))^n,$$

où l'égalité ♯ résulte de l'indépendance de $X_1, \dots, X_n$ de N, et l'égalité ♭ résulte de l'indépendance mutuelle des $X_1, \dots, X_n$. En posant $y := G_X(s)$, on a donc $G_T(s) = \mathbb{E}y^N = G_N(y) = G_N \circ G_X(s)$.

Théorème 7.4 (admis) *Soient $\overline{X}$, X_1, X_2, …une suite de v.a.. Si, pour tout $|s| \le 1$, $\lim_{n\to\infty} G_{X_n}(s) = G_{\overline{X}}(s)$, alors la suite des X_n* tend en loi *vers $\overline{X}$, c'est-à-dire que*

$$\lim_{n\to\infty}\mathbb{P}\{X_n = k\} = \mathbb{P}\{\overline{X} = k\} \text{ pour tout } k.$$

Dans l'exemple $Y_n \rightsquigarrow \mathcal{U}(1..n)$ (loi *discrète* uniforme sur $1..n$) on devine que si X est la limite des $X_n = \frac{1}{n}Y_n$, X suivra une loi *continue* uniforme $\mathcal{U}[0,1]$. Nous souhaitons donc une notion de fonction génératrice qui s'applique aussi au cas de v.a. absolument continues.

7.2 Fonction génératrice des moments (v.a. continue ou discrète)

Soit $X \rightsquigarrow \mathcal{E}(\lambda)$ une v.a. suivant une loi exponentielle. Son k-ième moment $\boxed{\mu_X(k) := \mathbb{E}X^k}$, se calcule facilement au moyen de k intégrations par parties et est donc égal à

$$\mu_X(k) = \mathbb{E}X^k = \int_0^\infty x^k \lambda e^{-\lambda x} dx = \frac{k\,!}{\lambda^k}.$$

On observe que les nombres $\frac{1}{k\,!}\mu_X(k)$ sont précisément les coefficients du développement en série entière de la fonction

$$\begin{aligned} M_X(t) &:= \frac{\lambda}{\lambda - t} = \frac{1}{1 - (\frac{t}{\lambda})} = \sum_{k=0}^{\infty} \frac{t^k}{\lambda^k} = \sum_{k=0}^{\infty} \frac{\mu_X(k)}{k\,!} t^k \\ &= \sum_{k=0}^{\infty} \frac{\mathbb{E}X^r}{k\,!} t^k = \sum_{k=0}^{\infty} \mathbb{E}\left[\frac{X^r}{k\,!} t^k\right] \overset{\natural}{=} \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{X^r}{k\,!} t^k\right] = \mathbb{E}e^{tX}, \end{aligned}$$

où l'égalité surmontée d'un bécarre $\natural$ impose la circonspection usuelle vis-à-vis du caractère infini de la somme considérée. La dernière expression, $\mathbb{E}e^{tX}$, se prête à la généralisation :

Définition : On appelle *fonction génératrice des moments* de X la fonction M_X définie par

$$\boxed{M_X(t) := \mathbb{E}e^{tX}.}$$

Théorème 7.5 *Supposons que tous les moments $\mu_X(k) := \mathbb{E}X^k$ d'une v.a. X soient définis, et que la série $\sum_{k=0}^{\infty} \frac{\mu_X(k)}{k\,!} t^k$ converge, avec un rayon de convergence R non nul. Alors, pour $|t| < R$, $\mathbb{E}e^{tX}$ existe et est égal à $\sum_{k=0}^{\infty} \frac{\mathbb{E}X^r}{k\,!} t^k$, et*

$$M_X(t) = \mathbb{E}e^{tX} = \sum_{k=0}^{\infty} \frac{\mu_X(k)}{k\,!} t^k \ , \ (|t| < R).$$

Remarques :

- Nous admettons ce théorème et retenons sa conclusion sous la forme que $\mathbb{E}e^{tX} =: M_X(t) = \sum_{k=0}^{\infty} \frac{\mathbb{E}X^r}{k\,!} t^k$ pour $|t| < R$. Sa preuve est simple dès lors qu'on dispose de la théorie de Lebesgue de l'intégrale.
- Si X est à valeurs entières positives, en posant $s = e^t$, nous voyons que $\boxed{M_X(t) = G_X(e^t)}$, puisque

  $$M_X(t) = \mathbb{E}e^{tX} = \mathbb{E}(e^t)^X = \mathbb{E}s^X = G_X(s) = G_X(e^t).$$

  Cette dernière expression est définie pour $|s| \leq 1$, donc $M_X(t)$ est définie pour $t \leq 0$ (au moins) ; les hypothèses sur la convergence de la série formée à partir des moments n'est utile que pour assurer une représentation de M_X par une série en puissances entières de t.
- Si X est absolument continue, de (fonction de) densité de probabilité (fdp) f_X, nous avons

  $$M_X(t) = \mathbb{E}e^{tX} = \int_{-\infty}^{+\infty} e^{tx} f_X(x) dx.$$

  L'opération fonctionnelle associant à une fonction $(x \mapsto f(x))$ la fonction $(t \mapsto M(-t) = \int_{-\infty}^{+\infty} f(t) e^{-tx} dx)$ s'appelle la *transformation de Laplace (bilatère)* et est utilisée dans biens des branches des Mathématiques, allant de méthodes pour l'ingénieur aux théories les plus récentes de sommation des séries divergentes qui apparaissent notamment en physique théorique.
- Par la formule de Taylor nous savons que le coefficient de t^k dans la série entière de M_X n'est autre que la valeur de $\frac{M_X^{(k)}(0)}{k\,!}$, donc $\mu_X(k) = M^{(k)}(0)$, d'où

  $$\boxed{\mathbb{E}X^k = M_X^{(k)}(0).} \tag{7.2}$$

Exemples :

– *Fonction génératrice des moments d'une v.a. uniforme* continue $X \rightsquigarrow \mathcal{U}[a,b]$:
On a $f_X(x) = \frac{1}{b-a}\mathbb{I}_{[a,b]}$, d'où $\mu_X(k) = \int_a^b \frac{x^k}{b-a}dx = \frac{b^{k+1}-a^{k+1}}{b-a}$, une expression guère commode pour calculer $M_X(t)$ à partir de sa série entière. En fait

$$M_X(t) = \mathbb{E}e^{tX} = \int e^{tx}\frac{1}{b-a}\mathbb{I}_{[a,b]}(x)dx = \frac{1}{b-a}\int_a^b e^{tx}dx = \frac{1}{b-a}\left[\frac{1}{t}e^{tx}\right]_{x=a}^{b} = \frac{e^{bt}-e^{at}}{(b-a)t}.$$

– *Fonction génératrice des moments d'une v.a. uniforme* discrete $X_n \rightsquigarrow \mathcal{U}]0..1]_{\frac{1}{n}} = \frac{1}{n}Y_n$ avec $Y_n \rightsquigarrow \mathcal{U}]1..n]$:
On a $p_{X_n}(\frac{i}{n}) = \mathbb{P}\{X = \frac{i}{n}\} = \mathbb{P}\{Y = i\} = \frac{1}{n}$, pour $i = 1..n$, d'où

$$M_{X_n}(t) = \mathbb{E}e^{tX} = \sum_{i=1}^{n}\frac{1}{n}e^{t\frac{i}{n}} = \frac{1}{n}\sum_{i=1}^{n}\left(e^{\frac{t}{n}}\right)^i = \frac{1}{n}e^{\frac{t}{n}}\frac{\left(e^{\frac{t}{n}}\right)^n - 1}{e^{\frac{t}{n}}-1} = \frac{e^t-1}{n(1-e^{-\frac{t}{n}})}.$$

– *Fonction génératrice des moments d'une v.a. exponentielle* $X \rightsquigarrow \mathcal{E}(\lambda)$:
On a $f_X(x) = \lambda e^{-\lambda x}\mathbb{I}_{[0,+\infty[}(x)$. Là encore, inutile de calculer les moments. Pour $t < \lambda$ on a

$$M_X(t) = \mathbb{E}e^{tX} = \int e^{tx}\lambda e^{-\lambda x}\mathbb{I}_{[0,+\infty[}(x) = \int_0^{+\infty}\lambda e^{(t-\lambda)x}dx = \frac{\lambda}{t-\lambda}\left[e^{(t-\lambda)x}\right]_{x=0}^{+\infty} = \frac{\lambda}{\lambda-t}.$$

– *Fonction génératrice des moments d'une v.a. normale* $X \rightsquigarrow \mathcal{N}(0,1)$:
On a $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, d'où

$$M_X(t) = \mathbb{E}e^{tX} = \int_{-\infty}^{+\infty}e^{tx}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}+tx}dx.$$

C'est une intégrale gaussienne ; il convient de mettre l'exposant de l'intégrand sous la forme $-\frac{v^2}{2}+c$. On a

$$-\frac{x^2}{2}+tx = -\frac{1}{2}(x^2-2tx) = -\frac{1}{2}\left[(x-t)^2-t^2\right] = -\frac{1}{2}(x-t)^2+\frac{t^2}{2} = -\frac{v^2}{2}+\frac{t^2}{2}$$

pour $v = x-t$, et finalement

$$M_X(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{x^2}{2}+tx}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{v^2}{2}+\frac{t^2}{2}}dv = e^{+\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}e^{-\frac{v^2}{2}}dv = e^{+\frac{t^2}{2}}.$$

On pourrait procéder de manière similaire pour $X \rightsquigarrow \mathcal{N}(\mu,\sigma)$ et obtenir

$$M_X(t) = e^{\mu t}e^{\frac{1}{2}\sigma^2t^2}.$$

Les calculs sont toutefois plus laborieux, et sont plus aisés avec le recul des propriétés de la transformation de Laplace. Notons qu'on trouve $M'_X(t) = (\mu+\sigma^2t)M_X(t)$ et $M''_X(t) = (\sigma^2+(\mu+\sigma^2t)^2)M_X(t)$, d'où $M'_X(0) = \mu$ et $M''_X(0) = \sigma^2+\mu^2$, d'où $\mathbb{E}X = M'_X(0) = \mu$, $\mathbb{E}X^2 = M''_X(0) = \sigma^2+\mu^2$, et $\mathrm{Var}\,X = \sigma^2+\mu^2-\mu^2 = \sigma^2$. On retrouve bien l'espérance et la variance d'une v.a. suivant une loi $\mathcal{N}(\mu,\sigma)$.

