Statistique Descriptive

Mesure ponctuelle associée à une série statistique

Les observations

Notons $x_1, x_2, \ldots, x_n$, la *série statistique observée* vue comme une suite numérique de valeurs d'un certaine variable quantitative x. Il y a donc ici n observations et l'indice i de x_i ($1 \leqslant i \leqslant n$) permet d'identifier l'observation (par l'ordre de relevé pour des observations chronologiques, par l'identité de l'individu sur lequel on a mesuré la variable x, etc.). Les n valeurs observées n'étant pas toutes forcément distinctes, on peut les regrouper en valeurs distinctes que l'on ordonne : $x_{(1)} < x_{(2)} < \cdots < x_{(j_{\max})}$ en associant à chaque valeur $x_{(j)}$ pour $1 \leqslant j \leqslant j_{\max}$, l'effectif n_j égal au nombre d'apparitions de cette valeur dans la série initiale des x_i. La fréquence de la valeur $x_{(j)}$ est $f_j = n_j/n$ et la somme des fréquences de $j = 1$ à $j_{\max}$ vaut 1. Toute l'information sur la série ordonnée peut alors être présentée sous la forme d'un tableau des $(x_{(j)}, f_j)_{j\in J}$, où l'on note

$$J = [\![1, j_{\max}]\!].$$

Mesure associée aux observations

Pour faire le lien avec la théorie des probabilités, il est commode de représenter cette information par l'objet mathématique suivant :

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} = \sum_{j\in J} f_j\, \delta_{x_{(j)}}.$$

Il s'agit d'une mesure (de probabilité) ponctuelle à support fini. La plus simple des mesures de ce type est la masse de Dirac δ_x où x est un réel, définie par :

$$\forall A \subset \mathbb{R}, \quad \delta_x(A) = \mathbf{1}_A(x) = \begin{cases} 1 & \text{si } x \in A, \\ 0 & \text{si } x \notin A. \end{cases}$$

La mesure μ est une combinaison linéaire à coefficients positifs de masses de Dirac et la mesure $\mu(A)$ d'une partie A de $\mathbb{R}$ est donc

$$\mu(A) = \sum_{j\in J} f_j\, \delta_{x_{(j)}}(A) = \sum_{j\in J} f_j\, \mathbf{1}_A(x_{(j)}).$$

Dans cette somme, les seuls termes non nuls sont ceux dont l'indice j vérifie $x_{(j)} \in A$. Ainsi $\mu(A)$ est simplement la fréquence des observations qui sont dans l'ensemble A.

Une façon imagée de se représenter la mesure μ est de voir $\mathbb{R}$ comme un fil de pêche tendu et de masse nulle sur lequel sont disposés $j_{\max}$ plombs (masses ponctuelles) de lestage, le j^e plomb de masse f_j étant localisé au point d'abscisse $x_{(j)}$. La quantité $\mu(A)$ est alors la masse totale de plomb contenue dans A.

Intégration par rapport à une mesure ponctuelle

À partir de la mesure ponctuelle définie ci-dessus, on peut reconstruire tous les objets mathématiques associés à la série statistique. Mais il convient d'abord de définir l'intégrale par rapport à cette mesure.

Cadre général

L'ordonnancement des x_i ne jouant aucun rôle dans cette notion d'intégrale, il est commode de considérer plus généralement les mesures ponctuelles à support fini

$$\mu = \sum_{k \in K} m_k \delta_{t_k}$$

où K est un ensemble fini et tous les m_k sont des réels positifs (pas forcément de somme 1). Si tous les réels t_k sont distincts et si chaque m_k est strictement positif, on parle de représentation canonique de μ. On vérifiera que ni la définition de l'intégrale ni ses propriétés exposées ci-dessous ne dépendent du choix de la représentation de μ qui n'est évidemment pas unique.

Pour toute fonction $g : \mathbb{R} \to \mathbb{R}$, on définit son intégrale par rapport à μ par :

$$\int_{\mathbb{R}} g \, \mathrm{d}\mu = \int_{\mathbb{R}} g(t) \, \mathrm{d}\mu(t) = \sum_{k \in K} m_k g(t_k).$$

Notons qu'en particulier : $\int_{\mathbb{R}} g \, \mathrm{d}\delta_a = g(a)$. L'intégrale ainsi définie a les propriétés immédiates suivantes.

1. Si g est constante ($g(t) = c$ pour tout $t \in \mathbb{R}$), $\int_{\mathbb{R}} c \, \mathrm{d}\mu = c\mu(\mathbb{R})$.
2. Intégrale d'une indicatrice : pour toute partie A de $\mathbb{R}$, $\int_{\mathbb{R}} \mathbf{1}_A \, \mathrm{d}\mu = \mu(A)$.
3. Linéarité par rapport à l'intégrande : pour tous a, b réels et toutes fonctions g, h, $\int_{\mathbb{R}} (ag + bh) \, \mathrm{d}\mu = a \int_{\mathbb{R}} g \, \mathrm{d}\mu + b \int_{\mathbb{R}} h \, \mathrm{d}\mu$.
4. Linéarité par rapport à la mesure : pour tous a, b réels positifs [1], toutes mesures ponctuelles à support fini μ et ν et toute fonction $g : \mathbb{R} \to \mathbb{R}$,
$$\int_{\mathbb{R}} g \, \mathrm{d}(a\mu + b\nu) = a \int_{\mathbb{R}} g \, \mathrm{d}\mu + b \int_{\mathbb{R}} g \, \mathrm{d}\nu.$$
5. Croissance : si $g \leqslant h$, $\int_{\mathbb{R}} g \, \mathrm{d}\mu \leqslant \int_{\mathbb{R}} h \, \mathrm{d}\mu$ et en conséquence, $\left|\int_{\mathbb{R}} g \, \mathrm{d}\mu\right| \leqslant \int_{\mathbb{R}} |g| \, \mathrm{d}\mu$.

1. Cette restriction n'est là que pour rester dans le cadre des mesures positives.

Intégrale par rapport à la mesure associée aux observations

Revenons maintenant et jusqu'à la fin de cette partie au cas particulier où μ est la mesure de probabilité $\sum_{j\in J} f_j \delta_{x_{(j)}}$ associée à la série statistique $x_1, x_2, \ldots, x_n$. Certains des paramètres importants de cette série peuvent s'exprimer comme intégrales par rapport à μ.

Moyenne arithmétique

Si g est l'identité sur $\mathbb{R} : t \mapsto t$, l'intégrale $\int_{\mathbb{R}} g \, \mathrm{d}\mu$ s'écrit

$$\int_{\mathbb{R}} t \, \mathrm{d}\mu(t) = \sum_{j\in J} f_j x_{(j)} = \frac{1}{n} \sum_{i=1}^{n} x_i = \overline{x}.$$

La *moyenne arithmétique* de la série statistique apparaît ainsi comme l'intégrale par rapport à μ de la fonction identité.

Écart moyen absolu

En choisissant $g : t \mapsto |t - \overline{x}|$, on obtient

$$\int_{\mathbb{R}} |t - \overline{x}| \, \mathrm{d}\mu(t) = \sum_{j\in J} f_j \left| x_{(j)} - \overline{x} \right| = \frac{1}{n} \sum_{i=1}^{n} \left| x_i - \overline{x} \right|.$$

On trouve ainsi l'*écart moyen absolu* de la série statistique.

Variance et écart-type

En choisissant $g : t \mapsto (t - \overline{x})^2$, on obtient

$$\int_{\mathbb{R}} (t - \overline{x})^2 \, \mathrm{d}\mu(t) = \sum_{j\in J} f_j (x_{(j)} - \overline{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

On trouve ainsi la *variance* V_n de la série. Le formalisme de l'intégration par rapport à μ permet d'ailleurs de retrouver facilement la formule de Koenig-Huygens. En effet,

$$V_n = \int_{\mathbb{R}} (t - \overline{x})^2 \, \mathrm{d}\mu(t) = \int_{\mathbb{R}} (t^2 - 2\overline{x}t + \overline{x}^2) \, \mathrm{d}\mu(t).$$

En utilisant la linéarité de l'intégrale (propriété 3) et la propriété 1, on obtient :

$$\begin{aligned} V_n = \int_{\mathbb{R}} t^2 \, \mathrm{d}\mu(t) - 2\overline{x} \int_{\mathbb{R}} t \, \mathrm{d}\mu(t) + \overline{x}^2 \int_{\mathbb{R}} 1 \, \mathrm{d}\mu(t) &= \int_{\mathbb{R}} t^2 \, \mathrm{d}\mu(t) - \overline{x}^2 \\ &= \sum_{j\in J} f_j x_{(j)}^2 - \overline{x}^2 \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2. \end{aligned}$$

La variance V_n donne une idée de la dispersion des observations autour de la moyenne. Si ces observations sont des grandeurs physiques exprimées dans une certaine unité (par exemple des longueurs en cm), la variance s'exprime dans le carré de cette unité (pour notre exemple en cm^2). Pour avoir à partir de V_n un paramètre de dispersion exprimé dans les mêmes unités que les observations, il convient donc de prendre la racine carrée de V_n, ce qui nous donne *l'écart-type* σ_n.

$$\sigma_n = \left(\int_{\mathbb{R}} (t - \overline{x})^2 \, \mathrm{d}\mu(t) \right)^{1/2} = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \right)^{1/2} .$$

Les « connaisseurs » reconnaîtront en σ_n la norme $L^2(\mu)$ de la fonction g.

Attention, on trouve aussi en statistique la notation σ_{n-1} associée à un échantillon de n observations (et non pas $n-1$ observations comme on pourrait le croire si les notations statistiques étaient cohérentes). C'est une autre quantité que σ_n dont nous parlerons ultérieurement.

Fonction de répartition et médiane(s)

Courbe des fréquences cumulées croissantes

La série statistique $x_1, x_2, \ldots, x_n$ peut être représentée graphiquement par sa courbe des fréquences cumulées croissantes qui contient toute l'information relative à la série ordonnée. Cette courbe en escalier est la représentation graphique de la fonction de répartition F de la mesure μ, définie par

$$F : \mathbb{R} \to \mathbb{R}, \quad t \mapsto \mu(]-\infty, t]).$$

La fonction $G = 1 - F : t \mapsto \mu(]t, +\infty[)$ est appelée fonction de survie de μ. Chacune de ces fonctions caractérise μ. La fonction F est une fonction en escaliers croissante, dont les sauts ont pour abscisses les $x_{(j)}$ et pour amplitude les f_j. Elle est continue à droite en chacun de ces sauts (et continue partout ailleurs). Elle vaut 1 à droite de $x_{(j_{\max})}$ et 0 à gauche de $x_{(1)}$. Notons aussi que la limite à gauche de F en un point quelconque t est égale à $\mu(]-\infty, t[)$.

Médiane(s) et intervalle médian

On appelle *médiane* de la série $x_1, \ldots, x_n$ ou de la mesure μ associée, tout réel m tel qu'au moins la moitié des valeurs x_i soient inférieures ou égales à m et au moins la moitié des valeurs x_i soient supérieures ou égales à m. Autrement dit, m est une médiane si et seulement si $\mu(]-\infty, m]) \geqslant 1/2$ et $\mu([m, +\infty[) \geqslant 1/2$. L'ensemble des réels m satisfaisant cette propriété est un intervalle appelé *intervalle médian*. Cet intervalle peut être réduit à un point, dans ce cas il est légitime de parler de *la* médiane, alors qu'en général il convient plutôt de dire *une* médiane. La représentation graphique de F permet de visualiser facilement ces deux cas,

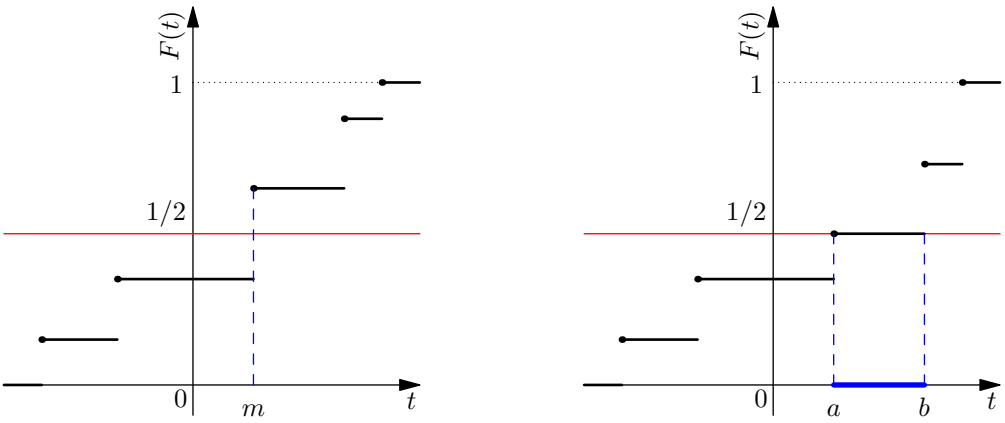

FIGURE 1 – Médianes et fonction de répartition, les deux cas possibles

cf. figure 1. On trace la droite horizontale d'équation $y = 1/2$. Soit cette droite n'intersecte pas la courbe représentative de F. Soit son intersection avec la courbe représentative de F est une marche d'escalier. Dans le premier cas la droite traverse un unique saut de F dont l'abscisse est *la* médiane de la série. Dans le deuxième cas, la projection de la marche d'ordonnée $1/2$ sur l'axe des abscisses, complétée par sa borne droite est l'intervalle médian $[a, b]$ (en bleu sur la figure).

Fonction de répartition et moyenne

Moyenne arithmétique

On peut interpréter graphiquement la moyenne $\overline{x}$ à l'aide des fonctions F et G. Cette interprétation correspond en un certain sens à une formule « d'intégration par parties » dont la généralisation ultérieure nous permettra de calculer l'espérance d'une variable aléatoire, lorsqu'elle existe, à partir de la seule connaissance de sa fonction de répartition. Voici comment procéder. On part de la formule

$$\overline{x} = \sum_{j \in J} f_j x_{(j)}$$

que l'on découpe suivant le schéma $\sum_{j\in J} = \sum_{i\in J^-} + \sum_{k\in J^+}$, où

$$J^- = \{i \in J \ ; \ x_{(i)} < 0\} \quad \text{et} \quad J^+ = \{k \in J \ ; \ x_{(k)} > 0\},$$

en remarquant que si l'un des $x_{(j)}$ vaut 0, sa contribution au calcul de $\sum_{j\in J}$ est nulle. Considérons maintenant la représentation graphique de F dans un repère orthonormé. Pour chaque $j \in J$, définissons le rectangle R_j obtenu en traçant les deux lignes horizontales joignant l'axe des ordonnées aux extrémités du segment vertical de saut de F en $x_{(j)}$. Si $k \in J^+$, $x_{(k)}$ est positif et $\text{aire}(R_k) = x_{(k)} f_k$ puisque f_k est l'amplitude du saut. Si $i \in J^-$, $x_{(i)}$ est négatif et $\text{aire}(R_i) = -x_{(i)} f_i$, cf. figure 2. Il ne nous reste plus qu'à remarquer que $\sum_{k\in J^+} f_k x_{(k)}$ est l'aire du domaine

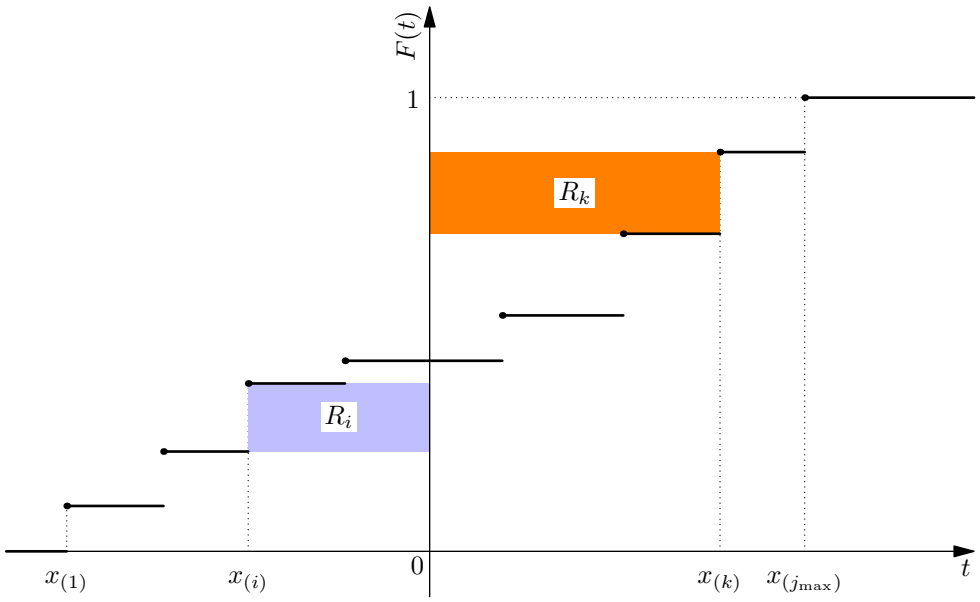

FIGURE 2 – Rectangles R_i, $i \in J^-$ et R_k, $k \in J^+$.

$R^+ = \bigcup_{k\in J^+} R_k$ délimité par l'axe des ordonnées, les marches d'escalier (à abscisses positives) et la droite horizontale d'équation $y = 1$, soit

$$\sum_{k\in J^+} f_k x_{(k)} = \int_0^{x_{(j_{\max})}} (1 - F(t))\,\mathrm{d}t = \int_0^{+\infty} (1 - F(t))\,\mathrm{d}t,$$

alors que $\sum_{i\in J^-} f_i x_{(i)}$ est l'opposé de l'aire du domaine $R^- = \bigcup_{i\in J^-} R_i$, délimité par l'axe des ordonnées, les marches d'escalier (à abscisses négatives) et l'axe des abscisses, soit

$$\sum_{i\in J^-} f_i x_{(i)} = -\int_{x_{(1)}}^{0} F(t)\,\mathrm{d}t = -\int_{-\infty}^{0} F(t)\,\mathrm{d}t.$$

Finalement nous obtenons la formule suivante :

$$\overline{x} = \int_0^{+\infty} (1 - F(t))\,\mathrm{d}t - \int_{-\infty}^{0} F(t)\,\mathrm{d}t = \int_0^{+\infty} G(t)\,\mathrm{d}t - \int_{-\infty}^{0} F(t)\,\mathrm{d}t,$$

qui peut se réécrire comme une formule d'« intégration par parties » pour $\int_{\mathbb{R}} t\,\mathrm{d}\mu(t)$:

$$\int_{\mathbb{R}} t\,\mathrm{d}\mu(t) = \int_0^{+\infty} \mu(]t, +\infty[)\,\mathrm{d}t - \int_{-\infty}^{0} \mu(]-\infty, t])\,\mathrm{d}t.$$

Graphiquement, $\overline{x}$ est la différence de l'aire coloriée en orange (domaine R^+) et de l'aire coloriée en bleu (domaine R^-) sur la figure 3.

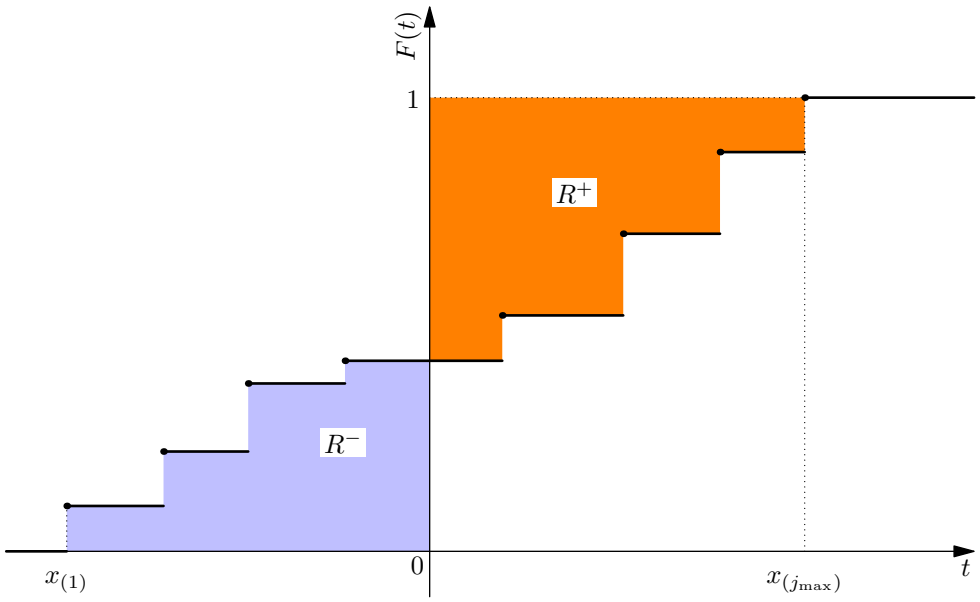

FIGURE 3 – Interprétation de $\overline{x}$ à l'aide de F : $\overline{x} = \text{aire}(R^+) - \text{aire}(R^-)$.

Moyenne des valeurs absolues

Comme sous-produit de l'étude précédente, nous obtenons immédiatement pour la moyenne des valeurs absolues de la série, l'interprétation graphique :

$$\frac{1}{n}\sum_{i=1}^{n}|x_i| = \sum_{j\in J} f_j \left|x_{(j)}\right| = \text{aire}(R^-) + \text{aire}(R^+),$$

ainsi que la formule intégrale

$$\frac{1}{n}\sum_{i=1}^{n}|x_i| = \sum_{j\in J} f_j \left|x_{(j)}\right| = \int_{-\infty}^{0} F(t)\,\mathrm{d}t + \int_{0}^{+\infty} (1-F(t))\,\mathrm{d}t.$$

Résumés d'une série statistique et minimisation

Quelle valeur numérique résume le mieux l'information contenue dans une série statistique ? Cette question n'est pas assez précise pour avoir une réponse unique. Tout dépend de la façon dont on mesure la perte d'information liée à ce « résumé ».

Choix d'une norme

Il est courant de mesurer cette perte à l'aide de distances associées à des normes. Les plus connues sont la norme L^1 et la norme L^2. Pour ces deux façons de mesurer la perte d'information, nous pouvons reformuler le problème en cherchant quelle est la constante c qui est la plus proche de la série statistique au sens L^1 ou au sens L^2.

On recherche donc quelles valeurs de c minimisent :

$$T_1(c) = \sum_{j\in J} f_j \left|x_{(j)} - c\right| = \int_{\mathbb{R}} |t-c| \, \mathrm{d}\mu(t),$$
$$T_2(c) = \sum_{j\in J} f_j (x_{(j)} - c)^2 = \int_{\mathbb{R}} (t-c)^2 \, \mathrm{d}\mu(t).$$

Ces deux problèmes pourraient se résoudre sans utiliser la mesure μ mais au prix d'écritures assez fastidieuses. La motivation essentielle pour l'utilisation de l'outillage introduit ci-dessus est la possibilité de généralisation ultérieure au cas des variables aléatoires.

Minimisation au sens de la norme 2

La minimisation de $T_2(c)$ est la plus facile et s'obtient en n'utilisant que les propriétés élémentaires de l'intégrale par rapport à μ. La solution est donnée par le calcul suivant :

$$\begin{aligned}\int_{\mathbb{R}} (t-c)^2 \, \mathrm{d}\mu(t) &= \int_{\mathbb{R}} \big((t-\overline{x}) + (\overline{x}-c)\big)^2 \, \mathrm{d}\mu(t) \\ &= \int_{\mathbb{R}} \big((t-\overline{x})^2 + 2(\overline{x}-c)(t-\overline{x}) + (\overline{x}-c)^2\big) \, \mathrm{d}\mu(t) \\ &= \int_{\mathbb{R}} (t-\overline{x})^2 \, \mathrm{d}\mu(t) + 2(\overline{x}-c)\int_{\mathbb{R}} (t-\overline{x}) \, \mathrm{d}\mu(t) + \int_{\mathbb{R}} (\overline{x}-c)^2 \, \mathrm{d}\mu(t).\end{aligned}$$

Pour la dernière égalité, nous avons utilisé la linéarité de l'intégrale par rapport à l'intégrande. La première intégrale ci-dessus n'est autre que la variance V_n de la série. Pour le deuxième intégrale, en utilisant la linéarité, l'intégrale de l'identité et l'intégrale d'une constante (avec ici $\mu(\mathbb{R}) = 1$) on obtient

$$\int_{\mathbb{R}} (t-\overline{x}) \, \mathrm{d}\mu(t) = \int_{\mathbb{R}} t \, \mathrm{d}\mu(t) - \overline{x}\int_{\mathbb{R}} \mathrm{d}\mu(t) = \overline{x} - \overline{x} = 0.$$

La troisième intégrale est celle d'une constante, cf. propriété 1, et comme $\mu(\mathbb{R}) = 1$,

$$\int_{\mathbb{R}} (\overline{x}-c)^2 \, \mathrm{d}\mu(t) = (\overline{x}-c)^2.$$

En résumé, nous avons montré que pour tout c réel,

$$T_2(c) = \int_{\mathbb{R}} (t-c)^2 \, \mathrm{d}\mu(t) = V_n + (\overline{x}-c)^2.$$

On en déduit que $T_2(c) \geqslant V_n$ pour tout c, avec égalité possible si et seulement si $(\overline{x}-c)^2 = 0$, c'est-à-dire $c = \overline{x}$.

En conclusion, $T_2(c) = \sum_{j\in J} f_j(x_{(j)} - c)^2$ a un unique minimum global atteint si et seulement si c est la moyenne arithmétique $\overline{x}$ de la série. On dit que $\overline{x}$ réalise la meilleure approximation de cette série au sens L^2 (ou des « moindres carrés »).

Dans ce contexte, l'écart-type $\sigma_n = \big(T_2(\overline{x})\big)^{1/2}$ s'interprète comme la distance, au sens L^2, entre la série statistique et sa meilleure approximation par une constante. Il est donc naturel de prendre σ_n comme *indicateur de dispersion* de la série.

Minimisation au sens de la norme 1

Pour résoudre la minimisation de $T_1(c)$, nous allons utiliser une interprétation graphique analogue à celle obtenue pour l'expression de $\overline{x}$ en fonction de F. Une simple adaptation de l'étude faite à cette occasion nous permet d'obtenir pour $\int_{\mathbb{R}} |t-c|\, \mathrm{d}\mu(t)$ la représentation graphique donnée par la figure 4. La formule in-

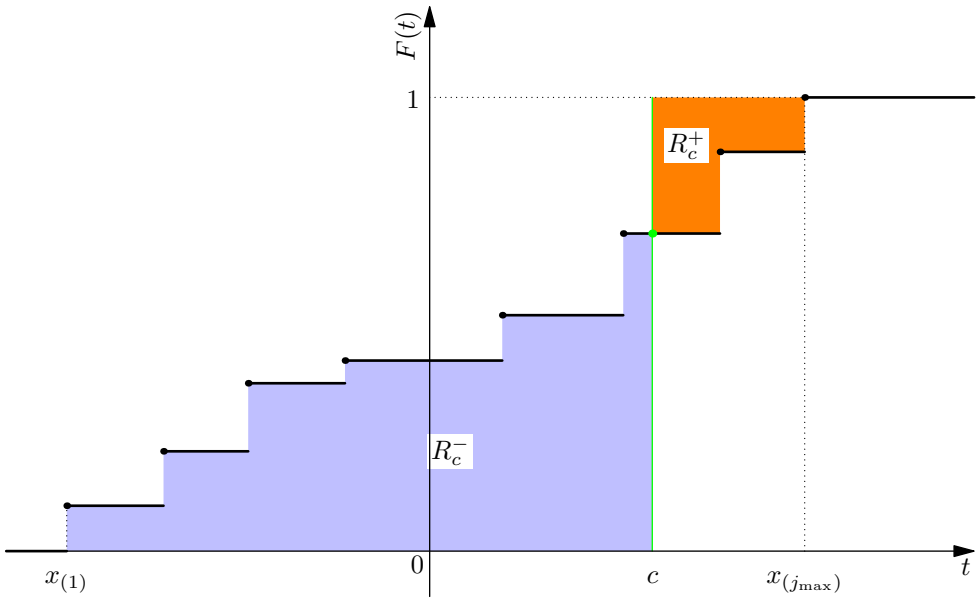

FIGURE 4 – Interprétation de $T_1(c)$ à l'aide de F : $T_1(c) = \mathrm{aire}(R_c^-) + \mathrm{aire}(R_c^+)$.

tégrale correspondante s'écrit :

$$T_1(c) = \sum_{j\in J} f_j \left|x_{(j)} - c\right| = \int_{\mathbb{R}} |t-c|\, \mathrm{d}\mu(t) = \int_{-\infty}^{c} F(t)\, \mathrm{d}t + \int_{c}^{+\infty} (1-F(t))\, \mathrm{d}t.$$

On peut avoir l'intuition du résultat de la minimisation de $T_1(c)$ en regardant comment varie la somme des aires de R_c^- et R_c^+ lorsque l'on déplace c.

Plus formellement, on commence par remarquer que la fonction T_1 est continue et même lipschitzienne de rapport 1 sur $\mathbb{R}$. En effet, en utilisant la croissance de μ (propriété 5), l'inégalité triangulaire et la propriété 1, on a pour tous réels c et c',

$$|T_1(c) - T_1(c')| \leqslant \int_{\mathbb{R}} \big|\, |t-c| - |t-c'| \,\big|\, \mathrm{d}\mu(t) \leqslant \int_{\mathbb{R}} |(t-c) - (t-c')|\, \mathrm{d}\mu(t) = |c-c'|\,.$$

Cette inégalité $|T_1(c') - T_1(c)| \leqslant |c - c'|$ aurait d'ailleurs pu s'obtenir graphiquement en interprétant $T_1(c) - T_1(c')$ comme une différence d'aires des deux zones délimitées par l'escalier dans le rectangle $[c, c'] \times [0, 1]$ et en majorant la valeur absolue de cette différence par la somme des aires. Considérons maintenant l'ensemble $I^+ = \{c \in \mathbb{R} \ ; \ F(c) > 1/2\}$. En raison de la croissance de F, I^+ est un intervalle de borne droite $+\infty$. La fonction T_1 est strictement croissante sur cet intervalle car si $c, c' \in I^+$ avec $c < c'$, quand on passe de c à c',

— l'aire de R_c^+ *diminue d'au plus* $(1 - F(c))(c' - c)$, autrement dit :

$$\text{aire}(R_c^+) - \text{aire}(R_{c'}^+) \leqslant (1 - F(c))(c' - c);$$

— l'aire de R_c^- *augmente d'au moins* $F(c)(c' - c)$, autrement dit :

$$\text{aire}(R_{c'}^-) - \text{aire}(R_c^-) \geqslant F(c)(c' - c).$$

Par conséquent,

$$\begin{aligned} T_1(c') - T_1(c) &= \big(\text{aire}(R_{c'}^+) - \text{aire}(R_c^+)\big) + \big(\text{aire}(R_{c'}^-) - \text{aire}(R_c^-)\big) \\ &\geqslant -(1 - F(c))(c' - c) + F(c)(c' - c) \\ &= (2F(c) - 1)(c' - c). \end{aligned}$$

Ce minorant est strictement positif car $c \in I^+$ et $c < c'$. Comme c et $c' > c$ étaient quelconques dans I^+, nous avons vérifié la stricte croissance de T_1 sur I^+.

Ensuite on considère $I^- = \{c \in \mathbb{R} \ ; \ F(c) < 1/2\}$. C'est un intervalle de borne gauche $-\infty$. Comme ci-dessus (détails laissés au lecteur), on vérifie que T_1 est strictement décroissante sur I^-. Par construction des intervalles I^- et I^+, la borne droite de I^- est inférieure ou égale à la borne gauche de I^+.

Si elles sont confondues, leur valeur commune est l'unique médiane m de la série (cf. le cas 1 dans la figure 1). Comme T_1 est strictement décroissante sur $I^- =]-\infty, m[$, strictement croissante sur $I^+ = [m, +\infty[$ et continue en tout point donc en m, elle atteint son unique minimum global en m.

Sinon, on est dans le cas 2 de la figure 1 : il y a un intervalle médian $[a, b]$ sur lequel $F(c)$ reste constamment égale à $1/2$, sauf en b. Sur cet intervalle, T_1 reste constante car lorsque c varie dans cet intervalle, tout gain ou perte de l'aire de R_c^+ est exactement compensé par la perte ou le gain de l'aire de R_c^-. Dans ce cas, T_1 est strictement décroissante sur $]-\infty, a[$, constante sur $[a, b]$ et strictement croissante sur $[b, +\infty[$. Comme elle est continue sur $\mathbb{R}$, donc en particulier en a, on en déduit que la valeur constante de T_1 sur $[a, b]$ est le minimum global de T_1 sur $\mathbb{R}$.

En conclusion, $T_1(c) = \sum_{j \in J} f_j \left| x_{(j)} - c \right|$ *a un unique minimum global atteint si et seulement si* c *est l'une quelconque des médianes de la série (ou* la *médiane s'il n'y en a qu'une). Chaque médiane réalise la meilleure approximation de la série au sens* L^1.

La valeur minimale $T_1(m)$ est donc la distance au sens L^1 de la série statistique à sa meilleure approximation par une constante. Il serait alors naturel comme nous l'avons fait pour l'écart-type dans le cas L^2, de prendre $T_1(m)$ comme indicateur de dispersion associé à la médiane. Il se trouve que les statisticiens préfèrent en général un autre paramètre qui est *l'écart interquartiles* $q_3 - q_1$. Le quartile q_i de la série est défini pour $i = 1, 2, 3$ comme la plus petite valeur de la série supérieure ou égale à au moins $i \times 25\%$ des observations. Notons que q_2 est la plus petite des médianes. L'intérêt de l'écart interquartiles est d'être, comme la médiane, insensible aux valeurs aberrantes, contrairement à $T_1(m)$. Par exemple si on observe la série d'effectif total 5 avec les valeurs : 1, 2, 3, 4, b, où $b > 4$, les quartiles sont $q_1 = 2$, $q_2 = 3$ (médiane unique dans ce cas), $q_3 = 4$. L'écart interquartiles est donc égal à 2 quelle que soit la valeur de $b > 4$. Par contre, $T_1(3) = 1 + b$ peut être très grand si b l'est.

Probabilités

Dans cette partie, on s'efforce de donner une présentation synthétique des connaissances probabilistes susceptibles d'être utiles aux enseignants de lycée. Pour des raisons de concision, on évitera les démonstrations que l'on peut trouver dans la plupart des ouvrages de probabilités. On insistera davantage sur des points problématiques ou peu connus.

Espaces probabilisés

Univers et tribus

La théorie moderne des probabilités représente les évènements observables à l'occasion d'une expérience aléatoire comme des sous-ensembles d'un même ensemble appelé parfois *univers* et noté traditionnellement Ω. Notons au passage que l'axiomatique de Kolmogorov (1933) qui fonde cette théorie ne définit pas la notion de hasard. De plus, tous les sous-ensembles de Ω ne sont pas forcément considérés comme des évènements. On se limite à une famille $\mathcal{F}$ de sous-ensembles de Ω appelée tribu. Cette famille sera le domaine de définition d'une *fonction d'ensembles* $P : \mathcal{F} \to [0,1]$, $A \mapsto P(A)$, appelée *probabilité*.

Nous allons expliciter cela dans un instant, mais auparavant arrêtons nous un instant sur cette idée peut-être un peu choquante de sous-ensembles de Ω qui ne seraient pas des évènements. On rencontre une situation analogue en géométrie lorsque l'on veut définir l'aire d'une région A du plan. On définit d'abord l'aire d'un carré (ou d'un rectangle). On peut alors quadriller le plan avec des carrés de côté c et coincer A entre la réunion des carrés du quadrillage qui sont inclus dans A et la réunion des carrés du quadrillage ayant une intersection non vide avec A. En faisant tendre c vers 0, si les deux aires encadrantes obtenues ont même limite, A aura une aire égale à cette limite. Mais ceci n'est pas possible pour toute région A du plan. Il y a des régions dont la structure est trop complexe pour qu'elles aient une aire. Lorsque l'on veut définir une probabilité sur la famille des sous-ensembles de Ω, il est rare que l'on dispose d'une formule explicite donnant immédiatement $P(A)$ en fonction de A. Le plus souvent, on a une idée précise de ce que devrait valoir P pour des ensembles bien particuliers et on essaie de construire P par divers prolongements faisant intervenir opérations ensemblistes et passages à la limite. Il

n'y a donc pas de raison que cette procédure permette d'attribuer une probabilité à tout sous-ensemble de Ω.

Bien sûr, si Ω est un ensemble fini et si on sait définir P sur les singletons, alors on peut prendre comme tribu $\mathcal{F} = \mathcal{P}(\Omega)$ famille de *tous* les sous-ensembles de Ω. $P(A)$ sera alors définie pour tout $A \subset \Omega$ comme la somme des $P(\{\omega\})$ pour tous les ω appartenant à A.

Le recours à un Ω infini s'impose pourtant souvent, même pour modéliser une expérience aussi simple que le jet répété d'un dé jusqu'à la première obtention du chiffre 6. D'autre part, une variable aléatoire susceptible de prendre une infinité de valeurs distinctes ne peut être définie que sur un Ω infini : si Ω est fini, pour toute application $X : \Omega \to \mathbb{R}$, l'ensemble image $X(\Omega) = \{X(\omega)\ ;\ \omega \in \Omega\}$ est forcément fini et de cardinal inférieur ou égal à celui de Ω. Enfin, les grands théorèmes limite du calcul des probabilités comme la loi des grands nombres et le théorème limite central fournissent le comportement limite de sommes de n variables aléatoires lorsque n tend vers l'infini. Sauf cas dégénérés, pour que ces sommes soient définies sur le même Ω, il est nécessaire que Ω soit infini. Pour ne prendre qu'un exemple, on ne peut pas définir une suite infinie de variables aléatoires de Bernoulli $(X_i)_{i\geqslant 1}$ indépendantes et de même loi (jeu de pile ou face infini) sur un Ω fini et il en va de même pour la suite $(S_n)_{n\geqslant 1}$ des sommes partielles $S_n = X_1 + \cdots + X_n$.

Nous avons évoqué à propos de la notion d'aire, les opérations ensemblistes. Elles jouent un rôle vital en théorie des probabilités. Les trois opérations basiques [2] sont

— la réunion : $A \cup B = \{\omega \in \Omega\ ;\ \omega \in A \text{ ou } \omega \in B\}$;
— l'intersection : $A \cap B = \{\omega \in \Omega\ ;\ \omega \in A \text{ et } \omega \in B\}$;
— le complémentaire : $A^c = \{\omega \in \Omega\ ;\ \omega \notin A\}$.

En réécrivant la définition de la réunion et de l'intersection sous la forme suivante,

— $\omega \in A \cup B$ si et seulement si ω appartient à l'un *au moins* des ensembles A et B,
— $\omega \in A \cap B$ si et seulement si ω appartient à *chacun* des ensembles A et B,

on peut étendre ces opérations à une famille quelconque, y compris infinie, $(A_i)_{i\in I}$ de sous-ensembles de Ω :

$$\begin{aligned}\bigcup_{i\in I} A_i &= \{\omega \in \Omega\ ;\ \omega \text{ appartient à au moins l'un des } A_i \text{ pour } i \in I\}\\ &= \{\omega \in \Omega\ ;\ \exists i \in I,\ \omega \in A_i\}\,,\end{aligned}$$

$$\begin{aligned}\bigcap_{i\in I} A_i &= \{\omega \in \Omega\ ;\ \omega \text{ appartient à chacun des } A_i \text{ pour } i \in I\}\\ &= \{\omega \in \Omega\ ;\ \forall i \in I,\ \omega \in A_i\}\,.\end{aligned}$$

On voit ainsi que les opérations ensemblistes intersection et réunion permettent de traduire les opérations logiques « et », « ou », ainsi que les quantificateurs $\forall$ et $\exists$.

2. Deux quelconques d'entre elles permettent de reconstruire toutes les autres.

Il importe de remarquer que lorsque I est infini, les définitions généralisées ci-dessus sont globales et s'obtiennent directement sans aucun passage à la limite.

Remarque. Dans la théorie des probabilités, on se limite en général aux opérations ensemblistes sur des familles d'évènements finies ou dénombrables. Rappelons qu'un ensemble infini I est dit *dénombrable* s'il est en bijection avec $\mathbb{N}$. Nous dirons qu'un ensemble est *au plus dénombrable* s'il est fini ou dénombrable. Parmi les ensembles dénombrables, citons $\mathbb{N}$ et chacun de ses sous-ensembles infinis, $\mathbb{Z}$, $\mathbb{Q}$, toute union d'une famille dénombrable d'ensembles dénombrables, tout produit cartésien d'une famille finie d'ensembles dénombrables (en particulier $\mathbb{N}^d$). Parmi les ensemble infinis non dénombrables, citons $\mathbb{R}$, $[a, b]$ (avec $a < b$), $\{0, 1\}^{\mathbb{N}}$ ensemble des suites binaires infinies et tout ensemble contenant un sous-ensemble infini non dénombrable. ◁

Nous pouvons maintenant formaliser la définition d'une tribu.

Définition (tribu). Une famille $\mathcal{F}$ de parties de Ω est appelée *tribu* sur Ω si elle
a) possède l'ensemble vide : $\varnothing \in \mathcal{F}$;
b) est stable par passage au complémentaire : $\forall A \in \mathcal{F}$, $A^c \in \mathcal{F}$;
c) est stable par union dénombrable : pour toute famille dénombrable $(A_i)_{i\in\mathbb{N}^*}$ dans $\mathcal{F}$, $\bigcup_{i\in\mathbb{N}^*} A_i \in \mathcal{F}$.

Le couple $(\Omega, \mathcal{F})$ est appelé espace mesurable ou probabilisable. ◁

On vérifie à partir de cette définition qu'une tribu est stable par unions finies (prendre tous les A_i vides à partir d'un certain rang), intersections finies et par intersections dénombrables (combiner *b)* et *c)*).

Les trois exemples les plus simples de tribus sont les suivants.
— La tribu triviale sur Ω est $\mathcal{F} = \{\Omega, \varnothing\}$.
— $\mathcal{P}(\Omega)$ famille de toutes les parties de Ω est une tribu.
— Si A est une partie de Ω, alors $\mathcal{F} = \{\Omega, \varnothing, A, A^c\}$ est une tribu. C'est la *plus petite* tribu possédant A comme élément, au sens où toute tribu $\mathcal{G}$ telle que $A \in \mathcal{G}$ contient $\mathcal{F}$. On dit que $\mathcal{F}$ est la tribu *engendrée* par A.

