Introduction aux probabilités et statistiques pour la science des données

C. Levrard, Notes de cours 2023

1 Le minimum de probas requis **2**
1.1 Rappels . . . 2
1.1.1 Espace probabilisé et variable aléatoire . . . 2
1.1.2 Variables discrètes et continues . . . 5
1.1.3 Couple de variable aléatoires réelles . . . 12
1.2 Convergences de suites de variables aléatoires réelles . . . 18
1.2.1 Point de vue suite de fonctions : convergence en probabilité et presque sûre . . . 18
1.2.2 Point de vue suite de lois : convergence en loi . . . 23
1.3 Outils pour la statistique et l'analyse de données . . . 29
1.3.1 Échantillon de variables i.i.d. . . . 29
1.3.2 Fluctuations autour de la moyenne : cadre non asymptotique . 31
1.3.3 Fluctuations autour de la moyenne : point de vue asymptotique 35

2 Le point de vue des statistiques **42**
2.1 Des statistiques descriptives à l'inférence statistique . . . 43
2.2 Estimation (ponctuelle) . . . 46
2.3 Méthodes classiques d'estimation . . . 50
2.3.1 Méthode des moments . . . 50
2.3.2 Méthode du maximum de vraisemblance . . . 54
2.4 Intervalles de confiance . . . 57
2.4.1 Non asymptotique . . . 58
2.4.2 Asymptotique . . . 59
2.5 Tests . . . 64
2.5.1 Recettes de construction de test . . . 66
2.5.2 p-valeur . . . 67

3 Deux modèles classiques (et méthodes qui vont avec) **70**
3.1 Modèle lin gaussien . . . 70
3.2 Modèles multinomiaux . . . 70

Chapitre 1

Le minimum de probas requis

1.1 Rappels

1.1.1 Espace probabilisé et variable aléatoire

On commence par rappeler comment les probas permettent de modéliser une expérience aléatoire. Le premier concept est celui *d'espace probablisé*.

Definition 1.1 : Espace probabilisé

Un espace probabilisé est un triplet $(\Omega, \mathcal{A}, \mathbb{P})$ tel que

1. Ω est un ensemble, appelé *univers*.
2. $\mathcal{A}$ est une *tribu* sur Ω (famille de sous-ensembles de Ω contenant Ω, stable par complément et union dénombrable). Un élément de $\mathcal{A}$ est appelé évènement.
3. $\mathbb{P}$ est une *mesure de probabilité* sur $(\Omega, \mathcal{A})$, c'est à dire une fonction σ-additive de $\mathcal{A}$ dans $[0, 1]$ telle que $\mathbb{P}(\Omega) = 1$.

Intuitivement parlant :

1. Ω peut correspondre à l'ensemble des réalisations possibles de votre expérience, ou être purement abstrait.
2. $\mathcal{A}$ est l'ensemble des évènement que vous pouvez observer, mesurer.
3. $\mathbb{P}$ est une répartition de masse sur l'ensemble des choses que vous pouvez observer/mesurer (plus la masse de $A \subset \Omega$ est grande, plus l'évènement associé est probable).

Exemple 1.2 *: Main de départ au poker.* Vous voulez essayer de quantifier le fait qu'une main de départ au poker soit favorable ou non. Vous pouvez modéliser cela de la manière suivante :

1. $\Omega = \{\text{ensemble des mains de départ possible}\} \simeq \mathcal{C}_2^{54}$ (ensemble des combinaisons de 2 éléments choisis parmi 54).
2. $\mathcal{A} = \mathcal{P}(\Omega)$ (ensemble des sous-ensembles de Ω).
3. Si $\omega \in \Omega$ représente une main possible de départ, $\mathbb{P}(\{\omega\}) = 1/|\Omega| = \binom{54}{2}^{-1}$. On suppose alors que toutes les mains de départ sont équiprobables (a priori il n'y a pas de raison de supposer une autre répartition).

De manière générale, lorsque on vous décrit une expérience 'physiquement', c'est à vous de poser le modèle (dans un cadre discret $\mathcal{A}$ sera souvent l'ensemble des parties de Ω et la mesure de probabilité dessus sera le plus souvent uniforme si on tire au hasard, de sorte que seule la détermination de Ω importera vraiment).

Exemple 1.3 *: Exemple dans un cadre non discret.* Vous êtes un artiste contemporain et lancez au hasard une goutte de peinture sur une toile carrée. On peut modéliser votre oeuvre de la manière suivante :

1. $\Omega = [0, L]^2$ (où L est la dimension de votre toile).
2. $\mathcal{A} = \mathcal{B}([0, L]^2) = \sigma\left(\{[a_1, b_1] \times [a_2, b_2]\}_{0 \leq a_1 \leq b_1 \leq L, 0 \leq a_2 \leq b_2 \leq L}\right)$, la tribu engendrée par l'ensemble des rectangles (aussi appelée tribu borélienne). On peut aussi la voir comme l'ensemble des réunions dénombrables d'intersections et d'unions finies de rectangles et complémentaires de rectangles).
3. $\mathbb{P}$ est la mesure uniforme sur $[0, L]^2$, définie par $\mathbb{P}([a_1, b_1] \times [a_2, b_2]) = (b_2 - a_2) \times (b_1 - a_1)/L^2$.

Deux remarques s'imposent au vu de cet exemple :

1. Pourquoi ne pas prendre comme ensemble d'évènements mesurable $\mathcal{P}([0, L]^2)$, c'est à dire l'ensemble des sous-parties de $[0, L]^2$ (qui est bien une tribu) ? La réponse vient de la théorie de la mesure en général : on ne peut pas construire sur cette tribu une mesure qui soit invariante par translation (pas de mesure uniforme dessus donc, cela découle du paradoxe de Banach-Tarski pour les lecteurs intéressés). On doit donc partir du principe qu'on ne peut pas tout mesurer, et qu'il faut donc spécifier les évènements de base qu'on veut pouvoir mesurer (ici les rectangles), et prendre la tribu engendrée par ces évènements de base. La plupart du temps, pour un $\Omega \subset \mathbb{R}^d$, on prendra comme tribu la tribu Borélienne (celle qui permet de mesurer tous les rectangles, ouverts, fermés, etc.).
2. On s'est contenté de définir $\mathbb{P}$ sur les rectangles, et cela suffit. En effet si deux mesures coïncident sur les évènements de base (et que ces derniers sont stables par intersections finies), alors elles sont égales. Une mesure de proba est donc totalement définie par sa donnée sur ces évènements de base.

Une justification propre de ces deux assertions sort du cadre d'un cours de L3 introductif, vous êtes donc priés d'attendre le M1.

Passage du continu au discret : On peut passer de cet exemple continu à un espace probabilisé discret : si on découpe la toile en K morceaux $(A_j)_{j=1,\dots,K}$, et qu'on suppose n'observer seulement que l'appartenance de la goutte de peinture à un de ces morceaux, un nouvel espace probabilisé est alors

— $\Omega' = [\![1, K]\!]$ (numéros des morceaux).

— $\mathcal{A}' = \mathcal{P}([\![1, K]\!])$ (là on a le droit de mesurer toutes les sous-parties).

— $\mathbb{P}' : \{j\} \mapsto \mathbb{P}(A_j)$.

Ce deuxième espace probabilisé (où on suppose que l'on observe moins de chose) s'obtient moralement comme 'fonction' du premier. De fait, pour passer de l'un à l'autre, on a utilisé la fonction

$$X : \begin{cases} [0, L]^2 & \to & [\![1, K]\!] \\ x & \mapsto & \sum_{j=1}^K j \mathbb{1}_{x \in A_j} \quad \text{(numéro de la case de } x\text{)}, \end{cases}$$

et transféré la masse de la case A_j sous $\mathbb{P}$ au numéro j correspondant. C'est exactement la définition d'une variable aléatoire et de sa loi.

Definition 1.4 : Variable aléatoire - Loi d'une variable aléatoire

Soit $(\mathcal{X}, \mathcal{A}_\mathcal{X})$ un espace muni d'une tribu. Une variable aléatoire (à valeurs dans $(\mathcal{X}, \mathcal{A}_\mathcal{X})$) est une fonction *mesurable* d'un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ à valeurs dans $(\mathcal{X}, \mathcal{A}_\mathcal{X})$. On rappelle qu'une telle fonction f est mesurable si, pour tout $B \in \mathcal{A}_\mathcal{X}$, $f^{-1}(B) \in \mathcal{A}$.

Si X est une variable aléatoire (à valeurs dans $(\mathcal{X}, \mathcal{A}_\mathcal{X})$), sa **loi** P_X est la mesure de probabilité sur $(\mathcal{X}, \mathcal{A}_\mathcal{X})$ définie par

$$\forall B \in \mathcal{A}_\mathcal{X} \quad P_X(B) = \mathbb{P}(X^{-1}(B)).$$

En d'autres termes, une variable aléatoire est juste une fonction partant d'un espace probabilisé, et sa loi est la manière dont elle transfère la mesure/masse de départ sur l'espace d'arrivée. FAIRE DESSIN

Remarque importante : On a deux points de vue possible sur une variables aléatoires, celui le considérant comme une fonction, et celui de sa loi (la ramenant juste à une répartition de masse). Les deux points de vue ne sont pas en bijection : si deux variables (sur un même espace) ont des lois différentes, alors forcément ces variables sont différentes, mais la réciproque est fausse.

Exemple 1.5 *: Contre-exemple.* Dans le modèle du peintre, si on note X la variable qui envoie x (position de la goutte de peinture) dans $[\![1, K]\!]$ (numéro de la case dans laquelle tombe la goutte), et si toutes les cases A_j ont même masse via $\mathbb{P}$, alors, pour tout $\sigma \in \mathcal{S}_K$ (permutation des K numéros), $\sigma \circ X$ a même loi que X (on ne change pas la loi en permutant les issues d'une variable aléatoire uniforme discrète).

Dans la plupart des applications, l'importance de l'espace probabilisé de départ $(\Omega, \mathcal{A}, \mathbb{P})$ est à relativiser. En effet, si on peut se ramener à l'observation de variables aléatoires dont on connaît la loi on peut directement travailler sur $(\mathcal{X}, \mathcal{A}_\mathcal{X}, P_X)$.

Exemple 1.6 *: Jeux de dés.* On lance 2 dés au hasard. Dans une première situation on s'intéresse au résultat X du premier dé (par exemple dépasse-t-il 3 ?). Vous avez alors deux options :

— Poser $\Omega = [\![1, 6]\!]^2$ et $\mathbb{P}(\{(i, j)\}) = 1/36$ (mesure uniforme sur les issues possibles pour les deux dés). Exprimer ensuite X comme fonction de Ω dans $[\![1, 6]\!]$ (celle qui a (i, j) associe i). Calculer ensuite

$$\mathbb{P}(X = 3) = \frac{|X^{-1}(\{3, 4, 5, 6\})|}{36} = \frac{|\{(i, j) \in [\![1, 6]\!]^2 \mid i \geq 3\}|}{36} = \frac{4 \times 6}{36} = \frac{2}{3}.$$

— Modéliser directement la loi de X. Le premier dé étant non truqué, on a que X suit une loi uniforme sur $[\![1, 6]\!]$ ($X \sim \mathcal{U}([\![1, 6]\!])$), et donc

$$\mathbb{P}(X = 3) = P_X([\![3, 6]\!]) = \frac{2}{3}.$$

Avec la deuxième option on n'a défini ni $\mathbb{P}$ ni Ω mais ce n'est pas grave, vu que $\mathbb{P}(X = 3)$ est par définition uniquement déterminé par P_X. Une autre manière de voir est de considérer $\Omega = [\![1, 6]\!]$ et $\mathbb{P} = P_X$ la loi uniforme dessus, mais

ce n'est pas un point de vue conseillé : si après on s'intéresse au résultat du second dé il va falloir changer Ω et on préfère éviter. Il vaut mieux donc garder un $\Omega, \mathbb{P}$ imaginaire dont on se fiche.

Si on s'intéresse maintenant a Z la somme des deux dés, là on ne peut pas court-circuiter la définition du $\Omega, \mathbb{P}$: en effet cela ne correspond pas à une loi classique (correspondant à la modélisation d'expériences classiques). On est donc condamnés à poser $\Omega = [\![1, 6]\!]^2$ et $\mathbb{P}(\{(i, j)\}) = 1/36$, et, pour $k \in [\![2, 12]\!]$, calculer éventuellement

$$P_Z(\{k\}) = \mathbb{P}(Z = k) = \frac{|\{(i, j) \mid i + j = k\}|}{36}.$$

On peut toutefois s'en sortir quand même (c'est à dire ne pas expliciter $(\Omega, \mathbb{P})$) en introduisant X_1 et X_2 les variables aléatoires correspondant aux deux tirages (que l'on suppose indépendants), et calculer

$$\begin{aligned}\mathbb{P}(Z = k) = \mathbb{P}(X_1 + X_2 = k) &= \sum_{j=1}^{k-1} \mathbb{P}(\{X_1 = j\} \cap \{X_2 = k - j\}) \\ &= \sum_{j=1}^{k-1} \mathbb{P}(X_1 = j)\mathbb{P}(X_1 = k - j) = \sum_{j=1}^{k-j} P_{X_1}(\{j\})P_{X_2}(\{k - j\}),\end{aligned}$$

où $P_{X_1} = P_{X_2} = \mathcal{U}([\![1, 6]\!])$.

La morale de l'exemple précédent est la suivante : si on peut s'épargner la définition d'un $(\Omega, \mathcal{A}, \mathbb{P})$ général en ne manipulant que des variables aléatoires dont on connaît la loi, **on le fait** (c'est souvent plus rapide, et ce sera toujours le cas dans la partie stats). D'où l'intérêt de connaître les lois classiques (correspondant à des expériences classiques elles aussi). Dans ce cas de figure on écrira encore $\mathbb{P}(X \in B)$, mais sans définir le $\mathbb{P}$ (ce sera un $\mathbb{P}$ abstrait, non nécessaire au bon déroulement des opérations).

Bien sûr il y aura des situations où on sera obligé de définir proprement $(\Omega, \mathcal{A}, \mathbb{P})$ (notamment dans les problèmes de jeux de cartes, urnes, ou autres expériences de probabilistes combinatoires).

1.1.2 Variables discrètes et continues

On se place dans le cadre général où X est une variable aléatoire à valeurs dans $\mathbb{R}$. Le distinguo entre variable discrète et variable continue se fait en regardant l'ensemble des valeurs possibles pour X.

— Si $X(\Omega)$ est dénombrable (on l'assimilera alors à $\mathbb{Z}$ pour simplifier), la variable est dite discrète.

— Si $X(\Omega)$ n'est pas dénombrable, X n'est pas nécessairement continue : il faut que, pour tout $a < b$, $\mathbb{P}(X \in [a, b])$ s'exprime comme

$$\mathbb{P}(X \in [a, b]) = \int_a^b f(u)du,$$

où f est une fonction (mesurable) positive d'intégrale 1 (appelée alors densité de X).

Bien qu'il existe des variables non discrètes et non continues, pour ce cours on se cantonnera à ces deux cas (discret/continu).

Caractérisation des lois

Vous pouvez caractériser les lois de variables au moyen de 3 outils. On peut commencer par repartir de la définition d'une loi, qui est juste une manière de mettre de la masse sur des sous-ensembles de $\mathbb{R}$. Suivant le caractère continu ou discret on ne regardera pas les mêmes ensembles.

Proposition 1.7 : Caractérisation "brute" des lois

Si $X : \Omega \to \mathbb{Z}$ est une variable discrète, alors sa loi est entièrement déterminée par

$$k \mapsto \mathbb{P}(X = k),$$

pour $k \in \mathbb{Z}$.

Si $X : \Omega \to \mathbb{R}$ est une variable continue, alors sa loi est entièrement déterminée par sa densité f_X.

Cette dichotomie de cas découle de la même caractérisation de départ : une loi est totalement déterminée par la donnée des $P_X(B)$, où B parcourt l'ensemble de tous les évènements de $\mathcal{B}$ (pas facile à manipuler), ou tous les évènements **de base** (ici les intervalles de type $[a, b]$). La caractérisation suivante (par fonction de répartition) donne juste une autre type d'évènements de base suffisants à la caractérisation des lois.

Definition 1.8 : Fonction de répartition

Pour X une variable aléatoire (réelle), sa fonction de répartition F_X est définie par

$$F_X = \begin{cases} \mathbb{R} & \to & [0,1] \\ t & \mapsto & \mathbb{P}(X \leq t)(= P_X(]-\infty, t])). \end{cases}$$

Elle caractérise entièrement la loi P_X, au sens que si $F_X = F_Y$ (à un ensemble dénombrable près), alors $P_X = P_Y$.

Si X est discrète, la fonction de répartition s'écrit

$$F_X(t) = \sum_{k \leq \lceil t \rceil} \mathbb{P}(X = k),$$

et si X est continue, elle s'écrit

$$F_X(t) = \int_{-\infty}^{t} f_X(t)dt.$$

Comme auparavant, on voit que la disjonction de cas ne concerne que la manière de calculer la fonction de répartition. D'un point de vue conceptuel, il n'y a pas de différences, et d'un point de vue pratique, on pourra retrouver la densité en dérivant la fonction de répartition dans le cas continu (elle sera toujours dérivable par morceaux dans nos cas d'applications, dans un cadre plus général elle sera toujours dérivable presque sûrement).

Une dernière caractérisation des lois de variables aléatoires réelles est celle donnée par la fonction caractéristique.

Definition 1.9 : Fonction caractéristique

Soit X une variable aléatoire réelle. La fonction caractéristique de X, notée ϕ_X, est définie par

$$\phi_X : \begin{cases} \mathbb{R} & \to & \mathbb{C}, \\ t & \mapsto & \mathbb{E}(e^{itX}). \end{cases}$$

Elle caractérise la loi de X.

Là encore, suivant que vous avez une variable discrète ou continue, le calcul se fera d'une manière ou d'une autre. Cette caractérisation peut sembler d'une nature différente des deux autres, il n'en n'est rien si on se place du point de vue des *fonctions tests*. Heuristique parlant, on a vu que $P_X = P_Y$ ssi ces deux lois coïncident sur les évènements de type $[a, b]$, où $a \leq b$.

On peut reformuler ça de la manière suivante : P_X et P_Y coïncident ssi, pour tout $f \in \mathcal{F}$, $\mathbb{E}(f(X)) = \mathbb{E}(f(Y))$, où f est l'ensemble des fonctions de type $\mathbb{1}_{[a,b]}$, $a \leq b$. On voit alors que l'égalité des lois est caractérisée par l'égalité des espérances d'une famille de fonctions tests, et vous pouvez réécrire les deux premières caractérisations en termes de caractérisation d'espérances de fonctions tests. Dès lors, la caractérisation par fonctions caractéristiques revient à prendre une famille de fonctions tests bien spécifiques : l'ensemble des $x \mapsto e^{itx}$, $t \in \mathbb{R}$.

D'autres familles de fonctions tests caractérisent les lois : par exemple les fonctions positives, les fonctions continues bornées, les fonctions C^∞ à support compact, etc.. Pour creuser plus avant ces caractérisations il est nécessaire d'aller suivre un cours d'intégration standard.

Espérance et variance

De manière générale, pour une variable aléatoire réelle X, l'espérance de X est définie par

$$\int_\Omega X(\omega)\mathbb{P}(d\omega),$$

ce qui nécessite d'expliciter $(\Omega, \mathbb{P})$. De manière générale là encore, on peut écrire

$$\int_\Omega X(\omega)\mathbb{P}(d\omega) = \int_\mathbb{R} xP_X(dx),$$

c'est à dire que "l'intégrale de la fonction X sur Ω par rapport à la mesure $\mathbb{P}$" est la même chose que "l'intégrale de la fonction identité (x) sur $\mathbb{R}$ par rapport à la mesure P_X" : on parle de formule de transfert. L'espérance est donc totalement caractérisée par la loi, et là encore suivant le cas (discret ou continu) on aura des modes de calculs différents.

— Si X est discrète, alors

$$\mathbb{E}(X) = \sum_{k\in\mathbb{Z}} \mathbb{P}(X = k)k = \sum_{k\in\mathbb{Z}} P_X(\{k\})k.$$

— Si X est continue, alors

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f_X(x) dx.$$

On voit alors qu'une condition suffisante d'existence de ces espérances est :

— pour le cas discret : $\sum_{k\in\mathbb{Z}} |k|\mathbb{P}(X = k) < +\infty$.

— pour le cas continu : $\int_{\mathbb{R}} |x| f_X(x)dx < +\infty$.

On parle alors *d'intégrabilité* de X (on dit aussi que $X \in L_1(\mathbb{P})$, ce qui correspond à la notion d'intégrabilité de X vue comme fonction). On rappelle ci-dessous les propriétés de base de l'espérance.

Proposition 1.10 : Propriétés de l'espérance

Soient X, Y des variables aléatoires réelles et intégrables, et a,b dans $\mathbb{R}$. On a alors

— $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$ *(linéarité de l'espérance).*

— *Si $X \geq 0$, alors $\mathbb{E}(X) \geq 0$ (positivité de l'espérance).*

— *Si $X \geq 0$ et $\mathbb{E}(X) = 0$, alors $X = 0$ presque sûrement (au sens que $\mathbb{P}(X = 0) = 1$.*

— *Si ϕ est une fonction convexe, alors $\mathbb{E}(\phi(X)) \geq \phi(\mathbb{E}(X))$ (inégalité de Jensen).*

Un exemple de calcul pratique.

Exemple 1.11 *: Jeux de dés suite.* Dans l'expérience où on lance 2 dés et on regarde leur somme Z, si on cherche à calculer l'espérance de Z, deux options (voir trois).

— Poser proprement $\Omega = [\![1, 6]\!]^2$, $\mathbb{P}(\{(i, j)\}) = 1/36$, et

$$\mathbb{E}(Z) = \sum_{i=1}^{6}\sum_{j=1}^{6} \frac{i + j}{36}.$$

— Ne bosser qu'avec les variables aléatoires (recommandé). On pose X_1 et X_2 les résultats des deux dés, qui chacun suivent une loi uniforme sur $[\![1, 6]\!]$. Par linéarité on a alors

$$\mathbb{E}(Z) = \mathbb{E}(X_1) + \mathbb{E}(X_2).$$

Arrivé ici soit on connaît ses lois classiques (recommandé encore) et on sait que $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 7/2$. Soit on retrouve rapidement ce résultat en utilisant un argument de symétrie : la variable $7 - X_1$ est elle aussi uniforme sur $[\![1, 6]\!]$ (met la même répartition de masse dessus). On en déduit alors que $\mathbb{E}(X_1) = \mathbb{E}(7 - X_1)$, et donc que $2\mathbb{E}(X_1) = 7$ par linéarité.

On comprend bien l'intérêt d'éviter le $(\Omega, \mathbb{P})$.

La variance mesure l'étalement de la loi autour de son espérance, et informellement vaut l'espérance des écarts à l'espérance au carré. Plus précisément, si X^2 est intégrale (c'est à dire $\mathbb{E}(X^2) < +\infty$), la variance est définie par

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

Au passage, on peut remarquer que $E(X^2) < +\infty$ implique $E(|X|) < +\infty$: en effet, l'inégalité de Jensen donne $\mathbb{E}(|X|)^2 \leq \mathbb{E}(X^2)$. Pour une variable discrète, on calculera donc

$$\mathrm{Var}(X) = \sum_{k\in\mathbb{Z}} \mathbb{P}(X=k)(k-\mathbb{E}(X))^2 = \sum_{k\in\mathbb{Z}} \mathbb{P}(X=k)k^2 - \left(\sum_{k\in\mathbb{Z}} \mathbb{P}(X=k)k\right)^2,$$

et pour une variable continue

$$\mathrm{Var}(X) = \int_{\mathbb{R}} (x-\mathbb{E}(X))^2 f_X(x)dx = \int_R x^2 f_X(x)dx - \left(\int_{\mathbb{R}} x f_X(x)dx\right)^2.$$

Proposition 1.12 : Propriétés de la variance

Soit X une variable aléatoire telle que $\mathbb{E}(X^2) < +\infty$, et a, b dans $\mathbb{R}$. On a alors

- $\mathrm{Var}(aX+b) = a^2\mathrm{Var}(X)$.
- *Si* $\mathrm{Var}(X) = 0$, *alors* $X = \mathbb{E}(X)$ *presque sûrement.*

On remarque au passage que, en toute généralité, la variance n'est **pas** linéaire. Par exemple si on prend $Y = -X$, alors $\mathrm{Var}(X+Y) = 0 \neq \mathrm{Var}(X) + \mathrm{Var}(Y) = 2\mathrm{Var}(X)$ (si $\mathrm{Var}(X) \neq 0$). Un cas standard où la linéarité de la variance apparaît est celui de l'indépendance.

Lois classiques

On liste ici les lois classiques (à connaître), avec leurs fonctions de répartition, fonctions caractéristiques, espérance, variance, et expériences auxquelles elles correspondent.

