Probabilités et statistiques

Cours de Master 1

Bernard Delyon[1]

24 octobre 2013

1. IRMAR, Université Rennes I, Campus de Beaulieu, 35042 Rennes cédex, France.

Table des matières

I

Espaces probabilisés, variables aléatoires

L'objectif de ce chapitre est de repréciser les concepts de base de théorie de la mesure et de particulariser à théorie des probabilités. On suppose connue la théorie des probabilités sur un espace fini.

I.1 Mesure et probabilité

L'espace fondamental pour modéliser une expérience aléatoire est un triplet $(\Omega, \mathscr{F}, P)$. Ω contient toutes les réalisations possibles. Par exemple si l'expérience consiste à jeter deux dés, $\Omega = \{1,2,3,4,5,6\}^2$. $\mathscr{F}$ est un ensemble de parties $A \subset \Omega$ appelées événements, pour lesquelles est définie la probabilité $P(A)$. Dans cet exemple simple, $\mathscr{F}$ est l'ensemble de toutes les parties de Ω, et si le dé n'est pas pipé, on a $P(A) = \#A/36$ (nombre de cas favorables sur nombre de cas possibles). L'événement «une paire apparaît» a pour probabilité 1/6.

Dans le cas où Ω est dénombrable, on aura toujours encore $\mathscr{F} = \mathscr{P}(\Omega)$ et P sera caractérisée par la probabilité de chaque réalisation : $P(A) = \sum_{\omega \in A} P(\{\omega\})$. Par exemple $\Omega = \mathbb{N}$ et $P(\{n\}) = e^{-\lambda}\lambda^n/n!$ (loi de Poisson de paramètre λ).

Dans le cas où Ω n'est pas dénombrable, on ne peut malheureusement pas généralement avoir $\mathscr{F} = \mathscr{P}(\Omega)$ (ensemble des parties de Ω) ; c'est ce qui fait toute la complication de la théorie de la mesure. Par exemple la mesure uniforme sur $\Omega = [0,1]$ (mesure de Lebesgue), celle qui associe à toute réunion finie disjointe d'intervalle la somme de leurs longueurs, ne peut être étendue à tout $\mathscr{P}(\Omega)$ sans que ne soient violés des conditions naturelles d'invariance par translation ou d'additivité (définition I.2) ; c'est étonnant mais c'est ainsi (Ceci a été essentiellement mis en évidence par Giuseppe Vitali in 1905, par la construction de ce que l'on appelle aujourd'hui les ensembles de Vitali). Notons que dans cet exemple $P(\{\omega\}) = 0$ pour tout $\omega \in [0,1]$. Il se trouve que le bon cadre est le suivant :

Définition I.1 *(Tribu, espace mesurable) Une tribu (ou σ-algèbre) $\mathscr{F}$ sur E est une classe de parties de E vérifiant :*
(i) $E \in \mathscr{F}$
(ii) si $A \in \mathscr{F}$ alors $A^c \in \mathscr{F}$
(iii) si $A_n \in \mathscr{F}$ pour tout $n \in \mathbb{N}$, alors $\cup_n A_n \in \mathscr{F}$
La paire $(E, \mathscr{F})$ est appelée «espace mesurable». Les parties de E qui n'appartiennent pas à $\mathscr{F}$ sont dites non-mesurables.

Il s'ensuit que $\emptyset \in \mathscr{F}$, que $\mathscr{F}$ est stable par réunion finie, et que $\mathscr{F}$ est stable par intersection finie ou dénombrable.

Tribu engendrée. Il est simple de montrer qu'une intersection quelconque de tribu sur E est encore une tribu. Si $\mathscr{A}$ est une famille de parties de E, on note $\sigma(\mathscr{A})$ l'intersection de toutes les tribus contenant $\mathscr{A}$. C'est la plus petite tribu contenant $\mathscr{A}$ et on l'appelle «tribu engendrée par $\mathscr{A}$».

La tribu borélienne sur $\mathbb{R}$, notée $\mathscr{B}(\mathbb{R})$, est la tribu engendré par les intervalles. C'est également la tribu engendrée par les intervalles de la forme $]-\infty, t]$.

La tribu borélienne sur $\mathbb{R}^d$, notée $\mathscr{B}(\mathbb{R}^d)$, est la tribu engendrée par les pavés $\prod_{i=1}^d]s_i, t_i]$. Comme ces pavés s'expriment comme réunions et intersections d'ensembles $A_{i,t}$ de la forme $A_{i,t} = \{x : x_i \leq t\}$, c'est également la tribu engendrée par les $A_{i,t}$ (spécifiquement : $\prod_{i=1}^d]s_i, t_i] = \cap_i (A_{i,t_i} \backslash A_{i,s_i})$; faire un dessin en dimension 2).

Il est en réalité très difficile de fabriquer des ensembles qui ne sont pas $\mathscr{B}(\mathbb{R}^d)$-mesurables ; il faut toujours faire une construction assez alambiquée ; c'est presque un théorème. C'est pourquoi souvent la propriété de mesurabilité ne sera pas vérifiée dans les exemples de ce cours, même si ce n'est pas réellement difficle.

DÉFINITION I.2 *(Espace mesuré, probabilisé) Une mesure sur un espace mesurable $(E, \mathscr{F})$ est une application μ de $\mathscr{F}$ dans $\mathbb{R} \cup \{+\infty\}$ vérifiant*

(i) $\mu(\emptyset) = 0$

(ii) σ*-additivité : si* $(A_n)_{n\geq 1}$ *est une suite d'ensembles disjoints de* $\mathscr{F}$*, alors* $\mu(\cup_n A_n) = \sum_n \mu(A_n)$.

Si $\mu(E) = 1$*, il s'agit d'une mesure de probabilité.*

La théorie des probabilité n'est donc que le cas particulier de la théorie de la mesure où la masse totale vaut 1. Cette définition contient le strict nécessaire mais d'autres propriétés en découlent :

PROPOSITION I.3 *Si $(E, \mathscr{F}, \mu)$ est un espace mesuré, et $A, B \in \mathscr{F}$*

$$\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B).$$

Soit $B_1, B_2, ... \in \mathscr{F}$:

1. $\mu(\cup_n B_n) \leq \sum_n \mu(B_n)$.
2. *Si pour tout n $B_n \subset B_{n+1}$, alors* $\mu(\cup_n B_n) = \lim_n \mu(B_n)$.
3. *Si* $\mu(B_1) < \infty$*, et si pour tout n $B_{n+1} \subset B_n$, alors* $\mu(\cap_n B_n) = \lim_n \mu(B_n)$.

Démonstration: Notons que le point (ii) de la definition I.2 reste vrai pour un nombre fini d'ensembles (il suffit de prendre $A_n = \emptyset$ pour $n \geq n_0$) ; on en déduit alors facilement le premier point. Les deux points suivants s'obtiennnent en exprimant les B_n à l'aide des ensembles disjoints $A_n = B_n \backslash (B_1 \cup ... \cup B_{n-1})$. Le dernier s'obtient en appliquant le précédent aux ensemble $B_1 \backslash B_n$ et en utilisant que $\mu(B_1 \backslash B_n) = \mu(B_1) - \mu(B_n)$. ∎

La condition $\mu(B_1) < \infty$ du point 3 n'est pas superflue : penser à $B_n = [n, +\infty[$ sur $\mathbb{R}$ muni de la mesure de Lebesgue. Elle est toujours satisfaite si μ est une probabilité.

COROLLAIRE I.4 *Si $(\Omega, \mathscr{F}, P)$ est un espace probabilisé, et A_n une famille dénombrable d'ensembles de probabilité 1 alors*

$$P(\cap_n A_n) = 1.$$

Démonstration: $P(\cap_n A_n) = 1 - P((\cap_n A_n)^c) = 1 - P(\cup_n A_n^c) \geq 1 - \sum_n P(A_n^c) = 1$ ∎

Existence de mesures. Un des objectifs de la théorie de la mesure est de répondre au problème suivant : Soit E un espace, $\mathscr{A}$ une classe d'ensembles et P une fonction définie sur $\mathscr{A}$, peut-on étendre P de façon unique à $\mathscr{F} = \sigma(\mathscr{A})$ de sorte à obtenir un espace mesuré (probabilisé si $P(E) = 1$) ?

Par exemple $E = \mathbb{R}$, et $\mathscr{A}$ est la classe des intervalles.

Le théorème d'extension de Carathéodory est l'outil clé pour avoir une réponse positive ; la classe $\mathscr{A}$ et la fonction P de départ doivent satisfaire certaines conditions.

Nous n'aborderons pas ces questions. En particulier la mesure de Lebesgue est désormais supposée construite.

Exemples.
MESURE DE LEBESGUE. C'est la seule mesure sur $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ telle que la mesure d'un pavé $\prod_i]s_i, t_i]$ soit égale à son volume $\prod_i (t_i - s_i)$.

MASSE DE DIRAC. Soit $x \in E$, la masse de Dirac en x, notée δ_x, est la mesure $A \mapsto 1_{x \in A}$.

MESURE DE COMPTAGE SUR $\mathbb{N}$. C'est la mesure qui affecte le poids 1 à chaque entier : Si $A \subset \mathbb{N}$, $\mu(A) = Card(A)$.

Tribu produit, mesure produit. Si $(E_1, \mathscr{F}_1)$ et $(E_2, \mathscr{F}_2)$ sont deux espaces mesurables on définit la tribu produit sur $E = E_1 \times E_2$ comme la tribu $\mathscr{F}$ engendrée par les ensembles de la forme $A_1 \times A_2$ avec $A_1 \in \mathscr{F}_1$ et $A_2 \in \mathscr{F}_2$. Elle est notée $\mathscr{F}_1 \otimes \mathscr{F}_2$. Si les espaces sont mesurés avec des mesures μ_1 et μ_2, on définit la mesure produit comme la seule mesure μ sur $(E, \mathscr{F})$ satisfaisant $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$. L'existence et l'unicité de cette mesure est garantie par les théorèmes généraux.

Dans le cas d'espaces de probabilités Ω_1 et Ω_2, les éléments de $\Omega = \Omega_1 \times \Omega_2$, c-à-d les paires (ω_1, ω_2), représentent la combinaison des deux expériences, et la mesure produit postule leur indépendance, comme dans l'exemple $\Omega = \{1, 2, 3, 4, 5, 6\}^2$ présenté plus haut ; $\mathscr{F} = \mathscr{P}(\Omega)$.

On définit de même la tribu produit et la mesure produit sur un produit fini d'espaces. La mesure produit de la mesure de Lebesgue sur $\mathbb{R}^d$ est bien la mesure présentée précédemment.

Tribu complétée. Une mesure μ étant donnée, soit $\mathscr{N}$ la classe des ensembles négligeables, c-à-d contenus dans un ensemble de mesure nulle. On vérifie que la classe formées des ensembles de la forme $F \cup N$, $F \in \mathscr{F}$, $N \in \mathscr{N}$ forme une tribu ; elle s'appelle tribu complétée. D'un point de vue applicatif F et $F \cup N$ sont indistinguables et cette augmentation peut paraître purement cosmétique ; d'un point de vue mathématique elle a des justifications qui viendront en leur temps [1]. La complétée de la tribu de Borel pour la mesure de Lebesgue sur $\mathbb{R}^d$ s'appelle la tribu de Lebesgue.

I.2 Intégration

La mesure étant définie, on peut définir l'intégrale des fonctions étagées, c-à-d celles qui sont une combinaison linéaire finie d'indicateurs d'ensemble qu'on peut toujours supposer disjoints :

$$f(x) = \sum_i a_i 1_{A_i}(x)$$

par

$$\mu(f) = \int f(x)\mu(dx) = \sum_i a_i \mu(A_i). \tag{I.1}$$

Ce qui est conforme à la notion habituelle d'intégrale. Dans le cas d'un espace probabilisé, on note plutôt

$$E[f] = \sum_i a_i P(A_i).$$

Ce paragraphe a pour objet de montrer comment cette intégrale s'étend à une classe bien plus grande de fonctions, dites fonctions intégrables. Ceci généralise l'intégrale de Riemann.

1. Notons ici, pour le lecteur très avancé, deux points rarement soulignés à propos de la complétion :
1/ Il faut être prudent avec la complétion : Si l'on note $\mathscr{C}$ la complétée de $\mathscr{B}(\mathbb{R})$ pour la mesure de Lebesgue, il existe une fonction continue f (même bijective croissante) et des ensembles de $\mathscr{C}$ dont l'image réciproque par f n'appartient pas à $\mathscr{C}$; la fonction f, bien que continue n'est donc pas mesurable de $(\mathbb{R}, \mathscr{C})$ dans $(\mathbb{R}, \mathscr{C})$! Elle est bien entendu mesurable de $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ dans $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ et a fortiori de $(\mathbb{R}, \mathscr{C})$ dans $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. C'est pour cela que $\mathbb{R}$ en tant qu'espace d'arrivée est toujours muni de sa tribu de Borel. En revanche $\mathbb{R}$ comme espace de départ est souvent affecté de la tribu complétée, dite tribu de Lebesgue. Voir *Counterexamples in analysis*, par B.R. Gelbaum et J.M.H. Olmsted, Holden-Day, 1964, § 1.8.16 et 1.8.38.
2/ La complétion est parfois utile : considérons le carré unité muni de la tribu de Borel et les deux sous-tribus $\mathscr{X}$ et $\mathscr{Y}$ des ensembles ne dépendant respectivement que de la première et de la deuxième coordonnée. Soient $X(\omega)$ et $Y(\omega)$ les coordonnées du pont ω sur les deux axes, f une fonction continue bornée et $Z = f(X, Y)$. Si l'on considère la mesure uniforme sur la diagonale, alors $E[Z|\mathscr{X}] = f(X, X)$ et $E[Z|\mathscr{Y}] = f(Y, Y)$. Ces deux fonctions sont égales avec probabilité 1 à Z et pourtant $\mathscr{X} \cap \mathscr{Y} = \{\emptyset, \Omega\}$, en particulier $E[Z|\mathscr{X} \cap \mathscr{Y}] = E[Z]$; en résumé, l'espérance conditionnelle de Z sachant $\mathscr{X}$ ou $\mathscr{Y}$ est Z mais sachant $\mathscr{X} \cap \mathscr{Y}$ c'est $E[Z]$! En revanche les deux tribus complétées sont bien chacune égales à la complétée de la tribu de Borel, et donc le paradoxe disparaît si l'on complète au préalable. Ne pas compléter une tribu conditionnante peut donc conduire à des situations très contraires à l'intuition.

Fonctions mesurables

DÉFINITION I.5 *Soit $(E, \mathscr{F})$ un espace mesurable. Et $f : E \to \mathbb{R}^d$. On dit que f est mesurable, ou $\mathscr{F}$-mesurable, ou borélienne, si l'image réciproque par f de tout ensemble borélien appartient à $\mathscr{F}$.*

Tout indicateur d'ensemble mesurable, $1_A(x)$, est donc une fonction mesurable.

Si $(E, \mathscr{F}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$: Toute fonction continue par morceaux de $\mathbb{R}$ dans $\mathbb{R}$ est $\mathscr{B}(\mathbb{R})$-mesurable. Les fonctions intégrables au sens de Riemann sont $\mathscr{B}(\mathbb{R})$-mesurables ; il est en fait très difficile de fabriquer des fonctions de $\mathbb{R}$ dans $\mathbb{R}$ qui ne le soient pas.

PROPOSITION I.6 *Soit $f : E \to \mathbb{R}$ une fonction telle que l'image réciproque de tout intervalle $]-\infty, t]$ est un ensemble $\mathscr{F}$-mesurable, alors f est mesurable.*

Soit $f : E \to \mathbb{R}^d$ une fonction telle que chaque composante soit mesurable alors f est mesurable.

Démonstration: Pour le premier point, notons que la classe des ensembles dont l'image réciproque est $\mathscr{F}$-mesurable forme une tribu (à vérifier !) et contient les intervalles $]-\infty, t]$, elle contient donc la tribu engendrée par ces intervalles. Le deuxième point se démontre de façon similaire. ∎

La proposition suivante est conséquence directe de la définition :

PROPOSITION I.7 *Si $f : E \to \mathbb{R}^d$ est mesurable et $\varphi : \mathbb{R}^d \to \mathbb{R}^n$ est borélienne (l'image réciproque d'un borélien est borélien), alors $\varphi(f) : E \to \mathbb{R}^n$ est mesurable*

Rappelons que les fonctions continues par morceaux sont boréliennes. On obtient donc en particulier que les combinaisons linéaires finies d'indicateurs d'ensembles mesurables, appelées fonction étagées, sont encore des fonctions mesurables.

Concernant les suites de fonctions mesurables :

PROPOSITION I.8 *Si $f_n : E \to \mathbb{R}^d$ est une suite de fonctions mesurables alors $\sup f_n(x)$ et $\overline{\lim} f_n(x)$ sont mesurables. En particulier si $f_n(x)$ converge vers $f(x)$, alors $f(x)$ est une fonction mesurable.*

Démonstration: Comme

$$\overline{\lim} f_n(x) = \lim_n \sup_{p>n} f_p(x) = \inf_n \sup_{n<p} f_p(x)$$

il suffit de montrer que le sup est mesurable (on l'obtient pour l'inf en notant que $\inf f_n = -\sup(-f_n)$). Mais

$$\{x : \sup_n f_n \leq t\} = \cap_n \{x : f_n \leq t\}$$

est bien $\mathscr{F}$-mesurable. ∎

Intégration. Soit μ une mesure sur E. On peut définir de façon cohérente l'intégrale de toute fonction positive mesurable, notée $\int_E f(x)\mu(dx)$, ou $\int_E f(x)d\mu(x)$, ou $\mu(f)$ et f est dite intégrable si cette intégrale est finie. Une fonction mesurable g de signe quelconque est dite intégrable si $|g|$ est intégrable et l'on définit son intégrale par

$$\mu(g) = \int_E g(x)\mu(dx) = \int_E g_+(x)\mu(dx) - \int_E g_-(x)\mu(dx).$$

Rappelons la démarche entreprise pour définir l'intégrale : l'intégrale d'une fonction positive $f(x)$ étagée est définie par (I.1). Il est simple d'approximer une fonction mesurable positive f par une suite de fonctions étagées f_n : il suffit d'arrondir $\min(f, n)$ au plus proche multiple de $1/n$ inférieur (on a au plus n^2 valeurs) ; l'intégrale de f est alors définie comme la limite croissante

$$\mu(f) = \lim_n \uparrow \mu(f_n).$$

Sous l'hypothèse que f est intégrable on montre alors que f_+ et f_- le sont également et l'on définit $\mu(f) = \mu(f_+) - \mu(f_-)$. On vérifie que cette intégrale correspond bien à celle définie pour les v.a. étagées et possède bien les vertus attendue pour une intégrale (linéarité...).

Une conséquence de cette construction est que pour toute fonction intégrable f il existe une suite de fonctions étagées f_n telles que $\mu(|f_n - f|)$ tendent zero 0 ; cette propriété est très utile pour étendre à toutes les fonctions intégrables un résultat déjà démontré pour les fonctions étagées. Noter encore que toute fonction positive est limite croissante de fonctions étagées.

Rappelons également qu'une fonction mesurable positive peut être infinie en certains points sans que les règles de calcul ne soient affectées tant que l'on ne tombe pas sur des expressions indéterminées comme $+\infty - \infty$. Si $\mu(x : f(x) = +\infty) > 0$ alors son intégrale est infinie. Si la fonction n'est plus positive, $\mu(f)$ n'est défini que si $\mu(|f|) < +\infty$.

Classes de fonctions. Soit f une fonction mesurable, l'ensemble des fonctions g telles que $\mu(\{x : f(x) \neq g(x)\}) = 0$ est la classe de fonctions (relativement à μ) associée à f [2]. Dans la suite on ne distinguera pas les fonctions d'une même classe d'équivalence (de même qu'on ne distingue pas $1,6999999...$ et $1,7$). Ceci permet par exemple que l'application $(f,g) \mapsto \int |f(x) - g(x)|\mu(dx)$ soit effectivement une distance. On dit qu'une fonction f n'est définie que *presque partout*.

Théorèmes de base. Les résultats de ce paragraphe sont des outils clé pour beaucoup de calculs, particulièrement le troisième, tant en analyse (pour l'intégration des fonctions) qu'en théorie des probabilités

THÉORÈME I.9 *Soit f_n une suite de fonctions mesurables, on a les trois résultats suivants :*

THÉORÈME DE CONVERGENCE MONOTONE. *Si pour tout $n \geq 1$, $0 \leq f_n \leq f_{n+1}$, alors*

$$\int_E \sup_n f_n(x)\ \mu(dx) = \sup_n \int_E f_n(x)\ \mu(dx).$$

THÉORÈME DE CONVERGENCE DOMINÉE DE LEBESGUE. *Si pour tout x $f_n(x) \to f(x)$ et s'il existe une fonction mesurable g telle que $|f_n| \leq g$ et $\mu(g) < \infty$ alors*

$$\lim_n \int_E f_n(x)\ \mu(dx) = \int_E f(x)\ \mu(dx).$$

Mentionnons également le

LEMME I.10 *(LEMME DE FATOU). Si pour tout $n \geq 1$, $f_n \geq 0$, alors*

$$\int_E \underline{\lim}_n f_n(x)\ \mu(dx) \leq \underline{\lim}_n \int_E f_n(x)\ \mu(dx).$$

Pour se souvenir du sens de l'inégalité dans le lemme de Fatou, on peut utiliser l'exemple $f_n(x) = 1_{n<x<n+1}$ sur $E = \mathbb{R}$ muni de la mesure de Lebesgue : le membre de gauche est nul et le membre de droite vaut 1.

Une conséquence (ou une réécriture) du théorème de convergence monotone est que si les g_i sont des fonctions mesurables ≥ 0 on a

$$\mu\left(\sum_n g_n\right) = \sum_n \mu(g_n)$$

les deux membres pouvant être infinis. Le théorème de convergence dominée a deux conséquences importantes classiques

1. CONTINUITÉ D'UNE INTÉGRALE DÉPENDANT D'UN PARAMÈTRE. Si pour tout $t \in]a,b[$, $x \mapsto f_t(x)$ est mesurable. et pour tout x, $f_t(x) \to f_c(x)$ lorsque $t \to c$, et si de plus il existe une fonction g intégrable telle que

   $$\forall x \in E, \forall t \in]a,b[,\ |f_t(x)| \leq g(x),$$

   alors

   $$\lim_{t\to c} \int_E f_t(x)\mu(dx) = \int_E f_c(x)\mu(dx).$$

   (Car pour toute suite $t_n \to c$, $\mu(f_{t_n}) \to \mu(f_c)$.)

2. On pourra vérifier à l'aide des théorèmes précédents que $\{x : f(x) \neq g(x)\}$ est bien mesurable (considérer $h(x) = (f(x), g(x))$).

2. Dérivation sous le signe intégral. Si pour tout $t \in]a,b[$, $x \mapsto f_t(x)$ est intégrable et pour tout x, $t \to f_t(x)$ est dérivable sur $]a,b[$, et si de plus il existe une fonction g intégrable telle que

$$\forall x \in E, \forall t \in]a,b[,\ |f'_t(x)| \leq g(x),$$

alors pour $a < t < b$ la fonction $t \mapsto \int_E f_t(x)\mu(dx)$ est dérivable en t et :

$$\frac{d}{dt}\int_E f_t(x)\mu(dx) = \int_E \left(\frac{d}{dt}f_t(x)\right)\mu(dx).$$

(Par application du résultat précédent à $g_h(x) = h^{-1}(f_{t+h}(x) - f_t(x))$ et $c = 0$.)

N'oublions pas le théorème de permutation des intégrales :

Théorème I.11 *(Fubini-Tonelli) Soient $(E,\mathscr{F},\mu)$ et $(F,\mathscr{G},\nu)$ deux espaces mesurés munis de mesures σ-finies*[3] *et $(E\times F, \mathscr{F}\otimes\mathscr{G}, \mu\otimes\nu)$ l'espace produit muni de la tribu produit et de la mesure produit. Si $f : E\times F \to \mathbb{R}$ est une fonction borélienne ≥ 0 de deux variables alors, la fonction d'une variable $x \mapsto \int f(x,y)\ \nu(dy)$ est borélienne et*

$$\int_E \left(\int_F f(x,y)\ \nu(dy)\right)\ \mu(dx) = \int_F \left(\int_E f(x,y)\ \mu(dx)\right)\ \nu(dy) \tag{I.2}$$

les deux termes pouvant être infinis. Cette quantité est également l'intégrale par rapport à la mesure produit $\mu\otimes\nu$.

Soit g une fonction borélienne telle que (I.2) soit fini pour $f(x,y) = |g(x,y)|$, alors g est intégrable pour la mesure produit, pour presque tout x la fonction $g(x,y)$ est intégrable en y, la fonction $\int_F g(x,y)dy$ est intégrable en x, et g satisfait (I.2), qui représente l'intégrale par rapport à la mesure produit.

La première partie est le théorème de Tonelli et la seconde celui de Fubini. A priori ce théorème s'applique donc en deux temps, une première fois avec $|f|$ pour vérifier l'intégrabilité puis une seconde avec f. Tout ceci reste vrai pour un produit de plus de deux espaces.

Ce théorème est en particulier très utilisé dans le cas $E = F = \mathbb{R}$ et μ et ν sont la mesure de Lebesgue. Avec la mesure de comptage (celle qui affecte la masse 1 à chaque entier) on obtient

$$\sum_i \sum_j f(i,j) = \sum_j \sum_i f(i,j) \tag{I.3}$$

pour $f \geq 0$ et sinon, cela est vrai si $|f|$ est sommable. On a de même

$$\int \left(\sum_i f(x,i)\right) dx = \sum_i \int f(x,i)dx$$

pour toute fonction positive. Noter que dans (I.3), si $|f|$ n'est pas sommable, les deux membres peuvent exister et être différents : si $f(i,j) = -1_{0\leq i<j}2^{-(j-i)} + 1_{i=j}$ la somme sur i vaut 2^{-j} et la somme sur j vaut 0 pour tout i (le plus simple pour le voir est d'écrire la «matrice» de terme $f(i,j)$).

Caractérisation des mesures de probabilité sur $\mathbb{R}^d$. Pour fixer les idées, rappelons que les mesures de probabilités *usuelles* sur $\mathbb{R}$ se séparent en deux classes (il en existe bien d'autres...) :

1. Les mesures ponctuelles : pour toute fonction f borélienne bornée, $\int f(x)\mu(dx) = \sum_i p_i f(x_i)$; ou encore $\mu = \sum_i p_i \delta_{x_i}$.
2. Les mesures à densité : pour toute fonction f borélienne bornée, $\int f(x)\mu(dx) = \int f(x)p(x)dx$

On appelle fonction de répatition d'une mesure de probabilité μ sur $\mathbb{R}$ la fonction

$$F(t) = \mu(]-\infty, t]).$$

3. Cette condition signifie qu'il existe une suite $E_n \in \mathscr{F}$ telle que $\mu(E_n) < +\infty$ et $E = \cup E_n$, et de même pour $(F,\mathscr{G},\nu)$.