En raisonnant comme dans le cas de la fonction génératrice des probabilités on voit qu'on a

Proposition 7.6 *Si X et Y sont deux v.a. indépendantes, $M_{X+Y}(t) = M_X(t)M_Y(t)$.*

Théorème 7.7 (admis) *Soient $\overline{X}$, X_1, X_2, ... une suite de v.a.. Si, pour tout $|t| \le R$, $\lim_{n\to\infty} M_{X_n}(t) = M_{\overline{X}}(t)$, alors la suite des X_n* tend en loi *vers $\overline{X}$, c'est-à-dire que*

$$\lim_{n\to\infty} M_{X_n}(t) = M_{\overline{X}}(t) \text{ pour tout } t \text{ implique que } \lim_{n\to\infty} F_{X_n}(x) = F_{\overline{X}}(x) \text{ pour tout } x,$$

où $F_{\overline{X}}(x) := \mathbb{P}\{\overline{X} \le x\}$ et $F_{X_n}(x) := \mathbb{P}\{X_n \le x\}$ désignent les fonctions caractérisques des v.a. considérées. En particulier pour deux v.a. X et Y, $\boxed{\text{si on a } M_X = M_Y, \text{ alors } X \text{ et } Y \text{ ont même loi.}}$

Reprenons l'exemple $Y_n \rightsquigarrow \mathcal{U}[1..n]$ (loi *discrète* uniforme sur $1..n$). Nous avons vu que $M_{X_n}(t) = \frac{e^t-1}{n(1-e^{-\frac{t}{n}})}$. Un développement limité à l'ordre 1 du dénominateur $n(1-e^{-\frac{t}{n}}) = n(1-(1-\frac{t}{n}+\ldots))$ montre que $\lim_n M_{X_n}(t) = \frac{e^t-1}{t}$, la fonction génératrice des moments d'une v.a. continue uniforme sur $[0,1]$, ce qui montre que la suite $(Y_n)_{n\ge1}$ converge en loi vers une loi $\mathcal{U}[0,1]$.

Chapitre 8

Le theorème limite central

Avec la loi des grands nombres nous avons vu le lien qui existe entre la notion abstraite de probabilité d'un évènement, et la fréquence de l'occurence de cet évenement dans une suite de réalisations indépendantes et identiques d'une expérience pouvant provoquer cet évènement (Théorème de Bernoulli). Plus génralement, la loi des grands nombre montre que la moyenne $M_n := \frac{1}{n}\sum_{i=1}^n X_i$ d'une suite de v.a. X_i i.i.d. (ayant espérance et variance) tends vers un nombre : l'espèrance commune $\mu := \mathbb{E}(X_i)$ de ces v.a. ; bien entendu, M_n n'est que de variance petite et reste donc aléatoire : si l'on veut observer la variabilité de M_n (qui est proche de μ) il est nécéssaire d'agrandir l'écart $M_n - \mu$ en le multipliant par une grandeur $\lambda(n)$ suffisament grande avec n. Le choix de $\lambda(n)$ peut se faire simplement de manière à retrouver dans la v.a. "amplifiée" $Z_n := \lambda(n)(M_n - \mu)$ soit de variance 1. Calculons : nous voulons

$$\begin{aligned} 1 &= \mathrm{Var}\,(Z_n) = \mathrm{Var}\,(\lambda(n)(M_n-\mu)) = \lambda^2(n)\mathrm{Var}\,((M_n-\mu)) = \lambda^2(n)\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) \\ &= \frac{\lambda^2(n)}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) \text{ par indépendance des } X_i \\ &= \frac{\lambda^2(n)}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{\lambda^2(n)}{n^2}\sum_{i=1}^n \mathrm{Var}\,(X_i) = \frac{\lambda^2(n)}{n^2} n\mathrm{Var}\,(X_i) = \frac{\lambda^2(n)}{n}\sigma^2. \end{aligned}$$

Il suffit donc de choisir $\boxed{\lambda(n) := \frac{\sqrt{n}}{\sigma}}$ pour que Z_n soit de variance 1. Nous avons alors

$$Z_n := \lambda(n)(M_n-\mu) = \frac{\sqrt{n}}{\sigma}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right) = \frac{1}{\sigma\sqrt{n}}\left(\sum_{i=1}^n X_i - n\mu\right)$$

qui est visiblement d'espérance nulle. Et là se produit une petite merveille : pour n grand, cette "remise à l'échelle" de $M_n - \mu$ produit une v.a. Z_n *dont la loi est proche de la loi normale* (centrée et réduite, puisque tel est le cas pour la loi de tous les Z_n), et ceci sans autre hypothèse : ceci explique pourquoi cette loi surgit si souvent en statistique. Cette merveille porte le nom de théorème limite central[1] (central limit theorem, en anglais)

8.1 Le théorème

Théorème 8.1 (théorème limite central) *Avec les notations et les hypothèses ci-dessus,* $Z_n := \frac{1}{\sigma\sqrt{n}}(X_1 + \ldots + X_n - n\mu)$ *tend en loi vers une v.a.* Z *de loi* $\mathcal{N}(0,1)$

Preuve : L'idée de la preuve que nous allons donner est simple : nous allons montrer que la fonction génératrice des moments M_{Z_n} des Z_n tends vers la fonction génératrice des moments M_Z de Z. Le théorème 7.7 permettra alors de conclure, puis qu'il assure que ceci implique la convergence en loi des Z_n vers Z. Nous utiliserons le lemme général suivant, dont nous laissons la preuve, élémentaire, en exercice :

[1] Nous adoptons ici la traduction *thèorème limite central* de "central limit theorem" donnée par Jean Jacod et Philip Protter dans *L'essentiel en théorie des probabilités* (Cassini, Paris 2003) et qui est la traduction de leur *Probability Essentials* (Springer, Berlin et al. 2000). La traduction théorème central limit est souvent adoptée et a donné l'acronyme "TCL" pour désigner ce théorème ; celle-ci est malheureuse, car il s'agit bien ici d'un théorème limite qui a un rôle central dans la théorie : le nom original est donc à comprendre par (central (limit theorem)) et non ((central limit) theorem) ; gâre à la non-associativité en linguistique !

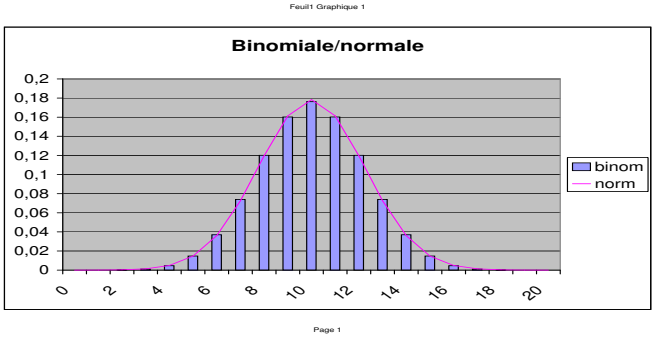


FIG. 8.1 – Le théorème limite central a été découvert par de Moivre et démontré par Laplace dans le cas d'une suite de v.a. de Bernoulli i.i.d. ; l'approximation d'une loi binômilale par une loi normale qui en résulte porte dès-lors le nom de théorème de de Moivre-Laplace. Ci dessus un histogramme de la loi binômiale $\mathcal{B}(20, 0.5)$ et de la densité de la loi normale $\mathcal{N}(\mu, \sigma)$ de même espérance $\mu = np = 10$ et même écart-type $\sigma = \sqrt{np(1-p)} = \sqrt{5}$.

Lemme 8.2 *Pour tout v.a. X pour laquelle M_X est définie, et tout $a \in \mathbb{R}$, on a* $\boxed{M_{aX}(t) = M_X(at)}$.