Lois discrètes

- **Loi de Bernoulli** : $X \sim \mathcal{B}(p)$, où $p \in [0,1]$.
  - Expérience de pile ou face avec probabilité de succès p.
  - Loi sur $\{0,1\}$, caractérisée par $\mathbb{P}(X=1) = p$.
  - Fonction de répartition : $F_X(t) = (1-p)\mathbb{1}_{t\in[0,1[} + \mathbb{1}_{t\geq 1}$.
  - Fonction caractéristique : $\phi_X(t) = (1-p) + pe^{it}$.
  - Espérance : p. Variance : $p(1-p)$.
- **Loi Binomiale** : $X \sim \mathcal{B}(n,p)$, $n \in \mathbb{N}^*$, $p \in [0,1]$.
  - Expérience de n pile ou face à n lancer où on compte le nombre de succès (avec une proba de succès p).
  - Loi sur $[\![0,n]\!]$, donnée par

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

  - Fonction de répartition : ignoble.
  - Fonction caractéristique : $((1-p) + pe^{it})^n$.
  - Espérance : np. Variance : $np(1-p)$.

— **Loi uniforme** : $X \sim \mathcal{U}(\llbracket 1, n \rrbracket)$, où $n \in \mathbb{N}^*$.
  — Expérience de tirage au hasard dans un ensemble à n éléments.
  — Loi sur $\llbracket 1, n \rrbracket$, caractérisée par

$$\forall k \in \llbracket 1, n \rrbracket \quad \mathbb{P}(X = k) = \frac{1}{n}.$$

  — Fonction de répartition : $F_X(t) = \left[\left(\frac{\lfloor t \rfloor}{n}\right) \wedge 1\right] \mathbb{1}_{t \geq 0}$.
  — Fonction caractéristique : $\phi_X(t) = \frac{e^{it}}{n} \frac{e^{int}-1}{e^{it}-1}$.
  — Espérance : $(n+1)/2$. Variance : $(n^2-1)/12$.
— **Loi Géométrique** : $X \sim \mathcal{G}(p)$, où $p \in]0, 1]$.
  — Modélise la loi du temps d'arrivée du premier succès dans une succession de pile ou face avec probabilité de succès individuel p.
  — Loi sur $\mathbb{N}^*$ donnée par

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

  — Fonction de répartition : $F_X(t) = (1 - (1-p)^{\lfloor t \rfloor}) \mathbb{1}_{t \geq 1}$.
  — Fonction caractéristique : $\phi_X(t) = \frac{pe^{it}}{1-(1-p)e^{it}}$.
  — Espérance : $1/p$. Variance $(1-p)/p^2$.
— **Loi Hypergéométrique** : $X \sim \mathcal{H}(n, n_1, N)$, où $n \leq n_1 \leq N \in \mathbb{N}^*$.
  — Si on tire n éléments sans remise dans un grand ensemble à N éléments (disons des poissons) avec une sous-population à n_1 éléments (disons des carpes), cela modélise le nombre de carpes tirées (loi du nombre d'éléments tirés dans la sous-population).
  — Loi sur $\llbracket 0, n \rrbracket$, donnée par

$$\forall k \in \llbracket 0, n \rrbracket \quad \mathbb{P}(X = k) = \frac{\binom{n_1}{k}\binom{N-n_1}{n-k}}{\binom{N}{k}}.$$

  — Fonction de répartition : ignoble.
  — Fonction caractéristique : ignoble.
  — Espérance : np. Variance : $np(1-p)\frac{N-n_1}{N-1}$, où $p = n_1/N$.
— **Loi de Poisson** : $X \sim \mathcal{P}(\lambda)$, où $\lambda > 0$.
  — Correspond à un modèle limite du nombre de succès dans une expérience de N pile ou face avec proba de succès individuel p, lorsque $Np \to \lambda$ et $N \to +\infty$.
  — Loi sur $\mathbb{N}$, donnée par

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

  — Fonction de répartition : ignoble.
  — Fonction caractéristique : $\phi_X(t) = e^{\lambda(e^{it}-1)}$.
  — Espérance : λ. Variance : λ.

Lois continues

— **Loi uniforme** : $X \sim \mathcal{U}(]a,b[)$, $a<b$.

— Densité :

$$f_X(t) = \frac{1}{b-a}\mathbb{1}_{a<t<b}.$$

— Fonction de répartition :

$$F_X(t) = \left(\left(\frac{t-a}{b-a}\right) \wedge 1\right)\mathbb{1}_{t>a}.$$

— Fonction caractéristique : $\phi_X(t) = \frac{e^{itb}-e^{ita}}{it(b-a)}$.

— Espérance : $(a+b)/2$. Variance : $(b-a)^2/12$.

— **Loi exponentielle** : $X \sim \mathcal{E}(\lambda)$, $\lambda > 0$.

— Modélise souvent des temps d'attente.

— Densité :

$$f_X(t) = \lambda e^{-\lambda t}\mathbb{1}_{t>0}.$$

— Fonction de répartition : $F_X(t) = (1-e^{-\lambda t})\mathbb{1}_{t>0}$.

— Fonction caractéristique : $\phi_X(t) = \frac{\lambda}{\lambda - it}$.

— Espérance : $1/\lambda$. Variance : $1/\lambda^2$.

— **Loi normale** (ou Gaussienne) : $X \sim \mathcal{N}(\mu, \sigma^2)$, où $\mu \in \mathbb{R}$ et $\sigma > 0$.

— Apparaît comme une loi limite de fluctuation autour de moyennes (cf la suite).

— Densité :

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

— Fonction de répartition : ignoble.

— Fonction caractéristique : $\phi_X(t) = e^{it\mu - \sigma^2 t^2/2}$.

— Espérance : μ. Variance : σ^2.

Recentrage et mise à l'échelle :

On a vu précédemment qu'on pouvait assez facilement calculer espérance et variance de $aX+b$, pour a, b dans $\mathbb{R}$. La loi de $aX+b$ se déduit facilement des fonctions de répartitions et caractéristiques :

— $F_{aX+b}(t) = F_X((t-b)/a)$ (si $a>0$),

— $\phi_{aX+b}(t) = e^{itb}\phi_X(at)$.

On peut aussi facilement donner les densités et probabilités (cas continu et discret).

— Si X est discrète, et $Y = aX+b$, avec $a \neq 0$, alors

$$\mathbb{P}(Y=y) = \mathbb{P}\left(X = \frac{y-b}{a}\right).$$

— Si X est continue, et $Y = aX + b$, avec $a \neq 0$, alors

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right).$$

Exemple 1.13 *: Lois normales.* Soit $X \sim \mathcal{N}(0,1)$, $\mu \in \mathbb{R}$ et $\sigma > 0$. Alors $Y = \mu + \sigma X$ a pour densité

$$\begin{aligned} f_Y(y) &= \frac{1}{\sigma} f_X\left(\frac{(y-\mu)}{\sigma}\right) \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, \end{aligned}$$

ce dont on déduit $Y \sim \mathcal{N}(\mu, \sigma^2)$. Pour ne pas s'embêter dans les calculs, on raisonne souvent en sens inverse : pour une $Y \sim \mathcal{N}(\mu, \sigma^2)$ donnée, on pose $X = (Y - \mu)/\sigma$, fait les calculs sur X et remonte après à Y.

Pareillement, on peut vérifier que $\lambda\mathcal{E}(1) \sim \mathcal{E}(1/\lambda)$, pour $\lambda > 0$ (si vous avez un doute, regardez les espérances).

1.1.3 Couple de variable aléatoires réelles

Vu qu'on peut définir une variable aléatoire X à valeur dans n'importe quel espace $\mathcal{X}$ muni d'une tribu $\mathcal{B}$, on peut définir un couple de variable aléatoires réelles (X, Y) comme une variable aléatoire Z à valeurs dans $\mathbb{R}^2$.

La loi d'une telle variable est alors entièrement déterminée par la masse des évènements de base **rectangles**, c'est à dire

$$P_{(X,Y)}([a_1, b_1] \times [a_2, b_2]) = \mathbb{P}(X \in [a_1, b_1] \cap Y \in [a_2, b_2]),$$

pour $a_j \leq b_j$. On peut détailler un peu suivant les cas :

— *continue/continue* : la loi est déterminée par $P_{(X,Y)}([a_1, b_1] \times [a_2, b_2])$ (loi des rectangles). Si le couple a une densité $f_{(X,Y)}$, cela s'écrit

$$\int_{a_1}^{b_1} dx \int_{a_2}^{b_2} dy f_{(X,Y)}(x, y).$$

— *continue/discrète* : la loi est déterminée par

$$\mathbb{P}(X \in [a_1, b_1] \cap Y = k), \qquad \text{pour } k \in \mathbb{Z}.$$

— *discrète/discrète* : la loi est déterminée par

$$\mathbb{P}(X = k_1 \cap Y = k_2), \qquad \text{pour } k_1, k_2 \in \mathbb{Z}.$$

Comme dans le cas d'une seule v.a.r., on peut définir une **fonction de répartition** multivariée qui caractérise entièrement la loi :

$$F_{(X,Y)}(t_1, t_2) = \mathbb{P}(X \leq t_1 \cap Y \leq t_2).$$

Dans le cas continue/continue, la densité du couple sera donnée par

$$f_{(X,Y)}(x, y) = (\partial_x \partial_y F_{(X,Y)})(x, y).$$

On peut aussi définir une fonction caractéristique multivariée, mais on n'utilisera pas ce concept pour ce cours.

Le point à bien comprendre est qu'il ne suffit pas d'avoir la loi de X et la loi de Y (les lois marginales) pour connaître la loi du couple (X, Y) : il faut aussi la structure de dépendance entre X et Y.

Une illustration en est donnée lorsqu'on essaye de calculer la variance de $X+Y$. On remarque que pour calculer une variance faisant intervenir X et Y, il faut une information sur la loi du couple (X, Y) (ou a minima sur leur covariance). En effet, on a

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

En définissant

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

on a la formule

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

Le terme traduisant la dépendance entre X et Y dans la variance de la somme est alors le terme de covariance.

Exemple 1.14 *: Lois marginales ne déterminent pas la loi de couple.* On prend $X \sim \mathcal{N}(0, 1)$, $Y_1 = X$, et $Y_2 = -X$. On a alors que $Y_1 \sim Y_2$. Or (X, Y_1) ne peut avoir même loi que (X, Y_2) :

$$\mathrm{Var}(X+Y_1) = 4 \neq 0 = \mathrm{Var}(X+Y_2).$$

Cela est dû au fait que $\mathrm{Cov}(X, Y_1) = 1$ et $\mathrm{Cov}(X, Y_2) = -1$.

Pour spécifier la loi du couple (X, Y), il faut donc spécifier, en plus des marginales P_X et P_Y une structure de dépendance entre X et Y. La plus simple de cette structure est **l'indépendance**.

Definition 1.15 : Indépendance de v.a.r.

> Soient X et Y deux v.a.r.. X et Y sont dites indépendantes ssi, pour tous $a_1 \leq b_1$ et $a_2 \leq b_2$,
>
> $$\mathbb{P}(X \in [a_1, b_1] \cap Y \in [a_2, b_2]) = \mathbb{P}(X \in [a_1, b_1])\mathbb{P}(Y \in [a_2, b_2]).$$

Là encore on peut décliner suivant les caractères discret/continu des marginales. Grosso modo, les deux variables sont indépendantes si la "densité/fonction de masse" du couple est produit des "densités/fonctions de masse" des lois marginales. On peut caractériser l'indépendance en prenant d'autres fonctions tests que des indicatrices de rectangle.

Proposition 1.16 : Caractérisations de l'indépendance

Soient X et Y deux v.a.r.. Les propositions suivantes sont équivalentes.

1. *X et Y sont indépendantes.*
2. *Pour toutes fonctions $g, h : \mathbb{R} \to \mathbb{R}$ pour lesquelles $\mathbb{E}(g(X))$ et $\mathbb{E}(h(Y))$ sont bien définies*
$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y)).$$
3. *Pour tous $t_1, t_2 \in \mathbb{R}$, $F_{(X,Y)}(t_1, t_2) = F_X(t_1)F_Y(t_2)$.*
4. *Pour tous $t_1, t_2 \in \mathbb{R}$, $\mathbb{E}(e^i t_1 X + t_2 Y) = \phi_X(t_1)\phi_Y(t_2)$.*

Pour le deuxième point, on peut regarder les fonctions de X (et Y) intégrables, ou positives, ou encore les fonctions continues et bornées, voire $\mathcal{C}^\infty$ à support compact. Pour le dernier point, le terme de gauche dans l'égalité est de fait la fonction caractéristique multivariée du couple (X, Y).

En pratique, si le fait que $X \perp\!\!\!\perp Y$ vous est spécifié, alors la caractérisation en $\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$ s'avère assez utile.

Exemple 1.17 *: Loi d'une somme de variables Gaussiennes indépendantes.* Soit $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ et $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. On s'intéresse à la loi de $X + Y$. De manière générale, la loi d'une somme de variables indépendantes se calcule assez bien via fonction caractéristiques :

$$\begin{aligned}
\phi_{X+Y}(t) &= \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itX}e^{itY}) \\
&= \mathbb{E}(e^{itX})\mathbb{E}(e^{itY}) \qquad \text{(par indépendance)} \\
&= e^{it\mu_1 - \sigma_1^2 t^2} e^{it\mu_2 - \sigma_2^2 t^2} \\
&= e^{it(\mu_1+\mu_2)} e^{-(\sigma_1^2+\sigma_2^2)t^2/2}.
\end{aligned}$$

On en déduit alors que $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

En revanche, si l'on vous demande de prouver que X et Y sont indépendantes, il est souvent plus commode de regarder les fonctions de masse/densité du couple. Ci-dessous les recettes standard en fonction des modalités discret/continu.

Exemple 1.18 *: Recettes standard pour preuve d'indépendance.*

— **Continu/continu**

— *Principe général* : On prend X et Y deux v.a.r. continues. La plupart du temps il s'agit soit de prouver que la densité $f_{(X,Y)}(x, y)$ (si elle existe) se met sous la forme $g(x)h(y)$, soit de montrer que la fonction de répartition se met sous la forme produit $F_{(X,Y)}(t_1, t_2) = G(t_1)H(t_2)$. A normalisation près cela suffit.

— *Exemple* : Soit (X, Y) un couple de densité
$$f(x, y) = \frac{\lambda\mu}{y} e^{-\lambda\frac{x}{y} - \mu y} \mathbb{1}_{x>0} \mathbb{1}_{y>0},$$
où $\lambda, \mu > 0$. On va montrer que $\frac{X}{Y} \perp\!\!\!\perp Y$, en regardant la densité du couple $(X/Y, Y)$. Soient alors g, h deux fonction mesurables positives (ou

C^∞ à support compact),

$$\mathbb{E}(g(X/Y)h(Y)) = \int_0^{+\infty}\int_0^{+\infty} g(x/y)h(y)\frac{\lambda\mu}{y}e^{-\lambda\frac{x}{y}-\mu y}dxdy.$$

Soit

$$\psi : \begin{cases}]0,+\infty[^2 & \to &]0,+\infty[^2 \\ (x,y) & \mapsto & (x/y,y). \end{cases}$$

On remarque que ψ est un C^1 difféomorphisme (gaffe aux espaces de départ et d'arrivée), de Jacobien

$$|J_\psi|_{x,y} = \begin{vmatrix} 1/y & -x/y^2 \\ 0 & 1 \end{vmatrix} = 1/y.$$

La formule de changement de variable donne alors (en posant $(u,v) = \psi(x,y)$),

$$\begin{aligned}\mathbb{E}(g(X/Y)h(Y)) &= \int_0^{+\infty}\int_0^{+\infty} g(u)h(v)\lambda\mu\,|J_\psi|_{x,y}\,e^{-\lambda u-\mu v}dxdy \\ &= \int_0^{+\infty}\int_0^{+\infty} g(u)h(v)(\lambda e^{-\lambda u})(\mu e^{-\mu v})dudv.\end{aligned}$$

On en déduit que la densité de $(X/Y,Y)$ est $(\lambda e^{-\lambda u})(\mu e^{-\mu v})$, qui se met bien sous forme d'un produit des densités marginales. On a donc que $X/Y \perp\!\!\!\perp Y$. On voit qu'au passage on récupère les lois marginales $X/Y \sim \mathcal{E}(\lambda)$, $Y \sim \mathcal{E}(\mu)$.

— **Discret/Discret**

— *Principe général* : La plupart du temps on se donne k_1, k_2 dans $\mathbb{Z}$ et on regarde $\mathbb{P}(\{X = k_1\} \cap \{Y = k_2\})$. Si cela ce met sous la forme $g(k_1)h(k_2)$ c'est gagné. De temps en temps on peut jouer au même jeu avec les fonctions de répartition.

— *Exemple* : On se donne (X,Y) de loi jointe sur $\mathbb{N}^2$ donnée par

$$\mathbb{P}((X,Y) = (k_1,k_2)) = \left(\frac{\lambda}{\mu}\right)^{k_1}\mu^{k_2}e^{-(\lambda+\mu)}\frac{1}{k_1!(k_2-k_1)!}\mathbb{1}_{k_2\geq k_1},$$

où $\lambda,\mu > 0$. On va montrer que $X \perp\!\!\!\perp Y - X$. Soient $n_1, n_2 \in \mathbb{N}$,

$$\begin{aligned}\mathbb{P}\left(\{X = n_1\} \cap \{Y - X = n_2\}\right) &= \mathbb{P}((X,Y) = (n_1, n_1+n_2)) \\ &= \left(\frac{\lambda}{\mu}\right)^{n_1}\mu^{n_1+n_2}e^{-\lambda}e^{-\mu}\frac{1}{n_1!n_2!}\mathbb{1}_{n_1+n_2\geq n_1} \\ &= \left(e^{-\lambda}\frac{\lambda^{n_1}}{n_1!}\right)\left(e^{-\mu}\frac{\mu^{n_2}}{n_2!}\right).\end{aligned}$$

Au passage, on a choisi n_1 et n_2 dans $\mathbb{N}$ parce qu'on savait à l'avance que c'étaient les options de masse positive. Si vous n'avez pas cette intuition, il est plus prudent de conserver n_1, n_2 dans $\mathbb{Z}$ et de garder l'indicatrice $\mathbb{1}_{n_1+n_2\geq n_1}\mathbb{1}_{n_1\geq 0}\mathbb{1}_{n_2\geq 0}$. On a bien une forme produit, et donc $Y - X \perp\!\!\!\perp X$ (et on a encore les lois marginales, $\mathcal{P}(\lambda)$ et $\mathcal{P}(\mu)$.

— **Discret/Continu**

— *Principe général* : Si X est discrète et Y continue, la plupart du temps il faut calculer $\mathbb{P}(X = k_1 \cap Y \leq t_2)$ et montrer que cela se met sous forme $g(k_1)H(t_2)$ (cela suffit). Pour les plus avancés, on peut aussi calculer la densité du couple en regardant $\mathbb{E}(f(X)g(Y))$, pour des f, g positives et montrer que cela se met sous la forme produit.

— *Exemple* : On se donne $Z \sim \mathcal{E}(\lambda)$, et on pose $X = \lfloor Z \rfloor$ (partie entière), $Y = Z - X$ (partie fractionnaire). On va montrer que $X \perp\!\!\!\perp Y$. De manière immédiate X est à valeurs dans $\mathbb{N}$ et Y à valeurs dans $[0, 1[$. Soit donc $k_1 \in \mathbb{N}$, et $t_2 \in [0, 1[$. On a

$$\begin{aligned}\mathbb{P}(\{X = k_1\} \cap \{Y \leq t_2\}) &= \mathbb{P}(Z \in [k_1, k_1 + t_2]) \\ &= \int_{k_1}^{k_1+t_2} \lambda e^{-\lambda t} dt \\ &= e^{-\lambda k_1} - e^{-\lambda(k_1+t_2)} \\ &= e^{-\lambda k_1}(1 - e^{-\lambda t_2}).\end{aligned}$$

On a donc un produit du type $g(k_1)H(t_2)$, et donc $X \perp\!\!\!\perp Y$. Là il faut bosser un peu plus pour avoir les lois marginales. Comme $Y \leq 1$ presque sûrement, on a

$$\mathbb{P}(X = k_1) = \mathbb{P}(\{X = k_1\} \cap \{Y \leq 1\}) = e^{-\lambda k_1}(1 - e^{-\lambda}).$$

On en déduit que $X \sim \mathcal{G}(e^{-\lambda} - 1)$. Une autre manière de faire est de se rendre compte que au vu de la formule $\mathbb{P}(X = k_1)$ est de la forme $e^{-\lambda k_1} \times C$, où C est une constante, et de calculer la constante en utilisant $\sum_{k_1} \mathbb{P}(X = k_1) = 1$.
Pour Y, on a (pour $t_2 \in [0, 1[$),

$$\begin{aligned}\mathbb{P}(Y \leq t_2) &= \sum_{k_1=0}^{+\infty} \mathbb{P}(\{X = k_1\} \cap \{Y \leq t_2\}) \qquad \text{(probabilités totales)} \\ &= \frac{1 - e^{-\lambda t_2}}{1 - e^{-\lambda}},\end{aligned}$$

ce qui caractérise entièrement la loi de Y (on peut dériver pour retrouver la densité sur $[0, 1[$).

L'indépendance n'est qu'un cas particulier de structure de dépendance entre deux variables X et Y. En toute généralité, on peut définir une notion de *loi conditionnelle* de Y sachant X, notée $P_{Y|X}$, qui est une loi aléatoire (elle dépend de X). L'indépendance se traduit alors en $P_{Y|X} = P_Y$ (la loi de Y sachant X ne dépend pas de X, et est donc constante et égale à la loi marginale de Y toute seule). Lorsque X est discrète, la loi conditionnelle se définit facilement.

Definition 1.19 : Loi conditionnelle - Conditionnement discret

Soit (X, Y) un couple de v.a.r., où X est discrète. Pour $k \in \mathbb{Z}$ tel que $\mathbb{P}(X = k) \neq 0$, la *loi de Y sachant $X = k$* est définie par :

— si Y est discrète

$$P_{Y|X=k}(\ell) = \frac{\mathbb{P}(\{X = k\} \cap \{Y = \ell\})}{\mathbb{P}(X = k)},$$

pour tout $\ell \in \mathbb{Z}$.

— si Y est continue, $P_{Y|X=k}$ est la loi de fonction de répartition

$$F_{Y|X=k}(t) = \frac{\mathbb{P}(\{X = k\} \cap \{Y \leq t\})}{\mathbb{P}(X = k)},$$

pour $t \in \mathbb{R}$. Pour trouver la densité de $P_{Y|X=k}$, il suffit de dériver.

Dans les deux cas cela définit bien une loi de probabilité, à k fixé.

La *loi conditionnelle* de Y sachant X est alors une loi de probabilité aléatoire, $P_{Y|X}$, qui vérifie

$$P_{Y|X} \mathbb{1}_{X=k} = \mathbb{1}_{X=k} P_{Y|X=k}.$$

En d'autre termes c'est une loi aléatoire qui vaut $P_{Y|X=k}$ lorsque X vaut k.

Lorsque X est continue, on peut encore définir une loi conditionnelle de Y sachant X, en passant par la notion d'espérance conditionnelle (qui définit l'espérance par rapport à la loi conditionnelle). Cela sort du cadre de ce cours. Un exemple de calcul de loi conditionnelle.

Exemple 1.20 *: Un terme sachant la somme.* Soient X et Y deux variables aléatoires indépendantes de lois respectives $\mathcal{P}(\lambda)$ et $\mathcal{P}(\mu)$. On s'intéresse à la loi de $X \mid X+Y$.

On commence par remarquer que $\{k \mid \mathbb{P}(X+Y = k) > 0\} = \mathbb{N}$. Soit donc $k \in \mathbb{N}$, et $\ell \in \mathbb{N}$. Si $\ell > k$, on a $\mathbb{P}(\{X + Y = k\} \cap X = \ell) = 0$, et donc $P_{X|X+Y=k}(\ell) = 0$. Si $\ell \leq k$,

$$\begin{aligned} P(\{X = \ell\} \cap \{X + Y = k\}) &= \mathbb{P}(\{X = \ell\} \cap \{Y = k - \ell\}) \\ &= \mathbb{P}(X = \ell)\mathbb{P}(Y = k - \ell) \qquad \text{(indépendance)} \\ &= e^{-\lambda} \frac{\lambda^\ell}{\ell!} e^{-\mu} \frac{\mu^{k-\ell}}{(k-\ell)!}. \end{aligned}$$

Par ailleurs, $X + Y \sim \mathcal{P}(\lambda + \mu)$ (somme de deux lois de Poisson indépendantes), et on a alors

$$\mathbb{P}(X + Y = k) = e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!}.$$

On aurait aussi pu retrouver cette formule en regardant $\sum_\ell P(\{X = \ell\} \cap \{X + Y = k\})$. En divisant, on obtient

$$\begin{aligned} \mathbb{P}(X = \ell \mid X + Y = k) &= e^{-\lambda} \frac{\lambda^\ell}{\ell!} e^{-\mu} \frac{\mu^{k-\ell}}{(k-\ell)!} \times \left(e^{-(\lambda+\mu)} \frac{(\lambda+\mu)^k}{k!} \right)^{-1} \\ &= \binom{k}{\ell} \left(\frac{\lambda}{\lambda+\mu} \right)^\ell \left(\frac{\mu}{\lambda+\mu} \right)^{k-\ell}. \end{aligned}$$

On en déduit alors que $X \mid X+Y=k \sim \mathcal{B}(k,\rho)$, avec $\rho = \lambda/(\lambda+\mu)$, et on peut écrire $X \mid (X+Y) \sim \mathcal{B}(X+Y,\rho)$ (en notation compacte).