Cette notion de tribu engendrée se généralise en remarquant que si $(\mathcal{G}_i)_{i\in I}$ est une famille quelconque de tribus sur Ω, $\mathcal{G} := \bigcap_{i\in I} \mathcal{G}_i$ est une tribu sur Ω (vérification immédiate à partir de la définition).

Définition (tribu engendrée). Soit $\mathcal{C}$ une famille de parties d'un ensemble Ω. On appelle *tribu engendrée* par $\mathcal{C}$, et on note $\sigma(\mathcal{C})$, la plus petite tribu contenant $\mathcal{C}$. C'est l'intersection de toutes les tribus sur Ω contenant $\mathcal{C}$. ◁

Définition (tribu borélienne). On appelle *tribu borélienne* sur $\mathbb{R}^d$ la tribu engendrée par la famille $\mathcal{O}$ des ensembles *ouverts*[3] de $\mathbb{R}^d$. On la notera $\mathrm{Bor}(\mathbb{R}^d)$. Ainsi

3. Un ensemble ouvert de $\mathbb{R}^d$ est une réunion (quelconque) de *pavés* ouverts $\prod_{k=1}^d]a_k, b_k[$. Un *fermé* est le complémentaire d'un ouvert.

$\mathrm{Bor}(\mathbb{R}^d) = \sigma(\mathcal{O})$. Les sous-ensembles de $\mathbb{R}^d$ qui sont éléments de sa tribu borélienne sont appelés boréliens de $\mathbb{R}^d$ ou boréliens tout court quand il n'y a pas d'ambiguïté. ◁

Remarque. On peut démontrer que $\mathrm{Bor}(\mathbb{R})$ est aussi engendrée par les fermés de $\mathbb{R}$, ou par les intervalles ouverts, ou les intervalles fermés, ou les semi-ouverts, ou les intervalles (ouverts ou fermés) à extrémités rationnelles, ou les intervalles $]-\infty, a]$, ou les intervalles $[a, +\infty[$. De même, $\mathrm{Bor}(\mathbb{R}^d)$ est engendrée par les pavés ouverts ou par les pavés de la forme $\prod_{k=1}^{d}]a_k, b_k]$. ◁

Les trois remarques suivantes tentent de répondre à des questions légitimes de lecteurs ayant déjà une certaine familiarité avec la théorie des probabilités. Elles ne sont pas destinées à une réutilisation directe dans l'enseignement secondaire.

Remarque. Pour une expérience aléatoire donnée, il existe généralement plusieurs choix possibles d'un espace probabilisable $(\Omega, \mathcal{F})$ utilisables pour modéliser cette expérience. Voir à ce sujet la section 1.5 *Remarques sur le choix d'un modèle* dans [8]. Ces choix peuvent donner des résultats identiques ou différents pour le calcul de la probabilité d'un évènement donné. C'est une des difficultés de l'enseignement de la théorie des probabilités (cf. aussi le célèbre *paradoxe de Bertrand*). Il importe de comprendre que dans la résolution d'un problème de probabilités, les mathématiques commencent seulement une fois spécifié le modèle utilisé, c'est-à-dire quel Ω ? quelle tribu $\mathcal{F}$? et comme nous le verrons ci-dessous, quelle probabilité P ? ou plus largement quelle famille de probabilités P ? Bien sûr, la connaissance de la théorie peut apporter une aide dans le choix de $(\Omega, \mathcal{F})$. Par exemple dans le cas où l'on veut traduire une certaine indépendance pour une suite de n expériences, il sera commode de prendre pour Ω le produit cartésien $\Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$, où $(\Omega_i, \mathcal{F}_i)$ est une représentation de l'expérience n$^\text{o}$ i et pour tribu $\mathcal{F}$ la tribu produit $\mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \cdots \otimes \mathcal{F}_n$ c'est-à-dire la tribu engendrée par les ensembles produits $A_1 \times A_2 \times \cdots \times A_n$, où pour tout $i \in [\![1, n]\!]$, $A_i \in \mathcal{F}_i$. ◁

Remarque. En fait, dès que l'on a trouvé un espace probabilisable $(\Omega, \mathcal{F})$ pour représenter une expérience aléatoire donnée, on peut en trouver une infinité, ne serait-ce qu'en considérant les espaces produits $(\Omega^n, \mathcal{F}^{\otimes n})$, où Ω^n représente le produit cartésien de n facteurs Ω et $\mathcal{F}^{\otimes n}$ la tribu produit (définie comme dans la remarque précédente) de $\mathcal{F}$ par elle même en n facteurs. Si $A \in \mathcal{F}$ représente une évènement associé à l'expérience dans l'espace probabilisable initial $(\Omega, \mathcal{F})$, il est représenté dans l'espace probabilisable $(\Omega^n, \mathcal{F}^{\otimes n})$ par $A' = A \times \Omega^{n-1}$. ◁

Remarque. Dans l'axiomatique de Kolmogorov, les « évènements élémentaires » $\{\omega\}$ ne sont pas forcément membres de la tribu $\mathcal{F}$. Dit autrement, les évènements élémentaires ne sont pas forcément des évènements observables ! Cela peut paraître très choquant, mais en voici un exemple élémentaire. On jette deux dés indistinguables (même couleur, mêmes dimensions, même matériau, réception sur la même

table, etc.) et on s'intéresse à la somme de points indiqués par chacun des deux dés. On peut prendre pour Ω le produit cartésien $[\![1,6]\!]^2$. Cela semble un peu maladroit puisqu'on n'a pas de moyen de distinguer entre première et deuxième composante d'un couple (i,j) dans cet Ω. Mais on peut prendre en compte cette indistinguabilité en choisissant pour tribu d'évènements observables, au lieu de $\mathcal{P}(\Omega)$, la tribu $\mathcal{F}$ des sous-ensembles *symétriques* de Ω. Un sous-ensemble A de Ω est dit symétrique si pour tout $(i,j) \in A$, son couple symétrique (j,i) appartient aussi à A. On laisse en exercice la vérification du fait que la famille des sous-ensembles symétriques de Ω est une tribu. Dans ce modèle, les singletons $\{(i,i)\}$ sont membres de la tribu $\mathcal{F}$, mais pas les singletons $\{(i,j)\}$ pour $i \neq j$. Le plus petit élément de la tribu $\mathcal{F}$ contenant un tel singleton est l'ensemble symétrique $\{(i,j),(j,i)\}$. ◁

Mesures de probabilité

Définition. Soit Ω un ensemble et $\mathcal{F}$ une tribu sur Ω. On appelle probabilité sur $(\Omega,\mathcal{F})$ toute application P de $\mathcal{F}$ dans $[0,1]$ vérifiant :

(i) $P(\Omega) = 1$.

(ii) P est σ-additive : pour toute suite $(A_j)_{j\geqslant 1}$ d'évènements de $\mathcal{F}$ deux à deux disjoints (incompatibles) :

$$P\Big(\bigcup_{j\in\mathbb{N}^*} A_j\Big) = \sum_{j=1}^{+\infty} P(A_j).$$

Le triplet $(\Omega,\mathcal{F},P)$ s'appelle espace probabilisé. ◁

Définir une probabilité sur $(\Omega,\mathcal{F})$ c'est en quelque sorte attribuer une « masse » à chaque évènement observable, avec par convention une masse totale égale à 1 pour l'évènement certain Ω.

Remarque. Lorsque Ω est un ensemble fini, on vérifie facilement que la définition ci-dessus se réduit à (i) et

$$\forall A,B \in \mathcal{F} \text{ tels que } A\cap B = \varnothing, \quad P(A\cup B) = P(A) + P(B).$$

Le point clé est de remarquer que si Ω a n éléments, toute suite $(A_j)_{j\geqslant 1}$ d'évènements de $\mathcal{F}$ deux à deux disjoints a au plus n évènements non vides. La σ-additivité se réduit alors à l'additivité finie soit $P(\bigcup_{i\in K} A_i) = \sum_{i\in K} P(A_i)$ pour K partie finie de $\mathbb{N}^*$ et les A_i deux à deux disjoints. Il suffit que cette propriété soit vraie pour K de cardinal 2 pour qu'elle le soit pour tout K fini ayant au moins 2 éléments (pourquoi ?). ◁

Revenons au cas général où Ω est quelconque. À partir de la définition axiomatique ci-dessus, on démontre facilement les propriétés suivantes.

Proposition (propriétés générales d'une probabilité).
Toute probabilité P sur $(\Omega, \mathcal{F})$ vérifie les propriétés suivantes :

1. $P(\varnothing) = 0$.
2. *Additivité.*
   a) *Si* $A \cap B = \varnothing$, $P(A \cup B) = P(A) + P(B)$.
   b) *Si les* A_i $(1 \leqslant i \leqslant n)$ *sont deux à deux disjoints :* $P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$.
3. $\forall A \in \mathcal{F}$, $P(A^c) = 1 - P(A)$.
4. $\forall A \in \mathcal{F}$, $\forall B \in \mathcal{F}$, $A \subset B \;\Rightarrow\; P(A) \leqslant P(B)$.
5. $\forall A \in \mathcal{F}$, $\forall B \in \mathcal{F}$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
6. *Continuité monotone séquentielle.*
   a) *Si* $(B_n)_{n \geqslant 0}$ *est une suite croissante d'évènements de* $\mathcal{F}$ *convergente*[4] *vers* $B \in \mathcal{F}$, *alors* $P(B) = \lim_{n \to +\infty} P(B_n)$. *En abrégé :*
   $$B_n \uparrow B \Rightarrow P(B_n) \uparrow P(B) \quad (n \to +\infty).$$
   b) *Si* $(C_n)_{n \geqslant 0}$ *est une suite décroissante d'évènements de* $\mathcal{F}$ *convergente*[5] *vers* $C \in \mathcal{F}$, *alors* $P(C) = \lim_{n \to +\infty} P(C_n)$. *En abrégé :*
   $$C_n \downarrow C \Rightarrow P(C_n) \downarrow P(C) \quad (n \to +\infty).$$
7. *Sous-additivité et sous-*σ*-additivité.*
   a) $\forall A \in \mathcal{F}, \forall B \in \mathcal{F}, \quad P(A \cup B) \leqslant P(A) + P(B)$.
   b) $\forall A_1, \ldots, A_n \in \mathcal{F}, \quad P\left(\bigcup_{i=1}^{n} A_i\right) \leqslant \sum_{i=1}^{n} P(A_i)$.
   c) $\forall A_1, \ldots, A_n, \ldots \in \mathcal{F}$, $P\left(\bigcup_{i \in \mathbb{N}^*} A_i\right) \leqslant \sum_{i=1}^{+\infty} P(A_i)$.

Le calcul de probabilités de réunions ou d'intersections est une question cruciale. La propriété 5 montre qu'en général on ne peut pas calculer $P(A \cup B)$ à partir de la *seule connaissance* de $P(A)$ et $P(B)$ et qu'on se heurte à la même difficulté pour $P(A \cap B)$. Le calcul des probabilités d'intersections sera discuté plus tard, à propos du conditionnement. Pour les probabilités de réunions, on peut se demander comment se généralise la propriété 5 lorsqu'on réunit plus de deux évènements. Il est facile de vérifier que :

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

Le cas général est donné par la formule de Poincaré qui exprime $P(A_1 \cup \cdots \cup A_n)$ à l'aide des probabilités de *toutes les intersections* des A_i : 2 à 2, 3 à 3, etc.

Proposition (formule de Poincaré).
Pour tout entier $n \geqslant 2$ *et tous évènements* $A_1, \ldots, A_n$ *:*

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) + \sum_{k=2}^{n} (-1)^{k+1} \sum_{1 \leqslant i_1 < i_2 < \ldots < i_k \leqslant n} P(A_{i_1} \cap \cdots \cap A_{i_k}).$$

4. Ce qui signifie : $\forall n \geqslant 0$, $B_n \subset B_{n+1}$ et $B = \bigcup_{n \geqslant 0} B_n$.
5. Ce qui signifie : $\forall n \geqslant 0$, $C_{n+1} \subset C_n$ et $C = \bigcap_{n \geqslant 0} C_n$.

Exemples de probabilités

Exemple (Masse de Dirac). L'exemple le plus simple de probabilité sur $(\Omega, \mathcal{F})$ est la masse de Dirac δ_{ω_0} en un point ω_0 fixé de Ω. Elle est définie par

$$\forall A \in \mathcal{F}, \quad \delta_{\omega_0}(A) = \mathbf{1}_A(\omega_0) = \begin{cases} 1 & \text{si } \omega_0 \in A, \\ 0 & \text{si } \omega_0 \notin A. \end{cases}$$

Cet exemple est presque trivial mais fort utile comme brique de base pour construire d'autres probabilités. ◁

Exemple (probabilités définies sur un Ω fini, muni de la tribu $\mathcal{P}(\Omega)$). Soit $\Omega = \{\omega_1, \dots, \omega_n\}$ un univers fini, muni de la tribu $\mathcal{P}(\Omega)$ de tous ses sous-ensembles. Cette tribu contient en particulier les n singletons $\{\omega_i\}$. Si P est une probabilité sur $(\Omega, \mathcal{P}(\Omega))$ alors par additivité de P (propriété 2 *b)* page 18), pour tout $A \in \mathcal{P}(\Omega)$, c'est-à-dire pour tout $A \subset \Omega$,

$$\begin{aligned} P(A) = P\left(\bigcup_{\omega_i \in A} \{\omega_i\}\right) &= \sum_{\omega_i \in A} p_i \quad \text{en posant } p_i = P(\{\omega_i\}) \\ &= \sum_{i=1}^{n} p_i \delta_{\omega_i}(A) \\ &= \left(\sum_{i=1}^{n} p_i \delta_{\omega_i}\right)(A). \end{aligned}$$

On en déduit que P est la mesure ponctuelle $\sum_{i=1}^n p_i \delta_{\omega_i}$. Remarquons que les p_i sont positifs ou nuls et que $\sum_{i=1}^n p_i = P(\Omega) = 1$. Réciproquement, on vérifie facilement que toute mesure ponctuelle de la forme $\sum_{i=1}^n p_i \delta_{\omega_i}$ avec les $p_i \geqslant 0$ et de somme 1 est une probabilité sur $(\Omega, \mathcal{P}(\Omega))$. Nous avons ainsi une caractérisation de *toutes* les probabilités sur $(\Omega, \mathcal{P}(\Omega))$. ◁

Remarque. Même sur Ω fini, on n'est pas obligé de prendre $\mathcal{F} = \mathcal{P}(\Omega)$. Voici un exemple simple. On représente le jet de deux dés cubiques indiscernables (même matériau, mêmes dimensions, même couleur) par $\Omega = [\![1, 6]\!]^2$. Pour prendre en compte l'indiscernabilité des couples (i, j) et (j, i), on munit Ω de la tribu $\mathcal{F}$ de ses parties symétriques, cf. la dernière remarque de la page 17. Si les deux dés sont équilibrés, on construira P en lui attribuant la valeur $2/36$ sur les évènements $\{(i, j), (j, i)\}$ avec $i \neq j$ et la valeur $1/36$ sur les $\{(i, i)\}$. Ceci permet par additivité finie de définir P sur tout évènement symétrique $A \in \mathcal{F}$. Avec cette modélisation, on retrouve les mêmes probabilités pour les évènements symétriques (par exemple « la somme des points est paire ») qu'avec la modélisation « classique » où on suppose les dés discernables et on choisit le même Ω, muni de la tribu $\mathcal{P}(\Omega)$ et de l'équiprobabilité. ◁

Remarque. Lorsque Ω est fini, la façon la plus simple de construire une probabilité sur $(\Omega, \mathcal{P}(\Omega))$ est de choisir $P(\{\omega\}) = 1/\operatorname{card}\Omega$. On parle alors d'*équiprobabilité* ou

de *probabilité uniforme* sur $(\Omega, \mathcal{P}(\Omega))$. C'est la modélisation qui s'impose naturellement lorsqu'on n'a pas de raison de penser *a priori* qu'un résultat élémentaire de l'expérience soit favorisé ou défavorisé par rapport aux autres. La situation est radicalement différente lorsque Ω est infini *dénombrable*. Sur un tel ensemble, *il ne peut pas y avoir d'équiprobabilité*. Imaginons que l'on veuille tirer une boule *au hasard* dans une urne contenant une infinité de boules numérotées de manière bijective par les entiers naturels. Soit $\{\omega_i\}$ l'évènement *tirage de la boule numérotée* i ($i \in \mathbb{N}$) et p_i sa probabilité. Par σ-additivité, les p_i vérifient nécessairement :

$$\sum_{i\in\mathbb{N}} p_i = 1.$$

Mais si les p_i sont égaux, tous les termes de la série ci-dessus valent p_0. Sa somme est alors $+\infty$ si $p_0 > 0$ ou 0 si $p_0 = 0$, il y a donc une contradiction. ◁

Voici une caractérisation des probabilités sur les espaces Ω au plus dénombrables.

Proposition. *Soit $\Omega = \{\omega_i \ ; \ i \in I\}$ un ensemble au plus dénombrable. La donnée d'une probabilité sur $(\Omega, \mathcal{P}(\Omega))$ équivaut à la donnée d'une famille $(p_i)_{i\in I}$ dans $\mathbb{R}_+$ telle que :*

$$\sum_{i\in I} p_i = 1$$

et des égalités

$$P(\{\omega_i\}) = p_i, \quad i \in I.$$

La probabilité P s'écrit alors $P = \sum_{i\in I} p_i \delta_{\omega_i}$, où δ_{ω_i} désigne la masse de Dirac (ou mesure de Dirac) au point ω_i. Par conséquent, pour tout $A \subset \Omega$, $P(A) = \sum_{\omega_i \in A} p_i$.

Remarque. Lorsque I est infini dénombrable, les écritures $\sum_{i\in I} p_i$ et $\sum_{i\in I} p_i \delta_{\omega_i}$ nécessitent quelques éclaircissements. Considérons une numérotation bijective particulière de I par $\mathbb{N}$: $\varphi : \mathbb{N} \to I$, $k \mapsto \varphi(k) = i$. Alors comme tous les $p_{\varphi(k)}$ sont positifs, la convergence et la somme de $\sum_{k=0}^{+\infty} p_{\varphi(k)} \delta_{\omega_{\varphi(k)}}$ ne dépendent pas du choix de la numérotation φ. En particulier, l'écriture $\sum_{i\in I} p_i = 1$ signifie que pour toute numérotation bijective φ, la série $\sum_{k=0}^{+\infty} p_{\varphi(k)} \delta_{\omega_{\varphi(k)}}$ converge et a pour somme 1. Cette convergence implique la convergence de la série à termes positifs $\sum_{k=0}^{+\infty} p_{\varphi(k)} \delta_{\omega_{\varphi(k)}}(A) = \sum_{k=0}^{+\infty} p_{\varphi(k)} \mathbf{1}_A(\omega_{\varphi(k)})$ vers une somme inférieure ou égale à 1. Comme tous les termes sont positifs, cette convergence et cette somme ne dépendent pas du choix de φ. Ceci justifie à nouveau l'écriture $\sum_{i\in I} p_i \delta_{\omega_i}(A)$. Et comme ceci vaut pour tout $A \in \mathcal{F}$, la définition de la probabilité P par la formule $P = \sum_{i\in I} p_i \delta_{\omega_i}$ est justifiée. Nous utiliserons librement dans la suite de ce document cette possibilité d'indexer des sommes de réels positifs ou de mesures par un ensemble dénombrable sans préciser le choix de la numérotation. ◁

Exemple (une probabilité définie sur $\big(\mathbb{N}, \mathcal{P}(\mathbb{N})\big)$). Soit a un réel strictement positif fixé. On pose :

$$\forall k \in \mathbb{N}, \qquad p_k = \frac{\mathrm{e}^{-a} a^k}{k!}.$$

On remarque que p_k est le terme général positif d'une série convergente :

$$\sum_{k=0}^{+\infty} \frac{\mathrm{e}^{-a} a^k}{k!} = \mathrm{e}^{-a} \sum_{k=0}^{+\infty} \frac{a^k}{k!} = \mathrm{e}^{-a}\mathrm{e}^{a} = 1.$$

Pour tout $A \subset \mathbb{N}$, on définit :

$$P(A) = \sum_{k \in A} p_k = \sum_{k \in \mathbb{N}} p_k \delta_k(A).$$

D'après la proposition précédente, P est une probabilité sur $(\mathbb{N}, \mathcal{P}(\mathbb{N}))$. On l'appelle *loi de Poisson de paramètre a.* ◁

Exemple (loi uniforme sur un segment de $\mathbb{R}$). Prenons $\Omega = \mathbb{R}$, $\mathcal{F} = \mathrm{Bor}(\mathbb{R})$ et notons λ_1 la mesure de Lebesgue sur $\mathbb{R}$ (c'est l'unique mesure sur la tribu borélienne de $\mathbb{R}$ qui donne pour mesure d'un segment quelconque la longueur de ce segment). Soit $[a, b]$ un segment fixé de $\mathbb{R}$. On définit une probabilité P sur $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$ en posant :

$$\forall A \in \mathrm{Bor}(\mathbb{R}), \quad P(A) = \frac{\lambda_1(A \cap [a,b])}{\lambda_1([a,b])} = \frac{\lambda_1(A \cap [a,b])}{b-a}.$$

Cette probabilité P est appelée *loi uniforme sur* $[a, b]$. Remarquons que pour cette probabilité, tout singleton est de probabilité nulle ($\forall x \in \mathbb{R}$, $P(\{x\}) = 0$), ce qui résulte de la propriété analogue de λ_1. On voit sur cet exemple que la probabilité d'une union infinie *non dénombrable* d'évènements deux à deux disjoints n'est pas forcément égale à la somme de la famille correspondante de probabilités d'évènements. En effet,

$$1 = P([a,b]) = P\Big(\bigcup_{x \in [a,b]} \{x\} \Big) \neq \sum_{x \in [a,b]} P(\{x\}) = 0.$$

◁

Exemple (lois uniformes dans $\mathbb{R}^d$). On peut généraliser l'exemple précédent en prenant au lieu d'un segment, un borélien B de $\mathbb{R}^d$ vérifiant $0 < \lambda_d(B) < +\infty$. Ici λ_d désigne la mesure de Lebesgue de $\mathbb{R}^d$ qui étend la notion de volume d'un pavé à tous les boréliens de $\mathbb{R}^d$. On définit alors une probabilité P sur $(\mathbb{R}^d, \mathrm{Bor}(\mathbb{R}^d))$, en posant :

$$\forall A \in \mathrm{Bor}(\mathbb{R}^d), \quad P(A) = \frac{\lambda_d(A \cap B)}{\lambda_d(B)}.$$

Cette probabilité est appelée *loi uniforme sur B.* ◁

Les lois uniformes des deux exemples précédents apparaissent naturellement dans des problèmes de probabilités géométriques qui permettent dès le lycée d'utiliser sans développement théorique un Ω infini (problème de la tige brisée, jeu de franc carreau). Elles apparaissent aussi naturellement dans les calculs d'aire ou de volume par la méthode de Monte-Carlo.

Nous allons donner maintenant une caractérisation de *toutes les probabilités* P sur $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$. Nous verrons ultérieurement que ces probabilités sont exactement les lois des variables aléatoires réelles. Comme le montre l'exemple de la loi uniforme sur un segment, il est illusoire d'espérer caractériser ces probabilités par les $P(\{x\})$. La situation est donc radicalement différente du cas d'un espace Ω au plus dénombrable. Au lieu des $P(\{x\})$, nous allons utiliser les $P(]a, b])$, ou les $P(]-\infty, x])$. Nous aurons besoin pour cela de la notion de fonction de répartition.

Définition (fonction de répartition). Soit P une probabilité sur $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$. On appelle fonction de répartition de P, l'application

$$F : \mathbb{R} \to [0, 1], \quad x \mapsto F(x) = P(]-\infty, x]).$$

◁

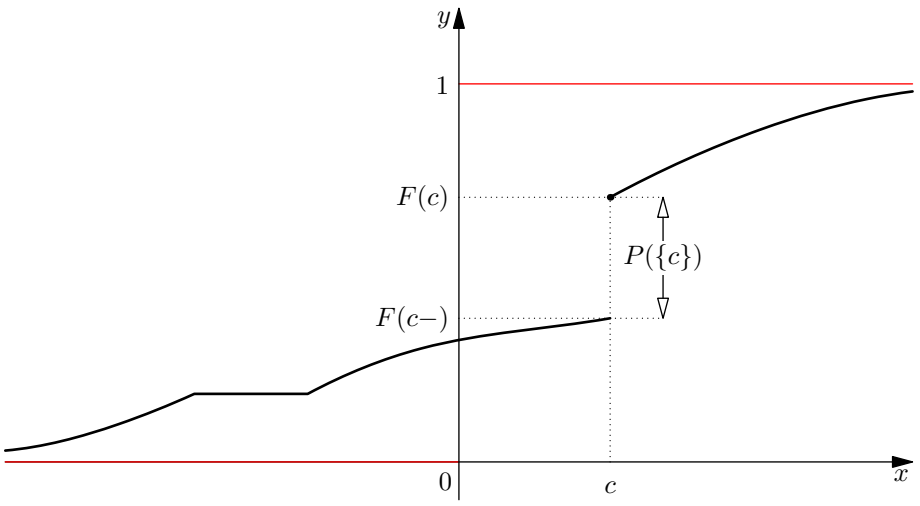

FIGURE 5 – Une fonction de répartition avec une discontinuité au point $x = c$

Voici les propriétés générales des fonctions de répartition.

Proposition. *La fonction de répartition F d'une probabilité P sur $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$ a les propriétés suivantes.*

a) F est croissante sur $\mathbb{R}$.

b) F a pour limite 0 en $-\infty$ et 1 en $+\infty$.

c) F est continue à droite sur $\mathbb{R}$ et a une limite à gauche en tout point $x \in \mathbb{R}$.

d) En notant $F(x-) = \lim_{\varepsilon \downarrow 0} F(x-\varepsilon)$ *la limite à gauche*[6] *au point* x*, on a*

$$\forall x \in \mathbb{R}, \quad P(\{x\}) = F(x) - F(x-).$$

De plus l'ensemble des $x \in \mathbb{R}$ *tels que* $F(x) \neq F(x-)$ *est au plus dénombrable.*

e) Si deux probabilités P_1 *et* P_2 *sur* $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$ *ont même fonction de répartition, elles sont égales, autrement dit* $P_1(B) = P_2(B)$ *pour tout* $B \in \mathrm{Bor}(\mathbb{R})$*.*

Le théorème suivant permet de construire toutes les probabilités sur $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$.

Théorème. *Soit* F *une fonction croissante et continue à droite sur* $\mathbb{R}$*, de limites* 0 *en* $-\infty$ *et* 1 *en* $+\infty$*. Il existe une unique probabilité* P *sur* $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$ *telle que*

$$\forall a, b \in \mathbb{R}, \text{ avec } a \leqslant b, \quad P(]a, b]) = F(b) - F(a).$$

Alors F *est la fonction de répartition de* P*.*

Probabilités conditionnelles

Prise en compte d'une information supplémentaire

La notion de probabilité conditionnelle permet de prendre en compte l'information de la réalisation d'un certain évènement pour modifier la probabilité de chaque évènement.

Définition (probabilité conditionnelle). Soit H un évènement de probabilité *non nulle*. Pour tout évènement A, on définit :

$$P(A \mid H) = \frac{P(A \cap H)}{P(H)},$$

appelée *probabilité conditionnelle* de l'évènement A *sous l'hypothèse* H. ◁

Remarquons que pour l'instant, il ne s'agit que d'un jeu d'écriture. On a simplement défini un réel $P(A \mid H)$ pour que :

$$P(A \cap H) = P(A \mid H)P(H).$$

Ce qui fait l'intérêt du concept de probabilité conditionnelle, c'est qu'il est souvent bien plus facile d'attribuer *directement* une valeur à $P(A \mid H)$ en tenant compte des conditions expérimentales (liées à l'information supplémentaire de la réalisation de H) et d'en déduire ensuite la valeur de $P(A \cap H)$. Le raisonnement implicite alors utilisé est : tout espace probabilisé modélisant correctement la réalité expérimentale devrait fournir cette valeur pour $P(A \mid H)$.

6. Cette notation est un peu abusive, puisqu'il ne s'agit pas forcément d'une valeur prise par la fonction F. Attention à ne pas confondre le « $x-$ » dans $F(x-)$ avec le « x^- » partie négative de x.

Remarque. Il importe de bien comprendre que l'écriture « $A \mid H$ » ne désigne pas un nouvel évènement [7] différent de A. Quand on écrit $P(A \mid H)$, ce que l'on a modifié, ce n'est pas l'évènement A, mais la valeur numérique qui lui était attribuée par la fonction d'ensembles P. Il serait donc en fait plus correct d'écrire $P_H(A)$ que $P(A \mid H)$. On conserve néanmoins cette dernière notation essentiellement pour des raisons typographiques : $P(A \mid H_1 \cap H_2 \cap H_3)$ est plus lisible que $P_{H_1 \cap H_2 \cap H_3}(A)$. ◅

Propriétés

Proposition. *Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé et H un évènement* fixé *tel que $P(H) \neq 0$. Alors la fonction d'ensembles $P(\,.\mid H)$ définie par :*

$$P(\,.\mid H) : \quad \mathcal{F} \to [0,1] \quad B \mapsto P(B \mid H)$$

est une nouvelle probabilité sur $(\Omega, \mathcal{F})$.

En conséquence, la fonction d'ensembles $P(\,.\mid H)$ vérifie toutes les propriétés générales d'une probabilité exposées page 18.

Nous n'avons vu jusqu'ici aucune formule permettant de calculer la probabilité d'une intersection d'évènements à l'aide des probabilités de ces évènements. Une telle formule *n'existe pas dans le cas général*. Les probabilités conditionnelles fournissent une méthode générale tout à fait naturelle pour calculer une probabilité d'intersection.

Proposition (règle des conditionnements successifs).
Soit n un entier supérieur ou égal à 2. Si $A_1, \ldots, A_n$ sont n évènements tels que $P(A_1 \cap \ldots \cap A_{n-1}) \neq 0$, on a :

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \times \cdots \times P(A_n \mid A_1 \cap \ldots \cap A_{n-1}).$$

Les probabilités conditionnelles permettent aussi de calculer la probabilité d'un évènement en conditionnant par *tous les cas possibles*. Du point de vue ensembliste, ces *cas possibles* correspondent à une *partition* de Ω.

Définition (partition). On dit qu'une famille $(H_i)_{i \in I}$ de parties de Ω est une partition de Ω si elle vérifie les trois conditions suivantes.

— $\forall i \in I,\ H_i \neq \varnothing$.
— $\Omega = \bigcup_{i \in I} H_i$.
— Les H_i sont deux à deux disjoints ($i \neq j \Rightarrow H_i \cap H_j = \varnothing$).

◅

7. En fait cette écriture prise isolément (sans le P) *n'a pas de sens* et ne devrait *jamais* être utilisée. Le symbole $\mid$ ne représente pas une opération sur les évènements qui l'entourent.

Proposition (conditionnement par les cas possibles). [8]

(i) Si H est tel que $P(H) \neq 0$ et $P(H^c) \neq 0$, on a

$$\forall A \in \mathcal{F}, \quad P(A) = P(A \mid H)P(H) + P(A \mid H^c)P(H^c).$$

(ii) Si $(H_i)_{i\in[\![1,n]\!]}$ est une partition finie de Ω en évènements de probabilité non nulle,

$$\forall A \in \mathcal{F}, \quad P(A) = \sum_{i=1}^{n} P(A \mid H_i)P(H_i).$$

(iii) Si $(H_i)_{i\in\mathbb{N}}$ est une partition de Ω telle qu'aucune $P(H_i)$ ne soit nulle,

$$\forall A \in \mathcal{F}, \quad P(A) = \sum_{i=0}^{+\infty} P(A \mid H_i)P(H_i).$$

La combinaison des deux propositions précédentes justifie la construction d'arbres pondérés.

Lorsqu'on a une partition de Ω en n hypothèses ou cas possibles H_i et que l'on sait calculer les $P(H_i)$ et les $P(A \mid H_i)$, on peut se poser le problème inverse : calculer $P(H_j \mid A)$ à l'aide des quantités précédentes. La solution est donnée par la formule suivante quelquefois appelée (abusivement) formule des probabilités des causes.

Proposition (formule de Bayes).
Soit A un évènement de probabilité non nulle. Si les évènements H_i $(1 \leqslant i \leqslant n)$ forment une partition de Ω et si aucune $P(H_i)$ n'est nulle, on a pour tout $j \in [\![1, n]\!]$:

$$P(H_j \mid A) = \frac{P(A \mid H_j)P(H_j)}{\sum_{i=1}^{n} P(A \mid H_i)P(H_i)}.$$

La formule se généralise au cas d'une partition dénombrable.

Indépendance

Indépendance de deux évènements

Soient A et B deux évènements de probabilité non nulle. Il arrive que la connaissance de la réalisation de A ne modifie pas notre information sur celle de B, autrement dit que $P(B \mid A) = P(B)$. C'est le cas par exemple lorsque l'on fait deux tirages avec remise et que la réalisation de A ne dépend que du résultat du premier tirage, celle de B que du deuxième. Symétriquement on aura dans cet exemple $P(A \mid B) = P(A)$. Cette remarque se généralise comme suit.

8. Ou formule des probabilités totales.

Proposition. *Si A et B sont des évènements de probabilité non nulle, les trois égalités suivantes sont équivalentes :*
(i) $P(B \mid A) = P(B)$,
(ii) $P(A \mid B) = P(A)$,
(iii) $P(A \cap B) = P(A)P(B)$.

D'autre part la relation (iii) est toujours vérifiée dans le cas dégénéré où $P(A) = 0$ ou $P(B) = 0$. En effet, on a alors à la fois $P(A)P(B) = 0$ et $0 \leqslant P(A \cap B) \leqslant \min\big(P(A), P(B)\big) = 0$ d'où $P(A \cap B) = 0$. Ainsi la relation (iii) est un peu plus générale que (i) et (ii). Elle a aussi sur les deux autres l'avantage de la symétrie d'écriture. C'est elle que l'on retient pour définir l'indépendance.

Définition. Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé. Deux évènements A et B de cet espace sont dits indépendants lorsque :

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

◁

Remarques.

1. Si A est un évènement tel que $P(A) = 0$ ou $P(A) = 1$, alors il est indépendant de tout évènement, *y compris de lui-même* (c'est le cas en particulier pour Ω et $\varnothing$).
2. Deux évènements *incompatibles* A et B avec $P(A) > 0$ et $P(B) > 0$ ne sont *jamais indépendants*. En effet $A \cap B = \varnothing$ implique $P(A \cap B) = 0$ or $P(A)P(B) \neq 0$.
3. L'indépendance de deux évènements A et B n'est pas une propriété intrinsèque aux évènements, elle est toujours relative au modèle $(\Omega, \mathcal{F}, P)$ que l'on a choisi. Voici un petit exercice pour l'illustrer. Une urne contient 6 boules numérotées de 1 à 6. On en tire une au hasard et on considère les évènements :

   $A = \{\text{le numéro de la boule tirée est pair}\}$
   $B = \{\text{le numéro de la boule tirée est multiple de 3}\}$.

   Vérifier que A et B sont indépendants. On recommence l'expérience avec une urne contenant 7 boules numérotées de 1 à 7. Les évènements A et B définis comme précédemment restent-ils indépendants ?

◁

Proposition. *Si A et B sont indépendants, il en est de même pour les paires d'évènements A et B^c, A^c et B, A^c et B^c.*

Indépendance mutuelle

On se propose de généraliser la notion d'indépendance à plus de deux évènements. Examinons d'abord la situation suivante.

Exemple. Alice et Bernard jettent chacun une pièce de monnaie équilibrée. On définit les évènements :

$$\begin{aligned} A &= \{\text{la pièce d'Alice indique pile}\},\\ B &= \{\text{la pièce de Bernard indique pile}\},\\ C &= \{\text{les deux pièces indiquent le même résultat}\}. \end{aligned}$$

En représentant une issue élémentaire de cette expérience par un couple dont la première composante donne le résultat du lancer d'Alice et la deuxième le résultat du lancer de Bernard, on peut prendre $\Omega = \{f,p\}^2$, $\mathcal{F} = \mathcal{P}(\Omega)$ et P l'équiprobabilité sur Ω. Alors $A = \{(p,f),(p,p)\}$, $B = \{(f,p),(p,p)\}$ et $C = \{(f,f),(p,p)\}$.

Il est clair que $P(A) = P(B) = P(C) = 2/4 = 1/2$. D'autre part :

$$P(A \cap B) = P(\{(p,p)\}) = \frac{1}{4} = P(A)P(B)$$

et de même $P(B \cap C) = 1/4 = P(B)P(C)$, $P(C \cap A) = 1/4 = P(C)P(A)$. Ainsi les évènements A, B, C sont *deux à deux indépendants*.

D'autre part $P(C \mid A \cap B) = 1$ car $A \cap B = \{(p,p)\}$. Donc la connaissance de la réalisation *simultanée* de A *et* B modifie notre information sur C. La notion d'indépendance deux à deux n'est donc pas suffisante pour traduire l'idée intuitive d'indépendance de plusieurs évènements. Ceci motive la définition suivante. ◁

Définition. Trois évènements A, B, C sont dits mutuellement indépendants lorsqu'ils vérifient les quatre conditions :

$$\begin{aligned} P(A \cap B) &= P(A)P(B),\\ P(B \cap C) &= P(B)P(C),\\ P(C \cap A) &= P(C)P(A),\\ P(A \cap B \cap C) &= P(A)P(B)P(C). \end{aligned}$$

◁

Avec cette définition de l'indépendance des évènements A, B et C on vérifie[9] que $P(A \mid B) = P(A)$, $P(A \mid B \cap C) = P(A)$, ainsi que toutes les égalités qui s'en déduisent par permutation sur les lettres A, B, C. On peut généraliser cette définition comme suit.

9. Sous réserve d'existence des probabilités conditionnelles.

Définition. Les n évènements $A_1, \ldots, A_n$ sont dits mutuellement indépendants si pour toute sous-famille $A_{i_1}, \ldots, A_{i_k}$ avec $1 \leqslant i_1 < \ldots < i_k \leqslant n$,

$$P(A_{i_1} \cap \ldots \cap A_{i_k}) = P(A_{i_1}) \times \cdots \times P(A_{i_k}).$$

◁

L'indépendance mutuelle implique évidemment l'indépendance deux à deux et la réciproque est fausse comme le montre l'exemple p. 27 : si A, B, C étaient mutuellement indépendants, ils devraient vérifier $P(C \mid A \cap B) = P(C)$. Un petit exercice de dénombrement permet de se convaincre que l'indépendance mutuelle est beaucoup plus forte que l'indépendance deux à deux. Par exemple l'indépendance deux à deux de dix évènements peut se décrire explicitement en écrivant 45 égalités, leur indépendance mutuelle nécessite pas moins de 1013 égalités (pourquoi ?).

Dans toute la suite, lorsque nous parlerons d'une famille de plusieurs évènements indépendants sans autre précision, nous sous-entendrons systématiquement *mutuellement* indépendants.

Proposition. *Si $\{A_1, \ldots, A_n\}$ est une famille de n évènements indépendants, toute famille obtenue en remplaçant certains des A_i par leur complémentaire est encore indépendante.*

Définition (indépendance d'une suite d'évènements). Une suite infinie d'évènements est dite indépendante si toute sous-suite finie comportant au moins deux évènements est formée d'évènements mutuellement indépendants. ◁

Remarque. Compte-tenu de la proposition précédente, on en déduit immédiatement que si $(A_i)_{i\in\mathbb{N}}$ est une suite indépendante d'évènements, toute suite formée en remplaçant certains des A_i (éventuellement tous) par leur complémentaire est encore indépendante. ◁

Le modèle des épreuves répétées indépendantes

Considérons une suite d'*épreuves* réalisées dans les mêmes conditions expérimentales, par exemple tirages avec remise dans la même urne, lancers successifs d'un dé, etc. Il est alors raisonnable de supposer que les résultats de tout sous-ensemble fini d'épreuves n'ont aucune influence sur ceux des autres épreuves.

Définition. On dit que les épreuves sont indépendantes si toute suite $(A_i)_{i\geqslant 1}$ où la réalisation de chaque A_i est déterminée uniquement par le résultat de la i^{e} épreuve est une suite indépendante d'évènements. ◁

Exemple. On réalise une suite d'épreuves indépendantes. Chaque épreuve résulte en un succès avec probabilité $p \in]0,1[$ ou en un échec avec probabilité $q = 1 - p$. Quelle est la probabilité des évènements suivants ?

a) $A = \{$Au moins un succès au cours des n premières épreuves$\}$.

b) $B = \{$Exactement k succès au cours des n premières épreuves$\}$.

c) $C = \{$Toutes les épreuves donnent un succès$\}$. ◁

Solution. Notons pour tout $i \geqslant 1$: $S_i = \{$succès à la i^{e} épreuve$\}$, S_i^c est alors l'évènement $\{$échec à la i^{e} épreuve$\}$.

a) $A = \bigcup_{i=1}^n S_i$, d'où $A^c = \bigcap_{i=1}^n S_i^c$. Les S_i^c étant indépendants, on en déduit :

$$P(A^c) = \prod_{i=1}^n P(S_i^c) = (1-p)^n = q^n,$$

d'où $P(A) = 1 - q^n$.

b) Traitons d'abord le cas $0 < k < n$. L'évènement B est la réunion disjointe de tous les évènements du type :

$$B_I = \Big(\bigcap_{i\in I} S_i\Big) \cap \Big(\bigcap_{j\in J} S_j^c\Big),$$

où I est une partie de cardinal k de $[\![1, n]\!]$ et J son complémentaire dans $[\![1, n]\!]$. L'ensemble d'indices I représente un choix possible des k épreuves donnant un succès, les autres épreuves indexées par J donnant alors un échec. En considérant tous les choix possibles de l'ensemble I, il y en a $\binom{n}{k}$, on obtient une partition de B par les B_I. Par indépendance des épreuves, pour tout I on a :

$$P(B_I) = \prod_{i\in I} P(S_i) \times \prod_{j\in J} P(S_j^c) = p^k q^{n-k}.$$

On voit ainsi que $P(B_I)$ ne dépend pas de I. On en déduit :

$$P(B) = \sum_{\substack{I\subset[\![1,n]\!] \\ \operatorname{card} I = k}} P(B_I) = \binom{n}{k} p^k q^{n-k}.$$

La vérification de la validité de la formule $P(B) = \binom{n}{k} p^k q^{n-k}$ dans les cas $k = 0$ et $k = n$ est laissée au lecteur.

c) Pour $n \geqslant 1$, soit $C_n = \{$succès aux n premières épreuves$\}$. Clairement C est inclus dans C_n donc $0 \leqslant P(C) \leqslant P(C_n)$. En utilisant l'indépendance des S_i on obtient :

$$P(C_n) = P\Big(\bigcap_{i=1}^n S_i\Big) = \prod_{i=1}^n P(S_i) = p^n.$$

donc pour tout $n \geqslant 1$, $0 \leqslant P(C) \leqslant p^n$. En faisant tendre n vers $+\infty$, on en déduit $P(C) = 0$. □

L'algorithme du rejet

Nous allons voir maintenant comment utiliser des épreuves répétées pour « fabriquer » des probabilités conditionnelles. La méthode exposée ci-dessous est le coeur de l'algorithme du rejet utilisé pour la simulation de variables ou de vecteurs aléatoires.