Rappelons que

$$Z_n := \frac{1}{\sigma\sqrt{n}}(X_1 + \ldots + X_n - n\mu) = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu),$$

et comme les v.a. $X_i - \mu$ sont indépendantes, par la proposition 7.6 on a

$$M_{Z_n}(t) = M_{\frac{1}{\sigma\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)}(t) = M_{\sum_{i=1}^{n}(X_i-\mu)}(\frac{t}{\sigma\sqrt{n}}) = \prod_{i=1}^{n} M_{X_i-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) = \left(\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n,$$

où l'on a posé $\varphi(\tau_n) := M_{X_i-\mu}(\tau_n)$ (qui sont toutes égales puisque les v.a. ont même loi), et $\tau_n = \frac{t}{\sigma\sqrt{n}}$; observons que $\lim_{n\to+\infty} \tau_n = \lim_{n\to+\infty} \frac{t}{\sigma\sqrt{n}} = 0$ pour tout t. Pour calculer la limite des $M_{Z_n}(t)$, nous alons utiliser un développement de Taylor à l'ordre 2 de φ ; pour cela, il nous faut connaître $\varphi(0)$, $\varphi'(0)$, et $\varphi''(0)$, que nous calculons en utilisant la formule (7.2). On a

$$\begin{aligned}
\varphi(0) &= \mathbb{E}((X_i-\mu)^0) = \mathbb{E}(1) = 1,\\
\varphi'(0) &= \mathbb{E}(X_i-\mu) = \mathbb{E}(X_i) - \mu = 0,\\
\varphi''(0) &= \mathbb{E}((X_i-\mu)^2) = \mathrm{Var}\,(X_i) = \sigma^2,
\end{aligned}$$

d'où finalement $\varphi(\tau) = \varphi(0) + \tau\varphi'(0) + \frac{\tau^2}{2}(\varphi''(0) + \varepsilon(\tau))$, où $\lim_{\tau\to 0} \varepsilon(\tau) = 0$. Nous allons également utiliser un développement de Taylor de $\ln(1+u)$; nous s'avons qu'il existe une fonction η telle que $\ln(1+u) = u(1+\eta(u))$, avec $\lim_{u\to 0} \eta(u) = 0$. Nous avons donc

$$\begin{aligned}
M_{Z_n}(t) &= \left(\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n = \left(1 + \frac{t^2}{2\sigma^2 n}(\sigma^2 + \varepsilon_n)\right)^n \text{ , où } \varepsilon_n := \varepsilon\left(\tfrac{t}{\sigma\sqrt{n}}\right)\\
&= \exp\left(n\ln\left(1 + \frac{t^2}{2\sigma^2 n}(\sigma^2+\varepsilon_n)\right)\right)\\
&= \exp\left(n\left(\frac{t^2}{2\sigma^2 n}(\sigma^2+\varepsilon_n)\right)(1+\eta_n)\right) \text{ , avec } \eta_n = \eta\left(\tfrac{t^2}{2\sigma^2 n}(\sigma^2+\varepsilon_n)\right)\\
&= \exp\left(\left(\frac{t^2}{2\sigma^2}(\sigma^2+\varepsilon_n)\right)(1+\eta_n)\right).
\end{aligned}$$

Donc

$$\lim_{n\to+\infty} M_{Z_n}(t) = \lim_{n\to+\infty} \exp\left(\left(\frac{t^2}{2\sigma^2}(\sigma^2+\varepsilon_n)\right)(1+\eta_n)\right) = e^{\frac{t^2}{2\sigma^2}\sigma^2} = e^{\frac{t^2}{2}} = M_Z(t).$$

On conclut, comme annoncé, en appliquant le théoreme 7.7.

□

8.2 Pratique du théorème limite central

Exemple typique : Une compagnie aérienne donne des réservations sur le vol d'un appareil de 400 places. La probabilité qu'un passager ayant réservé pour ce vol ne se présente pas est de $0.08 = 8\%$. Si la compagnie accorde 420 réservations sur ce vol, quel est le risque de "surbooking" (c'est-à-dire qu'il se présente plus de passagers que les 400 qui pourront embarquer) ?

Résolution par approximation normale : Soit X_i la v.a. de Bernoulli modélisant la présence $(X_i = 1)$ ou non $(X_i = 0)$ du i-ème passager réservé, $i = 1..n$, avec $n = 420$; par hypothèse $X_i \rightsquigarrow \mathcal{B}(1, p)$, avec $p = 1 - 8\% = 0.92\%$, et on suppose (implicitement ...) que les X_i sont indépendants. On a donc $\mu = \mathbb{E}X_i = p$, et $\sigma = \sqrt{\mathrm{Var}\,(X_i)} = \sqrt{p(1-p)}$.

Soit $X := X_1 + .. + X_{420}$ le nombre (aléatoire) de passagers réservés se présentant effectivement à l'enregistrement. Sous nos hypothèses $\mathbb{E}X = np = 420 \cdot 0.92 = 384.4$ et $\mathrm{Var}\,(X) = np(1-p) = 420 \cdot 0.92 \cdot 0.08 = 30.912$. Soit

$$\boxed{Z_n = \frac{X - \mathbb{E}X}{\sqrt{\mathrm{Var}\,(X)}}} \left(= \frac{X_1 + \ldots + X_{420} - n\mu}{\sigma\sqrt{n}} \right)$$

la v.a. considérée dans le théorème limite central, qui n'est autre que X centrée et réduite. L'*application du theorème consiste à assimiler* Z_{420} à $Z \rightsquigarrow \mathcal{N}(0,1)$.

L'évènement dont nous recherchons la probabilité est

$$E := \{X \leq 400\} = \left\{ Z_{420} \leq \frac{400 - \mathbb{E}X}{\sqrt{\mathrm{Var}\,(X)}} \right\} \simeq \left\{ Z \leq \frac{400 - \mathbb{E}X}{\sqrt{\mathrm{Var}\,(X)}} \right\} = \{Z \lessapprox 2.45\},$$

puisque $\frac{400 - \mathbb{E}X}{\sqrt{\mathrm{Var}\,(X)}} = \frac{400 - np}{\sqrt{np(1-p)}} = \frac{400 - 384.4}{\sqrt{30.912...}} = 2.446\ldots$.

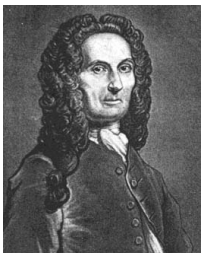
Comme $Z \rightsquigarrow \mathcal{N}(0,1)$, on recherche la valeur $2.45 = 2.40 + 0.05 = u_1 + u_2$ dans la table ci-dessous, et on trouve $\mathbb{P}(E) = 0.992857\ldots$. Il y a donc moins de 1% de risque qu'il se présente plus de 400 passagers à l'enregistrement[2].

[2] Sur la loi européenne sur la surréservation, voir par exemple :
http ://www.europe.gouv.fr/europe_7/europe_au_quotidien_25/surreservation_surbooking_avion_136.html

Voici quelques valeurs de la loi normale centrée réduite $\mathcal{N}(0,1)$ calculée au moyen d'Excel :