1.2 Convergences de suites de variables aléatoires réelles

On se donne maintenant une suite de variables aléatoires réelles (continues ou discrètes) $X_1, \ldots, X_n$. On a trois points de vue là-dessus :

1. On voit $X_1, \ldots, X_n$ comme une suite de fonctions de Ω dans $\mathbb{R}$, et on peut étudier leur convergence d'un point de vue fonctionnel. C'est l'objet des notions de convergence presque sûre, en probabilité, et dans les espaces $L_p(\Omega)$.
2. On regarde plutôt les lois $P_{X_1}, \ldots, P_{X_n}$ séquentiellement, on a donc une suite de lois, et on peut regarder leur convergence. C'est l'objet de la *convergence en loi*.
3. On ne regarde plus les lois une par une, mais la loi du n-uplet $(X_1, \ldots, X_n)$, $P_{X_{1:n}}$. En toute généralité on peut calculer la loi du n-uplet si on a les dépendances, via $P_{X_{1:n}} = P_{X_1} \times \prod_{i=2}^n P_{X_i|X_{1:i-1}}$ (dans le cas discret vous pouvez maintenant le faire). Pour le volet statistique de ce cours et les deux résultats centraux (loi des grands nombres et théorème central limite), on se placera dans le cadre où les variables sont supposées indépendantes. Dans ce cas $P_{X_{1:n}} = \bigotimes P_{X_i}$, ce qui signifie
   — pour des variables discrètes,
   $$\mathbb{P}(X_{1:n} = x_{1:n}) = \prod_{i=1}^n \mathbb{P}(X_i = x_i).$$
   — pour des variables continues,
   $$f_{X_{1:n}}(x_{1:n}) = \prod_{i=1}^n f_{X_i}(x_i),$$
   où les f désignent les densité.

   Dans les deux cas la loi du n-uplet est le 'produit' des n lois marginales.

1.2.1 Point de vue suite de fonctions : convergence en probabilité et presque sûre

Ici on se place du point de vue fonctionnel : on voit $X_1, \ldots, X_n$ comme une suite de fonctions de $(\Omega, \mathbb{P})$ à valeurs dans $\mathbb{R}$. La première notion de convergence est celle de la convergence presque sûre :

Definition 1.21 : CVPS

Soit $(X_n)_{n\geq 1}$ une suite de v.a.r., et X une v.a.r.. On dit que $(X_n)_{n\geq 1}$ converge *presque sûrement* vers X si et seulement si

$$\mathbb{P}\left(\{\omega \mid X_n(\omega) \nrightarrow X(\omega)\}\right) = 0.$$

On notera alors

$$X_n \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} X.$$

En d'autre termes, la convergence presque sûre est la convergence ponctuelle de $(X_n)_{n\geq 1}$ vue comme une suite de fonctions, à un ensemble de probabilité nulle près. FAIRE DESSIN.

Une deuxième notion de convergence d'un point de vue fonctionnel qui sera utilisé en statistiques (si pas dans ce cours, dans le suivant) est celle de la convergence en probabilité.

Definition 1.22 : CV en proba

Soit $(X_n)_{n\geq 1}$ une suite de v.a.r., et X une v.a.r.. On dit que $(X_n)_{n\geq 1}$ converge *en probabilité* vers X si et seulement si

$$\forall \varepsilon > 0 \quad \mathbb{P}\left(\{|X_n - X| \geq \varepsilon\}\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

On notera alors

$$X_n \underset{n\to+\infty}{\overset{\mathbb{P}}{\longrightarrow}} X.$$

Avec des mots, la convergence en proba stipule que la masse des endroits où X_n et X diffèrent doit tendre vers 0. FAIRE DESSIN

Même si la masse des zones où X_n et X diffèrent tend vers 0, cette zone peut bouger, ce qui peut empêcher la convergence presque sûre de la suite de v.a.r.

Exemple 1.23 *: CV en proba n'implique pas CVPS.* On pose $\Omega = [0, 1[$, et $\mathbb{P}$ la probabilité uniforme dessus. On prend $X_1 = \mathbb{1}_{[0,1[}$, $X_2 = \mathbb{1}_{[0,1/2[}$, $X_3 = \mathbb{1}_{[1/2,1[}$, $X_4 = \mathbb{1}_{[0,1/4[}$, etc.. FAIRE DESSIN. D'un point de vue formel on pose, pour $n \geq 1$, $u_n = \lfloor \log_2(n) \rfloor$, et $v_n = n - 2^{u_n} \in [\![0, 2^{u_n} - 1[\![$, de sorte que $n = 2^{u_n} + v_n$, et on définit

$$X_n : \omega \mapsto \mathbb{1}_{[2^{-u_n} v_n, 2^{-u_n}(v_n+1)[}(\omega).$$

On a alors $X_n(\omega) = 1$ si et seulement si $\log_2(v_n) \leq u_n + \log_2(\omega) < \log_2(v_n + 1)$, et donc, pour la sous-suite donnée par

$$\phi(n) = 2^n + \lfloor \log_2(2^n \omega) \rfloor,$$

on a $X_{\phi(n)}(\omega) = 1$. On a donc

$$X_n \underset{n\to+\infty}{\overset{p.s.}{\nrightarrow}} 0.$$

Par ailleurs, si $\varepsilon < 1$,

$$\mathbb{P}(|X_n| \geq \varepsilon) = 2^{-u_n} < 2^{-(\log_2(n)-1)} = \frac{2}{n} \underset{n\to+\infty}{\longrightarrow} 0.$$

On en déduit

$$X_n \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} 0.$$

En revanche, dans l'autre sens c'est vrai.

Proposition 1.24 : CVps implique CV en proba

Soit $(X_n)_{n\geq 1}$ une suite de v.a.r., et X une v.a.r. telles que $(X_n)_{n\geq 1}$ converge presque sûrement vers X. Alors

$$X_n \underset{n\to+\infty}{\xrightarrow{\mathbb{P}}} X.$$

Démonstration. Soit $\varepsilon > 0$. On note $U_{n,\varepsilon}$ l'évènement

$$U_{n,\varepsilon} = \bigcap_{k\geq n} \{|X_n - X| < \varepsilon\}.$$

Comme X_n converge presque sûrement vers X, on a que

$$\mathbb{P}\left(\bigcup_{n\geq 1} U_{n,\varepsilon}\right) = 1,$$

où au passage on reconnaît une limite inférieure d'évènements. En prenant le complémentaire on trouve

$$\mathbb{P}\left(\bigcap_{n\geq 1} U^c_{n,\varepsilon}\right) = 0.$$

Or $U^c_{n,\varepsilon} = \{\exists j \geq n \mid |X_n - X| \geq \varepsilon\}$ est une suite décroissante d'évènements, on a donc

$$\lim_{n\to+\infty} \mathbb{P}(U^c_{n,\varepsilon}) = \mathbb{P}\left(\bigcap_{n\geq 1} U^c_{n,\varepsilon}\right) = 0.$$

Enfin, remarquons que $\{|X_n - X| \geq \varepsilon\} \subset U^c_{n,\varepsilon}$, et donc

$$0 \leq \mathbb{P}\left(\{|X_n - X| \geq \varepsilon\}\right) \leq \mathbb{P}\left(U^c_{n,\varepsilon}\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Au passage, si vous connaissez le théorème de convergence dominée, la preuve est immédiate. □

Enfin, on peut quand même passer d'une convergence en proba à une convergence presque sûre, mais via une sous-suite. Pour cela on admettra le Lemme suivant.

Lemme 1.25 : Lemme de Borel-Cantelli

Si $(A_n)_{n\geq 1}$ est une suite d'évènements satisfaisant

$$\sum_{n\geq 1} \mathbb{P}(A_n) < +\infty,$$

alors

$$\mathbb{P}\left(\limsup_n A_n\right) := \mathbb{P}\left(\bigcap_{n\geq 1}\bigcup_{k\geq n} A_k\right) = 0.$$

En d'autres termes, si les probas de la suites d'évènements sont sommables, la probabilité de l'ensemble des ω qui sont dans une infinité de ces évènements est nulle.

Preuve du Lemme 1.25. On note $B_n = \bigcup_{k\geq n} A_k$, et on remarque que B_n est une suite décroissante d'évènements. On a alors

$$\mathbb{P}(\limsup_n A_n) = \mathbb{P}\left(\bigcap_{n\geq 1} B_n\right) = \lim_{n\to+\infty} \mathbb{P}(B_n).$$

Par ailleurs on a

$$\mathbb{P}(B_n) \leq \sum_{k\geq n} \mathbb{P}(A_k) \underset{n\to+\infty}{\longrightarrow} 0,$$

car $\sum_{k\geq 1} \mathbb{P}(A_k)$ est finie. Cela conclut la preuve. □

Proposition 1.26 : Convergence en proba implique CVps d'une sous-suite

Soit $(X_n)_{n\geq 1}$ une suite de v.a.r., et X une v.a.r. telles que $(X_n)_{n\geq 1}$ converge en proba vers X. Alors il existe une sous-suite $(X_{\varphi(n)})_{n\geq 1}$ telle que

$$X_{\varphi(n)} \underset{n\to+\infty}{\overset{p.s.}{\longrightarrow}} X.$$

Démonstration. Pour $n \in \mathbb{N}^*$, on définit $\varphi(n)$ par récursion comme

$$\inf\left\{k > \varphi(k-1) \mid \forall p \geq k \quad \mathbb{P}(|X_p - X| \geq 1/n) \leq 2^{-n}\right\}.$$

Cela définit bien une sous-suite car $\mathbb{P}(|X_p - X| \geq 1/n) \to 0$ lorsque $p \to +\infty$. Par construction, on a, pour tout n,

$$\mathbb{P}\left(\{|X_{\varphi(n)} - X| \geq 1/n\}\right) \leq 2^{-n}.$$

En notant A_n l'évènement $\{|X_{\varphi(n)} - X| \geq 1/n\}$, on a alors que

$$\sum_{n\geq 1} \mathbb{P}(A_n) < +\infty.$$

En appliquant le Lemme de Borel-Cantelli, on en déduit que

$$\mathbb{P}\left(\bigcap_{n\geq 1}\bigcup_{k\geq n} A_k\right) = 0.$$

En notant $C = \bigcap_{n\geq 1}\bigcup_{k\geq n} A_k$, on a $\mathbb{P}(C^c) = 1$, et $C^c = \{\omega \mid \exists n \quad \forall k \geq n \quad |X_{\varphi(k)} - X|(\omega) < 1/k\}$. On a donc que pour tout $\omega \in C^c$, $X_{\varphi(n)}(\omega) \to X(\omega)$. □

La convergence en proba est une convergence au sens classique : elle est compatible avec l'addition, la multiplication, et plus généralement elle est transférable par continuité.

Proposition 1.27 : Propriétés de la convergence en proba

Soient $(X_n)_{n\geq 1}$ et $(Y_n)_{n\geq 1}$ deux suite de variables aléatoires réelles convergeant en probabilité respectivement vers X et Y. Alors, pour toute fonction continue $g : \mathbb{R}^2 \to \mathbb{R}$, on a

$$g(X_n, Y_n) \xrightarrow[n\to+\infty]{\mathbb{P}} g(X, Y).$$

En particulier, cela implique :

- *si $h : \mathbb{R} \to \mathbb{R}$ est continue, alors $h(X_n) \to^{\mathbb{P}} h(X)$,*
- *$aX_n + bY_n \to^{\mathbb{P}} aX + bY$, où $a, b \in \mathbb{R}$, et $X_nY_n \to^{\mathbb{P}} XY$.*

Démonstration. Soit $\varepsilon > 0$, on veut borner $\mathbb{P}(|g(X_n, Y_n) - g(X, Y)| \geq \varepsilon)$. g étant continue, si (X_n, Y_n) est proche de (X, Y), alors $g(X_n, Y_n)$ devrait être proche de $g(X, Y)$. Le problème est que (X, Y) est aléatoire, pour y remédier un argument de continuité **uniforme** serait l'idéal.

Pour se ramener à un argument de continuité uniforme : on se donne $M > 0$, et $\delta_{M,\varepsilon}$ le module d'uniforme continuité de g sur $[-M-1, M+1]^2$ (existe par le théorème de Heine, on peut le prendre plus petit que 1), de sorte que si $(x, y) \in [-M, M]^2$ et (x', y') est tel que $|(x, y) - (x', y')| \leq \delta_{M,\varepsilon}$, alors $|g(x, y) - g(x', y')| \leq \varepsilon$ (on voit ici que prendre $\delta_{M,\varepsilon}$ plus petit que 1 assure que $(x', y') \in [-M-1, M+1]^2$).

On a alors

$$\begin{aligned}
&\mathbb{P}\left(|g(X, Y) - g(X_n, Y_n)| \geq \varepsilon\right) \\
&\quad = \mathbb{P}\left(\{|g(X, Y) - g(X_n, Y_n)| \geq \varepsilon\} \cap \{(X, Y) \in [-M, M]^2\}\right) \\
&\qquad + \mathbb{P}\left(\{|g(X, Y) - g(X_n, Y_n)| \geq \varepsilon\} \cap \{(X, Y) \notin [-M, M]^2\}\right) \\
&\quad \leq \mathbb{P}\left(\{|g(X, Y) - g(X_n, Y_n)| \geq \varepsilon\} \cap \{(X, Y) \in [-M, M]^2\}\right) + \mathbb{P}((X, Y) \notin [-M, M]^2) \\
&\quad \leq \mathbb{P}(|(X, Y) - (X_n, Y_n)| > \delta_{M,\varepsilon}) + \mathbb{P}(|X| > M) + \mathbb{P}(|Y| > M).
\end{aligned}$$

Le premier terme s'écrit

$$\mathbb{P}(|(X, Y) - (X_n, Y_n)| > \delta_{M,\varepsilon}) \leq \mathbb{P}(|X - X_n| > \delta_{M,\varepsilon}/2) + \mathbb{P}(|Y - Y_n| > \delta_{M,\varepsilon}/2) \xrightarrow[n\to+\infty]{} 0,$$

par convergence de $(X_n)_{n\geq 1}$ et $(Y_n)_{n\geq 1}$ vers X et Y en probabilité. Pour les deux autres termes, comme $\mathbb{P}(|X| = +\infty) = 0 = \lim \downarrow_M \mathbb{P}(|X| > M)$, on en déduit, pour

tout $\eta > 0$, l'existence de M_η tel que $\mathbb{P}(|X| > M_\eta) \leq \eta/2$ et $\mathbb{P}(|Y| > M_\eta) \leq \eta/2$. On en déduit alors que pour tout $\eta > 0$,

$$\mathbb{P}\left(|g(X,Y) - g(X_n,Y_n)| \geq \varepsilon\right) \leq 2\eta,$$

à partir d'un certain rang (en utilisant le $\delta_{M_\eta,\varepsilon}$ défini précédemment). Cette inégalité étant valable pour tout $\eta > 0$, on en déduit que $\mathbb{P}\left(|g(X,Y) - g(X_n,Y_n)| \geq \varepsilon\right) \to_{n\to+\infty} 0$. □

Pour conclure sur la convergence en probas, on mentionne ici que la convergence en proba n'implique **pas** la convergence des espérances (si définies).

Exemple 1.28 *: Contre-exemple.* On définit X_n par $\mathbb{P}(X_n = n^2) = 1/n$, et $\mathbb{P}(X_n = 0) = 1 - 1/n$. On a alors, pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(|X_n| \geq \varepsilon\right) = 1/n \underset{n\to+\infty}{\longrightarrow} 0,$$

et donc X_n tend vers $X \equiv 0$ (variable aléatoire constante) en proba. En revanche, on a $\mathbb{E}(X_n) = n^2/n = n \to +\infty$, et donc $\mathbb{E}(X_n)$ ne tend pas vers $\mathbb{E}(X) = 0$.

Point de vue suite de fonctions hors programme : convergence L_p

CV Lp + sous suite
autre sens CV monotone, CV dom, Fatou

1.2.2 Point de vue suite de lois : convergence en loi

On se place maintenant du point de vue des lois, c'est à dire qu'on regarde uniquement $P_{X_1}, \ldots, P_{X_n}$, et on aimerait bien définir une notion de convergence de ces lois. En effet, la plupart du temps lorsque l'on tire $X_1, \ldots, X_n$ de manière indépendante, il n'y a quasiment aucune chance pour que $(X_n)_{n\geq 1}$ converge d'un point de vue fonctionnel. En revanche, les lois elles peuvent converger.

Exemple 1.29 *: Pile ou face infini.*

On se donne une suite $(X_n)_{n\geq 1}$ de lois de Bernoulli, que l'on supposera indépendante (tirage d'une infinité de pile ou face indépendants). On aura toujours $X_n \sim \mathcal{B}(p)$, pour $p \in]0,1[$, donc en terme de lois P_{X_n} est constante de loi $X \sim \mathcal{B}(p)$. Montrons que $(X_n)_{n\geq 1}$ ne peut converger vers une variable X en probabilité.

En effet, supposons qu'un telle variable X existe. On aurait alors $X_\varphi(n) \to X$ p.s. pour une sous-suite, et comme $X_\varphi(n) \in \{0,1\}$, on aurait alors $X \in \{0,1\}$ presque sûrement. Par ailleurs, on aurait aussi

$$\mathbb{P}(X_n \neq X) = \mathbb{P}(|X_n - X| \geq 1) \underset{n\to+\infty}{\longrightarrow} 0,$$

et donc

$$\begin{aligned}\mathbb{P}(X = 1) &= \mathbb{P}(X_n = 1 \cap X_n = X) + \mathbb{P}(X_n \neq X \cap X_n = 0)\\ &= \mathbb{P}(X_n = 1) - \mathbb{P}(X_n = 1 \cap X_n \neq X) + \mathbb{P}(X_n \neq X \cap X_n = 0).\end{aligned}$$

Comme $\mathbb{P}(X_n \neq X) \to 0$, on en déduit que $\mathbb{P}(X = 1) \to p$, donc vaut p, et donc que $X \sim \mathcal{B}(p)$. Jusque là on a prouvé d'une certaine manière que si X_n tend vers

X d'un point de vue fonctionnel, la convergence a aussi lieu en termes de lois (on verra plus loin que c'est un phénomène général).

Prouvons maintenant que X doit être nécessairement constante presque sûrement. Quitte à prendre une sous-suite supposons que la convergence a lieu presque sûrement. Regardons alors l'évènement

$$A_1 = \{X_n(\omega) \to X(\omega)\} \cap \{X(\omega) = 1\}.$$

On peut réécrire

$$A_1 = \bigcup_{n\geq 1} \bigcap_{k\geq n} \{X_k = 1\},$$

et donc

$$\mathbb{P}(A_1) = \lim_{n\to+\infty} \mathbb{P}(\bigcap_{k\geq n} \{X_k = 1\}).$$

Or $\mathbb{P}(\bigcap_{k\geq n}\{X_k = 1\}) = \lim_{\ell\to+\infty} p^\ell = 0$. On en déduit alors $\mathbb{P}(A_1) = 0$, et comme $X_n \to X$ presque sûrement $\mathbb{P}(X = 1) = 0 = p$, d'où la contradiction.

De manière plus générale, la loi du 0/1 de Kolmogorov vous apprendra que si un évènement se décrit comme limite d'évènements indépendants, (comme par exemple l'ensemble de convergence de variables indépendantes), alors cet évènement est nécessairement de probabilité 0 et 1. On en déduit alors que si une suite de variable indépendantes X_n converge fonctionnellement vers X, alors X est nécessairement constante presque sûrement.

Pour caractériser la convergence de suite de variables indépendantes, il va donc falloir trouver autre chose que la convergence fonctionnelle. Une manière d'envisager une telle convergence est de raisonner par dualité, en considérant les lois de variables aléatoires comme des opérateurs linéaires sur des fonctions intégrables sur $\mathbb{R}$.

De ce point de vue, on peut définir la convergence en loi comme suit.

Definition 1.30 : Convergence en loi - Définition standard

Soit $(X_n)_{n\geq 1}$ une suite de v.a.r., et X une v.a.r.. On dit que $(X_n)_{n\geq 1}$ *converge en loi* vers X si et seulement si

$$\forall f \in C_b(\mathbb{R}) \quad \mathbb{E}(f(X_n)) \underset{n\to+\infty}{\longrightarrow} \mathbb{E}(f(X)),$$

où $C_b(\mathbb{R})$ désigne l'espace des fonctions continues bornées sur $\mathbb{R}$. On notera alors

$$X_n \underset{n\to+\infty}{\rightsquigarrow} X.$$

On peut remarquer qu'il y a un léger abus : X_n converge en loi veut plutôt dire que P_{X_n} converge. Cette définition générale est assez peu utilisée hors champ des probabilités théoriques. En pratique on utilise

— soit une caractérisation par les fonctions de répartitions,

— soit une caractérisation par les fonctions caractéristiques,

— soit une caractérisation ad-hoc variables discrètes/variables continues.

On va passer ces trois méthodes en revue.

Proposition 1.31 : Cas des variables discrètes

Soient $(X_n)_{n\geq 1}$ et X des v.a.r. discrètes. On a alors

$$X_n \underset{n\to+\infty}{\rightsquigarrow} X \quad \Leftrightarrow \quad \forall k \in \mathbb{Z} \quad \mathbb{P}(X_n = k) \underset{n\to+\infty}{\longrightarrow} \mathbb{P}(X = k).$$

Démonstration. On fait uniquement le sens direct. Soit $k \in \mathbb{Z}$, on note f_k la fonction tente $f_k : x \mapsto (1 - 2|x - k|) \vee 0$, qui est bien continue et bornée. On a alors

$$\mathbb{P}(X_n = k) = \mathbb{E}(f_k(X_n)) \to_{n\to+\infty} \mathbb{E}(f_k(X)) = \mathbb{P}(X = k).$$

Si l'autre sens vous intéresse, il suit les grandes lignes des méthodes par densité : vous avez la convergence des $\mathbb{E}(f(X_n))$ vers $\mathbb{E}(f(X))$ tout d'abord pour les fonctions indicatrices autour des entiers, puis indicatrices d'ouvert tout court, donc fonctions étagées puis par densité fonctions continues bornées. □

Cette caractérisation est assez intuitive : X_n converge vers X si la masse que donne X_n à chaque élément converge vers celle de X.

Exemple 1.32 *: Classique à connaître.* On se donne X_n de loi $\mathcal{B}(n, p_n)$, avec $np_n \to_{n\to+\infty} \lambda$. Le résultat à connaître est

$$X_n \underset{n\to+\infty}{\rightsquigarrow} \mathcal{P}(\lambda),$$

c'est à dire que, pour n assez grand, si on compte le nombre de succès dans une expérience de pile ou face où la probabilité de succès est faible, alors à la limite ce nombre de succès se comportera comme une loi de Poisson.

Une preuve directe est la suivante. Soit $k \in \mathbb{N}$. On a

$$\begin{aligned}\mathbb{P}(X_n = k) &= \binom{n}{k} p_n^k (1 - p_n)^{n-k} \mathbb{1}_{k\leq n} \\ &= \frac{\binom{n}{k}}{n^k} (np_n)^k (1 - p_n)^{n-k} \mathbb{1}_{k\leq n}.\end{aligned}$$

D'une part

$$\frac{\binom{n}{k}}{n^k} = \frac{1}{k!} \left((1)(1 - 1/n) \dots (1 - (k - 1)/n)\right) \underset{n\to+\infty}{\longrightarrow} \frac{1}{k!}.$$

D'autre part

$$(np_n)^k \underset{n\to+\infty}{\longrightarrow} \lambda^k.$$

Ensuite, en remarquant que $np_n \to_{n\to+\infty} \lambda$ implique $p_n \to_{n\to+\infty} 0$,

$$(1 - p_n)^{n-k} = \exp\left[(n - k)\log(1 - p_n)\right] = \exp\left[-p_n(n - k) + o(np_n)\right] \underset{n\to+\infty}{\longrightarrow} e^{-\lambda},$$

puis de manière évidente $\mathbb{1}_{k\leq n} \to_{n\to+\infty} 1$. On déduit de tout cela

$$\mathbb{P}(X_n = k) \underset{n\to+\infty}{\longrightarrow} e^{-\lambda} \frac{\lambda^k}{k!} = \mathbb{P}(X = k),$$

où $X \sim \mathcal{P}(\lambda)$. On en déduit que $X_n \rightsquigarrow_{n\to+\infty} X$.

On a une caractérisation similaire pour les variables continues, en remplaçant "masse en un point" par masse des boréliens.