THÉORÈME I.12 *Toute mesure de probabilité sur $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ est caractérisée par sa fonction de répartition.*

Elle est également caractérisée par les valeurs de $\int f(x)\mu(dx)$ lorsque f varie dans l'ensemble des fonctions positives continues bornées.

Démonstration: (Schéma) Le premier résultat vient de la propriété générale suivante appliquée à la classe $\mathscr{C}$ des intervalles $]-\infty, t]$: *si $\mathscr{C}$ est une classe de parties de Ω stable par intersection dénombrable, toute mesure sur $(\Omega, \sigma(\mathscr{C}))$ est caractérisée par ses valeurs sur $\mathscr{C}$.* Pour le deuxième résultat, noter que si f_n est une suite de fonctions continues positives bornées par 1 convergeant ponctuellement vers la fonction indicatrice de $]-\infty, t]$, alors en vertu du théorème de convergence dominée on a $\mu(]-\infty, t]) = \lim_n \int f_n(x)\mu(dx)$. ∎

Il est facile de vérifier que

1. si μ a une densité continue, elle est donnée par la dérivée de F.
2. si $\mu = \sum_i p_i \delta_{x_i}$ est ponctuelle, F est constante par morceaux et l'amplitude de son saut en x_i est p_i.

On définit également une fonction de répartition pour les mesures sur $\mathbb{R}^d$ mais elle est moins utilisée : $F(t_1, ...t_d) = \mu(]-\infty, t_1] \times ... \times]-\infty, t_d])$.

EXEMPLE 1 : Si la mesure est la distribution exponentielle de paramètre λ, c-à-d de densité $1_{x>0}\lambda e^{-\lambda x}$, on trouve $F(t) = 1 - e^{-\lambda t}$.

EXEMPLE 2 : Pour la distribution géométrique de paramètre p, c-à-d la mesure qui affecte à chaque entier $n \geq 0$ la masse $p(1-p)^n$, on trouve $F(t) = \sum_{n \leq t} p(1-p)^n = 1 - (1-p)^{[t]+1}$.

THÉORÈME I.13 *Toute mesure de probabilité sur $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$ est caractérisée par sa fonction caractéristique :*

$$\varphi(t) = \int e^{i\langle x,t\rangle}\mu(dx), \qquad t \in \mathbb{R}^d.$$

EXEMPLE 1 : Distribution exponentielle $\mathcal{E}(\lambda)$:

$$\varphi(t) = \int_0^{+\infty} e^{ixt}e^{-\lambda x}\lambda dx = \frac{\lambda}{\lambda - it}.$$

EXEMPLE 2 : Distribution géométrique $\mathcal{G}(p)$:

$$\varphi(t) = \sum_{n \geq 0} p(1-p)^n e^{int} = \frac{p}{1-(1-p)e^{it}}.$$

EXEMPLE 3 : Mesure gaussienne $\mathcal{N}(m, \sigma^2)$: C'est la mesure de densité

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-m)^2}{2\sigma^2}}. \tag{I.4}$$

On admettra que sa fonction caractéristique vaut

$$\varphi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} \int e^{ixt} e^{-\frac{(x-m)^2}{2\sigma^2}} dx = e^{itm} e^{-t^2\sigma^2/2}.$$

EXEMPLE 4 : Mesure gaussienne sur $\mathbb{R}^d$, $\mathcal{N}(m, R)$: C'est la mesure de densité

$$p(x) = \frac{1}{\sqrt{2\pi}^d \sqrt{\det(R)}} e^{-\frac{1}{2}(x-m)^T R^{-1}(x-m)}. \tag{I.5}$$

Sa fonction caractéristique vaut

$$\varphi(t) = e^{itm} e^{-t^T Rt/2}.$$

FONCTION GÉNÉRATRICE. Pour une distribution portée par les entiers on préfère parfois considérer la fonction génératrice

$$G(z) = \sum_{n=0}^{+\infty} z^n \mu(\{n\})$$

qui est naturellement définie pour $|z| \leq 1$ car dans ce cas la série est absolument (et même normalement) convergente, et l'on a $\varphi(t) = G(e^{it})$, ce qui fait qu'elle caractérise effectivement la distribution. Par souci de simplicité, nous n'en parlerons pas davantage.

Densité. Soit deux mesures μ et ν sur $(E, \mathscr{F})$; s'il existe une fonction $p(x) \geq 0$ telle que pour tout $A \in \mathscr{F}$

$$\nu(A) = \int_E 1_A(x) p(x) \mu(dx)$$

on dit que p est la densité de ν par rapport à μ. Cette relation équivaut à dire que pour toute fonction positive $\mathscr{F}$-mesurable

$$\int_E f(x)\nu(dx) = \int_E f(x)p(x)\mu(dx).$$

Clairement, si $p > 0$, alors p^{-1} est la densité de μ par rapport à ν (appliquer l'idéntité à $g(x) = f(x)p(x)^{-1}$).

EXEMPLES. La loi exponentielle $\mathcal{E}(\lambda)$ a pour densité $\lambda e^{-\lambda x}$ par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$. La loi de Poisson $\mathcal{P}(\lambda)$ a pour densité $\lambda^n e^{-\lambda x}/n!$ par rapport à la mesure de comptage sur $\mathbb{N}$. La mesure $\mathcal{B}(n, q)$ a pour densité $q^k(1-q)^{n-k}p^{-k}(1-p)^{-n+k}$ par rapport à $\mathcal{B}(n, p)$.

I.3 Variables aléatoires

DÉFINITION I.14 *Une variable aléatoire sur $(\Omega, \mathscr{F}, P)$ est une fonction mesurable à valeurs dans $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$.*

On a vu que des sommes de v.a., des limites de v.a. sont encore des v.a. Si on les compose par des fonctions mesurables (p.ex. continues par morceaux), on obtient encore des v.a. Comme on l'a vu plus haut, les v.a. sont en fait des classes de v.a.

Une fonction mesurable à valeurs dans un espace fini dénombrable muni de la tribu de ses parties sera également considérée comme une variable aléatoire, par identification de cet espace à $\{1, \ldots n\} \subset \mathbb{R}^d$ ou $\mathbb{N} \subset \mathbb{R}^d$.

> Mentionnons que pour les raisons données à la note 1 page 7, $\mathbb{R}^d$ n'est pas muni de sa tribu complétée. En revanche Ω l'est généralement. C'est pour cette raison que par souci de clarté nous considérons ici que les v.a. sont forcément à valeurs dans $\mathbb{R}^d$ (ou bien dans un espace dénombrable que l'on peut toujours identifier à $\mathbb{N}$).

Ensembles de mesure nulle, classes de variables et propriétés presque sûres. Une propriété sera dite presque sûrement vérifiée, ou encore vérifiée «avec probabilité 1», si l'ensemble des ω où elle n'est pas satisfaite est de mesure 0. Par exemple obtenir infiniment de fois 1 quand on lance un dé infiniment de fois, ou ne pas obtenir $1/2$ quand on tire un nombre au hasard sur l'intervalle $[0, 1]$.

Le corollaire I.4 assure qu'un nombre dénombrable de propriétés presque sûrement vérifiées est presque sûrement simultanément vérifié.

On dira en particulier que X_n converge presque sûrement vers X s'il existe un ensemble A de mesure 1 telle que

$$\forall \omega \in A,\ X_n(\omega) \longrightarrow X(\omega). \tag{I.6}$$

En fait il ne peut en être autrement puisque les X_n ne sont définies que presque sûrement. Notons que cette propriété ne dépend pas des représentants choisis (A change pour chaque choix de représentants) et signifie que les variables $X_n 1_A$ (équivalente à X_n) convergent vers $X 1_A$ (équivalente à X).

On verifie que le théorème I.9 reste valide si l'on remplace les inégalités et les limites par des inégalités et des limites presques sûres.

Par exemple la loi forte des grands nombres est vraie presque sûrement mais n'est généralement pas vérifiée pour tout ω : représentons une suite infinie de jets d'une pièce par un élément de $\Omega = \{0,1\}^{\mathbb{N}}$ avec la probabilité produit (on admet l'existence d'une telle construction) telle que chaque coordonnée vaille 1 avec probabilité p; si $X_n(\omega)$ est la n-ième coordonnée, les X_n sont des v.a.i.i.d. valant 1 avec probabilité p et 0 sinon, on a bien que $\frac{1}{n}(X_1(\omega) + ... + X_n(\omega))$ converge vers p avec probabilité 1 ; la trajectoire $\omega = (1,1,1,...)$, n'en existe pas moins ; il y a même des ω pour lesquel il n'y a pas convergence.

De manière générale toutes les égalités ou inégalités ponctuelles (i.e. pour chaque ω) apparaissant dans les calculs ne seront valides que presque sûrement. En vertu de ce qui vient d'être dit, cette démarche reste cohérente tant que l'on ne manipule qu'un nombre au plus dénombrable d'inégalités, ce qui sera toujours le cas.

Cas des variables symboliques. Une variable aléatoire peut prendre ses valeurs dans un alphabet fini donné $\{a_1, ...a_n\}$, par exemple un nom de département ou une marque de voiture. D'un point de vue théorique, tout se passe comme si elle prenait ses valeurs dans $\{1, ...n\}$, la condition de variable aléatoire étant que $\{\omega : X(\omega) = a_i\}$ soit mesurable pour tout i.

Bilan. Les résultats présentés ici montrent que les manipulations habituelles conservent la mesurabilité des variables. Cette question est donc finalement, en première approche, tout à fait secondaire.

Quand au caractère «presque sûr» des propriétés, on peut considérer cela, d'un point de vue pratique ou applicatif, comme une subtilité de mathématicien qui ne modifie pas l'interprétation.

Bien entendu, dans une approche plus avancée (comme en théorie des processus) ces questions peuvent soulever de réels problèmes.

I.4 Espérance

Une v.a. X est intégrable (plus précisément P-intégrable) si

$$\int_\Omega |X(\omega)| P(d\omega) < +\infty$$

et son espérance n'est autre que son intégrale pour la mesure P

$$E[X] = \int_\Omega X(\omega) P(d\omega).$$

Si elle n'est pas intégrable mais positive, on convient que $E[X] = +\infty$. Tous les théorème vus en rappel d'intégration sont bien entendu valides. On vérifie par le même calcul que $E[Y] = +\infty$.

Ajoutons un résultat qui est classique et très utilisé :

LEMME I.15 *(Borel-Cantelli. Sens facile) Soit A_n une suite d'ensembles $\mathscr{F}$-mesurables tels que $\sum_n P(A_n) < +\infty$ alors*

$$P(\{\omega : \omega \in A_n \ i.o.\}) = 0. \tag{I.7}$$

La notation «i.o.» signifie «infiniment souvent» (infinitely often) et l'ensemble $\{\omega : \omega \in A_n \ i.o.\}$ est l'ensembles des ω pour lesquel $\omega \in A_n$ pour une infinité de n. (I.7) signifie donc qu'avec probabilité 1, il n'existe qu'un nombre fini de valeurs de n pour lesquelles $\omega \in A_n$.

Démonstration: On a

$$\{\omega : \omega \in A_n \ i.o.\} = \{\omega : \sum_n 1_{A_n}(\omega) = +\infty\}.$$

Mais

$$E\left[\sum_n 1_{A_n}(\omega)\right] = \sum_n E[1_{A_n}(\omega)] = \sum_n P(A_n) < +\infty.$$

Il s'ensuit que la variable $\sum_n 1_{A_n}$ est presque sûrement finie ce qui est bien (I.7). ∎

Les théorèmes de continuité d'espérance. Le théorème I.9 se réécrit

THÉORÈME I.16 *Soit X_n une suite de v.a., on a les trois résultats suivants :*

Théorème de convergence monotone. *Si pour tout $n \geq 1$, $0 \leq X_n \leq X_{n+1}$, alors*

$$E[\sup_n X_n] = \sup_n E[X_n].$$

Théorème de convergence dominée de Lebesgue. *Si pour tout ω $X_n(\omega) \to X(\omega)$ et s'il existe une fonction mesurable Y telle que $|X_n| \leq Y$ et $E[Y] < \infty$ alors*

$$\lim_n E[X_n] = E[X].$$

Moyenne et variance. Soit X une v.a. réelle. La moyenne m_x de la loi de X est son espérance. C'est un indice de localisation de ses valeurs ; la médiane en est un autre.

La variance est un indice de variabilité de X, elle vaut

$$Var(X) = \sigma_X^2 = E[X^2] - E[X]^2 = E[(X - E[X])^2].$$

σ_X est l'écart-type. Elle est nulle si X est constante et pour tout $a \in \mathbb{R}$, $Var(X + a) = Var(X)$. On considère qu'un intervalle de valeurs typiques de X est $[m_X - 2\sigma_X, m_X + 2\sigma_X]$ (le facteur 2 est assez arbitraire). La variance de la somme satisfait :

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

où la covariance vaut

$$Cov(X, Y) = E[XY] - E[X]E[Y] = E[(X - E[X])(Y - E[Y])].$$

Deux v.a. seront dites décorrélées si leur covariance est nulle. Si X est un vecteur $Var(X)$ est la matrice telle que

$$Var(X)_{ij} = Cov(X_i, X_j)$$

Si l'on remarque que pour un vecteur u vu comme une matrice colonne $n \times 1$, le produit uu^T est une matrice $n \times n$ de terme général $u_i u_j$, on vérifie facilement que

$$Var(X) = E[XX^T] - E[X]E[X]^T = E[(X - E[X])(X - E[X])^T].$$

Si X et Y sont deux vecteurs $Cov(X, Y)$ est la matrice de terme général $Cov(X_i, Y_j)$

$$Cov(X, Y) = E[XY^T] - E[X]E[Y^T] = E[(X - E[X])(Y - E[Y])^T].$$

On a

$$Cov(X, Y) = Cov(Y, X)^T$$

et la formule de variance de la somme devient

$$Var(X + Y) = Var(X) + Var(Y) + Cov(X, Y) + Cov(Y, X)$$

Donc si X et Y sont décorrélées, la variance de leur somme est la somme des variances. Ceci implique que si les v.a. $X_1, \ldots X_n$ sont décorrélées deux à deux alors

$$Var(X_1 + \cdots + X_n) = Var(X_1) + \cdots + Var(X_n).$$

THÉORÈME I.17 *(Inégalité de Markov) Soit Y une v.a. réelle ≥ 0, pour tout $\varepsilon > 0$*

$$P(Y \geq \varepsilon) \leq \frac{E[Y]}{\varepsilon}$$

Démonstration: Noter que pour $y \geq 0$ on a $1_{y \geq \varepsilon} \leq y/\varepsilon$, puis appliquer à Y et prendre l'espérance. ∎

THÉORÈME I.18 *(Inégalité de Chebyshev) Soit X une v.a. réelle intégrable, pour tout $\varepsilon > 0$*

$$P(|X - E[X]| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2}$$

Démonstration: Appliquer l'inégalité de Markov $Y = |X - E[X]|^2$. ∎

Calcul des moments à partir de la fonction caractéristique. Les moments de la v.a. réelle X sont les quantités $E[X^n]$, $n \geq 1$. Si l'on dérive informellement la fonction caractéristique de la loi de X

$$\varphi(t) = E[e^{itX}]$$

par rapport à t, on trouve :

$$\frac{d^n}{dt^n}\varphi(t) = i^n E[X^n e^{itX}]$$

et donc en $t = 0$

$$\varphi^{(n)}(0) = i^n E[X^n]$$

qui est, au facteur i^n près, le moment d'ordre n de X. Le théorème de dérivation sous le signe intégral assure que cela est légitime si $E[|X|^n] < +\infty$. Si X est à valeur dans $\mathbb{R}^d$, $d > 1$, on obtient

$$\frac{\partial^n \varphi(0)}{\partial t_1 ... \partial t_n} = i^n E[X_{t_1} ... X_{t_n}].$$

Plutôt que de dériver, il peut être plus simple d'écrire le développement de φ en puissances de t ; en effet en développant l'exponentielle de l'espérance on obtient

$$\varphi(t) = \sum_n t^n \frac{i^n}{n!} E[X^n] \tag{I.8}$$

(la permutation de $\int$ et $\sum$ peut se justifier avec le théorème de Fubini sous des conditions très générales). Pour la variable gaussienne centrée, c'est le plus facile car la dérivation peut paraître compliquée et en revanche on a tout de suite

$$\varphi(t) = \sum_n t^{2n} \frac{(i\sigma)^{2n}}{2^n n!} \tag{I.9}$$

et l'identification des développements (I.8) et (I.9) donne les moments $E[X^{2n}] = (2n)!/(2^n n!)$. Les moments impairs sont nuls car la variable est symétrique.

I.5 Digression sur Ω

On a vu que Ω est «l'espace des possibles». Par exemple si l'expérience consiste à jeter 4 fois un dé non pipé, Ω sera $\{1, 2, 3, 4, 5, 6\}^4$ et pour tout $\omega \in \Omega$

$$P(\{\omega\}) = \frac{1}{6^4}$$

Les applications coordonnées, $X_i : \omega = (\omega_1, \omega_2, \omega_3, \omega_4) \mapsto X_i(\omega) = \omega_i$ sont des variables aléatoires ; elles permettent de construire toutes les autres.

On pourrait aussi considérer que «l'espace des possibles» est bien plus gros et qu'il faudrait prendre en compte la vitesse de lancement, les frottements sur le tapis au cours du lancer, la force du vent, la volonté divine.... Tous ces phénomènes restant aléatoires. À la limite les 4 tirages peuvent être considérés comme fonctions de toutes ces incertitudes. L'espace devient énorme mais ce qui nous intéresse c'est les 4 résultats qui sont 4 variables aléatoires $X_1(\omega), ...X_4(\omega)$. On voit se dessiner un autre point de vue qui consiste a postuler l'existence d'un gros espace Ω contenant *tous* les possibles et de variables aléatoires X_1, X_2, X_3, X_4 décrivant les observations ; mais il n'est reste pas moins que, comme le souligne le premier point de vue, tout ce que l'on observe se restreint à ces quatre variables et la construction du gros espace peut paraître superflue.

Deuxième exemple : On veut modéliser la durée de vie d'une ampoule. L'option 1 consiste à dire que ω est la durée de vie et, par exemple, $P(\omega \leq t) = e^{-\lambda t}$ (loi exponentielle de paramètre λ). L'option 2 consiste à dire que Ω est l'espace de toutes les réalisations de conditions d'utilisation possibles, de défauts de fabrication, etc., et que la durée de vie est une v.a. $X(\omega)$ telle que $P(X(\omega) \leq t) = e^{-\lambda t}$.

Dans ces deux exemples on voit deux tendances : (1) Ω est l'espace de toutes les observations possibles pour l'expérience en cours et l'aléatoire est représenté par une mesure sur cet espace, (2) il existe un espace probabilisé Ω tel que les observations sont des variables aléatoires sur cet espace. Le premier point de vue correspond exactement à ce qui est fait mathématiquement, le deuxième est conceptuellement plus parlant, et peut être adopté sans difficulté puisqu'il généralise le premier.

Troisième exemple. Si l'expérience consite à jeter un dé une infinité de fois (pour vérifier la loi des grand nombres...), l'espace Ω et $\{1,2,3,4,5,6\}^{\mathbb{N}}$. C'est un espace assez compliqué à manipuler (fabriquer une tribu, une mesure...). Il est plus simple de postuler l'existence d'un espace $(\Omega, \mathscr{F}, P)$ abstrait et d'une suite de variables aléatoires indépendantes $X_n, n \geq 1$, c-à-d telles que

$$P(X_1 = n_1, X_2 = n_2, ..., X_k = n_k) = \frac{1}{6^k} \tag{I.10}$$

pour toute suite finie $n_1, \dots n_k$.

La vérité mathématique essentielle n'en reste pas moins que l'espace est celui des trajectoires $\{1, \dots 6\}^{\mathbb{N}}$, et que tout le modèle de l'expérience est contenu dans le mesure que l'on met sur cet espace.

De plus l'existence d'un espace avec des v.a. satisfaisant certaines conditions n'est pas toujours garantie : dans le cas de (I.10) la construction de $(\Omega, \mathscr{F}, P)$ n'est pas immédiate (dans le cas des chaînes de Markov c'est encore pire), et donc le postuler peut paraître abusif ; il y a donc en toute rigueur un travail de construction probabiliste à faire.

I.6 Loi d'une variable aléatoire

Définition I.19 *Soit X une v.a. sur $(\Omega, \mathscr{F}, P)$ à valeurs dans $\mathbb{R}^d$. La loi de X est la mesure de probabilité P_X sur $\mathbb{R}^d$ définie par*

$$P_X(B) = P(X(\omega) \in B). \tag{I.11}$$

C'est encore la probabilité sur $\mathbb{R}^d$, image de P par X.

Une variable aléatoire X qui ne prend que les valeurs 0 ou 1 est un indicateur d'ensemble : $X = 1_A$ avec $A = \{\omega : X(\omega) = 1\}$.

Dans la suite on ne considérera que des v.a. discrète ou ayant une densité par rapport à la mesure de Lebesgue, et éventuellelement des variables combinant ces deux cas (exemple 3 du § I.6) ; la théorie des probabilité permet d'en construire d'autres sortes encore mais qu'on peut considérer ici comme pathologiques.

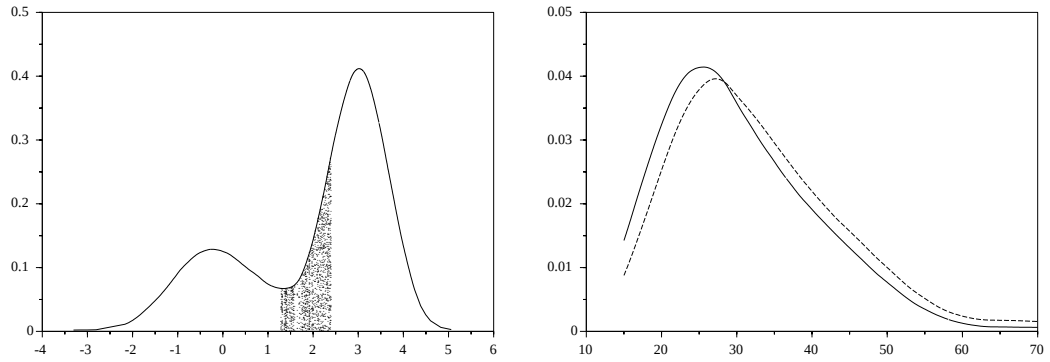


Figure I.1 – Un exemple de densité : la probabilité de l'intervalle $[1,3\ ;\ 2,4]$ est la surface de la zone noircie ; la surface totale (intégrale) fait 1. A droite : Densités des variables « âge du (de la) marié(e) le jour du mariage » en Alaska en 1995 (estimation à partir des données de l'Alaska Bureau of Vital Statistics). Les femmes sont en trait plein et les hommes en pointillés.

Si la variable est discrète et prend les valeurs sur $\{v_1, ..v_J\}$, sa répartition est détérminée par l'ensemble des $p_j = P(X = v_j)$. On a $\sum_j p_j = 1$.

Si elle est scalaire et admet une densité $p(x)$, la probabilité qu'elle tombe dans $[a, b]$ vaut l'intégrale de cette densité entre les deux points a et b, cf figure I.1.

PROPOSITION I.20 *Si X a pour loi P_X, alors pour toute fonction borélienne P_X-intégrable $f : \mathbb{R}^d \to \mathbb{R}$*

$$E[f(X)] = \int f(x) P_X(dx). \tag{I.12}$$

Sans donner de démonstration précise, notons simplement que l'équation (I.12) est vraie si f est un indicateur d'ensemble (car dans ce cas c'est simplement (I.11)) ; le problème est donc d'étendre (I.12) aux fonctions étagées puis aux fonctions boréliennes bornées, et enfin aux fonctions P_X-intégrables, ce qui est une démarche classique en théorie de la mesure.

LOI MARGINALE. Soit (X, Y) un couple de v.a. réelles de densité $p(x, y)$. Pour toute fonction $x \mapsto f(x)$ continue positive on peut calculer $E[f(X)]$ en considérant f comme fonction de X et Y et l'on a

$$E[f(X)] = \int \left(\int f(x) p(x, y) dy \right) dx = \int f(x) \left(\int p(x, y) dy \right) dx$$

ce qui signifie que la loi de X a pour densité $\int p(x, y) dy$.

Par exemple si (X, Y) un couple de v.a. réelles de densité $1_{y>0} 1_{0<x<1} x e^{-xy}$. La loi de X a pour densité sur $[0, 1]$:

$$\int_0^{+\infty} 1_{0<x<1} x e^{-xy} dy = 1$$

et donc X suit la loi uniforme sur $[0, 1]$.

CALCUL D'UNE ESPÉRANCE À L'AIDE DU THÉORÈME DE FUBINI. Soit (X, Y) le couple de v.a. réelles de l'exemple prédédent ; calculons $E[XY]$. Comme X et Y sont positives on applique directement le théorème de Tonelli pour intégrer dans l'ordre qui nous arrange

$$E[XY] = \int \left(\int 1_{y>0} 1_{0<x<1} x^2 y e^{-xy} dy \right) dx = \int \left(\int 1_{z>0} 1_{0<x<1} e^{-z} dz \right) dx = 1$$

Les distributions classiques. Pour les distribution sur les entiers on donne $p_k = P(X = k)$ et pour les distributions sur $\mathbb{R}$ on donne la densité. Une écriture abrégée $\mathcal{E}(\lambda) \sim \lambda^{-1} \mathcal{E}(1)$ signifie «$\mathcal{E}(\lambda)$ est la loi d'une v.a. $\mathcal{E}(1)$ divisée par λ».