Loi Normale N(0,1)
Probabilité que X soit inférieure à u1+u2

u1\u2	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000000	0,5039894	0,5079784	0,5119665	0,5159535	0,5199389	0,5239223	0,5279032	0,5318814	0,5358565
0,1	0,5398279	0,5437954	0,5477585	0,5517168	0,5556700	0,5596177	0,5635595	0,5674949	0,5714237	0,5753454
0,2	0,5792597	0,5831661	0,5870644	0,5909541	0,5948348	0,5987063	0,6025681	0,6064198	0,6102612	0,6140918
0,3	0,6179114	0,6217195	0,6255158	0,6293000	0,6330717	0,6368306	0,6405764	0,6443087	0,6480272	0,6517317
0,4	0,6554217	0,6590970	0,6627572	0,6664021	0,6700314	0,6736448	0,6772419	0,6808225	0,6843863	0,6879331
0,5	0,6914625	0,6949743	0,6984682	0,7019441	0,7054015	0,7088403	0,7122603	0,7156612	0,7190427	0,7224047
0,6	0,7257469	0,7290692	0,7323712	0,7356528	0,7389138	0,7421540	0,7453732	0,7485712	0,7517478	0,7549030
0,7	0,7580364	0,7611480	0,7642376	0,7673050	0,7703501	0,7733727	0,7763728	0,7793501	0,7823046	0,7852362
0,8	0,7881447	0,7910300	0,7938920	0,7967307	0,7995459	0,8023375	0,8051055	0,8078498	0,8105704	0,8132671
0,9	0,8159399	0,8185888	0,8212136	0,8238145	0,8263912	0,8289439	0,8314724	0,8339768	0,8364569	0,8389129
1,0	0,8413447	0,8437523	0,8461358	0,8484950	0,8508300	0,8531409	0,8554277	0,8576903	0,8599289	0,8621434
1,1	0,8643339	0,8665004	0,8686431	0,8707618	0,8728568	0,8749280	0,8769755	0,8789995	0,8809998	0,8829767
1,2	0,8849303	0,8868605	0,8887675	0,8906514	0,8925122	0,8943502	0,8961653	0,8979576	0,8997274	0,9014746
1,3	0,9031995	0,9049020	0,9065824	0,9082408	0,9098773	0,9114919	0,9130850	0,9146565	0,9162066	0,9177355
1,4	0,9192433	0,9207301	0,9221961	0,9236414	0,9250663	0,9264707	0,9278549	0,9292191	0,9305633	0,9318879
1,5	0,9331928	0,9344783	0,9357445	0,9369916	0,9382198	0,9394292	0,9406200	0,9417924	0,9429466	0,9440826
1,6	0,9452007	0,9463011	0,9473839	0,9484493	0,9494974	0,9505285	0,9515428	0,9525403	0,9535214	0,9544861
1,7	0,9554346	0,9563671	0,9572838	0,9581849	0,9590705	0,9599409	0,9607961	0,9616365	0,9624621	0,9632731
1,8	0,9640697	0,9648522	0,9656206	0,9663751	0,9671159	0,9678433	0,9685573	0,9692582	0,9699460	0,9706211
1,9	0,9712835	0,9719335	0,9725711	0,9731967	0,9738102	0,9744120	0,9750022	0,9755809	0,9761483	0,9767046
2,0	0,9772499	0,9777845	0,9783084	0,9788218	0,9793249	0,9798179	0,9803008	0,9807739	0,9812373	0,9816912
2,1	0,9821356	0,9825709	0,9829970	0,9834143	0,9838227	0,9842224	0,9846137	0,9849966	0,9853713	0,9857379
2,2	0,9860966	0,9864475	0,9867907	0,9871263	0,9874546	0,9877756	0,9880894	0,9883962	0,9886962	0,9889894
2,3	0,9892759	0,9895559	0,9898296	0,9900969	0,9903582	0,9906133	0,9908625	0,9911060	0,9913437	0,9915758
2,4	0,9918025	0,9920237	0,9922397	0,9924506	0,9926564	0,9928572	0,9930531	0,9932443	0,9934309	0,9936128
2,5	0,9937903	0,9939634	0,9941322	0,9942969	0,9944574	0,9946138	0,9947664	0,9949150	0,9950600	0,9952012
2,6	0,9953388	0,9954729	0,9956035	0,9957307	0,9958547	0,9959754	0,9960929	0,9962074	0,9963188	0,9964274
2,7	0,9965330	0,9966358	0,9967359	0,9968332	0,9969280	0,9970202	0,9971099	0,9971971	0,9972820	0,9973645
2,8	0,9974448	0,9975229	0,9975988	0,9976725	0,9977443	0,9978140	0,9978817	0,9979476	0,9980116	0,9980737
2,9	0,9981341	0,9981928	0,9982498	0,9983051	0,9983589	0,9984111	0,9984617	0,9985109	0,9985587	0,9986050
3,0	0,9986500	0,9986937	0,9987361	0,9987772	0,9988170	0,9988557	0,9988932	0,9989296	0,9989649	0,9989991
3,1	0,9990323	0,9990645	0,9990957	0,9991259	0,9991552	0,9991836	0,9992111	0,9992377	0,9992636	0,9992886
3,2	0,9993128	0,9993363	0,9993590	0,9993810	0,9994023	0,9994229	0,9994429	0,9994622	0,9994809	0,9994990
3,3	0,9995165	0,9995335	0,9995499	0,9995657	0,9995811	0,9995959	0,9996102	0,9996241	0,9996375	0,9996505
3,4	0,9996630	0,9996751	0,9996868	0,9996982	0,9997091	0,9997197	0,9997299	0,9997397	0,9997492	0,9997584
3,5	0,9997673	0,9997759	0,9997842	0,9997922	0,9997999	0,9998073	0,9998145	0,9998215	0,9998282	0,9998346
3,6	0,9998409	0,9998469	0,9998527	0,9998583	0,9998636	0,9998688	0,9998739	0,9998787	0,9998834	0,9998878
3,7	0,9998922	0,9998963	0,9999004	0,9999042	0,9999080	0,9999116	0,9999150	0,9999184	0,9999216	0,9999247
3,8	0,9999276	0,9999305	0,9999333	0,9999359	0,9999385	0,9999409	0,9999433	0,9999456	0,9999478	0,9999499
3,9	0,9999519	0,9999538	0,9999557	0,9999575	0,9999592	0,9999609	0,9999625	0,9999640	0,9999655	0,9999669
4,0	0,9999683	0,9999696	0,9999709	0,9999721	0,9999733	0,9999744	0,9999755	0,9999765	0,9999775	0,9999784
4,1	0,9999793	0,9999802	0,9999810	0,9999819	0,9999826	0,9999834	0,9999841	0,9999848	0,9999854	0,9999860
4,2	0,9999866	0,9999872	0,9999878	0,9999883	0,9999888	0,9999893	0,9999898	0,9999902	0,9999906	0,9999911
4,3	0,9999915	0,9999918	0,9999922	0,9999925	0,9999929	0,9999932	0,9999935	0,9999938	0,9999941	0,9999943
4,4	0,9999946	0,9999948	0,9999951	0,9999953	0,9999955	0,9999957	0,9999959	0,9999961	0,9999963	0,9999964
4,5	0,9999966	0,9999968	0,9999969	0,9999970	0,9999972	0,9999973	0,9999974	0,9999976	0,9999977	0,9999978
4,6	0,9999979	0,9999980	0,9999981	0,9999982	0,9999983	0,9999983	0,9999984	0,9999985	0,9999986	0,9999986
4,7	0,9999987	0,9999988	0,9999988	0,9999989	0,9999989	0,9999990	0,9999990	0,9999991	0,9999991	0,9999992
4,8	0,9999992	0,9999992	0,9999993	0,9999993	0,9999993	0,9999994	0,9999994	0,9999994	0,9999995	0,9999995
4,9	0,9999995	0,9999995	0,9999996	0,9999996	0,9999996	0,9999996	0,9999996	0,9999997	0,9999997	0,9999997
5,0	0,9999997	0,9999997	0,9999997	0,9999998	0,9999998	0,9999998	0,9999998	0,9999998	0,9999998	0,9999998



Abraham de Moivre (1667-1754)

Pierre-Simon de Laplace (1749-1827)

Johann Carl Friederich Gauss (1777-1855)

Chapitre 9

Estimateurs au maximum de vraisemblance

Avec ce chapitre nous commençons l'étude de quelques outils centraux de la statistique.

9.1 Estimateur

Définition : Soit $n > 0$ un entier. Nous appellerons n-échantillon d'une loi $\mathcal{L}$ toute suite $X_1, \ldots, X_n$ de v.a. indépendantes de loi $\mathcal{L}$.

La statistique-pratique est un ensemble de techniques de traitement de données qui, face à la donnée de n nombres (ou plus généralement vecteurs) $x_1, \ldots, x_n$ produits par "échantillonage" - c'est-à-dire selon un protocole expérimental propre au domaine considéré (sociologie, contrôle de qualité, etc.) - choisit un n-échantillon au sens de la définition ci-dessus pour modèle mathématique suggérant un traitement de ces données.

Prenons l'exemple d'un référendum (ou d'un plébicite) où les électeurs ne peuvent que répondre par "oui" ou "non" (les abstentions étant sans influence sur le résultat, ce qui exclut les cas où il y a un quorum à atteindre). Choisissons $n = 1000$, et posons $x_i = 1$ si la i-ème personne interrogée déclare savoir ce qu'elle ira voter et vouloir voter "oui" (si elle déclare ne pas savoir ou ne pas envisager de voter, on écarte cette réponse de la présente analyse) et $x_i = 0$ si elle déclare vouloir voter "non".

Cette situation simple est généralement modélisée par un 1000-échantillon $X_1, \ldots, X_{1000}$ d'une loi de Bernoulli $\mathcal{B}(1, p)$, et on considère que l'opinion est en faveur du "oui" si et seulement si $p \geq 0.5$.

On est alors confronté au problème "d'estimer" la valeur de p. Dans le modèle considéré ici (Bernouilli) la loi des grands nombres vient à notre secours : elle assure que $\lim_{n\to+\infty}(X_1 + \ldots + X_n)/n = \mathbb{E}(X_1) = p$; on dit dans ce cas que $\hat{p} := (X_1 + \ldots + X_n)/n$ est un *estimateur* du paramètre p ; en pratique, on choisit alors $p = p^* := (x_1 + \ldots + x_{1000})/1000$.

Nous nous intéresserons ici à la *statistique paramétrique*, où la loi $\mathcal{L} = \mathcal{L}(\theta)$ retenue peut être caractérisé par un paramètre θ, qui est un nombre ou un vecteur. Ainsi, par exemple, si $X_i \rightsquigarrow \mathcal{B}(1, p)$, alors $\theta = p$ est un nombre, mais si $X_i \rightsquigarrow \mathcal{N}(\mu, \sigma)$, alors $\theta = (\mu, \sigma)$ est un vecteur, tout comme dans le cas d'un dé pipé où l'on peut choisir $\theta = (p_1, \ldots, p_5)$ (et $p_6 = 1 - (p_1 + \ldots + p_5)$) et $p_k := \mathbb{P}_\theta(\{X_i = k\})$.

Définition : On dit que $\hat{\theta} : (x_1, \ldots, x_n) \mapsto \hat{\theta}_n := \hat{\theta}(x_1, \ldots, x_n)$ est un *estimateur* convergeant vers θ si et seulement si , en loi, on a $\theta = \lim_{n\to+\infty} \hat{\theta}(X_1, \ldots, X_n)$ pour toute suite de v.a. X_i indépendantes, de loi $\mathcal{L}(\theta)$.