Proposition 1.33 : Cas des variables continues

Soit $(X_n)_{n\geq 1}$ une suite de variables continues, et X une variable continue, on a alors

$$\begin{aligned} X_n \underset{n\to+\infty}{\rightsquigarrow} X \quad &\Leftrightarrow \forall A \in \mathcal{B}(\mathbb{R}) \quad \mathbb{P}(X_n \in A) \underset{n\to+\infty}{\longrightarrow} \mathbb{P}(X \in A) \\ &\Leftrightarrow \quad \forall a < b \quad \int_a^b f_{X_n}(u)du \underset{n\to+\infty}{\longrightarrow} \int_a^b f_X(u)du. \end{aligned}$$

Par ailleurs, si pour presque tout $x \in \mathbb{R}$, $f_{X_n}(x) \to_{n\to+\infty} f_X(x)$, alors $X_n \rightsquigarrow_{n\to+\infty} X$.

La preuve repose essentiellement sur de la théorie de l'intégration et l'approximation des fonction $\mathbb{1}_A$ par des fonctions continues bornées (par exemple $(1 - k\mathrm{d}(x, A)) \vee 0$ lorsque A est un fermé). Cette caractérisation n'est mentionnée que pour son caractère intuitif : elle n'est pas généralisable, et assez peu employée en pratique. Plusieurs remarques :

1. La première caractérisation devient fausse en toute généralité (il faut A tel que $\mathbb{P}(\partial A) = 0$), un lecteur intéressé est renvoyé au Théorème de Portmanteau.
2. La caractérisation par convergence des densités est dans un sens uniquement : on peut construire X_n et X a densités telles que $X_n \rightsquigarrow_{n\to+\infty} X$ et $f_{X_n}(u) \nrightarrow_{n\to+\infty} f_X(u)$ nulle part.

Il faut plutôt privilégier les deux caractérisations suivantes, qui elles sont généralisables à n'importe quel type de variable aléatoire.

Proposition 1.34 : Caractérisation par les fonctions de répartition

Soient $(X_n)_{n\geq 1}$ et X des v.a.r., et $D = \{t \in \mathbb{R} \mid F_X \text{ est continue en } t\}$. On a alors

$$X_n \underset{n\to+\infty}{\rightsquigarrow} X \quad \Leftrightarrow \quad \forall t \in D \quad F_{X_n}(t) \to F_X(t).$$

On remarque que si la loi cible X est continue, alors sa fonction de répartition l'est et $D = \mathbb{R}$. Cette caractérisation sera très utile, notamment lorsque l'on devra construire des intervalles de confiance asymptotiques. Enfin, cette caractérisation permet de prouver la convergence de variables discrètes vers une variables continue (et réciproquement).

Exemple 1.35 *: Lois uniformes.* Soit X_n de loi uniforme sur $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$. On peut montrer que $X_n \rightsquigarrow_{n\to+\infty} X$, où $X \sim \mathcal{U}(]0, 1])$.

Pour cela regardons les fonctions de répartition (on notera F celle de X et F_n celle de X_n). Premièrement

$$F_X(t) = 0 \vee (t \wedge 1),$$

FAIRE DESSIN, qui est bien continue partout. Par ailleurs,

$$F_n(t) = O \vee \left(\frac{\lfloor nt \rfloor}{n} \wedge 1 \right),$$

FAIRE DESSIN AUSSI. Soit donc $t \in \mathbb{R}$. Si $t \leq 0$, alors $F_n(t) = F(t) = 0$, et si $t \geq 1$, $F_n(t) = F(t) = 1$, la convergence est triviale dans ces deux cas. Maintenant si $t \in]0, 1[$,

$$\frac{nt - 1}{n} < \frac{\lfloor nt \rfloor}{n} \leq t.$$

Le terme de gauche tendant vers t, on a bien $F_n(t) \to_{n\to+\infty} F(t)$, et donc $X_n \rightsquigarrow_{n\to+\infty} X$.

L'autre caractérisation passe par les fonctions caractéristiques.

Théorème 1.36 : Théorème de Lévy

Soient $(X_n)_{n\geq 1}$ et X des v.a.r. telles que $X_n \rightsquigarrow_{n\to+\infty} X$. Alors,

$$\forall t \in \mathbb{R} \quad \phi_{X_n}(t) \underset{n\to+\infty}{\longrightarrow} \phi_X(t).$$

Réciproquement, si pour tout t dans $\mathbb{R}$, $\phi_{X_n}(t)$ converge vers $\phi(t)$, où ϕ est une fonction continue en 0, alors ϕ est une fonction caractéristique (d'une variable notée X), et on a $X_n \rightsquigarrow_{n\to+\infty} X$.

On peut remarquer qu'à $(X_n)_{n\geq 1}$ et X fixées, ce résultat implique l'équivalence entre $X_n \rightsquigarrow_{n\to+\infty} X$ et la convergence simple des fonctions caractéristiques. D'un point de vue conceptuel, ces deux caractérisations (fonctions de répartition et caractéristiques) collent parfaitement avec la caractérisation d'une loi via des espérances de fonctions tests : il faut que l'espérance des fonctions tests caractérisant la loi de X_n converge vers l'espérance des fonctions tests caractérisant la loi de X, où les fonctions tests sont de la forme $\mathbb{1}_{]-\infty,t]}(x)$ dans un cas et e^{itx} dans l'autre. Dans un sens, la convergence en loi peut être vue comme une convergence faible.

Cette caractérisation de la convergence en loi par fonctions caractéristiques est très utile lorsqu'il s'agit de montrer la convergence en loi de somme de variables indépendantes.

Exemple 1.37 *: Somme de variables uniformes.*

Soient $X_1, \ldots, X_n, \ldots$ une suite de variables indépendantes, de loi $\mathcal{U}(]-1, 1[)$. On va regarder la convergence en loi de $S_n = \sqrt{\frac{3}{n}} \sum_{i=1}^n X_i$.

La fonction caractéristique d'un X_i est

$$\phi(t) = \frac{e^{it} - e^{-it}}{2it} = \frac{\sin(t)}{t}.$$

On en déduit alors la fonction caractéristique de S_n, notée ϕ_n :

$$\begin{aligned}
\phi_n(t) &= \mathbb{E}\left(e^{it\sqrt{\frac{3}{n}}(\sum_{j=1}^n X_j)}\right) \\
&= \mathbb{E}\left(\prod_{j=1}^n e^{it\sqrt{\frac{3}{n}}X_j}\right) \\
&= \prod_{j=1}^n \phi\left(\frac{\sqrt{3}t}{\sqrt{n}}\right) \qquad \text{(par indépendance)} \\
&= \left(\sqrt{n}\frac{\sin\left(\frac{\sqrt{3}t}{\sqrt{n}}\right)}{\sqrt{3}t}\right)^n.
\end{aligned}$$

Maintenant, on a $\sin(\sqrt{3}t/\sqrt{n}) = \sqrt{3}t/\sqrt{n} - t^3 3^{3/2}/(6n^{3/2}) + o(n^{-3/2})$, ce dont on déduit

$$\sqrt{n}\frac{\sin\left(\frac{\sqrt{3}t}{\sqrt{n}}\right)}{\sqrt{3}t} = 1 - \frac{t^2}{2n} + o(n^{-1}).$$

Ensuite,

$$\begin{aligned}
\phi_n(t) &= \left(1 - \frac{t^2}{2n} + o(n^{-1})\right)^n \\
&= \exp\left[n\log\left(1 - \frac{t^2}{2n} + o(n^{-1})\right)\right] \\
&= \exp\left[n\left(-\frac{t^2}{2n} + o(n^{-1})\right)\right] \\
&\underset{n\to+\infty}{\longrightarrow} e^{-t^2/2}.
\end{aligned}$$

On en déduit que $S_n \rightsquigarrow_{n\to+\infty} X$, où $X \sim \mathcal{N}(0,1)$. On vient de prouver manuellement un théorème central limite pour la loi uniforme sur $]-1,1[$.

On a vu dans le premier exemple que la convergence en loi n'impliquait pas la convergence en proba : de fait, une suite de variables aléatoires indépendantes peut converger en loi, elle ne pourra jamais converger en probabilité. Par contre, l'autre sens est vrai.

Proposition 1.38 : Convergence en proba implique convergence en loi

Soit $(X_n)_{n\geq 1}$ une suite de v.a.r. et X une v.a.r. telles que $(X_n)_{n\geq 1}$ converge vers X en probabilité. Alors

$$X_n \underset{n\to+\infty}{\rightsquigarrow} X.$$

Démonstration. Si on veut éviter les théorèmes de convergence dominée, on peut passer par la caractérisation par la convergence des fonctions de répartition. Soit donc t un point de continuité de F_X (qu'on notera F par la suite), et $\delta > 0$. En notant F_n la fonction de répartition de X_n, on a

$$\begin{aligned}
F_n(t) &= \mathbb{P}(X_n \leq t) \leq \mathbb{P}(\{X_n \leq t\} \cap \{|X - X_n| \leq \delta\}) + \mathbb{P}(|X - X_n| > \delta) \\
&\leq \mathbb{P}(X \leq t + \delta) + \mathbb{P}(|X - X_n| > \delta) = F(t+\delta) + \mathbb{P}(|X - X_n| > \delta), \\
F_n(t) &\geq \mathbb{P}(\{X_n \leq t\} \cap \{|X - X_n| \leq \delta\}) \geq F(t - \delta).
\end{aligned}$$

On en déduit donc que

$$|F_n(t) - F(t)| \leq |F(t+\delta) - F(t)| \vee |F(t-\delta) - F(t)| + \mathbb{P}(|X - X_n| > \delta).$$

Soit donc $\varepsilon > 0$. Comme F est continue en t, il existe δ_ε tel que $|F(t+\delta_\varepsilon) - F(t)| \vee |F(t-\delta_\varepsilon) - F(t)| \leq \varepsilon/2$. On a alors

$$|F_n(t) - F(t)| \leq \varepsilon/2 + \mathbb{P}(|X - X_n| > \delta_\varepsilon) \leq \varepsilon \qquad \text{pour } n \text{ assez grand.}$$

Donc $X_n \rightsquigarrow_{n\to+\infty} X$. □

FAIRE DESSIN RESUMANT DIFFERENTS MODES DE CV.

1.3 Outils pour la statistique et l'analyse de données

1.3.1 Échantillon de variables i.i.d.

Un cadre standard en statistiques est celui où on considère $X_1, \ldots, X_n$ une suite de v.a.r. indépendantes et de même loi. On abrégera cette situation en *i.i.d.*, pour *indépendantes et identiquements distribués*. On se donne X de même loi que les X_i. La question de base en statistiques est : à quel point la *moyenne empirique*,

$$\bar{X}_n = \sum_{i=1}^{n} X_i$$

est-elle proche de la *moyenne* (tout court) $\mathbb{E}(X)$, si elle existe. Formellement parlant, $\bar{X}_n$ est une variable aléatoire, et $\mathbb{E}(X)$ un réel. Intuitivement, si vous tirez à pile ou face n fois, vous vous attendez à ce que la proportion de piles obtenus converge vers $1/2$ lorsque n augmente, pour une certaine notion de convergence. C'est tout l'objet des résultats qui vont suivre : essayer de quantifier l'écart de $\bar{X}_n$ à $\mathbb{E}(X)$.

On commence par regarder quelques propriétés d'une somme de variables indépendantes.

Proposition 1.39

Soient $X_1, \ldots, X_n$ des variables aléatoires réelles. On note $S_n = \sum_{i=1}^n X_i$.

1. *Si, pour tout $1 \leq i \leq n$, $\mathbb{E}(X_i)$ existe, on a*

$$\mathbb{E}(S_n) = \sum_{i=1}^{n} \mathbb{E}(X_i).$$

2. *Si les $(X_i)_{i=1,\ldots,n}$ sont **indépendantes**, et vérifient $\mathbb{E}(X_i^2) < +\infty$ pour tout $i \in [\![1, n]\!]$, alors*

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

3. *Si les $(X_i)_{i=1,\ldots,n}$ sont **indépendantes**, alors*

$$\forall t \in \mathbb{R} \quad \phi_{S_n}(t) = \prod_{i=1}^{n} \phi_{X_i}(t).$$

Preuve de la Proposition 1.39. Pour la partie espérance, cela découle directement de la linéarité (Proposition 1.10), par récurrence éventuellement.

Pour la partie variance, comme $S_n^2/n^2 \leq \sum_{i=1}^n X_i^2/n$ (convexité de la fonction $x \mapsto x^2$), on en déduit que $\mathbb{E}(S_n^2) < +\infty$, et S_n admet donc une variance. On calcule alors

$$\begin{aligned}
\mathrm{Var}(S_n) &= \mathbb{E}\left(\sum_{i=1}^n (X_i - \mathbb{E}(X_i))\right)^2 \\
&= \sum_{i=1}^n \mathbb{E}((X_i - \mathbb{E}(X_i))^2) + \sum_{i \neq j} \mathbb{E}\left[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))\right] \qquad \text{(linéarité de l'espérance)} \\
&= \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i \neq j} \mathbb{E}[(X_i - \mathbb{E}(X_i))]\mathbb{E}[(X_j - \mathbb{E}(X_j))] \qquad \text{(indépendance)} \\
&= \sum_{i=1}^n \mathrm{Var}(X_i).
\end{aligned}$$

Enfin, pour la partie fonction caractéristique, on a, pour $t \in \mathbb{R}$,

$$\begin{aligned}
\phi_{S_n}(t) &= \mathbb{E}\left[e^{it\sum_{j=1}^n X_j}\right] \\
&= \mathbb{E}\left[\prod_{j=1}^n e^{itX_j}\right] \\
&= \prod_{j=1}^n \mathbb{E}(e^{itX_j}) \qquad \text{(indépendance)} \\
&= \prod_{j=1}^n \phi_{X_j}(t).
\end{aligned}$$

□

Revenons maintenant à notre cadre de variables i.i.d. pour lesquelles on aimerait bien quantifier les fluctuations de la moyenne empirique autour de la vraie moyenne. On a un corollaire immédiat pour ce cas.

Corollaire 1.40

Soient $X_1, \ldots, X_n$ des v.a.r. i.i.d., de loi commune X.

- *Si $\mathbb{E}(|X|) < +\infty$ (X admet une espérance), alors $\mathbb{E}(\bar{X}_n) = \mathbb{E}(X)$.*
- *Si $\mathbb{E}(X^2) < +\infty$ (X admet une variance), alors*

$$\mathrm{Var}(\bar{X}_n) = \frac{\mathrm{Var}(X)}{n}.$$

- *Pour tout $t \in \mathbb{R}$,*

$$\phi_{\bar{X}_n}(t) = \phi_X(t/n)^n.$$

Preuve du Corollaire 1.40. Avec les notations de la Proposition 1.39, on écrit $\bar{X}_n = \frac{S_n}{n}$.

Pour l'espérance, la linéarité donne

$$\mathbb{E}(\bar{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \mathbb{E}(X) \qquad (X_i \sim X \text{ pour tout } i).$$

Pour la variance, on a que $\mathrm{Var}(S_n/n) = \frac{1}{n^2}\mathrm{Var}(S_n)$, et la deuxième partie de la Proposition 1.39 donne

$$\mathrm{Var}(S_n) = \sum_{i=1}^{n}\mathrm{Var}(X_i) = n\mathrm{Var}(X) \qquad (X_i \sim X \text{ pour tout } i).$$

Enfin, pour la fonction caractéristique, il suffit de remarquer que $\phi_{\bar{X}_n}(t) = \mathbb{E}(e^{i(t/n)S_n}) = \phi_{S_n}(t/n)$, et d'appliquer la dernière partie de la Proposition 1.39. □

On a alors que dans le cas i.i.d., la moyenne empirique $\bar{X}_n$ est une variable aléatoire de même espérance que celle de X, mais de variance $\mathrm{Var}(X)/n$. Ses fluctuations autour de $\mathbb{E}(X)$ décroissent donc avec n, de manière quantifiable.

1.3.2 Fluctuations autour de la moyenne : cadre non asymptotique

On va essayer de quantifier un peu cette notion de fluctuation de $\bar{X}_n$ autour de $\mathbb{E}(X)$, en utilisant les outils de la section précédente (avec quelques autres). Le cadre *non-asymptotique* s'entend comme *pas à la limite* (ce sera l'objet de la section d'après). Ici n est donc fixe, grand, mais ne tend pas vers $+\infty$.

Pour transformer les bornes sur la variances du Corollaire 1.40 en bornes sur les fluctuations de $\bar{X}_n$ autour de $\mathbb{E}(X)$, on aura besoin *d'inégalités de concentration*. On en verra deux (les plus simples).

Proposition 1.41 : Inégalité de Markov

Soit Y une variable aléatoire réelle ***positive****. On a alors, pour tout $t > 0$,*

$$\mathbb{P}(Y \geq t) \leq \frac{\mathbb{E}(Y\mathbb{1}_{Y\geq t})}{t} \leq \frac{\mathbb{E}(Y)}{t}.$$

Démonstration. Soit $t > 0$, si $\mathbb{E}(Y\mathbb{1}_{Y\geq t}) = +\infty$, alors l'inégalité est triviale. On suppose donc que $\mathbb{E}(Y\mathbb{1}_{Y\geq t}) < +\infty$, ce qui garantit au passage que

$$\mathbb{E}(Y) = \mathbb{E}(Y\mathbb{1}_{Y<t}) + \mathbb{E}(Y\mathbb{1}_{Y\geq t}) \leq t + \mathbb{E}(Y\mathbb{1}_{Y\geq t}) < +\infty.$$

Il suffit alors de remarquer que $t\mathbb{1}_{Y\geq t} \leq Y\mathbb{1}_{Y\geq t}$, ce qui en intégrant garantit que

$$t\mathbb{P}(Y \geq t) \leq \mathbb{E}(Y\mathbb{1}_{Y\geq t}),$$

et on a la première inégalité. Pour la seconde, on remarque que, comme Y est positive, $Y\mathbb{1}_{Y\geq t} \leq Y$, et donc que $\mathbb{E}(Y\mathbb{1}_{Y\geq t}) \leq \mathbb{E}(Y)$. □

Cette inégalité de Markov est l'inégalité de base pour tout ce qui concerne la concentration de moyennes empiriques autour de l'espérance. L'astuce consiste souvent à prendre pour Y une fonction bien choisie de la variable qui nous intéresse X.

Dans un cadre de v.a.r. i.i.d., on peut déjà en déduire un résultat sur les fluctuations de $\bar{X}_n$ autour de $\mathbb{E}(X)$.

Corollaire 1.42 : Fluctuations au moins "constantes"

Soient $X_1, \dots, X_n$ des v.a.r. i.i.d., de loi commune X, avec $\mathbb{E}(|X|) < +\infty$. On a alors, pour tout $t > 0$,

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}(X)\right| \geq t\right) \leq \frac{\mathbb{E}(|X - \mathbb{E}(X)|)}{t}.$$

Démonstration. En notant $Y = \left|\bar{X}_n - \mathbb{E}(X)\right|$ (qui est bien positive), pour appliquer l'inégalité de Markov, il s'agit de contrôler $\mathbb{E}(Y)$. On a

$$\begin{aligned}
\mathbb{E}(Y) &= \mathbb{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i - \mathbb{E}(X_i))\right|\right) && (X_i \sim X \text{ pour tout } i)\\
&\leq \mathbb{E}\left(\sum_{i=1}^{n}\frac{|X_i - \mathbb{E}(X_i)|}{n}\right) && \text{(inégalité triangulaire)}\\
&= \frac{n\mathbb{E}(|X - \mathbb{E}(X)|}{n} && \text{(linéarité de l'espérance).}
\end{aligned}$$

□

Application : Jeu de pile ou face.

Si $X_1, \dots, X_n$ sont i.i.d. $\mathcal{B}(1/2)$ (n tirage à pile ou face indépendants équilibré), on a

$$\mathbb{E}(|X - \mathbb{E}(X)|) = \mathbb{E}(|X - (1/2)|) = 1/2.$$

On en déduit alors que, pour tout $t > 0$, $\mathbb{P}(|\bar{X}_n - (1/2)| \geq t) \leq 1/(2t)$. Par exemple, la probabilité qu'il y ait plus de 3/4 de pile ou de face dans la série ($t = 1/4$) est majorée par 2, ce qui est stupide parce qu'étant une probabilité elle est forcément majorée par 1. On comprend alors les limitations de cette inégalité : la borne donnée par ce Corollaire est la même pour tout n, alors qu'on s'attendrait plutôt à une fonction décroissante en n en suivant l'intuition d'une concentration de plus en plus forte autour de $\mathbb{E}(X)$ au fur et à mesure que le nombre de tirages augmente.

Ici tout provient du fait que on ne peut pas espérer majorer $\mathbb{E}(|\bar{X}_n - \mathbb{E}(X)|)$ par autre chose que $\mathbb{E}(|X - \mathbb{E}(X)|)$ si on ne rajoute pas d'hypothèses sur X :

Exemple 1.43 *: Loi de Cauchy.* Si X suit une loi de Cauchy standard sur $\mathbb{R}$, de densité

$$f(x) = \frac{1}{\pi(1 + x^2)}.$$

En admettant que $\phi_X(t) = e^{-|t|}$ (cf un cours d'analyse complexe), on a, pour tout $t \in \mathbb{R}$,

$$\phi_{\bar{X}_n}(t) = (\phi_X(t/n))^n = \left(e^{-|t|/n}\right)^n = \phi_X(t).$$

On en déduit alors que $\bar{X}_n \sim X$ suit donc une loi de Cauchy standard, et ne se concentre pas mieux autour de 0 que X_1 toute seule.

Dans ce cas précis X n'admet même pas d'espérance, ce n'est donc pas un cas où on peut écrire $\mathbb{E}(|\bar{X}_n - \mathbb{E}(X)|) = \mathbb{E}(|X - \mathbb{E}(X)|)$, cela dit l'intuition générale est la bonne : pour que la moyenne empirique se concentre autour de la vraie moyenne lorsque n grandit, il faut généralement des conditions supplémentaires de type "moments" ($\mathbb{E}(|X|^p) < +\infty$, pour $p \geq 1$), voire moments exponentiels. La condition $\mathbb{E}(|X|) < +\infty$ toute seule sera certes suffisante d'un point de vue asymptotique (cf plus loin), mais on ne pourra pas en déduire une borne non asymptotique sur les fluctuations.

Une condition simple à rajouter est alors $\mathbb{E}(X^2) < +\infty$ (existence d'une variance). Dans ce cas, l'inégalité de Bienaymé-Cebicev permet de facilement borner les fluctuations autour de l'espérance.

Proposition 1.44 : Inégalité BT (Bienaymé-Cebicev)

Soit X une v.a.r. telle que $\mathbb{E}(X^2) < +\infty$. Alors, pour tout $t > 0$,

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq t\right) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

Démonstration. La preuve repose encore sur une inégalité de Markov. Soit $t > 0$, on a alors

$$\mathbb{P}\left(|X - \mathbb{E}(X)| \geq t\right) = \mathbb{P}\left(|X - \mathbb{E}(X)|^2 \geq t^2\right) \leq \frac{\mathbb{E}((X - \mathbb{E}(X))^2)}{t^2} = \frac{\mathrm{Var}(X)}{t^2},$$

en appliquant l'inégalité de Markov à $Y = (X - \mathbb{E}(X))^2$. □

On remarque alors qu'on peut généraliser l'inégalité de Bienaymé Cebicev si on a des conditions de moment d'ordre supérieur : si $\mathbb{E}(|X - \mathbb{E}(X)|^p) = M_p < +\infty$, on aura par les mêmes arguments $\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq M_p/t^p$. L'usage particulier de Bienaymé Cebicev est lié au fait que dans le cadre de variables indépendantes, $\mathrm{Var}(\bar{X}_n)$ se calcule bien (là où majorer $\mathbb{E}(|\bar{X}_n - \mathbb{E}(X)|^p)$ peut s'avérer plus compliqué).

Corollaire 1.45 : Fluctuations en $1/\sqrt{n}$

Soient $X_1, \ldots, X_n$ des v.a.r. i.i.d., de loi commune X, avec $\mathbb{E}(|X|^2) < +\infty$. On a alors, pour tout $t > 0$,

$$\mathbb{P}\left(\left|\bar{X}_n - \mathbb{E}(X)\right| \geq t\right) \leq \frac{\mathrm{Var}(X)}{nt^2}.$$

Démonstration. On applique l'inégalité de Bienaymé Cebicev à $\bar{X}_n$, en utilisant le Corollaire 1.40 qui donne $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X)/n$. □

Cette inégalité permet alors de montrer que $\bar{X}_n$ converge en probabilité vers $\mathbb{E}(X)$, et permet même de quantifier les fluctuations de $\bar{X}_n$: si X admet une variance, les fluctuations autour de $\mathbb{E}(X)$ seront de l'ordre au pire $1/\sqrt{n}$ (prendre $t = t'/\sqrt{n}$). Le Théorème central limite va montrer que cet ordre de grandeur est le bon (les fluctuations seront exactement d'ordre $1/\sqrt{n}$ d'un point de vue asymptotique).

Application : Jeu de pile ou face, suite.