Considérons une suite d'épreuves répétées où la i^{e} épreuve consiste à générer un point aléatoire M_i du plan [10]. Formellement, on a une suite $(M_i)_{i\geqslant 1}$ de points aléatoires telle que pour tout borélien A, les évènements $\{M_i \in A\} = \{\omega \in \Omega;\ M_i(\omega) \in A\}$ sont indépendants et de même probabilité ne dépendant que de A. On fixe un borélien B de $\mathbb{R}^2$ tel que $P(M_1 \in B) > 0$ et on définit l'indice $T(\omega)$ comme le premier indice i tel que $M_i(\omega)$ soit dans B, voir figure 6. Si $M_i(\omega)$ n'appartient à B pour aucun $i \in \mathbb{N}^*$, on pose $T(\omega) = +\infty$. On note M_T le point aléatoire ainsi obtenu (c'est le premier M_i à tomber dans B, mais attention son numéro est lui même aléatoire et peut changer avec ω). On convient par exemple que $M_T(\omega) = (0, 0)$ dans le cas particulier où $T(\omega) = +\infty$. On se propose de calculer $P(M_T \in A)$ pour A borélien quelconque de $\mathbb{R}^2$. Nous supposons implicitement ici que les écritures

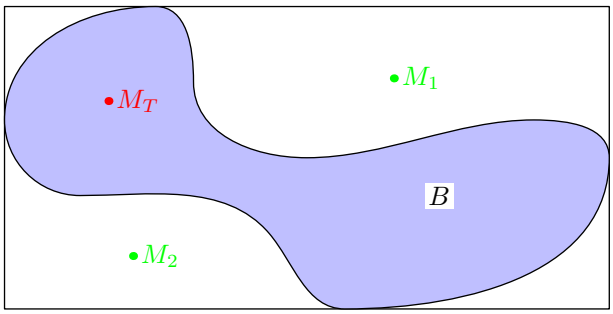

FIGURE 6 – Simulation par rejet, ici $T(\omega) = 3$

$\{M_i \in A\}$, $\{M_T \in A\}$, $\{T = k\}$ représentent effectivement des évènements [11].

Comme les évènements $\{T = k\}$, $k \in \overline{\mathbb{N}}^* = \mathbb{N}^* \cup \{+\infty\}$ réalisent une partition dénombrable de Ω, on a la décomposition en union dénombrable disjointe :

$$\forall A \in \mathrm{Bor}(\mathbb{R}^2), \quad \{M_T \in A\} = \bigcup_{k\in\overline{\mathbb{N}}^*} \{M_T \in A \text{ et } T = k\}.$$

Pour tout $k \in \mathbb{N}^*$, l'évènement $\{M_T \in A \text{ et } T = k\}$ peut s'écrire comme suit :

$$\{M_T \in A \text{ et } T = k\} = \{M_k \in A \text{ et } T = k\} = \{M_k \in A\} \cap \underbrace{\bigcap_{i<k} \{M_i \notin B\} \cap \{M_k \in B\}}_{\{T=k\}}$$

10. Tout ce qui suit est valable aussi avec des points aléatoires de $\mathbb{R}^d$ au lieu de $\mathbb{R}^2$.

11. On peut justifier ces suppositions avec la notion de mesurabilité qui est liée à l'étude des variables ou vecteurs aléatoires.

Ceci nous permet d'exploiter l'indépendance des épreuves répétées pour obtenir :

$$P(M_T \in A \text{ et } T = k) = (1 - P(M_1 \in B))^{k-1} P(M_k \in A \cap B).$$

En notant $p = P(M_1 \in B)$, $q = 1 - p$ et en appliquant cette formule avec $A = \mathbb{R}^2$, on obtient :

$$P(M_T \in \mathbb{R}^2 \text{ et } T = k) = P(T = k) = q^{k-1} p.$$

En rappelant que p est strictement positif, d'où $0 < q < 1$ et en sommant pour tous les $k \in \mathbb{N}^*$, il vient :

$$P(T \in \mathbb{N}^*) = \sum_{k \in \mathbb{N}^*} P(T = k) = \sum_{k \in \mathbb{N}^*} q^{k-1} p = \frac{p}{1 - q} = 1.$$

Par conséquent, $P(T = +\infty)$ est nulle, tout comme $P(M_T \in A \text{ et } T = +\infty)$ qu'elle majore. Donc pour calculer $P(M_T \in A)$ par σ-additivité à partir de la décomposition en union dénombrable de $\{M_T \in A\}$, on ne perd rien en laissant tomber le terme $P(M_T \in A \text{ et } T = +\infty)$. Ainsi :

$$\begin{aligned} P(M_T \in A) = \sum_{k \in \mathbb{N}^*} P(M_T \in A \text{ et } T = k) &= \sum_{k \in \mathbb{N}^*} q^{k-1} P(M_1 \in A \cap B) \\ &= \frac{P(M_1 \in A \cap B)}{1 - q} \\ &= \frac{P(M_1 \in A \cap B)}{P(M_1 \in B)}. \end{aligned}$$

Finalement, pour tout ensemble borélien A du plan $\mathbb{R}^2$:

$$P(M_T \in A) = P(M_1 \in A \mid M_1 \in B).$$

Application. Supposons que les M_i soient choisis au hasard suivant la loi uniforme sur un rectangle C (simulation informatique très facile) contenant B, comme celui de la figure 6, autrement dit que pour tout entier i,

$$P(M_i \in A) = \frac{\lambda_2(A \cap C)}{\lambda_2(C)} = \frac{\text{aire}(A \cap C)}{\text{aire}(C)}.$$

Alors pour tout borélien A,

$$P(M_T \in A) = \frac{\dfrac{\lambda_2(A \cap B \cap C)}{\lambda_2(C)}}{\dfrac{\lambda_2(B \cap C)}{\lambda_2(C)}} = \frac{\lambda_2(A \cap B)}{\lambda_2(B)},$$

ce qui montre que le point aléatoire M_T suit la loi uniforme sur B. ◁

Variables aléatoires réelles

Variables aléatoires et lois

Souvent ce qui intéresse l'observateur dans une expérience aléatoire est une quantité numérique. Par exemple dans de nombreux jeux, lorsqu'on jette deux dés on s'intéresse seulement à la somme de points indiqués par les faces supérieures après immobilisation des dés. Un tas d'autres informations que l'on pourrait relever (vitesse initiale des dés, hauteur maximale de la trajectoire, nombre de rebonds, intensité du bruit du choc, etc) sont délibérément ignorées. On peut aussi mentionner toutes les observations statistiques où on relève une valeur numérique : température, pression atmosphérique, hauteur de pluie dans des observations météorologiques, taille ou masse d'un individu, nombre d'accidents sur une portion d'autoroute, hauteur de crue d'un fleuve, consommation de carburant d'un avion pour un vol donné, durée de vie d'un composant électrique, etc. Si l'expérience aléatoire est modélisée par l'espace probabilisé $(\Omega, \mathcal{F}, P)$ où les évènements élémentaires ω représentent les issues élémentaires de l'expérience, la quantité numérique à laquelle on s'intéresse peut donc se représenter sous la forme $X(\omega)$, où X est une certaine application définie sur Ω et à valeurs dans $\mathbb{R}$. C'est ce que l'on appelle communément une *variable aléatoire*. Il y a ici une difficulté pédagogique liée au langage traditionnel, puisque cette « variable » est en fait une fonction. Il y a une autre difficulté, mathématique celle-là, qui est que toute application $\Omega \to \mathbb{R}$ ne mérite pas forcément le nom de variable aléatoire. Commençons par clarifier ce point.

Puisque la quantité X observée est « variable » on souhaite généralement décrire en un certain sens cette variabilité. On est ainsi naturellement conduit à étudier la probabilité que X prenne une valeur donnée ou plus généralement « tombe » dans un sous-ensemble donné de $\mathbb{R}$. Idéalement, on aimerait connaître pour tout borélien B de $\mathbb{R}$, la probabilité $P(X \in B)$, notation abrégée pour $P(\{\omega \in \Omega\ ;\ X(\omega) \in B\})$. C'est bien sûr totalement utopique puisque l'on ne sait même pas décrire explicitement tous les boréliens de $\mathbb{R}$. Heureusement, la tribu borélienne de $\mathbb{R}$ étant engendrée par certaines familles d'intervalles, il suffira en pratique de pouvoir calculer les $P(X \in I)$, pour I de la forme $]a, b]$ ou pour I de la forme $]-\infty, b]$. Il est commode ici d'introduire l'inverse ensembliste de X en notant pour tout sous-ensemble C de $\mathbb{R}$:

$$X^{-1}(C) = \{\omega \in \Omega\ ;\ X(\omega) \in C\}.$$

Il importe de remarquer que cette écriture « X^{-1} » ne suppose aucunement la bijectivité de X et utilise seulement le fait que X est une application de Ω dans $\mathbb{R}$. On s'intéresse donc aux quantités $P(X^{-1}(B))$. Encore faut-il qu'elles aient un sens, autrement dit que les sous-ensembles $X^{-1}(B)$ de Ω appartiennent à l'ensemble de définition de la fonction d'ensembles P, c'est-à-dire à la tribu $\mathcal{F}$ des évènements observables. On réservera donc le nom de variable aléatoire aux applications X ayant cette propriété :

$$\text{pour tout borélien } B \text{ de } \mathbb{R}, \quad X^{-1}(B) \in \mathcal{F}.$$

On démontre en théorie de la mesure, qu'une condition suffisante (et évidemment nécessaire) pour la réalisation de cette propriété est :

$$\text{pour tout intervalle } I \text{ de } \mathbb{R}, \quad X^{-1}(I) \in \mathcal{F}.$$

C'est cette version avec les images réciproques d'intervalles qui est retenue comme définition d'une variable aléatoire dans le programme de l'agrégation interne de mathématiques (2015). La notation I ci-dessus désigne un intervalle *quelconque* de $\mathbb{R}$, mais on pourrait tout aussi bien se limiter à l'un des familles d'intervalles qui engendrent la tribu borélienne, les deux plus utiles étant la famille des intervalles de la forme $]-\infty, b]$ où b est un réel quelconque et la famille des intervalles $]a, b]$, où a et $b > a$ sont des réels quelconques.

L'inverse ensembliste défini ci-dessus a des propriétés très commodes : il commute avec les réunions et les intersections de familles quelconques d'ensembles et avec le passage au complémentaire. En utilisant ces propriétés, on vérifie facilement la proposition suivante préalable à la définition de la loi d'une variable aléatoire.

Proposition. *Soit X une variable aléatoire sur $(\Omega, \mathcal{F})$ et P une probabilité sur $(\Omega, \mathcal{F})$. La fonction d'ensembles $P_X = P \circ X^{-1}$ définie sur* $\mathrm{Bor}(\mathbb{R})$ *par*

$$\forall B \in \mathrm{Bor}(\mathbb{R}), \quad P_X(B) = P\big(X^{-1}(B)\big) = P(X \in B)$$

est une probabilité sur $\big(\mathbb{R}, \mathrm{Bor}(\mathbb{R})\big)$.

Définition. La probabilité P_X ainsi définie est appelée *loi* de la variable aléatoire X *sous* P. ◁

C'est cette *mesure de probabilité* P_X qui *décrit précisément la variabilité* de X. Pour des raisons évidentes de simplification pédagogique, cet objet fondamental est assez systématiquement ignoré dans les manuels élémentaires de probabilités où la notion de loi n'est pas précisément définie ou alors confondue avec un tableau « valeurs possibles - probabilités d'obtention » dans le cas d'une loi discrète, voire avec une densité dans le cas d'une loi absolument continue.

Remarque. Dans les problèmes usuels de probabilités, on travaille souvent avec un seul $(\Omega, \mathcal{F}, P)$ et on se contente alors de l'appellation *loi de* X. Il n'en va pas de même en statistique où l'on met généralement en concurrence *plusieurs* modèles $(\Omega, \mathcal{F}, P_\theta)$, où θ est un paramètre inconnu et où on se propose de choisir un de ces modèles au vu des valeurs $X(\omega)$ observées. C'est là que l'appellation loi de X *sous* P_θ s'impose. Pour donner un exemple simple, considérons le problème du sondage d'un échantillon de 500 personnes avant le second tour d'une élection présidentielle opposant le candidat A au candidat B. Ici θ est la proportion *inconnue* d'électeurs votant A dans la population totale. Puisqu'elle est inconnue, cette proportion θ est susceptible *a priori* de prendre n'importe quelle valeur dans $\mathbb{Q} \cap [0, 1]$. On confronte

donc une infinité de modèles différents. Si X est le nombre de personnes interrogées favorables à A, la loi de X sous P_θ est la loi binomiale [12] $\mathrm{Bin}(500, \theta)$. ◁

Une fonctionnalité importante de la notion de loi d'une variable aléatoire est qu'elle permet de transformer le calcul des probabilités d'un espace $(\Omega, \mathcal{F}, P)$ qui peut être d'une très grande complexité, en un calcul de probabilités dans l'espace plus familier bâti sur la droite réelle, à savoir $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}), P_X)$. Si on ne s'intéresse qu'à des évènements dont la réalisation n'est déterminée que par les valeurs prises par X, on peut même complètement oublier $(\Omega, \mathcal{F}, P)$ et ne plus travailler qu'avec P_X.

Loi d'une variable aléatoire discrète

Nous donnons maintenant une formule explicite pour les lois d'une famille importante de variables aléatoires, les variables aléatoires discrètes.

Définition (variable aléatoire discrète). On appelle *variable aléatoire discrète* sur $(\Omega, \mathcal{F})$, toute application $X : \Omega \to \mathbb{R}$ vérifiant les deux conditions suivantes.

(*i*) L'ensemble des images $X(\Omega) = \{X(\omega) \ ; \ \omega \in \Omega\}$ est une partie au plus dénombrable de $\mathbb{R}$. On peut donc numéroter ses éléments par des indices entiers [13]

$$X(\Omega) = \{x_0, x_1, \ldots, x_k, \ldots\}.$$

(*ii*) Pour tout $x \in X(\Omega)$, $X^{-1}(\{x\}) \in \mathcal{F}$.

◁

Il est facile de vérifier que toute variable aléatoire discrète est une variable aléatoire réelle puisque pour tout intervalle I, $X(\Omega) \cap I$ est au plus dénombrable et la tribu $\mathcal{F}$ est stable par unions dénombrables. La réciproque est fausse.

Proposition. *Soient $(\Omega, \mathcal{F}, P)$ un espace probabilisé et X une variable aléatoire* discrète *sur $(\Omega, \mathcal{F})$. La loi de X sous P est la probabilité*

$$P_X = \sum_{x \in X(\Omega)} P(X = x)\delta_x.$$

La somme ci-dessus désigne une somme finie ou une série selon que $X(\Omega)$ est un ensemble fini ou dénombrable.

12. En fait c'est une loi hypergéométrique (tirages sans remise), mais en raison du théorème de convergence de la loi hypergéométrique vers la loi binomiale, on peut la remplacer en pratique par une binomiale si l'échantillon est de grande taille.

13. Pour tous les exemples classiques que nous rencontrerons, il est possible de les numéroter de manière *croissante* : $x_0 < x_1 < x_2 \ldots$. Mais ce n'est pas toujours le cas, car $X(\Omega)$ peut être par exemple, l'ensemble des décimaux (ou des rationnels) de $[0, 1]$.

Définition (loi discrète sur $\mathbb{R}$). On appelle loi discrète sur $\mathbb{R}$, toute mesure ponctuelle μ sur $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$ qui est aussi une probabilité. Une telle loi admet donc une représentation sous la forme

$$\mu = \sum_{i\in I} p_i \delta_{x_i} : \quad \mathrm{Bor}(\mathbb{R}) \to [0,1],\ B \longmapsto \mu(B) = \sum_{x_i \in B} p_i,$$

où I est un ensemble d'indices au plus dénombrable, $(x_i)_{i\in I}$ une famille de nombres réels et $(p_i)_{i\in I}$ une famille de réels positifs de somme $\sum_{i\in I} p_i = 1$. ◁

Avec cette définition il est clair que la loi, sous n'importe quelle probabilité P sur $(\Omega, \mathcal{F})$, d'une variable aléatoire discrète est toujours une loi discrète. La réciproque est fausse : une variable aléatoire réelle X peut avoir, sous une certaine probabilité P sur $(\Omega, \mathcal{F})$, une loi discrète tout en ayant un ensemble de valeurs $X(\Omega)$ infini non dénombrable, par exemple $X(\Omega) = \mathbb{R}$. L'exemple non dégénéré le plus simple qu'on puisse en donner est le suivant. Prenons $\Omega = \mathbb{R}$, $\mathcal{F} = \mathrm{Bor}(\mathbb{R})$ et $X : \omega \mapsto \omega$ l'identité sur $\mathbb{R}$. Munissons $(\Omega, \mathcal{F})$ de la mesure de probabilité $P = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ (les connaisseurs auront reconnu une modélisation un peu alambiquée du jeu de pile ou face en un coup avec une pièce équilibrée). Alors la loi de X, $P_X = P \circ X^{-1} = P$ est discrète. Mais comme $X(\Omega) = \mathbb{R}$ la variable aléatoire X n'est pas discrète.

Nous avons déjà rencontré des lois discrètes lors de l'introduction de la mesure μ associée aux observations d'une série statistique $x_1, \ldots, x_n$:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i} = \sum_{j\in J} f_j\, \delta_{x_{(j)}}.$$

μ est la loi de toute variable aléatoire Y vérifiant $P(Y = x_{(j)}) = f_j$ pour tout $j \in J$.

Remarques générales sur les lois

Remarque. Deux variables aléatoires peuvent *avoir même loi sans être égales*. Par exemple considérons le jet de deux dés, l'un bleu et l'autre rouge. Notons X le nombre de points indiqué par le dé bleu et Y celui du rouge. Les variables aléatoires X et Y sont définies sur le même espace probabilisé $\Omega = \{1,2,3,4,5,6\}^2$ muni de l'équiprobabilité. On a $X(\Omega) = Y(\Omega) = [\![1,6]\!]$ et :

$$\forall k \in [\![1,6]\!], \quad P(X = k) = \frac{1}{6}, \quad P(Y = k) = \frac{1}{6}.$$

Donc X et Y ont même loi : $P_X = P_Y = \sum_{k=1}^{6} \frac{1}{6}\delta_k$. Pour autant, on n'a pas l'égalité des variables aléatoires X et Y qui signifierait $X(\omega) = Y(\omega)$ *pour tout* $\omega \in \Omega$ (égalité de deux applications). Autrement dit, en lançant deux dés on obtiendrait *à coup sûr* un double. Par contre nous pouvons considérer *l'évènement* $\{X = Y\}$ dont la réalisation n'est pas certaine et calculer sa probabilité :

$$P(X = Y) = P\Big(\bigcup_{k=1}^{6} \{(X,Y) = (k,k)\}\Big) = \frac{6}{36} = \frac{1}{6}.$$

On en déduit : $P(X \neq Y) = 5/6$. ◁

Remarque. Deux variables aléatoires peuvent avoir même loi en étant *définies sur des espaces probabilisés différents* $(\Omega, \mathcal{F}, P)$ et $(\Omega', \mathcal{F}', P')$. Prenons par exemple pour X les points indiqués par un dé équilibré et posons $Z = 0$ si X est pair, $Z = 1$ si X est impair. Les variables aléatoires X et Z peuvent être définies sur $\Omega = \{1, 2, 3, 4, 5, 6\}$ muni de la tribu $\mathcal{P}(\Omega)$ et de l'équiprobabilité P sur Ω. La loi de Z est $P_Z = \frac{1}{2}(\delta_0 + \delta_1)$. Prenons maintenant $\Omega' = \{-1, 1\}$, muni de la tribu $\mathcal{P}(\Omega')$ et de l'équiprobabilité P' sur Ω' et posons pour $\omega' \in \Omega'$, $Z'(\omega') = (1 + \omega')/2$. Alors la loi de Z' est $P'_{Z'} = \frac{1}{2}(\delta_0 + \delta_1) = P_Z$. ◁

Remarquons que si X et Y sont définies sur des espaces probabilisés différents, *il n'y a pas* d'évènement $\{X = Y\}$, pas plus que de variable aléatoire « $X + Y$ ». Essayez d'en écrire la définition explicite pour vous en convaincre.

Remarque. Pour toute probabilité Q sur $\big(\mathbb{R}, \mathrm{Bor}(\mathbb{R})\big)$, il existe au moins un espace probabilisé $(\Omega, \mathcal{F}, P)$ et une variable aléatoire réelle X sur $(\Omega, \mathcal{F})$ dont la loi sous P soit égale à Q. Il suffit de prendre $\Omega = \mathbb{R}$, $\mathcal{F} = \mathrm{Bor}(\mathbb{R})$ et pour X l'application identité de $\mathbb{R} \to \mathbb{R}$. En prenant $P = Q$, on a clairement $P_X = Q$. Bien entendu, il y a une infinité d'autres solutions à ce problème. ◁

Il y a donc identité entre les mesures de probabilité sur $\big(\mathbb{R}, \mathrm{Bor}(\mathbb{R})\big)$ et les lois des variables aléatoires réelles. Comme nous savons caractériser une probabilité sur $\big(\mathbb{R}, \mathrm{Bor}(\mathbb{R})\big)$ par sa fonction de répartition, ceci va nous permettre de classifier les lois des variables aléatoires réelles.

Fonction de répartition (f.d.r.)

Définition (f.d.r. d'une variable aléatoire). Soient $(\Omega, \mathcal{F}, P)$ un espace probabilisé et X une variable aléatoire sur $(\Omega, \mathcal{F})$. On appelle fonction de répartition (f.d.r.) de X, la fonction F_X définie sur $\mathbb{R}$ par :

$$\forall x \in \mathbb{R}, \quad F_X(x) = P_X(\,]-\infty, x]) = P(X \leqslant x).$$

◁

La fonction F_X est la fonction de répartition de la probabilité P_X. Elle ne dépend donc que de la loi[14] de X.

Remarque. Deux variables aléatoires de même loi ont même fonction de répartition. ◁

Proposition. *La fonction de répartition F_X d'une variable aléatoire X est croissante sur $\mathbb{R}$, avec limite 0 en $-\infty$ et 1 en $+\infty$. Elle est continue à droite et limitée à gauche en tout point de $\mathbb{R}$. En notant $F_X(x-)$ la limite à gauche de F_X au point x,*

$$\forall x \in \mathbb{R}, \quad P(X = x) = F(x) - F(x-).$$

14. Il serait plus correct, mais plus long, de parler de f.d.r. de la loi de X ou même de f.d.r. de la loi de X sous P.

La fonction de répartition d'une variable aléatoire caractérise sa loi, autrement dit : $F_X = F_Y$ si et seulement si les variables aléatoires X et Y ont même loi.

On peut calculer les probabilités d'appartenance de X à un intervalle à l'aide de F_X comme suit :

$$\begin{aligned}
P(a < X \leqslant b) &= F_X(b) - F_X(a),\\
P(a \leqslant X \leqslant b) &= F_X(b) - F_X(a-),\\
P(a \leqslant X < b) &= F_X(b-) - F_X(a-),\\
P(a < X < b) &= F_X(b-) - F_X(a),\\
P(X \leqslant a) &= F_X(a),\\
P(X < a) &= F_X(a-),\\
P(X > b) &= 1 - F_X(b),\\
P(X \geqslant b) &= 1 - F_X(b-).
\end{aligned}$$

Médianes et quantiles

Définition (médianes). Soit X une variable aléatoire réelle sur $(\Omega, \mathcal{F}, P)$. On appelle médiane de X (sous P) ou de la loi de X sous P, tout réel m vérifiant $P(X \leqslant m) \geqslant 1/2$ *et* $P(X \geqslant m) \geqslant 1/2$. ◁

Définition (quantiles). Soit X une variable aléatoire réelle sur $(\Omega, \mathcal{F}, P)$, de fonction de répartition F. Soit $u \in]0, 1[$. On appelle u-quantile de X (ou de sa loi sous P) le plus petit réel t tel que $F(t) \geqslant u$. ◁

Preuve (Existence du u-quantile). Notons

$$I(u) = \{t \in \mathbb{R} \ ; \ F(t) \geqslant u\}.$$

Comme $\lim_{t\to+\infty} F(t) = 1 > u$, il existe au moins un $t \in \mathbb{R}$ tel que $F(t) > u$ donc l'ensemble de réels $I(u)$ n'est pas vide. D'autre part, comme F est croissante sur $\mathbb{R}$, pour tout $s \in I(u)$ et tout $t > s$, $F(t) \geqslant F(s) \geqslant u$, donc $t \in I(u)$. Ceci montre que pour tout $s \in I(u)$, l'intervalle $[s, +\infty[$ est inclus dans $I(u)$. On en déduit que $I(u)$ est un intervalle de la forme $]a, +\infty[$ ou $[a, +\infty[$ avec $a = -\infty$ ou $a \in \mathbb{R}$. Si a est fini, par continuité à droite de F au point a, $F(a) \geqslant u$, donc $a \in I(u)$. Si $a = -\infty$, cela signifierait que $F(t) \geqslant u > 0$ pour tout t réel, en contradiction avec le fait que $F(t)$ tend vers 0 quand t tend vers $-\infty$. Par conséquent, $I(u) = [a, +\infty[$, avec $a \in \mathbb{R}$. Ce réel a est donc le plus petit élément de $I(u)$, donc le plus petit réel t vérifiant $F(t) \geqslant u$. L'existence du u-quantile est ainsi complètement justifiée. □

La figure 7 illustre les trois cas pour la détermination graphique du u-quantile q.

Proposition. *Le $1/2$-quantile d'une variable aléatoire X est une médiane de X.*

Preuve. Notons m_0 le u-quantile pour $u = 1/2$. Il vérifie donc à la fois $F(m_0) \geqslant 1/2$, c'est à dire $P(X \leqslant m_0) \geqslant 1/2$, et pour tout $s < m_0$, $P(X \leqslant s) < 1/2$. Montrons alors que $P(X \geqslant m_0) \geqslant 1/2$ en distinguant les deux cas $P(X < m_0) \leqslant 1/2$ et $P(X < m_0) > 1/2$. Le premier cas est évident puisque

$$P(X \geqslant m_0) = 1 - P(X < m_0) \geqslant 1 - 1/2 = 1/2.$$

Supposons maintenant que $P(X < m_0) > 1/2$. Comme $P(X < m_0)$ est la limite à gauche de F au point m_0,

$$P(X < m_0) = \lim_{\substack{s \to m_0 \\ s < m_0}} F(s).$$

Comme nous l'avons noté ci-dessus, pour tout $s < m_0$, $F(s) \leqslant 1/2$, donc par passage à la limite à gauche en m_0, $P(X < m_0) \leqslant 1/2$. Ainsi, le cas $P(X < m_0) > 1/2$ ne peut se produire et ceci achève la vérification de l'inégalité $P(X \geqslant m_0) \geqslant 1/2$.

Nous venons de montrer que m_0 est une médiane de X et donc que toute variable aléatoire X a toujours au moins une médiane. □

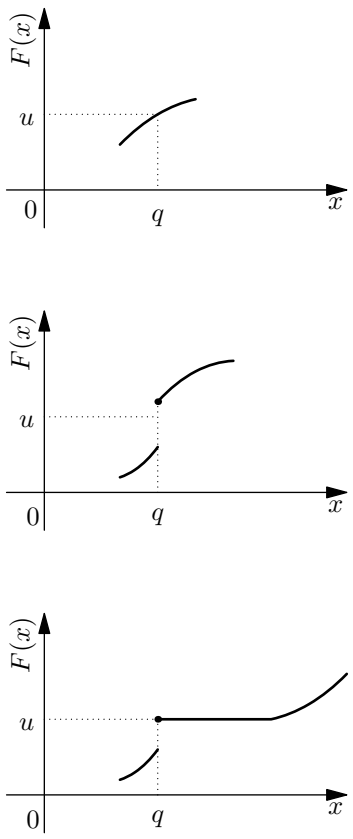

FIGURE 7 – Détermination graphique du u-quantile q

On peut se demander comment détecter au vu de F s'il y a plus d'une médiane. Nous avons déjà vu dans le cas d'une série statistique (et donc pour une variable aléatoire discrète à support fini) que s'il y a un palier horizontal d'altitude $1/2$ sur la représentation graphique de F, alors il y a une infinité de médianes qui sont les éléments d'un intervalle $[a, b]$ appelé intervalle médian. Voyons maintenant s'il y a une situation analogue pour une variable aléatoire quelconque.

Commençons par donner une caractérisation de la (des) médiane(s) de X à l'aide de la f.d.r. F. La condition $P(X \leqslant m) \geqslant 1/2$ se réécrit naturellement $F(m) \geqslant 1/2$. En rappelant que $F(m-)$ désigne la limite à gauche de F au point m, la condition $P(X \geqslant m) \geqslant 1/2$ se réécrit $1 - F(m-) \geqslant 1/2$ soit encore $F(m-) \leqslant 1/2$. En résumé :

$$m \text{ est médiane de } X \text{ si et seulement si } \quad F(m-) \leqslant \frac{1}{2} \leqslant F(m).$$

De cette équivalence on déduit en particulier que si $F(m) = 1/2$ ou si $F(m-) = 1/2$, m est une médiane. Si $F(m) < 1/2$ ou si $F(m-) > 1/2$, m n'est pas une médiane.

Notons

$$S = \left\{ t \in \mathbb{R} \; ; \; F(t) = \frac{1}{2} \right\}.$$

Nous allons distinguer trois cas, correspondant à la figure 7 avec $u = 1/2$.

Cas 1. $S = \varnothing$. Notons $m_0 = F^{-1}(1/2)$. Nous avons vu que m_0 est une médiane de X. Donc $F(m_0-) \leqslant 1/2 \leqslant F(m_0)$. Comme $S = \varnothing$, la dernière inégalité est en fait stricte. Soit $t \neq m_0$. Si $t > m_0$, $F(t-) \geqslant F(m_0) > 1/2$ donc t n'est pas une médiane de X. Si $t < m_0$, $F(t) \leqslant F(m_0-) \leqslant 1/2$, donc $F(t) \leqslant 1/2$ et comme $S = \varnothing$, $F(t) < 1/2$ donc t n'est pas une médiane. Par conséquent dans le cas 1, m_0 est l'unique médiane de X.

Cas 2. S est un singleton $\{t_0\}$. Par définition de m_0, on a $m_0 \leqslant t_0$ donc par croissance de F, $F(m_0) \leqslant F(t_0) = 1/2$. Mais comme $F(m_0) \geqslant 1/2$ par continuité à droite de F au point m_0, on en déduit que $F(m_0) = 1/2$, donc $m_0 \in S$, donc $m_0 = t_0$. Soit $t \neq m_0$. Si $t > m_0$, alors pour tout $s \in]m_0, t[$, $F(s) > 1/2$ donc $F(t-) > 1/2$ et t n'est pas une médiane de X. Si $t < m_0$, $F(t) \leqslant F(m_0) = 1/2$ et comme $S = \{m_0\}$, on a en fait $F(t) < 1/2$, d'où t n'est pas une médiane de X.

Cas 3. S a plus d'un élément. Si m et m' sont deux solutions distinctes ($m < m'$) de l'équation $F(t) = 1/2$, par croissance de F, $F(s) = 1/2$ pour tout $s \in [m, m']$. On en déduit que S est un intervalle de $\mathbb{R}$. Notons a sa borne inférieure et b sa borne supérieure. On a déjà vu que $a = F^{-1}(1/2) > -\infty$. D'autre part $b < +\infty$ car F tendant vers 1 en $+\infty$, il existe un $s \in \mathbb{R}$ tel que $1/2 < F(s) \leqslant 1$. Nous savons déjà que $a = F^{-1}(1/2)$ est une médiane de X. Pour tout $t \in]a, b[$, $F(t) = 1/2$, donc t est une médiane de X et $F(b-) = 1/2$. Cette dernière égalité implique que b est aussi une médiane de X. Si $t < a$, $F(t) \leqslant 1/2$ mais comme $S \subset [a, b]$, $t \notin S$ d'où $F(t) < 1/2$ donc t n'est pas une médiane de X. Si $t > b$, pour tout $s \in]b, t[$, $F(s) > 1/2$, d'où $F(t-) > 1/2$ donc t n'est pas une médiane.

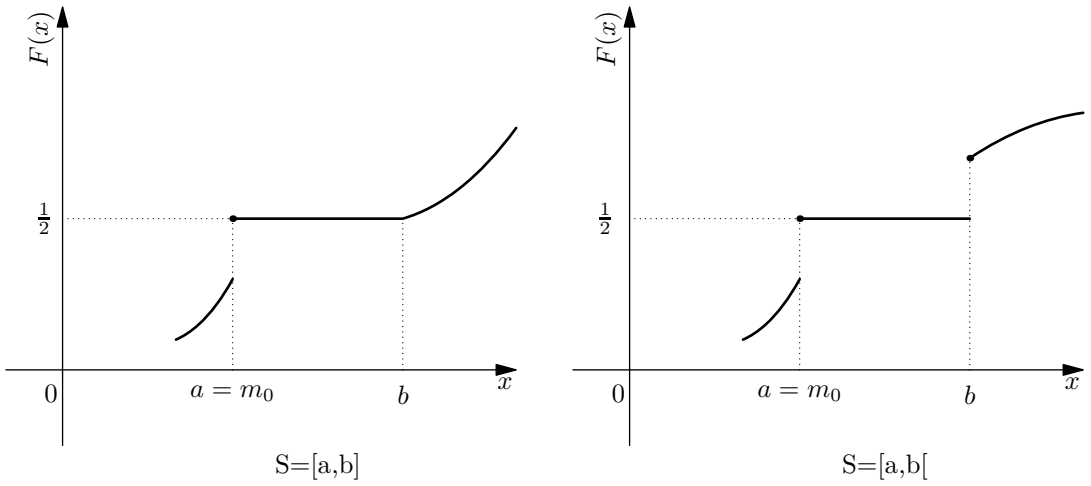

FIGURE 8 – Cas 3 : $S = [a, b]$ ou $S = [a, b[$

Nous pouvons conclure cette étude comme suit.

Proposition. *Soit X une variable aléatoire réelle et F sa fonction de répartition. Si l'équation $F(t) = 1/2$ a au plus une solution, X a une unique médiane qui est le quantile $m_0 = F^{-1}(1/2)$. Sinon l'ensemble des médianes de X est un intervalle $[a, b]$ avec $-\infty < a = m_0 < b < +\infty$, qui est la fermeture de l'intervalle S des solutions de l'équation $F(t) = 1/2$. Cet intervalle $[a, b]$ est appelé* intervalle médian *de X.*

Lois à densité

Si l'on veut esquisser une classification sommaire des lois des variables aléatoires, on peut commencer par les partager entre les lois à fonction de répartition continue sur $\mathbb{R}$ et les lois à fonction de répartition non continue[15] sur $\mathbb{R}$. On parle plus simplement de *lois continues* ou encore *lois diffuses* dans le premier cas et de lois non continues ou non diffuses dans le deuxième. Dans la famille des lois non continues, nous connaissons déjà la sous-famille des lois discrètes. Dans la famille des lois continues, une importante sous-famille est celle des *lois à densité* que nous allons examiner maintenant.

La loi d'une variable aléatoire X est à densité f si pour tout intervalle de $\mathbb{R}$, la probabilité d'appartenance de X à cet intervalle peut s'écrire comme l'intégrale de f sur cet intervalle. L'apparente simplicité de cette définition informelle est trompeuse. Dans le cadre de ces compléments, nous nous restreignons à l'intégration au sens de Riemann et il se trouve que cette notion n'est pas totalement satisfaisante pour les besoins de la théorie des probabilités. L'intégrale de Lebesgue donnerait une notion plus générale de densité permettant, entre autres, de caractériser les lois à densité

15. Rappelons que l'ensemble des points de discontinuité d'une f.d.r. quelconque est au plus dénombrable.

comme celles dont la f.d.r. est *absolument continue*, une notion bien plus restrictive que la continuité, que nous examinerons plus loin. La définition de densité plus restrictive que nous donnons ci-dessous est néanmoins suffisante pour la plupart des cas pratiques.

Définition (densité de probabilité). On appelle densité de probabilité sur $\mathbb{R}$ toute fonction f vérifiant

a) f est définie et positive sur $\mathbb{R}\backslash K$, où K est une partie finie (éventuellement vide) de $\mathbb{R}$;

b) f est Riemann intégrable sur tout intervalle $[a,b] \subset \mathbb{R}\backslash K$;

c) l'intégrale généralisée de f sur $]-\infty,+\infty[$ converge et

$$\int_{-\infty}^{+\infty} f(t)\,\mathrm{d}t = 1.$$

◁

Si f est une fonction positive définie seulement sur un intervalle $]a,b[$ de $\mathbb{R}$ et telle que $\int_a^b f(t)\,\mathrm{d}t = 1$, on peut en faire une densité en la prolongeant à tout $\mathbb{R}$ en posant $f(t) = 0$ pour $t \notin]a,b[$. Voici quatre exemples simples de densités :

$$f_1(t) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(t); \qquad f_2(t) = \frac{1}{2\sqrt{t}}\mathbf{1}_{]0,1]}(t);$$

$$f_3(t) = \mathrm{e}^{-t}\mathbf{1}_{[0,+\infty[}(t); \qquad f_4(t) = \frac{1}{\pi(1+t^2)}.$$

Remarque. La définition de f_2 repose sur un *abus d'écriture* d'usage courant. En effet il y a en toute rigueur un problème pour calculer $f_2(t)$ lorsque $t \leqslant 0$, puisqu'alors il nous faut former le produit de l'expression $1/(2\sqrt{t})$ non définie par 0. La convention adoptée est que si la formule de calcul d'une fonction contient le produit d'une indicatrice par une expression non définie lorsque cette indicatrice est nulle, le produit vaut 0 dans ce cas. Ceci permet de considérer que la « définition » de f_2 comme ci-dessus est un raccourci d'écriture commode pour :

$$f_2(t) = \begin{cases} \frac{1}{2\sqrt{t}} & \text{si } t \in]0,1], \\ 0 & \text{si } t \notin]0,1]. \end{cases}$$

◁

Définition. Soient $(\Omega, \mathcal{F}, P)$ un espace probabilisé et X une variable aléatoire réelle sur $(\Omega, \mathcal{F})$. La loi de X sous P a pour densité f si :

$$\forall a \in \mathbb{R},\ \forall b \geqslant a, \qquad P(X \in]a,b]) = \int_a^b f(t)\,\mathrm{d}t.$$

On dit aussi par abus de langage que X a pour densité f (lorsqu'il n'y a pas ambiguïté sur P). ◁

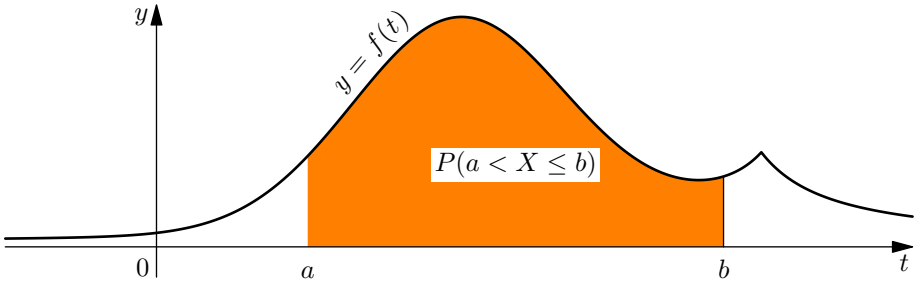

FIGURE 9 – $P(a < X \leqslant b) = \displaystyle\int_a^b f(t)\,\mathrm{d}t$ pour X de densité f

Remarque. Il est clair d'après cette définition que si Y est une autre variable aléatoire ayant même loi que X (donc mêmes probabilités d'appartenance aux intervalles), elle a aussi pour densité f. D'autre part, il n'y a pas unicité de la densité d'une variable aléatoire. Par exemple $g_1 = \mathbf{1}_{[0,1]}$ et $g_2 = \mathbf{1}_{]0,1[}$ sont deux densités de probabilité qui donnent les mêmes intégrales : $\int_a^b g_1(t)\,\mathrm{d}t = \int_a^b g_2(t)\,\mathrm{d}t$ pour toute paire de réels a et b. Ces deux fonctions peuvent chacune être prise comme densité de la loi uniforme sur $[0, 1]$. ◁

Examinons maintenant les relations entre densité (lorsqu'elle existe) et fonction de répartition (qui elle, existe toujours) d'une variable aléatoire.

Proposition. *Si la variable aléatoire X a pour densité f, sa fonction de répartition F vérifie :*

a) $\forall x \in \mathbb{R},\ F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$;

b) F *est continue sur* $\mathbb{R}$;

c) si f *est continue au point* x_0, *alors* F *est dérivable en* x_0 *et* $F'(x_0) = f(x_0)$.

Corollaire. *Si la variable aléatoire X a pour densité f, les égalités suivantes sont vérifiées pour tous $a, b \in \mathbb{R}$ tels que $a < b$:*

$$P(a < X < b) = P(a < X \leqslant b) = P(a \leqslant X < b) = P(a \leqslant X \leqslant b) = \int_a^b f(t)\,\mathrm{d}t,$$

$$P(X < a) = P(X \leqslant a) = \int_{-\infty}^{a} f(t)\,\mathrm{d}t,$$

$$P(X > b) = P(X \geqslant b) = \int_b^{+\infty} f(t)\,\mathrm{d}t.$$

Remarques.