9.2 Vraisemblance

9.2.1 Heuristique et définition

Nous avons vu que la loi des grands nombres fournit "spontanément" un estimateur de l'espérance d'une loi, mais si l'on recherche une méthode un peu générale pour deviner un estimateur, la *méthode du maximum de vraisemblance* est une stratégie souvent efficace. En voici le principe :

Si un échantillonage a produit la suite finie $x_1^*, \ldots, x_n^*$ de nombres et qu'on a choisit de modéliser cette situation par un n-échantillon $X_1, \ldots, X_n$ de v.a. indépendantes de loi $\mathcal{L}(\theta)$, et si le choix de la

valeur du paramètre θ est le problème auquel on est confronté, on peut considérer l'évènement $E^* = \{X_1 = x_1^*, \ldots, X_n = x_n^*\}$, et plus généralement

$$E(x_1, \ldots, x_n) = \{X_1 = x_1, \ldots, X_n = x_n\} = \{X_1 = x_1\} \cap \ldots \cap \{X_n = x_n\}$$

et sa probabilité, sous $\mathbb{P}_\theta$ telle que la loi des X_i est $\mathcal{L}(\theta)$, vaut

$$L(x_1, \ldots, x_n\,;\theta) := \mathbb{P}_\theta(E(x_1, \ldots, x_n)) = \mathbb{P}_\theta(\{X_1 = x_1\} \cap \ldots \cap \{X_n = x_n\}) = \mathbb{P}_\theta(\{X_1 = x_1\}) \cdot \ldots \cdot \mathbb{P}_\theta(\{X_n = x_n\}),$$

cette dernière égalité résultant de l'hypothèse d'indépendance des v.a. X_i. L'idée très heuristique est alors que le choix θ^* qu'il convient d'effecteur pour θ, est celui pour lequel cette probabilité est maximale pour les valeurs x_1^*, …, x_n^* obtenues, et donc de poser

$$\theta^* = \text{Argmax}_{\,\theta}\{L(x_1^*, \ldots, x_n^*\,;\theta)\},$$

c'est-à-dire la valeur (si elle existe et est unique) de θ pour laquelle la fonction $\theta \mapsto L(x_1^*, \ldots, x_n^*\,;\theta)$ est maximale[1]. Souvent, ceci peut se ramener à résoudre en θ l'équation $\frac{\partial L}{\partial \theta}(x_1^*, \ldots, x_n^*\,;\theta) = 0$.

Définition : La fonction $L_n : (x_1, \ldots, x_n\,;\theta) \mapsto \boxed{L_n(x_1, \ldots, x_n\,;\theta) = \prod_{i=1}^n \mathbb{P}_\theta(\{X_i = x_i\})}$ pour des $X_i \rightsquigarrow \mathcal{L}(\theta)$ s'appelle la vraisemblance de la loi $\mathcal{L}$.

La v.a. obtenue en appliquant la fonction $(x_1, \ldots, x_n) \mapsto \text{Argmax}_{\,\theta}\{L(x_1, \ldots, x_n\,;\theta)\}$ appliquée au n-échantillon $(X_1, \ldots, X_n)$ s'appelle l'*estimateur au maximum de vraisemblance* du paramètre θ de la loi discrète $\mathcal{L}(\theta)$.

9.2.2 Exemples

Referendum

Reprenons l'exemple où les X_i suivent une loi de Bernoulli $\mathcal{B}(1, p)$, et donc $\theta = p$. Introduisons la notation $s := x_1 + \ldots + x_{1000}$ pour la somme des valeurs observées sur l'échantillon x_1, …, x_{1000}, c'est-à-dire le nombre de personnes interrogées qui ont déclaré qu'elles voterons "oui". Nous avons donc $L_n(x_1, \ldots, x_n\,;\theta) = \prod_{i=1}^n \mathbb{P}_\theta(\{X_i = x_i\}) = \theta^s(1-\theta)^{n-s}$, pour $\theta = p$, $n = 1000$, et $s = x_1 + \ldots + x_n$, puisque $\theta = p = \mathbb{P}_\theta(\{X_i = 1\})$ et $1 - \theta = 1 - p = \mathbb{P}_\theta(\{X_i = 0\})$.

Les extrémités de l'intervalle $[0, 1]$ auquel appartient θ ne peuvent être des extrema (sauf si $s = 0$ ou $s = n$) et le maximum θ^* de la fonction concave $\theta^s(1-\theta)^{n-s}$ est donc un zéro de la dérivée $\frac{\partial}{\partial \theta} L_n(x_1, \ldots, x_n\,;\theta) = \theta^{s-1}(1-\theta)^{n-s-1}(s - n\theta)$, d'où $\theta^* = \frac{s}{n} = \frac{x_1 + \ldots + x_n}{n}$. En d'autres termes, l'estimateur au maximum de vraisemblance $\hat{p}$ de p est donc $\boxed{\hat{\theta} := \frac{X_1 + \ldots + X_n}{n}}$, c'est-à-dire le même estimateur que l'estimateur de l'espérance $\mathbb{E}(X_1)$ trouvé en appliquant la loi des grands nombres, ce qui convient, puisque $p = \mathbb{E}(X_i)$.

Variables poissoniennes

Supposons que le tirage d'un n-échantillon $X_1, \ldots, X_n$ de v.a. suivant une loi de Poisson $\mathcal{P}(\lambda)$, $\lambda > 0$ inconnu, ait produit l'échantillon $x_1, \ldots, x_n$. Ici $\theta = \lambda$, et $\mathbb{P}_\theta(\{X_i = x_i\}) = e^{-\theta}\frac{\theta^{x_i}}{x_i\,!}$; la vraisemblance de l'échantillon $x_1, \ldots, x_n$ est donc ici $L_n(x_1, \ldots, x_n; \theta) = \prod_{i=1}^n e^{-\theta}\frac{\theta^{x_i}}{x_i\,!} = e^{-n\theta}\frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i\,!}$, et donc $L_n(x_1, \ldots, x_n; \theta) = e^{-n\theta}\frac{\theta^s}{\prod_{i=1}^n x_i\,!}$, où l'on a une nouvelle fois posé $s := x_1 + \ldots + x_n$. Il est un peu plus commode de calculer avec le logarithme de cette expression, et comme ln est une fonction croissante, il nous suffit de rechercher le maximum θ^* de

$$l_n(x_1, \ldots, x_n; \theta) = \ln(L_n(x_1, \ldots, x_n; \theta)) = -n\theta + s\ln(\theta) - \sum_{i=1}^n \ln(x_i\,!).$$

Cette fonction est concave et son extremum θ^* est donc le zéro de la dérivée $\frac{\partial}{\partial \theta} l_n(x_1, \ldots, x_n; \theta) = -n + \frac{s}{\theta}$, c'est-à-dire $\theta^* = \frac{s}{n}$.

Nous trouvons donc une nouvelle fois $\boxed{\hat{\theta} := \frac{X_1 + \ldots + X_n}{n}}$ comme estimateur de λ, ce qui convient, puisque $\lambda = \mathbb{E}(X_i)$ pour toute v.a. $X_i \rightsquigarrow \mathcal{P}(\lambda)$.

[1] Il y a souvent malentendu sur le sens du mot "maximum". Pour une fonction $\theta \mapsto L(\theta)$, le *maximum* désigne une valeur θ^* de θ ; ainsi, $\theta^* := 0$ est le maximum de $L(\theta) := 1 - \theta^2$, et 1 est la *valeur maximale* $L(\theta^*)$. On peut aussi être amené, comme ici, à considérer un ensemble de valeurs, comme $\{L(\theta), \theta \in \mathbb{R}\}$; dans ce cas $\text{Max}\,\{L(\theta), \theta \in \mathbb{R}\}$ est une valeur de L ($1 = L(\theta^*)$, dans notre exemple). Comme l'ensemble est indexé par des $\theta \in \mathbb{R}$, on désigne par Argmax l'ensemble des θ^* donnant à L sa valeur maximale. Quand ce θ^* est unique, c'est ce θ^* que l'on note (abusivement) Argmax .

9.3 Cas d'une loi continue

9.3.1 Heuristique et définition

Si la loi $\mathcal{L}(\theta)$ suivie par les X_i est une loi continue, comme $\mathcal{U}_{[a,b]}$ ou $\mathcal{N}(\mu,\sigma)$, on a $\mathbb{P}_\theta(\{X_i = x_i\}) = 0$, et la vraisemblance que nous avons considérée jusqu'ici est tout bonnement (ou plutôt "mauvaisement") nulle, et tous les θ sont des extrema, ce qui ne nous avance guère. L'idée est alors de remplacer $\mathbb{P}_\theta(\{X_i = x_i\})$ par $\mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\})$ pour un $\varepsilon > 0$ suffisamment petit, puis de rechercher θ_ε maximisant $\prod_{i=1}^n \mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\})$. On peut se débarasser du ε qui est arbitraire par la remarque suivante dans le cas où la densité $x \mapsto f_\theta(x)$ caractérisant la loi $\mathcal{L}(\theta)$ est une fonction continue au point x_i : dans ce cas le théorème de la moyenne assure l'existence de fonctions $\varepsilon \mapsto \alpha_{i,\theta}(\varepsilon)$ telles que $\mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\}) = 2\varepsilon(f_\theta(x_i) + \alpha_{i,\theta}(\varepsilon))$, avec $\lim_{\varepsilon\to 0} \alpha_{i,\theta}(\varepsilon) = 0$; le (ou les) θ_ε rendant maximal $\prod_{i=1}^n \mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\})$ sont les même que ceux maximisant

$$\prod_{i=1}^n \frac{1}{2\varepsilon}\mathbb{P}_\theta(\{|X_i - x_i| \leq \varepsilon\}) = \prod_{i=1}^n (f_\theta(x_i) + \alpha_{i,\theta}(\varepsilon)) \ ;$$

en faisant tendre ε vers 0, cette expression devient

$$L_n(x_1,\ldots,x_n;\theta) := \prod_{i=1}^n f_\theta(x_i) \tag{9.1}$$

que nous adoptons comme vraisemblance dans ce cas :

Définition : Si la loi $\mathcal{L}(\theta)$ des X_i est une loi continue de densité f_θ, on appelle vraisemblance de l'échantillon $(x_1,\ldots,x_n)$ pour la loi continue $\mathcal{L}(\theta)$ la fonction $\boxed{L_n(x_1,\ldots,x_n;\theta) := \prod_{i=1}^n f_\theta(x_i)}$.