Dans l'exemple précédent de pile ou face, on a, pour tout $t > 0$,

$$\mathbb{P}(|\bar{X}_n - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{Var}(X)}{nt^2} = \frac{1}{4nt^2}.$$

Cela assure par exemple qu'après n lancers, la probabilité d'avoir plus de $3/4$ de pile ou de face est plus petite que $4/n$, par exemple 4% après 100 lancers.

Comme dit plus haut, on peut avoir des bornes de concentration plus précises (en t essentiellement) si X a des moments d'ordre supérieurs, des moments exponentiels, ou est bornée. Dans le cas du pile ou face par exemple, l'inégalité de Hoeffding permet d'affiner un peu (hors programme).

Dans le cadre de ce cours, le seul cas de figure où vous pourrez trouvez mieux que BT pour donner une borne non asymptotique sur les fluctuations est celui où vous **connaissez** la loi de $\bar{X}_n$ (ou S_n).

Exemple 1.46 *: Deux classiques.*
Lois normales : Soient $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{N}(\mu, 1)$. Alors, pour tout $t \in \mathbb{R}$,

$$\phi_{\bar{X}_n}(t) = \phi_X(t/n)^n = (e^{it\mu/n - t^2/(2n)})^n = e^{it\mu - (1/n)t^2/2}.$$

On en déduit alors que $\bar{X}_n \sim \mathcal{N}(\mu, 1/n)$, on encore que $\sqrt{n}(\bar{X}_n - \mu) \sim \mathcal{N}(0, 1)$. Si N désigne une loi $\mathcal{N}(0, 1)$ standard, on peut alors écrire

$$\mathbb{P}\left(\sqrt{n}|\bar{X}_n - \mu| \geq t\right) = \mathbb{P}(|N| \geq t) = F_N(-t) + (1 - F_N(t)) = 2F_N(-t).$$

Or, pour les lois dites "standard", on connaît (ou on sait approcher) les fonctions de répartition ou les fonctions quantiles. On a alors une **égalité**

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq t\right) = 2F_N(-\sqrt{n}t),$$

que l'on peut utiliser. Par exemple, on sait que $F_N(-t) \leq e^{-t^2/2}$, cela donne alors

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq t\right) \leq 2e^{-nt^2/2},$$

là où une application directe de BT aurait donné

$$\mathbb{P}\left(|\bar{X}_n - \mu| \geq t\right) \leq \frac{1}{nt^2},$$

qui est moins précise lorsque $n \to +\infty$.

Pile ou face (binomiale) : Dans le cas où $X_1, \ldots, X_n$ sont i.i.d. de loi $\mathcal{B}(1/2)$, on a que $S_n = \sum_{i=1}^n X_i \sim \mathcal{B}(n, 1/2)$, loi binomiale de paramètres $(n, 1/2)$ qui est une loi connue. Soit B une variable avec une telle loi. On a alors, pour tout $t > 0$,

$$\begin{aligned}\mathbb{P}(|\bar{X}_n - (1/2)| \geq t) &= \mathbb{P}(B \geq n/2 + nt) + \mathbb{P}(B \leq n/2 - nt) \\ &= F_B(n/2 - nt) + 1 - F_B(\lceil n/2 + nt \rceil - 1).\end{aligned}$$

On peut alors se référer aux tabulations de la loi B pour là encore avoir une **égalité**. Pour aller plus loin et en déduire une majoration plus fine que celle donnée par BT, il faudrait ici avoir des encadrements de F_B, ce qui est possible, mais un peu calculatoire.

1.3.3 Fluctuations autour de la moyenne : point de vue asymptotique

Ici on regarde ce qui se passe lorsque $n \to +\infty$. A minima, on espère que $\bar{X}_n$ converge vers $\mathbb{E}(X)$, dans un sens à définir. C'est l'objet de la loi des grands nombres.

Théorème 1.47 : Loi(s) des grands nombres

Soient $X_1, \dots, X_n$ une suite de v.a.r. i.i.d., de loi commune X vérifiant $\mathbb{E}(|X|) < +\infty$. Alors

- *(loi faible des grands nombres) $\bar{X}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \mathbb{E}(X)$.*
- *(loi forte des grands nombres) $\bar{X}_n \xrightarrow[n\to+\infty]{p.s.} \mathbb{E}(X)$.*

On remarque que la version forte implique la version faible. Pour le volet statistiques, la version faible suffira la plupart du temps, et la preuve est relativement simple.

Preuve du Théorème 1.47. On prouve la version faible (la version forte nécessitera d'autres outils que vous verrez peut-être plus tard).

On commence par remarquer que si $\mathbb{E}(X^2) < +\infty$, alors, pour tout $\varepsilon > 0$, l'inégalité de Bienaymé Cebicev donne

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}(X)| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(X)}{n\varepsilon^2} \xrightarrow[n\to+\infty]{} 0.$$

Passons maintenant au cas général, et, quitte à poser $Y_i = X_i - \mathbb{E}(X)$, supposons que les X_i sont centrés. Soit $\varepsilon > 0$. Comme $\mathbb{E}(|X|) < +\infty$, il existe M_0 tel que, pour tout $M \geq M_0$,

$$\mathbb{E}(|X|\mathbb{1}_{|X|>M}) \leq \frac{\varepsilon}{4}.$$

On pose alors, pour un $M \geq M_0$, $X_i^M = X_i \mathbb{1}_{|X_i|\leq M}$, de telle sorte que $\mathbb{E}((X_i^M)^2 \leq M^2 < +\infty$ pour pouvoir utiliser le résultat précédent. On remarque aussi que

$$\left|\mathbb{E}(X) - \mathbb{E}(X^M)\right| \leq E(|X|\mathbb{1}_{|X|>M}) \leq \frac{\varepsilon}{3},$$

où on rappelle qu'on a pris $\mathbb{E}(X) = 0$. On décompose alors

$$\frac{1}{n}\sum_{i=1}^n X_i = \frac{1}{n}\left(\sum_{i=1}^n (X_i^M - \mathbb{E}(X_i^M))\right) + \frac{1}{n}\left(\sum_{i=1}^n (X_i - X_i^M)\right) + \mathbb{E}(X^M),$$

On a alors

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - X_i^M)\right| \geq \varepsilon/3\right) + \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i^M - \mathbb{E}(X_i^M))\right| \geq \varepsilon/3\right).$$

Pour le premier terme, a

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n (X_i - X_i^M)\right| \geq \varepsilon/3\right) \leq \frac{3\mathbb{E}(|X|\mathbb{1}_{X>M})}{\varepsilon},$$

en utilisant l'inégalité de Markov. Pour le deuxième terme, on a

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i^M - \mathbb{E}(X_i^M))\right| \geq \varepsilon/3\right) \underset{n\to+\infty}{\longrightarrow} 0.$$

Soit alors $\delta > 0$. On se donne $M_\delta \geq M_0$ tel que $\frac{3\mathbb{E}(|X|\mathbb{1}_{X>M_\delta})}{\varepsilon} \leq \delta/2$, et on a alors que pour n assez grand

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X_i^{M_\delta} - \mathbb{E}(X_i^{M_\delta}))\right| \geq \varepsilon/3\right) \leq \delta/2.$$

Cela donne, pour n assez grand

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right| \geq \varepsilon\right) \leq \delta,$$

d'où le résultat. □

Ce résultat est asymptotique par nature : sous la condition $\mathbb{E}(|X|) < +\infty$, la convergence de $\bar{X}_n$ vers $\mathbb{E}(X)$ est assurée, en revanche cette convergence peut être arbitrairement lente (par exemple on peut avoir $\mathbb{P}(|\bar{X}_n - \mathbb{E}(X)| \geq t) \sim \log(n)^{-1}$). Pour avoir une idée des fluctuations de $\bar{X}_n$ autour de $\mathbb{E}(X)$, on ne peut donc s'en contenter, et il faut la plupart du temps demander plus de conditions sur X, notamment de type moments ($\mathbb{E}(X^2) < +\infty$ par exemple), comme dans le cadre non asymptotique.

La condition d'existence de variance donne, par Bienaymé Cebicev,

$$\mathbb{P}(\sqrt{n}|\bar{X}_n - \mathbb{E}(X)| \geq t) \leq \mathrm{Var}(X)/t^2,$$

ce dont on peut déduire que la variable $\sqrt{n}(\bar{X}_n - \mathbb{E}(X))$ a des fluctuations autour de 0 contrôlées (cette variable ne peut donc pas diverger vers un infini avec proba positive). C'est le sens de "fluctuations de $\bar{X}_n$ autour de $\mathbb{E}(X)$ en $1/\sqrt{n}$. A priori rien n'empêche que

$$\mathbb{P}(\sqrt{n}|\bar{X}_n - \mathbb{E}(X)| \geq t) \underset{n\to+\infty}{\longrightarrow} 0,$$

c'est à dire que les fluctuations de $\bar{X}_n$ autour de $\mathbb{E}(X)$ soit d'ordre plus petit que $1/\sqrt{n}$. Le résultat fondamental suivant montrer que, dès lors que l'on a une variance, $1\sqrt{n}$ est exactement le bon ordre de ces fluctuations.

Théorème 1.48 : Théorème Central Limite (ou de la limite centrale)

Soient $X_1, \ldots, X_n$ une suite de v.a.r. i.i.d., de loi commune X vérifiant $\mathbb{E}(X^2) < +\infty$. Alors

$$\sqrt{n}(\bar{X}_n - \mathbb{E}(X)) \rightsquigarrow \mathcal{N}(0, \mathrm{Var}(X)).$$

Ce théorème fondamental fait un peu plus que donner le bon ordre des fluctuations autour de la moyenne : il caractérise aussi la loi limite de ces fluctuations (loi Normale). Cela justifie le rôle particulier joué par la loi Normale : celui d'attracteur universel pour les fluctuations autour de la moyenne, si tant est que l'on ait des moments d'ordre 2. C'est une des raisons pour lesquelles un accent particulier est mis sur le cadre Gaussien dans les cours de Statistiques.

Preuve du Théorème 1.48. On aura besoin du résultat technique et donc admis suivant :

LEMME 1.49

Soit X une v.a.r. et $k \geq 1$. Si $\mathbb{E}(|X|^k) < +\infty$, alors ϕ_X est C^k, et

$$\phi^{(k)}(0) = i^k \mathbb{E}(X^k).$$

C'est une interversion dérivation/intégrale standard, la preuve est donc renvoyée au cours d'analyse et intégration.

Dans notre cas on a que $\mathbb{E}(X^2)$ est finie, et donc que ϕ_X est C^2. Pour prouver la convergence en loi de $\sqrt{n}(\bar{X}_n - \mathbb{E}(X))$, posons $Y = X - \mathbb{E}(X)$ (ϕ_Y reste C^2), et

$$Z_n = \sqrt{n}\bar{Y}_n,$$

et regardons sa fonction caractéristique. Pour $t \in \mathbb{R}$, on a

$$\begin{aligned}
\phi_{Z_n}(t) &= \mathbb{E}(e^{it\sqrt{n}(\bar{Y}_n)}) \\
&= \mathbb{E}\left(e^{it/\sqrt{n}(\sum_{j=1}^n Y_j)}\right) \\
&= \phi_Y(t/\sqrt{n})^n.
\end{aligned}$$

Comme ϕ_Y est C^2, on peut écrire

$$\phi_Y(t/\sqrt{n}) = \phi_Y(0) + \frac{t}{\sqrt{n}}\phi_Y'(0) + \frac{t^2}{2n}\phi_Y''(0) + o(1/n),$$

avec $\phi_Y(0) = 1$, $\phi_Y'(0) = i\mathbb{E}(Y) = 0$ et $\phi_Y''(0) = -\mathbb{E}(Y^2) = -\mathrm{Var}(X)$, d'après le Lemme 1.49. On en déduit alors que

$$\begin{aligned}
\phi_{Z_n}(t) &= \left(1 - \frac{t^2\mathrm{Var}(X)}{2n} + o(1/n)\right)^n \\
&= \exp\left(n\log\left(1 - \frac{t^2\mathrm{Var}(X)}{2n} + o(1/n)\right)\right) \\
&= \exp\left(-\frac{t^2\mathrm{Var}(X)}{2} + o(1)\right) \\
&\underset{n\to+\infty}{\longrightarrow} e^{-t^2\mathrm{Var}(X)/2} = \phi_{\mathcal{N}(0,\mathrm{Var}(X))}(t),
\end{aligned}$$

ce qui prouve $Z_n \rightsquigarrow \mathcal{N}(0, \mathrm{Var}(X))$.

□

Exemple 1.50 *: Pile ou face toujours.* Dans notre expérience de pile ou face i.i.d. où $X_i \sim \mathcal{B}(1/2)$ (1 vaut Pile), pour un t quelconque, on a

$$\begin{aligned}
\mathbb{P}(\sqrt{n}|\bar{X}_n - \frac{1}{2}| \leq t) &\underset{n\to+\infty}{\longrightarrow} \mathbb{P}(|\mathcal{N}(0,1/4)| \leq t) \\
&= \mathbb{P}\left(|\mathcal{N}(0,1)| \leq 2t\right) \\
&= 1 - 2F(-2t) = 2F(2t) - 1,
\end{aligned}$$

où F est la fonction de répartition d'une loi normale standard (connue). FAIRE DESSIN. En particulier, ce résultat stipule que la probabilité que la proportion de pile obtenus soit comprise entre $1/2 - t/\sqrt{n}$ et $1/2 + t\sqrt{n}$ converge vers $2F(2t) - 1$.

Si on reprend la question initiale (probabilité que la proportion de pile soit comprise entre 1/4 et 3/4), il faut prendre $t/\sqrt{n} = 1/4$, c'est à dire un t_n qui bouge avec n, et le Théorème central limite seul ne peut vous aider (rien ne garantit a priori que $F_{Z_n}(\sqrt{n}/4)$ converge, où $Z_n = \sqrt{n/\mathrm{Var}(X)}(\bar{X}_n - \mathbb{E}(X))$).

Comme expliqué au dessus, dans certains cas on peut avoir besoin de relier $F_{Z_n}(t_n)$ à $F(t_n)$ (où F est la fonction de répartition d'une loi Gaussienne standard et $Z_n = \sqrt{n/\mathrm{Var}(X)}(\bar{X}_n - \mathbb{E}(X))$), ce que ne permet pas de faire le Théorème central limite seul (il donne la convergence à t fixé). On peut montrer que ça marche encore : de fait, $\|F_{Z_n} - F\|_\infty$ converge bien vers 0 sous les mêmes hypothèses.

Proposition 1.51 : Théorème Central Limite "Uniforme"

Soient $X_1, \ldots, X_n$ une suite de v.a.r. i.i.d., de loi commune X vérifiant $\mathbb{E}(X^2) < +\infty$. Alors

$$\|F_{Z_n} - F\|_\infty \underset{n\to+\infty}{\longrightarrow} 0,$$

avec $Z_n = \sqrt{n/\mathrm{Var}(X)}(\bar{X}_n - \mathbb{E}(X))$, et F est la fonction de répartition d'une loi $\mathcal{N}(0,1)$.

Démonstration. Pour ceux qui ont déjà vu ça, c'est une conséquence du Théorème de Dini (passage d'une convergence simple à une convergence uniforme). On peut toutefois le refaire "à la main" rapidement.

Soit $\varepsilon > 0$. Il existe alors T tel que $F(-T) \leq \varepsilon$. Soit alors $t \in]-\infty, -T[$, on a que

$$|F_{Z_n}(t) - F_Z(t)| \leq F_{Z_n}(T) \vee F_Z(T) \leq F_Z(T) + |F(T) - F_{Z_n}(T)| \leq 2\varepsilon,$$

si $n \geq n_T$. Pareillement, pour $t \in]T, +\infty[$,

$$|F_{Z_n}(t) - F_Z(t)| \leq 1 - F_Z(T) \vee 1 - F_{Z_n}(T) \leq 1 - F_Z(T) + |F(T) - F_{Z_n}(T)| \leq 2\varepsilon,$$

pour $n \geq n_T$. On a donc, pour $n \geq n_T$,

$$\sup_{[-T,T]^c} |F_{Z_n}(t) - F(t)| \leq 2\varepsilon.$$

Regardons maintenant la partie sur $[-T, T]$. F étant continue, elle est uniformément continue sur $[-T, T]$ par le Théorème de Heine Borel. On peut alors découper $[-T, T]$ en $t_0 = -T < t_1 < \ldots < t_N = T$ de sorte que, si $|i - j| \leq 1$, $|F(t_i) - F(t_j)| \leq \varepsilon$. On va alors montrer qu'on peut se ramener à de la convergence simple sur ces t_i.

On se donne alors n_ε tel que pour tout $n \geq n_\varepsilon$, pour tout $j = 1, \ldots, N$, $|F_{Z_n}(t_j) - F(t_j)| \leq \varepsilon$. Soit alors $n \geq n_\varepsilon$, et $t \in [-T, T]$. Il existe $j \in [\![1, N]\!]$ tel que $t \in [t_{j-1}, t_j]$, et on peut écrire

$$\begin{aligned}
|F_{Z_n}(t) - F(t)| &\leq (F_{Z_n}(t_j) - F(t_{j-1})) \vee (F(t_j) - F_{Z_n}(t_{j-1})) \\
&\leq |F_{Z_n}(t_{j-1}) - F(t_{j-1})| \vee |F_{Z_n}(t_j) - F(t_j)| + |F(t_{j-1}) - F(t_j)| \\
&\leq 2\varepsilon.
\end{aligned}$$

On en déduit alors que, pour tout $n \geq n_T \vee n_\varepsilon$,

$$\|F_{Z_n} - F\|_\infty \leq 2\varepsilon,$$

ce qui permet de conclure.. □

Exemple 1.52 *: Pile ou face encore.* Si on revient à notre question initiale qui est de majorer la probabilité d'avoir une proportion de pile ou de face plus grande que 3/4, on réécrit cette probabilité sous la forme

$$\begin{aligned}
\mathbb{P}\left(|\bar{X}_n - 1/2| \geq 1/4\right) &= \mathbb{P}\left(|2\sqrt{n}(\bar{X}_n - 1/2)| \geq \sqrt{n}/2\right) \\
&= F_{Z_n}(-\sqrt{n}/2) + (1 - F_{Z_n}(\sqrt{n}/2)) \\
&\leq 2F(-\sqrt{n}/2) + 2\|F - F_{Z_n}\|_\infty \\
&\leq 2e^{-n/2} + 2\|F - F_{Z_n}\|_\infty,
\end{aligned}$$

ce qui garantit qu'elle converge vers 0 (ce qu'on savait déjà en utilisant Bienaymé Cebicev).

Cet exemple illustre bien un problème inhérent aux résultats asymptotiques : pour en tirer une majoration de probabilité de déviation non asymptotique on est obligés d'avoir une idée de la vitesse de la converge vers la loi limite, c'est à dire une borne sur $\|F - F_{Z_n}\|_\infty$. Comme pour la loi des grands nombres, cette vitesse peut être arbitrairement lente sous la seule condition $\mathbb{E}(X^2) < +\infty$. Si on rajoute encore des conditions (par exemple $\mathbb{E}(|X|^3) < +\infty$), on peut avoir des résultats sur ces vitesses (par exemple le Théorème de Berry-Esseen). On conclut sur une remarque d'ordre général : la plupart du temps, passer par un résultat asymptotique pour obtenir une garantie non asymptotique ne mène pas aux résultats les plus précis, il vaut souvent mieux utiliser des inégalités de concentration directement.

Un dernier outil utile dans un cadre asymptotique est le Lemme de Slutsky, qui permet dans un cas précis de considérer la convergence en loi comme une "vraie convergence". Plus précisément, on aimerait bien avoir un résultat du type "si $X_n \rightsquigarrow X$ et $Y_n \rightsquigarrow Y$, alors $X_n + Y_n \rightsquigarrow X + Y$, $X_n Y_n \rightsquigarrow XY$.

En toute généralité, ce genre de résultat est conceptuellement douteux : en effet, d'hypothèses sur la loi de X_n seule et Y_n seule on pourrait déduire quelque chose sur la loi de $X_n + Y_n$, qui fait aussi intervenir la **dépendance** entre X_n et Y_n.

Exemple 1.53 *: Contre exemple.* Soit $N \sim \mathcal{N}(0,1)$. On pose $X_n = N$ pour tout n, et $Y_n = -N$. Supposons que l'assertion "$X_n \rightsquigarrow X$ et $Y_n \rightsquigarrow Y$ implique $X_n + Y_n \rightsquigarrow X + Y$" soit vraie.

Alors on a $X_n \rightsquigarrow N$, et $Y_n \rightsquigarrow -N$, et donc $X_n + Y_n \rightsquigarrow 0$. D'un autre côté, on a aussi $Y_n \rightsquigarrow N$ (N et $-N$ ont même loi). Et donc $X_n + Y_n \rightsquigarrow 2N$. On aurait alors $2N \sim \mathcal{N}(0,4) \sim 0$, ce qui est impossible.

L'exemple précédent montre que pour statuer sur la convergence de $X_n + Y_n$, il s'agit de regarder la loi du **couple** (X_n, Y_n), c'est à dire les lois de X_n, Y_n, mais aussi la structure de dépendance entre les deux. Dans l'exemple précédent, on a $(X_n, Y_n) \rightsquigarrow (N, -N)$, et dans ce cas $X_n + Y_n \rightsquigarrow 0$ est juste : $(x, y) \mapsto x + y$ étant une fonction continue, la convergence en loi se transmet.

Comme vu plus haut, $X_n \rightsquigarrow X$ et $Y_n \rightsquigarrow Y$ ne suffisent pas à garantir que $(X_n, Y_n) \rightsquigarrow (X, Y)$: il manque la structure de dépendance. Si cette dépendance est spécifiée, on peut s'en sortir.

Théorème 1.54 : Lemme de Slutsky

Soit (X_n, Y_n) une suite de couples de v.a.r.. Si $X_n \rightsquigarrow X$ et $Y_n \rightsquigarrow c$, où c désigne la v.a.r. constante valant c, alors

1. $Y_n \xrightarrow[n\to+\infty]{\mathbb{P}} c$,
2. $(X_n, Y_n) \rightsquigarrow (X, c)$.

En particulier, $X_n + Y_n \rightsquigarrow X + c$, $X_n Y_n \rightsquigarrow cX$.

Démonstration. Soit F_c la fonction de répartition de la variable constante égale à c. On a alors

$$F_c(t) = \mathbb{1}_{t \geq c},$$

qui a donc pour ensemble de continuité $\mathbb{R} \setminus \{c\}$. Soit $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}(|Y_n - c| \geq \varepsilon) &= \mathbb{P}(Y_n \leq c - \varepsilon) + \mathbb{P}(Y_n \geq c + \varepsilon) \\ &\leq F_n(c - \varepsilon) + \mathbb{P}(Y_n > c + \varepsilon/2) \\ &= F_n(c - \varepsilon) + 1 - F_n(c + \varepsilon/2),\end{aligned}$$

où F_n désigne la fonction de répartition de Y_n. Comme $c - \varepsilon$ et $c + \varepsilon/2$ sont des points de continuités de F_c, on a

$$\begin{aligned}F_n(c - \varepsilon) &\xrightarrow[n\to+\infty]{} F(c - \varepsilon) = 0 \\ F_n(c + \varepsilon/2) &\xrightarrow[n\to+\infty]{} F(c + \varepsilon/2) = 1,\end{aligned}$$

ce qui suffit à montrer que $Y_n \xrightarrow[n\to+\infty]{\mathbb{P}} c$.

Pour la convergence du couple, on admettra que la caractérisation par la convergence des fonctions de répartitions (multivariées) sur l'ensemble de continuité de la fonction de répartition cible reste valide. Soit (X, Y) un couple de variable aléatoire tel que $Y = c$ presque sûrement. La loi du couple est alors entièrement caractérisée par la loi de X : en effet, si $(x, y) \in \mathbb{R}^2$,

$$F(x, y) := \mathbb{P}(\{X \leq x\} \cap \{Y \leq y\}) = \mathbb{1}_{y \geq c} F_X(x).$$

Dans ce cas précis il n'est pas nécessaire de spécifier la structure de dépendance entre X et Y (de fait elles sont indépendantes).

On en déduit aussi que l'ensemble des points de continuité de F est $D := D_X \times (\mathbb{R} \setminus \{c\})$, où D_X est l'ensemble des points de continuité de F_X. Soit $(x, y) \in D$, et supposons $y > c$. On a

$$\mathbb{P}(\{X_n \leq x\} \cap \{Y_n \leq y\}) = \mathbb{P}(X_n \leq x) - \mathbb{P}(\{X_n \leq x\} \cap \{Y_n > y\}),$$

donc

$$\mathbb{P}(X_n \leq x) - \mathbb{P}(Y_n > y) \leq \mathbb{P}(\{X_n \leq x\} \cap \{Y_n \leq y\}) \leq \mathbb{P}(X_n \leq x).$$

Les termes de gauche et droite convergeant vers $F_X(x)$, on en déduit

$$\mathbb{P}(\{X_n \leq x\} \cap \{Y_n \leq y\}) \xrightarrow[n\to+\infty]{} F_X(x),$$

si $y > c$. Maintenant, pour $y < c$,

$$\begin{aligned}\mathbb{P}(\{X_n \leq x\} \cap \{Y_n \leq y\}) &\leq \mathbb{P}(Y_n \leq y) \\ &\to 0.\end{aligned}$$

On en déduit, pour tout $(x, y) \in D$

$$\mathbb{P}(\{X_n \leq x\} \cap \{Y_n \leq y\}) \underset{n\to+\infty}{\longrightarrow} \mathbb{1}_{y>c} F_X(x) = F(x, y),$$

et donc $(X_n, Y_n) \rightsquigarrow (X, Y)$. Les corollaires se déduisent du petit lemme suivant.