1. D'après *b)* toute variable aléatoire à densité a une fonction de répartition continue. La réciproque *est fausse* : il existe des lois à fonction de répartition continue sans densité.

2. Par ailleurs si X a une densité, sa fonction de répartition n'est pas forcément dérivable en tout point. Par exemple la densité f_2 ci-dessus a pour fonction de répartition associée $F_2(x) = \sqrt{x}\mathbf{1}_{]0,1]}(x) + \mathbf{1}_{]1,+\infty[}(x)$; cette écriture condensée signifie que $F_2(x)$ est nul sur $\mathbb{R}^-$, vaut $\sqrt{x}$ entre 0 et 1 et reste constant égal à 1 sur $]1, +\infty[$. F_2 est dérivable en tout point sauf en 0 et en 1.

◁

La proposition suivante donne une règle pratique permettant de trouver la densité (lorsqu'elle existe !) à partir de la fonction de répartition dans les cas les plus courants.

Proposition. *On suppose que la fonction de répartition F de X est C^1 par morceaux au sens suivant : F est continue sur $\mathbb{R}$ et dérivable sur $\mathbb{R}$ privé (éventuellement) d'un ensemble fini de points $a_1 < \ldots < a_n$. Sur chacun des intervalles ouverts $]-\infty, a_1[$, $]a_i, a_{i+1}[$ $(1 \leqslant i < n)$, $]a_n, +\infty[$, la dérivée f de F est continue. Alors X a pour densité f.*

L'idée sous-jacente à cette proposition est que si la f.d.r. F d'une variable aléatoire X est *suffisamment régulière*, alors la loi de X a une densité. On a vu d'autre part que si la loi de X est à densité, sa f.d.r. F est continue. La question qui surgit alors naturellement est : peut-on caractériser l'existence d'une densité par la régularité de la fonction de répartition ? La réponse est oui et la notion de régularité de la f.d.r. qui équivaut à l'existence d'une densité est celle d'*absolue continuité*.

Définition. Une fonction $F : \mathbb{R} \to \mathbb{R}$ est dite *absolument continue* sur $\mathbb{R}$, si pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour tout $n \geqslant 1$ et pour toute famille finie de n intervalles $[a_k, b_k] \subset \mathbb{R}$, $1 \leqslant k \leqslant n$, d'intérieurs $]a_k, b_k[$ deux à deux disjoints,

$$\sum_{k=1}^{n}(b_k - a_k) < \delta \implies \sum_{k=1}^{n}|F(b_k) - F(a_k)| < \varepsilon.$$

◁

Notons que si F est croissante sur $\mathbb{R}$, les valeurs absolues sont superflues dans la dernière inégalité ci-dessus. La continuité absolue sur $\mathbb{R}$ implique évidemment la continuité uniforme sur $\mathbb{R}$ (se restreindre à $n = 1$). Par ailleurs toute f.d.r. continue sur $\mathbb{R}$ est *uniformément* continue sur $\mathbb{R}$ (exercice !). Nous donnerons ci-dessous un exemple de f.d.r. continue sur $\mathbb{R}$ mais non absolument continue. La continuité absolue est donc une propriété strictement plus forte que la continuité uniforme.

Proposition. *La fonction de répartition F d'une loi à densité f est absolument continue sur $\mathbb{R}$.*

La preuve de cette proposition est immédiate dans le cas particulier où f est bornée sur $\mathbb{R}$: $f \leqslant C$ pour une constante C. Il suffit alors de prendre $\delta = \varepsilon/C$. Pour le cas général, cf. [9] pp. 219–220.

Voyons maintenant un exemple de variable aléatoire ayant une fonction de répartition *continue mais sans densité*.

Exemple (une v.a. à f.d.r. continue sans densité). On génère les chiffres décimaux d'un nombre réel X de $[0, 1]$ en effectuant une suite infinie de tirages avec remise d'une boule dans une urne qui en contient 9 numérotées de 0 à 8. Autrement dit, X peut s'écrire :

$$X = \sum_{j=1}^{+\infty} \frac{X_j}{10^j},$$

les variables aléatoires X_j étant indépendantes et de même loi uniforme sur $[\![0, 8]\!]$. L'indépendance des variables aléatoires sera vue ultérieurement. Pour notre propos, il suffit de raisonner en termes d'épreuves répétées pour disposer de l'indépendance de la famille d'évènements $(\{X_j = i_j\})_{j\in\mathbb{N}*}$ pour toute suite d'entiers $(i_j)_{j\in\mathbb{N}*}$.

On pourrait certes objecter que cet exemple n'est pas réaliste puisqu'en pratique on sera bien obligé de s'arrêter au bout d'un nombre fini de tirages et la variable aléatoire ainsi générée ne sera qu'une variable aléatoire discrète, donc sans densité mais avec fonction de répartition discontinue. Il faut néanmoins noter qu'il en va de même pour toute simulation de la loi uniforme sur $[0, 1]$ par tirage de boules (en rajoutant dans la même urne une boule numérotée 9) ou par la fonction `rand` d'une calculatrice. Plus généralement lorsqu'on simule une variable aléatoire de façon informatique on n'en obtient jamais qu'une approximation discrète.

Concernant X défini formellement ci-dessus par une série de variables aléatoires, on remarque que cette série converge sur tout Ω. En effet, pour ω quelconque, chaque $X_j(\omega)$ étant dans $[\![0, 8]\!]$, la série $\sum_{j=1}^{+\infty} X_j(\omega)10^{-j}$ est à termes positifs et son terme général est majoré par 8×10^{-j} qui est le terme général d'une série géométrique de raison $1/10$ convergente.

Montrons que la fonction de répartition F de X est continue sur $\mathbb{R}$. Comme X est positive, $F(t) = 0$ pour tout $t < 0$. D'autre part, $X(\omega) \leqslant \sum_{j=1}^{+\infty} \frac{8}{10^j} = \frac{8}{9}$ pour tout $\omega \in \Omega$, d'où $F(\frac{8}{9}) = P(X \leqslant \frac{8}{9}) = 1$, donc F est constante égale à 1 sur $[\frac{8}{9}, +\infty[$. Pour prouver que F est continue sur $\mathbb{R}$, il nous suffit donc de vérifier que pour tout réel $x \in [0, \frac{8}{9}]$, $P(X = x) = 0$, autrement dit, que la fonction de répartition F n'a aucun saut dans cet intervalle. Soit donc x quelconque dans cet intervalle et $(x_j)_{j\geqslant 1}$ la suite des chiffres de son développement décimal illimité propre. Notons que si x admet un développement décimal illimité impropre, c'est-à-dire ne comportant que des 9 à partir d'un certain rang, ce développement ne peut être généré par la suite $(X_j(\omega))_{j\geqslant 1}$ puisqu'il n'y a aucune boule numérotée 9 dans l'urne. Nous avons ainsi pour tout $\omega \in \Omega$, équivalence entre « $X(\omega) = x$ » et « pour tout $j \geqslant 1$, $X_j(\omega) = x_j$ »,

ce qui se traduit par l'égalité d'évènements :

$$\{X = x\} = \bigcap_{j\in\mathbb{N}^*} \{X_j = x_j\}.$$

Notons $A_n = \bigcap_{j=1}^{n}\{X_j = x_j\}$. Pour tout $n \geqslant 1$, $\{X = x\}$ est inclus dans A_n, d'où $0 \leqslant P(X = x) \leqslant P(A_n)$. Or par indépendance des X_j,

$$P(A_n) = \prod_{j=1}^{n} P(X_j = x_j) \leqslant 9^{-n},$$

car $P(X_j = x_j)$ vaut $1/9$ si $x_j \in [\![0, 8]\!]$ et 0 si $x_j = 9$. Par conséquent, $P(A_n)$ tend vers zéro quand n tend vers l'infini et en passant à la limite dans l'encadrement $0 \leqslant P(X = x) \leqslant P(A_n)$, on en déduit que $P(X = x) = 0$. Ceci établit la continuité de F.

Pour prouver que la loi de X n'a pas de densité, on démontre que F n'est pas *absolument continue*. C'est un peu plus délicat que ce qui précède et nous renvoyons le lecteur intéressé à [9] pp. 240–243. ◁

Lois classiques

On trouve dans tous les manuels de probabilités un catalogue plus ou moins riche de lois dites classiques. On ne mentionnera ici que les incontournables en indiquant brièvement ce qui fait leur intérêt.

Lois classiques discrètes

Lois de Bernoulli

Une variable aléatoire X suit la loi de Bernoulli de paramètre p, ($p \in [0, 1]$) si

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p = q.$$

Notation : $X \sim \text{Bern}(p)$. P_X est la mesure ponctuelle $q\delta_0 + p\delta_1$.

Les cas $p = 0$ et $p = 1$ sont dégénérés puisqu'alors X est presque-sûrement constante.

L'exemple fondamental de variable aléatoire suivant une loi de Bernoulli est celui des v.a. discrètes $X = \mathbf{1}_A$, où A est un évènement : $\mathbf{1}_A \sim \text{Bern}(P(A))$.

Loi uniforme sur un ensemble fini de réels

Soit $x_1, \ldots, x_n$, n réels distincts. Une variable aléatoire X suit la loi uniforme sur $\{x_1, \ldots, x_n\}$ si P_X est l'équiprobabilité sur cet ensemble :

$$\forall k \in [\![1, n]\!], \quad P(X = x_k) = \frac{1}{n}.$$

Notation : $X \sim \text{Unif}\{x_1, \ldots, x_n\}$. La loi P_X est la mesure ponctuelle $\frac{1}{n}\sum_{k=1}^{n} \delta_{x_k}$.

Par exemple, le nombre de points indiqué par un dé équilibré suit la loi uniforme sur $[\![1, 6]\!]$.

Lois binomiales

Une variable aléatoire X suit la loi binomiale de paramètres n et p, $n \in \mathbb{N}^*$ et $p \in [0,1]$, notation $X \sim \mathrm{Bin}(n,p)$, si

$$\forall k \in [\![0,n]\!], \qquad P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

La formule ci-dessus définit bien une loi de probabilité puisque les $\binom{n}{k} p^k (1-p)^{n-k}$ sont positifs et :

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = \big(p + (1-p)\big)^n = 1^n = 1,$$

en appliquant la formule du binôme de Newton (d'où le nom de la loi). La loi binomiale $\mathrm{Bin}(n,p)$ est la loi du nombre de succès obtenus en une suite de n épreuves répétées indépendantes avec pour chaque épreuve une probabilité de succès p.

De même, soit $A_1, \ldots, A_n$ une famille d'évènements mutuellement indépendants tous de même probabilité p. La variable aléatoire $S_n = \sum_{i=1}^n \mathbf{1}_{A_i}$ suit la loi $\mathrm{Bin}(n,p)$.

Lois hypergéométriques

La loi hypergéométrique intervient dans les tirages sans remise. Dans une population totale de N individus, dont M ont une certaine caractéristique $\mathcal{C}$, on prélève au hasard un échantillon de n individus (tirage sans remise). On note X le nombre d'individus ayant la caractéristique $\mathcal{C}$ dans cet échantillon. Un peu de dénombrement nous amène à :

$$P(X=k) = \frac{\binom{M}{k} \times \binom{N-M}{n-k}}{\binom{N}{n}} \quad \text{si} \quad \begin{cases} 0 \leqslant k \leqslant M, \\ 0 \leqslant n-k \leqslant N-M. \end{cases}$$

La loi de X définie par cette égalité et ces conditions s'appelle loi hypergéométrique de paramètres N, M et n. Notation : $X \sim \mathrm{Hypg}(N,M,n)$. Le paramètre N est l'effectif de la population totale, M celui de la sous-population d'intérêt et n la taille de l'échantillon observé.

Pour une taille d'échantillon n *fixée*, plus N et M sont grands, moins les tirages sans remise diffèrent des tirages avec remise. La justification théorique de cette affirmation est fournie par le théorème suivant.

Théorème (convergence de l'hypergéométrique vers la binomiale). *On suppose que quand N tend vers $+\infty$, $M = M(N)$ tend vers $+\infty$ en vérifiant la condition :*

$$\lim_{N \to +\infty} \frac{M}{N} = p \quad \textit{avec} \quad 0 < p < 1.$$

Alors, n restant fixé, la loi hypergéométrique $\mathrm{Hypg}(N,M,n)$ *« converge » vers la loi binomiale* $\mathrm{Bin}(n,p)$ *au sens suivant. Si la suite de v.a. $(X_N)_{N \geqslant 1}$ vérifie pour tout N $X_N \sim \mathrm{Hypg}(N,M,n)$ et si Y est une v.a. de loi binomiale* $\mathrm{Bin}(n,p)$*, alors :*

$$\forall k \in [\![0,n]\!], \qquad \lim_{N \to +\infty} P(X_N = k) = P(Y=k),$$

autrement dit :

$$\forall k \in [\![0, n]\!], \qquad \lim_{N\to+\infty} \frac{\binom{M}{k} \times \binom{N-M}{n-k}}{\binom{N}{n}} = \binom{n}{k} p^k (1-p)^{n-k}.$$

Lois géométriques

Une variable aléatoire X suit la loi géométrique de paramètre $p \in]0, 1[$, si

$$\forall k \in \mathbb{N}^*, \quad P(X = k) = (1-p)^{k-1} p.$$

Notation : $X \sim \text{Geom}(p)$. La loi $\text{Geom}(p)$ est la mesure ponctuelle $\sum_{k\in\mathbb{N}^*} q^{k-1} p \delta_k$, où $q = 1 - p$.

La loi géométrique est la loi du *temps d'attente* (mesuré en nombre d'épreuves) du premier succès dans une suite infinie d'épreuves répétées indépendantes avec même probabilité de succès $p \in]0, 1[$. On vérifie facilement que

$$\forall n \in \mathbb{N}^*, \quad P(X > n) = (1-p)^n.$$

Les lois géométriques se caractérisent par la propriété d'*absence de mémoire en temps discret* au sens suivant : la v.a. discrète X à valeurs dans $\mathbb{N}$ suit une loi géométrique si et seulement si

$$\forall n \in \mathbb{N}, \forall k \in \mathbb{N}, \quad P(X > n + k \mid X > n) = P(X > k).$$

Lois géométriques tronquées

Une variante de la loi géométrique qui a l'avantage de ne pas recourir à l'outil des séries et de prendre en compte la durée toujours finie du temps d'observation est la loi géométrique tronquée, cf. le document *Ressources pour la classe de première générale et technologique, Statistiques et probabilités*, juin 2011. Il y a cette fois 2 paramètres $p \in]0, 1[$ probabilité de succès lors d'une épreuve et n durée maximale d'observation comptée en nombre d'épreuves. Si au bout de n épreuves on n'a toujours pas obtenu de succès, on arrête l'observation et on donne à X la valeur 0. On en déduit immédiatement que :

$$P(X = 0) = (1-p)^n, \quad \text{et} \quad \forall k \in [\![1, n]\!], \quad P(X = k) = (1-p)^{k-1} p.$$

Lois de Poisson

On dit qu'une variable aléatoire X suit la loi de Poisson de paramètre $\alpha > 0$ si

$$\forall k \in \mathbb{N}, \quad P(X = k) = \frac{\mathrm{e}^{-\alpha} \alpha^k}{k!}.$$

Notation : $X \sim \text{Pois}(\alpha)$. La loi $\text{Pois}(\alpha)$ est la mesure ponctuelle $\sum_{k\in\mathbb{N}} \frac{\mathrm{e}^{-\alpha}\alpha^k}{k!} \delta_k$.

On modélise souvent par une variable aléatoire suivant une loi de Poisson des nombres de réalisations d' « évènements rares » : accidents sur une portion de route pendant une durée donnée, tremblements de terre dans une région, fautes sur une page imprimée, chutes de météorites dans une région donnée, erreurs de numérotation téléphonique, nombre de clients dans une file d'attente à un instant donné, centenaires dans la population d'un village, etc.

Une des raisons de l'importance de cette loi est le théorème de convergence de la loi binomiale vers la loi de Poisson.

Théorème (convergence de la loi binomiale vers la loi de Poisson).
Si $(p_n)_{n\geqslant 1}$ *est une suite de réels de* $[0,1]$ *vérifiant*

$$np_n \xrightarrow[n\to+\infty]{} \alpha \in]0,+\infty[,$$

alors pour tout $k \in \mathbb{N}$,

$$\binom{n}{k} p_n^k (1-p_n)^{n-k} \xrightarrow[n\to+\infty]{} \frac{\mathrm{e}^{-\alpha}\alpha^k}{k!}.$$

Ce théorème sert de justification théorique à la règle pratique suivante : lorsque n est « grand » et np « petit », on peut remplacer la loi $\mathrm{Bin}(n,p)$ par la loi $\mathrm{Pois}(\alpha)$ où $\alpha = np$. En général on considère que n de l'ordre de quelques centaines et np de l'ordre de quelques unités donnent une bonne approximation. Sous cette forme, cette règle relève plus de la cuisine que des mathématiques, même si les comparaisons numériques la supportent fortement. Pour des résultats plus précis, voir les exercices 3.18 et 3.19 dans [8]. Sans entrer dans les détails techniques, on peut en extraire les inégalités suivantes facilement mémorisables. Notons S_n une v.a. suivant la loi binomiale de paramètres n et p et X une v.a. suivant la loi de Poisson de paramètre $\alpha = np$. On a alors les inégalités :

$$\forall k \geqslant 2\alpha - 1, \quad P(X > k) < P(X = k),$$

$$\forall k \geqslant 2\alpha + 1, \quad P(S_n = k) \leqslant P(X = k),$$

dont on déduit :

$$\forall k \geqslant 2\alpha + 1, \quad P(S_n > k) < P(X = k).$$

Ceci montre que lorsque $\alpha = np$ est petit, la loi $\mathrm{Pois}(\alpha)$ comme la loi $\mathrm{Bin}(n,p)$ ont leur masse concentrée sur un petit nombre de valeurs. Par exemple pour $\alpha = 3$, la probabilité pour chacune des variables X et S_n de tomber dans l'intervalle $[\![0,8]\!]$ est supérieure à 0,991 898.

Lois classiques à densité

Lois uniformes

Une variable aléatoire réelle X suit la loi uniforme sur $[a,b]$, $-\infty < a < b < +\infty$, si

$$\forall B \in \mathrm{Bor}(\mathbb{R}), \quad P(X \in B) = P_X(B) = \frac{\lambda_1([a,b] \cap B)}{\lambda_1([a,b])},$$

où λ_1 désigne la mesure de Lebesgue sur $\mathbb{R}$ (en particulier $\lambda_1([a,b]) = b-a$). Notation : $X \sim \mathrm{Unif}[a,b]$.

La loi $\mathrm{Unif}[a,b]$ a pour fonction de répartition F donnée par

$$F(x) = P_X(\,]-\infty,x]) = \frac{\lambda_1([a,b] \cap]-\infty,x])}{\lambda_1([a,b])} = \begin{cases} 0 & \text{si} \quad x < a; \\ \dfrac{x-a}{b-a} & \text{si} \quad a \leqslant x < b; \\ 1 & \text{si} \quad b \leqslant x. \end{cases}$$

Une densité de cette loi est $f = \frac{1}{b-a}\mathbf{1}_{[a,b]}$.

La loi uniforme sur un segment est l'extension de la notion d'équiprobabilité quand on passe du discret au continu. On la retrouve naturellement dans de nombreux problèmes de *probabilités géométriques*, par exemple le problème de la tige brisée. Dans les calculs sur la loi uniforme, il est conseillé d'utiliser chaque fois que possible les rapports de longueurs plutôt que les intégrales de f.

La loi uniforme sur $[0,1]$ est simulée par les calculatrices et les logiciels de calcul scientifique. Une des raisons de son importance réside dans le théorème suivant.

Théorème. *Si X est une variable aléatoire réelle de fonction de répartition continue strictement croissante F et si U est une variable aléatoire de loi uniforme sur $[0,1]$, alors la variable aléatoire $Y = F^{-1}(U)$ a même loi que X.*

Rappelons qu'avoir même loi que X ne signifie aucunement être égale à X. Ce théorème permet de réduire la simulation informatique de la loi de X à celle de U. Nous verrons ultérieurement que ce résultat s'étend à toutes les fonctions de répartition, sans hypothèse de continuité ni de croissance stricte, *via* une redéfinition de F^{-1} (fonction quantile).

Lois exponentielles

Soit a un réel strictement positif. Une variable aléatoire réelle X suit la loi exponentielle de paramètre a si elle admet pour densité

$$f : \mathbb{R} \to \mathbb{R}^+, \quad t \longmapsto f(t) = a\mathrm{e}^{-at}\mathbf{1}_{[0,+\infty[}(t).$$

La fonction de survie G d'une loi exponentielle a une expression particulièrement simple :

$$G : \mathbb{R} \to [0,1], \quad x \longmapsto G(x) = P(X > x) = 1 - F(x) = \begin{cases} 1 & \text{si } x \leqslant 0, \\ \mathrm{e}^{-ax} & \text{si } x > 0. \end{cases}$$

Les lois exponentielles sont souvent choisies pour modéliser des temps d'attente : temps d'attente à partir de maintenant du prochain tremblement de terre, du prochain faux numéro sur une ligne téléphonique, de la prochaine désintégration d'un atome de radium, durée de vie d'un composant électronique, temps entre deux arrivées consécutives dans une file d'attente, etc. La raison de ce choix est la propriété *d'absence de mémoire* en temps continu qui caractérise la famille des lois exponentielles.

Théorème (absence de mémoire).

i) Si la variable aléatoire X suit une loi exponentielle, elle vérifie la propriété d'absence de mémoire :

$$\forall s \in \mathbb{R}^+, \forall t \in \mathbb{R}^+, \quad P(X > t + s \mid X > t) = P(X > s).$$

ii) Réciproquement si une variable aléatoire X vérifie la propriété d'absence de mémoire ci-dessus, elle suit une loi exponentielle.

Lois gaussiennes

On dit qu'une variable aléatoire X suit la loi gaussienne ou normale $\mathfrak{N}(\mu, \sigma)$ si elle a pour densité la fonction :

$$f_{\mu,\sigma} : \mathbb{R} \longrightarrow \mathbb{R}^+ \quad t \longmapsto \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right).$$

La loi $\mathfrak{N}(0, 1)$ est appelée loi normale standard.

Ces lois jouent un rôle capital dans l'étude des lois limites de sommes de variables aléatoires indépendantes. Par exemple, théorème de de Moivre Laplace, si S_n suit la loi $\mathrm{Bin}(n, p)$, alors pour tout $x \in \mathbb{R}$, $P\big(S_n - np \leqslant x\sqrt{np(1-p)}\big)$ converge quand n tend vers l'infini vers $\Phi(x)$, où Φ est la f.d.r. de la loi gaussienne $\mathfrak{N}(0, 1)$.

Tous les calculs de probabilités concernant une variable aléatoire de loi $\mathfrak{N}(\mu, \sigma)$ peuvent se ramener à des calculs sur une variable de loi normale standard par centrage et normalisation.

Proposition. *Si la variable aléatoire X suit la loi $\mathfrak{N}(\mu, \sigma)$, $Y = (X - \mu)/\sigma$ suit la loi $\mathfrak{N}(0, 1)$. Autrement dit, toute v.a. gaussienne X de loi $\mathfrak{N}(\mu, \sigma)$ peut s'écrire $X = \mu + \sigma Y$ avec Y de loi $\mathfrak{N}(0, 1)$.*

Remarque. La famille des lois gaussiennes est *stable par transformations affines* : si X a pour loi $\mathfrak{N}(\mu, \sigma)$, alors pour tout $(a, b) \in \mathbb{R}^* \times \mathbb{R}$, la v.a. $aX + b$ est encore gaussienne, de loi $\mathfrak{N}(a\mu + b, |a|\sigma)$. ◁

La figure 10 illustre la signification du paramètre de position μ et du paramètre de dispersion σ pour la loi gaussienne $\mathfrak{N}(\mu, \sigma)$. Cette concentration de pratiquement toute la probabilité dans l'intervalle $[\mu-3\sigma, \mu+3\sigma]$ permet l'utilisation des lois gaussiennes pour modéliser des grandeurs aléatoires qui *a priori* prennent leurs valeurs

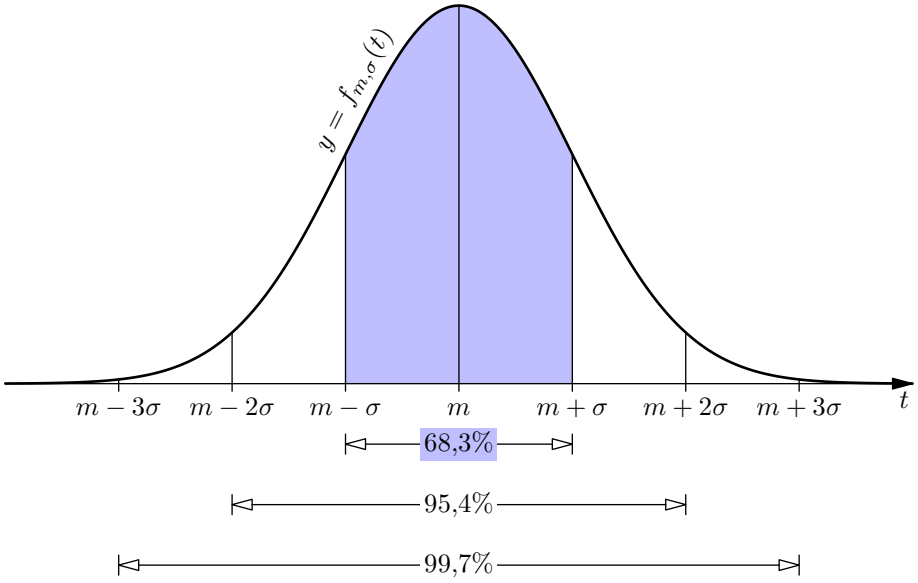

FIGURE 10 – Concentration de la loi $\mathfrak{N}(\mu, \sigma)$ autour de m

seulement dans un petit intervalle de $\mathbb{R}^+$: taille, poids, ..., même si théoriquement une variable gaussienne peut prendre toute valeur entre $-\infty$ et $+\infty$.

Il n'existe pas d'expression d'une primitive de la densité gaussienne $f_{\mu,\sigma}$ à l'aide des fonctions usuelles. Les valeurs de la fonction de répartition Φ de $\mathfrak{N}(0,1)$ sont tabulées. L'utilisation de cette table est progressivement supplantée par celle des logiciels de calcul scientifique qui incluent des programmes de calcul numérique des fonctions de répartition gaussiennes.

Lois de Cauchy

Une variable aléatoire X suit la loi de Cauchy (ou loi de Cauchy de paramètres 0 et 1) si elle admet pour densité :

$$f : \mathbb{R} \to \mathbb{R}^+, \quad t \longmapsto f(t) = \frac{1}{\pi(1+t^2)}.$$

Notation : $X \sim \mathrm{Cau}(0,1)$.

Cette loi est *symétrique*, ce qui signifie que X et $-X$ ont même loi, ceci résultant ici de la parité de f. La fonction de répartition F est donnée par :

$$F : \mathbb{R} \to [0,1], \quad x \longmapsto F(x) = \int_{-\infty}^{x} \frac{\mathrm{d}t}{\pi(1+t^2)} = \frac{1}{\pi}\Big(\frac{\pi}{2} + \arctan x\Big),$$

où $\arctan x$ est l'unique réel $y \in]-\pi/2, \pi/2[$ tel que $\tan y = x$.

Si $Y = a + bX$, avec X de loi $\mathrm{Cau}(0,1)$, $a \in \mathbb{R}$ et $b \in \mathbb{R}_+^*$, on dit encore que Y suit une loi de Cauchy, de paramètres (a,b), notation $Y \sim \mathrm{Cau}(a,b)$. La densité est

alors

$$f : \mathbb{R} \to \mathbb{R}^+, \quad t \longmapsto f_{a,b}(t) = \frac{1}{\pi b} \frac{1}{1 + \left(\frac{t-a}{b}\right)^2}.$$

Les lois de Cauchy fournissent des exemples simples de v.a. qui ne vérifient pas la loi forte des grands nombres car elles n'ont pas d'espérance. Une propriété remarquable de la loi $\mathrm{Cau}(0,1)$ est que si $X_1, \ldots, X_n$ suivent cette loi et sont indépendantes, alors leur moyenne arithmétique suit encore la loi $\mathrm{Cau}(0,1)$.

Espérance

Définitions « pragmatiques » de l'espérance

L'espérance $\mathrm{E}\, X$ d'une variable aléatoire X est, *lorsqu'elle existe*, la *moyenne des valeurs de cette variable, pondérées par leurs probabilités de réalisation*. On voit bien comment traduire cette « définition » *informelle* dans le cas le plus simple d'une v.a. discrète X prenant ses valeurs dans l'ensemble fini $X(\Omega) = \{x_1, \ldots, x_n\}$ en posant :

$$\mathrm{E}\, X = \sum_{k=1}^{n} x_k P(X = x_k).$$

On étend naturellement cette définition au cas des variables aléatoires discrètes dont l'ensemble des valeurs $X(\Omega) = \{x_k \ ; \ k \in \mathbb{N}^*\}$ est infini dénombrable en remplaçant la somme ci-dessus par la série

$$\mathrm{E}\, X = \sum_{k=1}^{+\infty} x_k P(X = x_k),$$

sous réserve que celle-ci soit convergente. Ici se présente une première difficulté. On veut pouvoir disposer de l'additivité de l'espérance des variables aléatoires discrètes : si les v.a. discrètes X et Y ont une espérance, alors la v.a. discrète $X + Y$ a une espérance et $\mathrm{E}(X + Y) = \mathrm{E}\, X + \mathrm{E}\, Y$. Ceci nous amène à restreindre la définition ci-dessus au cas où la série est *absolument* convergente, ce qui s'écrit aussi $\sum_{k=1}^{+\infty} |x_k| P(X = x_k) < +\infty$. Pour s'en convaincre, voir par exemple la preuve de cette propriété d'additivité dans [8], Prop. 5.2 où l'on utilise la propriété de sommabilité par paquets des séries doubles.

Voyons maintenant comment traduire la définition informelle de $\mathrm{E}\, X$ dans le cas d'une variable aléatoire à densité f. Remarquons en préalable que la fonction de répartition F de la v.a. X est continue sur $\mathbb{R}$ et qu'en conséquence pour tout réel x, $P(X = x) = 0$. L'utilisation d'une série comme ci-dessus est donc sans espoir. Pour dépasser cette difficulté adoptons, juste pour un instant, un point de vue de physicien. Pour tout réel x, la probabilité que la v.a. X tombe dans l'intervalle « infinitésimal » $[x, x + \mathrm{d}x]$ est en première approximation $f(x)\,\mathrm{d}x$. On obtient alors

$\mathrm{E}\,X$ en « sommant » pour tous x réels les quantités infinitésimales $xf(x)\,\mathrm{d}x$ et ceci nous amène à définir $\mathrm{E}\,X$ par l'intégrale généralisée

$$\mathrm{E}\,X = \int_{-\infty}^{+\infty} xf(x)\,\mathrm{d}x,$$

sous réserve qu'elle converge. Là aussi on impose par analogie avec le cas discret (une explication détaillée sortirait du cadre de cette introduction) la restriction de convergence *absolue* de cette intégrale.

À ce stade, nous avons obtenu les deux définitions de l'espérance qui figurent (implicitement) dans le programme de l'Agrégation interne de mathématiques et dans la réunion des programmes de l'enseignement secondaire, des classes de technicien supérieur et des classes préparatoires aux grandes écoles, réunion qui constitue le programme du CAPES externe de mathématiques. L'avantage de ces définitions est de permettre très rapidement le calcul des espérances des variables aléatoires suivant la plupart des lois usuelles, en évitant la longue marche d'approche imposée par la théorie moderne des probabilités, à savoir la construction d'une intégrale « abstraite » sur l'espace $(\Omega, \mathcal{F}, P)$, voir par exemple [7].

On peut néanmoins mentionner au moins deux inconvénients majeurs de ces définitions « pragmatiques » de l'espérance.

1. L'espérance n'est définie que pour des variables aléatoires discrètes ou à densité. Or on rencontre très naturellement dans des problèmes de modélisation des v.a. qui ne sont ni discrètes ni à densité (sans être forcément de loi singulière). Nous donnons ci-après un exemple dans le domaine de l'assurance.
2. Même si le programme de l'agrégation interne demande d'admettre l'additivité de l'espérance des v.a. à densité, il subsiste un sérieux problème, c'est que la somme de deux variables aléatoires X et Y à densité peut très bien ne pas être à densité (ni discrète). Dans ce cas comment interpréter $\mathrm{E}(X+Y)$? Si on décidait de la *définir* par l'égalité $\mathrm{E}(X+Y) = \mathrm{E}\,X + \mathrm{E}\,Y$, on se heurterait à un problème de cohérence car il y a une infinité de décompositions différentes d'une variable aléatoire Z en somme de deux variables aléatoires.

Exemple (prime d'assurance). Une compagnie d'assurance assure n clients contre un certain type de risque (incendie , accident automobile, etc.). Ces n clients ont le même profil et paieront la même prime annuelle. Si on note X_i le remboursement de sinistres annuel pour le client n$^{\mathrm{o}}$ i, les v.a. X_i ont donc même loi. En général le contrat prévoit un plafond M de remboursement (la compagnie ne pouvant supporter seule un coût exhorbitant qui la ruinerait). En pratique, on modélise souvent le coût d'un sinistre *sachant qu'il a lieu* par une loi à densité à support $[0, +\infty[$. Bien sûr, beaucoup d'assurés n'ont aucun sinistre dans l'année. La fonction de répartition F de la v.a. X_i a donc un saut à l'origine d'amplitude p_0 qui représente la probabilité pour un assuré de n'avoir aucun sinistre dans l'année [16]. Il y a un autre saut

16. S'il y a une franchise d'un montant m, ce saut est situé en m au lieu de 0 et représente la probabilité d'avoir un coût de sinistre inférieur à la franchise.

d'amplitude $1 - F(M-)$ au point M, cf. figure 11. La somme des sauts de F est inférieure à 1, donc la loi de X_i n'est pas discrète. F a deux points de discontinuité, donc la loi de X_i n'est pas à densité. Pourtant, l'espérance de X_i est le premier paramètre dont a besoin la compagnie pour calculer la prime qu'elle va demander à ses clients. Si cette prime est inférieure à cette espérance, la compagnie court à une ruine certaine. Si elle est trop supérieure à cette espérance, elle risque de perdre ses clients au profit de la concurrence. ◁

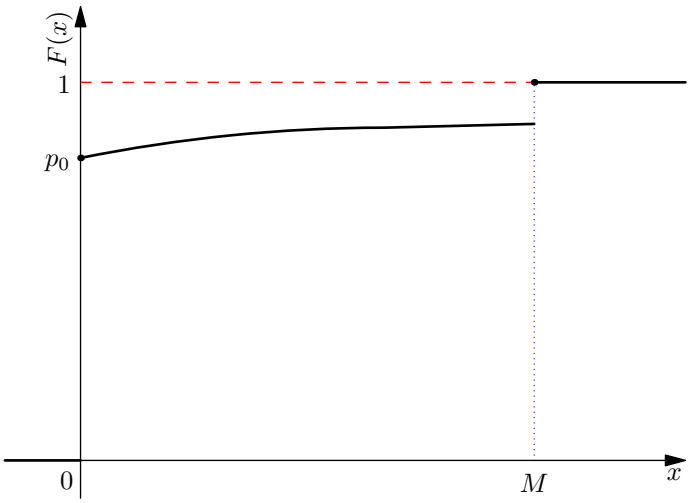

FIGURE 11 – Fonction de répartition du remboursement annuel d'un assuré.

Exemple (une somme de v.a. à densité, ni discrète ni à densité). On marque au hasard un point d'abscisse X sur le segment $[0, 1]$ (suivant la loi uniforme) et on note Y la distance de ce point à l'extrémité du segment la plus proche. Alors Y est à densité mais $X + Y$ n'est ni discrète ni à densité. Voyons cela en détail. $X : \Omega \to [0, 1]$ est une variable aléatoire de loi uniforme sur $[0, 1]$ et la v.a. Y s'exprime en fonction de X par

$$Y = \begin{cases} X & \text{si } 0 \leqslant X \leqslant 1/2, \\ 1 - X & \text{si } 1/2 < X \leqslant 1. \end{cases}$$

Vérifions d'abord que Y est à densité. Par construction, Y est définie sur Ω et à valeurs dans $[0, 1/2]$. Pour connaître sa loi, il suffit donc de calculer la restriction de sa fonction de répartition F à l'intervalle $[0, 1/2[$, puisque $F(t) = 0$ pour $t < 0$ et $F(t) = 1$ pour $t \geqslant 1/2$. En partitionnant Ω par les deux évènements complémen-

taires [17] $\{0 \leqslant X \leqslant 1/2\}$ et $\{1/2 < X \leqslant 1\}$, on obtient pour tout $t \in [0, 1/2[$:

$$\begin{aligned} P(Y \leqslant t) &= P(Y \leqslant t \text{ et } 0 \leqslant X \leqslant 1/2) + P(Y \leqslant t \text{ et } 1/2 < X \leqslant 1) \\ &= P(X \leqslant t \text{ et } 0 \leqslant X \leqslant 1/2) + P(1 - X \leqslant t \text{ et } 1/2 < X \leqslant 1) \\ &= t + P(1 - t \leqslant X \leqslant 1) \\ &= t + (1 - (1 - t)) = 2t. \end{aligned}$$

On reconnaît la f.d.r. de la loi Unif$[0, 1/2]$, donc Y a pour densité $2\,\mathbf{1}_{[0,1/2]}$.

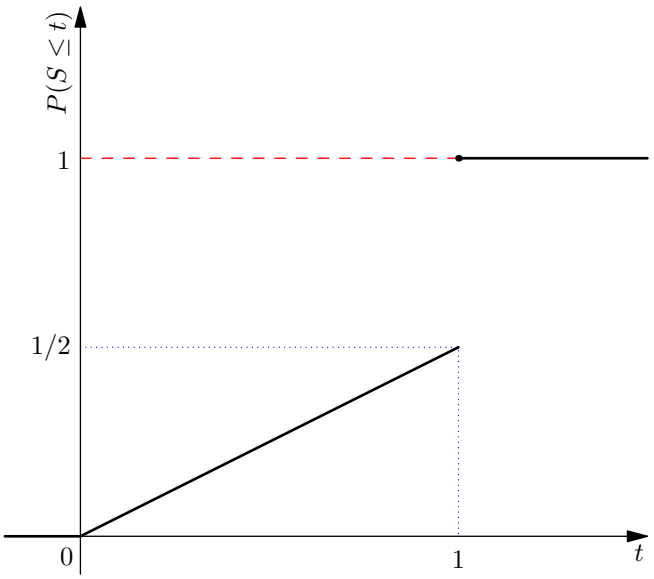

FIGURE 12 – Fonction de répartition de S.

Cherchons maintenant la f.d.r. de $S = X + Y$. La v.a. S prend toutes ses valeurs dans $[0, 1]$ puisque si X tombe dans $[0, 1/2]$, $0 \leqslant S = 2X \leqslant 1$ et si X tombe dans $]1/2, 1]$, $S = X + 1 - X = 1$. Calculons $P(S \leqslant t)$ pour $t \in [0, 1[$.

$$\begin{aligned} P(S \leqslant t) &= P(S \leqslant t \text{ et } 0 \leqslant X \leqslant 1/2) + P(S \leqslant t \text{ et } 1/2 < X \leqslant 1) \\ &= P(2X \leqslant t \text{ et } 0 \leqslant X \leqslant 1/2) + 0. \end{aligned}$$

En effet, puisque $t < 1$ et $S = 1$ sur $\{1/2 < X \leqslant 1\}$, $\{S \leqslant t\} \cap \{1/2 < X \leqslant 1\} = \varnothing$, d'où

$$\forall t \in [0, 1[, \quad P(S \leqslant t) = P(\{X \leqslant t/2\} \cap \{0 \leqslant X \leqslant 1/2\}) = P(0 \leqslant X \leqslant t/2) = t/2.$$

En rappelant que $P(S \leqslant t) = 1$ dès que $t \geqslant 1$, nous pouvons maintenant tracer la représentation graphique de la f.d.r. de S, cf. figure 12. Puisque cette f.d.r. est discontinue au point 1, la loi de S n'est pas à densité. Elle n'est pas non plus discrète car dans ce cas, la somme des amplitudes des sauts de sa f.d.r. devrait valoir 1, or il y a un seul saut d'amplitude 1/2. ◁

17. Notons que X prend *toutes* ses valeurs dans $[0, 1]$

Définition de l'espérance par la fonction de répartition

Revenons au problème de la définition de l'espérance. Il y a une voie intermédiaire entre l'approche pragmatique restreinte aux v.a. discrètes ou à densité et la construction d'une intégrale abstraite sur Ω. On peut s'inspirer de l'expression de la moyenne d'une série statistique à l'aide de la fonction de répartition (ou de la courbe des fréquences cumulées croissantes) pour proposer une définition de l'espérance de X entièrement basée sur la fonction de répartition. L'avantage est que la fonction de répartition existe pour toute v.a. et que l'on traite ainsi toutes les lois. De plus on reste dans le cadre de l'intégrale de Riemann. En prime, on obtient une interprétation graphique de l'espérance. On peut alors démontrer quasiment toutes les propriétés de l'espérance étudiées habituellement dans le cadre de l'intégrale abstraite sur Ω, cf. [9], chapitre 7.