9.3.2 Exemples

Distribution uniforme

On suppose que l'échantillon x_1, ..., x_n est tiré de manière uniforme entre a et b, mais a et b sont inconnus. On modélise donc le problème par une loi uniforme $\mathcal{U}[a,b]$ dont la densité est $f_{(a,b)} := \frac{1}{b-a}\mathbb{I}_{[a,b]}$ et on va chercher un estimateur de $\theta = (a,b)$ par la méthode du maximum de vraisemblance. La vraisemblance de l'échantillon x_1, ..., x_n est donc $L_n(x_1,\ldots,x_n;\theta) := \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{1}{b-a}\mathbb{I}_{[a,b]}(x_i) = \frac{1}{(b-a)^n}$ si tous les $x_i \in [a,b]$ et vaut 0 si un des $x_i \notin [a,b]$. On voit donc que $L_n(x_1,\ldots,x_n;\theta)$ est maximal si $\theta = \theta^* = (a^*,b^*) = (\mathrm{Min}\,\{x_1,\ldots,x_n\}, \mathrm{Max}\,\{x_1,\ldots,x_n\})$, puisque ceci nous donne la plus petite valeur de $b-a$ sans annuler la vraisemblance. Ceci nous conduit à considérer l'estimateur

$$\boxed{\hat\theta = (\hat a,\hat b) = (\mathrm{Min}\,\{X_1,\ldots,X_n\}, \mathrm{Max}\,\{X_1,\ldots,X_n\})}.$$

Il reste à montrer que si X_1, ..., X_n est un n-échantillon de loi $\mathcal{U}[a,b]$, alors $\mathrm{Min}\,\{X_1,\ldots,X_n\}$ converge bien, en probabilité, vers a et que $\mathrm{Max}\,\{X_1,\ldots,X_n\}$ converge en probabilité vers b. Considérons par exemple le cas de $\mathrm{Min}\,\{X_1,\ldots,X_n\}$. On a $\{a+\varepsilon < \mathrm{Min}\,\{X_1,\ldots,X_n\}\} = \{a+\varepsilon < X_1,\ldots,a+\varepsilon < X_n\}$, d'où, comme les X_i sont indépendants, $\mathbb{P}(\{a+\varepsilon < \mathrm{Min}\,\{X_1,\ldots,X_n\}\}) = \mathbb{P}(\{a+\varepsilon < X_1\} \cap \ldots \cap \{a+\varepsilon < X_n\}) = \mathbb{P}(\{a+\varepsilon < X_1\}) \cdot \ldots \cdot \mathbb{P}(\{a+\varepsilon < X_n\}) = \left(\frac{b-a-\varepsilon}{b-a}\right)^n$, qui tend bien vers 0 lorsque n tend vers $+\infty$. On montrerais de même que $\mathrm{Max}\,\{X_1,\ldots,X_n\}$ converge en probabilité vers b.

Variables normales

On suppose à présent que l'échantillon x_1, ..., x_n est tiré de manière normale avec une espérance μ et un écart-type σ, mais μ et σ sont inconnus. On modélise donc le problème par une loi normale $\mathcal{N}(\mu,\sigma)$ dont la densité est $f_{(\mu,\sigma)}(x) := \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ et on va chercher un estimateur de $\theta = (\mu,\sigma)$ par la méthode du maximum de vraisemblance. La vraisemblance de l'échantillon x_1, ..., x_n est donc

$$\begin{aligned} L_n(x_1,\ldots,x_n;\theta) : &= \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{\sum_{i=1}^n (x_i-\mu)^2}{2\sigma^2}}. \end{aligned}$$

Ici, il est une nouvelle fois plus agréable de considérer la log-vraisemblance

$$l_n(x_1,\dots,x_n;\theta) := \ln(L_n(x_1,\dots,x_n;\theta)) = -n(\ln(\sigma)+\ln(\sqrt{2\pi})) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Pour que $\theta^* = (\mu^*,\sigma^*)$ soit un extremum sur $\mathbb{R}\times\mathbb{R}_*^+$ il faut que les deux dérivées $\frac{\partial}{\partial\mu}l_n(x_1,\dots,x_n;\theta) = \frac{1}{\sigma^2}\sum_{i=1}^n(x_i-\mu) = \frac{1}{\sigma^2}(s-n\mu)$ où $s=\sum_{i=1}^n x_i$, et $\frac{\partial}{\partial\sigma}l_n(x_1,\dots,x_n;\theta) = -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^n(x_i-\mu)^2$ s'annulent pour $\theta=\theta^*$, ce qui implique que $\mu = \frac{s}{n} = \frac{1}{n}\sum_{i=1}^n x_i$, et $\sigma^2 = \frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2$. Ceci nous conduit donc à envisager l'estimateur

$$\boxed{\hat\theta = (\hat\mu,\hat\sigma) = \left(\frac{1}{n}\sum_{i=1}^n X_i, \left(\frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2\right)^{\frac{1}{2}}\right)}.$$

En ce qui concerne la première composante $\hat\mu$, nous retrouvons une nouvelle fois la moyenne comme estimateur de l'espérance $\mu=\mathbb{E}(X_i)$, quant-à la seconde composante, nous trouvons

$$\boxed{\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^n(x_i-\mu)^2)}$$

dont nous verrons qu'il s'agit bien, pour toute loi, d'un estimateur de la variance σ^2.

Chapitre 10

Convergence d'estimateurs

Au chapitre précèdent, nous avons introduit la notion d'estimateur $\hat{\theta}_n = \theta_n(X_1, \ldots, X_n)$ de la valeur θ du paramètre caractérisant la loi des $X_i \rightsquigarrow \mathcal{L}(\theta)$, et nous avons donné une stratégie (le *maximum de vraissemblance*) pour former des fonctions θ_n et, partant, suggérer des estimateurs $\hat{\theta}_n$. Il convient à présent de déterminer dans quelle mesure un estimateur $\hat{\theta}_n$ converge effectivement vers la (bonne) valeur θ. Comme $\hat{\theta}_n$ est une v.a., l'écart entre $\hat{\theta}_n$ et θ dépend à la fois de $\omega \in \Omega$ (le "hasard") et de n (l'"asymptotique"). Conformément à l'usage en analyse classique, nous exprimerons que $\hat{\theta}_n$ est proche de θ pour n grand en définissant ce que nous entendons par $\theta = \lim_{n\to+\infty} \hat{\theta}_n$. Voici deux manière, non equivalentes, d'exprimer qu'une v.a. $\overline{Y}$ est la limite d'une suite de v.a. Y_n.

10.1 Convergence d'une suite de v.a.

10.1.1 Convergence en probabilité

Définition : On dit qu'une suite de v.a. $(Y_n)_{n\geq 0}$ *converge en probabilité vers la v.a.* $\overline{Y}$, et on note $Y_n \xrightarrow{P} \overline{Y}$ si et seulement si $\forall \delta > 0$, $\lim_{n\to+\infty} \mathbb{P}(\{|Y_n - \overline{Y}| > \delta\}) = 0$.
On dit que l'estimateur $\hat{\theta}_n$ de θ est *consistent* si et seulement si $\hat{\theta}_n \xrightarrow{P} \theta$.

10.1.2 Convergence dans L^q

Rappelons que l'inégalité de Markov (6.1) assure que $\mathbb{P}(\{|X| \geq \delta\}) \leq \frac{1}{\varphi(\delta)}\mathbb{E}(\varphi(|X|))$ pour n'importe quelle fonction croissante $\varphi : \mathbb{R}_*^+ \longrightarrow \mathbb{R}_*^+$. Une stratégie pour montrer la convergence en probabilité des Y_n vers $\overline{Y}$ est alors de trouver une fonction φ telle qu'on puisse s'assurer que $\lim_{n\to+\infty} \mathbb{E}(\varphi(|Y_n - \overline{Y}|))=0$. Lorsqu'on peut choisir $\varphi(x) = x^q$ avec $q \geq 1$, on dit que les Y_n convergent *dans* L^q vers $\overline{Y}$:

Définition : Soit $q \geq 1$; on dit qu'une suite de v.a. $(Y_n)_{n\geq 0}$ *converge dans* L^q *vers la v.a.* $\overline{Y}$, et on note $Y_n \xrightarrow{L^q} \overline{Y}$ si et seulement si $\lim_{n\to+\infty} \mathbb{E}(|Y_n - \overline{Y}|^q) = 0$.

Nous venons de remarquer que l'inégalité de Markov montre qu'on a

Proposition 10.1 *Si* $Y_n \xrightarrow{L^q} \overline{Y}$, *alors* $Y_n \xrightarrow{P} \overline{Y}$.

Nous voyons que la convergence en probabilité ne dit rien des valeurs de $Y_n - \overline{Y}$ lorsque $|Y_n - \overline{Y}| > \delta$; ces valeurs peuvent donc être très grandes, quand bien même elles seraient de faible probabilité. Il est donc douteux que cette proposition admette une réciproque. Voici un exemple montrant que la réciproque de la proposition (10.1) est effectivement fausse en général.

Contre-exemple à la réciproque

Considérons une situation où $\overline{Y} = 0$; soit $q \geq 1$ quelconque fixé. Choisissons $Y_n := nZ_n$, avec $Z_n \rightsquigarrow \mathcal{B}(1, p_n)$. Observons que, comme $Z_n \rightsquigarrow \mathcal{B}(1, p_n)$, Z_n ne prend que les valeurs 0 et 1, et donc $|Z_n|^q = Z_n$. On a donc

$$\mathbb{P}(\{|Y_n - \overline{Y}|^q > \delta\}) = \mathbb{P}(\{n^q Z_n > \delta\}) = \mathbb{P}\left(\left\{Z_n > \frac{\delta}{n^q}\right\}\right) \overset{\sharp}{=} \mathbb{P}(\{Z_n = 1\}) = p_n,$$

où ($\sharp$) est vrai dès que $\frac{\delta}{n^q} \leq 1$.