Nom	Notation	p_k ou $p(x)$.	Commentaires
Bernoulli	$\mathcal{B}(1,p)$	$p_1 = 1 - p_0 = p$	p est l'espérance.
Binomiale	$\mathcal{B}(n,p)$	$\binom{n}{k} p^k (1-p)^{n-k}$	Loi de la somme de n v.a. $\mathcal{B}(1,p)$ indépendantes. $0 \leq k \leq n$.
Binomiale négative	$\mathcal{B}_-(n,p)$	$\binom{n+k-1}{n-1} p^n (1-p)^k$	Loi du nombre d'échecs avant le n-ième succès dans un schéma de Bernoulli (jeu de pile ou face). $k \geq 0$.
Géométrique	$\mathcal{G}(p)$	$p(1-p)^k$	Instant (moins 1) du premier succès dans un schéma de Bernoulli. $k \geq 0$.
Poisson	$\mathcal{P}(\lambda)$	$\frac{\lambda^k}{k!} e^{-\lambda}$	Ex : Loi du nombre de panne d'une machine sur une certaine durée.
Uniforme	$\mathcal{U}([a,b])$	$\frac{1}{b-a} 1_{a \leq x \leq b}$	
Exponentielle	$\mathcal{E}(\lambda)$	$1_{x \geq 0} \lambda e^{-\lambda x}$	Attention : L'espérance vaut $\frac{1}{\lambda}$. Ex. : Durée de vie d'une ampoule. $\mathcal{E}(\lambda) \sim \lambda^{-1} \mathcal{E}(1)$.
Laplace	$\mathcal{L}(\lambda)$	$\lambda e^{-\lambda \lvert x \rvert}/2$	$\mathcal{L}(\lambda) \sim \lambda^{-1} \mathcal{L}(1)$.
Cauchy	$\mathcal{C}(\lambda)$	$\frac{1}{\pi} \frac{\lambda}{x^2+\lambda^2}$	Espérance infinie. $\mathcal{C}(\lambda) \sim \lambda \mathcal{C}(1)$.
Normale (ou gaussienne)	$\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)}$	Est dite centrée si $\mu = 0$ et réduite si $\sigma = 1$. Une somme de v.a. gaussiennes indépendantes est encore gaussienne. $\mathcal{N}(\mu, \sigma^2) \sim \mu + \sigma \mathcal{N}(0,1)$.
Gamma	$\Gamma(\beta, p)$	$1_{x \geq 0} \left(\frac{x}{\beta}\right)^p \frac{e^{-\frac{x}{\beta}}}{x\Gamma(p)}$	Généralise les exponentielles. Une somme de v.a. indépendantes Gamma de même β suit encore une loi Gamma. $\Gamma(\beta, p) \sim \beta\Gamma(1,p)$.
Normale	$\mathcal{N}(\mu, R)$	$\frac{e^{-(x-\mu)^T R^{-1} (x-\mu)/2}}{(2\pi)^{d/2}\sqrt{\det(R)}}$	Sur $\mathbb{R}^d$.

Comment calculer la loi d'une v.a. Y fonction d'autres v.a. de loi connue ? L'idéal est de trouver la probabilité de chaque atome dans le cas discret et la densité dans le cas continu (ou sa fonction de répartition [4]). Parfois ce n'est pas possible. Donnons deux exemples où tout se passe bien :

Exemple 1 : Y est discrète. On décrit l'ensemble $\{y_1, y_2, ...\}$ de ses valeurs possibles et l'on calcule explicitement $P(Y = y_n)$.

Par exemple si X suit une loi exponentielle de paramètre λ (densité $1_{x>0}\lambda e^{-\lambda x}$), et $Y = [X]$ on a pour tout $n \geq 0$

$$P(Y = n) = P([X] = n) = P(n \leq X < n+1) = \int_n^{n+1} \lambda e^{-\lambda x} dx = e^{-\lambda n} - e^{-\lambda(n+1)}.$$

Y suit une loi géométrique de paramètre $p = 1 - e^{-\lambda}$ (rappel : $P(Y = n) = p(1-p)^n$). ∎

Exemple 2 : Y est continue. Si Y est à valeurs réelles le plus simple est généralement de se donner une fonction f abstraite continue positive bornée et de tenter de calculer $E[f(X)]$ avec des méthodes de changement de variables ce qui fera apparaître spontannément la densité. Par exemple si X est exponentielle de paramètre λ et $Y = 1 + \sqrt{X}$ il vient

$$E[f(Y)] = E[f(1+\sqrt{X})] = \int_0^{+\infty} f(1+\sqrt{x}) \lambda e^{-\lambda x} dx = \int_1^{+\infty} f(y) \lambda e^{-\lambda(y-1)^2} 2(y-1) dy.$$

La loi de Y est donc la loi de densité $1_{y \geq 1} \lambda e^{-\lambda(y-1)^2}(y-1)$ par rapport à la mesure de Lebesgue. ∎

Voici maintenant un cas un peu plus compliqué :

4. Il arrive que le calcul de la fonction de répartition soit significativement plus simple, mais c'est rare ; c'est le cas dans l'exemple classique où Y est un maximum de variables indépendantes.

EXEMPLE 3 : Y EST MIXTE. C'est rare. Supposons que l'on mesure la durée de vie d'une ampoule en la laissant allumée pendant 500 heures. Y correspond au moment où elle a claqué et vaut 500 si elle n'a pas claquée. C'est un exemple de donnée censurée à droite : $Y = \min(X, 500)$ où X est la durée de vie de l'ampoule. Si X est exponentielle de paramètre λ, on a

$$\begin{aligned} E[f(Y)] =& E[f(\min(X,500))] \\ =& \int_0^{+\infty} f(\min(x,500))\lambda e^{-\lambda x}dx \\ =& \int_0^{500} f(x)\lambda e^{-\lambda x}dx + f(500)\int_{500}^{\infty}\lambda e^{-\lambda x}dx \\ =& \int_0^{500} f(x)\lambda e^{-\lambda x}dx + f(500)e^{-500\lambda} \end{aligned} \tag{I.13}$$

ce qui exprime la loi de Y ; formellement la loi de Y est la somme de deux mesures

$$1_{y\leq 500}\lambda e^{-\lambda y}dy + e^{-500\lambda}\delta_{500}(dy).$$

Cette somme s'exprime comme la moyenne de deux probabilités

$$(1-e^{-500\lambda})\left[1_{y\leq 500}\frac{\lambda}{1-e^{-500\lambda}}e^{-\lambda y}dy\right] + e^{-500\lambda}\left[\delta_{500}(dy)\right].$$

Ce qui peut s'interpréter de la façon suivante : avec probabilité $p = e^{-500\lambda}$ l'ampoule n'a pas claqué, et avec probabilité $1-p$ l'ampoule claque (avant le temps 500) et alors la durée de vie observée suit la loi $1_{y\leq 500}\frac{\lambda}{1-p}e^{-\lambda y}dy$.

C'est un mélange de lois.

Mélange de lois. Si l'on suppose que la distribution du poids des boeufs suit une loi $\mathcal{N}(m,\sigma^2)$ et celle des taureaux est $\mathcal{N}(m',\sigma'^2)$ la distribution du poids des bovidés adultes mâles suit une loi de densité

$$q(x) = p\frac{e^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} + (1-p)\frac{e^{-(x-m')^2/(2\sigma'^2)}}{\sqrt{2\pi\sigma'^2}}$$

où p est la proportion de boeufs dans la population. Voir par exemple le premier graphique de la figure I.1 p.16. En effet si X est la variable de poids d'un individu tiré au hasard, on aura pour toute fonction bornée f, puisqu'on a une probabilité p de tomber sur un boeuf

$$\begin{aligned} E[f(X)] =& p\int f(x)\frac{e^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}dx + (1-p)\int f(x)\frac{e^{-(x-m')^2/(2\sigma'^2)}}{\sqrt{2\pi\sigma'^2}}dx \\ =& \int f(x)q(x)dx \end{aligned}$$

(ceci sera montré plus rigoureusement plus tard avec la forumle de Bayes). On dit que la loi de densité q est le mélange de $\mathcal{N}(m,\sigma^2)$ et $\mathcal{N}(m',\sigma'^2)$ avec les poids p et $1-p$. Si ces deux lois sont bien séparées, une pesée de l'animal permet de savoir à quel groupe il appartient avec une faible probabilité d'erreur.

I.7 Espaces de variables aléatoires

Pour tout $1 \leq p < \infty$ note $L_p(P)$ l'espace des v.a. X telles que

$$E[|X|^p] < \infty.$$

Par exemple, si Y suit la loi $\mathcal{E}(\lambda)$ et $X(\omega) = e^{Y(\omega)}$, on a

$$E[|X|^p] = \int e^{py}e^{-\lambda y}\lambda dy$$

qui n'est fini que si $p < \lambda$; cette condition est donc nécessaire et suffisante pour que $X \in L_p(P)$.

C'est un espace vectoriel, et on le muni de la norme

$$\|X\|_p = E[|X|^p]^{1/p}$$

(il se trouve que c'est une norme, ce qui est presque évident pour $p = 1$). Lorsque p tend vers l'infini, $\|X\|_p$ converge vers le sup-essentiel de $|X|$, c-à-d la plus grande valeur qui puisse être dépassée par $|X|$ avec probabilité non nulle

$$\|X\|_\infty = \sup\{x : P(|X| > x) > 0\}$$

(cette définition tient compte de ce que X n'est définie que presque sûrement : si X est la v.a. de $\Omega = [0, 1]$ muni de la mesure de Lebesgue telle que $X(\omega) = \omega$ sauf $X(1) = 10$, alors $\|X\|_\infty = 1$, car cette v.a. est indistinguable de celle qui vaut ω partout). L'inégalité de Jensen implique que pour tout $p \geq 1$

$$\|X\|_1 \leq \|X\|_p.$$

L'espace L_1 est celui pour lequel l'espérance est définie et est finie. Il est très rare de considérer des v.a. qui n'en font pas partie, cependant une v.a. suivant une loi de Cauchy n'est pas intégrable. L'inégalité de Hölder met en jeu deux espaces L_p et L_q pour deux exposants conjugués p et q :

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q, \qquad \frac{1}{p} + \frac{1}{q} = 1.$$

En particulier si $X, Y \in L_2$ alors $XY \in L_1$ et

$$|E[XY]| \leq E[X^2]^{1/2} E[Y^2]^{1/2}$$

qui est une inégalité de Cauchy-Schwarz. Les espaces importants sont L_1, L_2 et L_∞.

II

Indépendance

II.1 Définitions. Propriétés de base

Définition II.1 *(Tribu engendrée par une v.a.) Soit f une fonction définie sur un espace E à valeurs dans $\mathbb{R}^d$. La tribu $\sigma(f)$ engendrée par f est la famille des ensembles de la forme $\{x : f(x) \in B\}$ où B varie dans l'ensemble de tous les boréliens.*

En particulier, si X est une v.a. définie sur un espace mesurable $(\Omega, \mathscr{F})$, la tribu $\sigma(X)$ est une sous tribu de $\mathscr{F}$.

Un ensemble de $\sigma(X)$ est donc un ensemble qui peut être décrit à l'aide de X seulement.

Exemples : 1/ Si $X = 1_A$ est un indicateur, alors $\sigma(X) = \{\emptyset, A, A^c, \Omega\}$.

2/ Si $\Omega = \mathbb{R}^2$ et $X = \|\omega\|$ un ensemble de $\sigma(X)$ sera un ensemble qui ne dépend que la distance à 0, c'est donc exactement un ensemble mesurable invariant par toute rotation de centre 0.

3/ Si $\Omega = [0,1]^2$ muni de la tribu produit et $X(\omega) = \omega_1$, $\sigma(X)$ est la tribu des ensembles de la forme $A_1 \times [0,1]$, où A_1 est mesurable.

Théorème II.2 *Une variable aléatoire Y est mesurable par rapport la tribu $\sigma(X)$ si et seulement si elle peut s'écrire sous la forme $Y = \varphi(X)$ pour une certaine fonction borélienne φ.*

> La démonstration n'est pas très difficile ; elle suit une démarche typique de théorie de la mesure : il faut noter que c'est vrai si Y est un indicateur d'ensemble. Cela reste vrai si Y est une fonctions étagée, car les ensembles indicateurs qui la compose se réécrivent $\{\omega : Y(\omega) \in [c_i, d_i]\}$ et sont donc des ensembles de $\sigma(X)$. Il ne reste plus qu'à écrire Y comme limite de fonctions étagées (en arrondissant Y, cf page 8).

Dans la suite du cours on parlera souvent de tribus. Il faut avoir présent à l'esprit que l'on peut toujours considérer qu'une tribu est de la forme $\sigma(X)$ pour une certaine v.a. X (c'est un résultat mathématique). En pratique on aura différentes tribus engendrées par différentes v.a., $\mathscr{A} = \sigma(1_{X>0})$, $\mathscr{B} = \sigma(X)$, $\mathscr{C} = \sigma(X, Y)$; ici $\mathscr{A} \subset \mathscr{B}$ ce qui peut s'interpréter par le fait qu'il y a plus d'information dans $\mathscr{B}$ que dans $\mathscr{A}$ (connaître X c'est mieux que de connaître seulement son signe).

Définition II.3 *Soit $(\Omega, \mathscr{F}, P)$ un espace probabilisé et $\mathscr{F}_1, \dots \mathscr{F}_n$ des sous-tribus de $\mathscr{F}$; ces tribus sont dites indépendantes si pour tous ensembles $A_1 \in \mathscr{F}_1, \dots A_n \in \mathscr{F}_n$ on a*

$$P(A_1 \cap \dots \cap A_n) = P(A_1) \dots P(A_n).$$

Des v.a. $X_1, \dots X_n$ sont dites indépendantes si les tribus $\sigma(X_1), \dots \sigma(X_n)$ sont indépendantes, c-à-d si pour tous boréliens de $\mathbb{R}^d$ $B_1, \dots B_n$

$$P(X_1 \in B_1, \dots X_n \in B_n) = P(X_1 \in B_1) \dots P(X_n \in B_n). \tag{II.1}$$

En particulier, des variables de la forme $Y_k = f_k(X_k)$ sont indépendantes, puisque les tribus en jeu sont plus petites.

La définition habituelle de l'indépendance de deux événements

$$P(A \cap B) = P(A)P(B)$$

coïncide avec l'indépendance de $\sigma(1_A) = \{\emptyset, \Omega, A, A^c\}$ et $\sigma(1_B)$ (à vérifier !).

Noter que l'équation (II.1) exprime que les v.a. sont indépendantes si et seulement si leur mesure image est une mesure produit.

THÉORÈME II.4 *Si la loi des v.a. $X_1, \dots X_n$ a une densité p par rapport à la mesure de Lebesgue, alors ces v.a. sont indépendantes si et seulement si cette densité a la forme produit*

$$p(x_1, \dots x_n) = p_1(x_1) \dots p_n(x_n).$$

DÉFINITION II.5 *Des v.a. $X_1, \dots X_n \dots$ sont dites indépendantes si pour tout $n \geq 1$ les v.a. $X_1, \dots X_n$ sont indépendantes.*

THÉORÈME II.6 *Si les v.a. $X_1, X_2 \dots$ sont indépendantes et si $J_1, \dots J_n$ sont des parties finies disjointes de $\mathbb{N}$, alors des variables de la forme $Y_k = f_k(X_i, i \in J_k)$ sont indépendantes.*

THÉORÈME II.7 *Si les v.a. $X_1, \dots X_n$ sont indépendantes et si $E[|X_i|] < \infty$ pour tout i, alors*

$$E[|X_1 \dots X_n|] < \infty \quad \text{et} \quad E[X_1 \dots X_n] = E[X_1] \dots E[X_n].$$

Pour la démonstration, noter que c'est vrai pour les indicateurs, et cela s'étend facilement aux fonction étagées, il ne reste plus qu'à approcher les X_i par des fonctions étagées et à passer à la limite en utilisant le théorème de convergence dominée de Lebesgue.

En particulier, si $f_1 \dots f_n$ sont des fonctions mesurables et si les espérances existent (p.ex. les f_i sont bornées) :

$$E[f_1(X_1) \dots f_n(X_n)] = E[f_1(X_1)] \dots E[f_n(X_n)].$$

Réciproquement, si cette relation est vérifiée pour toutes fonctions f_k boréliennes bornées positives, alors les X_k sont indépendantes (tout simplement parce que (II.1) est satisfait).

II.2 Exemples

Indépendance deux à deux. Attention, ce n'est pas courant, mais des variables peuvent être deux à deux indépendantes sans être indépendantes : Considérons un tirage Y selon la loi uniforme sur $\{1, 2, 3, 4\}$. Soit les v.a. $X_1 = 1_{Y \text{pair}}$, $X_2 = 1_{Y \leq 2}$ et $X_3 = 1_{2 \leq Y \leq 3}$. Il est facile à vérifier, et naturellement intuitif, que chaque paire (X_i, X_j), $1 \leq i < j \leq 3$, est une paire de v.a. indépendantes (c-à-d que les ensembles correspondants sont indépendants), mais que (X_1, X_2, X_3) n'est pas un triplet de v.a. indépendantes (car chacune est fonction des deux autres. Calculer $E[X_1 X_2 X_3]$).

Variables décorrélées. Soient X et Y deux v.a. de $L_2(P)$, X et Y sont dites décorrélées si

$$E[XY] = E[X]E[Y].$$

Dans le cas vectoriel :

$$E[XY^T] = E[X]E[Y]^T.$$

Si elles sont indépendantes elles sont donc décorrélées, mais l'inverse est généralement faux : Soit X une v.a. gaussienne, $Y = |X|$. Alors les v.a. X et Y sont décorrélées mais pas indépendantes (exercice).

La variance d'une somme de v.a. indépendantes est donc la somme des variances (cf page 14).

Fonction caractéristique d'une somme de v.a. indépendantes. Si $X_1, \dots X_n$ sont des v.a. indépendantes réelles, on peut écrire

$$E[e^{it(X_1 + \dots X_n)}] = E[e^{itX_1}] \dots E[e^{itX_n}].$$

La fonction caractéristique de la somme est donc le produit des fonctions caractéristiques. Mentionnons que pour les densités par rapport à la mesure de Lebesgue (si elles existent) c'est bien plus compliqué : la densité de la somme est le produit de convolution des densités.

La loi multinomiale $\mathcal{M}(n, p_1, \dots p_q)$. Soit $X_1, \dots X_n$ des v.a.i.i.d. à valeurs dans un ensemble de q modalités, disons $\{1, \dots q\}$; on posera

$$p_j = P(X_1 = j)$$

si bien que $\sum p_j = 1$. Soit les variables

$$N_j = \sum_i 1_{X_i = j} = \#\{i : X_i = j\}.$$

Par exemple : il y a $q = 3$ candidats à une élection, avec une proportion p_1, p_2, p_3 de votes favorables ; $p_1 + p_2 + p_3 = 1$. On tire n personnes avec remise dans la population pour faire un sondage et l'on mesure le nombre N_k, $1 \leq k \leq 3$, de personnes déclarant voter pour k. La distribution du vecteur (N_1, N_2, N_3) permet d'étudier la représentativité du sondage.

Le vecteur $(N_1, \dots N_q)$ est une v. a. à valeurs dans $\{1 \dots n\}^q$ avec

$$\sum_j N_j = n.$$

Calculons la loi de ce vecteur. On a

$$\{N_1 = n_1, \dots N_q = n_q\} = \cup \{X_1 = i_1, \dots X_n = i_n\}$$

où la réunion est prise sur toutes les suites d'indices $(i_1, \dots i_n)$ telles qu'exactement n_j d'entre eux vallent j. Notons cet ensemble I ; il vient

$$\begin{aligned} P(N_1 = n_1, \dots N_q = n_q) &= \sum_I P(X_1 = i_1, \dots X_n = i_n) \\ &= \sum_I P(X_1 = i_1) \dots P(X_n = i_n) \\ &= \sum_I p_{i_1} \dots p_{i_n} \\ &= |I|\, p_1^{n_1} \dots p_q^{n_q} \end{aligned}$$

car chaque terme de la somme a la même valeur. Une formule classique de dénombrement pour le calcul du cardinal de I conduit alors à

$$P(N_1 = n_1, \dots N_q = n_q) = \frac{n_1! \dots n_p!}{n!}\, p_1^{n_1} \dots p_q^{n_q}.$$

Si $p = 2$, c'est la loi binomiale $\mathcal{B}(n, p_1)$.

Statistiques d'ordre. Soit $X_1, \dots X_n$ des v.a.i.i.d. réelles de loi de densité $p(x)$; l'hypothèse de densité implique que pour $i \neq j$

$$P(X_i = X_j) = \int\int 1_{x=y} p(x) p(y)\, dx\, dy = 0$$

Il s'ensuit qu'avec probabilité 1 ces v.a. sont toutes différentes (corollaire I.4). On note $X_{(1)}, \dots X_{(n)}$ les mêmes variables réarrangées par ordre croissant. Calculons la loi de ce vecteur : commençons par $n = 3$ pour simplifier :

$$E[f(X_{(1)}, X_{(2)}, X_{(3)})] = E[f(X_1, X_2, X_3) 1_{X_1 \leq X_2 \leq X_3}] + E[f(X_1, X_3, X_2) 1_{X_1 \leq X_3 \leq X_2}] + \dots$$

où les points marquent les 4 possibilités restantes pour l'ordre des variables. Tous ces termes sont égaux, par symétrie. Donc

$$E[f(X_{(1)}, X_{(2)}, X_{(3)})] = 6E[f(X_1, X_2, X_3) 1_{X_1 \leq X_2 \leq X_3}].$$

Avec n termes on a de même

$$E[f(X_{(1)}, \dots X_{(n)})] = n! E[f(X_1, \dots, X_n) 1_{X_1 \leq \dots \leq X_n}] = n! \int f(x_1, \dots, x_n) 1_{x_1 \leq \dots \leq x_n} dx_1 \dots dx_n.$$

La loi de $(X_{(1)}, \dots X_{(n)})$ a par conséquent la densité sur $\mathbb{R}^n$

$$n!\; 1_{x_1 \leq \dots \leq x_n} p(x_1) \dots p(x_n).$$

Indépendance et fonction caractéristique Soit $X = (X_1, \dots X_n)$ un vecteur aléatoire de fonction caractéristique

$$\varphi_X(t) = E[e^{it_1X_1 + \dots it_nX_n}].$$

Clairement si les X_k sont indépendantes, on a

$$\varphi_X(t) = \varphi_{X_1}(t_1) \dots \varphi_{X_n}(t_n)$$

où φ_{X_i} est la fonction caractéristique de X_i. Mais réciproquement si φ a une forme produit

$$\varphi_X(t) = \psi_1(t_1) \dots \psi_n(t_n).$$

alors en prenant pour t un vecteur dont toutes les composantes sont nulles sauf la i-ième, on voit que $\psi_i = \varphi_{X_i}$, ce qui signifie que X a même fonction caractéristique que le vecteur $X^* = (X_1^*, \dots X_n^*)$ où les X_i^* sont indépendantes avec chacune la loi de X_i (c-à-d que X^* suit la loi produit). Donc X et X^* ont même loi, ce qui prouve l'indépendance :

THÉORÈME II.8 *Si la fonction caractéristique d'un vecteur aléatoire $X = (X_1, \dots X_n)$ suit une forme produit, alors ses composantes sont indépendantes.*

III

Dépendance

III.1 Espérance conditionnelle

Un exemple. Une poule pond des œufs en nombre poissonnien X de paramètre λ. Chaque œuf donne naissance à un poussin avec probabilité p. Soit Y le nombre total de poussins. Il est clair qu'une fois que la ponte a eu lieu, l'espérance du nombre de poussins est Xp. Cette espérance n'est cependant pas une véritable espérance car c'est une variable aléatoire ; c'est ce qu'on appelle l'espérance de Y sachant X, notée $E[Y|X]$. L'espérance du nombre de poussins est l'espérance de cette variable c'est-à-dire λp qui est le nombre de poussins moyen que l'on escompte avant d'avoir vu le résultat de la ponte.

On peut se demander à l'inverse la valeur de l'espérance du nombre d'œuf sachant le nombre de poussins. En utilisant la formule de Bayes on trouve

$$P(X = n|Y = k) = \frac{P(X = n, Y = k)}{P(Y = k)} = \cdots = 1_{n\geq k} e^{-\lambda(1-p)} \frac{(\lambda(1-p))^{n-k}}{(n-k)!}$$

C'est une loi de poisson $\mathcal{P}(\lambda(1-p))$ décalée de k ce qui fait que

$$E[X|Y = k] = k + \lambda(1-p).$$

Le nombre d'œufs escompté est le nombre de poussins observé décalé d'une constante fixe, ce qui n'avait rien d'évident au départ.

Un autre exemple. Soit C_j la valeur d'un actif, ou de l'indice du CAC 40 au jour j. Il peut être intéressant de savoir prédire C_j au vu des valeurs passées $C_{j-1}, C_{j-2} \ldots$. La valeur C_{j-1} peut être un bon prédicteur, mais on peut arguer que si la suite augmente régulièrement, un prédicteur plus élevé, tenant compte de la pente moyenne sera meilleur ; on pourrait également trouver intéressant de prendre en compte C_{j-2} et C_{j-3}... Cela semble sans fin, et le problème est effectivement difficile.

Une méthode couramment utilisée en économétrie est de mettre un modèle statistique sur cette suite et de prendre comme estimée la variable aléatoire de la forme $\varphi(C_{n-1}, C_{n-2} \ldots)$ la plus proche (en une certain sens) de C_n. C'est $E[C_n|C_{n-1}, C_{n-2} \ldots]$, l'espérance de C_n sachant les valeurs antérieures.

Espérance conditionnelle à un événement. Soit A un ensemble mesurable non trivial et Y une v.a. intégrable, on définit l'espérance de Y sachant A par la v.a.

$$E[Y|A] = \begin{cases} \frac{E[Y 1_A]}{P(A)} & \text{si} \quad \omega \in A \\ \frac{E[Y 1_{A^c}]}{P(A^c)} & \text{si} \quad \omega \notin A. \end{cases}$$

C'est simplement la valeur moyenne que prend Y sur l'ensemble A ou sur A^c selon que $\omega \in A$ ou non. Donc quand on dit «sachant A» cela signifie «sachant si ω appartient à A ou non». Par exemple Y est la durée de vie et A est l'ensemble des fumeurs.

Si Y est un indicateur, on retrouve bien la formule de Bayes à cela près que «sachant A» s'y interprète par «sachant que ω appartient à A».

Soit X une autre v.a. prenant un nombre fini de valeur, il est naturel de définir $E[Y|X]$ par

$$E[Y|X] = \frac{E[Y\ 1_{\{X=c\}}]}{P(X=c)} \quad \text{si} \quad X(\omega) = c$$

ou en d'autres termes

$$E[Y|X] = \sum_k \frac{E[Y\ 1_{\{X=c_k\}}]}{P(X=c_k)} 1_{X(\omega)=c_k}.$$

où les c_k sont les valeurs possibles prises par X ; noter que c'est une fonction de X.

Si X est maintenant une v.a. réelle, il est naturel de définir $E[Y|X]$ par quelque chose de ressemblant à

$$E[Y|X] = E[Y \mid \{X \in I_k\}] \quad \text{si} \quad X(\omega) \in I_k. \tag{III.1}$$

où les intervalles I_k forment une partition fine de $\mathbb{R}$, p.ex. $I_k = [\frac{k}{n}, \frac{k+1}{n}[$; il se trouve que ceci est possible au sens où le membre de droite

$$Z_n = \sum_k E[Y \mid \{X \in I_k\}] 1_{X(\omega)\in I_k}$$

converge vers une limite lorsque que l'on raffine la partition[1], c-à-d $n \to \infty$; cette limite est $E[Y|X]$. Cette construction s'étend à X vectorielle. Bref :

$E[Y|X]$ est une variable aléatoire $E[Y|X](\omega)$ qui est une sorte de d'espérance de Y restreinte à l'ensemble des ω' tels que $X(\omega') = X(\omega)$.