Lemme 1.55 : Convergence en loi transmissible par fonction continue

Si $X_n \rightsquigarrow X$ (où X peut être un vecteur aléatoire), et $g : \mathcal{X} \to \mathcal{Y}$ est une fonction continue, alors

$$g(X_n) \rightsquigarrow g(X).$$

Démonstration. En repartant de la Définition 1.30 qui reste valide en dehors du cadre v.a.r. (c'est la définition générale de la convergence en loi), si $f : \mathcal{Y} \to \mathbb{R}$ est une fonction continue bornée, alors $f \circ g \in C_b(\mathcal{X})$, et donc par la même définition

$$\mathbb{E}(f(g(X_n)) = \mathbb{E}((f \circ g)(X_n)) \underset{n\to+\infty}{\longrightarrow} \mathbb{E}(f \circ g)(X)) = \mathbb{E}(f(g(X))).$$

□

On conclut alors en remarquant que $(x, y) \mapsto x + y$ et $(x, y) \mapsto xy$ sont des fonctions continues. □

Chapitre 2

Le point de vue des statistiques

À gros traits, la différence entre le point de vue probabiliste et le point de vue statistique se résume à ceci :

— le probabiliste prédit mais n'observe pas,

— le statisticien observe et infère.

FAIRE DESSIN

Si on se place du point de vue de la fluctuation de la moyenne empirique $\bar{X}_n$ autour de $\mathbb{E}(X)$:

— le probabiliste connaît $\mathbb{E}(X)$ et prédit qu'avec grosse proba $\bar{X}_n$ va être proche de $\mathbb{E}(X)$.

— le statisticien **ne connaît pas** $\mathbb{E}(X)$, observe $\bar{X}_n$, et en déduit une plage de valeurs plausibles pour $\mathbb{E}(X)$.

Exemple 2.1 *: Terrasse et jours de soleils à Rennes.* Un gérant de brasserie hésite à investir dans une terrasse. Pour cela, il a besoin de savoir combien de jours par an il peut espérer ouvrir en terrasse en 2024, c'est à dire essayer de prévoir le nombre de jours de beau temps.

Pour le volet prédictif, il va faire appel à un probabiliste. Ce probabiliste va modéliser le nombre de jours de beau temps en 2024 par une loi binomiale $\mathcal{B}(n, \theta)$, où $n = 365$, et θ est supposé connu. Si on note N_{2024} le nombre de jours de beau temps à Rennes en 2024 (non observé), le probabiliste peut alors dire, en utilisant l'inégalité de BT, que

$$\mathbb{P}\left(|N_{2024} - n\theta| \geq \varepsilon\right) \leq \frac{n\theta(1-\theta)}{\varepsilon^2},$$

par exemple on va être sûrs à 90% que $N_{2024} \in [n\theta - \sqrt{1/(10n\theta(1-\theta)}, n\theta + \sqrt{1/(10n\theta(1-\theta)}]$. En se basant sur cette prédiction, le gérant va faire ses calculs et décider ou non d'investir dans une terrasse.

Le problème est que dans la plupart des situations le probabiliste **ne connaît pas** θ, et est donc bien embêté pour répondre aux attentes du gérant. Il va alors faire appel à un statisticien pour **estimer** ce θ. Pour ce faire, le statisticien va collecter $(N_j)_{j=2000,\ldots,2023}$, où N_j est le nombre de jours d'ensoleillement **observés** l'année j. Il va ensuite **supposer que les** N_j **sont i.i.d.** $\mathcal{B}(n, \theta)$, avec le même θ que pour l'année 2024. Il va enfin utiliser BT lui aussi, mais dans l'autre sens :

$$\mathbb{P}\left(\left|\frac{\bar{N}}{n} - \theta\right| \geq \varepsilon\right) \leq \frac{\theta(1-\theta)}{n \times 24 \times \varepsilon^2} \leq \frac{1}{4 \times n \times 24 \times \varepsilon^2},$$

où $\bar{N} = \frac{1}{24}\sum_{j=2000}^{2023} \frac{N_j}{n}$ est la proportion moyenne d'ensoleillement observée entre 2000 et 2023 (on appellera ça un **estimateur** de θ). Par exemple le statisticien peut déduire qu'on est sûr à 90% que

$$\theta \in [\bar{N}/n - \sqrt{1/(960n)}, \bar{N}/n + \sqrt{1/(960n)}].$$

On appellera cela un intervalle de confiance. En se basant sur cette estimation de θ, le probabiliste pourra répondre au gérant de brasserie. Bon, souvent probabiliste et statisticien sont la même personne.

2.1 Des statistiques descriptives à l'inférence statistique

On l'aura compris, les statistiques se basent sur des observations $X_1, \dots, X_n$ (par exemple le nombre de jours ensoleillés sur l'année i). Vous avez probablement dû rencontrer dans votre scolarité des **descripteurs** de ces observations, comme

1. la moyenne empirique : $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ qui mesure la valeur moyenne **des observations**.
2. La variance empirique : $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$, qui mesure la dispersion moyenne des observations autour de leur valeur moyenne.
3. D'autres choses empiriques (médiane, quantiles, fonction de répartition).

Calculer ces quantités relève de la **statistique descriptive** (on se contente de décrire ce qu'on a sous les yeux). La **statistique inférentielle** elle se base sur les observations pour essayer de dire des choses sur le processus qui les a généré, par exemple le θ pour les jours d'ensoleillement, pour éventuellement après en déduire des prédictions.

Pour passer de la description pure à l'inférence on a besoin de dire que ce que l'on observe est tiré suivant un processus caché qui a une certaine forme, c'est à dire de *poser un modèle*. En effet, si on reprend l'exemple des jours d'ensoleillement, extrapoler N_{2024} à partir des N_j, $j = 2000, \dots, 2023$ requiert de supposer que tout le monde est généré à partir du même processus (une binomiale $\mathcal{B}(n, \theta)$), de manière indépendante.

Si on ne fait pas cette hypothèse, rien n'empêche de dire que les N_j observés ont été tirés suivant le processus déterministe qui met la masse 1 sur les N_j observés, et on ne peut rien en déduire pour N_{2024}.

De manière moins théorique, si entre 2000 et 2023 vous avez eu une éruption volcanique importante (mettons en l'année 2010, par exemple en Islande), il vous semblerait bizarre de dire que le nombre de jours d'ensoleillement en 2010 suit le comportement "normal" du nombre de jours annuel d'ensoleillement. Si vous voulez prédire N_{2024}, il serait plus sage de ne pas prendre en compte N_{2010}.

Toutes ces considérations reviennent à se questionner sur la nature de ce que l'on observe, et comment on peut le modéliser. Cela revient à **poser un modèle**.

Definition 2.2 : Modèle statistique

Un modèle statistique est un triplet $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, où

- $\mathcal{X}$ est l'espace dans lequel vivent les observations.
- $\mathcal{A}$ est une tribu sur $\mathcal{X}$.
- $(P_\theta)_{\theta\in\Theta}$ est une famille de lois sur $\mathcal{X}$, indexée par $\theta \in \Theta$ (appelé espace des paramètres).

Il est implicitement supposé que nos observations sont modélisées par X, variable sur $\mathcal{X}$, qui a pour loi P_θ, pour un θ inconnu que l'on va chercher à déterminer.

Le cas le plus courant est le suivant : on observe $X_1, \dots, X_n$ des v.a.r. i.i.d., de loi standard indexée par $\theta \in \Theta$ ($[0, 1]$ pour les jours d'ensoleillement).

- Si X_1 est une v.a. discrète, alors le modèle associé aux observations de $X_1, \dots, X_n$ est $(\mathbb{N}^n, \mathcal{P}(\mathbb{N}^n), (P_\theta^{\otimes n})_{\theta\in\Theta})$, où P_θ désigne la loi d'un seule observation X_1, et $P_\theta^{\otimes n}$ la loi d'un n-uplet i.i.d., définie par

$$P_\theta^{\otimes n}(k_1, \dots, k_n) = \mathbb{P}(\{X_1 = k_1\} \cap \dots \cap \{X_n = k_n\}) = \prod_{i=1}^{n} P_\theta(\{k_i\}).$$

- Si X_1 est une variable continue, alors le modèle associé aux observations est $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (P_\theta^{\otimes n})_{\theta\in\Theta})$, où cette fois ci $P_\theta^{\otimes n}$ est définie par sa densité (par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$)

$$f_\theta^{\otimes n}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_\theta(x_i).$$

L'étape de définition d'un modèle étant cruciale (sans ça on ne peut rien faire), quelques exemples pour s'assurer de la bonne compréhension de cette notion.

Exemple 2.3 *: Exemples simples/standard.*
Jours d'ensoleillements : Si on ne constate aucun phénomène particulier susceptible de modifier le comportement normal du nombre de jours d'ensoleillements, on peut supposer que les N_j ($j = 2000, \dots, 2023$) sont i.i.d., de loi $\mathcal{B}(n, \theta)$, pour un θ inconnu. Le modèle est alors

$$(\llbracket 0, n\rrbracket^{24}, \mathcal{P}(\llbracket 0, n\rrbracket^{24}, (\mathcal{B}(n, \theta)^{\otimes 24})_{\theta\in[0,1]}.$$

On va alors chercher à dire des choses sur θ (par exemple pour essayer de prévoir N_{2024}.

Temps de trajets domicile travail : Vous collectez vos n derniers temps de trajet entre votre domicile et ici, et les notez $X_1, \dots, X_n$. A problème de métro près, vous pouvez considérer que ces $X_1, \dots, X_n$ sont i.i.d.. Une modélisation possible pour un temps de trajet est $X_1 \sim \mathcal{E}(\theta)$ (loi exponentielle de paramètre θ). Le modèles correspondant à ces observations est alors

$$(]0, +\infty[^n, \mathcal{B}(]0, +\infty[^n), (\mathcal{E}(\theta)^{\otimes n})_{\theta>0}).$$

On va alors chercher à dire des choses sur θ pour prévoir à quelle heure il faut partir de chez vous pour arriver en retard avec probabilité moins de 1%.

Espacement des rangées en train, taille des gens : La SNCF construit de nouvelles rames, et a besoin de déterminer un espacement entre ses rangées de sièges, de telle manière à ce que seuls les 1% les plus grands de ses usagers puissent être incommodés. Pour cela, elle doit se faire une idée de la distribution des tailles parmi ses usagers.

À cette fin, elle demande à ses contrôleurs de mesurer les passagers pendant 2 mois, et récolte les observations $X_1, \dots, X_n$. Comme il n'y a pas de raison de supposer que la taille d'un passager influe sur celle de son voisin, on peut supposer que les X_i sont i.i.d.. Enfin, une modélisation raisonnable de la taille d'un individu est $X_1 \sim \mathcal{N}(\mu, \sigma^2)$, où $\mu > 0$ et $\sigma^2 > 0$. Le modèle correspondant à ces observations est alors

$$(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), (\mathcal{N}(\mu, \sigma^2)^{\otimes n})_{\mu>0,\sigma^2>0}).$$

On remarque que bien qu'on ne puisse pas observer de taille négative en pratique, de part la modélisation $X_1 \sim \mathcal{N}(\mu, \sigma^2)$ on est obligé de prendre comme espace d'observation $\mathbb{R}^n$. C'est un phénomène général : les modèles sont toujours un peu faux.

On remarque aussi que ce modèle est indexé par deux paramètres réels : $\mu > 0$ et $\sigma^2 > 0$, ou de manière équivalente par le vecteur paramètre (μ, σ^2). Pour dire des choses sur la taille moyenne des gens en général, on cherchera à estimer μ. Pour dire des choses sur la dispersion de la taille des gens autour de la taille moyenne en général on cherchera à estimer σ^2.

Si on revient sur le problème de la SNCF : on a besoin de prévoir la taille minimale des 1% les plus grands en général, ce qui revient à estimer le quantile d'ordre 99% d'une loi $\mathcal{N}(\mu, \sigma^2) \sim \mu + \sigma\mathcal{N}(0, 1)$. Si q est le quantile d'ordre 99% d'une loi $\mathcal{N}(0, 1)$ (loi standard, donc quantiles connus), la valeur que cherche à estimer la SNCF est alors

$$\mu + \sigma q,$$

et on aura donc besoin d'estimer les deux paramètres.

Ces exemples suivent le schéma classique de formalisation d'un problème d'inférence statistique basé sur des observations.

1. Étape 1 : Modéliser vos observations.
2. Étape 2 : Traduire votre problème en terme de paramètres de votre modèle (quel paramètre vous intéresse pour répondre à votre question).
3. Étape 3 : Estimer votre paramètre, donner un intervalle de confiance, un test.
4. Étape 4 (optionnelle) : Se servir de l'étape précédente pour faire une prévision sensée.

Les deux premiers points ont été abordés, il reste maintenant à détailler l'étape 3, correspondant aux 3 différentes choses que la statistique inférentielle vous permettra de faire :

1. *Estimation ponctuelle* : estimer le paramètre θ, ou sa partie qui vous intéresse, c-a-d trouver une valeur approchée.
2. *Intervalle de confiance* : Donner une zone de Θ dans laquelle le vrai paramètre θ a des chances de se trouver.
3. *Faire un test sur θ* : Répondre à une question binaire sur θ, par exemple "θ est-il positif ?".

2.2 Estimation (ponctuelle)

On commence par définir ce que l'on entend par "estimateur".

Definition 2.4 : Estimateur

Dans le modèle $(\mathcal{X}^n, \mathcal{A}, (P_\theta^{\otimes n})_{\theta\in\Theta})$, un estimateur de θ, est une fonction $T : \mathcal{X}^n \to \Theta$.

En d'autres termes, un estimateur T est une fonction de l'espace des observations dans l'espace des paramètres. Il est tacitement admis que $T(X_{1:n})$ est censé approcher θ.

Exemple 2.5.

— Dans le modèle $(\mathcal{N}(\theta, \sigma_0^2)^{\otimes n})_{\theta\in\mathbb{R}}$, un estimateur standard de θ est la moyenne empirique

$$T(X_{1:n}) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

— Dans le modèle $(\mathcal{U}(]0, \theta[)^{\otimes n})_{\theta>0}$ (tirage de n lois uniformes i.i.d.), deux estimateurs raisonnables de θ sont

$$T_1(X_{1:n}) = 2 * \bar{X}_n,$$
$$T_2(X_{1:n}) = \max_{i=1\dots n} X_i.$$

Le point important est qu'un estimateur ne peut pas prendre θ en argument. Ses valeurs ne doivent dépendre du paramètre θ qu'au travers de $P_\theta^{\otimes n}$. Lorsque cela a du sens, pour une statistique f, on notera

$$E_\theta(f(X_{1:n})) = \int_{\mathcal{X}^n} f(u) P_\theta^{\otimes n}(du),$$

correspondant à l'espérance de $f(X_{1:n})$ lorsque $X_{1:n} \sim P_\theta^{\otimes n}$.

On peut aussi trouver les notations $\hat{\theta}$, $\hat{\theta}_n$ pour désigner un estimateur de θ (le n rappelle qu'il est basé sur un n-échantillon). Par convention le chapeau est réservé aux statistiques/estimateurs (observables à partir des données).

Biais et Risque quadratique

Pour un estimateur $T(X_{1:n})$ donné, on cherche à savoir s'il est proche de la cible θ ou non. Une manière de mesurer cela est de voir si en moyenne $T(X_{1:n})$ vaut θ (biais), et de mesurer l'écart quadratique moyen entre $T(X_{1:n})$ et θ (risque quadratique).

Une manière d'évaluer la qualité d'estimation ponctuelle est de considérer le *risque quadratique* de l'estimateur T.

Definition 2.6

Dans le modèle $(\mathcal{X}^n, \mathcal{A}, (P_\theta^{\otimes n})_{\theta\in\Theta})$, le *risque quadratique* de T est

$$R_T(\theta) = E_\theta(T(X_{1:n}) - \theta)^2.$$

Si $E_\theta T(X_{1:n})^2 = +\infty$, le risque quadratique de T est aussi infini. Ce n'est pas la seule manière d'évaluer la qualité d'un estimateur, on verra plus finement les manières de comparer les estimateurs dans le chapitre bayésien.

Décomposition biais/variance : On peut décomposer le risque quadratique de la manière suivante :

$$R_T(\theta) = (E_\theta(T(X_{1:n})) - \theta)^2 + \mathrm{Var}_\theta(T(X_{1:n})),$$

où

— le terme $E_\theta(T(X_{1:n})) - \theta$ est appelé *biais* de l'estimateur T,

— $\mathrm{Var}_\theta(T(X_{1:n})) = E_\theta\left[(T(X_{1:n}) - E_\theta(T(X_{1:n})))^2\right]$ est la variance de l'estimateur T sous $P_\theta^{\otimes n}$.

Un estimateur T tel que

$$\forall \theta \in \Theta \quad E_\theta(T(X)) = \theta$$

est dit *non-biaisé*, son risque quadratique se résume alors à sa variance.

Exemple 2.7. Dans le modèle $(\mathcal{U}(]0,\theta[)^{\otimes n})_{\theta\in\Theta}$,

— l'estimateur $T_1(X_{1:n}) = 2\bar{X}_n$ est sans biais, son risque quadratique est

$$R_{T_1}(\theta) = \frac{4}{n}\mathrm{Var}(\mathcal{U}(]0,\theta[)) = \frac{4\theta^2}{n}\mathrm{Var}(\mathcal{U}(]0,1[) = \frac{\theta^2}{3n}.$$

— l'estimateur $T_2(X_{1:n})$ est biaisé. Un calcul simple montre que T_2 a pour densité $nt^{n-1}\theta^{-n}$, on en déduit

  — biais : $\theta - \frac{n\theta}{n+1} = \frac{\theta}{n+1}$,

  — risque quadratique : $R_{T_2}(\theta) = \frac{\theta^2}{(n+1)(n+2)}$.

Dans cet exemple le risque de T_2 est sensiblement meilleur que celui de T_1. De manière générale, l'absence de biais est une propriété souhaitable mais ne garantit pas l'optimalité.

Comportements asymptotiques souhaitables

Lorsque l'on parle de "convergence" d'estimateurs, on se place dans le cas où on observe un n-échantillon i.i.d.. On prendra donc le modèle $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta}$.

Un estimateur étant une fonction mesurable des observations (et cet espace d'observation variant avec n), il serait plus précis de parler de suite d'estimateurs. Encore un abus de langage et une convention.

Une propriété minimale des estimateurs est que lorsque l'information disponible (n, taille d'échantillon) croît, l'estimateur converge vers la valeur souhaitée.

Definition 2.8 : Consistance

Un(e) (suite d') estimateur(s) T (de $\theta \in \mathbb{R}^k$) est dit *consistant* si

$$\forall \theta \in \Theta \quad T(X_{1:n}) \xrightarrow[n\to+\infty]{\mathbb{P}} \theta.$$

Lorsque la convergence a lieu p.s. on parle de consistance forte.

Recettes de consistance : Prouver une consistance peut se faire

1. avec la loi des grands nombres,
2. en utilisant la convergence du risque quadratique vers 0.

Exemple 2.9. En reprenant le modèle $(\mathcal{U}(]0,\theta[^{\otimes n})_{\theta>0}$, avec les deux estimateurs T_1 et T_2.

— Comme $E_\theta|X_1| < +\infty$, la loi des grands nombres donne $\bar{X}_n \to_{n\to+\infty} \theta/2$ p.s., on en déduit que T_1 est (fortement) consistant.

— Comme $R_{T_2}(\theta) \to_{n\to\infty} 0$, T_2 converge vers θ dans $L_2(P_\theta)$, donc en proba (P_θ), il est lui aussi consistant. On peut montrer la forte consistance en utilisant Borel-Cantelli (exercice).

Definition 2.10 : Normalité asymptotique

Dans le cas où $\theta \in \mathbb{R}$, un estimateur T est dit *asymptotiquement normal en* θ s'il existe une suite r_n positive et $\sigma_\theta^2 > 0$ tels que

$$r_n(T(X_{1:n}) - \theta) \rightsquigarrow_{n\to+\infty} \mathcal{N}(O, \sigma_\theta^2).$$

La normalité asymptotique en θ est la convergence en loi de l'estimateur renormalisé vers une loi normale non dégénérée. La normalité asymptotique désigne la même propriété lorsqu'elle est valide pour tout $\theta \in \Theta$. Enfin on peut étendre la définition en dimension supérieure en requérant une matrice de covariance non-nulle.

La normalité asymptotique n'est pas intéressante en elle-même : l'idée est de chercher le comportement asymptotique des fluctuations de l'estimateur autour de sa cible pour pouvoir en déduire ultérieurement des garanties en terme de risque asymptotique ou d'intervalle de confiance. Le théorème central limite indique que le comportement asymptotique normal est relativement fréquent.

Exemple 2.11. On reprend notre modèle favori $(\mathcal{U}(]0,\theta[^{\otimes n})$, et nos deux estimateurs.

— Comme $E_\theta X_1^2 < +\infty$, le théorème central limite donne

$$\sqrt{n}\,(T_1(X_{1:n}) - \theta) \rightsquigarrow \mathcal{N}\left(0, \frac{\theta^2}{3}\right).$$

— Pour T_2, un calcul rapide donne, pour $t > 0$,

$$P_\theta(n(\theta - T_2(X_{1:n})) > t) = \left(1 - \frac{t}{n\theta}\right)^n \mathbb{1}_{t\le n\theta} \longrightarrow_{n\to\infty} \exp(-t/\theta),$$

et donc $n(\theta - T_2(X_{1:n})) \rightsquigarrow \mathcal{E}(\theta^{-1})$. Pas de normalité asymptotique donc, mais on a quand même un comportement asymptotique.

Recettes pour normalité asymptotique : La plupart des normalités asymptotiques se prouvent à l'aide du théorème central limite et de deux outils : le lemme de Slutsky et la Δ-méthode. On rappelle le Lemme de Slutsky ci-dessous :

Théorème : Lemme de Slutsky

Soit (X_n, Y_n) une suite de couples de v.a.r.. Si $X_n \rightsquigarrow X$ et $Y_n \rightsquigarrow c$, où c désigne la v.a.r. constante valant c, alors

1. $Y_n \xrightarrow[n\to+\infty]{\mathbb{P}} c$,
2. $(X_n, Y_n) \rightsquigarrow (X, c)$.

En particulier, $X_n + Y_n \rightsquigarrow X + c$, $X_n Y_n \rightsquigarrow cX$.

Le lemme de Slutsky autorise certaines opérations sur les limites en loi. Par exemple $X_n \rightsquigarrow \mathcal{N}(0, \sigma^2)$ et $\hat{\sigma}_n \xrightarrow{\mathbb{P}} \sigma$ implique $X_n/\hat{\sigma}_n \rightsquigarrow \mathcal{N}(0, 1)$, ce qui sera assez utile pour les intervalles de confiance. Une conséquence du lemme de Slutsky est la "Méthode Δ", permettant de transférer la propriété de normalité asymptotique via fonctionnelle différentiable.

Théorème 2.12 : Δ-méthode

Soit $(X_n)_{n\geq 1}$ une suite de variable aléatoires, et $(r_n)_{n\geq 1}$ suite de réels positifs telles que

1. $r_n \xrightarrow[n\to+\infty]{} +\infty$,
2. $r_n(X_n - x) \rightsquigarrow_{n\to+\infty} X$,

pour un $x \in \mathbb{R}$ et X une variable aléatoire sur $\mathbb{R}$. Soit $g : \mathbb{R} \to \mathbb{R}$ une fonction différentiable en x, on a alors

$$r_n(g(X_n) - g(x)) \rightsquigarrow_{n\to+\infty} g'(x)X.$$

Preuve du Théorème 2.12. Comme $r_n \to +\infty$, une première application du Lemme de Slutsky à $(r_n^{-1}, r_n(X_n - x))$ permet de montrer $X_n \xrightarrow{\mathbb{P}} x$. On peut alors déduire de la différentiabilité de g en x que

$$\frac{g(X_n) - g(x)}{X_n - x} \xrightarrow{\mathbb{P}} g'(x).$$

Le Lemme de Slutsky garantit alors que $(r_n(X_n - x), (g(X_n) - g(x))/(X_n - x)) \rightsquigarrow (X, g'(x))$, et par continuité du produit $r_n(g(X_n) - g(x)) \rightsquigarrow g'(x)X$.