Nous avons donc convergence en probabilité des Y_n vers $\overline{Y} = 0$ dès lors qu'on suppose que $\lim p_n = 0$.

En revanche $\mathbb{E}(|Y_n - \overline{Y}|^q) = \mathbb{E}(n^q Z_n) = n^q p_n$ qui sera constamment égale à 1 (et ne convergera donc pas vers 0) si l'on pose $p_n = \frac{1}{n^q}$ qui tend bien vers 0, ce qui, comme nous l'avons vu, assure la convergence en probabilité. Nous avons donc bien formé ainsi un exemple de suite $(Y_n)_{n\geq 1}$ qui converge en probabilité vers $\overline{Y} = 0$ et ne converge pas dans L^q.

10.1.3 Cas L^2 : convergence en moyenne quadratique

Approfondissons le cas important où $q = 2$ déjà abordé au paragraphe 6.3.4 ; c'est le cas qui est le plus commode du point de vue des calculs, et c'est celui où la variabilité est mesurée par la variance σ^2 qui nous est à présent familière. Un peu de terminologie :

Définition : On dit que l'estimateur $\hat{\theta}_n$ *converge en moyenne quadratique vers* θ si et seulement si $\hat{\theta}_n \xrightarrow{L^2} \theta$. On dit aussi que *l'erreur quadratique* $(\mathbb{E}(|\hat{\theta}_n - \theta|^2))^{\frac{1}{2}}$ tends vers 0.

Remarquons tout d'abord que l'identité du Huygens assure que

$$\mathbb{E}\left(|\hat{\theta}_n - \theta|^2\right) = \mathrm{Var}\,(\hat{\theta}_n) + \left(\theta - \mathbb{E}(\hat{\theta}_n)\right)^2. \tag{10.1}$$

Par cette identité, nous voyons que pour assurer la convergence en moyenne quadratique de $\hat{\theta}_n$ vers θ il faut et il suffit qu'à la fois $\lim_{n\to+\infty} \mathrm{Var}\,(\hat{\theta}_n) = 0$ et $\lim_{n\to+\infty} \mathbb{E}(\hat{\theta}_n) = \theta$.

Exemple : Nous avons déjà souligné que la Loi des Grands Nombres assure, pour toute suite $(X_n)_{n\geq 0}$ de v.a. i.i.d. d'espérance $\mu := \mathbb{E}(X_i)$ et de variance $\sigma^2 := \mathrm{Var}\,(X_i)$, la convergence en probabilité de $\hat{\theta} := M_n := \frac{1}{n}(X_1 + \ldots + X_n)$ vers l'espérance commune $\mu := \mathbb{E}(X_i)$. Montrons que nous avons aussi convergence dans L^2. En vertu de (10.1) il suffit que nous vérifions à la fois que $\lim_{n\to+\infty} \mathbb{E}(M_n) = \mu$ et $\lim_{n\to+\infty} \mathrm{Var}\,(M_n) = 0$. On a clairement $\mathbb{E}(M_n) = \mu$ pour tout n ; de plus

$$\mathrm{Var}\,(M_n) = \mathrm{Var}\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n^2}(\mathrm{Var}\,(X_1) + \ldots + \mathrm{Var}\,(X_n)) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}\ ;$$

donc $\hat{\theta} := M_n$ converge bien, dans L^2, vers μ.

10.2 Biais d'un estimateur

Nous venons de voir que pour l'estimateur $\hat{\theta}_n := M_n$ de μ, on a $\mathbb{E}(\hat{\theta}_n) = \mu$; on dit que la moyenne M_n est un estimateur *sans biais* de l'espérance μ ; plus généralement :

Définition : On appelle *biais* d'un estimateur $\hat{\theta}_n$ de θ le nombre $\theta - \mathbb{E}(\hat{\theta}_n)$. On dit que l'estimateur $\hat{\theta}_n$ de θ est *sans biais* si et seulement si son biais $\theta - \mathbb{E}(\hat{\theta}_n)$ est nul pour tout n.

Remarque : Au vu de (10.1) on comprend l'intérêt, pour réduire l'erreur en moyenne quadratique $\mathbb{E}(|\hat{\theta}_n - \theta|^2)$ de l'estimateur $\hat{\theta}_n$ de θ, de préférer des estimateurs sans biais. Toutefois on peut préférer un estimateur avec biais dans le cas où le choix d'un tel estimateur avec biais permet d'avoir une convergence plus rapide de $\mathrm{Var}\,(\hat{\theta}_n)$ vers 0.

Au chapitre 9 nous avons vu que, pour la loi $\mathcal{L}(a,b) := \mathcal{U}[a,b]$, $\hat{a} := \mathrm{Min}\,\{X_1, \ldots, X_n\}$ est un estimateur consistent de a et $\hat{b} := \mathrm{Max}\,\{X_1, \ldots, X_n\}$ est un estimateur consistent de b. Calculons le biais $b - \mathbb{E}(\mathrm{Max}\,\{X_1, \ldots, X_n\})$ de $\hat{b}$.

Nous avons vu que $F(x) := \mathbb{P}(\{\mathrm{Max}\,\{X_1, \ldots, X_n\} \leq x\}) = \left(\frac{x-a}{b-a}\right)^n$ pour $x \in [a,b]$. Donc, pour $x \in]a,b[$, la densité f est donnée par $f(x) = F'(x) = \frac{n}{(b-a)^n}(x-a)^{n-1}$, d'où, en intégrant par parties

$$\begin{aligned}
\mathbb{E}(\mathrm{Max}\,\{X_1, \ldots, X_n\}) &= \int_a^b x f(x) dx = \frac{1}{(b-a)^n}\int_a^b x n (x-a)^{n-1} dx \\
&= \frac{1}{(b-a)^n}\left([x(x-a)^n]_{x=a}^b - \int_a^b (x-a)^n dx\right) \\
&= \frac{1}{(b-a)^n}\left(b(b-a)^n - \frac{1}{n+1}[(x-a)^{n+1}]_{x=a}^b\right) \\
&= \frac{1}{(n+1)(b-a)^n}[(n+1)b(b-a)^n - (b-a)^{n+1}] = \frac{1}{n+1}(nb+a).
\end{aligned}$$

Finalement $\boxed{\overline{b} := \mathbb{E}(\text{Max}\{X_1, \ldots, X_n\}) = b - \frac{b-a}{n+1}}$.

On montrerait de même que $\boxed{\overline{a} := \mathbb{E}(\text{Min}\{X_1, \ldots, X_n\}) = a + \frac{b-a}{n+1}}$. Nous voyons donc que le biais $(a, b) - \hat{\theta}_n$ de cet estimateur est égal $(\frac{b-a}{n+1}, -\frac{b-a}{n+1})$.

10.2.1 Elimination du biais d'un estimateur

Supposons que $a = 0$ soit connu et que seul b soit à estimer. On vérifie aisément que le biais est inchangé : $b - \mathbb{E}(\text{Max}\{X_1, \ldots, X_n\}) = \frac{b}{n+1}$, et donc $\mathbb{E}(\text{Max}\{X_1, \ldots, X_n\}) = \frac{n}{n+1}b$, d'où encore $b = \frac{n+1}{n}\mathbb{E}(\text{Max}\{X_1, \ldots, X_n\}) = \mathbb{E}(\frac{n+1}{n}\text{Max}\{X_1, \ldots, X_n\})$, d'où l'idée de considérer le nouvel estimateur $\boxed{\hat{b}' := \frac{n+1}{n}\text{Max}\{X_1, \ldots, X_n\}.}$

Compte tenu de ce qui précède, $\hat{b}'$ est, lui, sans biais.

Exercice : On ne suppose plus connue la valeur de a.

1. Trouver $\alpha(n)$ et $\beta(n)$ tel que $\hat{b}'' := \alpha(n)\hat{a} + \beta(n)\hat{b}$ soit un estimateur sans biais de b.
2. Calculer Var $(\hat{b})$ et en déduire Var $(\hat{b}')$.
3. En déduire que les estimateurs $\hat{b}$ et $\hat{b}'$ sont des estimateurs consistents de b et qu'ils convergent en moyenne quadratique.

.

Chapitre 11

Intervalles de confiance

Nous allons examiner la notion de domaine de confiance (généralement un intervalle) sur une problématique toujours sensible : celle des sondages. Je prendrai l'exemple du sondage IFOP–Wanadoo-actu effectué les 12 et 13 mai 2005 (OUI : 46%, NON : 54%). Rappelons tout-de-suite qu'il s'agissait d'un cliché de l'opinion à ces dates, et pas d'un pronostic du résultat, ne serais-ce que parce que seuls 67% des partisans du OUI se déclaraient sûrs de leur choix et 33% pouvoir encore changer d'avis (et 76%-24% pour les "déclarants" du NON). Nous allons aborder la question de ce que l'on pouvait conclure de ce sondage avec un peu de certitude sur l'opinion à cette date.

11.1 Modélisation

Parmi les N personnes de la population qui sont inscrites sur les listes electorales et déclarant aller voter le 29 mai, il y a une *proportion* p voulant voter OUI et $1-p$ voulant noter NON.