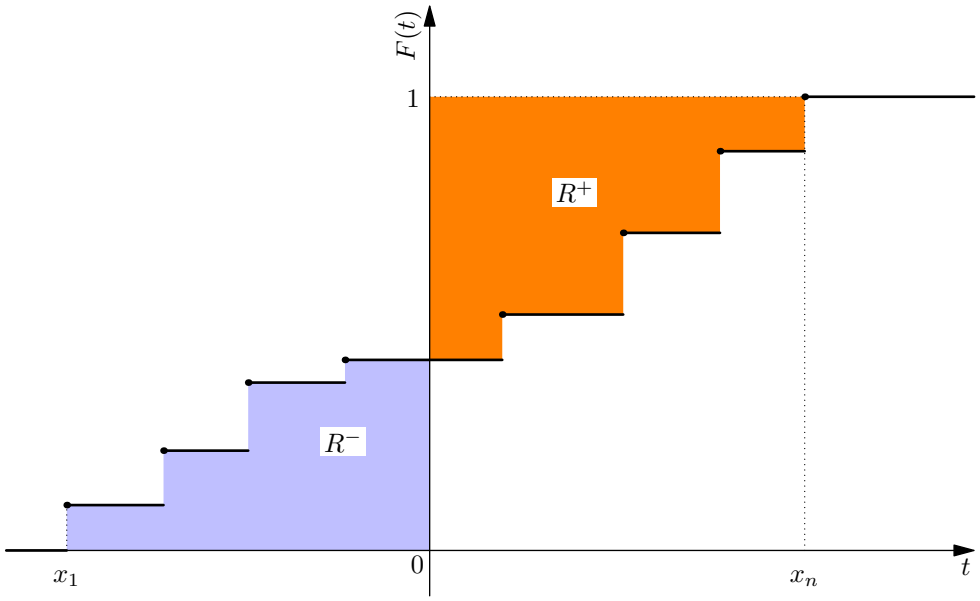

FIGURE 13 – Si X est discrète à support fini, $\mathrm{E}\,X = \mathrm{aire}(R^+) - \mathrm{aire}(R^-)$.

Partons du cas d'une v.a. discrète dont l'ensemble des valeurs $X(\Omega) = \{x_1, \ldots, x_n\}$ est fini et notons F sa fonction de répartition. L'espérance de X, moyenne des valeurs de X pondérées par leurs probabilités de réalisation est alors

$$\mathrm{E}\,X = \sum_{k=1}^{n} x_k P(X = x_k).$$

Quitte à réindexer les x_k, on peut supposer par commodité d'écriture que leur numérotation est croissante. Une adaptation immédiate de l'étude faite à propos de la moyenne d'une série statistique nous donne la formule

$$\mathrm{E}\,X = \int_0^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t - \int_{-\infty}^{0} F(t)\,\mathrm{d}t,$$

ainsi que l'interprétation graphique de la figure 13.

Notons que les intégrales généralisées dans la formule ci-dessus se réduisent à des intégrales sur un segment de fonctions monotones et sont donc des intégrales de Riemann ordinaires sans problème de convergence.

Pour définir l'espérance de X v.a. quelconque, il suffit d'oublier que F est une fonction en escalier et de considérer les intégrales comme de véritables intégrales généralisées, cf. figure 14. Il faut poser une restriction pour qu'aucune de ces intégrales ne vaille $+\infty$ (noter qu'on intègre à chaque fois une fonction positive, donc soit l'intégrale converge dans $\mathbb{R}_+$ soit elle vaut $+\infty$). En effet si elles étaient toutes deux infinies, on ne saurait comment définir leur différence et si une seule valait $+\infty$, admettre la valeur $-\infty$ ou $+\infty$ pour une espérance poserait le même problème pour l'étude de l'additivité de l'espérance.

Définition (espérance via la fonction de répartition). Soit X une variable aléatoire réelle sur $(\Omega, \mathcal{F}, P)$ et F sa fonction de répartition (sous P). On dit que X est intégrable si

$$\int_{-\infty}^{0} F(t)\,\mathrm{d}t < +\infty \quad \text{et} \quad \int_{0}^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t < +\infty.$$

Si X est intégrable, on appelle espérance de X, notation $\mathrm{E}\,X$, le réel

$$\mathrm{E}\,X = \int_{0}^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t - \int_{-\infty}^{0} F(t)\,\mathrm{d}t.$$

◁

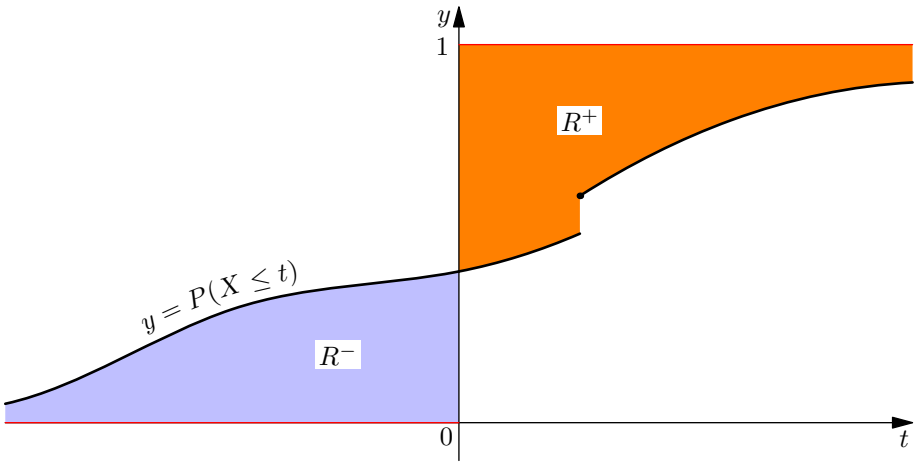

FIGURE 14 – $\mathrm{E}\,X = \mathrm{aire}(R^+) - \mathrm{aire}(R^-) = \int_0^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t - \int_{-\infty}^{0} F(t)\,\mathrm{d}t$

Remarque. $\mathrm{E}\,X$ ne dépend que de la loi de X sous P, il serait plus correct de parler d'espérance de X sous P. Notons que si X et Y ont même loi (sous P) et si X est intégrable (sous P), alors Y l'est aussi et $\mathrm{E}\,X = \mathrm{E}\,Y$. ◁

Remarque. Si on considère directement une mesure de probabilité Q sur $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$, on peut parler d'espérance de Q, définie à partir de sa f.d.r. F, sous réserve que cette dernière vérifie les conditions d'intégrabilité de la définition précédente. Cette espérance de Q lorsqu'elle existe, peut s'interpréter comme le *centre de gravité* ou *barycentre* de la répartition de masse sur la droite réelle associée à Q. ◁

Remarque. Lorsque Z est une v.a. positive, $\mathrm{E}\, Z$ se réduit à $\int_0^{+\infty} P(Z > t)\,\mathrm{d}t$, la convergence de cette intégrale étant la condition d'intégrabilité de Z. ◁

Remarque. On peut montrer de façon élémentaire, cf. par exemple [9] prop. 7.14, que $\int_0^{+\infty} P(X > t)\,\mathrm{d}t = \int_0^{+\infty} P(X \geqslant t)\,\mathrm{d}t$ et $\int_{-\infty}^{0} P(X \leqslant t)\,\mathrm{d}t = \int_{-\infty}^{0} P(X < t)\,\mathrm{d}t$. Ceci permet, entre autres, de justifier les égalités

$$\mathrm{E}\,|X| = \int_0^{+\infty} P(|X| > t)\,\mathrm{d}t = \int_{-\infty}^{0} F(t)\,\mathrm{d}t + \int_0^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t.$$

◁

Lien avec les définitions « pragmatiques »

Avant d'explorer les possibilités offertes par notre nouvelle définition de l'espérance, il convient de vérifier qu'elle contient bien comme cas particuliers les définitions « pragmatiques » de l'espérance des v.a. de loi discrète ou à densité.

Le cas des lois discrètes à support fini est évident puisqu'il nous a servi pour introduire la définition de l'espérance.

Cas des lois discrètes à support dénombrable. Supposons donc que la loi P_X est à support exactement dénombrable $\{x_k \ ; \ k \in \mathbb{N}^*\}$, donc $P_X = \sum_{k\in\mathbb{N}^*} p_k \delta_{x_k}$, avec les x_k distincts et $p_k = P(X = x_k)$. Regardons d'abord le cas particulier, mais qui recouvre la plupart des lois discrètes connues à support dénombrable, où les x_k sont positifs et où la numérotation $k \to x_k$ peut être choisie croissante. Notons ℓ la limite finie ou infinie de la suite *croissante* $(x_n)_{n\geqslant 1}$ et remarquons que si ℓ est fini, alors $F(t) = 1$ pour tout $t \geqslant \ell$ et donc $\int_0^{\ell} \big(1 - F(t)\big)\,\mathrm{d}t = \int_0^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t$, cette dernière intégrale étant alors trivialement convergente en $+\infty$. Par une adaptation immédiate de l'étude faite à propos de la moyenne d'une série statistique, on a pour tout $n \geqslant 1$,

$$\sum_{k=1}^{n} p_k x_k = \int_0^{x_n} \big(1 - F(t)\big)\,\mathrm{d}t - x_n\big(1 - F(x_n)\big). \qquad (\star)$$

Il s'agit de vérifier que la série $\sum_{k=1}^{+\infty} p_k x_k$ converge si et seulement si l'intégrale généralisée $\int_0^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t$ converge et que si ces convergences ont lieu, alors $\sum_{k=1}^{+\infty} p_k x_k = \int_0^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t$. On note d'abord que $\sum_{k=1}^{n} p_k x_k \leqslant \int_0^{x_n} \big(1 - F(t)\big)\,\mathrm{d}t$, donc si l'intégrale converge, toutes les sommes partielles de la série sont majorées

par $\int_0^\ell \big(1-F(t)\big)\,\mathrm{d}t$, ce qui implique la convergence de la série. Réciproquement, si la série converge, puisque la numérotation des x_k est croissante,

$$0 \leqslant x_n\big(1-F(x_n)\big) = x_n \sum_{k=n+1}^{+\infty} p_k \leqslant \sum_{k=n+1}^{+\infty} p_k x_k$$

et ce reste de série convergente tend vers zéro quand n tend vers l'infini. Donc $\int_0^{x_n} \big(1-F(t)\big)\,\mathrm{d}t = \sum_{k=1}^{n} p_k x_k + x_n\big(1-F(x_n)\big)$ converge quand n tend vers l'infini vers $\sum_{k=1}^{+\infty} p_k x_k$ et ceci entraîne la convergence de $\int_0^{+\infty} \big(1-F(t)\big)\,\mathrm{d}t$. Nous avons ainsi montré l'équivalence des convergences de la série à termes *positifs* $\sum_{k=1}^{+\infty} p_k x_k$ et de l'intégrale généralisée $\int_0^{+\infty} \big(1-F(t)\big)\,\mathrm{d}t$, c'est à dire l'équivalence des conditions d'intégrabilité de X selon les deux définitions de $\mathrm{E}\,X$. Si cette intégrabilité a lieu, nous avons vu que $x_n\big(1-F(x_n)\big)$ tend vers 0 et passant à la limite dans $(\star)$, on en déduit que $\sum_{k=1}^{+\infty} p_k x_k = \int_0^{+\infty} \big(1-F(t)\big)\,\mathrm{d}t$.

On pourrait au prix de quelques alourdissements d'écritures, étendre la démonstration ci-dessus au cas où la suite $(x_k)_{k\geqslant 1}$ peut être découpée en une sous-suite décroissante convergente vers $\ell' \in [-\infty, 0[$ et une sous-suite croissante convergente vers $\ell \in]0, +\infty]$. Cela ne couvre pas tous les cas possibles, en particulier le cas où les x_k sont par exemple tous les rationnels de $[0,1]$. Pour une preuve en toute généralité [18], nous renvoyons à [9] corollaire 7.27.

Cas des lois à densité. Nous allons vérifier les égalités

$$\int_0^{+\infty} P(X>t)\,\mathrm{d}t = \int_0^{+\infty} x f(x)\,\mathrm{d}x \quad \text{et} \quad \int_{-\infty}^{0} P(X\leqslant t)\,\mathrm{d}t = \int_{-\infty}^{0} (-x) f(x)\,\mathrm{d}x,$$

où ces intégrales généralisées de fonctions positives peuvent être convergentes ou valoir $+\infty$. Ceci nous donnera l'équivalence des conditions d'intégrabilité suivant les deux définitions de $\mathrm{E}\,X$ et en cas d'intégrabilité, l'égalité des deux expressions de $\mathrm{E}\,X$. Pour la première égalité, on note que si X admet pour densité f, alors pour tout t, $P(X>t) = \int_t^{+\infty} f(x)\,\mathrm{d}x$. Par conséquent :

$$\int_0^{+\infty} P(X>t)\,\mathrm{d}t = \int_0^{+\infty} \left\{ \int_t^{+\infty} f(x)\,\mathrm{d}x \right\} \mathrm{d}t = \int_0^{+\infty} \left\{ \int_0^{+\infty} f(x) \mathbf{1}_{[t,+\infty[}(x)\,\mathrm{d}x \right\} \mathrm{d}t.$$

Notons que pour $t \geqslant 0$, $\mathbf{1}_{[t,+\infty[}(x) = \mathbf{1}_{[0,x]}(t)$. L'intégrande $(x,t) \mapsto \mathbf{1}_{[0,x]}(t) f(x)$ étant *positive*, le théorème de Fubini-Tonelli légitime l'interversion des intégrations [19], d'où :

$$\int_0^{+\infty} P(X>t)\,\mathrm{d}t = \int_0^{+\infty} \left\{ \int_0^{+\infty} f(x) \mathbf{1}_{[0,x]}(t)\,\mathrm{d}t \right\} \mathrm{d}x = \int_0^{+\infty} f(x) \left\{ \int_0^{+\infty} \mathbf{1}_{[0,x]}(t)\,\mathrm{d}t \right\} \mathrm{d}x.$$

18. En fait dans le cas des v.a. positives. Pour le cas des v.a. réelles, il suffit, sous réserve d'intégrabilité, de séparer partie positive et partie négative de X pour en déduire le résultat.

19. Même si les intégrales valent $+\infty$.

Comme pour $x \geqslant 0$, $\int_0^{+\infty} \mathbf{1}_{[0,x]}(t)\,\mathrm{d}t = \int_0^x \mathrm{d}t = x$, la première égalité annoncée est établie. La vérification de la deuxième repose aussi sur le théorème de Fubini-Tonelli.

$$\begin{aligned}\int_{-\infty}^0 P(X \leqslant t)\,\mathrm{d}t &= \int_{-\infty}^0 \left\{\int_{-\infty}^t f(x)\,\mathrm{d}x\right\}\,\mathrm{d}t \\ &= \int_{-\infty}^0 \left\{\int_{-\infty}^0 \mathbf{1}_{]-\infty,t]}(x) f(x)\,\mathrm{d}x\right\}\,\mathrm{d}t \\ &= \int_{-\infty}^0 \left\{\int_{-\infty}^0 \mathbf{1}_{[x,0]}(t) f(x)\,\mathrm{d}t\right\}\,\mathrm{d}x \\ &= \int_{-\infty}^0 \left\{\int_{-\infty}^0 \mathbf{1}_{[x,0]}(t)\,\mathrm{d}t\right\} f(x)\,\mathrm{d}x.\end{aligned}$$

Comme pour $x \leqslant 0$, $\int_{-\infty}^0 \mathbf{1}_{[x,0]}(t)\,\mathrm{d}t = \int_x^0 \mathrm{d}t = -x$, on obtient bien la deuxième égalité annoncée.

Lien avec l'intégrale abstraite

Après avoir payé son tribut aux définitions pragmatiques de l'espérance, l'auteur de ces lignes se sent moralement obligé d'en faire autant pour la définition de l'espérance par l'intégrale abstraite sur $(\Omega, \mathcal{F}, P)$. Le lecteur qui ignore cette théorie peut sauter sans inconvénient ce passage.

Dans cette théorie, la condition d'intégrabilité de X s'écrit $\int_\Omega |X|\,\mathrm{d}P < +\infty$. Sous cette condition, l'espérance de X est définie par $\mathrm{E}\,X = \int_\Omega X\,\mathrm{d}P$. Nous allons vérifier qu'en partant de cette définition, on peut retrouver la formule exprimant $\mathrm{E}\,X$ à l'aide de la fonction de répartition F de X.

On commence par le découpage :

$$\mathrm{E}\,X = \int_\Omega X\mathbf{1}_{\{X>0\}}\,\mathrm{d}P + \int_\Omega X\mathbf{1}_{\{X\leqslant 0\}}\,\mathrm{d}P.$$

On remarque ensuite que si $X \leqslant 0$, $[0, X[= \varnothing$ et si $X > 0$, $[X, 0] = \varnothing$, ce qui nous permet d'écrire en notant λ la mesure de Lebesgue sur $\mathbb{R}$:

$$X\mathbf{1}_{\{X>0\}} = \lambda([0, X[) = \int_{\mathbb{R}^+} \mathbf{1}_{[0,X[}(t)\,\mathrm{d}\lambda(t) = \int_{\mathbb{R}^+} \mathbf{1}_{]t,+\infty[}(X)\,\mathrm{d}\lambda(t)$$

et

$$X\mathbf{1}_{\{X\leqslant 0\}} = -\lambda([X, 0]) = -\int_{\mathbb{R}^-} \mathbf{1}_{[X,0]}(t)\,\mathrm{d}\lambda(t) = -\int_{\mathbb{R}^-} \mathbf{1}_{]-\infty,t]}(X)\,\mathrm{d}\lambda(t).$$

Puis en intégrant sur Ω les v.a. $X\mathbf{1}_{\{X>0\}}$ et $X\mathbf{1}_{\{X\leqslant 0\}}$ et en appliquant le théorème de Fubini-Tonelli aux intégrales doubles $\int_\Omega \int_{\mathbb{R}^+}$ et $\int_\Omega \int_{\mathbb{R}^-}$, on obtient :

$$\begin{aligned}\mathrm{E}\,X &= \int_{\mathbb{R}^+} \left\{\int_\Omega \mathbf{1}_{]t,+\infty[}(X)\,\mathrm{d}P\right\}\,\mathrm{d}\lambda(t) - \int_{\mathbb{R}^-} \left\{\int_\Omega \mathbf{1}_{]-\infty,t]}(X)\,\mathrm{d}P\right\}\,\mathrm{d}\lambda(t) \\ &= \int_{\mathbb{R}^+} P(X > t)\,\mathrm{d}\lambda(t) - \int_{\mathbb{R}^-} P(X \leqslant t)\,\mathrm{d}\lambda(t).\end{aligned}$$

Pour conclure, on rappelle que si g est une fonction monotone positive sur un intervalle quelconque I de $\mathbb{R}$, d'extrémités $a \in \overline{\mathbb{R}}$ et $b \in \overline{\mathbb{R}}$, son intégrabilité au sens de Lebesgue sur I équivaut à la convergence de son intégrale de Riemann généralisée sur I et que les intégrales $\int_I g \,\mathrm{d}\lambda$ et $\int_a^b g(t)\,\mathrm{d}t$ sont égales. Par conséquent,

$$\mathrm{E}\, X = \int_0^{+\infty} P(X > t)\,\mathrm{d}t - \int_{-\infty}^{0} P(X \leqslant t)\,\mathrm{d}t,$$

ce qui nous redonne l'expression de $\mathrm{E}\, X$ à l'aide de la fonction de répartition F de X.

Le lecteur attentif aura remarqué que nous avons oublié de justifier l'équivalence des conditions d'intégrabilité de X pour la définition par $\int_\Omega X \,\mathrm{d}P$ et celle par la fonction de répartition. Cet oubli est facile à réparer en remarquant que

$$\int_\Omega |X|\,\mathrm{d}P = \int_\Omega X\mathbf{1}_{\{X>0\}}\,\mathrm{d}P - \int_\Omega X\mathbf{1}_{\{X\leqslant 0\}}\,\mathrm{d}P.$$

Il suffit alors de relire les calculs ci-dessus en faisant le changement de signe adéquat pour voir que

$$\int_\Omega |X|\,\mathrm{d}P = \int_0^{+\infty} P(X > t)\,\mathrm{d}t + \int_{-\infty}^{0} P(X \leqslant t)\,\mathrm{d}t.$$

Par conséquent :

$$\int_\Omega |X|\,\mathrm{d}P < +\infty \quad\Longleftrightarrow\quad \int_0^{+\infty} \big(1 - F(t)\big)\,\mathrm{d}t < +\infty \text{ et } \int_{-\infty}^{0} F(t)\,\mathrm{d}t < +\infty.$$

Propriétés de l'espérance

Dans tout ce qui suit, nous adoptons pour l'espérance la définition basée sur la fonction de répartition. Toutes les propriétés que nous allons examiner peuvent se démontrer sans autre bagage mathématique qu'une bonne connaissance de l'intégration au sens de Riemann [20].

Nous avons vu que la seule condition pour l'existence de l'espérance d'une variable aléatoire X sur $(\Omega, \mathcal{F}, P)$ est *l'intégrabilité* de X. Cette intégrabilité équivaut à la condition

$$\int_0^{+\infty} P(|X| > t)\,\mathrm{d}t < +\infty.$$

Proposition (intégrabilité des variables aléatoires bornées). *Si la variable aléatoire X est* bornée, *c'est-à-dire s'il existe une constante c telle que pour tout $\omega \in \Omega$, $|X(\omega)| \leqslant c$, ou plus généralement si X est P-presque-sûrement bornée c'est-à-dire s'il existe c réel tel que $P(|X| \leqslant c) = 1$, alors X est intégrable.*

20. Niveau Bac+2 selon les illusions de l'auteur de ces lignes.

En effet dans les deux cas c est positif ou nul et $P(|X| > t) = 0$ pour tout $t \geqslant c$ ce qui réduit l'intégrale généralisée $\int_0^{+\infty} P(|X| > t)\,\mathrm{d}t$ à une intégrale de Riemann ordinaire $\int_0^c P(|X| > t)\,\mathrm{d}t$ donc finie (et majorée par c).

Proposition (conditions suffisantes d'intégrabilité d'une variable aléatoire). *Si X vérifie pour une constante positive c, $P(|X| > t) \leqslant ct^{-\alpha}$ pour un certain $\alpha > 1$ et tout $t \geqslant t_0 > 0$, ou si $P(|X| > t) \leqslant ct^{-1}(\ln t)^{-\beta}$ pour un $\beta > 1$ et tout $t \geqslant t_0 > 0$, alors X est intégrable.*

Réciproquement, l'intégrabilité de X nous donne un renseignement sur la vitesse de convergence[21] vers 0 de $P(|X| > t)$ quand t tend vers $+\infty$. C'est l'inégalité de Markov que nous verrons ci-après page 64.

Voici maintenant deux propriétés qui découlent immédiatement de la formule de calcul de l'espérance d'une v.a. discrète à support fini. Il n'est pas inutile de retrouver les résultats en utilisant la formule de calcul de $\mathrm{E}\,X$ par la f.d.r.

Proposition (espérance d'une constante, d'une indicatrice).

a) Si X est une v.a. constante égale à c, ou plus généralement P-presque-sûrement constante égale à c, c'est-à dire $P(X = c) = 1$, $\mathrm{E}\,X = c$, ce que l'on peut encore écrire :

$$\mathrm{E}\,c = c, \textit{ pour toute constante } c.$$

b) Pour tout évènement $A \in \mathcal{F}$,

$$\mathrm{E}\,(\mathbf{1}_A) = P(A).$$

Proposition (linéarité de l'espérance).

a) L'espérance des v.a. réelles intégrables est additive *: si X et Y v.a. réelles définies sur le même $(\Omega, \mathcal{F}, P)$ sont intégrables, alors $X + Y$ l'est aussi et*

$$\mathrm{E}(X + Y) = \mathrm{E}\,X + \mathrm{E}\,Y.$$

b) Si X est intégrable, cX l'est aussi pour toute constante réelle c et

$$\mathrm{E}(cX) = c\,\mathrm{E}\,X.$$

Un exemple bien connu d'application de l'additivité de l'espérance est le calcul de l'espérance d'une v.a. X de loi binomiale $\mathrm{Bin}(n, p)$ en écrivant que X *a même loi* que la somme S_n de n v.a. de Bernoulli indépendantes de même paramètre p (ou d'indicatrices $\mathbf{1}_{A_1}, \ldots \mathbf{1}_{A_n}$ de n évènements indépendants de même probabilité p). On obtient ainsi $\mathrm{E}\,X = np$.

Un exemple moins connu mais où la simplification du calcul d'espérance due à l'additivité est encore plus spectaculaire est le cas des lois hypergéométriques.

21. Pour n'importe quelle variable aléatoire X, $P(|X| > t)$ tend vers 0 quand t tend vers $+\infty$, car $P(|X| > t) \leqslant F(-t) + 1 - F(t)$, où F est la f.d.r. de X qui tend toujours vers 0 en $-\infty$ et vers 1 en $+\infty$.

Exemple (espérance d'une loi hypergéométrique).
Si X suit la loi hypergéométrique $\mathrm{Hypg}(N, M, n)$, $\mathrm{E}\, X = n\dfrac{M}{N}$.
L'espérance d'une v. a. ne dépendant que de sa loi, on ne perd pas de généralité en supposant que X est le nombre d'objets défectueux observé dans un échantillon de taille n prélevé sans remise dans une population totale de N objets dont M sont défectueux. En numérotant de 1 à M tous les objets défectueux, on a alors

$$X = X_1 + \cdots + X_M,$$

où

$$X_i = \begin{cases} 1 & \text{si le } i^{\text{e}} \text{ objet défectueux est prélevé,} \\ 0 & \text{sinon.} \end{cases}$$

Chaque X_i est une variable de Bernoulli de paramètre $p_i = P(X_i = 1)$. Attention, contrairement au cas des prélèvements avec remises, les X_i n'ont aucun raison ici d'être indépendantes (autrement dit ici les évènements $A_i = \{$ prélèvement du i^{e} objet défectueux $\}$ ne sont pas indépendants). Le calcul de p_i se ramène au dénombrement de tous les échantillons possibles contenant le i^{e} objet défectueux. Un tel échantillon est constitué en prenant le i^{e} défectueux et en complétant par $n-1$ objets (défectueux ou non) choisis dans le reste de la population. D'où :

$$p_i = P(X_i = 1) = \frac{1 \times \binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}.$$

Ainsi les X_i ont même loi et même espérance $\mathrm{E}\, X_i = n/N$. Par linéarité on en déduit :

$$\mathrm{E}\, X = \sum_{i=1}^{M} \mathrm{E}\, X_i = M\frac{n}{N}.$$

Remarquons que le résultat obtenu est le même que pour un prélèvement de n objets *avec remise* : dans ce cas, le nombre Y d'objets défectueux dans l'échantillon suit la loi binomiale $\mathrm{Bin}(n, M/N)$ et $\mathrm{E}\, Y = nM/N$. ◁

Proposition (espérance et ordre).

a) L'espérance des v.a. réelles intégrables est croissante *: si X et Y v.a. réelles définies sur le même $(\Omega, \mathcal{F}, P)$ sont intégrables et vérifient $X \leqslant Y$, c'est-à-dire pour tout $\omega \in \Omega$, $X(\omega) \leqslant Y(\omega)$, alors $\mathrm{E}\, X \leqslant \mathrm{E}\, Y$.*

b) Si X est intégrable, $|X|$ l'est aussi et

$$|\mathrm{E}\, X| \leqslant \mathrm{E}\, |X|.$$

Proposition (inégalité de Markov). *Si X est une variable aléatoire positive,*

$$\forall x > 0, \quad P(X \geqslant x) \leqslant \frac{\mathrm{E}\, X}{x}.$$

Une preuve « muette[22] » de l'inégalité de Markov est donnée par la figure 15.

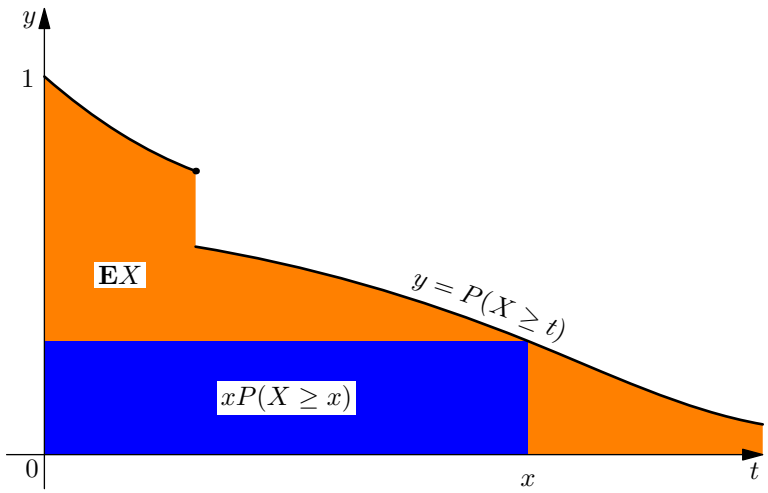

FIGURE 15 – Inégalité de Markov : $xP(X \geqslant x) \leqslant \displaystyle\int_0^{+\infty} P(X \geqslant t)\,\mathrm{d}t = \mathrm{E}\, X$.

Remarques.

1. Cette inégalité n'a d'intérêt que lorsque le second membre est inférieur à 1, c'est-à-dire lorsque $\mathrm{E}\, X < +\infty$ *et* $x > \mathrm{E}\, X$.
2. En pratique, l'inégalité de Markov est souvent appliquée à des variables aléatories réelles (par forcément positives) Y sous la forme :
$$\forall x > 0, \quad P(|Y| \geqslant x) \leqslant \frac{\mathrm{E}\,|Y|}{x}.$$
3. Il peut sembler un peu incongru de vouloir contrôler $P(X \geqslant x)$ à l'aide de $\mathrm{E}\, X$, puisque le calcul de cette espérance par la définition via la f.d.r. présuppose la connaissance des $P(X > t)$ pour $t \geqslant 0$, dont on déduit facilement les $P(X \geqslant t)$. Il se trouve qu'il arrive souvent que l'on sache calculer $\mathrm{E}\, X$ sans connaître, ou sans avoir besoin de calculer, la loi de X. C'est le cas par exemple quand X est une somme finie de variables aléatoires d'espérances connues. On peut aussi savoir majorer $\mathrm{E}\, X$ sans connaître la loi de X. Dans ces situations, l'inégalité de Markov est très utile. Pour ne citer qu'un exemple, l'inégalité de Markov est l'un des outils pour établir des « lois des grands nombres ».

◁

22. Enfin presque, puisqu'il convient de rappeler ici que $\int_0^{+\infty} P(X \geqslant t)\,\mathrm{d}t = \int_0^{+\infty} P(X > t)\,\mathrm{d}t$, cf. la remarque de la page 58 ou [9] prop. 7.14.

Moments d'une variable aléatoire réelle

On étudie dans cette section les $\mathrm{E}\, h(X)$, où h est une fonction réelle et X une variable aléatoire réelle sur $(\Omega, \mathcal{F}, P)$. Pour que l'expression $\mathrm{E}\, h(X)$ ait un sens, il est nécessaire que $Y = h(X)$ soit une variable aléatoire réelle. Cette condition sera réalisée si $h : \mathbb{R} \to \mathbb{R}$ est *borélienne*, c'est-à-dire si $B' = h^{-1}(B) \in \mathrm{Bor}(\mathbb{R})$ pour tout $B \in \mathrm{Bor}(\mathbb{R})$. Alors en effet,

$$Y^{-1}(B) = \{\omega \in \Omega\ ;\ h(X(\omega)) \in B\} = \{\omega \in \Omega\ ;\ X(\omega) \in h^{-1}(B)\} = X^{-1}(B') \in \mathcal{F},$$

puisque X est une variable aléatoire sur $(\Omega, \mathcal{F})$. Comme le borélien B ci-dessus est quelconque, ceci montre que Y est elle aussi une variable aléatoire sur $(\Omega, \mathcal{F})$. Il semble difficile d'exhiber un exemple de fonction h non borélienne sans faire appel à l'axiome du choix. C'est dire que la restriction « h borélienne » n'en est pas vraiment une en pratique.

Si h est borélienne, $\mathrm{E}\,|h(X)| = \int_0^{+\infty} P(|h(X)| > t)\,\mathrm{d}t$ existe toujours comme élément de $\overline{\mathbb{R}}_+$ et si cette intégrale converge dans $\mathbb{R}_+$, $\mathrm{E}\, h(X)$ existe comme nombre réel. On pourra désigner $\mathrm{E}\,|h(X)|$ et $\mathrm{E}\, h(X)$ respectivement par l'appellation h-moment absolu de X et h-moment de X. Bien entendu si h est borélienne positive, h-moment absolu et h-moment sont confondus et ce dernier existe toujours dans $\overline{\mathbb{R}}_+$. Nous utiliserons aussi l'appellation générique de *moments fonctionnels* pour désigner les h-moments[23].

Le cas le plus utile est celui où h est une fonction puissance, $h(x) = x^r$, on parle alors de moment d'ordre r de X.

Définition. Soit r un réel positif. On appelle *moment absolu d'ordre* r de la variable aléatoire réelle X la quantité $\mathrm{E}(|X|^r)$, élément de $\overline{\mathbb{R}}_+$. Si r est *entier* et X^r intégrable, donc si le moment absolu d'ordre r de X est *fini*, on appelle *moment d'ordre* r de X le réel $\mathrm{E}(X^r)$. On notera $\mathrm{E}\,|X|^r$ pour $\mathrm{E}(|X|^r)$ et $\mathrm{E}\, X^r$ pour $\mathrm{E}(X^r)$ en prenant garde de ne pas confondre ces quantités avec $(\mathrm{E}\,|X|)^r$ et $(\mathrm{E}\, X)^r$ respectivement. ◁

Remarquons qu'on ne définit pas le moment d'ordre r *non entier* pour X v.a. réelle, même si $\mathrm{E}\,|X|^r < +\infty$. En effet dans ce cas, X^r n'est pas définie sur l'évènement $\{X < 0\}$ et si $P(X < 0) \neq 0$, X^r ne peut être égale presque sûrement à une v.a. définie sur tout Ω. Bien entendu, si X est une v.a. positive, $\mathrm{E}\, X^r$ existe toujours dans $\overline{\mathbb{R}}_+$.

Proposition. *Si la variable aléatoire X a un moment absolu d'ordre r fini, elle a aussi un moment absolu d'ordre p fini pour tout $p \in [0, r]$.*

L'existence d'un moment absolu d'ordre r fini donne un renseignement sur la vitesse de convergence vers 0 de $P(|X| \geqslant t)$ quand t tend vers $+\infty$. On a alors $P(|X| \geqslant t) = O(t^{-r})$ par le corollaire suivant de l'inégalité de Markov.

23. Ces appellations ne sont pas standard, nous les adoptons par confort de rédaction.

Proposition (inégalité de Markov avec moment). *Pour toute variable aléatoire réelle X, pour tout réel $r > 0$,*

$$\forall t > 0, \quad P(|X| \geqslant t) \leqslant \frac{\mathrm{E}\,|X|^r}{t^r}.$$

Bien entendu, cette inégalité n'a d'intérêt que si $\mathrm{E}\,|X|^r < +\infty$ et $t^{-r}\,\mathrm{E}\,|X|^r < 1$.

Proposition (moments d'une v.a. discrète). *Si X est une variable aléatoire discrète, pour tout réel $r \geqslant 0$,*

$$\mathrm{E}\,|X|^r = \sum_{x \in X(\Omega)} |x|^r P(X = x).$$

De plus si r est entier et $\mathrm{E}\,|X|^r < +\infty$,

$$\mathrm{E}\,X^r = \sum_{x \in X(\Omega)} x^r P(X = x).$$

Rappelons que si X est une v.a. discrète, l'ensemble de ses valeurs $X(\Omega)$ est fini ou dénombrable. Par conséquent l'écriture $\sum_{x \in X(\Omega)}$ désigne soit une somme d'un nombre fini de termes, soit une série. Lorsque cette série est à termes positifs sa somme (dans $\overline{\mathbb{R}}_+$) ne dépend pas de l'ordre d'indexation de ses termes. Si cette série est absolument convergente pour une indexation particulière, elle le reste pour toute autre indexation et la somme ne dépend pas de l'indexation. Si la série des valeurs absolues diverge pour une indexation particulière, elle diverge pour toute autre indexation et dans ce cas, l'espérance de X^r n'existe pas.

Proposition (moments d'une v.a. à densité). *Si X est une variable aléatoire réelle à densité f, pour tout réel $r \geqslant 0$,*

$$\mathrm{E}\,|X|^r = \int_{-\infty}^{+\infty} |x|^r f(x)\,\mathrm{d}x.$$

Si de plus r est entier et $\mathrm{E}\,|X|^r < +\infty$,

$$\mathrm{E}\,X^r = \int_{-\infty}^{+\infty} x^r f(x)\,\mathrm{d}x.$$

Proposition (moments fonctionnels d'une v.a. discrète). *Si X est une variable aléatoire discrète et $h : \mathbb{R} \to \mathbb{R}$ une application borélienne,*

$$\mathrm{E}\,|h(X)| = \sum_{x \in X(\Omega)} |h(x)| P(X = x).$$

De plus, si $\mathrm{E}\,|h(X)| < +\infty$,

$$\mathrm{E}\,h(X) = \sum_{x \in X(\Omega)} h(x) P(X = x).$$

Proposition (moments fonctionnels d'une v.a. à densité). *Si X est une variable aléatoire de densité f et $h : \mathbb{R} \to \mathbb{R}$ une application réglée sur tout intervalle fermé borné de $\mathbb{R}$,*

$$\mathrm{E}\,|h(X)| = \int_{-\infty}^{+\infty} |h(x)| f(x)\,\mathrm{d}x.$$

De plus, si $\mathrm{E}\,|h(X)| < +\infty$,

$$\mathrm{E}\,h(X) = \int_{-\infty}^{+\infty} h(x) f(x)\,\mathrm{d}x.$$

Une application h est réglée sur $[a, b]$ si elle est limite uniforme sur $[a, b]$ d'une suite de fonctions en escaliers [24]. On démontre en analyse que h est réglée sur $[a, b]$ si et seulement si elle admet en tout point de $]a, b[$ une limite à gauche et une limite à droite (finies) ainsi qu'une limite finie à droite en a et à gauche en b. La classe des fonctions réglées, sans être aussi grande que celle des fonctions boréliennes, devrait donc suffire à nos besoins. Elle contient en particulier les fonctions continues et les fonctions monotones par morceaux.

Nous présentons maintenant une formule injustement méconnue, permettant de calculer des moments fonctionnels à partir de la fonction de survie. Son intérêt est de permettre un tel calcul pour des variables aléatoires qui ne sont ni discrètes ni à densité.

Proposition (moments fonctionnels et fonction de survie). *Soient X une variable aléatoire positive et g une application continue strictement croissante $\mathbb{R}_+ \to \mathbb{R}$, de classe C^1 sur $\mathbb{R}_+^*$. Alors*

$$\mathrm{E}\,g(X) = g(0) + \int_0^{+\infty} P(X > s) g'(s)\,\mathrm{d}s.$$

Variance d'une variable aléatoire

Le h-moment $\mathrm{E}\,h(X)$ pour $h : x \mapsto (x - \mathrm{E}\,X)^2$ occupe une place particulière dans la théorie des probabilités.

Définition (variance et écart type). Si X est de carré intégrable ($\mathrm{E}\,X^2 < +\infty$), on appelle *variance* de X le réel positif noté $\mathrm{Var}\,X$ défini par

$$\mathrm{Var}\,X = \mathrm{E}\,(X - \mathrm{E}\,X)^2.$$

On appelle alors *écart type* de X le réel $\sigma(X) = (\mathrm{Var}\,X)^{1/2}$. ◁

24. On en déduit que l'application réglée h est borélienne en montrant qu'elle est limite simple *sur* $\mathbb{R}$ d'une suite de fonctions en escalier, donc boréliennes. Ainsi $h(X)$ est bien une variable aléatoire par composition d'une application borélienne et d'une variable aléatoire.

Nous savons déjà que si $\mathrm{E}\,X^2 = \mathrm{E}\,|X|^2$ est fini, $\mathrm{E}\,|X|$ l'est aussi, donc $\mathrm{E}\,X$ est bien défini. De plus $(X - \mathrm{E}\,X)^2 = X^2 - 2(\mathrm{E}\,X)X + (\mathrm{E}\,X)^2$ apparaît alors comme une combinaison linéaire de trois variables[25] intégrables, donc est aussi intégrable. Ainsi la v.a. positive $(X - \mathrm{E}\,X)^2$ est intégrable et $\mathrm{E}(X - \mathrm{E}\,X)^2$ est bien un réel positif, ce qui justifie la définition de $\mathrm{Var}\,X$. Notons aussi que si X représente une grandeur physique, X, $\mathrm{E}\,X$ et $\sigma(X)$ ont la même unité, mais pas $\mathrm{Var}\,X$.

Lorsqu'elle existe, la variance de X est une façon de mesurer la *dispersion* de la loi de X autour de l'espérance. Les raisons de l'importance de la variance apparaîtront ultérieurement (inégalité de Tchebycheff, théorème limite central).

Soit $h : \mathbb{R} \to \mathbb{R}_+$, $x \mapsto (x - \mathrm{E}\,X)^2$. L'application des formules de calcul du h-moment pour une v.a. discrète ou à densité nous donne, sous réserve d'intégrabilité de X^2, les formules respectives :

$$\mathrm{Var}\,X = \sum_{x \in X(\Omega)} (x - \mathrm{E}\,X)^2 P(X = x) \quad \text{(si la loi de } X \text{ est discrète)},$$

$$\mathrm{Var}\,X = \int_{-\infty}^{+\infty} (x - \mathrm{E}\,X)^2 f(x)\,\mathrm{d}x \quad \text{(si la loi de } X \text{ est à densité } f).$$

Dans la pratique, ces formules sont rarement utilisées, on leur préfère la formule suivante qui simplifie les calculs.

Proposition (formule de Koenig-Huygens pour la variance). *Si la variable aléatoire X est de carré intégrable,*

$$\mathrm{Var}\,X = \mathrm{E}\,X^2 - (\mathrm{E}\,X)^2.$$

Preuve. Rappelons que nous notons $\mathrm{E}\,X^2$ pour $\mathrm{E}(X^2)$ et que le second membre de la formule ci-dessus n'est donc généralement pas nul. On pose $c = \mathrm{E}\,X$.

$$\begin{aligned} \mathrm{Var}\,X &= \mathrm{E}\,(X - c)^2 = \mathrm{E}\,(X^2 - 2cX + c^2) \\ &= \mathrm{E}\,X^2 - 2c\,\mathrm{E}\,X + \mathrm{E}\,c^2 \\ &= \mathrm{E}\,X^2 - 2c^2 + c^2 = \mathrm{E}\,X^2 - c^2, \end{aligned}$$

en utilisant la linéarité de l'espérance et l'espérance d'une constante. □

Proposition (translation et changement d'échelle). *Si* $\mathrm{E}\,X^2 < +\infty$,

$$\forall a \in \mathbb{R},\ \forall b \in \mathbb{R}, \quad \mathrm{Var}(aX + b) = a^2\,\mathrm{Var}\,X, \quad \sigma(aX + b) = |a|\sigma(X).$$

Preuve. En utilisant la définition de la variance, la linéarité de l'espérance et le fait que l'espérance d'une constante est cette constante :

$$\begin{aligned} \mathrm{Var}(aX + b) &= \mathrm{E}\left(aX + b - \mathrm{E}\,(aX + b)\right)^2 = \mathrm{E}\,(aX + b - a\,\mathrm{E}\,X - b)^2 \\ &= \mathrm{E}\left(a(X - \mathrm{E}\,X)\right)^2 = \mathrm{E}\left(a^2(X - \mathrm{E}\,X)^2\right) \\ &= a^2\,\mathrm{E}\left((X - \mathrm{E}\,X)^2\right) = a^2\,\mathrm{Var}\,X, \end{aligned}$$

ce qui nous donne les formules annoncées. □

25. À savoir X^2, X et la v.a. constante $(\mathrm{E}\,X)^2$.