Par exemple si U et V sont deux v.a. indépendantes $\mathcal{E}(1)$; soit f une fonction, $Y = f(U,V)$, et $X = \sqrt{U^2+V^2}$ (la norme du vecteur (U,V)) ; soit la limite de moyenne prises sur des couronnes très fines :

$$f(x) = \lim_{\varepsilon\to 0} \frac{E[Y 1_{x-\varepsilon\le\sqrt{U^2+V^2}\le x+\varepsilon}]}{P(x-\varepsilon \le \sqrt{U^2+V^2} \le x+\varepsilon)}$$

l'espérance de Y sachant X sera $f(X)$.

Une autre définition moins intuitive mais plus précise fait l'objet du paragraphe suivant.

Une définition. Le théorème suivant permet de donner une définition rigoureuse générale au concept d'espérance conditionnelle à certaines observations. Elle est d'apparence abstraite, mais il faut garder présent à l'esprit que l'on peut toujours considérer qu'une tribu $\mathscr{A}$ est de la forme $\sigma(X)$ pour une certaine v.a. (vectorielle) X et qu'une v.a. $\mathscr{A}$-mesurable est une v.a. de la forme $f(X)$; par exemple $\mathscr{A} = \sigma(C_1, \dots C_{n-1})$ dans l'exemple précédent. Ce sont les deux théorèmes suivants qui donneront des procédés de calcul effectifs.

Théorème et définition III.1 *Soit $(\Omega, \mathscr{F}, P)$ un espace probabilisé et $\mathscr{A}$ une sous tribu de $\mathscr{F}$. Alors pour toute v.a. $Y \in L_1(P)$ il existe une unique v.a. $\mathscr{A}$-mesurable $Z \in L_1(P)$, telle que pour toute autre v.a. $\mathscr{A}$-mesurable bornée U on ait*

$$E[YU] = E[ZU].$$

C'est l'espérance conditionnelle de Y sachant $\mathscr{A}$, que l'on note $Z = E[Y|\mathscr{A}]$. En particulier si Y est $\mathscr{A}$-mesurable alors $E[Y|\mathscr{A}] = Y$.

Si $\mathscr{A}$ est la tribu engendrée par des v.a. $X_1, \dots X_q$, on note aussi $E[Y|X_1, \dots X_q]$; dans ce cas Z est de la forme $Z = f(X_1, \dots X_q)$ pour une certaine fonction f.

1. La théorie des martingales permet d'y arriver facilement, du moins pour la sous-suite Z_{2^i}.

L'interprétation est la suivante : si pour fixer les idées $\mathscr{A} = \sigma(X_1, \dots X_q)$, alors $E[Y|\mathscr{A}]$ est la meilleure estimation que l'on puisse faire en utilisant les v.a. $(X_1, \dots X_q)$. Si $\mathscr{A}$ est la tribu triviale $\{\emptyset, \Omega\}$, on observe rien du tout, les seules variables $\mathscr{A}$-mesurables sont les constantes, et nécessairement $E[Y|\mathscr{A}] = E[Y]$. Si à l'opposé $\mathscr{A} = \mathscr{F}$, Y elle-même est $\mathscr{A}$-mesurable et l'on a $E[Y|\mathscr{A}] = Y$.

L'espérance conditionnelle peut être vue comme une projection sur l'espace des v.a. $\mathscr{A}$-mesurables (noter le caractère involutif). Plus précisément, dans l'espace $L_2(P)$, l'opérateur $Y \mapsto E[Y|\mathscr{A}]$ est en fait une projection orthogonale, $E[Y|\mathscr{A}]$ étant la v.a. $\mathscr{A}$-mesurable la plus proche de Y. En effet, supposons que $Y \in L_2(P)$ et calculons la distance de Y à toute autre v.a. $\mathscr{A}$-mesurable $Z' \in L_2(P)$; on vérifie facilement en exploitant la relation $E[YZ'] = E[ZZ']$ que

$$E[(Y - Z')^2] = E[(Y - Z)^2] + E[(Z - Z')^2].$$

La distance de Y à Z' est bien toujours supérieure à la distance de Y à Z. Notons également le théorème de Pythagore obtenu avec $Z' = 0$:

$$E[Y^2] = E[(Y - Z)^2] + E[Z^2].$$

Dans le théorème qui suit, l'inclusion de tribus $\mathscr{A} \subset \mathscr{B}$ réfère typiquement à la situation $\mathscr{A} = \sigma(X)$ et $\mathscr{B} = \sigma(X, X')$, la formule se lisant $E[E[Y|X, X']|X] = E[Y|X]$.

THÉORÈME III.2 *L'espérance conditionnelle satisfait les propriétés de linéarité*

$$E[\lambda Y|\mathscr{A}] = \lambda E[Y|\mathscr{A}]$$
$$E[Y + Z|\mathscr{A}] = E[Y|\mathscr{A}] + E[Z|\mathscr{A}]$$

Et si $\mathscr{A} \subset \mathscr{B}$

$$E[E[Y|\mathscr{B}]|\mathscr{A}] = E[Y|\mathscr{A}]. \tag{III.2}$$

Si Z *est* $\mathscr{A}$*-mesurable, on a*

$$E[ZY|\mathscr{A}] = ZE[Y|\mathscr{A}]$$

(sous le signe «$|\mathscr{A}$*», les v.a.* $\mathscr{A}$*-mesurables se comportent comme des constantes). Si* Y *est indépendante de* $\mathscr{A}$

$$E[Y|\mathscr{A}] = E[Y].$$

Toutes ces formules sont conformes à l'intuition. Le formule (III.2) est associé au fait que prour projeter orthogonalement sur un espace A, on peut commencer par projeter orthogonalement sur un espace B qui le contient : $P_A = P_A P_B$.

THÉORÈME III.3 *Soit* (X, Y) *une paire de v.a. On suppose que* Y *peut s'écrire sous la forme*

$$Y = f(X, U)$$

pour une certaine fonction f *et une certaine v.a.* U *indépendante de* X*. Alors*

$$E[Y|X] = \int f(X, u) P_U(du). \tag{III.3}$$

Si la loi de (X, Y) *a pour densité* $p(x, y)$ *par rapport à la mesure de Lebesgue, alors*

$$E[Y|X] = \int y\, p(y|X)\, dy, \qquad p(y|x) = \frac{p(x, y)}{\int p(x, y')dy'} = \frac{p(x, y)}{p_X(x)}. \tag{III.4}$$

La démonstration de ces deux théorèmes n'est pas très difficile en utilisant la définition : il suffit de vérifier par calcul que les candidats pour Z proposés dans (III.3) et (III.4) satisfont bien les conditions du théorème III.1.

Le deuxième point montre que dans le cas d'une paire (X, U) de v.a. *indépendantes* le calcul de l'espérance conditionnelle se fait en faisant perdre à X son caractère aléatoire (puisqu'il est connu) et en le considérant comme une constante.

Noter que si $f(x, u)$ est de la forme $g(x)h(u)$, alors (III.3) est immédiat ; ceci donne une autre méthode pour montrer le premier résultat par approximation de f par une somme de tels produits. Remarquons également que la formule (III.1) conduit également assez facilement à (III.4) par un calcul informel où l'on fait tendre la largeur de l'intervalle vers 0.

EXEMPLE : PROCESSUS AUTORÉGRESSIF. Soit U_n une suite i.i.d. centrée , $a \in \mathbb{R}$, et les v.a. définies par

$$X_{n+1} = a + bX_n + U_{n+1}, \quad n \geq 0 \tag{III.5}$$

$X_0 = x_0$ étant donné déterministe. Cette équation peut modéliser par exemple l'évolution d'actifs d'un jour au suivant. Posons $\mathscr{F}_n = \sigma(U_1, \dots U_n)$; il convient de remarquer que d'une part X_k est fonction de $(U_1, \dots U_k)$ et donc $\sigma(X_1, \dots X_n) \subset \mathscr{F}_n$ (en d'autres termes : «toute fonction de $(X_1, \dots X_n)$ peut s'exprimer comme en fonction de $(U_1, \dots U_n)$»), et d'autre part comme $U_k = X_k - a - bX_{k-1}$, on a $\mathscr{F}_n \subset \sigma(X_1, \dots X_n)$; finallement $\mathscr{F}_n = \sigma(X_1, \dots X_n)$. On a

$$E[X_{n+1}|\mathscr{F}_n] = a + bX_n + E[U_{n+1}|\mathscr{F}_n] = a + bX_n$$

qui est la prédiction à un jour. Pour obtenir la prédiction à deux jours, on conditionne l'expression précédente par rapport à $\mathscr{F}_{n-1}$:

$$\begin{aligned} E[X_{n+1}|\mathscr{F}_{n-1}] =& E[E[X_{n+1}|\mathscr{F}_n]|\mathscr{F}_{n-1}] \\ =& E[a + bX_n|\mathscr{F}_{n-1}] \\ =& a + bE[X_n|\mathscr{F}_{n-1}] \\ =& a + ab + b^2 X_{n-1}. \end{aligned}$$

On voit donc par récurrence que

$$E[X_n|\mathscr{F}_{n-k}] = a + ab + ... + ab^{k-1} + b^k X_{n-k}$$

qui est la prédiction à k jours. On peut également trouver de façon similaire des formules analogues pour $E[X_n^2|\mathscr{F}_{n-k}]$, faisant intervenir la variance commune aux U_k.

Terminons par un théorème assez intuitif :

THÉORÈME III.4 *Soit $Y \in L_1(P)$, X une autre v.a. à valeurs de $\mathbb{R}^d$ et φ une fonction injective sur $\mathbb{R}^d$:*

$$E[Y|\varphi(X)] = E[Y|X].$$

Si φ n'est pas injective, c'est bien entendu faux ; par exemeple si φ est la fonction nulle, on a $E[Y|\varphi(X)] = 0.$ pour tout x

III.2 Loi conditionnelle

L'espérance conditionnelle correspond à l'espérance selon une certaine loi ; cette loi dépend du point ω. La définition mathématique rigoureuse est un peu compliquée mais le principe est simple : l'application $A \mapsto E[1_A|X]$ est, pour chaque $\omega \in \Omega$, une mesure de probabilité, la loi conditionnelle.

Par exemple, sachant $N(\omega) = n$ le nombre de poussins suit une $\mathcal{B}(n, p)$; la probabilité d'avoir k poussins sachant N est donc $p^k(1-p)^{N(\omega)-k}\binom{N(\omega)}{k}$, cette probabilité est une variable aléatoire.

Pour obtenir l'espérance de $\varphi(Y)$ sous la loi conditionnelle, il suffit de remplacer Y par $\varphi(Y)$ dans les équations précédentes ; explicitons :

THÉORÈME III.5 *Soit (X, Y) une paire de v.a. et φ une fonction borélienne bornée.*

1/ Si X est discrète prenant les valeurs c_k :

$$E[\varphi(Y)|X] = \sum_k \frac{E[\varphi(Y)\ 1_{\{X=c_k\}}]}{P(X=c_k)} 1_{X(\omega)=c_k}.$$

2/ Si Y peut s'écrire sous la forme

$$Y = f(X,U)$$

pour une certaine fonction f et une certaine v.a. U indépendante de X, alors

$$E[\varphi(Y)|X] = \int \varphi(f(X,u)) P_U(du).$$

ce qui peut se dire : «la loi de $f(X,U)$ sachant $X=x$ est la loi de $f(x,U)$».

3/ Si la loi de (X,Y) a pour densité $p(x,y)$ par rapport à la mesure de Lebesgue, alors

$$E[\varphi(Y)|X] = \int \varphi(y)\, p(y|X)\, dy, \qquad p(y|x) = \frac{p(x,y)}{\int p(x,y')dy'} = \frac{p(x,y)}{p_X(x)}.$$

Dans le cas où Y ne prend un nombre fini (ou dénombrable) de valeurs, $y_1, y_2, \ldots$, la loi conditionnelle et déderminée par les v.a. $P(Y = y_k|X)$. Si X est elle même discrète, tout se réduit au calcul de $P(Y = y_k|X = c_j)$; pour chaque valeur de j on a un distribution de probabilité. C'est par exemple la méthode à suivre si l'on s'intéresse à la loi du nombre d'œufs sachant le nombre de poussins (exemple du début de chapitre).

Seul le troisième point donne une formule explicite pour la loi conditionnelle, c'est la loi dont la densité est

$$p(y|X) = \frac{p(X,y)}{\int p(X,y')dy'}.$$

Le dénominateur est désormais une «constante» puisque X est connu et seul y varie, si bien qu'en réalité, la loi de Y sachant X se *lit* directement sur $p(x,y)$.

Si l'on veut simuler par ordinateur la paire (X,Y) on peut simuler d'abord $X \sim p_X(x)$ puis $Y \sim p(y|x)$ où x est la valeur obtenue à la première étape.

Soit la paire (X,Y) de densité $\frac{e^{-y^2/2x}e^{-x}}{\sqrt{2\pi x}} 1_{x>0}$. On vérifie facilement que $X \sim \mathcal{E}(1)$ et que, comme $\frac{p(x,y)}{p_X(x)} = \frac{e^{-y^2/2x}}{\sqrt{2\pi x}}$, conditionnellement à $X=x$, Y suit une $\mathcal{N}(0,x)$.

Application concrète du théorème. On utilisera essentiellement les deuxième et troisième points. Le deuxième est, comme on va le voir, typiquement intéressant lorsqu'on considère des v.a. définies comme un système dynamique

$$X_n = f(X_{n-1}, U_n), \quad X_0 = x_0$$

où f est une fonction connue, les U_k sont i.i.d., et ici la condition initiale est déterministe x_0. Le troisième est utile dans des situations où la loi est donnée en bloc par sa densité.

Retour aux actifs. Revenons au processus autorégressif (III.5). Si $U_n \sim \mathcal{N}(0,\sigma^2)$, alors conditionnellement à $\mathscr{F}_n$ (c-à-d à $U_1, \ldots U_n$), X_{n+1} suit la loi $\mathcal{N}(a+bX_n, \sigma^2)$: il suffit d'appliquer le deuxième point du théorème en notant que X_{n+1} est de la forme $f(X_n, U_{n+1})$.

L'espérance conditionnelle de X_{n+1} sachant le passé est $a+bX_n$. La variance conditionnelle σ^2 quantifie l'incertitude autour de cette valeur : comme la gaussienne $\mathcal{N}(0,1)$ a une probabilité 0,01 d'être dans l'intervalle $[-2,6;2,6]$ la variable X_{n+1} qui est $a+bX_n+\sigma\mathcal{N}(0,1)$ se trouvera dans l'intervalle $[a+bX_n-2,6\sigma; a+bX_n+2,6\sigma]$ avec probabilité 0,01. σ^2 est la variance conditionnelle, elle permet de mesurer la qualité de la prédiction donnée par l'espérance conditionnelle. Il se trouve que dans cet exemple la variance conditionnelle ne dépend pas du passé mais en général elle peut être fonction de $(X_1, \ldots X_n)$, tout comme l'espérance conditionnelle.

VARIANCE CONDITIONNELLE. La variance conditionnelle de Y sachant X est

$$E[Y^2|X] - E[Y|X]^2 = E[(Y - E[Y|X])^2|X].$$

On peut le voir simplement comme une variance où le signe $E[.]$ est remplacé par $E[.|X]$. Elle représente l'incertitude de la prédiction de Y par $E[Y|X]$, après observation de X.

On pourrait croire que l'apport de l'information donnée par X va diminuer l'incertitude et que donc la variance conditionnelle sera toujours inférieure à la variance. C'est vrai en moyenne seulement, et donc inexact : il se peut que X apporte essentiellement comme information que l'incertitude actuelle est bien plus grande que celle que l'on observe en moyenne auquel cas la variance conditionnelle sera grande. Soit par exemple Z un actif et U une variable $\mathcal{B}(1,p)$ qui indique, si elle vaut 1, l'imminence d'une forte agitation des marchés. Un modèle (naïf!) pour la valeur Y de Z le lendemain est

$$Y = Z + (\sigma_1 + \sigma_2 U)V, \qquad V \sim \mathcal{N}(0,1)$$

où σ_1 et σ_2 sont deux paramètres positifs. La variable V est indépendante du reste. Le jour j, on ne connaît que $X = (Z,U)$. La loi conditionnelle de Y est donc $\mathcal{N}(Z, (\sigma_1 + \sigma_2 U)^2)$. Si U vaut 1, la variance conditionnelle de Y est $(\sigma_1 + \sigma_2)^2$. La variance de Y est en revanche

$$Var(Y) = Var(Z) + E[(\sigma_1 + \sigma_2 U)^2] = Var(Z) + (1-p)\sigma_1^2 + p(\sigma_1 + \sigma_2)^2.$$

On voit donc que si $Var(Z)$ est très petit, alors $(\sigma_1 + \sigma_2)^2$ dépasse la variance de Y.

Quelqu'un de moins bien informé ne connaît pas U. Pour cette personne, c'est la loi de Y sachant Z qui intervient. Nous laissons au lecteur le soin de vérifier qu'il s'agit d'un mélange de gaussiennes en utilisant le deuxième point du théorème.

IV

Variables gaussiennes

IV.1 Loi gaussienne sur $\mathbb{R}$.

IV.1.1 Définition

Les lois gaussiennes, ou normales, sur $\mathbb{R}$ sont les lois de densité

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

notées $\mathcal{N}(\mu, \sigma^2)$, ou les masses de Dirac δ_μ. Son espérance est μ et sa variance σ^2.

La loi est dité centrée si $\mu = 0$ et réduite si $\sigma = 1$.

IV.1.2 Homothéties et translations

Il est essentiel de remarquer que si $X \sim \mathcal{N}(0,1)$, alors la variable $Y = \mu + \sigma X$ suit la loi $\mathcal{N}(\mu, \sigma^2)$. En effet, pour toute fonction continue bornée f, on obtient par un changemement de variables élémentaire :

$$E[f(Y)] = E[f(\mu + \sigma X)] = \int f(\mu + \sigma x) e^{-\frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} = \int f(y) e^{-\frac{(y-\mu)^2}{2\sigma^2}} \frac{dy}{\sqrt{2\pi\sigma^2}}$$

On fabrique donc toute les loi gaussiennes réelles par ce procédé, les masses de Dirac étant obtenues en prenant $\sigma = 0$.

Noter également que l'on peut faire l'inverse si $\sigma > 0$: si $Y \sim \mathcal{N}(\mu, \sigma^2)$, et $X = (Y - \mu)/\sigma$, alors $X \sim \mathcal{N}(0,1)$.

IV.1.3 Fonction caractéristique

Un calcul un peu compliqué montre que la fonction caractéristique de $\mathcal{N}(0,1)$ est

$$\varphi_{0,1}(t) = e^{-t^2/2}.$$

On en déduit facilement la fonction caractéristique de $\mathcal{N}(\mu, \sigma^2)$; en effet, si $Y = \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$ avec $X \sim \mathcal{N}(0,1)$, on a

$$\varphi_{\mu,\sigma}(t) = E[e^{itY}] = E[e^{it(\mu+\sigma X}] e^{it\mu} E[e^{i(t\sigma)X}] = e^{it\mu} e^{-\sigma^2 t^2/2}.$$

Par conséquent, si $X_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$, $1 \leq j \leq n$, sont indépendantes, la fonction caractéristique de leur somme S est

$$\begin{aligned}
\varphi_S(t) = E[e^{it(X_1 + \dots X_n)}] &= E[e^{itX_1}] \dots E[e^{itX_n}] \\
&= e^{it\mu_1} e^{-\sigma_1^2 t^2/2} \dots e^{it\mu_n} e^{-\sigma_n^2 t^2/2} \\
&= e^{it\sum \mu_j} e^{-(\sum \sigma_j^2) t^2/2}.
\end{aligned}$$

qui est la fonction caractéristique de $\mathcal{N}(\sum \mu_j, \sum \sigma_j^2)$: *Une somme de gaussienne indépendantes est gaussienne.* Forcément les moyennes s'additionnent les moyennes, et les variances galement en raison de l'indépendance.

On a également pour tout t réel, et même $t \in \mathbb{C}$,

$$E[e^{tX}] = e^{t\mu} e^{\sigma^2 t^2/2}$$

formule qui redonne la fonction caractéristique en y remplaçant t par it.

IV.2 Rappels sur les variances

Notations. Soient x et y deux vecteurs de $\mathbb{R}^d$, si on les considère comme des matrices colonne, leur produit scalaire s'écrit

$$\langle x, y \rangle = \sum_i x_i y_i = x^T y.$$

Si A est une matrice, on a

$$x^T A y = \langle x, Ay \rangle = \sum_{ij} x_i A_{ij} y_j.$$

Soit $X = (X_1, ...X_d)^T$ une v.a. à valeurs dans $\mathbb{R}^d$ (on a mis le signe de transposition pour souligner qu'il s'agit d'un vecteur colonne). On a

$$E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$$

La matrice de covariance de X, parfois appelée variance de X, est par définition la matrice

$$Var(X) = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & \dots & Cov(X_1, X_d) \\ Cov(X_2, X_1) & Var(X_2) & \dots & Cov(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(X_d, X_1) & Cov(X_d, X_2) & \dots & Var(X_d) \end{pmatrix}$$

où

$$Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j] = E[(X_i - E[X_i])(X_j - E[X_j])].$$

Si l'on remarque que

$$XX^T = \begin{pmatrix} X_1^2 & X_1 X_2 & \dots & X_1 X_d \\ X_2 X_1 & X_2^2 & \dots & X_2 X_d \\ \vdots & \vdots & \ddots & \vdots \\ X_d X_1 & X_d X_2 & \dots & X_d^2 \end{pmatrix}$$

il apparaît que

$$Var(X) = E[XX^T] - E[X]E[X]^T = E[(X - E[X])(X - E[X])^T].$$

Soit S une matrice symétrique $d \times d$, la forme quadratique associée est l'application sur $\mathbb{R}^d$

$$u \mapsto \sum_{ij} u_i u_j S_{ij} = u^T S u.$$

La forme quadratique associée à $Var(X)$ est, en notant $Y = X - E[X]$

$$u^T Var(X) u = u^T E[YY^T] u = E[u^T YY^T u] = E[(u^T Y)(Y^T u)] = E[\langle u, Y\rangle^2]$$

qui est la variance de $\langle u, X\rangle$ car $E[\langle u, X\rangle] = \langle u, E[X]\rangle$. La matrice de covariance permet donc de calculer par simple produit la variance de toute combinaison linéaire des variables.

Si X et Y sont deux vecteurs, on définit la covariance entre X et Y par

$$Cov(X, Y) = E[XY^T] - E[X]E[Y]^T = E[(X - E[Y])(X - E[Y])^T].$$

C'est la matrice, a priori rectangulaire, des $Cov(X_i Y_j)$. Noter que

$$Cov(Y, X) = Cov(X, Y)^T.$$

Notons aussi que si X est décomposé en deux sous-vecteurs

$$X = \begin{pmatrix} Y \\ Z \end{pmatrix}$$

on a

$$Var(X) = \begin{pmatrix} Var(Y) & Cov(Y, Z) \\ Cov(Z, Y) & Var(Z) \end{pmatrix}.$$

IV.3 Vecteurs gaussiens

La définition formelle est la suivante

DÉFINITION IV.1 *Un vecteur aléatoire X sur $\mathbb{R}^d$ est dit gaussien si pour tout $u \in \mathbb{R}^d$ le produit scalaire $\langle u, X\rangle$ suit une loi gaussienne sur $\mathbb{R}$.*

Par conséquent, pour toute matrice M carrée ou rectangulaire à d colonnes, $Y = MX$ est encore un vecteur gaussien.

Si l'on connaît l'espérance et la variance de X, alors on connaît l'espérance et la variance de $\langle u, X\rangle$ pour tout u (qui valent $\langle u, \mu\rangle = u^T \mu$ et $u^T V u$), et donc la loi de $\langle u, X\rangle$, puisque la loi gaussienne est caractérisée par son espérance et sa variance. On peut donc calculer la fonction caractéristique de X :

$$\varphi_X(u) = E[e^{i\langle u, X\rangle}].$$

Ceci démontre que la loi d'une v.a. gaussienne est entièrement caractérisée par son espérance et sa variance et un calcul immédiat (si l'on connaît déjà la fonction caractéristique de la gaussienne réelle) donne

$$\varphi_X(u) = e^{iu^T \mu} e^{u^T V u/2}, \quad V = Var(X)$$

Comme la fonction caractéristique caractérise la loi, cette expression de φ_X permet de montrer le

THÉORÈME IV.2 *Soit X un vecteur gaussien d'espérance μ et de variance V supposée inversible, alors sa loi possède une densité par rapport à la mesure de Lebesgue, qui vaut*

$$p(x) = \frac{e^{-(x-\mu)^T V^{-1}(x-\mu)/2}}{(2\pi)^{d/2}\sqrt{\det(V)}}.$$

On note cette loi $\mathcal{N}(\mu, V)$.

Si V n'est pas inversible il n'y a pas de densité car la loi de X est portée par un sous-espace, c'est par exemple le cas des vecteurs gaussiens

$$\begin{pmatrix} X \\ X \end{pmatrix}, \quad \begin{pmatrix} X \\ Y \\ X + Y \end{pmatrix}$$

où X et Y sont deux gaussiennes indépendantes.

Un vecteur de v.a. gaussiennes n'est pas nécessairement un vecteur gaussien ; par exemple si $X \sim \mathcal{N}(0,1)$ et $Y \sim \mathcal{B}(1,1/2)$ est indépendant de X, il est simple de vérifier que $Z = (2Y-1)X$ est $\mathcal{N}(0,1)$ et que pourtant $X+Z$ ne peut pas être gaussien, car nul avec probabilité exactement $1/2$.

En revanche on vérifie facilement en utilisant la définition qu'un vecteur de v.a. gaussiennes indépendantes est un vecteur gaussien.

Expression à l'aide de gausienne centrées réduites. On sait qu'une gaussienne scalaire $Z \sim \mathcal{N}(\mu, \sigma^2)$ se réécrit comme translatée dilatée d'une certaine gausienne standard :

$$Z = \mu + \sigma U, \quad U \sim \mathcal{N}(0,1)$$

il suffit de prendre $U = (Z-\mu)/\sigma$. On peut montrer de même que tout vecteur gaussien Z peut se mettre sous la forme

$$Z = \mu + AU, \quad U \sim \mathcal{N}(0, Id) \tag{IV.1}$$

où A est une matrice carrée. Notons que U est simplement un vecteurs de gaussiennes indépendantes centrées réduites. Par exemple le deuxième vecteur de l'exemple ci-dessus s'écrit si $X \sim N(0,1)$ et $Y \sim N(0,4)$

$$\begin{pmatrix} X \\ Y \\ X+Y \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} X \\ Y/2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 2 & 0 \end{pmatrix} \begin{pmatrix} X \\ Y/2 \\ Z \end{pmatrix}$$

où Z est indépendante $\mathcal{N}(0,1)$; dans cet exemple, on voit que le fait que V soit non inversible fait que A peut être choisie rectangulaire plus petite.