On procède alors à un *sondage*, un protocole (subtile) cherchant à tirer "au hasard" et de façon "indépendantes" n personnes a qui on demandera leur intention de vote x_1, x_2, ..., x_n, avec $x_i = 1$ si la i-ème personne interrogée déclare vouloir voter OUI.

On modélise cette interrogation de n personnes par un n-échantillon X_1, ..., X_n de v.a. i.i.d., avec $X_i = \mathcal{B}(1,p)$. Il est naturel de choisir pour modèle des $X_i \rightsquigarrow \mathcal{B}(1,q)$, avec $q \in]0,1[$, puisque seule deux réponses sont possibles (on a écarté les réponses différentes de OUI et NON), codée par 1 et 0, respectivement. Le choix de $q = p$, la proportion d'electeurs voulant voter OUI peut se motiver par le choix suivant : on définit $X_i : \Omega_i \longrightarrow \{0,1\}$, avec $\Omega_i = \{\omega_1, \ldots, \omega_N\}$, chacun des $\omega_j \in \Omega_i$ représantant un des N électeurs inscrits ; on postule que chacun des électeurs a la même chance d'être interrogé lors de la i-ème interrogation, et on choisit donc une loi uniforme sur Ω_i, d'où $\mathbb{P}(\{\omega\}) = \frac{1}{N}$ pour tout $\omega \in \Omega_i$. Soient $A := \{X_i = 1\}$ les électeurs voulant voter OUI et $a := \mathrm{Card}\,(A)$ ne nombre de ces électeurs voulant voter OUI. ; on a alors $\mathbb{P}(\{X_i = 1\}) = \mathbb{P}(A) = \frac{a}{N} = p$, par définition de p. Ceci permet de choisir la *loi* de $X_i \rightsquigarrow \mathcal{B}(1,p)$. A noter que pour avec un modèle avec n v.a. $X_1, \ldots, X_n$ indépendantes, il faut alors choisir un autre Ω ; par exemple $\Omega = \Omega_1 \times \ldots \times \Omega_n$, ou plus simplement $\Omega = \{0,1\}^n$, avec $\mathbb{P}(\{(\omega_i, \ldots, \omega_n)\}) = p^{\omega_1+\ldots+\omega_n}(1-p)^{(1-\omega_1)+\ldots+(1-\omega_n)}$ et $X_i(\omega_i, \ldots, \omega_n) = \omega_i$. Notons que toutefois toutes ces précisions de sont pas indispensables : tout ce dont nous avons besoin, c'est que les X_i soient i.i.d. de loi $\mathcal{B}(1,p)$, où p est la proportion d'électeurs voulant voter OUI.

11.2 Domaine de confiance pour l'estimation de p

La question maintenant est d'estimer la valeur de la proportion p et de juger de la qualité de cette estimation. Dans notre modèle p est l'espérance commune des X_i et nous avons vu que la moyenne $\boxed{P_n := \frac{1}{n}(X_1 + \ldots + X_n)}$ est un estimateur qui converge (en probabilité) vers cette espérance commune : c'est essentiellement la loi des grands nombres (LGN).

Si les n personnes interrogées ont répondu $x_1, \ldots, x_n$, en pratique on donne l'estimation

$$\boxed{\hat{p} = p_n = \frac{1}{n}(x_1 + \ldots + x_n)}.$$

La question est alors de savoir quelle confiance attacher à cette estimation $\hat{p}$. C'est l'objet de la définition d'un domaine de confiance au seuil α, par exemple $\alpha = 5\%$.

Définition : On dit que l'intervalle (aléatoire) $\mathcal{D}_n = [P_n - \Delta^-, P_n + \Delta^+]$ est un domaine de confiance pour l'estimation de p au seuil α si et seulement si $\boxed{\mathbb{P}(\{p \notin \mathcal{D}_n\}) \leq \alpha}$.

En d'autres termes, il y a une probabilité au plus égale à α de se tromper si l'on affirme que $p \in \mathcal{D}_n$.
Pour déterminer un tel intervalle de confiance, il convient de réécrire l'évènement

$$E_n := \{p \in \mathcal{D}_n\} = \{p \in [P_n - \Delta^-, P_n + \Delta^+]\}.$$

Tout d'abord nous avons

$$\begin{aligned} E_n &= \{p \in [P_n - \Delta^-, P_n + \Delta^+]\} = \{p \geq P_n - \Delta^-, p \leq P_n + \Delta^+\} = \{P_n \leq p + \Delta^-, P_n \geq p - \Delta^+\} \\ &= \{P_n \in [p - \Delta^+, p + \Delta^-]\}. \end{aligned}$$

Par ailleur, nous allons réécrire $P_n = \mu + \sigma_n Z_n$ où $\mu = \mathbb{E}(P_n)$ et $\sigma_n^2 = \mathrm{Var}\,(P_n)$; ainsi Z_n sera une v.a. centrée ($\mathbb{E}(Z_n) = 0$) et réduite ($\mathrm{Var}\,(Z_n) = 1$). Pour cela, il suffit de poser

$$\mu = \mathbb{E}(P_n) = \mathbb{E}\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n}(\mathbb{E}(X_1) + \ldots + \mathbb{E}(X_n)) = \frac{n}{n}p = p.$$

$$\sigma_n^2 = \mathrm{Var}\,(P_n) = \mathrm{Var}\,\left(\frac{1}{n}(X_1 + \ldots + X_n)\right) = \frac{1}{n^2}(\mathrm{Var}\,(X_1) + \ldots + \mathrm{Var}\,(X_n)) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

où $\sigma^2 = \mathrm{Var}\,(X_i) = p(1-p)$, d'où finalement $\boxed{\sigma_n = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}}$.

A présent, il est facile de voir que l'évènement E_n peut encore s'écrire

$$E_n = \{Z_n \in [z_-, z_+]\} \text{ avec } z_- = -\frac{\sqrt{n}}{\sigma}\Delta^+ \text{ et } z_+ = +\frac{\sqrt{n}}{\sigma}\Delta^-.$$

11.3 Approximation normale

La préparation que nous avons faite de P_n en l'écrivant $P_n = p + \frac{\sigma}{\sqrt{n}}Z_n$ est dictée par l'usage que nous allons faire du théorème limite central (TLC) ; en effet, ce théorème assure précisément que la v.a. Z_n définie par cette relation tends en loi vers $Z \rightsquigarrow \mathcal{N}(0,1)$. Ainsi, $\mathbb{P}(E_n) = \mathbb{P}(\{Z_n \in [z_-, z_+]\}) = F_{Z_n}(z_+) - F_{Z_n}(z_-) \approx F_Z(z_+) - F_Z(z_-)$ dès que n est assez grand.

Choisissons à présent $\Delta^+ = \Delta^- =: \Delta$, ce qui implique que $z_- = -z_+$. Nous cherchons $\Delta = \Delta_\alpha$ tel que $\mathbb{P}(E_n^c) = \alpha$, c'est à dire $\mathbb{P}(E_n) = 1 - \alpha$. Comme la densité de Z est paire, ceci revient à choisir z_+ tel que $F_Z(z_+) = 1 - \frac{\alpha}{2}$. Comme $z_+ = \frac{\sqrt{n}}{\sigma}\Delta$, nous obtenons

$$\boxed{\Delta_\alpha = \frac{\sigma}{\sqrt{n}}F_Z^{-1}(1-\tfrac{\alpha}{2}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}F_Z^{-1}(1-\tfrac{\alpha}{2}) \leq \frac{1}{2\sqrt{n}}F_Z^{-1}(1-\tfrac{\alpha}{2}) =: \Delta'_\alpha,}$$

l'inégalité résultant du fait que pour $p \in [0,1]$ on a $p(1-p) \leq \frac{1}{4}$.

En utilisant le tableau de la fonction de répartition F_Z de la loi normale centrée réduite, nous obtenons les valeurs suivantes pour Δ'_α, et donc que $\mathbb{P}(p \notin [P_n - \Delta'_\alpha, P_n + \Delta'_\alpha]) \leq \alpha$.

α	1%	2%	3%	4%	5%	6%	7%	8%	9%	10%
$A = F_Z^{-1}(1-\frac{\alpha}{2})$	2,576	2,326	2,170	2,054	1,960	1,881	1,812	1,751	1,695	1,645
$n = 400$	0,064	0,058	0,054	0,051	0,049	0,047	0,045	0,044	0,042	0,041
$n = 850$	0,044	0,040	0,037	0,035	0,034	0,032	0,031	0,030	0,029	0,028
$n = 1000$	0,041	0,037	0,034	0,032	0,031	0,030	0,029	0,028	0,027	0,026
$n = 2000$	0,029	0,026	0,024	0,023	0,022	0,021	0,020	0,020	0,019	0,018

Valeurs de $\Delta'_\alpha := \frac{1}{2\sqrt{n}}F_Z^{-1}\left(1-\frac{1}{\alpha}\right) \geq \frac{\sigma}{\sqrt{n}}F_Z^{-1}\left(1-\frac{1}{\alpha}\right)$

Exercice : Soient X_1, ..., X_n i.i.d. avec $X_i \rightsquigarrow \mathcal{B}(1,p)$ et $P_n := \frac{1}{n}(X_1 + \ldots + X_n)$ l'estimateur usuel de p. Montrer que si n est suffisamment grand pour justifier l'approximation normale donnée par le TLC ($n \geq 30$), on a que $\alpha := \mathbb{P}(p \notin [P_n - \frac{1}{\sqrt{n}}, P_n + \frac{1}{\sqrt{n}}]$ est inférieur à (et proche de) 5%.

Index