Il est clair, d'après la définition de la variance, que la variance d'une constante est nulle. La réciproque est *presque* vraie :

Proposition (nullité de la variance et constance p.s.).

$$\operatorname{Var} X = 0 \Leftrightarrow P(X = \operatorname{E} X) = 1 \Leftrightarrow X \text{ est presque sûrement constante.}$$

Deux problèmes de minimisation (suite de la partie 1)

Nous avons déjà cherché quelle constante approchait le mieux une série statistique donnée. Nous avons vu que la réponse dépend du choix de la distance pour mesurer cette proximité. Nous allons voir que l'on peut résoudre le problème analogue pour une variable aléatoire sur $(\Omega, \mathcal{F}, P)$. Nous cherchons donc à résoudre les deux problèmes de minimisation suivants : trouver la ou les constantes c qui réalisent le minimum de $T_1(c)$ ou $T_2(c)$ définis par

$$T_1(c) = \operatorname{E}|X - c|, \quad T_2(c) = \operatorname{E}(X - c)^2.$$

Pour T_1, on suppose que X est intégrable, pour T_2 on suppose que X est de carré intégrable.

Minimisation de $T_2(c)$

Le problème de la minimisation de $T_2(c)$ se résout exactement comme dans le cas d'une série statistique, modulo le changement de notations. Puisque X est de carré intégrable, elle est *a fortiori* intégrable donc $\operatorname{E} X$ existe. Posons pour alléger les écritures, $\operatorname{E} X = m$. En utilisant la linéarité de l'espérance et le fait que l'espérance d'une constante est égale à cette constante, on obtient pour c réel quelconque

$$\begin{aligned}
\operatorname{E}(X - c)^2 &= \operatorname{E}\big((X - m) + (m - c)\big)^2 \\
&= \operatorname{E}\big((X - m)^2 + 2(m - c)(X - m) + (m - c)^2\big) \\
&= \operatorname{E}(X - m)^2 + 2(m - c)\operatorname{E}(X - m) + \operatorname{E}\big((m - c)^2\big) \\
&= \operatorname{Var} X + 2(m - c)(\operatorname{E} X - \operatorname{E} m) + (m - c)^2 \\
&= \operatorname{Var} X + 2(m - c)(\operatorname{E} X - m) + (m - c)^2,
\end{aligned}$$

d'où

$$\forall c \in \mathbb{R}, \quad \operatorname{E}(X - c)^2 = \operatorname{Var} X + (m - c)^2.$$

On en déduit immédiatement que $\operatorname{E}(X - c)^2$ est toujours supérieur ou égal à $\operatorname{Var} X$, avec égalité si et seulement $c = m = \operatorname{E} X$.

Proposition. *Si X est de carré intégrable, $T_2(c) = \operatorname{E}(X - c)^2$ a un unique minimum global atteint si et seulement si c est l'espérance de X.*

Dans ce contexte, l'écart-type $\sigma = \big(T_2(\operatorname{E} X)\big)^{1/2} = \big(\operatorname{Var} X\big)^{1/2}$ s'interprète comme la distance, au sens L^2, entre la variable aléatoire réelle X et sa meilleure approximation par une constante. Il est donc naturel de prendre σ comme *indicateur de dispersion* de X (ou de sa loi).

Minimisation de $T_1(c)$

Pour minimiser $T_1(c)$, commençons par l'exprimer à l'aide d'intégrales. La fonction de répartition de la v.a. $X - c$ est $t \mapsto F(t + c)$, en notant F la f.d.r. de X. En utilisant la formule pour l'espérance de la valeur absolue d'une v.a., voir la remarque page 58, on en déduit que

$$T_1(c) = \int_{-\infty}^{0} F(t + c)\,\mathrm{d}t + \int_{0}^{+\infty} \big(1 - F(t + c)\big)\,\mathrm{d}t.$$

En effectuant dans ces deux intégrales le changement de variable[26] « translation » $s = t + c$, on obtient

$$T_1(c) = \int_{-\infty}^{c} F(s)\,\mathrm{d}s + \int_{c}^{+\infty} \big(1 - F(s)\big)\,\mathrm{d}s.$$

Ceci justifie l'interprétation graphique donnée par la figure 16. À partir de là, on

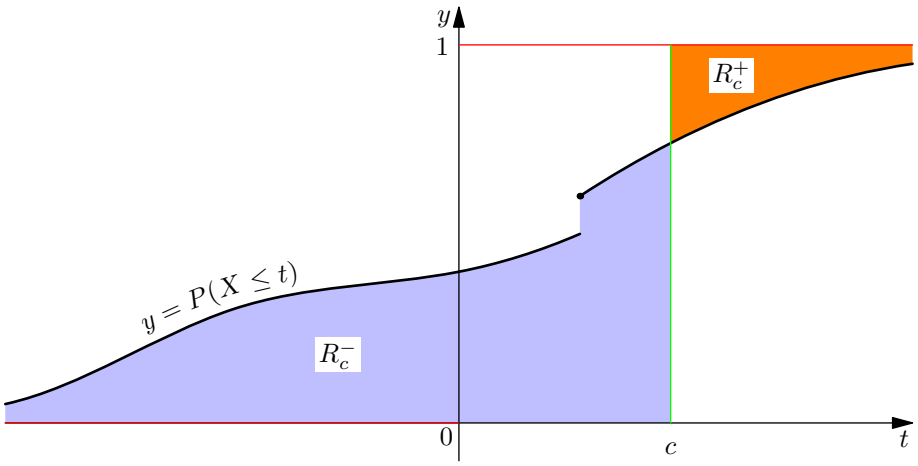

FIGURE 16 – $T_1(c) = \mathrm{E}\,|X - c| = \mathrm{aire}(R_c^-) + \mathrm{aire}(R_c^+)$

peut reproduire presque mot pour mot la résolution du problème de minimisation de $T_1(c)$ pour une série statistique. Pour la continuité de T_1 on utilise l'inégalité triangulaire et la croissance de l'espérance. Pour l'étude du sens de variation de T_1, une relecture attentive de la preuve donnée dans le cas d'une série statistique montre que l'on n'utilise pas le fait que la représentation graphique de F est une fonnction en escaliers mais seulement la croissance de F.

Proposition. *Si X est intégrable, $T_1(c) = \mathrm{E}\,|X - c|$ a un unique minimum global, atteint si et seulement si c est une médiane de X.*

26. Dans les cours classiques sur l'intégrale de Riemann, on suppose pour faire le changement de variable que la fonction à intégrer est continue, ce qui n'est pas forcément vérifié ici puisque F et $1 - F$ peuvent avoir des sauts. Néanmoins le changement de variable translation fonctionne avec des fonctions qui sont seulement Riemann intégrables, ce qui est le cas des fonctions monotones, voir par exemple [9] prop. 4.41 p. 122.

Variance d'une somme, covariance

Le calcul de la variance d'une somme joue un rôle essentiel en théorie des probabilités. En guise d'exemple introductif, nous allons voir une façon instructive de calculer « à la main » la variance d'une loi binomiale. Le résultat suivant est bien connu.

Proposition. *Si X suit la loi* $\mathrm{Bin}(n,p)$, $\operatorname{Var} X = np(1-p)$.

Preuve. On rappelle que X a même loi que la variable aléatoire

$$S_n = \sum_{i=1}^{n} \mathbf{1}_{A_i},$$

où les n évènements A_i sont mutuellement indépendants et de même probabilité p. Comme S_n est bornée ($\forall \omega \in \Omega,\ 0 \leqslant S_n(\omega) \leqslant n$), il n'y a aucun problème d'existence de $\operatorname{Var} S_n$ ni donc de $\operatorname{Var} X$ puisque la variance ne dépend que de la loi. Pour la même raison, $\operatorname{Var} X = \operatorname{Var} S_n$. Nous posons $Y_i = \mathbf{1}_{A_i} - \mathrm{E}\,\mathbf{1}_{A_i} = \mathbf{1}_{A_i} - P(A_i) = \mathbf{1}_{A_i} - p$. On a alors :

$$S_n - \mathrm{E}\,S_n = \sum_{i=1}^{n} \mathbf{1}_{A_i} - \sum_{i=1}^{n} \mathrm{E}\,\mathbf{1}_{A_i} = \sum_{i=1}^{n} (\mathbf{1}_{A_i} - \mathrm{E}\,\mathbf{1}_{A_i}) = \sum_{i=1}^{n} Y_i.$$

D'où :

$$(S_n - \mathrm{E}\,S_n)^2 = \sum_{i=1}^{n} Y_i^2 + \sum_{\substack{1 \leqslant i,j \leqslant n \\ i \neq j}} Y_i Y_j.$$

Par linéarité de l'espérance, on en déduit :

$$\operatorname{Var} S_n = \mathrm{E}(S_n - \mathrm{E}\,S_n)^2 = \sum_{i=1}^{n} \mathrm{E}\,Y_i^2 + \sum_{\substack{1 \leqslant i,j \leqslant n \\ i \neq j}} \mathrm{E}(Y_i Y_j).$$

Comme la v.a. $\mathbf{1}_{A_i}$ ne prend que les valeurs 0 et 1, elle est égale à son carré, d'où

$$\begin{aligned} \mathrm{E}\,Y_i^2 = \mathrm{E}(\mathbf{1}_{A_i} - p)^2 &= \mathrm{E}(\mathbf{1}_{A_i} - 2p\mathbf{1}_{A_i} + p^2) \\ &= P(A_i) - 2pP(A_i) + p^2 \\ &= p - p^2 = p(1-p). \end{aligned}$$

Notons au passage que nous venons de calculer la variance d'une v.a. de Bernoulli de paramètre p. D'autre part,

$$\begin{aligned} \mathrm{E}(Y_i Y_j) &= \mathrm{E}\big[(\mathbf{1}_{A_i} - p)(\mathbf{1}_{A_j} - p)\big] \\ &= \mathrm{E}(\mathbf{1}_{A_i}\mathbf{1}_{A_j}) - p\,\mathrm{E}\,\mathbf{1}_{A_j} - p\,\mathrm{E}\,\mathbf{1}_{A_i} + p^2 \\ &= \mathrm{E}(\mathbf{1}_{A_i}\mathbf{1}_{A_j}) - p^2. \end{aligned}$$

Il reste donc à calculer les $\mathrm{E}(\mathbf{1}_{A_i}\mathbf{1}_{A_j})$ pour $i \neq j$. Comme les indicatrices ne peuvent prendre que les valeurs 0 ou 1, il en est de même pour les produits $\mathbf{1}_{A_i}\mathbf{1}_{A_j}$ qui sont donc des v.a. de Bernoulli de paramètre p' donné par :

$$p' = P(\mathbf{1}_{A_i}\mathbf{1}_{A_j} = 1) = P(A_i \cap A_j) = P(A_i)P(A_j) = p^2,$$

par *indépendance* de A_i et A_j lorsque $i \neq j$. On en déduit que

$$\text{si } i \neq j, \quad \mathrm{E}(Y_iY_j) = p^2 - p^2 = 0,$$

d'où finalement,

$$\operatorname{Var} S_n = \sum_{i=1}^{n} \mathrm{E}\, Y_i^2 = np(1-p).$$

□

Venons en maintenant à la variance d'une somme

$$S_n = \sum_{i=1}^{n} X_i$$

de variables aléatoires. Le calcul ci-dessus de la variance d'une loi binomiale illustre le rôle clé joué par des quantités comme $\mathrm{E}(X_iX_j)$ et $\mathrm{E}\big[(X_i - \mathrm{E}\, X_i)(X_j - \mathrm{E}\, X_j)\big]$. Nous allons généraliser l'étude de ces quantités. La première question qui se pose alors est celle de l'intégrabilité d'un produit de deux variables aléatoires X et Y. En utilisant l'inégalité $|XY| \leqslant X^2 + Y^2$, on voit qu'une condition suffisante pour l'existence de $\mathrm{E}(XY)$ est que X et Y soient de carré intégrable. On peut d'ailleurs majorer $\mathrm{E}(XY)$ à l'aide des moments d'ordre 2 de X et Y.

Théorème (inégalité de Cauchy-Schwarz).
Si X et Y ont des moments d'ordre 2 :

$$|\,\mathrm{E}(XY)| \leqslant (\mathrm{E}\, X^2)^{1/2}(\mathrm{E}\, Y^2)^{1/2}.$$

La preuve est similaire à celle de l'inégalité de Cauchy-Schwarz en géométrie euclidienne. Il suffit d'écrire que $\mathrm{E}(X + tY)^2$ est positif pour tout t ... On vérifie aussi que l'inégalité est une égalité si et seulement si les variables aléatoires X et Y, vues comme des vecteurs de l'espace vectoriel des v.a. de carré intégrable, sont presque sûrement colinéaires, au sens où il existe $(\alpha, \beta) \neq (0, 0)$ tel que $P(\alpha X + \beta Y = 0) = 1$.

Regardons maintenant comment calculer $\operatorname{Var}(X + Y)$ lorsque X et Y ont des moments d'ordre 2 (ce qui entraîne l'existence de cette variance). En utilisant la définition de la variance et la linéarité de l'espérance on obtient :

$$\begin{aligned}
\operatorname{Var}(X+Y) &= \mathrm{E}\big[X + Y - \mathrm{E}(X+Y)\big]^2 \\
&= \mathrm{E}\big[(X - \mathrm{E}\, X) + (Y - \mathrm{E}\, Y)\big]^2 \\
&= \mathrm{E}\big[(X - \mathrm{E}\, X)^2 + (Y - \mathrm{E}\, Y)^2 + 2(X - \mathrm{E}\, X)(Y - \mathrm{E}\, Y)\big] \\
&= \operatorname{Var} X + \operatorname{Var} Y + 2\,\mathrm{E}\big[(X - \mathrm{E}\, X)(Y - \mathrm{E}\, Y)\big].
\end{aligned}$$

Ainsi on voit qu'en général $\operatorname{Var}(X+Y) \neq \operatorname{Var} X + \operatorname{Var} Y$ et qu'il y a un terme correctif. Nous le noterons $2\operatorname{Cov}(X,Y)$.

Définition (covariance).
Si les v.a. X et Y ont des moments d'ordre 2, on appelle covariance du couple aléatoire (X,Y) la quantité :

$$\operatorname{Cov}(X,Y) = \mathrm{E}\big[(X - \mathrm{E}\,X)(Y - \mathrm{E}\,Y)\big].$$

En particulier, $\operatorname{Cov}(X,X) = \operatorname{Var} X$. ◁

Proposition (propriétés de la covariance).
Pour tout couple (X,Y) de v.a. ayant des moments d'ordre 2 :
(*i*) $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$.
(*ii*) *Pour tous réels a, b, c, d :* $\operatorname{Cov}(aX+b, cY+d) = ac\operatorname{Cov}(X,Y)$.
(*iii*) $|\operatorname{Cov}(X,Y)| \leqslant \sigma(X)\sigma(Y)$.

Définition (coefficient de corrélation).
Si X et Y sont des v.a. non p.s. constantes et de carré intégrable, on appelle coefficient de corrélation entre X et Y la quantité :

$$\rho(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma(X)\sigma(Y)}.$$

◁

D'après (*iii*) on a toujours $-1 \leqslant \rho(X,Y) \leqslant 1$. D'autre part il résulte facilement du cas d'égalité dans l'inégalité de Cauchy-Schwarz que $|\rho|$ est maximal lorsque Y est une fonction affine de X : $Y = aX + b$. Quand $\rho = 0$, on dit que X et Y sont *non corrélées*.

Proposition (formule de Koenig).
Si la covariance de X et Y existe, elle peut se calculer par :

$$\operatorname{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}\,X\,\mathrm{E}\,Y.$$

Proposition (variance d'une somme, cas général).
Si les v.a. $X_1, \ldots, X_n$ ont des moments d'ordre 2 :

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i,j=1}^{n} \operatorname{Cov}(X_i, X_j) \\ &= \sum_{i=1}^{n} \operatorname{Var} X_i + \sum_{\substack{i,j=1 \\ i\neq j}}^{n} \operatorname{Cov}(X_i, X_j).\end{aligned}$$

Preuve. Nous avons déjà rencontré le cas $n = 2$ pour lequel la formule précédente s'écrit :

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

Pour n quelconque, l'identité algébrique

$$\left(\sum_{i=1}^{n} Y_i\right)^2 = \sum_{i,j=1}^{n} Y_i Y_j$$

utilisée avec $Y_i = X_i - \mathrm{E}\, X_i$ et la linéarité de l'espérance nous donnent :

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \mathrm{E}\left\{\sum_{i=1}^{n} X_i - \mathrm{E}\left(\sum_{i=1}^{n} X_i\right)\right\}^2 &= \mathrm{E}\left\{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathrm{E}\, X_i\right\}^2 \\ &= \sum_{i,j=1}^{n} \mathrm{E}(Y_i Y_j) \\ &= \sum_{i,j=1}^{n} \operatorname{Cov}(X_i, X_j).\end{aligned}$$

□

Cette expression de la variance d'une somme peut s'interpréter géométriquement comme le développement du carré scalaire d'une somme dans l'espace vectoriel des variables aléatoires[27] centrées et de carré intégrables (ici les Y_i). Dans cet espace vectoriel, le produit scalaire de Y_i par Y_j est défini par $\mathrm{E}(Y_i Y_j)$. Dire que X_i et X_j sont non corrélées signifie que Y_i et Y_j sont orthogonales pour ce produit scalaire. En particulier,

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var} X_i \quad \text{si les } X_i \text{ sont deux à deux non corrélées.}$$

C'est en quelque sorte le *théorème de Pythagore* de la théorie des probabilités. Sous une forme intuitive qui sera précisée ultérieurement (inégalité de Tchebycheff, loi des grands nombres, théorème limite central), ce résultat nous dit que pour une somme de n variables aléatoires de carré intégrable, de même loi et non corrélées, l'ordre de grandeur de la dispersion autour de l'espérance est $\sqrt{n}$ au lieu de n. On pourra regarder à titre d'exercice le cas où les X_i suivent la loi uniforme sur $[0, 1]$.

Espérance et indépendance

Ce qui précède montre l'intérêt de disposer de conditions suffisantes pour que deux variables aléatoires X_i et X_j soient non corrélées. En revenant à la définition de la covariance, il suffirait que l'on puisse écrire

27. En toute rigueur, des classes d'équivalence modulo l'égalité presque-sûre des v.a.

$$\text{« } \mathrm{E}\big((X_i - \mathrm{E}\,X_i)(X_j - \mathrm{E}\,X_j)\big) = \big(\mathrm{E}(X_i - \mathrm{E}\,X_i)\big)\big(\mathrm{E}(X_j - \mathrm{E}\,X_j)\big) \text{ »},$$

pour que $\mathrm{Cov}(X_i, X_j) = 0$ puisque $\mathrm{E}(X_i - \mathrm{E}\,X_i) = \mathrm{E}\,X_i - \mathrm{E}(\mathrm{E}\,X_i) = \mathrm{E}\,X_i - \mathrm{E}\,X_i$. Plus généralement, on peut se demander sous quelles conditions on dispose de la sympathique formule $\mathrm{E}(XY) = (\mathrm{E}\,X)(\mathrm{E}\,Y)$. De même que nous notons $\mathrm{E}\,X^2$ pour $\mathrm{E}(X^2)$, nous noterons désormais $\mathrm{E}\,XY$ pour $\mathrm{E}(XY)$. L'écriture $\mathrm{E}\,X\,\mathrm{E}\,Y$ n'est pas ambigüe en raison de la propriété $\mathrm{E}(Xc) = c\,\mathrm{E}\,X$ pour c constante.

Il convient d'abord de noter que la formule $\mathrm{E}\,XY = \mathrm{E}\,X\,\mathrm{E}\,Y$ est trop belle pour être toujours vraie. Le contre exemple le plus simple [28] est le cas où X est l'indicatrice d'un évènement de probabilité $p \in]0,1[$ et où $Y = X$. Alors $XY = X^2 = X$, puisque X ne peut prendre que les valeurs 0 ou 1, d'où $\mathrm{E}(XY) = p$ tandis que $(\mathrm{E}\,X)(\mathrm{E}\,Y) = p^2 \neq p$.

En dehors du cas trivial où l'une des v.a. X et Y est constante, l'exemple le plus simple de couple de v.a. (X, Y) vérifiant $\mathrm{E}\,XY = \mathrm{E}\,X\,\mathrm{E}\,Y$ est celui où $X = \mathbf{1}_A$ et $Y = \mathbf{1}_B$, les évènements A et B étant indépendants. Cet exemple a déjà été traité au cours du calcul de la variance d'une loi binomiale, mais étant donné son rôle clé, il n'est pas inutile d'y revenir en détail. On remarque d'abord que

$$\mathbf{1}_A \mathbf{1}_B = \mathbf{1}_{A \cap B}.$$

Pour justifier cette égalité entre variables aléatoires définies sur le même Ω, il nous faut montrer que

$$\forall \omega \in \Omega, \quad \mathbf{1}_A(\omega)\mathbf{1}_B(\omega) = \mathbf{1}_{A\cap B}(\omega).$$

Dans cette égalité à démontrer, le premier membre comme le deuxième ne peuvent prendre que les valeurs 0 ou 1. Le premier membre vaut 1 si et seulement si les deux facteurs du produit valent 1, autrement dit si et seulement si $\omega \in A$ et $\omega \in B$, ce qui équivaut à $\omega \in A \cap B$. Finalement, pour tout $\omega \in \Omega$,

$$(\mathbf{1}_A\mathbf{1}_B)(\omega) = \mathbf{1}_A(\omega)\mathbf{1}_B(\omega) = \begin{cases} 1 & \text{si } \omega \in A \cap B \\ 0 & \text{sinon} \end{cases} \quad \text{d'où} \quad (\mathbf{1}_A\mathbf{1}_B)(\omega) = \mathbf{1}_{A\cap B}(\omega).$$

Ensuite, puisque l'espérance de l'indicatrice d'un évènement est égale à la probabilité de cet évènement,

$$\mathrm{E}\,XY = \mathrm{E}\,\mathbf{1}_A\mathbf{1}_B = \mathrm{E}\,\mathbf{1}_{A\cap B} = P(A \cap B).$$

Pour l'instant nous n'avons pas encore utilisé l'indépendance et donc les égalités ci-dessus sont valables pour toute paire d'évènements A et B. Si de plus A et B sont indépendants, alors

$$P(A \cap B) = P(A)P(B) = \mathrm{E}\,\mathbf{1}_A\,\mathrm{E}\,\mathbf{1}_B = \mathrm{E}\,X\,\mathrm{E}\,Y.$$

28. Il est un peu frustrant de ne donner qu'un contre exemple avec Y fonction de X, mais pour l'instant nous ne savons pas calculer $\mathrm{E}(XY)$ si on connaît seulement la loi de X et celle de Y.

Nous avons ainsi prouvé que lorsque $X = \mathbf{1}_A$ et $Y = \mathbf{1}_B$, $\mathrm{E}\, XY = \mathrm{E}\, X\, \mathrm{E}\, Y$ si A et B sont indépendants [29].

Ce premier exemple élémentaire va nous permettre d'en construire un plus riche avec des v.a. X et Y discrètes à support fini. Supposons que $X(\Omega) = \{x_1, \dots, x_m\}$ et $Y(\Omega) = \{y_1, \dots, y_n\}$. Alors on peut écrire les décompositions

$$X = \sum_{i=1}^{m} x_i \mathbf{1}_{A_i}, \quad Y = \sum_{j=1}^{n} y_j \mathbf{1}_{B_j},$$

où

$$A_i = X^{-1}(\{x_i\}) = \{\omega \in \Omega\ ;\ X(\omega) = x_i\} \text{ et } B_j = Y^{-1}(\{y_j\}) = \{\omega \in \Omega\ ;\ Y(\omega) = y_j\}.$$

Les deux familles d'évènements $\{A_i,\ i \in [\![1, m]\!]\}$ et $\{B_j,\ j \in [\![1, n]\!]\}$ constituent chacune une *partition* de Ω. Le produit XY s'écrit alors

$$XY = \sum_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}} x_i y_j \mathbf{1}_{A_i} \mathbf{1}_{B_j} = \sum_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}} x_i y_j \mathbf{1}_{A_i \cap B_j},$$

d'où par linéarité de l'espérance,

$$\mathrm{E}\, XY = \sum_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}} x_i y_j\, \mathrm{E}\, \mathbf{1}_{A_i \cap B_j} = \sum_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}} x_i y_j P(A_i \cap B_j).$$

Supposons que le couple de v.a. (X, Y) vérifie la propriété suivante :

$$\forall (i, j) \in [\![1, m]\!] \times [\![1, n]\!],\ A_i = X^{-1}(\{x_i\}) \text{ et } B_j = Y^{-1}(\{y_j\}) \text{ sont indépendants.} \quad (\star)$$

Dans ce cas on déduit du calcul ci-dessus que

$$\begin{aligned} \mathrm{E}\, XY = \sum_{\substack{1 \leqslant i \leqslant m \\ 1 \leqslant j \leqslant n}} x_i y_j P(A_i) P(B_j) &= \left(\sum_{i=1}^{m} x_i P(A_i) \right) \left(\sum_{j=1}^{n} y_j P(B_j) \right) \\ &= \left(\sum_{i=1}^{m} x_i\, \mathrm{E}\, \mathbf{1}_{A_i} \right) \left(\sum_{j=1}^{n} y_j\, \mathrm{E}\, \mathbf{1}_{B_j} \right) = \mathrm{E}\, X\, \mathrm{E}\, Y. \end{aligned}$$

En conclusion, si le couple de v.a. discrètes à support fini (X, Y) vérifie la propriété $(\star)$, $\mathrm{E}\, XY = \mathrm{E}\, X\, \mathrm{E}\, Y$.

Pour généraliser davantage, il est commode d'interpréter $(\star)$ en faisant un détour par les tribus. Notons $\mathcal{F}_X$ la sous-tribu de $\mathcal{F}$ engendrée par la famille d'évènements

29. En fait « si et seulement si » comme on peut le voir en recollant de manière adéquate les deux lignes d'égalités $\mathrm{E}\, XY = \dots$ et $\dots = \mathrm{E}\, X\, \mathrm{E}\, Y$ dans la preuve.

$\{A_i,\ i \in [\![1,m]\!]\}$. Comme cette famille est une partition finie de Ω, les éléments de $\mathcal{F}_X$ sont exactement tous les ensembles de la forme $\bigcup_{i\in I} A_i$ pour $I \subset [\![1,m]\!]$, vérification laissée en exercice. On définit de même la tribu $\mathcal{F}_Y$ à partir de la partition $\{B_j,\ j \in [\![1,n]\!]\}$. Nous allons montrer que si $(\star)$ est vérifiée, alors pour tout $A \in \mathcal{F}_X$ et tout $B \in \mathcal{F}_Y$, A et B sont indépendants. D'après la forme générale des éléments de $\mathcal{F}_X$ et $\mathcal{F}_Y$, il existe $I \subset [\![1,m]\!]$ et $J \subset [\![1,n]\!]$ tels que $A = \bigcup_{i\in I} A_i$ et $B = \bigcup_{j\in J} B_j$. On en déduit que

$$\begin{aligned}
P(A\cap B) &= P\left(\left(\bigcup_{i\in I} A_i\right)\cap\left(\bigcup_{j\in J} B_j\right)\right)\\
&= P\left(\bigcup_{(i,j)\in I\times J}(A_i\cap B_j)\right) \text{ union finie disjointe}\\
&= \sum_{(i,j)\in I\times J} P(A_i\cap B_j)\\
&= \sum_{(i,j)\in I\times J} P(A_i)P(B_j)\\
&= \left(\sum_{i\in I} P(A_i)\right)\left(\sum_{j\in J} P(B_j)\right)\\
&= P\left(\bigcup_{i\in I} A_i\right) P\left(\bigcup_{j\in J} B_j\right) = P(A)P(B),
\end{aligned}$$

ce qui établit l'indépendance annoncée.

Proposition (tribu engendrée par une variable aléatoire). *Soit Z une variable aléatoire réelle sur l'espace probabilisable $(\Omega,\mathcal{F})$. La famille des évènements $Z^{-1}(C) = \{Z \in C\}$ pour C borélien de $\mathbb{R}$ est une sous-tribu de $\mathcal{F}$ appelée tribu engendrée par Z.*

La vérification de l'affirmation ci-dessus est un exercice facile qui exploite le fait que l'inverse ensembliste Z^{-1} commute avec réunion et intersection. On peut donner une interprétation probabiliste de la tribu engendrée par Z en disant que c'est la famille des évènements dont la réalisation ne dépend que des valeurs prises par Z.

À ce stade, le lecteur attentif est en train de se demander si la tribu $\mathcal{F}_X$ introduite ci-dessus est la tribu engendrée par X. La réponse est positive et pour le voir, le point clé est de remarquer que si C est un borélien quelconque de $\mathbb{R}$, en notant $C_X = \{x_i \in X(\Omega)\ ;\ x_i \in C\}$,

$$X^{-1}(C) = \bigcup_{x_i\in C_X} A_i.$$

Donc $X^{-1}(C)$ est une union finie d'évènements A_i donc un élément de $\mathcal{F}_X$. Il en résulte que la tribu engendrée par X est une sous-tribu de $\mathcal{F}_X$. L'inclusion inverse entre ces deux tribus est évidente puisque chaque A_i est l'image réciproque par X du borélien $\{x_i\}$.

Définition (tribus indépendantes). Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé et $\mathcal{G}$, $\mathcal{H}$ deux sous-tribus de $\mathcal{F}$. On dit qu'elles sont indépendantes[30] si pour tout $A \in \mathcal{G}$ et tout $B \in \mathcal{H}$, les évènements A et B sont indépendants. $\lhd$

Avec ce nouveau langage, nous pouvons dire que les deux tribus $\mathcal{F}_X$ et $\mathcal{F}_Y$ introduites ci-dessus sont les tribus engendrées par X et Y et que si $(\star)$ est vérifiée, elles sont indépendantes. Réciproquement, l'indépendance de $\mathcal{F}_X$ et $\mathcal{F}_Y$ implique $(\star)$.

Nous sommes maintenant en mesure d'aborder la généralisation finale.

Définition (indépendance de deux variables aléatoires réelles). On dit que deux variables aléatoires réelles définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$ sont indépendantes si les tribus qu'elles engendrent sont indépendantes. $\lhd$

Nous admettrons qu'une condition nécessaire et suffisante pour que deux v.a. X et Y soient indépendantes est que pour tous intervalles I_1, I_2 de $\mathbb{R}$,

$$P(X \in I_1 \text{ et } Y \in I_2) = P(X \in I_1)P(Y \in I_2).$$

Cette CNS est la définition « pragmatique » proposée par le programme de l'Agrégation interne.

Théorème. *Soient X et Y deux variables aléatoires réelles définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$, indépendantes et intégrables. Alors XY est intégrable et $\mathrm{E}\,XY = \mathrm{E}\,X\,\mathrm{E}\,Y$.*

Notons au passage que lorsque X et Y sont indépendantes, nous n'avons pas besoin de supposer X et Y de carré intégrable pour assurer l'intégrabilité de XY. Ce théorème se démontre en approximant X et Y par des suites de v.a. discrètes à support fini $(X_n)_{n\geqslant 1}$ et $(Y_n)_{n\geqslant 1}$ qui « héritent » de l'indépendance de X et Y et en utilisant un théorème d'interversion limite-espérance.

Avant d'explorer plus avant les propriétés de l'indépendance des variables aléatoires, fixons une réponse au problème initial qui a motivé l'introduction de cette notion.

Corollaire. *Si X et Y sont deux variables aléatoires réelles définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$ de carré intégrable et indépendantes,* $\mathrm{Cov}(X, Y) = 0$.

Il convient de noter immédiatement que le réciproque est fausse. Deux v.a. de carré intégrables peuvent être non corrélées sans être indépendantes.

Une propriété extrêmement commode de l'indépendance est sa conservation par image par fonction borélienne (hérédité de l'indépendance par image borélienne).

30. On devrait dire en toute rigueur « P-indépendantes ».

Proposition. *Soient X et Y deux variables aléatoires indépendantes définies sur $(\Omega, \mathcal{F}, P)$ et soient f et g deux fonctions boréliennes $\mathbb{R} \to \mathbb{R}$. Alors les v.a. $f(X)$ et $g(Y)$ sont indépendantes.*

La preuve tient en une ligne et illustre la pertinence de la définition de l'indépendance des v.a. via les tribus engendrées. Elle se réduit au constat que la tribu engendrée par $f(X)$ est une sous-tribu de celle engendrée par X (et de même pour $g(Y)$ et Y).

En conséquence pratique, si X et Y sont indépendantes et si $f(X)$ et $g(Y)$ sont intégrables, $f(X)g(Y)$ est intégrable et $\mathrm{E}(f(X)g(Y)) = \mathrm{E}\, f(X)\, \mathrm{E}\, g(Y)$.

Suites de variables aléatoires indépendantes

Nous pouvons maintenant généraliser la notion d'indépendance à plus de deux variables aléatoires. Comme pour l'indépendance des évènements, on obtient ainsi une propriété plus forte que l'indépendance deux à deux.

Regardons d'abord le cas d'une famille finie de v.a. $X_1, \ldots, X_n$ définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$. Dans le cas $n = 2$, nous savons déjà que par définition, X_1 et X_2 sont indépendantes si les tribus engendrées $\mathcal{F}_{X_1}$ et $\mathcal{F}_{X_2}$ sont indépendantes, autrement dit si pour tous sous-ensembles boréliens B_1 et B_2 de $\mathbb{R}$, les évènements $X_1^{-1}(B_1)$ et $X_2^{-1}(B_2)$ sont indépendants. Voici la généralisation au cas $n > 2$.

Définition (indépendance d'une famille finie de variables aléatoires). Les variables aléatoires $X_1, \ldots, X_n$ définies sur le même $(\Omega, \mathcal{F}, P)$ sont indépendantes si les tribus engendrées $\mathcal{F}_{X_1}, \ldots, \mathcal{F}_{X_n}$ sont indépendantes, autrement dit si pour tous sous-ensembles boréliens $B_1, \ldots, B_n$ de $\mathbb{R}$ les n évènements $X_1^{-1}(B_1), \ldots, X_n^{-1}(B_n)$ sont *mutuellement* indépendants. ◁

Comme dans le cas $n = 2$, on peut réduire cette définition théorique à une définition pragmatique en remplaçant les B_i par des intervalles quelconques de $\mathbb{R}$. L'indépendance des n variables aléatoires $X_1, \ldots, X_n$ est évidemment conservée par images boréliennes. Par récurrence à partir du cas $n = 2$, on en déduit le théorème suivant.

Théorème. *Soient $X_1, \ldots, X_n$ variables aléatoires indépendantes définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$ et $h_1, \ldots, h_n$ des fonctions boréliennes $\mathbb{R} \to \mathbb{R}$. Alors les variables aléatoires $h_1(X_1), \ldots, h_n(X_n)$ sont indépendantes. Si de plus elles sont intégrables, alors leur produit $h_1(X_1) \ldots h_n(X_n)$ l'est aussi et*

$$\mathrm{E}\left(h_1(X_1) \ldots h_n(X_n)\right) = \mathrm{E}\, h_1(X_1) \ldots \mathrm{E}\, h_n(X_n).$$

Dans les problèmes de temps d'attente, comme dans ceux de convergence d'une suite de variables aléatoires (loi forte des grands nombres, TLC) on a besoin de la notion de suite infinie de variables aléatoires indépendantes. Connaissant déjà la définition de l'indépendance d'une famille finie de variables aléatoires indépendantes, la généralisation s'opère comme suit.

Définition (suite infinie de variables aléatoires indépendantes). Soit $(X_i)_{i\in\mathbb{N}^*}$ une suite infinie de variables aléatoires définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$. On dit que les X_i sont indépendantes si pour toute partie finie K de $\mathbb{N}^*$ ayant au moins deux éléments, $(X_i)_{i\in K}$ est une famille finie de variables aléatoires indépendantes. ◁

Loi des grands nombres

En théorie des probabilités, l'expression générique « loi des grands nombres » désigne un théorème qui établit la convergence en un certain sens des moyennes arithmétiques d'une suite de variables aléatoire X_i vers une constante. Cette convergence est très utile en statistique pour *estimer* des paramètres d'une loi inconnue, sur la base de l'observation d'un *échantillon* $X_1, \ldots, X_n$ de grande taille.

Convergences presque sûre et en probabilité

Commençons par préciser ce que l'on entend par convergence d'une suite de variables aléatoires (Z_n) vers une v.a. Z. Comme les Z_n sont des applications de Ω dans $\mathbb{R}$, le premier mode de convergence auquel on pense est la convergence *pour tout* $\omega \in \Omega$ de la suite de réels $Z_n(\omega)$ vers le réel $Z(\omega)$. Ceci correspond à la convergence simple d'une suite d'applications en analyse. Malheureusement pour le type de résultat que nous avons en vue, ce mode de convergence est trop restrictif. Pour la loi des grands nombres, même dans le cas le plus favorable[31], on ne peut empêcher que la suite étudiée diverge pour une infinité de ω. Ce qui sauve la situation est que l'ensemble de ces ω a une probabilité nulle. Ceci nous amène à définir la convergence *presque sûre*.

Définition (convergence presque sûre).
Soit $(Z_n)_{n\geqslant 1}$ une suite de variables aléatoires et Z une v.a. définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$. On dit que Z_n converge presque sûrement vers Z si l'ensemble des ω tels que $Z_n(\omega)$ converge vers $Z(\omega)$ a pour probabilité 1. Cette convergence sera notée :

$$Z_n \xrightarrow[n\to+\infty]{\text{p.s.}} Z.$$

◁

Rappelons qu'un événement de probabilité 1 n'est pas forcément égal à Ω, il peut même y avoir une infinité d'éléments dans son complémentaire. Remarquons aussi que l'ensemble Ω' des ω tels que $Z_n(\omega)$ converge vers $Z(\omega)$ est bien un événement observable, c'est-à-dire un événement de la famille $\mathcal{F}$, cf.[9] pp. 334–335. Il est donc légitime de parler de sa probabilité.

31. Voir la discussion à propos de la loi forte des grands nombres pour les fréquences dans [8] section 6.5.

Dans la convergence presque sûre, le rang n_0 à partir duquel on peut approximer $Z_n(\omega)$ par $Z(\omega)$ avec une erreur inférieure à ε dépend à la fois de ε et de $\omega \in \Omega'$: $n_0 = n_0(\varepsilon, \omega)$. On ne sait pas toujours expliciter la façon dont $n_0(\varepsilon, \omega)$ dépend de ω. D'autre part on peut très bien avoir $\sup\{n_0(\varepsilon, \omega),\ \omega \in \Omega'\} = +\infty$. Ceci fait de la convergence presque sûre en général un résultat essentiellement théorique.

Supposons que la valeur de Z_n dépende du résultat de n épreuves répétées (ou de n observations). Savoir que Z_n converge presque sûrement vers Z ne permet pas de *prédire* le nombre non aléatoire n d'épreuves (ou d'observations) à partir duquel on aura $|Z_n(\omega) - Z(\omega)| < \varepsilon$ (sinon pour tous les $\omega \in \Omega'$, du moins avec une probabilité supérieure à un seuil fixé à l'avance par exemple 95%, 99%,...). Or cette question a une grande importance pratique pour le statisticien. C'est l'une des raisons de l'introduction de la *convergence en probabilité* qui permet de répondre à cette question lorsque l'on connaît la vitesse de convergence selon ce mode.

Définition (convergence en probabilité).
Soit $(Z_n)_{n\geqslant 1}$ une suite de variables aléatoires et Z une v.a. définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$. On dit que Z_n converge en probabilité vers Z si :

$$\forall \varepsilon > 0, \quad \lim_{n\to+\infty} P(|Z_n - Z| \geqslant \varepsilon) = 0.$$

Notation : $Z_n \xrightarrow[n\to+\infty]{\Pr} Z$. ◁

La convergence presque sûre implique la convergence en probabilité, la réciproque est fausse. Pour cette raison, la convergence en probabilité de la suite des moyennes arithmétiques $M_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ évoquée en introduction s'appelle une loi *faible* des grands nombres, sa convergence presque sûre une loi *forte* des grands nombres.

Inégalité de Bienaymé-Tchebycheff

Le théorème suivant n'est, d'un point de vue moderne, qu'un cas particulier de l'inégalité de Markov avec moment, mais il lui est historiquement antérieur. Il fut publié une première fois en 1853 par le mathématicien français Jules Bienaymé qui l'utilisa pour prouver une loi faible des grands nombres. Il fut publié en russe et en français en 1867 par le mathématicien russe Pafnouti Tchebycheff.

Théorème (inégalité de Bienaymé-Tchebycheff). *Si la variable aléatoire réelle X est de carré intégrable,*

$$\forall t > 0, \quad P(|X - \mathrm{E}\,X| \geqslant t) \leqslant \frac{\mathrm{Var}\,X}{t^2}.$$

Cette inégalité jette un éclairage sur les notions de variance et d'écart-type.

Remarque. Si on pose $t = u\sigma(X)$ dans l'inégalité de Bienaymé-Tchebycheff ci-dessus, elle devient :

$$\forall u > 0, \quad P\big(|X - \mathrm{E}\,X| \geqslant u\sigma(X)\big) \leqslant \frac{1}{u^2}.$$

Sous cette forme, le majorant obtenu est indépendant de la loi de X. Ceci permet de comprendre pourquoi $\sigma(X)$ s'appelle *écart type* ou *unité d'écart*. Pour toute loi de probabilité ayant un moment d'ordre 2, la probabilité d'observer une déviation par rapport à l'espérance d'au moins u unités d'écart est majorée par u^{-2}. ◁

Exemple. On jette 3600 fois un dé. Minorer la probabilité que le nombre d'apparitions du 1 soit compris strictement entre 480 et 720.