Indépendance et matrice de covariance. Supposons V inversible et diagonale par bloc, par exemple deux blocs

$$V = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix}.$$

Ceci signifie exactement que les vecteurs Y et Z de la décomposition conforme aux dimensions de A et B

$$X = \begin{pmatrix} Y \\ Z \end{pmatrix}$$

sont décorrélés. Regardons les conséquences sur la densité $p(x)$; pour cela, décomposons de même tout vecteur x en deux sous vecteurs :

$$x = \begin{pmatrix} y \\ z \end{pmatrix}.$$

Comme

$$V^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix}$$

il vient (on suppose ici pour simplifier que $\mu = 0$)

$$x^T V^{-1} x = (y^T z^T) \begin{pmatrix} A^{-1} & 0 \\ 0 & B^{-1} \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = y^T A^{-1} y + z^T B^{-1} z$$

ce qui implique que $p(x)$ s'exprime comme le produit d'une fonction de y par une fonction de z, chacune de ces fonctions étant une densité gaussienne. Par conséquent Y et Z sont indépendantes. On a donc le théorème fondamental :

THÉORÈME IV.3 *Si* $X = \begin{pmatrix} Y \\ Z \end{pmatrix}$ *est un vecteur gaussien tel que* Y *et* Z *sont des vecteurs décorrélés, alors* Y *et* Z *sont indépendants.*

Loi conditionnelle. Soit $X = \begin{pmatrix} Y \\ Z \end{pmatrix}$ un vecteur gaussien de matrice de corrélation

$$C_{xx} = \begin{pmatrix} C_{yy} & C_{yz} \\ C_{zy} & C_{zz} \end{pmatrix}.$$

pour trouver la loi conditionnelle de Y sachant Z on va utiliser la décomposition suivante :

$$Y = (Y - C_{yz}C_{zz}^{-1}Z) + C_{yz}C_{zz}^{-1}Z = W + C_{yz}C_{zz}^{-1}Z. \tag{IV.2}$$

Calculons la corrélation entre W et Z, en supposant pour simplifier les écritures que les variables sont centrées :

$$E[WZ^T] = E[(Y - C_{yz}C_{zz}^{-1}Z)Z^T] = E[YZ^T] - C_{yz}C_{zz}^{-1}E[ZZ^T] = C_{yz} - C_{yz} = 0.$$

La première variable est donc indépendante de Z si bien que la formule (IV.2) décompose Y en une fonction de Z et une variable aléatoire W indépendante de Z ; la variance de W est (on utilise ici la décorrélation de W et de Z) :

$$E[WW^T] = E[(Y - C_{yz}C_{zz}^{-1}Z)W^T] = E[YW^T] = E[Y(Y - C_{yz}C_{zz}^{-1}Z)^T] = C_{yy} - C_{yz}C_{zz}^{-1}C_{zy}$$

et par conséquent W étant gaussienne

$$W \sim \mathcal{N}(E[Y] - C_{yz}C_{zz}^{-1}E[Z], C_{yy} - C_{yz}C_{zz}^{-1}C_{zy}).$$

La loi de $Y = W + C_{yz}C_{zz}^{-1}Z$ sachant $Z = z$ est donc la loi de $W + C_{yz}C_{zz}^{-1}z$

$$\boxed{Loi(Y|Z = z) = \mathcal{N}(C_{yz}C_{zz}^{-1}z, C_{yy} - C_{yz}C_{zz}^{-1}C_{zy}).}$$

Dans le cas non centré, on a de même

$$\boxed{Loi(Y|Z = z) = \mathcal{N}(E[Y] + C_{yz}C_{zz}^{-1}(z - E[Z]), C_{yy} - C_{yz}C_{zz}^{-1}C_{zy}).}$$

formule plus confuse qui se déduit aisément de la précédente. En particulier

$$E[Y|Z = z] = E[Y] + C_{yz}C_{zz}^{-1}(z - E[Z])$$

que l'on peut également écrire

$$E[Y|Z] = E[Y] + C_{yz}C_{zz}^{-1}(Z - E[Z])$$

ou plus simplement, en notant par un tilde les variables recentrées

$$E[\tilde{Y}|Z] = C_{yz}C_{zz}^{-1}\tilde{Z}$$

formule simple qui permet de retouver aussitôt la précédente. La variance conditionnelle ne fait pas intervenir le centrage et possède la propriété remarquable de ne pas dépendre de Z :

$$Var(Y|Z) = C_{yy} - C_{yz}C_{zz}^{-1}C_{zy}.$$

La loi log-normale. C'est simplement la loi de l'exponentielle d'une gausienne scalaire. Cette loi sert à modéliser certaines v.a. positives. Par exemple le modèle de Black-Scholes donne une loi log-normale à la valeur d'une action.

Exemple : Optimisation de portefeuille. Soit p placements donc chacun rapporte par unité de temps et d'argent un gain gaussien $X \sim \mathcal{N}(\mu, V)$: Si au jour j on place une somme unité avec les proportions $c_1, \dots c_p$ dans chaque placement, le gain le lendemain sera une v.a. $X \sim \mathcal{N}(\langle c, \mu\rangle, c^T V c)$. L'objectif est de maximiser le gain moyen sans que le risque ne dépasse une certaine valeur prescrite, soit

$$\max_c \langle c, \mu\rangle \quad \text{sous} \quad \langle c, \mathbf{1}\rangle \leq 1, \quad c^T V c \leq \sigma^2$$

Un calcul donne la solution suivante : posons

$$A = V^{-1}, \quad c_\mu = \frac{A\mu}{\mathbf{1}^T A \mu}, \quad c_1 = \frac{A\mathbf{1}}{\mathbf{1}^T A \mathbf{1}}, \quad \sigma_\mu^2 = c_\mu V c_\mu, \quad \sigma_1^2 = c_1 V c_1.$$

c_x est la stratégie de placement proportionnel à Ax et σ_x^2 le risque correspondant ; alors

1. Si $\sigma \leq \sigma_\mu$: $\quad c = \frac{\sigma}{\sigma_\mu} c_\mu$
2. Si $\sigma > \sigma_\mu$: $\quad c = c_1 + b\,(c_\mu - c_1), \quad b^2 = \frac{\sigma^2 - \sigma_1^2}{\sigma_\mu^2 - \sigma_1^2}$

Même dans le cas simple où $A = Id$, on voit que la solution reste compliquée : pour les petits σ, l'investissement dans chaque placement est proportionnel à ce qu'il rapporte (ne pas mettre tout ses œufs dans le même panier), et l'on a bizarrement $c^T \mathbf{1} < 1$ (du fait de l'interdiction de dépasse le risque σ^2), en revanche, pour σ grand, on fait une combinaison entre cette stratégie et la stratégie consistant à faire un investissement égal (cette dernière ayant un poids négatif, ce qui a pour effet de réduire, voire éliminer, les placements de faible rendement).

Exemple : Ruine d'une compagnie d'assurance. Un calcul déjà compliqué pour un problème simple. Une compagnie d'assurance ouvre boutique avec un capital initial c. On suppose que les recettes mensuelles (primes) valent p et que le montant des sinistres au mois n est une v.a. gaussienne $X_n \sim \mathcal{N}(\mu, \sigma^2)$. Au mois $n+1$, la compagnie a en réserve la somme

$$R_n = c + np - X_1 - X_2 \cdots - X_n$$

Les X_k sont supposées indépendantes. On se doute que si $p < \mu$, alors la compagnie fera faillite en un temps record ; on va supposer que $p \geq \mu$, ce qui va simplifier les écritures. Appelons $\tau(\omega)$ le premier instant que $R_n < 0$, c'est l'instant de ruine de la compagnie. Si la compagnie ne fait jamais faillite, alors $\tau = +\infty$. Formellement

$$\tau = \min\{n : R_n < 0\}.$$

Posons $S_n = Y_1 + \dots Y_n$, $Y_i = X_i - \mu$, on a également

$$\tau = \min\{n : S_n > c + n(p - \mu)\}.$$

Calculons l'expression suivante, pour tout $\lambda > 0$:

$$u_n = e^{-n\lambda^2\sigma^2/2} E[e^{\lambda S_n} 1_{\tau \leq n-1}] = e^{-n\lambda^2\sigma^2/2} E[e^{\lambda Y_n + \lambda S_{n-1}} 1_{\tau \leq n-1}]$$

Comme Y_n est indépendant de $(Y_1 + \dots Y_{n-1})$ et que $1_{\tau \leq n-1}$, l'indicateur d'une ruine avant $n-1$, ne dépend que de $(Y_1 + \dots Y_{n-1})$, on a

$$\begin{aligned}
u_n =& e^{-n\lambda^2\sigma^2/2} E[e^{\lambda Y_n}] E[e^{\lambda S_{n-1}} 1_{\tau \leq n-1}] \\
=& e^{-(n-1)\lambda^2\sigma^2/2} E[e^{\lambda S_{n-1}} (1_{\tau \leq n-2} + 1_{\tau = n-1})] \\
=& u_{n-1} + e^{-(n-1)\sigma^2/2} E[e^{\lambda S_{n-1}} 1_{\tau = n-1}] \\
=& \dots \\
=& \sum_{k=1}^{n-1} e^{-k\lambda^2\sigma^2/2} E[e^{\lambda S_k} 1_{\tau = k}]
\end{aligned}$$

Lorsque n tend vers l'infini, le terme de droite tend vers

$$\begin{aligned}\sum_{k=1}^{\infty} E[e^{\lambda S_k - k\sigma^2/2} 1_{\tau=k}] =& \sum_{k=1}^{\infty} E[e^{\lambda S_\tau - \tau\sigma^2/2} 1_{\tau=k}] \\ =& E[e^{\lambda S_\tau - \tau\lambda^2\sigma^2/2} \sum_{k=1}^{\infty} 1_{\tau=k}] \\ =& E[e^{\lambda S_\tau - \tau\lambda^2\sigma^2/2} 1_{\tau<\infty}]\end{aligned}$$

En ce qui concerne la limite de u_n, on a

$$\begin{aligned}u_n =& e^{-n\lambda^2\sigma^2/2} E[e^{\lambda S_n}(1 - 1_{\tau\geq n})] \\ =& 1 - e^{-n\lambda^2\sigma^2/2} E[e^{\lambda S_n} 1_{\tau\geq n}]\end{aligned}$$

Mais si $\tau \geq n$, $S_{n-1} \leq c + (n-1)(p-\mu)$. Le deuxième terme est donc, en valeur absolue, inférieur à

$$e^{-n\lambda^2\sigma^2/2} E[e^{c+(n-1)\lambda(p-\mu)+\lambda Y_n}] = e^{c-(n-1)(\lambda^2\sigma^2/2-\lambda p+\lambda\mu)}$$

qui tend vers 0 dès que $\lambda\sigma^2/2 - p + \mu > 0$, soit

$$\lambda > \lambda_0 = 2\,\frac{p-\mu}{\sigma^2}$$

ce que l'on va supposer dans la suite. On a donc pour de tels λ :

$$E[e^{\lambda S_\tau - \tau\lambda^2\sigma^2/2} 1_{\tau<\infty}] = 1$$

pour tout $\lambda > \lambda_0$. Comme $S_\tau \geq c + \tau(p-\mu)$, il vient

$$E[e^{\lambda(c+\tau(p-\mu)) - \tau\lambda^2\sigma^2/2} 1_{\tau<\infty}] \leq 1$$

soit exactement

$$E[e^{\lambda\tau(\lambda_0-\lambda)\sigma^2/2} 1_{\tau<\infty}] \leq e^{-\lambda c}$$

la fonction Puisque la fonction $\lambda \mapsto \lambda(\lambda_0 - \lambda)$ est décroissante au voisinage de λ_0, en faisant décroître λ vers λ_0, on a par le théorème de convergence croissante

$$P(\tau < \infty) \leq e^{-c\lambda_0}. \tag{IV.3}$$

Si l'on veut que cette borne soit $\leq 10^{-6}$ on trouve, comme $e^{-13,8} \simeq 10^{-6}$, qu'il suffit que capital initial dépasse

$$c > \frac{13,8}{\lambda_0} = 6,9\,\frac{\sigma^2}{p-\mu}.$$

Notons que l'on obtient assez facilement une borne dans l'autre sens : pour tout $n \geq 0$

$$\begin{aligned}P(\tau < \infty) \geq& P(S_n > c + n(p-\mu)) \\ \geq& E[1_{S_n - c - n(p-\mu) > 0}] \\ =& E[1_{S_n \geq c + n(p-\mu)}] \\ =& P(\mathcal{N}(0,1) > \frac{c + n(p-\mu)}{\sigma\sqrt{n}}) \\ =& P(\mathcal{N}(0,1) > 2\frac{\sqrt{c(p-\mu)}}{\sigma}) \quad \text{avec} \quad n = \frac{c}{p-\mu}\end{aligned}$$

et en utilisant la première borne de l'encadrement

$$\frac{1}{\sqrt{2\pi}A}\left(1 - \frac{1}{A}\right) e^{-\frac{A^2}{2}} \leq P(\mathcal{N}(0,1) > A) \leq \frac{1}{\sqrt{2\pi}A} e^{-\frac{A^2}{2}}.$$

avec $A = 2\sqrt{6,9}$, on trouve pour $c = 6,9\,\frac{\sigma^2}{p-\mu}$: $P(\tau < \infty) \geq 0,06.10^{-6}$. L'encadrement obtenu pour $P(\tau < \infty)$ n'est pas extraodinaire mais on voit bien que la clé de la survie de la compagnie est que le nombre $c(p-\mu)/\sigma^2$ soit assez grand, de l'ordre d'une dizaine.

V

Théorèmes limites

Il existe de nombreux théorèmes-limite avec des variantes également nombreuses en théorie des probabilités. On se restreint ici aux deux plus utilisés : la loi des grands nombres et le théorème-limite central.

V.1 Convergence d'une suite de variables aléatoires

V.1.1 Convergence vers une variable aléatoire spécifique

Convergence presque sûre. Si l'on voit une variable aléatoire comme une fonction de Ω dans $\mathbb{R}$ (ou $\mathbb{R}^d$), la notion de convergence la plus simple est la convergence ponctuelle :

Définition V.1 *On dit qu'une suite X_n de v.a. converge presque sûrement vers une v.a. Y si la convergence suivante a lieu avec probabilité 1*

$$\lim_{n\to\infty} X_n(\omega) = Y(\omega).$$

C'est-à-dire

$$P(\omega : \lim_{n\to\infty} X_n(\omega) = Y(\omega)) = 1.$$

Vu que l'on convient de ne pas distinguer deux v.a. qui diffèrent sur un ensemble de probabilité nulle seulement, la convergence en tout point ne fait pas sens.

La convergence presque sûre de X_n vers Y équivaut revient à dire que la suite a à la convergence presque sûre de $X_n - Y$ vers 0.

La convergence preque sûre est en pratique rarement simple à montrer, voici un critère un peu fort mais qui a le mérite d'être simple

Théorème V.2 *Soit X_n une suite de variables aléatoires telles que pour un $p > 0$ on ait*

$$\sum_{n=1}^{\infty} E[|X_n|^p] < \infty$$

alors X_n converge presque sûrement vers 0.

Démonstration: En effet en raison du théorème de convergence croissante

$$\begin{aligned} E[\sum_{n=1}^{\infty} |X_n|^p] =& E[\lim_k \sum_{n=1}^{k} |X_n|^p] \\ =& \lim_k E[\sum_{n=1}^{k} |X_n|^p] \quad \text{en raison du théorème de convergence croissante} \\ =& \lim_k \sum_{n=1}^{k} E[|X_n|^p] \\ =& \sum_{n=1}^{\infty} E[|X_n|^p] \\ <& \infty \end{aligned}$$

La variable aléatoire $\sum_{n=1}^{\infty} |X_n(\omega)|^p$ est donc d'espérance finie ; elle est donc presque sûrement finie. Ceci implique que la suite $|X_n(\omega)|^p$ converge vers 0, et donc $X_n(\omega)$ également. ∎

Convergence en probabilité. Il existe une notion plus faible de convergence :

DÉFINITION V.3 *On dit qu'une suite X_n de v.a. converge en probabilité vers une v.a. Y si pour tout $\varepsilon > 0$*

$$\lim_{n\to\infty} P(|X_n(\omega) - Y(\omega)| \geq \varepsilon) = 0.$$

Le théorème de convergence dominée de Lebesgue implique bien qu'une suite convergeant presque sûrement converge en probabilité. La réciproque est fausse en général.

La convergence en probabilité est souvent plus facile à montrer que la convergence presque sûre. Donnons un exemple : soit Z_n une suite de v.a. positives d'espérance 1 et

$$X_n(\omega) = \frac{Z_n(\omega)}{n}.$$

Il est naturel de se demander si X_n converge vers 0. On a par l'inégalité de Tchebychev

$$P(|X_n(\omega)| \geq \varepsilon) \leq \frac{E[|X_n|]}{\varepsilon} = \frac{E[|Z_1|]}{n\varepsilon}.$$

ce qui implique que X_n converge en probabilité vers 0 si Z_n est intégrable. La convergence presque sûre est bien plus compliquée à montrer (et peut même ne pas arriver si l'on choisit bien la suite Z_n).

V.1.2 Convergence en loi

DÉFINITION V.4 *On dit qu'une suite X_n de v.a. converge en loi vers une v.a. Y si l'on a convergence des fonctions caractéristiques : pour tout $u \in \mathbb{R}^d$*

$$\lim_{n\to\infty} E[e^{i\langle u, X_n\rangle}] = E[e^{i\langle u, Y\rangle}]$$

La variable Y n'intervient que par sa loi ; «X_n converge en loi vers Y» est en réalité un racourci pour «la loi de X_n converge vers la loi de Y». D'ailleurs on peut très bien dire «X_n converge en loi vers la distribution centrée réduite».

La convergence en probabilité implique la convergence en loi ; pour le montrer le plus simple est de majorer

$$|E[e^{i\langle u, X_n\rangle}] - E[e^{i\langle u, Y\rangle}]| \leq E[|e^{i\langle u, X_n\rangle} - e^{i\langle u, Y\rangle}|]$$

et de terminer en utilisant l'inégalité :

$$|e^{ia} - e^{ib}| \leq 2.1_{|b-a|\geq\varepsilon} + \varepsilon 1_{|b-a|\leq\varepsilon}$$

(car le membre de gauche est toujours ≤ 2 et vaut $|e^{i\frac{a+b}{2}}(e^{i\frac{a-b}{2}} - e^{-i\frac{a-b}{2}})| = 2|\sin((a-b)/2)|$).

Poursuivons ce paragraphe par un théorème un peu délicat à démontrer :

THÉORÈME V.5 *Soit X_n une suite de variables qui converge en loi vers Y, alors pour toute fonction continue bornée*

$$\lim_{n\to\infty} E[f(X_n)] = E[f(Y)].$$

Nous admettrons que si les v.a. sont à valeurs dans $\mathbb{N}$, la convergence en loi est simplement la convergence des probabilités individuelles $P(X_n = m)$ vers $P(Y = m)$, pour tout m.

On aura également besoin du théorème suivant dont la démonstration ne pose en revanche pas de difficulté particulière

THÉORÈME V.6 *Soit X_n une suite de variables qui converge en loi vers Y, et Z_n une autre suite qui converge en loi vers une constante c (c.-à-d. une masse de Dirac) alors X_nZ_n converge en loi vers cY.*

V.2 Loi des grands nombres

LEMME V.7 *(Une condition générale pour la loi des grands nombres) Soit $U_n, n \geq 1$ une suite de variables aléatoires et $S_n = U_1 + U_2 + ...U_n$ telles que :*

$$\begin{aligned} &U_n \geq 0 \quad w.p.1 \\ &n^{-1}E[S_n] \longrightarrow l \\ &Var(S_n) \leq cn \end{aligned}$$

pour des réels c et l, alors

$$\frac{S_n}{n} \longrightarrow l \quad w.p.1.$$

REMARQUE. Il est facile de voir que $Var(S_n) \leq cn^{2-\varepsilon}$ pour un $\varepsilon > 0$ suffit. Remplacer dans la démonstration la suite n^2 par $n^{2/\varepsilon}$.

Démonstration: On a

$$E\left[\sum_n \left(\frac{S_{n^2} - E[S_{n^2}]}{n^2}\right)^2\right] \leq \sum_n \frac{c^2}{n^2} < \infty.$$

Par conséquent $\frac{S_{n^2}-E[S_{n^2}]}{n^2}$ converge presque sûrement vers zero. Donc $\frac{S_{n^2}}{n^2}$ converge vers l. Remarquons que pour $n^2 \leq k \leq (n+1)^2$:

$$\frac{S_{n^2}}{n^2}\frac{n^2}{(n+1)^2} \leq \frac{S_k}{k} \leq \frac{S_{(n+1)^2}}{(n+1)^2}\frac{(n+1)^2}{n^2}$$

et comme les deux côtés tendent vers l, le résultat est démontré. ∎

THÉORÈME V.8 *Soit X_n une suite de variables indépendantes de même loi et d'espérance finie, alors, avec probabilité 1 :*

$$\lim_{n\to\infty} \frac{X_1(\omega) + X_2(\omega) + ...X_n(\omega)}{n} = E[X_1]. \tag{V.1}$$

Démonstration: Faisons la démonstration dans le cas où $E[X_1^2] < \infty$. Appliquons le lemme précédent à $U_n = X_n^+ = \max(X_n, 0)$. Alors, comme $Var(S_n) = nVar(X_1^+)$ et $E[S_n] = E[X_1^+]$, les hypothèses sont bien vérifiées et

$$\lim_{n\to\infty} \frac{X_1^+ + X_2^+ + ...X_n^+}{n} = E[X_1^+].$$

On a le même résultat en remplaçant X_n^+ par $X_n^- = \max(-X_n, 0)$, et comme pour tout x, $x = x^+ - x^-$, on obtient (V.1) en faisant la différence. ∎

V.3 Théorème-limite central

THÉORÈME V.9 *Soit X_n une suite de variables indépendantes de même loi, centrées, de variance R, alors les variables*

$$Y_n(\omega) = \frac{X_1(\omega) + X_2(\omega) + ...X_n(\omega)}{\sqrt{n}}$$

convergent en loi vers la variable $\mathcal{N}(0, R)$.

Mentionnons que ce résultat reste vrai sous des hypothèses plus faibles, par exemple en remplaçant l'hypothèse d'identité des lois par $\sup_i E[|X_i|^3] < +\infty$. Un théorème analogue existe également pour le cas où les variances sont distinctes. Ceci explique pourquoi la distribution gaussienne se rencontre souvent dans la nature, dès qu'un phénomène observé est la somme d'un grand nombre de facteurs indépendants (bruit thermique, prix d'un produit...).

Démonstration: On suppose ici pour simplifier que les X_n sont bornées, le cas général se traitant dans le même esprit mais avec des complications techniques non-négligeables. Commençons par le cas où ces v.a. sont scalaires ; il existe $m > 0$ tel que presque sûrement

$$|X_n(\omega)| \leq m.$$

On va calculer la fonction caractéristique de Y_n ; pour cela il faut noter que pour tout $z \in \mathbb{C}$

$$e^z - 1 - z - \frac{z^2}{2} = \sum_{k=3}^{\infty} \frac{z^k}{k!} = \sum_{k=0}^{\infty} \frac{z^{k+3}}{(k+3)!} = \frac{z^3}{6} \sum_{k=0}^{\infty} \frac{6z^k}{(k+3)!}.$$

et donc

$$|e^z - 1 - z - \frac{z^2}{2}| \leq \frac{|z|^3}{6} \sum_{k=0}^{\infty} \frac{|z|^k}{k!} \leq e^{|z|} \frac{|z|^3}{6}. \tag{V.2}$$

Par conséquent, en raison de l'indépendance, pour tout $t \in \mathbb{R}$

$$E[e^{itY_n}] = E[e^{iX_1/\sqrt{n}}]^n = E\left[\left(1 + it\frac{X_1}{\sqrt{n}} - t^2\frac{X_1^2}{2n} + Z_1 n^{-3/2}\right)\right]^n$$

où $Z_1(\omega)$ est borné par $\frac{m^3}{6}e^m$ en vertu de (V.2) appliqué à $z = iX_1(\omega)/\sqrt{n}$; donc

$$E[e^{itY_n}] = \left(1 - t^2\frac{R}{2n} + E[Z_1]n^{-3/2}\right)^n = \exp\left\{n \log\left(1 - t^2\frac{R}{2n} + E[Z_1]n^{-3/2}\right)\right\}$$

en utlisant que $h - h^2/2 \leq \log(1+h) \leq h$), on a facilement que

$$n \log\left(1 - t^2\frac{R}{2n} + E[Z_1]n^{-3/2}\right) \longrightarrow -\frac{t^2 R}{2}$$

et donc

$$E[e^{itY_n}] \longrightarrow \exp\left(-\frac{t^2 R}{2}\right)$$

qui est bien la fonction caractéristique de la gaussienne. Le théorème est donc démontré dans le cas scalaire borné.

Dans le cas où X_k est vectoriel borné, de covariance R, alors $\langle X_k, u\rangle$ est de variance $u^T R u$, et l'on déduit du résultat scalaire que $\langle Y_n, u\rangle$ converge en loi vers $\mathcal{N}(0, u^T R u)$, ce qui permet d'obtenir directement que la fonction caractéristique de Y_n converge vers celle de $\mathcal{N}(0, R)$. ∎

Exemple d'application : le semeur. On sait qu'une graine a une probablibité $p = 0,75$ de donner naissance à une plante. Combien doit on semer de graines pour être sûr à 99% d'obtenir au moins 50 plantes ?