□

Exemple 2.13. Dans le modèle $(\mathcal{E}(\theta)^{\otimes n})_{\theta>0}$ (observations de n v.a. exponentielles de paramètres θ indépendantes) où on cherche à estimer θ. On peut partir du théorème central limite

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\theta}\right) \rightsquigarrow \mathcal{N}\left(0, \frac{1}{\theta^2}\right),$$

et appliquer la méthode Δ avec la fonction $u \mapsto 1/u$ pour montrer que l'estimateur $T(X_{1:n}) = \bar{X}_n^{-1}$ vérifie une normalité asymptotique

$$\sqrt{n}\left(T(X_{1:n}) - \theta\right) \rightsquigarrow -\theta^2 \mathcal{N}\left(0, \frac{1}{\theta^2}\right) = \mathcal{N}(0, \theta^2).$$

2.3 Méthodes classiques d'estimation

On présente ici deux méthodes classiques d'estimation : la méthode des moment et celle du maximum de vraisemblance, qui ont toutes deux été utilisées dans les exemples.

2.3.1 Méthode des moments

Dans un modèle où on observe $X_1, \dots, X_n$ i.i.d. de loi commune P_θ, où $\theta \in \Theta \subset \mathbb{R}$, une approche "naturelle" consiste à regarder l'espérance de X_1 sous P_θ (si bien définie),

$$\psi(\theta) = E_\theta(X_1),$$

puis de voir si cette espérance caractérise entièrement θ (c'est à dire si $\psi : \Theta \to \mathbb{R}$ est injective). Si c'est bien le cas, on substitue la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ à $E_\theta(X)$ pour trouver un estimateur de θ, c'est à dire qu'on pose

$$\psi(\hat{\theta}_n) = \bar{X}_n,$$

ce qui par injectivité de ψ permet de définir

$$\hat{\theta}_n := \psi^{-1}(\bar{X}_n).$$

En d'autres termes on choisit comme estimateur le θ sous lequel $E_\theta(X_1)$ coïncide avec $\bar{X}_n$.

Exemple 2.14. Dans le modèle $(\mathcal{E}(\theta)^{\otimes n})_{\theta>0}$, cherchons encore à estimer θ. À θ fixé, on a

$$E_\theta X_1 = \frac{1}{\theta}.$$

Avec les notations précédentes,

$$\psi : \begin{cases}]0, +\infty[& \longrightarrow & \mathbb{R} \\ x & \mapsto & E_x(X_1) = 1/x \end{cases}$$

est bien injective. Un estimateur sensé de θ s'obtient alors en remplaçant $E_\theta X_1$ par $\bar{X}_n$, c'est à dire poser

$$\psi(\hat{\theta}) = \frac{1}{\hat{\theta}_n} = \bar{X}_n,$$

ou encore $\hat{\theta}_n = \frac{1}{\bar{X}_n}$.

On généralise cette heuristique dans deux directions :

1. il arrive que $\theta \mapsto E_\theta(X_1)$ ne soit pas injective. Il faut alors regarder une autre espérance caractérisant θ, c'est à dire une autre fonction test $f : \mathcal{X} \to \mathbb{R}$ telle que $\psi : \theta \mapsto E_\theta(f(X_1))$ soit injective, et remplacer $E_\theta(f(X_1))$ par $\bar{f} = \frac{1}{n}\sum_{i=1}^n f(X_i)$ (l'espérance "sur l'échantillon"). La plupart du temps, on prendra f telle que $E_\theta(f(X_1))$ soit facile à calculer, les fonctions puissance $f(x) = x^k$ sont souvent un choix pertinent (d'où la dénomination "méthode des moments").

2. Cette méthode s'étend facilement au cas où $\Theta \subset \mathbb{R}^k$ (il faut alors trouver k fonctions test).

Deux exemples avant de passer à la définition générale.

Exemple 2.15 *: Non injectivité de l'espérance.*

On se place dans le modèle $(\mathcal{U}(]-\theta,\theta[)^{\otimes n})_{\theta>0}$. Si $\psi_1 : x \mapsto E_x(X_1)$, on a alors, pour tout $x > 0$, $\psi_1(x) = \mathbb{E}(\mathcal{U}(]-x,x[)) = 0$, et donc ψ_1 n'est pas injective.

Dans un sens c'est assez moral : le paramètre θ dans ce modèle ne caractérise pas la tendance centrale des observations (toujours 0), mais plutôt sa dispersion. Une fonction test appropriée devrait refléter cette intuition. Un choix pertinent est alors

$$\psi_2 : \begin{cases}]0,+\infty[& \longrightarrow &]0,+\infty[\\ x & \mapsto & E_x(X_1^2), \end{cases}$$

c'est à dire aller regarder les moments d'ordre 2. Le calcul donne, pour $x > 0$,

$$\psi_2(x) = \mathrm{Var}(\mathcal{U}(]-x,x[)) = 4x^2\mathrm{Var}(\mathcal{U}(]0,1[)) = \frac{x^2}{3},$$

ψ_2 est donc bien injective (sur $]0,+\infty[$). Un estimateur par moments satisfait alors

$$\psi_2(\hat{\theta}_n) = \bar{X^2},$$

avec $\bar{X^2} = \frac{1}{n}\sum_{i=1}^n X_i^2$, ou encore

$$\hat{\theta}_n = \sqrt{3\bar{X^2}}.$$

Exemple 2.16 *: Avec deux paramètres.*

On se place dans le modèle Gaussien $(\mathcal{N}(\mu,\sigma^2)^{\otimes n})_{\mu\in\mathbb{R},\sigma^2>0}$, c'est à dire un modèle où on ne connaît ni μ ni σ^2. Il va donc nous falloir trouver deux fonctions tests f_1 et f_2 permettant d'estimer μ et σ^2 à partir de $E_{\mu,\sigma^2}(f_1(X_1))$ et $E_{\mu,\sigma^2}(f_2(X_1))$.

Le premier paramètre μ étant un paramètre de tendance centrale, il devrait être caractérisé par $(\mu,\sigma^2) \mapsto E_{\mu,\sigma^2}X_1 = \mu$, c'est à dire la fonction test $f_1 : x \mapsto x$ (c'est bien le cas).

Le deuxième paramètre σ^2 étant un paramètre de dispersion (il caractérise la variance de X_1), il semble pertinent d'aller regarder le moment d'ordre 2 : $f_2 : x \mapsto x^2$. En effet, on a

$$E_{\mu,\sigma^2}(X_1^2) = \mu^2 + \sigma^2,$$

de sorte que, à μ fixé (en fait estimé par la première équation), on peut retrouver σ^2. Mis bout à bout, on a le système d'équations

$$\begin{aligned} E_{\hat{\mu},\hat{\sigma}^2}f_1(X_1) &= \bar{X} \\ E_{\hat{\mu},\hat{\sigma}^2}f_2(X_1) &= \bar{X^2}, \end{aligned}$$

qui s'écrit

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{\mu}^2 + \hat{\sigma}^2 &= \bar{X^2}, \end{aligned}$$

ou encore

$$\hat{\mu} = \bar{X},$$
$$\hat{\sigma}^2 = \bar{X^2} - (\bar{X})^2,$$

on retrouve alors des estimateurs "naturels" des moyenne et variance : la moyenne empirique et la variance empirique (sur l'échantillon).

On peut maintenant définir la méthode des moments dans le cadre général où on observe $X_1, \dots, X_n$ i.i.d. de loi commune P_θ, où $\theta \in \Theta \subset \mathbb{R}^k$. Cela consiste en :

1. Trouver k fonctions tests (autant que de paramètres) $f_1, \dots, f_k$ telles que

$$\psi : \begin{cases} \Theta & \longrightarrow & \mathbb{R}^k \\ \theta & \mapsto & E_\theta(\mathbf{f}(X_1)) := (E_\theta(f_j(X_1)))_{j=1,\dots,k} \end{cases}$$

   soit injective.

2. Définir $\hat{\theta}$ en remplaçant $E_\theta(\mathbf{f})$ par $\bar{\mathbf{f}}$ (vraies espérances par espérances sur l'échantillon), c'est à dire caractériser $\hat{\theta}$ par le système de k équations

$$\psi(\hat{\theta}) = \bar{\mathbf{f}} = \left(\frac{1}{n} \sum_{i=1}^{n} f_j(X_i) \right)_{j=1,\dots,k}.$$

   Par injectivité de ψ, on a

$$\hat{\theta} = \psi^{-1}(\bar{\mathbf{f}}).$$

Un des avantages de la méthode des moments est qu'elle fournit des estimateurs dont le comportement asymptotique est assez simple à déterminer.

Proposition 2.17 : Consistance des estimateurs par moments

Soit $\theta \in \mathring{\Theta}$. Si $\psi : \theta \mapsto E_\theta(\mathbf{f})$ est localement injective et continue autour de θ, et $E_\theta \|f(X_1)\| < +\infty$, alors

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbb{P}} \theta,$$

où $\hat{\theta}_n$ désigne l'estimateur par moments associé au système d'équations ψ.

Preuve de la Proposition 2.17. La condition $E_\theta\|f(X_1)\| < +\infty$ permet d'invoquer la loi des grands nombres : on a, en regardant coordonnée par coordonnée, $\psi(\hat{\theta}) = \bar{\mathbf{f}} \xrightarrow[n\to+\infty]{\mathbb{P}} E_\theta(\mathbf{f}(X_1)) = \psi(\theta)$. Pour en déduire que $\hat{\theta} \xrightarrow[n\to+\infty]{\mathbb{P}} \theta$, montrer que ψ^{-1} est continue en $\psi(\theta)$ suffit.

Pour ce faire, par hypothèse on a U voisinage ouvert de θ dans Θ tel que $\psi_{|U}$ soit continue et injective. Comme $U \subset \mathbb{R}^k$ et ψ est à valeurs dans $\mathbb{R}^k$, le Théorème d'invariance du domaine de Brouwer montre alors que $V = \psi(U)$ est un ouvert de $\mathbb{R}^k$, et $\psi_{|U}$ est un homéomorphisme entre U et V. En particulier, ψ^{-1} est continue en $\psi(\theta)$. □

On peut remarquer que la régularité "minimale" de ψ garantissant la consistance est plutôt "ψ^{-1} est continue en $\psi(\theta)$". En pratique, on utilise assez peu ce résultat général : on calcule ψ^{-1} au cas par cas et il suffit de vérifier qu'il est continu au point d'intérêt. Concernant les hypothèses probabilistes, comme il s'agit d'appliquer la loi des grands nombres à une moyenne $\bar{\mathbf{f}}$, une condition d'intégrabilité suffit.

Comme dans la partie probabilité, la méthode des moments se basant sur une substitution moyenne/moyenne empirique, les fluctuations de l'estimateur par moments autour de la cible peuvent se déduire d'un Théorème central limite sur $\bar{f}$, transféré sur $\hat{\theta}$ par la méthode Δ. On se limitera ici au cas $\Theta \subset \mathbb{R}$.

Théorème 2.18 : Normalité asymptotique des estimateurs par moments

Soit $\theta \in \mathring{\Theta} \subset \mathbb{R}$. Si $\psi : \theta \mapsto E_\theta(\mathbf{f})$ est localement $\mathcal{C}^1$ autour de θ, $\psi'(\theta) \neq 0$, et $E_\theta f^2(X_1) < +\infty$, alors

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \rightsquigarrow \mathcal{N}\left(0, \frac{\mathrm{Var}_\theta(f(X_1))}{\psi'(\theta)^2}\right),$$

où $\hat{\theta}$ désigne l'estimateur par méthode des moments associé à l'équation f.

Démonstration. La condition $E_\theta f^2(X_1) < +\infty$ permet de faire appel à la loi des grands nombres. On a

$$\sqrt{n}(\bar{f} - \psi(\theta)) \rightsquigarrow \mathcal{N}(0, \mathrm{Var}_\theta(f(X_1))).$$

Pour en déduire les fluctuations de $\hat{\theta} = \psi^{-1}(\bar{f})$, il faut s'assurer que ψ^{-1} existe et est dérivable pour pouvoir appliquer la méthode Δ.

C'est une conséquence du Théorème d'inversion locale : comme ψ est $\mathcal{C}^1$ autour de θ, et $\psi'(\theta) \neq 0$, il existe U voisinage ouvert de θ dans Θ tel que $\psi_{|U}$ soit un $\mathcal{C}^1$ difféomorphisme (sur son image, qui contient $\psi(\theta)$). Une application de la méthode Δ donne alors

$$\sqrt{n}(\hat{\theta} - \theta) \rightsquigarrow (\psi^{-1})'(\psi(\theta))\mathcal{N}(0, \mathrm{Var}_\theta(f(X_1))).$$

On conclut en remarquant que $(\psi^{-1})'(\psi(\theta)) = \frac{1}{\psi'(\theta)}$. □

Là encore on utilise assez peu ce théorème en pratique : on calcule ψ^{-1} et on applique directement la méthode Δ au cas par cas. Ces deux résultats montrent surtout qu'il est relativement "facile" d'obtenir le comportement asymptotique d'estimateurs construits par la méthode des moments.

Exemple 2.19 *: Paramètre de loi exponentielle.*

On reprend le modèle $(\mathcal{E}(\theta)^{\otimes n})_{\theta>0}$, pour lequel l'estimateur par méthode des moments associé à la fonction test $x \mapsto x$ était

$$\hat{\theta} = \frac{1}{\bar{X}}.$$

Comme $X_1 \in L_1(P_\theta)$, on a que $\bar{X} \xrightarrow[n\to+\infty]{\mathbb{P}} E_\theta(X_1) = 1/\theta$, par la loi des grands nombres. Comme $\varphi : x \mapsto 1/x$ (définie sur $]0, +\infty[$) est continue, on en déduit immédiatement

$$\hat{\theta} = \varphi(\bar{X}) \xrightarrow[n\to+\infty]{\mathbb{P}} \varphi(E_\theta(X_1)) = \theta,$$

et donc que $\hat{\theta}$ est consistant.

Pour les fluctuations autour de la cible, on part du Théorème central limite : comme $E_\theta(X_1^2) < +\infty$,

$$\sqrt{n}(\bar{X} - 1/\theta) \rightsquigarrow \mathcal{N}(0, 1/\theta^2).$$

Comme φ est $\mathcal{C}^1$ en $1/\theta$, on en déduit

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}(\varphi(\bar{X}) - \varphi(1/\theta)) \rightsquigarrow \varphi'(1/\theta)\mathcal{N}(0, 1/\theta^2) \sim \mathcal{N}(0, \theta^2),$$

en utilisant la méthode Δ et $\varphi'(1/\theta) = -1/(1/\theta)^2 = -\theta^2$.

2.3.2 Méthode du maximum de vraisemblance

En toute généralité, les méthodes du maximum de vraisemblance se basent sur un modèle *dominé* (c'est à dire un modèle où toutes les lois envisageables pour les observations admettent une densité au sens de Radon par rapport à une mesure σ-finie). Les cas des v.a. discrètes ou continues sont des sous-cas (qui constituent la majorité des situations faciles).

Cas discret

Dans le modèle $(\mathcal{X}^n, \mathcal{A}, (P_\theta^{\otimes n})_{\theta\in\Theta})$, dire qu'on est dans le cas discret revient à dire $\mathcal{X} = \mathbb{Z}$, et, pour tout $\theta \in \Theta$,

$$\forall k \in \mathbb{Z} \quad P_\theta(\{k\}) := p_\theta(k).$$

En d'autres termes P_θ admet une densité par rapport à la mesure de comptage sur $\mathbb{Z}$, qu'on notera p_θ (qui est une fonction de $\mathbb{Z}$ à valeurs dans $[0, 1]$).

Dans le cas $n = 1$, du point de vue probabiliste, l'issue x la plus probable sous P_θ est

$$x^* \in \arg\max_{k\in\mathbb{Z}} p_\theta(k),$$

correspondant au mode de P_θ. La maximisation de vraisemblance consiste à renverser ce point de vue : à observation x fixée, le θ le plus vraisemblable est

$$\hat{\theta} \in \arg\max_{\theta\in\Theta} p_\theta(x).$$

FAIRE DESSIN. À x fixé, la fonction

$$V_x : \begin{cases} \Theta & \to & \mathbb{R}^+ \\ \theta & \mapsto & p_\theta(x) \end{cases}$$

est appelée **vraisemblance**. La méthode du maximum de vraisemblance consiste alors à choisir

$$\hat{\theta} \in \arg\max_{\theta\in\Theta} V_X(\theta),$$

c'est à dire le θ le plus vraisemblable si on a observé X.

On peut remarquer que formellement V_X est une fonction aléatoire (X l'est). Cela dit, comme un estimateur est une fonction de $\mathcal{X}$ vers Θ, pour caractériser $\hat{\theta}$

entièrement il suffit de calculer $\arg\max_{\theta\in\Theta} V_x(\theta)$ pour toutes les valeurs x que peut prendre X.

Dans le cas $n \geq 1$ quelconque, le principe reste le même : à $x_{1:n} \in \mathbb{Z}^n$ fixé, la **vraisemblance** devient la fonction

$$V_{x_{1:n}} : \begin{cases} \Theta & \to & \mathbb{R}^+ \\ \theta & \mapsto & \prod_{i=1}^n p_\theta(x_i), \end{cases}$$

c'est à dire la fonction qui à un θ détermine à quel point il est vraisemblable si on observe $x_1, \ldots, x_n$. L'estimateur du maximum de vraisemblance est alors donné par

$$\hat{\theta} \in \arg\max_{\theta\in\Theta} V_{X_{1:n}}(\theta).$$

Autre remarque importante : il peut arriver que $V_{x_{1:n}}$ admette plusieurs ou aucun maximum. C'est un des défauts de cette méthode de n'être pas proprement définie (sauf dans certains cas très précis comme les modèles exponentiels).

Remarque importante : À $x_{1:n}$ fixé, lorsque $p_\theta(x_i) > 0$ pour tout $\theta \in \Theta$, on maximise plutôt la **log-vraisemblance**, définie par

$$\ell_{x_{1:n}}(\theta) = \log(V_{x_{1:n}}(\theta)) = \sum_{i=1}^n \log(p_\theta(x_i)),$$

souvent plus facile à calculer et donnant le même estimateur. Insistons sur le fait que le passage à la log-vraisemblance est pertinent **uniquement dans le cas où** $p_\theta(x_i) > 0$ pour tout $\theta \in \Theta$.

Exemple 2.20 *: Lois géométriques.*

On se place dans le modèle $(\mathbb{N}^n, \mathcal{P}(\mathbb{N}^n), (\mathcal{G}(\theta)^{\otimes n})_{\theta\in]0,1[})$, où $\mathcal{G}(\theta)$ est la loi géométrique de paramètre θ, caractérisée par

$$\forall k \geq 1 \quad \mathcal{G}(\theta)(\{k\}) = (1-\theta)^{k-1}\theta := p_\theta(k).$$

Les valeurs d'observations possibles pour une $\mathcal{G}(\theta)$ sont les $k \geq 1$. Pour trouver un EMV, il suffit alors de déterminer

$$\arg\max_{\theta\in\Theta} \prod_{i=1}^n p_\theta(x_i),$$

pour tout $x_1, \ldots, x_n \geq 1$. Or, pour tout $x \geq 1$ et $\theta \in]0,1[$, $p_\theta(x) > 0$. On peut donc se contenter de maximiser la **log**-vraisemblance

$$\begin{aligned} \ell_{x_{1:n}}(\theta) &= \sum_{i=1}^n \log(p_\theta(x_i)) \\ &= \sum_{i=1}^n \log(\theta) + (x_i - 1)\log(1-\theta), \end{aligned}$$

pour $x_1, \ldots, x_n \geq 1$.

On voit apparaître un petit souci : si $x_1 = x_2 = \ldots = x_n = 1$, alors $\ell_{x_{1:n}}(\theta) = n\log(\theta)$ qui est maximale en $\theta = 1$ (qui n'est pas dans $]0,1[$).

En revanche, si $\sum_{i=1}^n (x_i - 1) > 0$, en dérivant une fois on obtient

$$\begin{aligned} \ell'_{x_{1:n}}(\theta) &= \frac{n}{\theta} - \frac{\sum_{i=1}^n (x_i - 1)}{1-\theta} \\ &= \frac{n - n\theta - n\theta\bar{x} + n\theta}{\theta(1-\theta)} \\ &= \frac{n - n\theta\bar{x}}{\theta(1-\theta)}. \end{aligned}$$

On en déduit alors que $\ell_{x_{1:n}}$ est maximale en

$$\frac{1}{\bar{x}} < 1.$$

Avec un léger abus, on peut écrire dans les deux cas le maximum de vraisemblance comme $1/\bar{x}$, et donc l'estimateur par maximum de vraisemblance est

$$\hat{\theta} = \frac{1}{\bar{X}}.$$

On peut remarquer le phénomène suivant : prenons la fonction test $x \mapsto x$. On a alors

$$E_\theta(X_1) = \frac{1}{\theta},$$

qui est bien injective en θ. Un estimateur par la méthode des moments vérifie alors l'équation

$$\bar{X} = \frac{1}{\hat{\theta}_{Moments}},$$

soit

$$\hat{\theta}_{Moments} = \frac{1}{\bar{X}} = \hat{\theta}_{EMV}.$$

De fait, dans certains modèles (les modèles exponentiels), on peut relier estimateurs par moments et estimateurs par maximum de vraisemblance.

Cas continu

Dans le cas de v.a.r. continues, on a $\mathcal{X} = \mathbb{R}$, et P_θ qui admet une densité p_θ. Le principe de la maximisation de vraisemblance reste le même : si $x_1, \ldots, x_n$ sont des valeurs d'observations possibles (c'est à dire telles que $p_\theta(x_i) > 0$ pour au moins un θ),

$$V_{x_{1:n}} : \begin{cases} \Theta & \to & \mathbb{R}^+ \\ \theta & \mapsto & \prod_{i=1}^n p_\theta(x_i) \end{cases}$$

est la fonction de vraisemblance, qui à θ associe un score décrivant à quel point θ est vraisemblable si on a observé $x_1, \ldots, x_n$ (FAIRE DESSIN CAS $n = 1$).

Un estimateur du maximum de vraisemblance est alors donné par

$$\hat{\theta} = \arg\max_{\theta \in \Theta} V_{X_{1:n}}(\theta).$$

Les mêmes deux remarques que dans le cas discret :

1. En pratique il suffit de calculer $\arg\max_{\theta\in\Theta} V_{x_{1:n}}(\theta)$ pour toutes les valeurs $x_1, \ldots, x_n$ que peut prendre $X_{1:n}$, c'est à dire les valeurs $x_1, \ldots, x_n$ telles que $\prod_{i=1}^n p_\theta(x_i) \neq 0$ pour **au moins un** $\theta \in \Theta$.
2. Si **pour tout** $\theta \in \Theta$ $\prod_{i=1}^n p_\theta(x_i) \neq 0$, alors maximiser $V_{x_{1:n}}$ revient à maximiser la **log**-vraisemblance

   $$\ell_{x_{1:n}}(\theta) = \log(V_{x_{1:n}}(\theta)) = \sum_{i=1}^n \log(p_\theta(x_i)),$$

   ce qui est souvent plus commode en pratique.

Exemple 2.21 *: Lois uniformes.*

On se place dans le modèle $(\mathcal{U}(]0,\theta])^{\otimes n})_{\theta>0}$. La densité (pour une observation est donnée par)

$$p_\theta(x) = \frac{1}{\theta}\mathbb{1}_{]0,\theta]}(x).$$

On remarque deux choses :

1. L'ensemble des valeurs possibles pour $x_{1:n}$ est $]0,+\infty[^n$: en effet pour $\theta \geq \max_{i=1,\dots,n} x_i$, $\prod_{i=1}^n p_\theta(x_i) > 0$.
2. Pour tout $x_{1:n} \in]0,+\infty[^n$, on a des valeurs de θ pour lesquelles $\prod_{i=1}^n p_\theta(x_i) = 0$: les $\theta < \max_{i=1,\dots,n} x_i$. On ne pourra donc pas utiliser la log-vraisemblance.

Calculons l'estimateur du maximum de vraisemblance. Soit donc $x_{1:n} \in]0,+\infty[^n$. On a

$$\begin{aligned} V_{x_{1:n}}(\theta) &= \prod_{i=1}^n p_\theta(x_i) \\ &= \theta^{-n} \prod_{i=1}^n \mathbb{1}_{]0,\theta]}(x_i) \\ &= \theta^{-n}\mathbb{1}_{m_n \leq \theta}, \end{aligned}$$

où $m_n = \max_{i=1,\dots,n} x_i$. FAIRE DESSIN. On a alors immédiatement

$$\arg\max_{\theta>0} V_{x_{1:n}}(\theta) = m_n.$$

On en déduit que

$$\hat{\theta}_{EMV} = \max_{i=1,\dots,n} X_i,$$

ce qui correspond à l'estimateur "naturel" dans cette situation.

L'exemple précédent est assez générique : lorsque bien défini, un estimateur du maximum de vraisemblance est souvent "le meilleur", au moins pour certains modèles assez réguliers (comme les modèles exponentiels). Il est toutefois plus technique à mettre en oeuvre que la famille des estimateurs par moments.