Notons S le nombre d'apparitions du 1. Cette variable aléatoire suit la loi binomiale $\mathrm{Bin}(3600, 1/6)$. La valeur exacte de la probabilité qui nous intéresse est :

$$P(480 < S < 720) = \sum_{k=481}^{719} \binom{3600}{k} \Big(\frac{1}{6}\Big)^k \Big(\frac{5}{6}\Big)^{3600-k}.$$

Le calcul de la valeur *exacte* de cette somme nécessiterait l'écriture d'un programme et une certaine puissance de calcul informatique. L'inégalité de Bienaymé-Tchebycheff est une alternative pratique à un calcul aussi déraisonnable. En effet on a :

$$\mathrm{E}\, S = 3600 \times \frac{1}{6} = 600, \qquad \mathrm{Var}\, S = 3600 \times \frac{1}{6} \times \frac{5}{6} = 500$$

et on remarque que $480 - 600 = -120$, $720 - 600 = 120$ d'où :

$$480 < S < 720 \Leftrightarrow -120 < S - 600 < +120 \Leftrightarrow |S - 600| < 120.$$

L'inégalité de Bienaymé-Tchebycheff s'écrit ici :

$$\forall t > 0, \quad P(|S - 600| \geqslant t) \leqslant \frac{500}{t^2}.$$

En particulier pour $t = 120$:

$$P(|S - 600| \geqslant 120) \leqslant \frac{500}{120^2} = \frac{5}{144}.$$

D'où en passant à l'événement complémentaire :

$$P(480 < S < 720) = P(|S - 600| < 120) \geqslant 1 - \frac{5}{144} \geqslant 0{,}965.$$

◁

Le lien entre l'inégalité de Bienaymé-Tchebycheff et la loi faible des grands nombres est le corollaire suivant obtenu en combinant l'inégalité appliquée à la variable $S_n = \sum_{i=1}^n X_i$ avec le calcul de la variance de S_n.

Corollaire. *Soient $X_1, \ldots, X_n$ des variables aléatoires réelles définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$, de carré intégrable et deux à deux non-corrélées. Notons $S_n = \sum_{i=1}^n X_i$. Alors*

$$\forall t > 0, \quad P\Big(\, |S_n - \mathrm{E}\, S_n| \geqslant t\Big) \leqslant \frac{1}{t^2} \sum_{i=1}^n \mathrm{Var}\, X_i.$$

Ceci s'applique en particulier au cas où les X_i sont indépendantes.

Loi faible des grands nombres

Théorème (loi faible des grands nombres). *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$, de même loi, de carré intégrable et deux à deux non-corrélées. Notons $S_n = \sum_{k=1}^{n} X_k$. On a la convergence en probabilité :*

$$\frac{S_n}{n} \xrightarrow[n\to+\infty]{\Pr} \mathrm{E}\, X_1.$$

Preuve. Comme les X_k ont même loi, $\mathrm{E}\, X_k = \mathrm{E}\, X_1$ et $\operatorname{Var} X_k = \operatorname{Var} X_1$ pour tout k. Comme elles sont aussi deux à deux non-corrélées, $\operatorname{Var} S_n = n \operatorname{Var} X_1$. Par linéarité de l'espérance on a aussi $\mathrm{E}\, S_n = n\, \mathrm{E}\, X_1$. L'inégalité de Bienaymé-Tchebycheff nous dit alors que :

$$\forall t > 0, \quad P\big(|S_n - \mathrm{E}\, S_n| \geqslant t\big) = P\big(|S_n - n\, \mathrm{E}\, X_1| \geqslant t\big) \leqslant \frac{n \operatorname{Var} X_1}{t^2}.$$

Posant $t = n\varepsilon$, on en déduit que pour tout $\varepsilon > 0$ et tout $n \in \mathbb{N}^*$,

$$P\big(|S_n - n\, \mathrm{E}\, X_1| \geqslant n\varepsilon\big) = P\left(\left|\frac{S_n}{n} - \mathrm{E}\, X_1\right| \geqslant \varepsilon\right) \leqslant \frac{n \operatorname{Var} X_1}{n^2\varepsilon^2} = \frac{\operatorname{Var} X_1}{n\varepsilon^2}.$$

Pour tout $\varepsilon > 0$ fixé, on a ainsi

$$P\left(\left|\frac{S_n}{n} - \mathrm{E}\, X_1\right| \geqslant \varepsilon\right) \leqslant \frac{\operatorname{Var} X_1}{n\varepsilon^2} \xrightarrow[n\to+\infty]{} 0,$$

ce qui établit la convergence en probabilité de la suite de variables aléatoires S_n/n vers la variable aléatoire constante $\mathrm{E}\, X_1$. ▫

Remarque. Nous avons en fait démontré un peu plus que la seule convergence en probabilité. Nous avons obtenu une vitesse de convergence vers 0 en $O(1/n)$ pour $P(|n^{-1}S_n - \mathrm{E}\, X_1| \geqslant \varepsilon)$. Si l'on connaît $\operatorname{Var} X_1$ ou si on sait le majorer, on peut donc répondre à la question posée page 81 lors de l'introduction de la convergence en probabilité[32]. ◁

Corollaire (loi faible des grands nombres pour les fréquences).
Si $(X_n)_{n\geqslant 1}$ est une suite de v.a. de Bernoulli indépendantes de même paramètre p, alors :

$$M_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to+\infty]{\Pr} p.$$

32. Attention à ne pas parler ici de *vitesse de convergence en probabilité* en $O(1/n)$. La vitesse de convergence en probabilité est définie autrement et est ici en $O(n^{-1/2})$, cf. page 87.

C'est un cas particulier de la loi faible des grands nombres en notant que des v.a. indépendantes et de carré intégrable[33] sont deux à deux non-corrélées et qu'ici $\mathrm{E}\, X_1 = p$. Ce qui fait l'importance de ce corollaire est l'application à la convergence des fréquences. Considérons une suite d'épreuves répétées indépendantes. Pour chaque épreuve la probabilité d'un « succès » est p. Notons X_i l'indicatrice de l'événement *succès à la i^e épreuve*. Alors $S_n = \sum_{i=1}^n X_i$ est le nombre de succès en n épreuves et $M_n = n^{-1}S_n$ est la *fréquence* des succès au cours des n premières épreuves.

Loi forte des grands nombres

Les lois fortes des grands nombres ne sont au programme ni du CAPES ni de l'Agrégation interne de Mathématiques. Pour des raisons culturelles et pour une meilleure compréhension des éléments de statistique mathématique contenus dans la troisième partie de ce livret, il est néanmoins bon de connaître l'énoncé suivant et sa réciproque, deux théorèmes dus à Kolmogorov et Khintchine.

Théorème (loi forte des grands nombres). *On suppose les X_k indépendantes,* de même loi *et* $\mathrm{E}\,|X_1| < +\infty$. *Alors*

$$\frac{S_n}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow[n\to+\infty]{\text{p.s.}} \mathrm{E}\, X_1.$$

Ce résultat est le meilleur possible en raison du théorème suivant.

Théorème (réciproque de la LFGN). *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires indépendantes et de même loi telle que S_n/n converge presque sûrement. Alors* $\mathrm{E}\,|X_1| < +\infty$ *et la limite* p.s. *de S_n/n est la constante* $\mathrm{E}\, X_1$.

TLC (convergence vers une loi normale)

Considérations terminologiques

Commencer une section avec un titre comportant un sigle (TLC) non déjà défini dans le texte n'est pas une pratique très recommandable. Si l'auteur s'y est laissé aller c'est qu'on trouve dans la littérature mathématique française plusieurs traductions différentes pour désigner le même théorème qui permet de justifier sous des conditions très générales l'approximation de la loi d'une somme de variables aléatoires indépendantes et de même loi par une loi gaussienne. Heureusement, toutes ces traductions peuvent se retrouver dans le sigle TLC. Le lecteur pressé peut donc sauter à la sous-section suivante et considérer que dans toute la suite de ce texte, TLC désigne le « théorème de la limite centrée » des programmes.

33. Les v.a. de Bernoulli ont des moments de tout ordre.

La première apparition d'une dénomination standardisée pour ce théorème semble être allemande sous la forme *Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung*. La traduction mot à mot de *Grenzwertsatz* donnerait « théorème de valeurs limites », celle de *Wahrscheinlichkeitsrechnung* donnerait « calcul de vraisemblance », mais tous les probabilistes savent qu'il s'agirait d'un faux sens et que la traduction correcte est « calcul des probabilités ».

Dans la littérature de langue anglaise (la langue internationale des scientifiques de notre époque), on parle de *central limit theorem* que l'on abrège usuellement en CLT. On peut mentionner que *limit theorem* que l'on traduit paresseusement en français par l'anglicisme « théorème limite » signifie en fait théorème de convergence (comme dans le théorème de convergence dominée ou celui de convergence monotone).

Ceci étant dit, on trouve dans la littérature mathématique en langue française les différentes traductions suivantes de *Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung* ou *central limit theorem*.

1. « Théorème limite central » ou « Théorème central limite ».
2. « Théorème de la limite centrale » ou « Théorème limite centrale ».
3. « Théorème de la limite centrée ».

Chacun est libre de se faire un opinion sur la pertinence de la mise en valeur de différents aspects du TLC par chacune de ces versions. L'auteur se sent néanmoins moralement obligé de mentionner qu'en dehors de la francophonie, c'est le sens de la version 1 qui est adopté par les mathématiciens, comme l'illustrent les deux citations suivantes.

> The term "central limit theorem" most likely traces back to Georg Pólya. As he recapitulated at the beginning of a paper published in 1920, it was "generally known that the appearance of the Gaussian probability density[34] e^{-x^2}" in a great many situations "can be explained by one and the same limit theorem," which plays "a central role in probability theory" [Pólya 1920, p.171]. Laplace had discovered the essentials of this fundamental theorem in 1810, and with the designation "central limit theorem of probability theory," which was even emphasized in the paper's title, Pólya gave it the name that has been in general use ever since.

Hans Fischer, A History of the Central Limit Theorem, Springer 2010.

> The appelation "central" is due to Polyá (1920) who used it because of the central role of the theorem in probability theory, not as the modern French do, because it describes the behavior of the center of the distribution as opposed to its tails.

34. Polyà was a bit careless with his phraseology here. Naturally he knew that e^{-x^2} is not a probability density (the norming factor $1/\sqrt{\pi}$ is missing).

Lucien Le Cam, The Central Limit theorem around 1935, Statistical science, 1986, Vol. 1, No 1, 78–96.

La référence bibliographique à l'article de Polyà publié en 1920 est :
Polyà, G. (1920). Über den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentproblem, Math. Z., 8 171–180.

TLC : dans le vif du sujet

Théorème (TLC). *Soit $(X_k)_{k\geqslant 1}$ une suite de variables aléatoires définies sur le même espace probabilisé $(\Omega, \mathcal{F}, P)$, indépendantes, de même loi et de carré intégrable et non* p.s. *constantes. Notons $\mu := \mathrm{E}\, X_1$, $\sigma^2 := \mathrm{Var}\, X_1$ avec $\sigma > 0$ et $S_n = \sum_{k=1}^n X_k$. Notons*

$$S_n^* := \frac{S_n - \mathrm{E}\, S_n}{\sqrt{\mathrm{Var}\, S_n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Alors

$$\forall x \in \mathbb{R}, \quad P(S_n^* \leqslant x) \xrightarrow[n\to+\infty]{} \Phi(x) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(\frac{-t^2}{2}\right) \mathrm{d}t.$$

Φ est la fonction de répartition de la loi $\mathfrak{N}(0,1)$.

Une conséquence pratique de la conclusion est que pour tous réels a, b, tels que $a < b$,

$$P\big(S_n^* \in I(a,b)\big) \xrightarrow[n\to+\infty]{} \Phi(b) - \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_a^b \exp\left(\frac{-t^2}{2}\right) \mathrm{d}t,$$

où $I(a,b)$ désigne n'importe lequel des quatre intervalles d'extrémités a et b.

Corollaire (théorème de de Moivre-Laplace). *Si B_n est une variable aléatoire de loi binomiale de paramètres n et $p \in]0,1[$, notons $q = 1 - p$ et*

$$B_n^* := \frac{B_n - np}{\sqrt{npq}} = \sqrt{\frac{n}{pq}}\left(\frac{B_n}{n} - p\right).$$

Alors

$$\forall x \in \mathbb{R}, \quad P(B_n^* \leqslant x) \xrightarrow[n\to+\infty]{} \Phi(x).$$

Le théorème de de Moivre-Laplace est une conséquence immédiate du TLC en remarquant que B_n a *même loi* que $S_n = X_1 + \cdots + X_n$, où les X_k sont des variables aléatoires de Bernoulli indépendantes et de même paramètre p et en rappelant que l'espérance et la variance de la loi $\mathrm{Bin}(n,p)$ sont respectivement np et npq.

La démonstration historique du théorème de de Moivre-Laplace repose sur un bon contrôle des coefficients binomiaux *via* la formule de Stirling[35]. L'intérêt de

35. Cette formule résulte d'une amicale compétition entre de Moivre et Stirling et fut découverte à cette occasion.

cette approche « élémentaire » est de donner une idée de la vitesse de convergence qui est en $O(n^{-1/2})$, voir [8] chap. 7.

Le TLC a de multiples applications, notamment en statistique. À ce stade, on peut souligner deux idées.

D'abord, on peut noter que le comportement asymptotique en loi de S_n^* *ne dépend pas de la loi* de X_1. La seule condition pour que la loi de S_n^* soit approximativement gaussienne pour les grandes valeurs de n est que X_1 soit de carré intégrable. Ceci donne un caractère universel aux lois gaussiennes et explique la fréquence de l'utilisation de ces lois en modélisation [36]. On peut dire que le comportement asymptotique en loi de sommes S_n^* et donc aussi de S_n « oublie » tout de la loi des X_i, sauf le paramètre de localisation $\mu = \mathrm{E}\, X_1$ et le paramètre de dispersion $\sigma^2 = \mathrm{Var}\, X_1$. C'est l'une des raisons de l'importance donnée à ces deux paramètres en théorie des probabilités.

La deuxième idée importante est que le TLC donne une idée de la vitesse de convergence dans la loi faible des grands nombres. *Grosso modo*, on peut dire que dans le bon cas où $\mathrm{E}\, X_1^2 < +\infty$, cette vitesse est en $O(n^{-1/2})$. Précisons le sens de cette affirmation. Par la conséquence pratique du TLC mentionnée p. 86 appliquée avec $[a, b] = [-t, t]$, on obtient :

$$\lim_{n\to+\infty} P\big(S_n^* \in [-t, t]\big) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1,$$

en utilisant la relation $\Phi(-t) = 1 - \Phi(t)$ due à la parité de la densité de $\mathfrak{N}(0, 1)$. En remarquant maintenant que

$$S_n^* = \frac{S_n - n\,\mathrm{E}\, X_1}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\Big(\frac{S_n}{n} - \mathrm{E}\, X_1\Big),$$

on peut réécrire la convergence précédente sous la forme

$$P\left(\left|\frac{S_n}{n} - \mathrm{E}\, X_1\right| \leqslant \frac{\sigma t}{\sqrt{n}}\right) = 2\Phi(t) - 1 + \varepsilon_n,$$

où ε_n est une suite de réels (pas forcément positifs), convergente vers 0. Pour tout $\delta > 0$, on peut choisir un $t = t(\delta)$ assez grand pour que $2\Phi(t) - 1 > 1 - \delta/2$ car $2\Phi(t) - 1$ tend vers 1 quand t tend vers $+\infty$. Ensuite pour $n \geqslant n_0(\delta)$, on aura $|\varepsilon_n| < \delta/2$ et finalement

$$\forall \delta > 0,\ \exists t(\delta), n(\delta),\ \forall n \geqslant n(\delta), \quad P\left(\left|\frac{S_n}{n} - \mathrm{E}\, X_1\right| \leqslant \frac{\sigma t(\delta)}{\sqrt{n}}\right) > 1 - \delta.$$

C'est en ce sens que l'on peut dire que S_n/n converge vers $\mathrm{E}\, X_1$ avec une vitesse en $O(n^{-1/2})$. On peut résumer ce résultat par l'écriture $|n^{-1}S_n - \mathrm{E}\, X_1| = O_P(n^{-1/2})$, dont le deuxième membre se lit « grand O en probabilité de $n^{-1/2}$ ».

36. D'autant plus qu'il existe de nombreuses généralisations du théorème limite central, avec des v.a. indépendantes mais de lois différentes, avec des vecteurs aléatoires, avec des v.a. « faiblement dépendantes » ...

Dans l'utilisation pratique du TLC, on travaille souvent avec n « grand » fixé et on approxime la loi de S_n^* par la loi $\mathfrak{N}(0,1)$, ou ce qui revient au même, *on approxime la loi* de S_n par la *loi gaussienne* $\mathfrak{N}(n\mu, \sigma\sqrt{n})$ *de même espérance et même variance* que S_n ($\mathrm{E}\, S_n = n\,\mathrm{E}\, X_1 = n\mu$ et $\operatorname{Var} S_n = n \operatorname{Var} X_1 = n\sigma^2$). Plus précisément, en notant que $g_n : x \mapsto \frac{x - n\,\mathrm{E}\, X_1}{\sigma\sqrt{n}}$ est une bijection croissante de $\mathbb{R}$ sur $\mathbb{R}$ et en posant pour $a < b$ réels,

$$a_n = g_n(a) = \frac{a - n\,\mathrm{E}\, X_1}{\sigma\sqrt{n}}, \quad b_n = g_n(b) = \frac{b - n\,\mathrm{E}\, X_1}{\sigma\sqrt{n}},$$

on a

$$P(a \leqslant S_n \leqslant b) = P(a_n \leqslant S_n^* \leqslant b_n) = \Phi(b_n) - \Phi(a_n) + \varepsilon_n.$$

On néglige alors le terme d'erreur ε_n et on termine le calcul en utilisant la table des valeurs de Φ, fonction de répartition de la loi normale standard.

Vitesse de convergence dans le TLC

La question qui se pose dans le calcul précédent est « que signifie n *grand* ? », ou encore « comment peut-on contrôler l'erreur ε_n ? », autrement dit, quelle est la vitesse de convergence vers 0 de ε_n ? La réponse est que dans le « bon cas » où X_1 a un moment d'ordre 3, la vitesse de convergence dans le théorème limite central est en $O(n^{-1/2})$.

Théorème (Berry-Esséen, 1941–42). *Soit $(X_i)_{i\geqslant 1}$ une suite de variables aléatoires indépendantes et de même loi telle que $\mathrm{E}\,|X_i|^3 < +\infty$. On note $\sigma^2 := \operatorname{Var} X_1$, $\rho^3 := \mathrm{E}\,|X_1 - \mathrm{E}\, X_1|^3$, avec $\sigma > 0$ et $\rho > 0$. Il existe alors une constante universelle $C > 0$ telle que pour tout $n \geqslant 1$,*

$$\Delta_n := \sup_{x\in\mathbb{R}} |P(S_n^* \leqslant x) - \Phi(x)| \leqslant C\frac{\rho^3}{\sigma^3}\frac{1}{\sqrt{n}}.$$

Pour une preuve du théorème, voir [2]. L'obtention de la meilleure constante C a été l'objet d'une longue quête. La valeur initiale de Esséen était $C = 7{,}59$. Une valeur plus moderne et proche de l'optimale est $C = 0{,}469\,1$ (Shevtsova (2013)).

Il est intéressant de regarder ce que donne le théorème de Berry-Esséen pour le cas de de Moivre-Laplace, donc avec des X_i suivant la loi de Bernoulli de paramètre p. En calculant ρ^3 on vérifie facilement que

$$\Delta_n \leqslant C\frac{p^2 + q^2}{\sqrt{pq}}\frac{1}{\sqrt{n}}, \quad q := 1 - p.$$

Il est intéressant de noter comment cette majoration se dégrade quand p est proche de 0 ou de 1. On dispose en fait de résultats plus précis d'Uspensky [10] concernant ce cas particulier, cf. [8, chap. 7].

Statistique Inférentielle

Intervalle de fluctuation

`Les_intervalles_de_fluctuation_de_la_seconde_a_la_terminale.pdf` présente à partir d'activités les trois versions de la notion d'intervalle de fluctuation dans les programmes et leurs applications, ainsi que des compléments théoriques. Dans ce qui suit, nous regardons les choses sous un angle un peu différent.

Concentration de la mesure

Le phénomène de concentration de la mesure (ou de la loi de probabilité d'une variable aléatoire) permet en statistique, d'obtenir de l'information, voire des quasi certitudes, à partir d'observations aléatoires. Pour l'illustrer graphiquement, comparons les diagrammes en bâtons de la loi uniforme sur $[\![0, 100]\!]$ et de la loi binomiale de paramètres 100 et 0,3 représentés figure 17. Rappelons que dans le diagramme en bâtons de la loi d'une variable aléatoire discrète X, le segment vertical d'abscisse x_k (ou « bâton ») a pour hauteur $P(X = x_k)$. Pour les deux lois considérées ici, il y a théoriquement un bâton de hauteur non nulle pour chacune des valeurs entières $x_k = k \in [\![0, 100]\!]$. Pour la loi uniforme, tous les bâtons ont la même hauteur $1/101$, tandis que pour la loi binomiale, les bâtons d'abscisse en dehors de l'intervalle $[\![12, 48]\!]$ ont une hauteur trop petite pour être visibles. On peut d'ailleurs vérifier par le calcul que la somme des hauteurs de tous ces bâtons « invisibles » vaut 0,000 057 5. Autrement dit, si la variable aléatoire S suit la loi binomiale de paramètres 100 et 0,3, $P(12 \leqslant S \leqslant 48) = 0{,}999\,942\,5$. Par contre, si U suit la loi uniforme sur $[\![0, 100]\!]$, $P(12 \leqslant U \leqslant 48) = 37/101 \simeq 0{,}366\,4$. Si avant d'observer une valeur de S ou de U, on parie qu'elle va tomber dans $[\![12, 48]\!]$, on est pratiquement sûr de gagner avec S et on a presque 2 chances sur 3 de perdre avec U. C'est cette concentration de la probabilité sur un intervalle court (relativement à la longueur du support) pour S qui permet de faire de l'estimation ou de la prise de décision à partir des observations.

Ce phénomène de concentration de la loi binomiale s'accentue quand n augmente. Pour le visualiser, regardons le diagramme en bâtons d'une loi binomiale avec $p = 0{,}3$ pour de grandes valeurs de n. On se contentera ici de $n = 1000$ (on peut encore calculer simplement les coefficients binomiaux pour cette valeur en Scilab, en utilisant le triangle de Pascal).

— Sur $[\![0, n]\!]$, cf. figure 18, ceci illustre la *loi faible des grands nombres*, à condi-

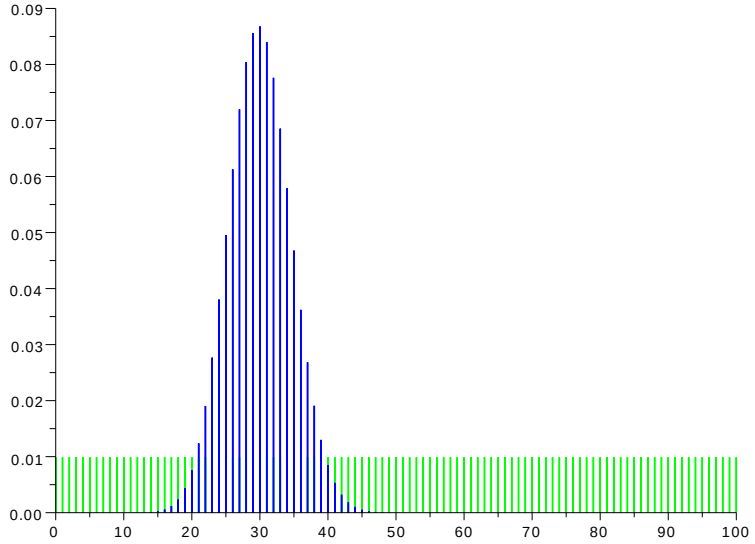

FIGURE 17 – Diagrammes en bâtons des lois $\mathrm{Bin}(100; 0,3)$ et $\mathrm{Unif}[\![0, 100]\!]$

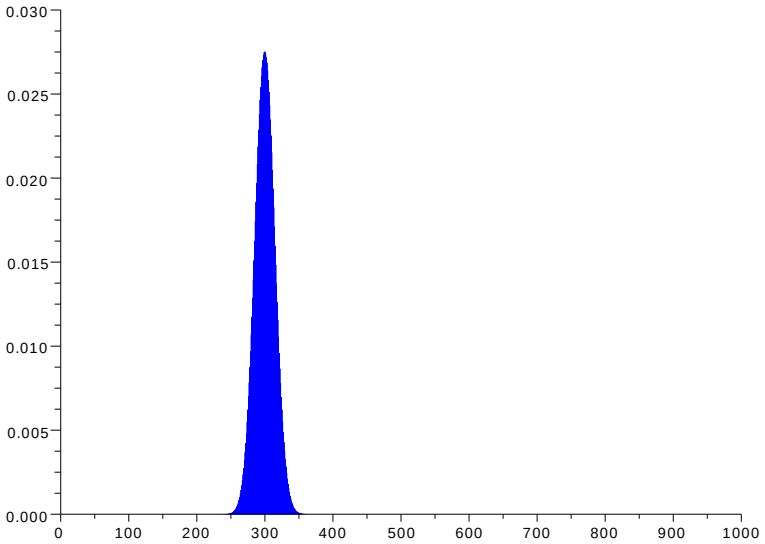

FIGURE 18 – Diagrammes en bâtons de $\mathrm{Bin}(1000; 0,3)$, loi faible des grands nombres

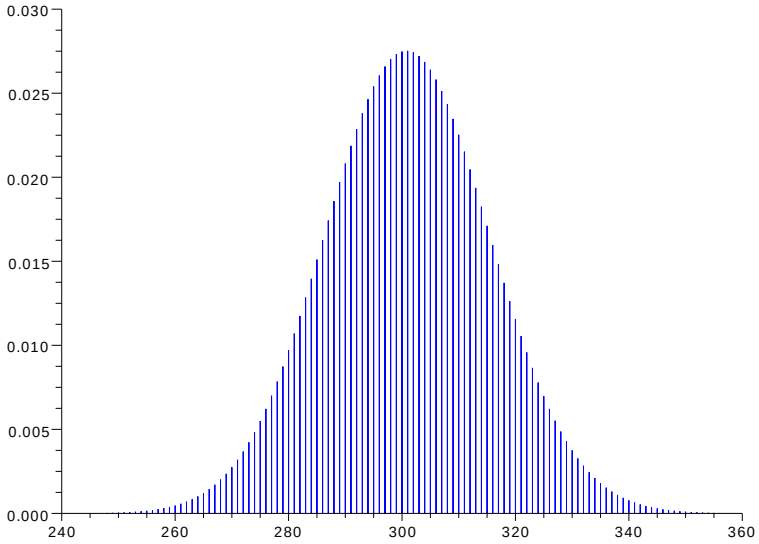

FIGURE 19 – Diagrammes en bâtons de Bin(1000; 0, 3), théorème limite central

tion de faire par la pensée une mise à l'échelle en $1/n$, ce qui revient ici à décréter que la graduation horizontale 1000 correspond à 1.

— Sur $[np - 4\sqrt{np(1-p)}, np + 4\sqrt{np(1-p)}]$, cf. figure 19, ceci illustre le théorème de de Moivre Laplace).

Notons que le passage de la figure 18 à la figure 19 n'est rien d'autre qu'un changement d'échelle horizontale. Le comportement mathématique que nous voulons illustrer par ces figures est le suivant. Si S_n est une variable aléatoire de loi binomiale de paramètres n et p, alors S_n/n converge *en probabilité* vers p quand n tend vers l'infini. C'est l'exemple le plus simple de loi faible des grands nombres et qui est la justification de l'estimation d'une probabilité inconnue par la fréquence de réalisation observée sur un grand échantillon. De manière équivalente, on peut aussi dire que $S_n/n - p$ converge en probabilité vers 0. En pratique, il est utile d'avoir une idée de la vitesse de convergence. Pour cela, on regarde le comportement de

$$Z_n = c_n \left(\frac{S_n}{n} - p \right)$$

où (c_n) est une suite de constantes (i.e. non aléatoires) tendant vers l'infini. Intuitivement, si c_n tend trop lentement vers $+\infty$, cela ne va rien changer et Z_n convergera vers 0 en probabilité. Si c_n tend trop vite vers l'infini, il y aura explosion avec oscillation indéfinie de Z_n entre $-\infty$ et $+\infty$ par amplification des fluctuations aléatoires de S_n/n autour de p. Il se trouve qu'il y a une situation intermédiaire quand c_n est de l'ordre de grandeur de $\sqrt{n}$, où Z_n ne converge plus en probabilité vers 0, mais n'ex-

plose pas pour autant. La loi de Z_n reste pour l'essentiel concentrée sur un intervalle de taille constante. Il y a en fait *convergence en loi* de Z_n vers une variable[37] Z de loi gaussienne de paramètres 0 et $\sigma = \sqrt{p(1-p)}$. Pour une telle variable aléatoire Z, $P(-4\sigma \leqslant Z \leqslant 4\sigma) \simeq 0{,}999\,946\,7$. En revenant au comportement de S_n plutôt qu'à celui de S_n/n (cela revient à tout multiplier par n) ceci explique pourquoi le choix de « zoomer » sur l'intervalle $[np - 4\sqrt{np(1-p)}, np + 4\sqrt{np(1-p)}]$ a permis de capturer et visualiser l'essentiel de la masse de la loi de S_n.

Si l'on doit retenir une idée simple de tout cela, c'est que l'amplitude des fluctuations de S_n autour de son espérance np est de l'ordre de $\sqrt{n}$. Cette idée est à la base de l'introduction de la notion d'intervalle de fluctuation pour une fréquence (de largeur de l'ordre de $1/\sqrt{n}$) dès la Seconde. Même si la définition de l'intervalle de fluctuation donnée en Seconde n'est pas totalement satisfaisante pour un mathématicien, elle donne d'assez bons résultats en pratique et permet d'aller directement à l'essentiel : l'exploitation du phénomène de concentration de la mesure (ici de la loi binomiale) comme aide statistique à la prise de décision.

Aide à la prise de décision

Commençons par un exemple un peu artificiel et volontairement simpliste.

Problème à deux urnes. On dispose de deux urnes d'apparence complètement identique, numérotées 1 et 2, le numéro étant masqué. On sait que l'urne 1 contient 30% de boules vertes et 70% de boules rouges tandis que l'urne 2 contient 78% de boules vertes et 22% de boules rouges. On en choisit une au hasard dans laquelle on effectue 100 tirages avec remise d'une boule. On note le nombre S d'apparitions des boules vertes lors de cette suite de tirages et on doit dire, au vu de ce nombre de boules vertes, si le numéro de l'urne choisie est ou non le 1.

Si les tirages ont lieu dans l'urne 1, la loi de S est la binomiale de paramètres 100 et 0,3, tandis qu'avec des tirages dans l'urne 2, la loi de S est la binomiale de paramètres 100 et 0,78. Il s'agit donc de mettre en concurrence ces deux lois possibles pour S, *au vu de l'observation d'une valeur de* S. Un coup d'oeil sur les diagrammes en bâtons (figure 20) montre qu'il n'y a pas photo et que même si ces deux lois ont le même support théorique à savoir $[\![0, 100]\!]$, en pratique elles ne vivent pas sur le même territoire. Les intervalles de fluctuation au seuil de 95% sont (approximativement) $[\![20, 40]\!]$ pour la première et $[\![68, 88]\!]$ pour la seconde[38]. Au vu du graphique, on se convaincra facilement que pour l'urne 1, la probabilité que S prenne ses valeurs dans $[\![10, 50]\!]$ est « pratiquement » égale à 1, tandis que pour l'urne 2 il en va de même avec l'intervalle $[\![60, 95]\!]$. Quel que soit le numéro de l'urne choisie, il est donc très peu vraisemblable que la valeur de S observée soit en dehors de l'un de ces deux intervalles. Si elle est dans le premier, on pourra conclure avec

37. L'expression « convergence en loi » d'une suite de *variables* aléatoires est un grossier (mais usuel) abus de langage. En fait, c'est de convergence de la *loi* de Z_n qu'il s'agit. Z_n converge tout aussi bien en loi vers n'importe quelle autre variable aléatoire Z' ayant même loi que Z.

38. Intervalles de fluctuation obtenus par la formule approchée $[np - \sqrt{n}, np + \sqrt{n}]$.

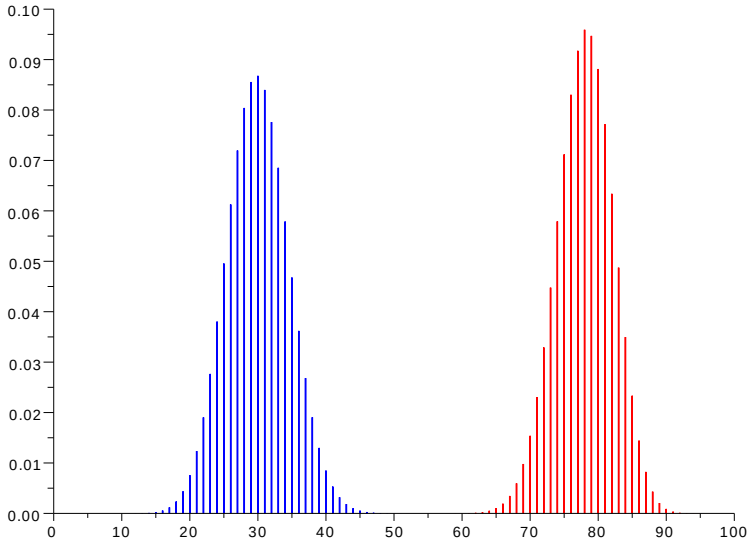

FIGURE 20 – Diagrammes en bâtons de $\mathrm{Bin}(100; 0,3)$ et $\mathrm{Bin}(100; 0,78)$

un risque d'erreur infime que l'urne choisie était la numéro 1. Si S est dans $[\![60, 95]\!]$, on concluera de même que l'urne choisie était la numéro 2. En remplaçant ces deux intervalles par les intervalles de fluctuation au seuil 95%, la règle de décision sera la même, avec un risque d'erreur un peu plus grand, proche de 5%. Attention, nous avons écrit « risque » et non pas « probabilité », car il y a ici deux probabilités et deux types d'erreur possible, chacune pouvant être mesurée avec l'une de ces deux probabilités.

Test sur la valeur d'une probabilité. Une situation plus complexe, mais plus proche de problèmes réels, est celle où on a une seule urne et *des raisons de croire* que sa composition est (par exemple) de 30% de boules vertes et 70% de rouges. On effectue à nouveau 100 tirages et au vu de la valeur du nombre S d'apparitions des boules vertes lors de cette suite de tirages, on doit décider si on conserve ou rejette cette hypothèse sur la composition de l'urne. C'est un problème de test sur la valeur d'une proportion (ou plus généralement d'une probabilité). Là encore, on peut utiliser l'intervalle de fluctuation pour se donner une règle de décision. Puisque l'intervalle de fluctuation au seuil de 95% est $[\![20, 40]\!]$, l'observation d'une valeur de S « trop petite » (ici strictement inférieure à 20) ou « trop grande » (ici strictement supérieure à 40) conduira à rejetter l'hypothèse d'une composition d'urne à 30% de boules vertes. On peut parler ici de zone d'acceptation *bilatérale*.

Dans certaines situations, il est naturel de rejetter l'hypothèse faite sur une proportion (ou une probabilité), seulement lorsque S prend une valeur « trop grande »

c'est le cas par exemple avec les problèmes de questionnaire à choix multiple. On prend comme hypothèse que le candidat a répondu au hasard à chaque question. Le nombre S de bonnes réponses est alors une variable aléatoire de loi binomiale avec pour paramètres n le nombre de questions et $p = c/d$ où d est le nombre de réponses proposées par question dont c sont correctes (le plus souvent, $c = 1$). On décidera alors de rejetter l'hypothèse si la valeur observée de S est supérieure ou égale à b, où b est le plus petit entier tel que $P(S \leqslant b) > 0{,}95$. Ici la zone d'acceptation sera $[\![0, b-1]\!]$, c'est une zone unilatérale[39].

Dans le problème à deux urnes décrit ci-dessus, la séparation des masses entre les deux lois binomiales mises en concurrence (cf. figure 20) était bien nette. Par contre dans le problème suivant avec une seule urne, la situation est bien plus complexe. En effet, il s'agit maintenant de *mettre en concurrence* la loi binomiale $\mathrm{Bin}(100; 0{,}3)$ correspondant à la composition d'urne que l'on cherche à « tester » avec *toutes les autres lois binomiales*, $\mathrm{Bin}(100, p)$ pour $p \neq 0{,}3$ (pour p rationnel de $]0, 1[$ puisqu'il s'agit d'une proportion[40]). Et là, on voit poindre une difficulté. Si p est trop « proche » de 0,3, les intervalles de fluctuation de $\mathrm{Bin}(100; 0{,}3)$ et de $\mathrm{Bin}(100, p)$ vont largement se recouvrir et la règle de décision proposée ci-dessus perdra beaucoup de sa pertinence. Alors que faire ?

La première réponse est d'augmenter le nombre de tirages n. En effet, en prenant n assez grand, on arrivera toujours à rendre disjoints les intervalles de fluctuation au seuil de 95% de $\mathrm{Bin}(n; 0{,}3)$ et de $\mathrm{Bin}(n, p)$:

— pour $p < 0{,}3$, il suffit que $np + \sqrt{n} < 0{,}3n - \sqrt{n}$;

— pour $p > 0{,}3$, il suffit que $np - \sqrt{n} > 0{,}3n + \sqrt{n}$.

Ces deux conditions peuvent se résumer par :

$$n > \frac{4}{(p - 0{,}3)^2}.$$

Par exemple avec $p = 0{,}4$, il faudrait au moins 400 tirages pour avoir des intervalles de fluctuation disjoints, avec $p = 0{,}31$, il en faudrait 40 000.

La seconde réponse est que la première n'est pas réaliste ! En effet, en pratique, il faut bien se fixer un nombre de tirages et comme on ne connaît pas p, on ne saura jamais si on a une bonne séparation des intervalles de fluctuation, comme dans le problème des deux urnes ci-dessus. Autrement dit, il faut se faire à l'idée que lorsque l'on prétend valider l'hypothèse que la proportion inconnue est 0,3, la valeur 0,3 est donnée avec une certaine précision, dépendant du nombre d'observations que l'on peut se permettre et que des valeurs proches seraient aussi acceptables.

39. En fait dans cet exemple, le caractère unilatéral vient de ce que l'examinateur se demande si le candidat a fait *mieux* que de répondre au hasard. Et si le nombre de bonnes réponses obtenu est vraiment petit, par exemple $S = 0$, il convient de s'interroger sur le comportement du candidat. Est-il particulièrement malchanceux ? Ou est-ce un « rebelle » qui a fait exprès de choisir une mauvaise réponse à chaque question ?

40. Donc cela fait une infinité de lois concurrentes. En réalité, si on s'accorde sur un volume maximal d'urne, par exemple 1 m^3 et un diamètre minimal des boules, par exemple 1 mm, cela n'en fait plus qu'un nombre fini, mais quand même très grand.

Intervalle de confiance

Une importante application du TLC en statistique est la construction d'intervalles de confiance pour un paramètre inconnu au vu d'un échantillon observé. Nous nous contenterons à ce stade du cas de l'estimation par intervalle de confiance d'une probabilité inconnue, présenté via l'exercice suivant.

Calibrage de pommes

Une coopérative agricole a un contrat de fourniture de pommes de catégorie A, c'est-à-dire dont le diamètre en mm est dans l'intervalle $[67, 73]$. Le gérant de la coopérative a besoin d'évaluer rapidement [41] la proportion p de pommes *hors catégorie* A dans la récolte qu'il vient d'emmagasiner. Pour cela, il prélève au hasard un échantillon de 400 pommes dont une calibreuse mécanique lui permet d'enregistrer les diamètres. On dénombre 70 pommes hors catégorie dans cet échantillon de taille 400. À partir de cette observation, nous allons construire deux intervalles de confiance au niveau 95% pour la proportion inconnue p de pommes hors catégorie dans la population totale.

Réduction à une situation binomiale. Remarquons d'abord que l'échantillon observé résulte d'un tirage *sans remise* de 400 individus dans une population de grande taille [42] N. En toute rigueur, la loi du nombre S_n de pommes hors catégorie dans l'échantillon de taille $n = 400$ est donc hypergéométrique de paramètres N, pN et n, mais vu la taille de la population, il est légitime d'approximer cette loi par la loi binomiale de paramètres $n = 400$ et p (inconnue) qui serait celle de S_n si les observations faites résultaient d'un tirage *avec remise*. Dans toute la suite, nous supposerons donc que S_n suit la loi binomiale $\mathrm{Bin}(400, p)$.

Méthode avec variance majorée. En notant

$$S_n^* := \frac{S_n - np}{\sqrt{np(1-p)}} = \sqrt{\frac{n}{p(1-p)}}\left(\frac{S_n}{n} - p\right),$$

la somme centrée réduite, le théorème de de Moivre-Laplace nous dit que

$$\forall t > 0, \quad P(|S_n^*| \leqslant t) \xrightarrow[n\to+\infty]{} P(|Z| \leqslant t) = \Phi(t) - \Phi(-t) = 2\Phi(t) - 1,$$

en rappelant que Φ désigne la f.d.r. de la loi normale standard $\mathfrak{N}(0,1)$ et en utilisant pour la dernière égalité la parité de la densité de $\mathfrak{N}(0,1)$. On peut réécrire cette convergence sous la forme :

$$\forall n \geqslant 1, \forall t > 0, \quad P(-t \leqslant S_n^* \leqslant t) = 2\Phi(t) - 1 + \varepsilon_n(t), \quad \varepsilon_n(t) \xrightarrow[n\to+\infty]{} 0,$$

41. Avant de lancer l'opération de calibrage et d'emballage.

42. En comptant 6 pommes au kg, cela donnerait environ $N = 600\,000$ pommes pour une récolte de 100 tonnes.

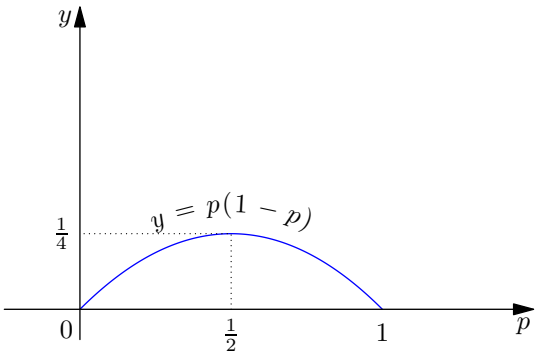

FIGURE 21 – Majoration de la variance d'une loi de Bernoulli

où $\varepsilon_n(t)$ est l'erreur d'approximation gaussienne.