Le réponse est la suivante : si l'on sème n graines et si X_i est la v.a. qui vaut 0 ou 1 selon l'échec ou la réussite de la i-ième graine, la nombre total de plantes sera

$$N = \sum_{i=1}^{n} X_i.$$

La variable N se réécrit

$$N = np + \sum_{i=1}^{n}(X_i - p) = np + \sqrt{n}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - p)\right).$$

On peut préférer normaliser le dernier terme en tenant compte de ce que chaque X_i a pour variance $p(1-p)$:

$$N = np + \sqrt{n}\sqrt{p(1-p)}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - p}{\sqrt{p(1-p)}}\right) = np + \sqrt{np(1-p)}\,Y_n$$

et Y_n est proche d'une $\mathcal{N}(0,1)$ en raison du théorème-limite central ; en d'autres termes la loi de N s'approxime par une $\mathcal{N}(np, np(1-p))$. On a donc

$$P(N > 50) = P\left(Y_n > \frac{50 - np}{\sqrt{np(1-p)}}\right) \simeq P\left(\mathcal{N}(0,1) > \frac{50 - np}{\sqrt{np(1-p)}}\right)$$

Soit λ la valeur telle que

$$P\left(\mathcal{N}(0,1) > -\lambda\right) = 0,99$$

si l'on veut que $P(N > 50) \geq 0,99$, il faut que

$$\frac{50 - np}{\sqrt{np(1-p)}} < -\lambda$$

On a $P\left(\mathcal{N}(0,1) < -\lambda\right) = 0,01$ et donc par symétrie de la loi normale $P\left(\mathcal{N}(0,1) > \lambda\right) = 0,01$. λ est donc la fonction quantile (inverse de la fonction de répartition) de la gaussienne standard au point $0,99$; on a $\lambda \simeq 2,3$. On obtient donc finalement puisque $p = 3/4$

$$50 - 3n/4 < -2,3\ \sqrt{3n}/4$$

soit

$$3n - 2,3\ \sqrt{3n} - 200 > 0$$

ou en posant $x = \sqrt{3n}$

$$x^2 - 2,3\ x - 200 > 0.$$

Comme x doit être positif et que ce polynôme est négatif en zéro, ceci équivaut à dire que x doit être plus grand que la racine positive :

$$x > \frac{2,3 + \sqrt{2,3^2 + 800}}{2} = 15,34$$

ce qui donne

$$n > 78.$$

Si l'on compte simplement que 3 graines sur 4 se développent, on arrive à $n = 50 \times 4/3 \simeq 67$; cce choix consiste à s'arranger pour que $E[N] = 50$, et il est probable qu'on obtienne en gros une fois sur deux moins de 50 plantes.

VI

Statistique exploratoire univariée et bivariée

VI.1 Les données en première analyse

VI.1.1 Introduction

En statistique *univariée*, les données sont constituées d'un tableau $(X_i)_{1\leq i\leq n}$, ensemble de valeurs qui peuvent être soit

- quantitatives continues. Ex : taille d'un individu.
- quantitatives discrètes. Ex : sortie d'un coup de dé.
- qualitatives. Ex : P ou F, sortie d'un jeu de pile ou face.

Les données quantitatives sont numériques. Les valeurs prises par une variable qualitative sont appelées les *modalités*.

La statistique *bivariée* se consacre à l'étude d'un tableau de données contenant deux variables, par exemple l'âge des mariés en Alaska en 1995. Ce tableau a donc deux colonnes, une par variable (âge de l'homme et âge de la femme), et un nombre arbitraire de lignes (une par individu).

L'idée qui guide les méthodes de statistique exploratoire est l'interprétation de ces données comme des réalisations indépendantes d'une variable aléatoire de loi inconnue.

VI.1.2 Tableaux et tables de contingence

La représentation la plus simple est la suite exhaustive des X_i.

Pour une variable qualitative il y a souvent peu de valeurs possibles vis-à-vis du nombre de données, et l'on peut représenter les données par un double tableau valeurs/nombre d'occurences : La table VI.1 permet de manipuler deux tableaux de longueur fixe 6 au lieu d'un grand tableau dont la longueur est celle de l'échantillon (ici 88). De même le tableau VI.1 représente une paire de variables à deux modalités.

v_j	1	2	3	4	5	6
n_j	12	16	21	10	13	16

	Hommes	Femmes
fumeurs	72	323
non-fumeurs	56	233

Table VI.1 – Résultats de 88 jets de dés. Table de contingence : Évaluation du tabagisme sur 684 individus classés par sexe.

Dans le tableau VI.2, plutôt que de lister l'ensemble des 5514 les paires (âge du marié, âge de la mariée), on a classifié ces paires en 81 classes (discrétisation) ce qui donne le tableau suivant :

F\ H	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55+	TOTAL
15-19	152	323	68	20	4	3	1			571
20-24	56	733	452	146	42	12	7	4	1	1453
25-29	3	157	417	312	136	49	31	7	1	1113
30-34	2	34	141	273	194	107	41	8	11	811
35-39		10	55	116	180	157	70	32	13	633
40-44		4	11	52	80	118	88	40	25	418
45-49			4	18	36	41	79	58	41	277
50-54			1	4	9	16	28	35	48	141
55+							7	8	82	97
TOTAL	213	1261	1149	941	681	503	352	192	222	5514

TABLE VI.2 – L'âge des mariés en Alaska en 1995 (Alaska Bureau of Vital Statistics). L'âge du marié varie en ligne et celui de la mariée en colonne. Chaque case indique un nombre de mariés.

VI.1.3 Histogrammes (variables quatitatives réelles)

Les tableaux ne sont pas très parlants, surtout s'ils sont grands. La représentation par histogramme pour les observations issues d'une variable continue, permet de visualiser aisément la distribution des données (distribution empirique).

Une méthode naturelle consiste à les discrétiser sur des intervalles et à tracer un **histogramme par effectifs**, également appelé « diagramme bâtons », où les ordonnées figurent le nombre n_i (ou la proportion p_i en %) de points observés dans la classe, fig.VI.1.

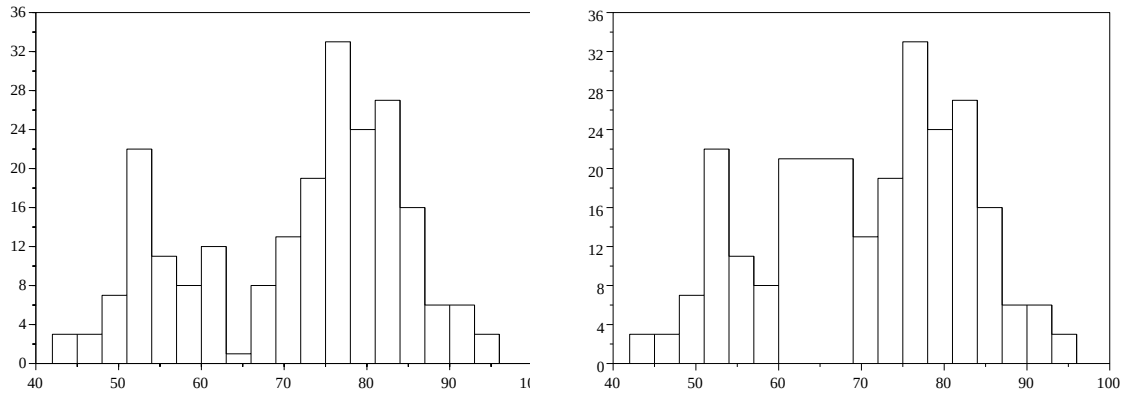


FIGURE VI.1 – Histogramme par effectif des intervalles de temps entre deux éruptions du geyser Old Faithful du parc de Yellowstone. Un total de 222 mesures ont été prises. Il y a 22 mesures entre 51 et 54 mn. Dans le second histogramme, on a fusionné 3 classes de sorte à mettre toutes les données de l'intervalle $[60, 69]$ dans une seule classe.

On voit tout de suite le défaut de cette méthode : le résultat ne correspond pas à ce que l'on attend si les intervalles sont inégaux. Pour pallier à cela on va diviser la hauteur par la largeur de l'intervalle, et pour une raison qui va apparaître, on va également diviser par le nombre total de points ; la hauteur au dessus du i-ième intervalle sera donc $h_i = n_i/(nl_i)$ où l_i est la largeur de l'intervalle. On obtient alors la figure VI.2.

Dans la représentation de la figure VI.2, noter que l'ordonnée h_i ne correspond pas à la probabilité empirique de la classe, mais c'est la surface ; on a donc $h_i l_i = p_i = n_i/n$. La surface totale fait donc 1. L'HISTOGRAMME FOURNIT À LA FOIS UN ESTIMATEUR DE LA DENSITÉ DE PROBABILITÉ ET UNE DESCRIPTION DES DONNÉES.

On voit deux modes se dessiner, laissant penser à deux types d'activité différents.

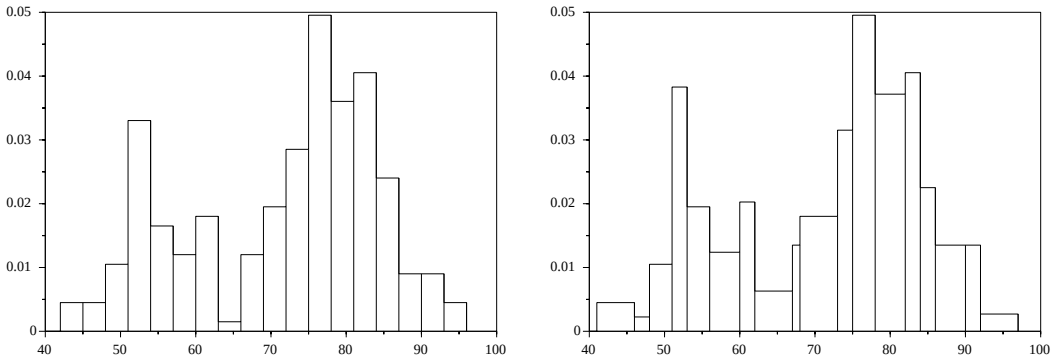


FIGURE VI.2 – Histogramme des données «geyser». Elles ont été discrétisées par intervalles de 3 mn. L'intégrale fait 1. On a observé $0, 01.3.222 = 7$ valeurs entre 48 et 51. On a placé à côté un histogramme avec des classes de taille variable.

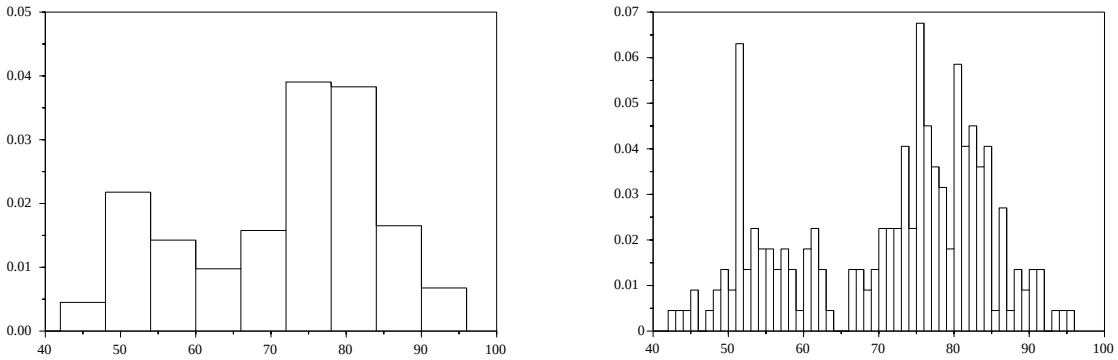


FIGURE VI.3 – Histogramme des données « geyser », discrétisées cette fois par intervalles de 6 mn ; on distingue toujours deux modes. Histogramme avec une discrétisation par intervalles d'1 mn ; les blocs de hauteur 0.005 correspondent à une seule donnée.

Effet de la fusion ou de la scission de classes. Si l'on fusionne deux classes, le bloc résultant aura une hauteur *moyenne* entre les deux blocs initiaux (et sa surface est la somme des surfaces). L'histogramme garde donc le même aspec, figure VI.3. Si les classes sont trop petites vis-à-vis du nombre d'échantillons, l'histogramme obtenu peut être assez mauvais car illusoirement précis, figure VI.3.

VI.1.4 Digression : un estimateur de la densité.

Pour une variable continue, les représentations du § VI.1.3 donnent une estimation de la densité de la variable, i.e. de la dérivée de $F(x)$. Plutôt que d'utiliser un histogramme (figure VI.2), qui donne une estimée très irrégulière, il est courant d'estimer la densité par une formule du type

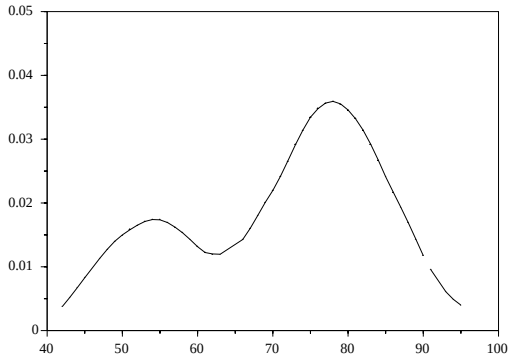
$$p_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

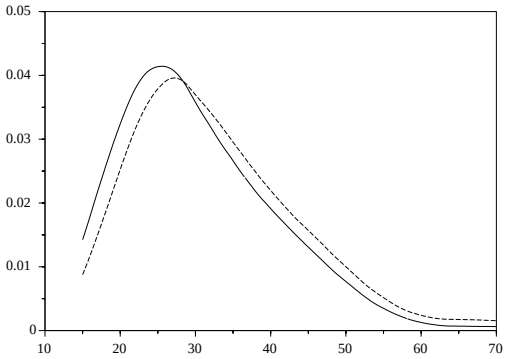
où K est une fonction positive d'intégrale 1 (de sorte que p_n est aussi d'intégrale 1) et h un réel, bien choisis. Si K est l'indicateur de $[-1/2, 1/2]$, $p_n(x)$ n'est autre que la proportion d'echantillons observés dans un voisinage de taille h de x. Cette formule peut s'interpréter comme une façon particulière de régulariser l'histogramme. Le choix de K et du réel h est un domaine difficile des statistiques. Pour K,

un bon choix est le noyau d'Epanechnikov

$$K(x) = \frac{3}{4\sqrt{5}}\left(1 - \frac{x^2}{5}\right), \quad |x| < \sqrt{5}.$$

Quant à h, il doit être assez petit pour n grand : $h \simeq sn^{-1/5}$ est préconnisé pour ce choix de K (s est l'écart-type empirique défini plus bas) ; toutefois en pratique, h est souvent choisi à la main à une valeur qui semble raisonnable. Les figures qui suivent représentent l'estimation de densité sur les données « geyser » (on a pris $h = sn^{-1/5} = 4,33$), ainsi que les densités des variables « âge du (de la) marié(e) le jour du mariage » en Alaska en 1995 (Les femmes sont en trait plein et les hommes en pointillés) :





VI.2 La distribution empirique

Comme nous l'avons dit plus haut, tout ce qui suit est guidé par le modèle qui fait des données $(X_i)_{1 \leq i \leq n}$ une suite d'observations indépendantes de même loi. L'hypothèse clef est donc l'**homogénéité** des données (identité des loi).

VI.2.1 Distribution et moyennes empiriques

La moyenne empirique d'une fonction f des données $(X_i)_{1 \leq i \leq n}$ est

$$\frac{1}{n}\sum_i f(X_i)$$

et la *probabilité empirique* d'un ensemble A est la proportion d'échantillons tombés dans A :

$$\frac{1}{n}\sum 1_A(X_i).$$

La *distribution empirique* est donc celle qui attribue à chaque valeur une probabilité égale à sa fréquence d'observation. C'est celle que l'on observe en tirant (avec remise) des échantillons au hasard dans l'ensemble des observations.

Exemple : dans le cas des 88 jets de dés (table VI.1) la loi empirique a les poids suivants

j	1	2	3	4	5	6
p_j	12/88	16/88	21/88	10/88	13/88	16/88

Dans le cas du tableau VI.1, en normalisant simplement par le nombre d'individus, on obtient le tableau :

	H	F
fumeur	0,105	0,472
non-fumeur	0,082	0,341

La loi des grands nombres assure que si les variables $(X_i)_{i=1,2...}$ sont indépendantes de même loi, alors, pour toute fonction f continue par morceaux bornée (et même pour bien d'autres...), leur moyenne empiriques convergent vers l'espérance de f sous cette loi, $E[f(X)]$.

Sous des conditions raisonnables, cette propriété reste vraie pour des variables non-indépendantes, l'hypothèse essentielle restant l'identité des lois.

Les moyennes empiriques sont donc l'approximation la plus naturelle des espérances à partir des données.

Cas de données discrétisées en intervalles. Il arrive que les variables originales soient discrétisées sur des intervalles, par exemple si l'âge du marié est donné par tranche d'âges (tableau VI.2) ; dans ce cas on ne peut pas retrouver la loi empirique des données originales. On prend alors conventionnellement comme loi empirique soit la loi *discrète* qui attribue à chaque milieu d'intervalle la probabilité de ce dernier (on fait comme si tous les mariés de l'intervalle 20-24 avaient 22ans), soit la loi *uniforme par morceaux* dont la densité est donnée par l'histogramme correspondant (on fait comme si les mariés de l'intervalle 20-24 sont uniformément répartis). Cette dernière solition est plus naturelle car la vraie loi a une densité mais rend le calcul des espérances empiriques plus difficles. Ces deux choix donnent des résultats différents en général sauf pour le calcul de la moyenne. Par exemple, si au lieu des données complètes, on ne dispose que du tableau VI.2, on estimera l'âge moyen du marié par :

$$\frac{17 \times 213 + 22 \times 1261 + ... + 52 \times 192 + 57 \times 222}{5514} = 32 \text{ ans et } 3 \text{ mois.}$$

La valeur 57 choisie ici est arbitraire et discutable.

VI.2.2 Fonction de répartition

La fonction de répartition empirique de l'échantillon est

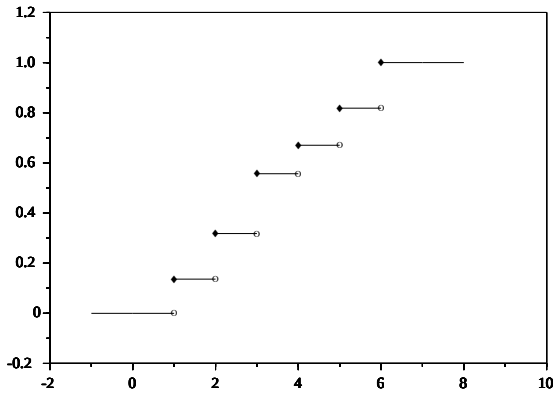
$$F_n(y) = \frac{1}{n} \sum_i 1_{x_i \leq y} = \frac{\text{nb de valeurs observées} \leq y}{\text{nb total de valeurs observées}}. \tag{VI.1}$$

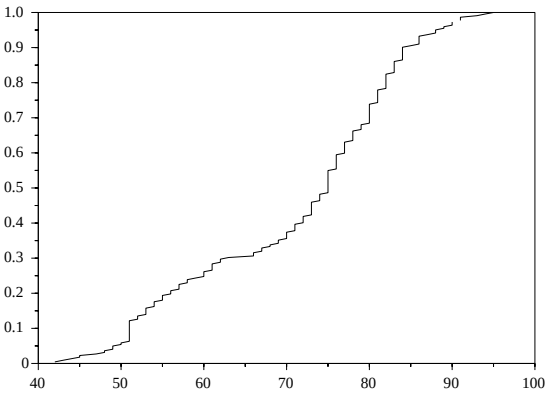
$F_n(y)$ est la fréquence empirique d'apparition de valeurs strictement inférieures à y. La fonction de répartition n'est donc définie que pour les variables prenant des valeurs numériques.

La loi des grands nombres implique que si les variables $(X_i)_{i=1,2...}$ sont de même loi P, alors pour tout y,

$$\lim_n F_n(y) = F(y) = P(X \leq y).$$

Si la fonction de répartition empirique est moins parlante que l'histogramme, son avantage est que la formule (VI.1) permet de représenter directement une variable continue sans faire de discrétisation préalable en intervalles (de taille arbitraire) ; voici les fonctions de répartition correspondant à la table VI.1 et des données « geyser » :





VI.2.3 Quantiles

La fonction quantile empirique est la fonction Q_n approximativement inverse de F_n (elle n'est pas inversible). $Q_n(\alpha)$ **est la valeur qui sépare les données en proportion** α **et** $1-\alpha$. Elle est définie précisément par :

$$Q_n(\alpha) = \check{x}_i \quad \text{si} \quad \frac{i-1}{n} \leq \alpha < \frac{i}{n}.$$

où la suite $(\check{x}_i)$ est la suite des x_i réordonnés par ordre croissant. On note généralement les quartiles : $Q1 = Q_n(0,25)$, $Q2 = Q_n(0,5)$, $Q3 = Q_n(0,75)$.

Dans la figure qui suit, les données « geyser » sont représentées sur l'axe réel par leur valeur. À gauche de la valeur $Q1$ (comme à droite de $Q3$) se trouve un quart des données :

Boîtes de dispersion. Il s'agit d'un format de représentation des données par une simple boîte de largeur (ou hauteur) $Q_3 - Q_1$, où Q_2 est indiqué par une séparation (chaque compartiment contient donc un quart des données) ; cette boîte est prolongée par deux traits : l'extrémité du premier correspond à la première donnée supérieure à $Q_1 - 1,5\Delta$, où $\Delta = Q_3 - Q_1$ et l'autre à la plus grande inférieure à $Q_3 + 1,5\Delta$. Les données extérieures à l'ensemble, c.-à-d. hors de $[Q_1 - 1,5\Delta, Q_3 + 1,5\Delta]$, sont représentées individuellement (s'il y en a). Cette représentation est surtout utilisée pour comparer graphiquement différents groupes.

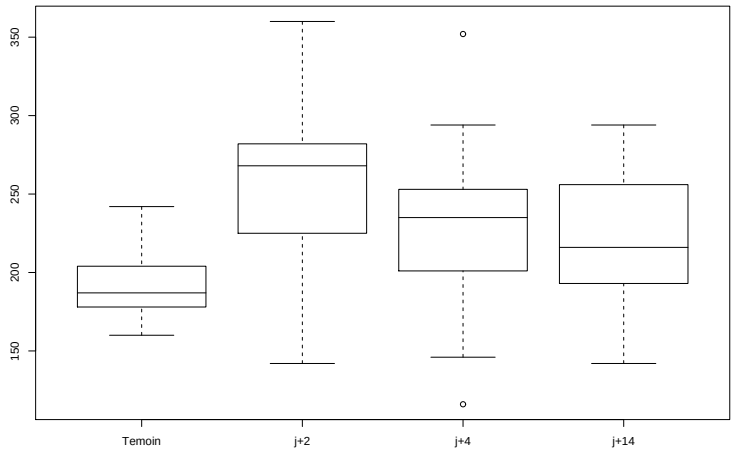


FIGURE VI.4 – On mesure le taux de cholesterol sur un groupe temoin et sur un autre groupe de patients ayant eut une crise cardiaque, 2,4 et 14 jours après la crise. D'après OzDASL.

POURQUOI $1,5\Delta$? Pour une v.a.gausienne, l'intervalle $[Q_1 - 1,5\Delta, Q_3 + 1,5\Delta]$ contient plus de 99% de la masse ; on a donc ainsi, même s'il y a peu de données, une estimation (plus ou moins réaliste) d'un intervalle en dehors duquel on ne s'attend pas à trouver plus de 1% des données.

VI.3 Indices synthétiques essentiels

VI.3.1 Mesures de localisation

La moyenne empirique est la moyenne arithmétique des données :

$$\bar{x} = \frac{1}{n} \sum x_i.$$

C'est aussi la valeur pour laquelle la somme des carrés des distances des données à cette valeur est minimale :

$$\bar{x} = \arg\min_y \sum_i (x_i - y)^2.$$

La moyenne des 222 données « geyser » est de 71mn.

Médiane. S'il y a un nombre impair de données, c'est la valeur centrale $\check{x}_{(n+1)/2}$. Sinon, c'est par convention la moyenne des deux valeurs centrales.

La médiane des 222 données « geyser » est de 75mn, ce que confirme la figurede la page 49. Sur cette figure, on trouve la première médiane à 3. Noter la dissymétrie de la distribution des données « geyser » (§ VI.1.3) qui déplace la médiane (par rapport à la moyenne) vers une zone de plus grande probabilité.

INVARIANCE PAR CHANGEMENT D'ÉCHELLE MONOTONE : Si $y_i = f(x_i)$ pour une certaine fonction *monotone* f et si Q_2 est la médiane des x_i, alors $f(Q_2)$ est la médiane des y_i. C'est une propriété importante qui n'est pas satisfaite par la moyenne.

VI.3.2 Mesures de dispersion

On cherche ici à quantifier l'étendue des données.

La variance empirique est la la variance de la distribution empirique, c-à-d la quantité définie par

$$Var(x) = \frac{1}{n}\sum_i (x_i - \bar{x})^2 = \frac{1}{n}\sum_i x_i^2 \; - \; \bar{x}^2.$$

On la note aussi s_x^2. Sa racine s_x est l'**écart-type empirique**, et sa dimension est celle des données. Sur les données « geyser », l'écart-type est $s = 12,8$.

On considère aussi des **intervalles interquantile**, par exemple $Q_3 - Q_1$ qui est l'étendue de la zone centrale contenant la moitié des données (cf § VI.2.3).

VI.3.3 Corrélation

Soient $x = (x_1, ...x_n)^T$ et $y = (y_1, ...y_n)^T$ deux vecteurs de $\mathbb{R}^n$, par exemple des paires (âge, revenu) pour des individus différents. On note leur moyennes respectives $\bar{x}$ et $\bar{y}$, et les vecteurs recentrés $\tilde{x}$ et $\tilde{y}$:

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \vdots \\ \tilde{x}_n \end{pmatrix}, \quad \bar{x} = \frac{1}{n}\sum_j x_j\,, \quad \tilde{x}_i = x_i - \bar{x}.$$

PROPRIÉTÉ VI.1 *La covariance empirique de x et y est*

$$Cov(x,y) = \frac{1}{n}\sum_i \tilde{x}_i\tilde{y}_i = \overline{xy} - \bar{x}\bar{y}$$

et le coefficient de corrélation linéaire vaut

$$Cor(x,y) = \frac{Cov(x,y)}{s_x s_y} = \frac{\langle \tilde{x}, \tilde{y}\rangle}{\|\tilde{x}\|_2 \; \|\tilde{y}\|_2}.$$

$Cor(x,y)$ est donc le cosinus de l'angle que forment les vecteurs des données centrées. $Cor(x,y)$ et $Cov(x,y)$ sont aussi notés parfois r_{xy} et c_{xy}.

PROPRIÉTÉ VI.2 $|Cor(x,y)| \leq 1$. $|Cor(x,y)| = 1$ *si et seulement s'il existe* $(a,b) \in \mathbb{R}_* \times \mathbb{R}$ *tels que* $y_i = ax_i + b$, $i = 1, ..n$ *; dans ce cas* $Cor(x,y) = signe(a)$.

Une forte corrélation signifie donc que les y sont quasiment fonction linéaire des x. En pratique une faible corrélation sera souvent interprétée (abusivement) comme de l'indépendance (par analogie avec le cas gaussien).

Ainsi, l'importance de l'exposition au soleil pour le cancer de la peau a été démontrée simplement en calculant la corrélation entre les taux de cancer dans certaines régions et la latitude.

La corrélation entre l'âge de la mariée et l'âge du marié (tableau VI.2) est de 0,8.

VI.3.4 Cas de données qualitatives

Soient $a_1, ...a_p$ (resp. $b_1, ...b_q$) les modalités pouvant être prises par x (resp. y). On peut alors construire le tableau de contingence (n_{ij}) de taille $p \times q$

$$n_{ij} = \#\{k : \ x_k = a_i, y_k = b_j\}.$$

Par exemple le tableau VI.1 p.45 ou encore le tableau VI.2 p.46 (si l'on considère la tranche d'âge comme une variable qualitative).

Il est en fait bien plus commode de raisonner sur les probabilités (fréquences) empiriques que l'on note

$$\hat{p}_{ij} = \frac{n_{ij}}{n}.$$

On pose également

$$\hat{p}_{i.} = \sum_j \hat{p}_{ij}, \quad \hat{p}_{.j} = \sum_i \hat{p}_{ij}.$$

Si x et y forment un tableau de données empiriquement indépendantes, on doit avoir pour toute paire i, j

$$\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j}.$$

La mesure la plus classique d'indépendance est le coefficient de contingence de Pearson :

Définition VI.3 *Le coefficient de contingence de Pearson entre x et y est*

$$\Phi^2 = n \sum_{ij} \frac{(\hat{p}_{ij} - \hat{p}_{i.}\hat{p}_{.j})^2}{\hat{p}_{i.}\hat{p}_{.j}}$$

Si x_i et y_i sont des variables aléatoires indépendantes, on vérifie (par la loi des grand nombres) que Φ^2 tend vers en loi vers une certaine distribution quand le nombre de données tend vers l'infini, et en revanche, en cas de dépendance, il est d'ordre n. La normalisation de chaque terme par $\hat{p}_{i.}\hat{p}_{.j}$ vient de considérations statistiques.

Dans le cas de l'exemple du tableau VI.1, on trouve $\Phi^2 = 0,14$. Si l'on remplace 72 par 152, une nette dépendance entre le sexe et le tabagisme apparaît et l'on trouve $\Phi^2 = 14$

VI.3.5 Corrélation partielle

Considérons l'exemple suivant : pour mesurer l'efficacité des pompiers, on calcule la corrélation entre le nombre de pompiers p envoyés sur un sinistre et le montant des dégâts d (en euros). On trouve

$$r_{pd} = 0,7.$$

Faut-il conclure de cette corrélation élevée que pour réduire les dégâts il faut diminuer les effectifs de pompiers ? Cette conclusion serait exacte si la statistique portait sur des incendies de même ampleur initiale. En d'autres termes, on voit bien que cette variable ampleur initiale a étant fortement corrélée à p et d introduit une corrélation « artificielle » entre le nombre de pompiers envoyés et les dégâts. Il faudrait calculer la corrélation sur des incendies d'ampleur initiale fixe, puis faire ensuite la moyenne sur cette variable.

La solution mathématique, justifiée plus loin, est de calculer la corrélation partielle entre p et d sachant a

Définition VI.4 *Le coefficient de corrélation partielle entre x et y sachant z est donnée par la formule*

$$r_{xy|z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}} \tag{VI.2}$$

Reprenons notre exemple. Si l'on a

$$r_{pa} = 0,9 \qquad \text{et} \qquad r_{va} = 0,9$$

alors $r_{pd|a} = -0,6$ et le signe est correct !

On interprète la nullité de $r_{xy|z}$ comme un indice d'indépendance de x et y conditionnellement à z : lorsque z est connu, y n'apporte aucune information supplémentaire sur x.

Voici deux exemples sur des données réelles où l'on voit la corrélation changer de signe (Kendall et Stuart, *Advanced Theory of Statistics*, Vol 2, Ex 27.1 & 27.2) :

1. Des statistiques portant sur 20 ans dans une région de Grande Bretagne

   $x =$ rendement à l'hectare
   $y =$ température moyenne au printemps
   $z =$ chutes de pluie

   et l'on trouve

   $$r_{yx} = -0,4, \quad r_{xz} = 0,8, \quad r_{yz} = -0,56, \quad r_{yx|z} = 0,1$$

2. Des statistiques portant sur 16 villes des Etats-Unis en 1935

   $x =$ criminalité
   $y =$ fréquentation des églises
   $z =$ nombre d'enfants par famille

   et l'on trouve

   $$r_{yx} = -0,31, \quad r_{xz} = -0,14, \quad r_{yz} = 0,85, \quad r_{yx|z} = 0,25.$$

La formule (VI.2) se justifie par le

THÉORÈME VI.5 *Si X, Y, Z forme un vecteur gaussien, et si r_{xy}, r_{xz}, r_{yz} désignent les trois corrélation, alors $r_{xy|z}$ donné par la formule (VI.2) est la corrélation entre X et Y pour la loi conditionnelle à Z :*

$$r_{xy|z} = \frac{E[XY|Z] - E[X|Z]E[Y|Z]}{\sqrt{E[X^2|Z] - E[X|Z]^2}\ \sqrt{E[Y^2|Z] - E[Y|Z]^2}}$$

(qui ne dépend pas de Z).

Démonstration: Soit $U = \begin{pmatrix} X \\ Y \end{pmatrix}$, on a vu que

$$Loi(U|Z = z) = \mathcal{N}(C_{uz}C_{zz}^{-1}z, C_{uu} - C_{uz}C_{zz}^{-1}C_{zu}).$$

La corrélation entre X et Y pour cette loi est donnée par le coefficient (1,2) de la matrice 2×2 de covariance de cette loi. Il vaut

$$C_{xy} - C_{xz}C_{zz}^{-1}C_{yz}.$$

On a donc

$$r_{xy|z} = \frac{C_{xy} - C_{zz}^{-1}C_{xz}C_{yz}}{\sqrt{C_{xx} - C_{zz}^{-1}C_{xz}^2}\ \sqrt{C_{yy} - C_{zz}^{-1}C_{yz}^2}}$$

qui se réarrange bien en (VI.2). ∎

VII

Estimation. Tests. Exemples

VII.1 Introduction

Le but de l'estimation est de calculer une certaine quantité dépendant de la distribution d'une variable aléatoire Y. Cette quantité peut être un moment :

$$\theta^* = E[Y^4]$$

un quartile

$$P(Y < \theta^*) = 0,25$$

ou autre....

Dans un problème d'estimation, on ignore la distribution de Y, mais on a à sa disposition une suite $Y_1, ...Y_n$ de v.a. indépendantes de loi commune identique à celle de Y.

La majeur partie des estimateurs connus d'une certaine quantité θ^* sont obtenus, à quelques modifications près, par un principe d'empirisme :

- exprimer θ^* comme une fonction de la distribution de Y
- définir l'estimateur $\hat{\theta}_n$ comme la valeur de cette fonction sur la distribution empirique.

Pour le premier exemple la fonction est « espérance de la puissance 4 » et l'on obtient

$$\hat{\theta}_n = \frac{1}{n} \sum Y_k^4$$

et dans le second l'estimée sera le premier quartile des données, c'est-à-dire la valeur qui sépare les 25% plus petites données de 75% plus grandes.

La convergence de θ_n vers θ^* sera généralement une conséquence, plus ou moins directe, de la loi des grands nombres.

Dans un cadre plus général de données dépendantes, on a plutôt recours à des estimateurs du type «maximum de vraisemblance» que l'on ne considèrera pas ici, mais qui souvent peuvent s'interpréter comme plus haut. Leur convergence est encore basée sur des versions de la loi des grands nombres.

VII.2 Quelques estimateurs. La loi des grands nombres

Moyenne. Soit une suite de variables aléatoires $Y_1, ...Y_n$ de même loi. On voudrait calculer leur espérance commune m et leur variance v. L'estimateur empirique est :

$$\hat{m}_n = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

$\hat{m}$ est l'espérance de la variable aléatoire obtenue par tirage uniforme dans l'ensemble $\{Y_1, ...Y_n\}$. La loi des grands nombres assure que ces quantités convergent vers m :

$$\lim_n \hat{m}_n = E[Y] = m.$$

Variance. De même, un estimateur de la variance est la variance empirique

$$\hat{v}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \hat{m}_n^2.$$

La loi des grands nombres assure que $\hat{v}_n$ est convergent :

$$\lim_n \hat{v}_n = \lim_n \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \lim_n \hat{m}_n^2 = E[Y^2] - E[Y]^2 = v.$$

L'estimateur de l'écart-type est

$$\hat{\sigma}_n = \sqrt{\hat{v}_n}$$

Corrélation. Soit une suite de variables aléatoires $(Y_1, Y_1'), ...(Y_n, Y_n')$ de même loi. Comme précédemment, des estimateurs convergents de leur corrélation et de leur covariance sont

$$\hat{c}_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}_n)(Y_i' - \hat{m}_n') = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i' - \hat{m}_n\hat{m}_n'$$
$$\hat{r}_n = \frac{\hat{c}_n}{\sqrt{\hat{v}_n\hat{v}_n'}}$$

où $\hat{m}_n$, $\hat{m}_n'$, $\hat{v}_n$ et $\hat{v}_n'$ sont les estimateurs définis précédemment de la moyenne et de la variance pour les deux lois. La convergence se montre de la même façon.

Fonction de répartition. De même, si l'on veut estimer la fonction de répartition en un point x, c'est-à-dire la probabilité que la variable Y soit inférieure à x, on utilisera la fonction de répartition empirique

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{Y_i \leq x} = \frac{1}{n}\ \#\{i :\ Y_i \leq x\}.$$

et

$$\lim_n \hat{F}_n(x) = E[1_{Y\leq x}] = P(Y \leq x) = F(x).$$

Quantile. On cherche la plus petite valeur A qui n'est dépassée par y qu'avec probabilité disons 5%, c'est-à-dire la solution de

$$P(Y > A) = 5\%.$$

On suppose pour simplifier que la fonction de répartition de Y est continue. Comme $A = F^{-1}(0,95)$ un estimateur naturel est

$$\hat{A}_n = F_n^{-1}(0,95).$$

C'est-à-dire que $\hat{A}_n$ est simplement la valeur telle que 5% des données lui sont supérieures. Cet estimateur découle également de l'utilisation de la loi des grands nombres, mais cette fois-ci de manière indirecte.

Probabilité d'un Bernoulli. Si chaque Y_i est $\mathcal{B}(p,1)$, c'est-à-dire vaut 1 avec probabilité p et 0 avec probabilité $1-p$, alors p est l'espérance de Y et une estimée est

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

DÉFINITION VII.1 *Soit $Y_1, Y_2, ...Y_n$ une suite i.i.d. et θ^* une quantité dépendant de la distribution commune aux Y_i (une fonction des moments, un quantile,...). Soit $\hat{\theta} = \hat{\theta}(Y_1, ...Y_n)$ une fonction de $Y_1, ..Y_n$. On dit que $\hat{\theta}$ est un estimateur non-biaisé de θ^* si*

$$E[\hat{\theta}] = \theta^*.$$

On dit que la suite d'estimateur $\hat{\theta}_n = \hat{\theta}_n(Y_1, ...Y_n)$ est un estimateur fortememnt convergent (ou consistant) de θ^* si avec probabilité 1

$$\lim_{n\to\infty} \hat{\theta}_n = \theta^*.$$

L'estimateur $\hat{m}_n$ est sans biais. En revanche, on vérifie que $\hat{v}_n$ est un estimateur biaisé de v. Les estimateurs qu'on a vu jusqu'à présent sont tous convergents.

VII.3 Loi asymptotique des estimateurs

Les estimateurs sont des variables aléatoires, puisque ce sont des fonctions des observations. Il sont construits de sorte à avoir une faible variance autour d'une valeur à trouver. On peut étudier cela de plus près.

VII.3.1 Normalité asymptotique

On va voir que les estimateurs étant généralement basés sur des moyennes empiriques (parfois de manière indirecte), leur comportement asymptotique est essentiellement gouverné par le théorème-limite central et leur distribution asymptotique, après normalisation sera gaussienne. On aura donc très souvent une vitesse de convergence en $n^{-1/2}$ avec la limite en loi

$$\sqrt{n}\ (\hat{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0, \sigma^2),$$

pour un σ à calculer.

Estimateur de la moyenne. En vertu du théorème-limite central,

- $\sqrt{n}\ (\hat{m}_n - m) \longrightarrow \mathcal{N}(0, v)$,

est asymptotiquement gaussien de variance v (variance de chaque Y_i).

Estimateur de la variance. La distribution asymptotique de $\hat{v}_n$ est plus difficile à obtenir puisque l'expression donnée plus haut pour $\hat{v}_n$ ne se présente pas comme une moyenne de v.a. indépendantes. Comme $\hat{v}_n$ est aussi la variance empirique de la suite $(Y_i - m)$ on a

$$\hat{v}_n = \frac{1}{n}\sum_{i=1}^{n}(Y_i - m)^2 - (\hat{m}_n - m)^2.$$

Il s'ensuit que

$$\sqrt{n}(\hat{v}_n - v) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[(Y_i - m)^2 - v] - \sqrt{n}(\hat{m}_n - m)^2.$$

En raison de ce qui précède, le deuxième terme tend vers 0 et le premier converge en loi vers une variable gaussienne de variance :

$$v_e = E[((Y_i - m)^2 - v)^2] = E[(Y_i - m)^4] - v^2.$$

et donc

- $\sqrt{n}\ (\hat{v}_n - v) \longrightarrow \mathcal{N}(0, v_e)$.

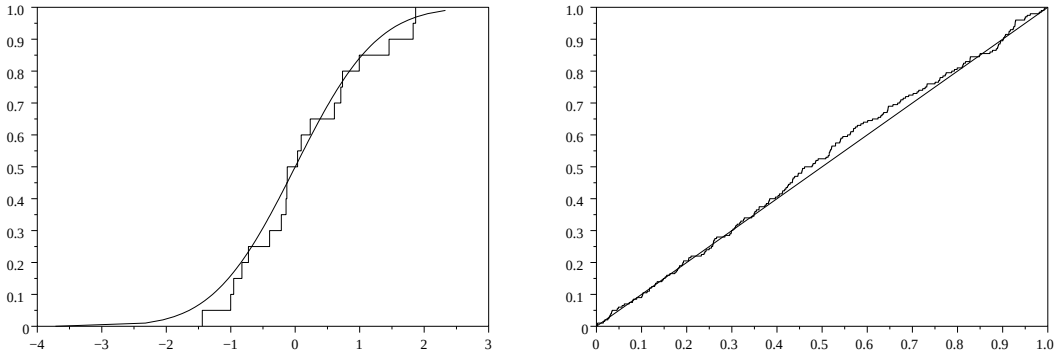


FIGURE VII.1 – Exemple de fonction de répartition empirique d'un échantillon de 20 valeurs gausiennes centrées réduites et fonction de répartition de la gaussienne. Même expérience avec 200 v.a. uniformes.

Estimateur de la covariance et de la corrélation. Par une méthode analogue on montre que **si les variables Y et Y' sont indépendantes** on a

- $\sqrt{n}\ \hat{c}_n \longrightarrow \mathcal{N}(0, \sigma^2 \sigma'^2)$, et $\sqrt{n}\ \hat{r}_n \longrightarrow \mathcal{N}(0,1)$.

S'il y a dépendance, on trouve des formules plus compliquées.

Bernoulli. Le théorème-limite central implique que

- $\sqrt{n}(\hat{p}_n - p) \longrightarrow \mathcal{N}(0, p(1-p))$.

Fonction de répartition. De la même façon, puisque $1_{Y_i<x}$ est un Bernoulli

- $\sqrt{n}(F_n(x) - F(x)) \longrightarrow \mathcal{N}(0, F(x)(1-F(x)))$.

Médiane. Soit m la médiane de la loi commune aux Y_i, et $\hat{m}_n \simeq \check{Y}_{n/2}$ la médiane de l'échantillon ; on peut montrer que

- $\sqrt{n}(\hat{m}_n - m) \longrightarrow \mathcal{N}(0, 1/(4f(m)^2))$.

où f est la densité de Y.

VII.3.2 Théorème de Kolmogorov

On a vu que la construction des estimateurs résultait souvent du remplacement de F (fonction de répartition de Y) par la fonction de répartition empirique F_n. Il existe deux théorèmes qui quantifient la proximité de ces deux fonctions. On posera

$$d_n = \sup_x |F_n(x) - F(x)|.$$

THÉORÈME VII.2 *(Glivenko-Cantelli) Avec probabilité 1, d_n converge vers 0.*

La figure VII.1 illustre ce phénomène. Le deuxième est une convergence en loi

THÉORÈME VII.3 *(Kolmogorov) Si $F(x)$ est continue, la loi de d_n est indépendante de F et la suite $\sqrt{n}d_n$ converge en loi :*

$$\lim_{n\to\infty} P(\sqrt{n}d_n < y) = 1 + 2\sum_{k=1}^{+\infty} (-1)^k e^{-2k^2y^2}.$$

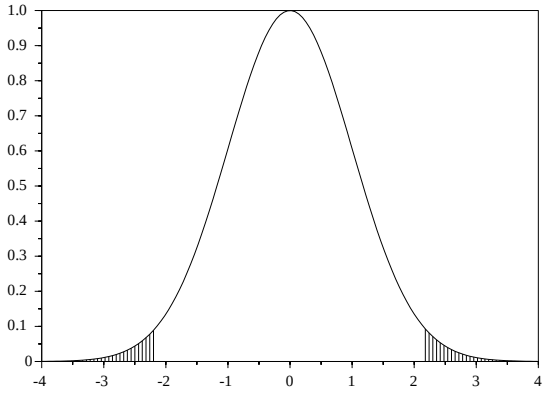


FIGURE VII.2 – Densité de la gaussienne. Chaque région hachurée est d'intégrale $\alpha/2$. Les abscisses correspondantes $(\pm 2, 2)$ sont $\pm M_\alpha$. La probabilité de tomber dans la région non-hachurée est $1-\alpha$.

Le fait que la loi de d_n est indépendante de la loi des Y_k est un fait simple à vérifier si F est strictement croissante :

$$\begin{aligned} d_n &= \sup_y |F_n(F^{-1}(y)) - y|, \quad \text{on a posé } F(x) = y \\ &= \sup_y |\frac{1}{n}\sum_{k=1}^{n} 1_{Y_k \leq F^{-1}(y)} - y| \\ &= \sup_y |\frac{1}{n}\sum_{k=1}^{n} 1_{F(Y_k) \leq y} - y| \end{aligned}$$

comme les $F(Y_k)$ sont i.i.d $\mathcal{U}([0,1])$ la loi de d_n est celle de

$$\sup_y |\frac{1}{n}\sum_{k=1}^{n} 1_{U_k \leq y} - y|$$

où les U_k sont i.i.d $\mathcal{U}([0,1])$.

VII.4 Intervalles de confiance

VII.4.1 Introduction. Définition

Plaçons nous dans le cas où l'on a la convergence en loi de l'estimateur $\hat{\theta}$:

$$\frac{\sqrt{n}}{\sigma}(\hat{\theta}_n - \theta^*) \longrightarrow \mathcal{N}(0,1), \quad \text{en loi.} \tag{VII.1}$$

Notons M_α la valeur telle que la variable gaussienne tombe dans l'intervalle $[-M_\alpha; M_\alpha]$ avec probabilité $1-\alpha$ (cf figure VII.2) :

$$\alpha = \frac{1}{\sqrt{2\pi}}\int_{\infty}^{-M_\alpha} e^{-x^2/2}dx + \frac{1}{\sqrt{2\pi}}\int_{M_\alpha}^{+\infty} e^{-x^2/2}dx = 2\frac{1}{\sqrt{2\pi}}\int_{M_\alpha}^{+\infty} e^{-x^2/2}dx.$$

Si l'on considère que $\sigma^{-1}\sqrt{n}(\hat{\theta}_n - \theta^*)$ a une loi gaussienne, alors cette variable est comprise entre $-M_\alpha$ et M_α avec probabilité α. On a donc la relation

$$-M_\alpha \leq \frac{\sqrt{n}}{\sigma}\,(\hat{\theta}_n - \theta^*) \leq M_\alpha \quad \text{avec probabilité } 1-\alpha$$

qui se réécrit

$$\theta^* \in \left[\hat{\theta}_n - \frac{\sigma M_\alpha}{\sqrt{n}}, \hat{\theta}_n + \frac{\sigma M_\alpha}{\sqrt{n}}\right], \quad \text{avec probabilité } 1-\alpha.$$

Intervalle à 99%. Pour un seuil $\alpha = 1\%$, on a $M_\alpha = 2,6$ ($P(|\mathcal{N}(0,1)| > 2,6) \simeq 0,01$)

$$\theta^* \in [\hat{\theta}_n - 2,6\sigma/\sqrt{n}, \hat{\theta}_n + 2,6\sigma/\sqrt{n}] \quad \text{avec probabilité } 0,99.$$

Cet intervalle de confiance est asymptotique, c'est-à-dire valide en théorie pour n grand seulement. On peut noter plus simplement

$$\boxed{\theta^* = \hat{\theta}_n \pm 2,6\ \sigma/\sqrt{n} \quad \text{avec probabilité de confiance 99\%.}}$$

Intervalle à 95%. De la même façon, on a $P(|\mathcal{N}(0,1)| > 1,96) \simeq 0,05)$ et

$$\boxed{\theta^* = \hat{\theta}_n \pm 1,96\ \sigma/\sqrt{n} \quad \text{avec probabilité de confiance 95\%.}}$$

On a la définition plus générale :

DÉFINITION VII.4 *Un intervalle de confiance pour θ^* de probabilité de confiance $1-\alpha$ est un intervalle aléatoire tel que*

$$\theta^* \in I \quad \textit{avec probabilité } 1-\alpha.$$

α est appelé le niveau.

Un peu de philosophie. Noter que l'équation (VII.1) conduit plus naturellement à

$$\hat{\theta}_n = \theta^* \pm 1,96\ \sigma/\sqrt{n} \quad \text{avec probabilité } 95\%$$

qui est un intervalle déterministe pour $\hat{\theta}_n$. C'est le point de vue des probabilités. Le renversement de cette équation correspond au passage des probabilités aux statistiques expliqué au chapitre 1.

VII.4.2 Intervalles exacts et intervalles approchés.

Les intervalles de confiances présentés au début du paragraphe précédent ne sont pas exacts car $\hat{\theta}_n - \theta^*$ ne suit pas rigoureusement une loi normale ; ils ont donc un niveau α_n différent de α ; on peut cependant dire que α_n tend vers α.

Malheureusement, on ne connaît pas la valeur de n pour laquelle on peut considérer ces approximations comme raisonnablement valide. C'est pourquoi il est déraisonnable de considérer des intervalles asymptotiques à probabilité de confiance très élevée (disons 99%) pour des n petits (disons 10).

VII.4.3 Un exemple d'intervalle exact

Soient $Y_1, ...Y_n$ des v.a. gaussiennes de variance connue σ^2 et de moyenne inconnue μ^*. On sait que leur moyenne empirique $\hat{\mu}_n$ a pour loi $\mathcal{N}(\mu^*, \sigma^2/n)$. La variable $\sqrt{n}(\mu_n - \mu^*)/\sigma$ suit donc une loi $\mathcal{N}(0,1)$. Donc

$$-2,6 \leq \frac{\sqrt{n}}{\sigma}\ (\hat{\mu}_n - \mu^*) \leq 2,6 \quad \text{avec probabilité } 99\%$$

ce qui se réécrit

$$\mu^* = \hat{\mu}_n \pm 2,6\ \frac{\sigma}{\sqrt{n}} \quad \text{avec probabilité de confiance } 99\%.$$

VII.4.4 Exemples d'intervalles approchés

Estimation de la moyenne. On observe 10 malades traités par un nouveau médicament. Pour chacun des malades le temps de guérison a été en jours ;

$$T : \quad 12 \quad 16 \quad 21 \quad 10 \quad 13 \quad 16 \quad 25 \quad 8 \quad 13 \quad 15$$

On voudrait savoir le temps moyen de guérison. La moyenne empirique est $\bar{T} = 14,9$ et la variance empirique 23, soit un écart type d'environ 4,8. La variance de l'estimateur de la moyenne étant égale à la variance de la variable elle-même divisée par le nombre de points, on a

$$E[T] = 14,9 \pm 1,96 \,.\, \sigma/\sqrt{10} \quad \text{avec probabilité de confiance } 95\%$$

ce qui donne en remplaçant σ par son estimée $4,8$:

$$E[T] = 14,9 \pm 3 \quad \text{avec probabilité de confiance } 95\%.$$

Estimation d'une proportion. On fait un sondage pour savoir qui de A ou B va gagner les élections. On obtient 1154 pour A et 1301 pour B et l'on suppose que l'échantillon est représentatif. La proportion p d'électeurs allant voter pour A s'estime à $\hat{p} = 1154/2455 = 0,47$. La variable $\sqrt{n}(\hat{p} - p)$ est approximativement $\mathcal{N}(0, \sigma^2)$, avec $\sigma^2 = p(1-p) \simeq \hat{p}(1-\hat{p})$, d'où $\sigma/\sqrt{n} \simeq 0,01$. On a donc l'intervalle de confiance à 95% :

$$p = \hat{p} \pm 1,96 \,.\, \sqrt{\hat{p}(1-\hat{p})}/\sqrt{n} = 0,47 \pm 1,96 \,.\, 0,01 = 0,47 \pm 0,02 \quad \text{à } 95\%.$$

La victoire de B est quasi certaine (si tant est que l'échantillon est représentatif).

Le remplacement de σ par $\hat{\sigma}$ est valide car il n'introduit qu'une erreur du deuxième ordre (c-à-d petite devant la largeur de l'intervalle).

VII.5 Tests de significativité

VII.5.1 Introduction

Commençons par un exemple volontairement simpliste. On veut tester si l'état de santé de certains malades s'améliore significativement à la suite d'un certain traitement. Pour cela on dispose de mesures de l'état de santé de 10 malades avant et après traitement

A	12	16	21	10	13	16	25	8	13	15
B	14	17	22	13	14	18	24	7	12	17

Il s'agit de tester l'hypothèse H_0 :«la variable $B - A$ a une moyenne nulle» (pas d'effet) contre son contraire H_1 qui assure d'un effet significatif du médicament. Il est clair qu'une priorité du test est de ne pas conclure H_1 si H_0 est vraie, ce qui entraînerait la mise sur le marché d'un médicament inefficace. Notons la disysmétrie : la décision H_1 doit être convaincante car elle a des conséquences importantes.

Le test sera ici simplement une fonction des 20 variables aléatoires observées.

DÉFINITION VII.5 *Soient $Y_1, ...Y_n$ une suite de v.a. Un test est une statistique $\varphi(Y_1, ...Y_n)$ dont la valeur, 0 ou 1, décide entre deux hypothèses H_0 et H_1 portant sur la distribution de l'échantillon.*

En toute généralité, H_0 et H_1 correspondent donc à deux ensembles de distributions de probabilités disjoints ; par exemple (Y_i sont supposées i.i.d)

- $H_0 : Y_i \sim \mathcal{N}(0,1), \quad H_1 : Y_i \sim \mathcal{N}(2,1)$
- $H_0 : Y_i \sim \mathcal{N}(0,1), \quad H_1 : Y_i \sim \mathcal{N}(\mu,1), \quad \mu > 0$

Dans ces deux exemples, l'hypothèse H_0 est dite **simple** car elle détermine complètement la loi de Y, contrairement à l'hypothèse H_1 du deuxième exemple, qui est dite **composite**. Dans la suite, on s'intéressera essentiellement au cas où H_0 est simple.