En résolvant par rapport à p l'encadrement $-t \leqslant S_n^* \leqslant t$, on voit que

$$-t \leqslant S_n^* \leqslant t \quad \Longleftrightarrow \quad \frac{S_n}{n} - t\sqrt{\frac{p(1-p)}{n}} \leqslant p \leqslant \frac{S_n}{n} + t\sqrt{\frac{p(1-p)}{n}}.$$

Cet encadrement n'est pas satisfaisant pour construire un intervalle de confiance pour p car *les bornes dépendent de* p via la quantité inconnue $p(1-p)$ qui est la variance d'une v.a. de Bernoulli de paramètre p. La méthode avec variance majorée consiste à se débarasser de cet inconvénient en notant que la fonction $g : p \mapsto p(1-p)$ atteint son maximum sur $[0,1]$ au point $p_0 = 1/2$ (figure 21) et est donc majorée par $g(p_0) = 1/4$. Ainsi pour tout $p \in [0,1]$, $\sqrt{p(1-p)} \leqslant \sqrt{1/4} = 1/2$, d'où :

$$-t \leqslant S_n^* \leqslant t \quad \Longrightarrow \quad \frac{S_n}{n} - \frac{t}{2\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t}{2\sqrt{n}}.$$

Cette implication se traduit par l'inclusion d'évènements $A_{n,t} \subset B_{n,t}$, en notant

$$A_{n,t} = \{-t \leqslant S_n^* \leqslant t\}, \quad B_{n,t} = \left\{ \frac{S_n}{n} - \frac{t}{2\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t}{2\sqrt{n}} \right\}.$$

On en déduit que

$$P(B_{n,t}) \geqslant P(A_{n,t}) = 2\Phi(t) - 1 + \varepsilon_n(t).$$

On cherche t tel que $2\Phi(t) - 1 = 0{,}95$, c'est-à-dire $\Phi(t) = 0{,}975$, d'où $t = 1{,}96$ (par lecture inverse de la table des valeurs de Φ). Considérant alors que $n = 400$ est suffisamment grand pour que l'on puisse négliger l'erreur d'approximation gaussienne $\varepsilon_n(t)$, on obtient finalement $P(B_{n,t}) \geqslant 0{,}95$ pour $n = 400$ et $t = 1{,}96$. En introduisant l'intervalle *aléatoire* :

$$I_{n,t} = \left[\frac{S_n}{n} - \frac{t}{2\sqrt{n}}, \frac{S_n}{n} + \frac{t}{2\sqrt{n}} \right]$$

on peut réécrire ceci sous la forme

$$P(p \in I_{n,t}) \geqslant 0{,}95 \quad \text{pour } t = 1{,}96 \text{ et } n = 400.$$

On pourra dire que $I_{n,t}$ est un intervalle de confiance *théorique* au niveau approximatif $2\Phi(t) - 1$, soit 95% pour $t = 1{,}96$.

En réalité, ce que l'on a observé, c'est *une réalisation particulière* $S_n(\omega_0) = 70$ de la variable aléatoire S_n et comme on ne connaît pas p, on ne peut pas dire avec certitude si le ω_0 sous-jacent [43] appartient ou non à $B_{n,t}$. On « parie » donc sur la réalisation de $B_{n,t}$ et d'après l'étude précédente, la probabilité de gagner ce pari est (approximativement) au moins 95%. L'intervalle $I = I_{n,t}(\omega_0)$ est un intervalle de confiance *numérique* [44] pour p au niveau 95% :

$$I = \left[\frac{70}{400} - \frac{1{,}96}{2\sqrt{400}}; \frac{70}{400} + \frac{1{,}96}{2\sqrt{400}}\right] = [0{,}126; 0{,}224].$$

Remarque. Attention à l'erreur classique « $P(p \in [0{,}126; 0{,}224]) = 0,95$ ». Quand on écrit « $p \in I_{n,t}$ », il s'agit de l'évènement $B_{n,t}$ auquel on peut attribuer une probabilité. Explicitement, $B_{n,t} = \{\omega \in \Omega;\ \frac{S_n(\omega)}{n} - \frac{t}{2\sqrt{n}} \leqslant p \leqslant \frac{S_n(\omega)}{n} + \frac{t}{2\sqrt{n}}\}$. Mais dès que l'on remplace $I_{n,t}$ par I, les bornes de l'intervalle ne dépendent plus de ω et comme p est inconnu mais pas aléatoire (p ne dépend pas de ω), l'ensemble $\{\omega \in \Omega;\ 0{,}126 \leqslant p \leqslant 0{,}224\}$ ne peut être que $\varnothing$ (si p ne vérifie pas l'encadrement) ou Ω (si p le vérifie). Si l'on considère cet ensemble comme un évènement, sa probabilité ne peut être que 0 ou 1, mais certainement pas 0,95. ◁

Méthode avec variance estimée. Au lieu de majorer $p(1-p)$, on l'estime par

$$V_n = \frac{S_n}{n}(1 - \frac{S_n}{n}).$$

On vérifie facilement grâce à la loi forte des grands nombres que V_n converge presque sûrement, donc aussi en probabilité, vers $p(1-p)$. En notant

$$C_{n,t} = \left\{\frac{S_n}{n} - \frac{t\sqrt{V_n}}{\sqrt{n}} \leqslant p \leqslant \frac{S_n}{n} + \frac{t\sqrt{V_n}}{\sqrt{n}}\right\},$$

on obtient

$$P(C_{n,t}) = 2\Phi(t) - 1 + \varepsilon'_n(t),$$

43. Ici ω_0 représente les résultats observés de la suite de tirages effectués. Cela peut être une suite binaire si on code par 1 une pomme hors-catégorie et par 0 une pomme de catégorie A, ou une suite de réels si on enregistre vraiment le diamètre de chaque pomme prélevée, etc.

44. Les appellations intervalle de confiance *théorique* ou *numérique* ne sont pas standard et la plupart du temps dans la littérature statistique, on n'explicite pas la distinction, le contexte permettant à un lecteur un peu familier du sujet de lever l'ambiguïté.

puis avec le choix $t = 1{,}96$, $P(C_{n,t}) \simeq 0{,}95$. L'intervalle de confiance numérique correspondant est

$$J = [0{,}137; 0{,}213].$$

Contrairement aux apparences, ce n'est pas de la cuisine, à cause du théorème suivant.

Théorème (TLC avec autonormalisation). *Soient $X_1, \dots, X_n, \dots$ des variables aléatoires indépendantes et de même loi telle que $\mathrm{E}\, X_1^2 < +\infty$ et $\sigma^2 := \mathrm{Var}\, X_1 > 0$. On note $S_n := X_1 + \cdots + X_n$. On suppose de plus que $(V_n)_{n\geqslant 1}$ est une suite de variables aléatoires positives qui converge en probabilité vers σ^2. Alors*

$$T_n := \sqrt{\frac{n}{V_n}} \left(\frac{S_n}{n} - \mathrm{E}\, X_1 \right) \xrightarrow[n\to+\infty]{\text{loi}} Z,$$

où Z suit la loi gaussienne standard $\mathfrak{N}(0,1)$. Comme la f.d.r. Φ de la loi $\mathfrak{N}(0,1)$ est continue sur $\mathbb{R}$, cette convergence en loi équivaut à

$$\forall x \in \mathbb{R}, \quad P(T_n \leqslant x) \xrightarrow[n\to+\infty]{} P(Z \leqslant x) = \Phi(x).$$

En général on applique ce théorème en prenant pour V_n la « variance empirique ». Cette variance empirique est pour chaque ω, la variance calculée sur la série statistique réellement observée $x_1 = X_1(\omega), \dots, x_n = X_n(\omega)$. C'est donc la variable aléatoire :

$$V_n = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - (\overline{X})^2,$$

où $\overline{X} = S_n/n$. En appliquant deux fois la loi forte des grands nombres, on voit facilement que la variance empirique converge presque sûrement (donc aussi en probabilité) vers la variance théorique $\sigma^2 = \mathrm{E}\, X_1^2 - (\mathrm{E}\, X_1)^2$.

Si on revient au cas où les X_i sont des variables de Bernoulli, on peut appliquer directement le théorème ci-dessus avec $V_n = \overline{X}(1-\overline{X})$ en notant que par la loi forte des grands nombres, V_n converge presque-sûrement vers $p(1-p) = \sigma^2$. On peut aussi remarquer que pour des variables de Bernoulli, la variance empirique n'est autre que $\overline{X}(1-\overline{X})$. La vérification de cette affirmation est facile une fois que l'on a noté que si X_i ne prend que les valeurs 0 et 1, $X_i = X_i^2$.

Épilogue

Nous terminerons ce livret en évoquant brièvement quelques questions qui prolongent naturellement sa lecture et renforcent l'unité de ses trois parties. Il s'agit essentiellement de préciser la problématique statistique. En deux mots, il s'agit à partir de l'observation d'un *échantillon* provenant d'une loi inconnue, d'obtenir de l'information sur cette loi (problème d'*estimation*) ou plus généralement de prendre

une décision en contrôlant au mieux les risques d'erreur (problème de *test*). Le résultat préliminaire important est le théorème de Glivenko-Cantelli, appelé quelquefois *théorème fondamental de la statistique*. Ce théorème permet de justifier l'idée intuitive que l'on peut reconstruire une loi inconnue à partir d'observations, avec une approximation d'autant meilleure que le nombre d'observations est plus grand. Nous introduirons ensuite le *modèle statistique* permettant de « mathématiser » les questions d'estimation et de prise de décision.

Mesure empirique et théorème de Glivenko-Cantelli

Définition (échantillon d'une loi). Soit $(\Omega, \mathcal{F}, P)$ un espace probabilisé et Q une mesure de probabilité sur $\mathbb{R}$ muni de sa tribu borélienne. Pour $n \geqslant 2$, on appelle n-échantillon de la loi Q, associé à $(\Omega, \mathcal{F}, P)$, toute suite finie $X_1, \ldots, X_n$ de variables aléatoires définies sur $(\Omega, \mathcal{F}, P)$, P-indépendantes et de même loi Q, c.-à-d. pour tout $i \in [\![1, n]\!]$, $P_{X_i} = P \circ X_i^{-1} = Q$. ◁

Quand il n'apparaît pas utile de préciser l'espace $(\Omega, \mathcal{F}, P)$ concerné, on parle plus simplement d'échantillon de la loi Q. Notons d'ailleurs qu'étant donné une loi quelconque Q, autrement dit une mesure de probabilité sur $\mathbb{R}$ muni de sa tribu borélienne, il est toujours possible de construire un espace probabilisé $(\Omega, \mathcal{F}, P)$ et une suite $X_1, \ldots, X_n$ de v.a. définies sur cet espace qui constituent un n-échantillon de la loi Q. Il suffit de prendre $\Omega = \mathbb{R}^n$, muni de sa tribu borélienne et de la probabilité produit $Q^{\otimes n}$ définie à partir des probabilités des pavés $C =]a_1, b_1] \times \cdots \times]a_n, b_n]$ en posant $Q^{\otimes n}(C) := Q(]a_1, b_1])\ldots Q(]a_n, b_n])$. En définissant pour $i = 1, \ldots, n$, X_i comme la projection $\mathbb{R}^n \to \mathbb{R}$ sur la i^{e} coordonnée, on vérifie facilement que $X_1, \ldots, X_n$ est un n-échantillon de la loi Q. Ce choix est essentiellement celui fait par les programmes du lycée. Il présente l'avantage d'éviter de parler de vecteur aléatoire $(X_1, \ldots, X_n)$ puisqu'en fait on ne considère que la moyenne arithmétique $n^{-1}(X_1 + \cdots + X_n)$ vue comme fonction sur l'espace des réalisations possibles $(x_1, \ldots, x_n)$ de l'échantillon[45]. L'inconvénient est de figer n, ce qui ne permet pas d'étudier ce qui se passe quand n tend vers l'infini *en conservant le même* Ω. En pratique cela ne permet pas de parler de loi *forte* des grands nombres.

Un outil fondamental en statistique est la *mesure empirique* que l'on peut décrire de manière informelle comme suit. On a des observations $x_1, \ldots, x_n$, que l'on interprète comme les $X_1(\omega), \ldots, X_n(\omega)$, où les v.a. X_i suivent la loi inconnue μ qui est une probabilité sur $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$. Faute d'information supplémentaire, on décide de « ne pas faire de jalouses » parmi les observations en attribuant à chacune la probabilité $1/n$. On construit ainsi sur l'espace $(\mathbb{R}, \mathrm{Bor}(\mathbb{R}))$ une *nouvelle* mesure de probabilité μ_n, dépendant des observations et on se sert de μ_n pour estimer la mesure de probabilité inconnue μ.

45. Dans cette modélisation, pour $\omega = (x_1, \ldots, x_n)$, $X_i(\omega) = x_i$ et donc le vecteur aléatoire $(X_1, \ldots, X_n) : \Omega \to \mathbb{R}^n$ est simplement l'identité sur $\mathbb{R}^n$.

Définition (mesure empirique). Soit $X_1, \ldots, X_n$ un n-échantillon. On appelle *mesure empirique* associée à cet échantillon, la « mesure aléatoire »

$$\mu_n : \omega \mapsto \mu_n(\omega) := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(\omega)}.$$

◁

Remarquons que pour chaque ω fixé, $\mu_n(\omega)$ est exactement la mesure associée aux observations $x_1, \ldots, x_n$ introduite dans la première partie de ce livret, avec $x_1 = X_1(\omega), \ldots, x_n = X_n(\omega)$. On en déduit que pour tout sous-ensemble borélien B de $\mathbb{R}$, $\mu_n(\omega, B) = \mu_n(\omega)(B)$ est la proportion de valeurs de l'échantillon appartenant à B ou *fréquence des observations* dans B.

Remarque (espérance et variance de la mesure empirique). Rappelons que si I est un ensemble fini, l'espérance de la loi discrète $\sum_{i\in I} p_i \delta_{x_i}$ est $m = \sum_{i\in I} p_i x_i$. En appliquant ceci à la loi discrète $\mu_n(\omega)$, on voit que celle-ci a pour espérance $n^{-1} \sum_{i=1}^{n} X_i(\omega) = S_n(\omega)/n$. Autrement dit : *l'espérance de la mesure empirique est la moyenne arithmétique des valeurs de l'échantillon ou moyenne empirique.* De même la variance de la loi discrète $\sum_{i\in I} p_i \delta_{x_i}$ est $s^2 = \sum_{i\in I} p_i (x_i - m)^2$. En appliquant ceci à $\mu_n(\omega)$, on voit qu'elle a pour variance $n^{-1} \sum_{i=1}^{n} \big(X_i(\omega) - S_n(\omega)/n\big)^2$. Ainsi *la variance de la mesure empirique est la variance empirique de l'échantillon.*

◁

Pour ω fixé, la mesure ponctuelle $\mu_n(\omega)$ a une fonction de répartition $F_n(\omega, .)$ (dont la représentation graphique est la courbe des fréquences cumulées croissantes). En laissant varier ω, on rend cette fonction de répartition aléatoire et on obtient la *fonction de répartition empirique.*

Définition (fonction de répartition empirique). Soit $X_1, \ldots, X_n$ un n-échantillon et μ_n la mesure empirique associée. On appelle *fonction de répartition empirique* de l'échantillon, la fonction de répartition F_n de μ_n. Plus formellement on pose pour tout $\omega \in \Omega$ et tout $x \in \mathbb{R}$:

$$F_n(\omega, x) := \mu_n(\omega,]-\infty, x]) = \frac{1}{n} \operatorname{card} \big\{ X_i(\omega);\ X_i(\omega) \leqslant x \big\}.$$

◁

Le théorème suivant nous dit que presque-sûrement, F_n converge uniformément sur $\mathbb{R}$ vers F, fonction de répartition de chacune des v.a. X_i. C'est en quelque sorte une loi forte des grands nombres fonctionnelle pour la suite des f.d.r. empiriques vues comme des fonctions aléatoires. La signification pratique est que si l'on a observé un échantillon de grande taille d'une loi inconnue de f.d.r. F, la fonction de répartition empirique peut être prise comme approximation de F.

Théorème (Glivenko-Cantelli). *Soit F_n la fonction de répartition empirique d'un échantillon $X_1, \ldots, X_n$, où les X_i ont pour f.d.r. F. Alors*

a) $\forall x \in \mathbb{R}, \quad F_n(x) \xrightarrow[n\to+\infty]{\text{p.s.}} F(x)$;

b) $\|F_n - F\|_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to+\infty]{\text{p.s.}} 0.$

À titre d'illustration, les figures 22 et 23 montrent la f.d.r. de la loi $\mathfrak{N}(3,2)$ et des f.d.r. empiriques construites sur un n-échantillon de la loi $\mathfrak{N}(3,2)$ pour $n = 100$ et $n = 400$, simulation réalisée avec Scilab. Les sauts des f.d.r. empiriques sont représentés ici par des segments verticaux en trait plein.

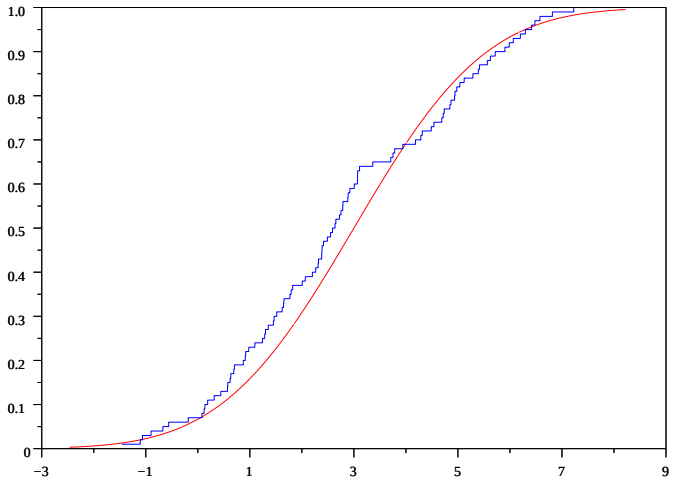

FIGURE 22 – Approximation de la f.d.r. de $\mathfrak{N}(3,2)$ par une f.d.r. empirique, $n = 100$

Modèle statistique

Définition (modèle statistique). On appelle *modèle statistique* la donnée d'un espace mesurable $(\Omega, \mathcal{F})$ et d'une famille $\big(P_\theta,\ \theta \in \Theta\big)$ de mesures de probabilité sur $(\Omega, \mathcal{F})$. Θ est appelé ensemble des paramètres. ◁

Cette définition situe d'emblée la statistique mathématique dans le prolongement de la théorie des probabilités, mais il convient de noter une différence fondamentale. En théorie des probabilités, on travaille généralement avec *un seul* espace probabilisé $(\Omega, \mathcal{F}, P)$ de référence, réputé modéliser correctement une expérience aléatoire. Ω est l'ensemble des « évènements élémentaires », $\mathcal{F}$ est une tribu de parties de Ω appelées par commodité « évènements » et P est une mesure de probabilité sur $(\Omega, \mathcal{F})$.

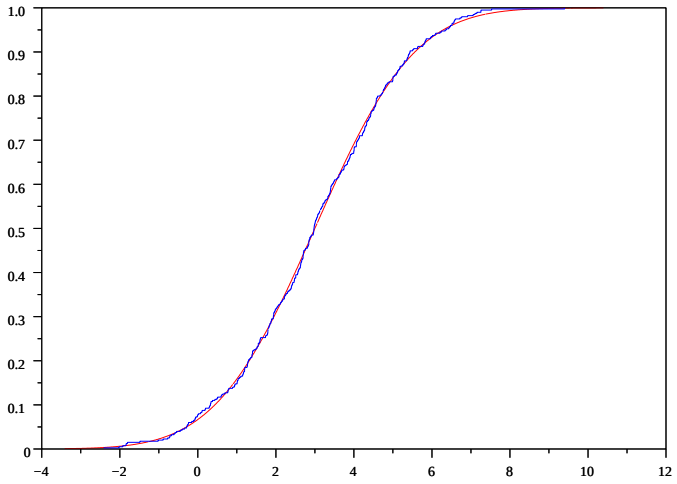

FIGURE 23 – Approximation de la f.d.r. de $\mathfrak{N}(3,2)$ par une f.d.r. empirique, $n = 400$

Les notions importantes de loi d'une v.a., d'espérance, d'indépendance sont toujours relatives à cette mesure P, même si on omet généralement de le préciser. Dans le modèle statistique, on dispose de *plusieurs espaces probabilisés* $(\Omega, \mathcal{F}, P_\theta)$, éventuellement une infinité, on ignore lequel est « le bon », et on les met en concurrence au vu des observations.

Pour compléter notre présentation du modèle statistique, il nous reste à définir les notions d'*échantillon* et de *statistique*. Nous avons déjà utilisé le mot « échantillon » dans le cadre probabiliste au sens de suite finie de variables aléatoires indépendantes et de même loi. Dans le cadre d'un modèle statistique, il convient de réécrire cette définition comme suit.

Définition (échantillon associé à un modèle statistique). Soit $\big(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ un modèle statistique. On appelle n-échantillon associé à ce modèle toute suite $X_1, \ldots, X_n$ de variables aléatoires définies sur $(\Omega, \mathcal{F})$ qui sont *pour tout* $\theta \in \Theta$, P_θ-indépendantes et de même loi Q_θ sous P_θ :

$$\forall \theta \in \Theta,\ \forall i \in [\![1, n]\!], \quad P_{\theta, X_i} := P_\theta \circ X_i^{-1} = Q_\theta.$$

◁

Définition (statistique). Soit $X_1, \ldots, X_n$ un n-échantillon associé à un modèle statistique $\big(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$. On appelle *statistique* associée à cette échantillon, toute v.a. Y_n de la forme

$$Y_n = f_n(X_1, \ldots, X_n)$$

où $f_n : \mathbb{R}^n \to \mathbb{R}$, $(t_1, \dots, t_n) \mapsto f_n(t_1, \dots, t_n)$ est une application borélienne *ne dépendant pas* de θ. ◁

Au risque d'insister lourdement, notons que le point important dans cette définition est la possibilité de calculer Y_n à partir des X_i *sans avoir besoin de connaître la valeur de* θ. Bien sûr la loi de Y_n (sous P_θ) dépend en général de θ, mais la fonction f_n qui elle, n'a rien à voir avec le modèle ni avec les X_i, ne doit pas dépendre de θ.

Exemples. Voici quatre exemples simples de statistiques.

1. La moyenne empirique $\overline{X} := \dfrac{X_1 + \cdots + X_n}{n} = \dfrac{S_n}{n}$.
2. La variance empirique $\dfrac{1}{n}\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2$.
3. $\min\limits_{1 \leqslant i \leqslant n} X_i$.
4. $\max\limits_{1 \leqslant i \leqslant n} X_i$.

◁

Par contre, la v.a. $n^{-1}\sum_{i=1}^{n}(X_i - \mathrm{E}\, X_i)^2$ n'est pas une statistique en général car $\mathrm{E}\, X_i$ doit en fait s'écrire $\mathrm{E}_\theta\, X_i$ et la fonction f_n correspondante donnée par $f_n(t_1, \dots, t_n) = n^{-1}\sum_{i=1}^{n}(t_i - \mathrm{E}_\theta\, X_i)^2$ n'est pas calculable sans la connaissance de $\mathrm{E}_\theta\, X_1$, donc du paramètre[46] θ (les X_i ayant même loi sous P_θ, $\mathrm{E}_\theta\, X_i = \mathrm{E}_\theta\, X_1$ pour tout i).

Estimation

Nous revenons maintenant à la question posée en introduction du chapitre : à partir de l'observation d'un échantillon $X_1, \dots, X_n$ associé à un modèle statistique $\big(\Omega, \mathcal{F}, (P_\theta)_{\theta \in \Theta}\big)$, comment « deviner quel est le bon θ » ? Plus précisément on s'intéresse généralement à un objet $\gamma = g(\theta)$ qui peut être de nature fonctionnelle (densité de la loi des X_i, f.d.r. des X_i, fonction de survie, ...) ou numérique (espérance, variance, médiane, ...) ou vectorielle fini-dimensionnelle (couple (μ, σ) dans un modèle gaussien, bornes d'un intervalle de support d'une loi, triplet des paramètres d'une loi de Weibull, ...). Dans le premier cas on parle *d'estimation fonctionnelle*. Nous en avons vu un exemple important avec la f.d.r. empirique qui est un estimateur fonctionnel de la f.d.r. F des X_i. Par le théorème de Glivenko Cantelli, cet estimateur fonctionnel converge vers F p.s. uniformément[47]. Nous nous occupons dans ce qui suit du cas où γ est fini-dimensionnel et pour simplifier, nous nous limitons dans les énoncés au cas où g est l'identité et $\gamma = \theta$ est un réel. En cas de besoin, par exemple si dans un modèle gaussien paramétré par $\theta = (\mu, \sigma)$, on s'intéresse à la variance $g(\theta) = \sigma^2$, il sera facile d'adapter les énoncés.

46. Sauf dans le cas particulier où $\mathrm{E}_\theta\, X_i$ est constante par rapport à θ.

47. Dans le cadre du modèle statistique $\big(\Omega, \mathcal{F}, (P_\theta)_{\theta \in \Theta}\big)$, F dépend bien sûr de θ et la conclusion du théorème de Glivenko Cantelli devient : « pour tout $\theta \in \Theta$, $\|F_n - F_\theta\|_\infty \to 0$ P_θ-p.s. ».

Avant de donner les définitions relatives à l'estimation, il est utile de passer en revue quelques exemples introductifs. Pour chacun de ces exemples, on suppose que l'on a un modèle statistique $\big(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ et on note $X_1, \dots, X_n$ un n-échantillon associé à ce modèle.

Exemple (estimation de l'espérance). Dans le cas où le paramètre θ inconnu est l'espérance de l'échantillon ($\mathrm{E}_\theta X_i = \theta$), un estimateur usuel de θ est la moyenne empirique

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Une des raisons de ce choix est que $\overline{X}$ converge presque sûrement quand n tend vers l'infini vers $\mathrm{E}_\theta X_1 = \theta$ par la loi forte des grands nombres. On dit que $\overline{X}$ est un estimateur *fortement consistant* de θ. Notons aussi que pour tout n, $\mathrm{E}_\theta \overline{X} = \theta$. On dit que $\overline{X}$ est un estimateur *sans biais* de θ. ◁

Exemple (estimation de la variance). Lorsque le paramètre inconnu est la variance $\theta = \sigma^2$, on peut l'estimer par la variance empirique :

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2.$$

En appliquant la loi forte des grands nombres aux suites aléatoires $(X_i)_{i\geqslant 1}$ et $(X_i^2)_{i\geqslant 1}$, on en déduit que S^2 converge p.s. vers σ^2. Là encore cet estimateur est fortement consistant. Par contre, on vérifie facilement en exercice que

$$\forall \theta \in \Theta, \quad \mathrm{E}_\theta(S^2) = \frac{n-1}{n}\operatorname{Var}_\theta X_1 \neq \operatorname{Var}_\theta X_1.$$

Ici l'espérance de l'estimateur n'est pas égale au paramètre à estimer, on dit que cet estimateur est *biaisé*. ◁

Remarque (à propos du σ_{n-1}). Il est facile de modifier S^2 pour en faire un estimateur sans biais de σ^2. En effet par linéarité de l'espérance, on a pour tout $\theta \in \Theta$,

$$\mathrm{E}_\theta\left(\frac{n}{n-1}S^2\right) = \frac{n}{n-1} \times \frac{n-1}{n}\operatorname{Var}_\theta X_1 = \operatorname{Var}_\theta X_1.$$

On en déduit que

$$\frac{n}{n-1}S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

est un estimateur sans biais de $\operatorname{Var}_\theta X_1$. Il reste évidemment fortement consistant puisque $n/(n-1)$ tend vers 1 et S^2 tend P_θ-p.s. pour tout θ vers $\operatorname{Var}_\theta X_1$. Lorsque la taille n de l'échantillon est fixée et « petite », on préfère cet estimateur à S^2. Pour estimer l'écart-type, on prend alors la racine carrée de cet estimateur sans biais, que l'on note traditionnellement (mais de façon incohérente) σ_{n-1}. Attention, σ_{n-1} n'est pas un estimateur sans biais de σ parce que l'espérance ne commute pas avec la racine carrée. ◁

Exemple (estimation du support d'une loi uniforme). Soit $X_1, \ldots, X_n$ un n-échantillon de la loi uniforme sur $[0, \theta]$, où $\theta \in]0, +\infty[$ est inconnu. Voici un premier estimateur T_n de θ défini par

$$T_n := \max_{1 \leqslant i \leqslant n} X_i.$$

Il est facile de vérifier que T_n est un estimateur fortement consistant et biaisé ($\mathrm{E}_\theta\, T_n < \theta$). Voici un deuxième estimateur T'_n fortement consistant et sans biais :

$$T'_n = 2\overline{X} = \frac{2}{n} \sum_{i=1}^{n} X_i.$$

En effet si X_i suit la loi uniforme sur $[0, \theta]$ son espérance vaut $\theta/2$. ◁

Dans les exemples vus ci-dessus, l'estimateur proposé est à chaque fois, une fonction des observations « proche », au moins pour les grandes valeurs de n, du paramètre qu'il est censé estimer. Cette idée de proximité est néanmoins trop imprécise pour être incorporée à la définition mathématique d'un estimateur, laquelle doit être valable pour toute valeur de n. On y renonce donc et il ne reste plus que la notion de fonction des observations, autrement dit de *statistique* au sens de la définition donnée page 102.

Définition (estimateur). Soit $X_1, \ldots, X_n$ un n-échantillon associé à un modèle statistique $\big(\Omega, \mathcal{F}, (P_\theta)_{\theta \in \Theta}\big)$, où Θ est une partie de $\mathbb{R}$. On appelle *estimateur* de θ associé à cet échantillon, toute v.a. T_n de la forme

$$T_n = f_n(X_1, \ldots, X_n)$$

où $f_n : \mathbb{R}^n \to \mathbb{R}$, $(t_1, \ldots, t_n) \mapsto f_n(t_1, \ldots, t_n)$ est une application borélienne ne dépendant pas de θ. ◁

D'un point de vue formel, les définitions des statistiques et des estimateurs sont équivalentes. La seule différence, non mathématique, est le *contexte* d'utilisation. L'expression *statistique* est plus générale car elle recouvre aussi bien les estimateurs que les statistiques de test.

Il faut bien avouer qu'à son niveau de généralité, la définition d'un estimateur a quelque chose de choquant car il semble finalement que T_n puisse n'avoir aucun rapport avec θ. À y regarder de plus près, on voit que le seul rapport de T_n avec θ, c'est que sa loi dépend de θ *via* la loi du vecteur aléatoire $(X_1, \ldots, X_n)$. C'est néanmoins bien peu et on se dépêche de compléter la définition 39 en définissant des propriétés qui permettent de dire que certains estimateurs sont moins mauvais que d'autres.

Définition (estimateur faiblement consistant). Soit T_n un estimateur de θ. On dit qu'il est *faiblement consistant* s'il converge en probabilité vers θ quand n tend vers l'infini. ◁

Bien sûr, cette définition contient un grossier abus de langage. Il faudrait dire « la *suite* d'estimateurs $(T_n)_{n\geqslant 1}$ est faiblement consistante si… ». D'autre part, rappelons que nous travaillons avec un modèle statistique $\big(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$. Dans ce cadre, la convergence en probabilité de T_n vers θ signifie très précisément :

$$\forall \theta \in \Theta, \quad \forall \varepsilon > 0, \quad P_\theta\big(|T_n - \theta| \geqslant \varepsilon\big) \xrightarrow[n\to+\infty]{} 0.$$

Avec le même abus de langage que ci-dessus, on définit la consistance forte.

Définition (estimateur fortement consistant). Soit T_n un estimateur de θ. On dit qu'il est *fortement consistant* s'il converge presque-sûrement vers θ quand n tend vers l'infini. ◁

Ici la convergence presque-sûre de T_n vers θ signifie :

$$\forall \theta \in \Theta, \quad P_\theta\Big(\lim_{n\to+\infty} T_n = \theta\Big) = 1.$$

Fixons pour un moment n et notons $T = T_n$. L'*erreur d'estimation* est la v.a. $T - \theta$. On suppose ici que $\mathrm{E}_\theta\,|T| < +\infty$ pour tout $\theta \in \Theta$, ce qui entraîne l'existence de $\mathrm{E}_\theta\, T$. On peut alors décomposer l'erreur d'estimation comme suit :

$$T - \theta = (T - \mathrm{E}_\theta\, T) + (\mathrm{E}_\theta\, T - \theta).$$

Le premier terme d'erreur $T - \mathrm{E}_\theta\, T$ est aléatoire et provient inévitablement des fluctuations de la v.a. T « autour » de son espérance. Le deuxième terme $(\mathrm{E}_\theta\, T - \theta)$ est déterministe et représente une *erreur systématique*, dont on pourrait se débarrasser en ajoutant une constante convenable[48] à T.

Définition (biais). Soit T un estimateur de θ. Si $\mathrm{E}_\theta\, T$ existe pour tout $\theta \in \Theta$, on appelle *biais* de l'estimateur T la quantité $(\mathrm{E}_\theta\, T - \theta)$. De plus

- si pour tout $\theta \in \Theta$, $\mathrm{E}_\theta\, T = \theta$, on dit que T est un estimateur *sans biais* de θ ;
- si $\mathrm{E}_\theta\, T \neq \theta$ pour au moins un $\theta \in \Theta$, on dit que T est un estimateur *biaisé* de θ ;
- si $(T_n)_{n\geqslant 1}$ est une suite d'estimateurs telle que pour tout $\theta \in \Theta$, $\mathrm{E}_\theta\, T_n$ converge vers θ quand n tend vers l'infini, on dit que T_n est *asymptotiquement sans biais*.

◁

Cette définition du biais contient encore un abus de langage, puisqu'on considère le biais comme un nombre réel, alors qu'il s'agit en fait de la fonction :

$$b : \Theta \to \mathbb{R}, \quad \theta \mapsto b(\theta) = \mathrm{E}_\theta\, T - \theta.$$

Dans l'exemple ci-dessus d'estimation de la variance, la variance empirique S^2 est un estimateur biaisé mais asymptotiquement sans biais de σ^2. Il en va de même pour l'estimateur T_n de la borne θ du support d'une loi uniforme.

48. À condition d'être capable de calculer $\mathrm{E}_\theta\, T - \theta$, ce qui n'est généralement pas le cas.

Estimation par intervalle de confiance

Au lieu de proposer directement une valeur calculée à partir de l'échantillon pour représenter (plus ou moins bien) la valeur inconnue de θ ou plus généralement de $g(\theta)$, une autre technique d'estimation consiste à l'encadrer avec une grande probabilité de succès entre des bornes calculées à partir de l'échantillon. On parle alors *d'estimation par intervalle de confiance*. Nous avons déjà vu un exemple de cette technique avec les intervalles de confiance pour une probabilité inconnue.

Définition (intervalles de confiance). Soit $\big(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta}\big)$ un modèle statistique, $(X_1, \dots, X_n)$ un n-échantillon associé à ce modèle, g une fonction $\Theta \to \mathbb{R}$, a_n et b_n des fonctions mesurables $\mathbb{R}^n \to \mathbb{R}$ telles que $a_n \leqslant b_n$ et $\varepsilon \in]0,1[$. On dit que l'intervalle aléatoire

$$I_n := \big[a_n(X_1, \dots, X_n), b_n(X_1, \dots, X_n)\big]$$

est un intervalle de confiance pour $g(\theta)$

— au niveau de confiance au moins $1-\varepsilon$ si

$$\forall \theta \in \Theta, \quad P_\theta\big(g(\theta) \in I_n\big) \geqslant 1-\varepsilon;$$

— au niveau de confiance $1-\varepsilon$ si

$$\inf_{\theta\in\Theta} P_\theta\big(g(\theta) \in I_n\big) = 1-\varepsilon.$$

◁

Au vu de cette définition, il convient de rectifier ce qui est dit dans l'introduction informelle ci-dessus : « encadrer avec une grande *probabilité* de succès entre des bornes calculées à partir de l'échantillon ». Si l'on parle de *niveau de confiance* pour I_n et pas de *probabilité* de recouvrement de $g(\theta)$, c'est parce qu'il n'y a pas une probabilité dans ce problème, mais toute une famille $(P_\theta)_{\theta\in\Theta}$. De plus, l'application $Q = \inf_{\theta\in\Theta} P_\theta : \mathcal{F} \to \mathbb{R}^+$, $A \mapsto Q(A) = \inf\{P_\theta(A),\, \theta \in \Theta\}$, n'est en général pas une probabilité sur $\mathcal{F}$.

En pratique, une fois que l'on dispose d'un intervalle aléatoire I_n vérifiant la définition ci-dessus, on propose comme intervalle de confiance pour $g(\theta)$, le $I_n(\omega)$ calculé à partir des observations $x_1 = X_1(\omega), \dots, x_n = X_n(\omega)$, c'est-à-dire

$$I_n(\omega) = [a_n(x_1, \dots, x_n), b_n(x_1, \dots, x_n)].$$

Cet intervalle n'a plus rien d'aléatoire et il n'est donc pas correct d'écrire des choses comme « $P_\theta(g(\theta) \in [a_n(x_1, \dots, x_n), b_n(x_1, \dots, x_n)]) \geqslant 1-\varepsilon$ ». La raison implicite du choix de $I_n(\omega)$ comme intervalle encadrant $g(\theta)$ est que la probabilité que l'intervalle aléatoire I_n recouvre $g(\theta)$ étant dans tous les cas (c'est-à-dire quel que soit θ) élevée (on prend généralement ε proche de 0, typiquement, $\varepsilon = 0{,}05$), on peut *croire* avec un risque d'erreur bien contrôlé que l'intervalle effectivement observé $I_n(\omega)$ recouvre bien $g(\theta)$.

Pour bien distinguer l'intervalle de confiance I_n défini ci-dessus et l'intervalle $I_n(\omega)$ calculé explicitement à partir des observations numériques, il serait souhaitable de parler d'intervalle de confiance théorique pour le premier et d'intervalle de confiance numérique pour le deuxième[49].

Formalisation de test

La statistique apporte une aide à la prise de décision via les tests. Le cas le plus simple d'un choix de décision binaire peut se formaliser comme suit. Dans le modèle statistique $(\Omega, \mathcal{F}, (P_\theta)_{\theta\in\Theta})$, on partage l'espace des paramètres en Θ_0 et Θ_1 son complémentaire[50]. Au vu d'un échantillon $X = (X_1, \ldots, X_n)$, on voudrait décider de croire que la vraie valeur de θ appartient à Θ_0, c'est ce que l'on appelle l'hypothèse nulle, ou de rejeter cette hypothèse. Pour cela on choisit une statistique $T_n = f_n(X_1, \ldots, X_n)$ et une fonction $\varphi : \mathbb{R} \to \{0, 1\}$. Si $\varphi(T_n) = 0$, on accepte l'hypothèse nulle, sinon on la rejette.

Avec cette procédure, on peut commettre deux types d'erreur :

1. rejeter à tort l'hypothèse nulle, c'est l'erreur de première espèce ;
2. accepter à tort l'hypothèse nulle, c'est l'erreur de deuxième espèce.

Par exemple lors de la mise au point d'une alarme de détection incendie, le premier souci du fabricant est d'éviter autant que possible les fausses alarmes car une alarme qui se déclenche intempestivement alors qu'il n'y a pas de départ de feu perd vite toute crédibilité et ne sera pas prise au sérieux lors d'un véritable départ de feu. Ici l'hypothèse nulle est « il n'y a pas d'incendie », l'hypothèse alternative est « il y a un départ de feu », l'erreur de première espèce est la fausse alarme et l'erreur de deuxième espèce est la non détection d'un départ de feu.

La procédure suivie par le fabriquant sera donc de s'assurer d'abord que son dispositif ne déclenche que rarement de fausses alarmes, donc de contrôler les probabilités $P_\theta(\varphi(T_n) = 1)$ pour tout $\theta \in \Theta_0$, par exemple en veillant à ce qu'elles soient inférieures à 5%. Ensuite, la qualité de l'alarme sera d'autant meilleure que $P_\theta(\varphi(T_n) = 1)$ sera élevée pour $\theta \in \Theta_1$. La fonction $\Theta_1 \to [0, 1]$, $\theta \mapsto P_\theta(\varphi(T_n) = 1)$ est appelée *puissance* du test.

49. Ces appellations ne sont pas standard.
50. Dans la théorie des tests on prend plus généralement Θ_1 *inclus* dans le complémentaire de Θ_0.

Bibliographie

[1] W. Feller. *An introduction to probability theory and its application*. 2e éd. T. 1. Wiley, 1957.

[2] W. Feller. *An introduction to probability theory and its application*. 2e éd. T. 2. Wiley, 1966.

[3] H. Fisher. *A history of the central limit theorem*. Springer, 2010.

[4] D. Foata, A. Fuchs et J. Franchi. *Calcul des probabilités*. 3e éd. Dunod, 2012. isbn : 978-2-10-057424-7.

[5] Le Cam L. « The central limit theorem around 1935 ». In : *Statistical Science* 1.1 (1986), p. 78–96.

[6] G. Polyà. « Über den Zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentproblem ». In : *Mathematische Zeitschrift* 8 (1920), p. 171–180.

[7] Ch. Suquet. *Intégration, Analyse de Fourier, Probabilités*. UFR de Mathématiques, Université Lille 1, fév. 2012. url : http://math.univ-lille1.fr/~suquet/Polys/IFP.pdf.

[8] Ch. Suquet. *Introduction au calcul des probabilités*. UFR de Mathématiques, Université Lille 1, fév. 2012. url : http://math.univ-lille1.fr/~suquet/Polys/ICP.pdf.

[9] Ch. Suquet. *Probabilités via l'intégrale de Riemann*. Ellipses, 2013.

[10] J.V. Uspensky. *Introduction to mathematical probability*. McGraw-Hill, 1937.