On appelle **probabilité d'erreur de première espèce** ou **niveau du test**, la probabilité α de décider H_1 si H_0 est vraie, c'est-à-dire la valeur maximum de $E[\varphi]$ sous H_0 (ou la valeur tout court si H_0

est simple). Noter que le test qui décide toujours H_0 a un niveau faible mais ne présente aucun interêt : sa probabilité d'erreur de seconde espèce (probabilité de décider H_0 sous H_1) est égale à 1.

Le but des tests de significativité est de déterminer si un ensemble de données permet d'invalider une hypothèse H_0. Dans l'exemple précédent, le laboratoire pharmaceutique cherchera à prouver qu'on a observé un effet *significatif* du nouveau traitement sur les malades, au sens où il est statstiquement très peu probable que H_0 soit valide.

Pour être fiable, un tel test devra avoir une très faible probabilité de décider H_1 si H_0 est vraie. On veut donc un petit niveau. Pour être intéressant, il devra être construit de sorte à décider H_1 le plus souvent possible quand H_1 est vraie (ceci est lié à la probabilité d'erreur de seconde espèce ; on dit que le test doit êter puissant).

Fin de l'exemple. Notons m l'espérance de $B - A$, on a l'estimateur

$$\hat{m} = \frac{1}{10} \sum_{i=1}^{10} B_i - A_i = 0,9.$$

Pour φ, on prend la v.a. qui vaut 0 si 0 est dans l'intervalle de confiance à 95% pour m basé sur $\hat{m}$ et 1 sinon. Le niveau de ce test est de 5% (par définition de l'intervalle de confiance).

La variance empirique de $B - A$ est 1,89, on obtient donc un intervalle de confiance à 95%

$$m = 0,9 \pm 1,96 \sqrt{1,89/10} = 0,9 \pm 0,85$$

On peut donc décider d'un effet significatif sur cet ensemble de 10 patients, pour un niveau de 5%. On vérifie qu'il n'est toutefois pas significatif à 1%.

VII.5.2 Tests basés sur un estimateur et un intervalle de confiance

Soit θ^* un certain paramètre de la loi de Y. On veut tester $H_0 : \theta^* = \theta_0$ contre son contraire, c'est-à-dire voir si les données permettent d'affirmer si θ^* est sensiblement différent de θ_0.

Soit un estimateur $\hat{\theta}$ de θ^* et $[\hat{\theta} - \delta, \hat{\theta} + \delta]$ un intervalle de confiance de probabilité de confiance $1 - \alpha$ pour θ^* :

$$\theta^* \in [\hat{\theta} - \delta, \hat{\theta} + \delta] \quad \text{avec probabilité } 1 - \alpha.$$

Considérons le test : Refuser H_0 si $\theta_0 \notin [\hat{\theta} - \delta, \hat{\theta} + \delta]$.

On sait que si H_0 est vraie $\theta^* = \theta_0$, le test décidera par erreur H_1 avec une probabilité ne dépassant pas α. Ce test a donc un niveau de α.

Noter que ce test est en fait

Refuser $\theta^* = \theta_0$ si $\lvert\hat{\theta} - \theta_0\rvert > \delta$.

VII.5.3 Approche générale basée sur une statistique

On base en général les tests sur une statistique S que l'on juge pertinente pour distinguer au mieux les deux hypothèses (p. ex. $S = |\hat{\theta} - \theta_0|$) ; par exemple S est plutôt petite sur H_0 et grande sous H_1 ; puis on se donne un seuil λ définissant le test

$$\varphi = 1_{S > \lambda}.$$

ou encore

Refuser H_0 si $S > \lambda$.

On tente de régler le seuil assez grand de sorte que la probabilité d'erreur de première espèce α ne dépasse pas une valeur prescrite :

$$P(S > \lambda) = \alpha, \quad \text{sous } H_0.$$

Toute valeur λ telle que $P(S > \lambda) \leq \alpha$ conviendrait, mais on refuserait H_0 moins souvent, ce qui est contraire à l'esprit du test. Il faut interpréter la conclusion du test avec précaution :

- Si le test refuse H_0 (décide H_1), on peut dire que cette conclusion est fausse avec probabilité au plus α, par définition du niveau.
- Si le test décide H_0, on ne peut en général rien dire. Pour éviter de décider H_0 quand H_1 est vraie, il faut choisir une bonne statistique et avoir suffisament d'échantillons.

Calcul du seuil. Si H_0 est simple, on connait, au moins théoriquement, la valeur de λ telle que $P(S > \lambda) = \alpha$ sous H_0 ; cette valeur est la fonction quantile en $1 - \alpha$. Si le calcul est trop difficile, on peut très bien l'estimer par simulation, en tirant sous H_0, par exemple 100000 réalisations indépendantes de S, et en l'estimant par la 95000ième valeur obtenue (par ordre croissant).

Cas où H_1 est non-H_0. Par exemple si H_0 est «$\theta^* = 0$» et H_1 est «$\theta^* \neq 0$» on conclura de la façon suivante :
- si le test refuse H_0 (décide H_1) : «θ^* est significativement (au niveau α) différent de 0»
- si le test décide H_0 : «les données ne permettent pas de conclure que θ^* est significativement différent de 0». Ceci peut arriver simplement parce que l'on a trop peu de données, ou parce que $\theta^* = 0$.

VII.5.4 Test de nullité d'une moyenne. Test de Student

On reprend l'exemple du paragraphe VII.5.1. Il s'agit de tester si une suite d'observations $Y_1, ...Y_n$ est issue d'une distribution de moyenne nulle Soit $\hat{m}$ la moyenne empirique de l'échantillon :

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Supposons que Y_i a pour moyenne m et variance σ. On peut alors assimiler $\hat{m}$ à une variable aléatoire gaussienne de moyenne m et de variance σ^2/n. Ceci implique l'intervalle de confiance (asymptotique)

$$-1,96 \leq \frac{\hat{m} - m}{\sigma/\sqrt{n}} \leq 1,96 \quad \text{avec probabilité de confiance 95\%.}$$

Comme la variance est inconnue, on la remplace par son estimée empirique, et l'on obtient l'intervalle de confiance asymptotique :

$$m = \hat{m} \pm 1,96\frac{\hat{\sigma}}{\sqrt{n}} \quad \text{avec probabilité de confiance 95\%.}$$

D'où le test (de probabilité de confiance 95%) :

► Refuser la nullité de la moyenne si $\frac{\sqrt{n}|\hat{m}|}{\hat{\sigma}} > 1,96$.

On retrouve bien la forme annoncée au § VII.5.3.

Si les Y_i sont gaussiens $\mathcal{N}(0, \sigma^2)$ (hypothèse H_0 un peut particulière), la loi de la statistique $\frac{\sqrt{n}|\hat{m}|}{\hat{\sigma}}$ est bien entendu indépendante de σ ; il se trouve que c'est une loi de Student à $n - 1$ degrés de libertés (elle dépend de n) ; on préfère souvent utiliser dans le test le quantile correspondant de cette loi plutôt que celui de la gaussienne (l'écart est souvent faible, par exemple si $n = 30$ 1,96 devient 2,04).

VII.5.5 Test d'identité de deux moyennes

On se donne deux échantillons de population d'origine différente et l'on voudrait décider si leur espérance de vie est différente. Soient $\hat{m}_1$ et $\hat{m}_2$ l'espérance de vie moyenne (empirique) dans chaque population :

$$\hat{m}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i$$

$$\hat{m}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} Z_i$$

où Y_i et Z_i sont les durées de vie des individus sélectionnés dans chaque population. On suppose que Y_i et Z_i ont pour moyenne m_1 et m_2 et pour variance σ_1 et σ_2. On peut alors assimiler $\hat{m}_1$ et $\hat{m}_2$ à deux variables aléatoires gaussiennes indépendantes de moyenne m_1 et m_2 et de variance σ_1^2/n_1 et σ_2^2/n_2. Sous cette approximation, la variable $\hat{m}_1 - \hat{m}_2$ a pour moyenne $m_1 - m_2$ et pour variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$. Ceci implique l'intervalle de confiance (asymptotique)

$$-1,96 \leq \frac{\hat{m}_1 - \hat{m}_2 - m_1 + m_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \leq 1,96 \quad \text{avec probabilité de confiance 95\%.}$$

Comme les variances sont inconnues, on les remplace par leurs estimées empiriques, et l'on obtient l'intervalle de confiance asymptotique :

$$m_1 - m_2 = \hat{m}_1 - \hat{m}_2 \pm 1,96\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} \quad \text{avec probabilité de confiance 95\%.}$$

D'où le test (de probabilité de confiance 95%) qui décide de la différence des moyennes si zéro sort de l'intervalle de confiance :

- Refuser l'égalité des moyennes si $\frac{|\hat{m}_1 - \hat{m}_2|}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} > 1,96.$

On retrouve encore la forme annoncée au § VII.5.3. Dans le cas $n_1 = n_2 = n$ et $\sigma_1 = \sigma_2 = \sigma$, la statistique de test devient $\sqrt{n/2}|\hat{m}_1 - \hat{m}_2|/\hat{\sigma}$; son interprétation est simple puisque c'est l'écart des moyennes empiriques normalisé par l'écart-type empirique, et par le facteur $\sqrt{n}$ du théorème-limite central.

VII.5.6 Test de comparaison de proportions

On veut tester l'efficacité d'un vaccin. Pour cela on se propose d'estimer si la probabilité d'attraper la maladie est inférieure si l'on a pris le vaccin. On considère deux populations, une non-vaccinée et une vaccinée. On est dans la situation du paragraphe précédent sauf que cette fois-ci Y_i est la variable aléatoire qui vaut 0 si l'on a pas été atteint et 1 sinon ; Z_i est la variable analogue observée sur la population vaccinée ; $m_i = p_i$ est la probabilité d'attraper la maladie et $\sigma_i^2 = p_i(1 - p_i)$; on a donc le test à 95%

- Refuser l'identité des lois si $\frac{|\hat{p}_1 - \hat{p}_2|}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} > 1,96.$

Exemple : passagers du Titanic. On compare la probabilité d'être survivant entre les trois classes à partir des données suivantes :

	1-ière	2-ième	3-ième	Total
Survivants	193	119	138	450
Morts	129	161	573	863
Total	322	280	711	1313

On trouve les probabilités empiriques de survie $\hat{p}_1 = 0,6$, $\hat{p}_2 = 0,425$ et $\hat{p}_3 = 0,2$. Les trois tests d'identités de loi ont pour statistique : $S_{12} = 4,35$, $S_{13} = 12,8$, et $S_{23} = 6,8$. Il y a bien une différence très significative.

Le risque relatif (RR) est le rapport p_1/p_2 des risques dans les deux populations. Le théorème limite central permet d'obtenir l'intervalle de confiance asymptotique à 95% suivant pour son logarithme

$$\log \frac{p_1}{p_2} = \log \frac{\hat{p}_1}{\hat{p}_2} \pm 1.96 \sqrt{\frac{1 - \hat{p}_1}{\hat{p}_1 n_1} + \frac{1 - \hat{p}_2}{\hat{p}_2 n_2}}$$

L'odds ratio (OR) est le rapport $OR = p_1(1-p_2)/(1-p_1)p_2$. Le théorème limite central permet d'obtenir l'intervalle de confiance asymptotique à 95% suivant pour son logarithme

$$\log(OR) = \log\left(\frac{\hat{p}_1(1-\hat{p}_2)}{(1-\hat{p}_1)\hat{p}_2}\right) \pm 1.96\sqrt{\frac{1}{\hat{p}_1 n_1} + \frac{1}{(1-\hat{p}_1)n_1} + \frac{1}{\hat{p}_2 n_2} + \frac{1}{(1-\hat{p}_2)n_2}}$$

Parenthèse : OR et RR en biostatistiques. De manière générale l'OR est souvent préféré pour les raisons suivantes :

- Si l'on remplace l'évènement «survie» par l'évènement «décès» pour le calcul du RR, on obtient $\frac{1-p_1}{1-p_2}$ qui n'est pas fonction du RR de départ, tandis que l'OR est simplement remplacé par son inverse. Il y a donc en fait deux RR mais un seul OR.
- Lors des «études de cas témoins» («case-control studies») on tire d'abord au hasard un nombre équivalent de personnes guéries (ayant survécu...) et d'autres malades (décédées...) afin d'avoir suffisament d'individus dans les deux situations et ensuite on sépare chaque groupe en deux (traitement/non-traitement, classe1/classe2...). L'exemple suivant [1] concerne les accidents veineux thrombo-emboliques en Europe selon l'utilisation ou non de contraceptifs oraux où l'on a tiré au hasard 433 personnes ayant eu un accident veineux et 1044 n'en ayant pas eu

	Contraceptifs	Pas de contraceptifs	Total
Cas d'accident	265	168	433
Contrôles	356	688	1044
Total	621	856	1477

Cette proportion de 433/1044 ne reflète ici aucune la réalité ; on ne peut pas estimer p_1, qui n'a rien à voir avec 265/621, et pas davantage RR. En revanche 265/433 est bien une estimation de la probabilité d'utiliser un contraceptif sachant que l'on a eu un accident veineux, et de même pour les trois autres rapports analogues ; par conséquent si l'on remarque que par la formule de Bayes (A=accident, C=contraceptif, $\bar{A}$=non-A)

$$OR = \frac{P(A|C)P(\bar{A}|\bar{C})}{P(\bar{A}|C)P(A|\bar{C})} = \frac{P(A,C)P(\bar{A},\bar{C})}{P(\bar{A},C)P(A,\bar{C})} = \frac{P(C|A)P(\bar{C}|\bar{A})}{P(C|\bar{A})P(\bar{C}|A)}$$

l'OR est correctement estimé par $265 \times 688/(356 \times 168) \simeq 3$.

VII.5.7 Test de corrélations

La loi asymptotique montrée plus haut dans le cas (hypothèse H_0) où les deux variables sont indépendantes conduit au test à 95%

- Refuser l'indépendance si $\sqrt{n}\,|\hat{r}_n| > 1,96$

Application : Test de dépendance de deux Bernoullis.

On voudrait savoir s'il existe chez les couples une corrélation entre le fait de posséder un animal domestique et ne pas avoir d'enfant. On mesure les deux variables U_i, qui vaut 1 si le couple numéro i a un animal domestique, et V_i qui vaut 1 si le couple numéro i a au moins un enfant. On a ici

$$\hat{r}_n = \frac{\hat{p}_{ae} - \hat{p}_a\hat{p}_e}{\sqrt{(1-\hat{p}_a)\hat{p}_a(1-\hat{p}_e)\hat{p}_e}}$$

où $\hat{p}_a$ (resp. $\hat{p}_e$, $\hat{p}_{ae}$) est la proportion de couples ayant un animal (resp. un enfant, un animal et un enfant).

VII.5.8 Un exemple

Voici le début du commentaire du docteur Serge Hercberg publié dans le Quotidien du Médecin (22 juin 2003) comcernant l'étude Suvimax effectuée sur un échantillon de 13017 personnes (7876 femmes et 5141 hommes) visant à évaluer l'importance de la consommation d'antioxydants (fruits et légumes) sur

1. Table 3 de l'article : "Venous thromboembolic disease and combined oral contraceptives", *The Lancet*, pp. 1575-1582, 1995

les risques de cancer. Cette étude a duré 8 ans pendant lesquels 6481 ont reçu des vitamines et minéraux antioxydant tandis que 6536 ont reçu un placebo.

« Les résultats sont très significatifs. Ils montrent nettement que l'apport de vitamines et de minéraux antioxydants à doses nutritionnelles réduit le risque de cancers ainsi que la mortalité globale chez les hommes. Cette baisse du taux de cancers de 31 % est très importante puisque près d'un cancer sur trois est évité en moins de huit ans (124 dans le groupe placebo contre 88 dans le groupe antioxydants ; RR=0,69, IC 95%=0,53-0,91 ; p<0,008). La différence entre les deux groupes est retrouvée pour la plupart des localisations de cancers, principalement digestifs, ORL, respiratoires et cutanés. La randomisation permet d'affirmer que la réduction observée a bien été causée par les antioxydants. Le nombre de décès chez les hommes était moindre dans le groupe antioxydants (40) que dans le groupe placebo (63) (p<0,02). En revanche, cet effet n'a pas été retrouvé chez les femmes (171 cancers dans le groupe placebo et 179 dans le groupe antioxydants ; 35 décès dans le groupe placebo, 36 dans le groupe antioxydants).

Comment expliquer l'absence d'effet chez la femme ?

Très probablement par un meilleur état du statut nutritionnel en antioxydants des femmes (bêta-carotène et vitamine C). En effet, les hommes avaient au départ de l'étude des taux sanguins de bêta-carotène plus bas que les femmes. Celles-ci consomment davantage de fruits et légumes. Or les niveaux sanguins de bêta-carotène sont corrélés positivement avec la consommation de fruits et légumes (r = 0,20 ; p<0,001). Autrement dit, les petits consommateurs de fruits et légumes ont les niveaux sanguins les plus faibles et réciproquement. Les femmes de l'étude Suvimax, n'étant pas carencées au départ, n'ont pas eu de bénéfice à être supplémentées.

A-t-on trouvé un bénéfice sur le plan cardio-vasculaire ?

Non. Nous avons comptabilisé 134 cardiopathies ischémiques dans le groupe antioxydants et 137 dans le groupe placebo. Il n'y avait donc pas de différence entre les deux groupes. [...]»

La différence de risque entre les deux populations (placebo et non-placebo) est jugée significative. La mention p<0,008 signifie qu'il faudrait faire un test de niveau inférieur à 0,8% pour ne plus être significatif. RR est le risque relatif. On est donc certain à 95% d'une diminution du risque comprise en 9% et 47%.

De même r=0,2 est l'estimée de la corrélation, et p<0,001 signifie qu'un test à 0,1% refuse H_0 :«r=0» (on trouve en fait un niveau limite bien plus petit).

Bien que certains sujets aient abandonnés ou aient été perdus de vue ce qui modifie les chiffres, on retrouve bien en gros les valeurs numériques annoncées pour les niveaux limites.

VII.5.9 Test du χ^2 : adéquation à une distribution discrète, comparaison

On veut savoir si un dé est pipé ou non. Il s'agit de voir si une variable discrète suit bien une distribution prescrite, ici la distribution uniforme $p_1 = p_2 = ... = p_6 = 1/6$. On fait un certain nombre n de tirages et l'on construit la statistique

$$S = n \sum_{i=1}^{k} \frac{(\hat{p}_i - p_i)^2}{p_i} \tag{VII.2}$$

où $\hat{p}_i$ et la fréquence d'apparition de la i-ième modalité et k le nombre de modalités (6 dans notre exemple), et p_i sa probabilité d'apparition sous H_0.

On peut montrer que sous H_0 :« les p_i correspondent bien à la distribution des données», la loi de S est (asymptotiquement pour n grand) celle d'un χ^2_{k-1}, c'est-à-dire la somme des carrés de $k-1$ variables normales standard indépendantes. On a donc le test (asymptotique) de niveau α

► Refuser la loi (p_i) si $S > Q_{k-1}(1-\alpha)$

où $Q_k(\alpha)$ est le quantile d'ordre α, c-à-d le nombre tel que $P(\chi^2_k > Q_k(1-\alpha)) = \alpha$. Ces quantités sont facilement disponibles sur ordinateur, et l'on a par exemple pour des seuil à 1% et 5% :

$\alpha \backslash k$	1	2	3	4	5
5%	3,8	6	7,8	9,5	11
1%	6,6	9,2	11,3	13,3	15

Valeur de $Q_k(\alpha)$ pour deux α et k=1,...5.

On peut très bien sinon estimer le seuil par simulation, comme expliqué au § VII.5.3, ce qui permet d'avoir un test exact (non-asymptotique) :

1. Tirer 100 000 (par ex.) échantillons (de taille n) sous H_0 (i.e. sous la loi $(p_1, ...p_k)$)
2. En déduire les 100 000 valeurs de S correspondantes et les ordonner
3. $Q(\alpha)$ est la p-ième valeur, avec $p = 100\,000(1-\alpha)$.

Un exemple un peu plus compliqué : le test de Hardy-Weinberg. Un allèle est une version d'un gène. Dans une cellule diploïde, il y a deux allèles pour chaque gène : un allèle transmis par chaque parent.

Soit A et a, deux allèles de fréquence respectivement p et $q = 1 - p$ dans la population. La loi de Hardy-Weinberg prévoit les fréquences suivantes pour les trois différents génotype :

- $p_0 = p^2$: la fréquence d'un génotype homozygote AA
- $p_1 = 2pq$: la fréquence d'un génotype hétérozygote Aa
- $p_2 = q^2$: la fréquence d'un génotype homozygote aa

Cette loi se base sur le modèle le plus simple de reproduction dans la population et est expérimentalement vérifiée. Pour voir si un gène intervient dans une certaine maladie, et plus précisément la présence d'un allèle particulier, une méthode consiste à sélectionner un certain nombre de malades et à regarder si la distribution du génotype satisfait la loi (hypothèse H_0) ou non (hypothèse H_1). Comme p n'est pas connu (à moins de l'estimer par une expérience antérieure), le test du χ^2 ne peut être appliqué tel quel.

Si $\hat{p}_0, \hat{p}_1, \hat{p}_2$ sont les proportions d'individus observés dans la population malade pour les génotypes AA, Aa, aa, on peut estimer p sous H_0 par

$$\hat{p} = \hat{p}_0 + \hat{p}_1/2, \quad \hat{q} = 1 - \hat{p}$$

et la statistique de test de Hardy-Weinberg est définie par

$$S_{HW} = n\left(\frac{(\hat{p}_0 - \hat{p}^2)^2}{\hat{p}^2} + \frac{(\hat{p}_1 - 2\hat{p}\hat{q})^2}{2\hat{p}\hat{q}} + \frac{(\hat{p}_2 - \hat{q}^2)^2}{\hat{q}^2}\right).$$

On voit qu'elle s'inspire de la statistique du χ^2 (VII.2). On peut montrer que sous H_0 cette statistique suit asymptotiquement un χ^2_1 ce qui permet de réaliser des tests.

Comparaison de deux échantillons. Pour décider si deux échantillons ont même loi (p.ex. taux de réussite au bac dans deux lycées différents, p_i^j étant la probabilité pour un lycéen du lycée j d'être reçu avec mention j), on peut utiliser la statistique suivante

$$S' = S(n_1, \hat{p}^1, \hat{p}^{12}) + S(n_2, \hat{p}^2, \hat{p}^{12})$$

où $S(n, \hat{p}, p)$ désigne la statistique (VII.2), $\hat{p}^1$ (resp. $\hat{p}^2$, $\hat{p}^{12}$) est le vecteur de probabilité estimée sur la base du premier échantillon (resp. du deuxième, des deux) et n_i la taille de l'échantillon i. On comparera cette statistique à un χ^2_{k-1} :

► Refuser l'identité des lois si $S' > Q_{k-1}(1-\alpha)$

VII.5.10 Tests d'indépendance

La statistique de Pearson peut être utiilsée pour tester l'indépendance. Si les deux caractères sont indépendants, la loi de Φ^2 est asymptotiquement un $\chi^2_{IJ-I-I+1}$.

Dans le cas des mariés de l'Alaska, les paires (x_k, y_k) sont la classe d'âge de chaque époux et l'on a $I = J = 9$. On trouve $\Phi^2 = 7195$, qui est bien plus grand que 64. Le test conclue à une corrélation significative, même pour des très petits niveaux.

VII.5.11 Tests de Kolmogorov et Smirnov : adéquation à une distribution continue, comparaison

Test d'une loi. On veut tester si les échantillons sont tirés selon une loi de fonction de répartition continue donnée $F(x)$ (hypothèse H_0). La statistique de test est

$$S_n = \sqrt{n} \sup_x |F_n(x) - F(x)|.$$

Au vu des résultats du § VII.3.2, la loi de S_n sous H_0 est indépendante de F. Définissons $q_n(\alpha)$ comme le quantile de cette loi connue :

$$P(S_n > q_n(1-\alpha)) = \alpha \quad \text{(sous } H_0\text{)}.$$

On a par exemple $q_{20}(5\%) = 1,57$ pour $n = 20$ (ces quantités sont tabulées). On a alors, par exemple, le test de probabilité de confiance 95% pour $n = 20$:

► La loi est significativement différente de F si $S_{20} > 1,57$.

Le niveau de ce test est la probabilité d'observer, sous H_0, que $\sqrt{n}\sup_x |F_n(x) - F(x)| > 1,57$; c'est 5% en raison du choix du seuil.

Comparaison de deux échantillons. On utilise le même principe pour comparer deux échantillons $Y_1, ...Y_{n_1}$ et $Z_1, ...Z_{n_2}$ de fonction de répartition F et G continues. On s'intéresse à l'hypothèse H_0 :«$F = G$». Soient F_{n_1} et G_{n_2} les fonctions de répartition empirique des deux échantillons. La statistique est cette fois :

$$S = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \sup_x |F_{n_1}(x) - G_{n_2}(x)|$$

où n_1 et n_2 sont les longueurs respectives des échantillons. En raisonnant sur le même type de principe, on a le test, ici à 95% :

► Les lois sont significativement différentes si $S > 1,36$

(1,36 est ici le seuil asymptotique pour n grand). Noter qu'on peut montrer comme au § VII.3.2 que la loi de S sous H_0 est également celle de

$$\sup_y |\frac{1}{n_1}\sum_{k=1}^{n_1} 1_{U_k \leq y} - \frac{1}{n_2}\sum_{k=1}^{n_2} 1_{V_k \leq y}|$$

où les U_k et les V_k sont toutes i.i.d $\mathcal{U}([0,1])$. On peut aisément estimer le quantile d'ordre α de cette variable par simulation.

Bibliographie

[1] J.-F. Delmas, *Introduction aux probabilités et à la statistique*, Presses de l'ENSTA, 2012.

[2] D. Foata et A. Fuchs, *Calcul des probabilités*, Dunod, 2012.

[3] P. Barbé et M. Ledoux, *Probabilité*, EDS Sciences, 2000.

[4] M. Cottrell et V. Genon-Catalot, *Exercices de probabilités*, Cassini, 2000.

[5] J.-F. Lecoutre, *Statistique et probabilités*, Dunod, 2003.

Cette bibliographie propose des livres dont l'objectif est proche de celui de ce cours. Il en existe de nombreux, ayant tous un contenu très analogue. Les livres choisis ici ont simplement l'avantage d'être disponibles à la bibliothèque universitaire de Rennes. Le livre de J.-F. Lecoutre est probablement le plus proche du cours, les autres sont plus avancés