2.4 Intervalles de confiance

Dans les sections précédents on a construit des estimateurs $\hat{\theta}$ de θ en se basant sur des n-uplets d'observations $X_1, \dots, X_n$. La partie probabiliste permet de dire à quel point $\hat{\theta}$ fluctue autour de θ, d'un point de vue non asymptotique ou asymptotique. Un intervalle de confiance renverse encore le point de vue : à partir de ces fluctuations on va déterminer à quel point la cible $\hat{\theta}$ est proche de θ, et donner un ensemble de Θ dans lequel on est à peu près sûr que θ se trouve (avec forte probabilité).

À partir d'ici on suppose que $\Theta \subset \mathbb{R}$ (un seul paramètre à estimer).

2.4.1 Non asymptotique

Definition 2.22 : Intervalle de niveau de confiance $1-\alpha$

Soit $(\mathcal{X}^n, \mathcal{A}, (P_\theta^{\otimes n})_{\theta\in\Theta})$ un modèle et $\alpha \in]0,1[$. Un intervalle de confiance (par défaut non asymptotique) de niveau $1-\alpha$ pour θ est un couple de statistiques (T^-, T^+) tel que

$$\forall \theta \in \Theta \quad P_\theta\left(T^- \leq \theta \leq T^+\right) \geq 1-\alpha.$$

— Un intervalle de confiance est dit **unilatère** si $T^- = -\infty$ ou $T^+ = +\infty$ (faire dessin).

— Un intervalle de confiance est dit **bilatère** si les deux bornes sont finies.

Lorsque l'inégalité est une égalité on parle de niveau de confiance *exact*. Prendre $T^- = -\infty$ et $T^+ = +\infty$ garantit toujours un niveau $1-\alpha$, le but implicite est de trouver des intervalles de confiance les plus petits possibles. En ce sens, des intervalles de confiances "croissants" en fonction du niveau de confiance sont naturels.

Dans le cas où Θ n'est pas un sous-ensemble de $\mathbb{R}$, on peut définir plus généralement des régions de confiances (plus nécessairement des intervalles) comme des sous-ensembles aléatoires de Θ (ce qui nécessite d'équiper Θ avec une tribu et de vérifier certaines hypothèses de mesurabilité, en dehors du cadre de ce cours).

Recettes pour les IC non asymptotiques : Il y a deux manières de faire, à partir d'un estimateur T de θ :

1. soit on connaît la loi d'une *quantité pivotale* (usuellement de type $(T-\theta)/a$, idéalement ne dépendant pas de θ) et on peut en inférer un intervalle de confiance (cas idéal),
2. soit on passe par des inégalités de concentration (à notre niveau BT ou Markov).

Exemple 2.23.

Dans le modèle $\mathcal{U}(]0,\theta[)^{\otimes n}$, avec $T_2(X_{1:n}) = \max_{i=1,\dots,n} X_i$, on peut remarque que la loi de $\frac{T_2}{\theta}$ ne dépend pas de θ, $\frac{T_2}{\theta}$ va jouer le rôle de quantité pivotale.

On sait que, pour $t \in]0,1[$,

$$P_\theta\left(t \leq \frac{T_2}{\theta} \leq 1\right) = 1 - t^n.$$

Pour $\alpha \in]0,1[$, en prenant $t_\alpha = \alpha^{\frac{1}{n}}$, on en déduit

$$\forall \theta > 0 \quad P_\theta\left(\alpha^{\frac{1}{n}} \leq \frac{T_2}{\theta} \leq 1\right) = 1-\alpha,$$

et donc que $\left[T_2, T_2\alpha^{-\frac{1}{n}}\right]$ est un intervalle de confiance pour θ au niveau de confiance α.

Considérons maintenant l'estimateur $T_1(X_{1:n}) = 2\bar{X}_n$ (estimateur par moments). Calculer sa loi sous P_θ s'avère compliqué, on utilisera alors l'inégalité de Bienaymé Cebicev pour construire un intervalle de confiance. En effet, on

$$\mathrm{Var}_\theta(T_1) = 4\mathrm{Var}_\theta(\bar{X}_n) = \frac{4}{n}\mathrm{Var}_\theta(X_1) = \frac{\theta^2}{12n}.$$

On en déduit alors que, pour tout $t > 0$,

$$\mathbb{P}\left(|T_1(X_{1:n}) - \theta| \geq t\right) \leq \frac{\theta^2}{12nt^2}.$$

Si $\alpha \in]0,1[$, on a alors

$$T_1 - \frac{\theta}{\sqrt{12n\alpha}} \leq \theta \leq T_1 + \frac{\theta}{\sqrt{12n\alpha}},$$

avec probabilité plus grande que $1-\alpha$. Malheureusement ce n'est **pas** un intervalle de confiance : les bornes font intervenir θ (qui est inconnu). Pour en déduire un intervalle de confiance, il faut retravailler les inégalités :

— de $\theta \leq T_1 + \frac{\theta}{\sqrt{12n\alpha}}$ on déduit que

$$\theta \leq \frac{T_1}{1 - (12n\alpha)^{-1/2}},$$

pour n assez grand pour que $12n\alpha > 1$.

— de $T_1 - \frac{\theta}{\sqrt{12n\alpha}} \leq \theta$ on déduit

$$\theta \geq \frac{T_1}{1 + (12n\alpha)^{-1/2}}.$$

Un intervalle de niveau de confiance (non asymptotique) $1-\alpha$ basé sur T_1 est donc

$$\left[\frac{T_1}{1 + (12n\alpha)^{-1/2}}; \frac{T_1}{1 - (12n\alpha)^{-1/2}}\right],$$

pour n assez grand pour que $12n\alpha > 1$.

Si on compare ces deux intervalles de confiance sur la base de leurs longueurs, le premier a une longueur en

$$T_2(\alpha^{-1/n} - 1) = T_2\frac{\log(1/\alpha)}{n} + o(1/n),$$

le second une longueur en

$$T_1\left(\frac{1}{1 - (12n\alpha)^{-1/2}} - \frac{1}{1 + (12n\alpha)^{-1/2}}\right) = \frac{T_1}{\sqrt{3n\alpha}} + o(n^{-1/2}).$$

En termes de dépendance en n (et α), le premier est plus précis et c'est moral : un intervalle de confiance basé sur une loi exacte des déviations à la cible sera toujours plus précis qu'un intervalle de confiance basé sur la majoration de ces déviations.

2.4.2 Asymptotique

Nous somme toujours dans le cadre simple $\theta \in \mathbb{R}$. Dans un cadre asymptotique on se place dans un cadre limite où $n \to +\infty$ (où n est la taille d'échantillon).

Definition 2.24 : Intervalle de niveau de confiance asymptotique $1-\alpha$

Dans un modèle i.i.d. $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta\in\Theta})$ et pour $\alpha \in]0,1[$, un intervalle de confiance *asymptotique* de niveau $1-\alpha$ pour θ est un couple de statistiques (T_n^-, T_n^+) tel que

$$\forall \theta \in \Theta \quad \lim_{n\to+\infty} P_\theta^{\otimes n}\left(T_n^- \leq \theta \leq T_n^+\right) \geq 1-\alpha.$$

— Un intervalle de confiance asymptotique est dit **unilatère** si $T_n^- = -\infty$ ou $T_n^+ = +\infty$ une infinité de fois.

— Un intervalle de confiance asymptotique est dit **bilatère** si les deux bornes sont finies sauf éventuellement un nombre fini de fois.

On peut là aussi étendre cette notion au delà de $\mathbb{R}$, et pour des suites de lois P_θ^n plus générales que le cadre i.i.d.. L'intérêt des intervalles de confiance asymptotique est de pouvoir baser des intervalles de confiance sur la loi (asymptotique), ce qui est souvent plus précis que via des inégalités de concentration.
Remarque importante : Un intervalle de confiance (non-asymptotique) sera toujours un intervalle de confiance asymptotique (au sens de cette définition), mais la réciproque est évidemment fausse. En pratique, on se place dans le cadre des intervalles de confiance asymptotique dès lors que l'on regarde une convergence en loi.

Recettes pour les ICA : Classiquement on construit des intervalles de confiance asymptotique à partir de convergence en lois, qui s'obtiennent de deux manières :

1. TCL,
2. à la main.

La construction effective s'opère ensuite à renfort parfois de méthode Δ et de Lemme de Slutsky.

Exemple 2.25 *: Taux d'éclosion des oeufs de pingouins.*

On collecte n oeufs de pingouins fécondés, et on note X_i la variable qui vaut 1 si l'oeuf i éclôt, 0 sinon. On peut modéliser cette expérience par un modèle $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathcal{B}(\theta)^{\otimes n}_{\theta\in]0,1[})$. Le paramètre d'intérêt est alors θ, taux d'éclosion "théorique" de ces oeufs.

Un estimateur sans biais de θ est donné par $T(X_{1:n}) = \bar{X}_n$ (moyenne empirique), qui sous P_θ a pour loi $\mathcal{B}(n,\theta)/n$.

Soit $\alpha > 0$, on cherche un IC de niveau de confiance $1-\alpha$ pour θ. Bienaymé Tchebychev donne

$$P_\theta\left(|T-\theta| \geq t\right) \leq \frac{\theta(1-\theta)}{nt^2} \leq \frac{1}{4nt^2}.$$

On en déduit que $[T \pm t_\alpha^{BT}]$ est un IC au niveau de confiance $1-\alpha$ pour θ, avec $t_\alpha^{BT} = \frac{1}{2\sqrt{n\alpha}}$.

Regardons maintenant ce qui se passe asymptotiquement. Comme $X_1 \in L_2(P_\theta)$, le Théorème Central Limite donne

$$\sqrt{n}\left(T-\theta\right) \rightsquigarrow \mathcal{N}(0, \theta(1-\theta)).$$

Par continuité de la loi limite, on en déduit que

$$\lim_{n\to+\infty} P_\theta\left(\theta \in \left[T \pm \frac{\sqrt{\theta(1-\theta)}}{\sqrt{n}} q_{\alpha/2}\right]\right) = 1-\alpha,$$

où q_u désigne le quantile d'ordre u d'une loi $\mathcal{N}(0,1)$. Comme le terme en $\theta(1-\theta)$ dépend évidemment de θ, l'intervalle ci-dessus n'est pas un intervalle de confiance. On peut y remédier de plusieurs manières.

1. En utilisant le lemme de Slutsky : la loi des grands nombres donne $T \xrightarrow{\mathbb{P}} \theta$, donc

$$\sqrt{n}\frac{(T-\theta)}{\sqrt{T(1-T)}} \rightsquigarrow \mathcal{N}(0,1),$$

   le terme de gauche est alors *quantité pivotale* asymptotique. Cela fournit l'intervalle asymptotique de niveau de confiance $1-\alpha$ suivant : $[T \pm \frac{\sqrt{T(1-T)}}{\sqrt{n}} q_{\alpha/2}]$.
2. En utilisant la méthode Δ : si on arrive à trouver une fonction G différentiable telle que $G'(\theta) = (\theta(1-\theta))^{-\frac{1}{2}}$, alors la méthode Δ donne

$$\sqrt{n}(G(T) - G(\theta)) \rightsquigarrow \mathcal{N}(0,1).$$

   L'intervalle de confiance asymptotique correspondant serait alors $G^{-1}\left([G(T) \pm \frac{q_{\alpha/2}}{\sqrt{n}}]\right)$ (dont on n'est même pas sûrs que ce soit un intervalle).
3. En majorant brutalement le terme de variance : comme $\theta(1-\theta) \leq 1/4$, on en déduit

$$\lim_{n\to+\infty} P_\theta\left(\theta \in \left[T \pm \frac{1}{2\sqrt{n}} q_{\alpha/2}\right]\right) \geq 1-\alpha,$$

   et donc $\left[T \pm \frac{1}{2\sqrt{n}} q_{\alpha/2}\right]$ est un ICA de niveau $1-\alpha$.

Ces trois méthodes permettent de couvrir beaucoup de situation, leur pertinence relève du cas par cas.

Pour comparer les deux intervalles obtenus, regardons la taille d'échantillon minimale requise pour obtenir une précision de 2% avec $\alpha = 0.1$. Si on fait des calculs, on obtient

— **Bienaymé-Tchebychev** : $n \geq 25000$,

— **Approximation normale** : $n \geq 6700$.

Remarque importante : Les deux intervalles obtenus ne sont pas de même nature : le premier est un intervalle de confiance non-asymptotique (valable pour tout n), le second un intervalle de confiance **asymptotique** valable à la limite $n \to +\infty$.

Dans le cadre particulier d'approximation d'une loi binomiale par une loi normale, si $n(\theta) \geq 5$ et $n(1-\theta) \geq 5$, alors les quantiles de $\sqrt{n}(\theta - T)/\sqrt{\theta(1-\theta)}$ et ceux de sa limite coïncident jusqu'au 2eme chiffre après la virgule (inclus).

En pratique, l'utilisation d'un ICA dans ce modèle binomial est considérée comme valide dès lors que $n(\theta \wedge (1-\theta)) \geq 5$.

L'utilisation pratique des intervalles de confiance asymptotique est tributaire de la rapidité de la convergence en loi qui la sous-tend. Pour certains modèles (comme

celui vu en exemple) des conventions existent. Dans un cadre plus général, on peut relier intervalles de confiance asymptotiques et non-asymptotiques en utilisant des résultats de type Berry-Esséen.

Théorème 2.26 : Théorème de Berry-Esséen

Soit $X_1, \dots, X_n$ une suite de variables i.i.d. telles que $\mathbb{E}(X_1) = 0$, $\mathrm{Var}(X_1) = \sigma^2$, et $\mathbb{E}(|X_1|^3) = \kappa < +\infty$. Si on note, pour $t \in \mathbb{R}$,

$$F_n(t) = \mathbb{P}\left(\frac{\sqrt{n}}{\sigma}\bar{X}_n \leq t\right)$$

la "vraie" fonction de répartition, et par Φ la fonction de répartition d'une loi $\mathcal{N}(0,1)$ (celle de la loi limite donc), on a

$$\sup_{t\in\mathbb{R}} |F_n(t) - \Phi(t)| \leq \frac{c\kappa}{\sigma^3\sqrt{n}},$$

où c est une constante numérique (à ce jour valant 0.4748).

Pour conclure cette partie, on comprend maintenant l'importance énorme des quantiles de la loi normale standard dans toutes les applications des statistiques (et du fameux 1.96). En pratique, ces quantiles sont tabulés avec une précision suffisante (avec des tables "manuscrites" ou logiciels). Une méthode plus matheuse d'encadrer ces quantiles se base sur le mini-résultat suivant

Lemme 2.27 : Encadrement des quantiles d'une loi normale standard

Pour $x \in \mathbb{R}$, on note $\phi(x) = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$ (densité d'une $\mathcal{N}(0,1)$), et

$$\bar{\Phi}(x) = \int_x^{+\infty} \phi(t)dt,$$

aussi appelée fonction de survie (de la loi $\mathcal{N}(0,1)$). On a alors, pour $x \geq 1$,

$$\frac{\phi(x)}{x}\left(1 - \frac{1}{x^2}\right) \leq \bar{\Phi}(x) \leq e^{-\frac{x^2}{2}} \wedge \frac{\phi(x)}{x}.$$

Démonstration. C'est essentiellement de l'intégration par parties. D'une part, on a

$$\begin{aligned}
\bar{\Phi}(x) &= \int_x^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}}dt \\
&= \left[\frac{-e^{-t^2/2}}{\sqrt{2\pi}t}\right]_x^{+\infty} - \int_x^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}t^2}dt \\
&\leq \frac{\phi(x)}{x}.
\end{aligned}$$

D'autre part, une inégalité de Markov donne

$$\bar{\Phi}(x) = \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda x}\right) \leq e^{-\lambda x + \frac{\lambda^2}{2}},$$

en choisissant $\lambda = x$ on obtient l'autre majoration de $\bar{\Phi}(x)$.

Pour la minoration, reprenons la précédente IPP, et remarquons que

$$\begin{aligned}\int_x^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}t^2}dt &= \left[\frac{-e^{-t^2/2}}{\sqrt{2\pi}t^3}\right]_x^{+\infty} - \int_x^{+\infty} \frac{3e^{-t^2/2}}{\sqrt{2\pi}t^4}dt \\ &\leq \frac{\phi(x)}{x^3}.\end{aligned}$$

□

Terminons sur un exemple où la convergence en loi ne découle pas du TCL.

Exemple 2.28 *: Lois uniformes encore.*

Dans le modèle $\mathcal{U}(]0,\theta[^{\otimes n}$, si on veut construire un intervalle de confiance **asymptotique** basé sur $T_2 = \max_{i=1,\dots,n} X_i$. Il s'agit de regarder la convergence en loi de T_2, ce qui ne peut pas s'obtenir via le théorème central limite. Il faut donc regarder la convergence en loi de T_2 autour de θ "manuellement".

Notons F_n la fonction de répartition de T_2. On a immédiatement $F_n(t) = 0$ pour tout $t \leq 0$ et $F_n(t) = 1$ si $t \geq \theta$. Soit maintenant $0 < t < \theta$, on a

$$\begin{aligned}F_n(t) = P_\theta(T_2 \leq t) &= P_\theta\left(\bigcap_{i=1}^n \{X_i \leq t\}\right) \\ &= \left(\frac{t}{\theta}\right)^n \qquad \text{(indépendance des } X_i\text{)}.\end{aligned}$$

Regardons maintenant les fluctuations de T_2 autour de θ. On sait d'une part que $T_2 \leq \theta$„ et d'autre part

$$P_\theta(\theta - T_2 \geq t) = \left(1 - \frac{t}{\theta}\right)^n \qquad \forall t \in [0,\theta].$$

En renormalisant correctement pour faire apparaître une convergence, pour $t \geq 0$ on obtient

$$\begin{aligned}P_\theta(n(\theta - T_2) \geq t) &= \left(1 - \frac{t}{n\theta}\right)^n \mathbb{1}_{t \leq n\theta} \\ &\underset{n\to+\infty}{\longrightarrow} \exp -(t/\theta).\end{aligned}$$

On en déduit alors que $n(\theta - T_2) \rightsquigarrow \mathcal{E}(1/\theta)$ (loi exponentielle de paramètre $1/\theta$). Maintenant, si $\alpha \in]0,1[$, on déduit

$$P_\theta(n(\theta - T_2) \geq \theta \log(1/\alpha)) \underset{n\to+\infty}{\longrightarrow} \alpha.$$

Pour obtenir un intervalle de confiance il faut retravailler l'inégalité :

$$P_\theta(n(\theta - T_2) \geq \theta \log(1/\alpha)) = P_\theta(\theta \geq \frac{T_2}{1 - n^{-1}\log(1/\alpha)},$$

pour n assez grand (ce qui est le cas dans un cadre asymptotique). Un intervalle de niveau de confiance **asymptotique** $1-\alpha$ est alors donné par

$$\left[T_2; \frac{T_2}{1 - n^{-1}\log(1/\alpha)}\right],$$

qui est de longueur

$$\frac{T_2 \log(1/\alpha)}{n} + o(1/n),$$

c'est à dire du même ordre que l'intervalle obtenu dans un cadre non asymptotique en se basant sur la loi de $M_n = \max_{i=1,\dots,n} X_i$.

Dans ce cas précis on ne gagne rien à passer dans un cadre asymptotique : si on connaît la loi exacte des fluctuations de l'estimateur autour de la cible, l'intervalle qui en découlera sera forcément "optimal", même dans un cadre asymptotique.

2.5 Tests

Le point de vue des tests est moins naturel que celui de l'estimation. Plutôt que d'estimer θ, on cherche plutôt à répondre à une question binaire dessus. Dans le cadre de l'estimation du taux d'éclosion des oeufs de pingouins, on s'intéresse à la question "est-il plus grand que 1/2" plutôt qu'à son estimation.

Evidemment, si on peut donner des garanties en estimation (via des intervalles de confiance par exemple), on pourra donner des garanties en test et la réciproque est fausse. En ce sens, tester est plus "facile" qu'estimer.

Formellement, une question binaire sur θ revient à choisir deux sous ensemble Θ_0 et Θ_1 de Θ. On parle alors *d'hypothèses*

$$\begin{aligned} H_0 &: \theta \in \Theta_0, \\ H_1 &: \theta \in \Theta_1. \end{aligned}$$

Par convention H_0 est appelée *hypothèse nulle*, et H_1 *hypothèse alternative*. On verra plus bas que ces deux rôles ne sont pas symétriques.

Tester revient donc à estimer $g : \theta \mapsto \mathbb{1}_{\theta \in \Theta_1}$, dès lors un test est juste un estimateur de cette quantité.

Definition 2.29 : Test

Dans un modèle $(\mathcal{X}^n, \mathcal{A}, (P_\theta^{\otimes n})_{\theta \in \Theta})$, un test T est une fonction mesurable de $\mathcal{X}^n$ dans $\{0, 1\}$.

Par convention toujours, lorsque la sortie du test est 0, on dit qu'on *accepte* (sous-entendu l'hypothèse nulle), tandis que $T = 1$ correspond au *rejet* de cette hypothèse.

Comme pour les estimateurs, il s'agit maintenant de mesurer la qualité d'un test, au vu des deux hypothèses que l'on cherche à discriminer. Une approche naturelle est d'équiper $\{0, 1\}$ avec une distance, par exemple $\mathbb{1}_{y \neq y'}$, et à calculer le risque d'un test défini par

$$R_T(\theta) = E_\theta \left(\mathbb{1}_{T(X_{1:n}) \neq g(\theta)} \right) = P_\theta(T(X_{1:n}) \neq g(\theta)),$$

qui quantifie la probabilité que notre test se trompe sous P_θ. Plutôt que de regarder $\sup_\theta R_T(\theta)$ (qui donnerait la même importance aux erreurs sous H_0 et sous H_1), on distingue les erreurs maximales sous les deux hypothèses.

Definition 2.30 : Erreurs de première et seconde espèce, puissance

Dans un modèle $(\mathcal{X}^n, \mathcal{A}, (P_\theta^{\otimes n})_{\theta\in\Theta})$, pour un test T et deux hypothèses Θ_0 et Θ_1, on définit

- *l'erreur de première espèce* de T : $\sup_{\theta\in\Theta_0} P_\theta(T(X_{1:n}) = 1)$ (probabilité max de rejeter à tort),
- *l'erreur de seconde espèce* de T : $\sup_{\theta\in\Theta_1} P_\theta(T(X_{1:n}) = 0)$.

On parle aussi de *puissance* (minimale) d'un test :

$$\inf_{\theta\in\Theta_1} P_\theta(T(X_{1:n}) = 1),$$

qui est juste 1 moins l'erreur de seconde espèce.

La plupart du temps, un type d'erreur est plus "grave" que l'autre. Dans le cadre d'un test qui prend en entrée divers paramètres d'un patient (par exemple le résultat d'un sondage nasal) et dont la sortie est 0 (patient sain) ou 1 (patient malade), les faux négatifs (patients malades dont le test indique la santé) sont plus préoccupants que les faux négatifs (patients sains que le test détecte comme malade). Par convention, un *test de niveau* α est un test qui contrôle l'erreur la plus grave.

Definition 2.31 : Test de niveau α

Dans un modèle $(\mathcal{X}^n, \mathcal{A}, (P_\theta^{\otimes n})_{\theta\in\Theta})$, pour deux hypothèses Θ_0 et Θ_1, et un paramètre $\alpha \in [0, 1]$, un test T est dit de niveau α si son erreur de première espèce est majorée par α, c-à-d

$$\sup_{\theta\in\Theta_0} P_\theta(T(X_{1:n}) = 1) \leq \alpha.$$

Le fait qu'un test soit de niveau α ne dépend que de son comportement sous H_0, le fait qu'il arrive à correctement détecter H_1 est de ce point de vue secondaire. Evidemment, le but va être, sous une contrainte de niveau α, de trouver des tests les plus puissants possibles. Comme exemples limites de tests de niveau α aveugles à H_1, on peut citer le test nul $T \equiv 0$, qui est de niveau 0 (mais aussi de puissance nulle), ou alors un test purement aléatoire, de loi $\mathcal{B}(\alpha)$ indépendante des observations (qui de niveau α et de puissance α).

La plupart des tests utilisés sont calibrés par leur niveau. On se rend compte alors que la seule certitude que l'on peut avoir à l'issue d'un tel test est dans le cas d'un rejet de H_0 (lorsque $T = 1$) : dans ce cas la probabilité de faire une erreur est majorée par α. En revanche, il n'y a aucune garantie sur l'erreur faite lorsque l'on accepte H_0.

Le choix (dissymétrique) des hypothèses H_0 et H_1 est alors crucial en pratique. De manière informelle, il faut mettre en H_0 le contraire de ce que l'on cherche à prouver.

Exemple 2.32 *: Oeufs de pingouins suite.*

Dans le modèle où on observe l'éclosion ou non de n oeufs de pingoins, on peut se poser la question de savoir si la fonte des glaces à un effet sur le taux d'éclosion θ, en supposant que le taux d'éclosion normal est $\theta = 1/2$.

Si vous êtes dans la pétrochimie et cherchez à prouver que non, le taux d'éclosion reste normal, il vous faudra prendre $H_0 : \theta < 1/2$ et $H_1 : \theta = 1/2$.

À l'inverse, si vous êtes plutôt écologiste et cherchez à prouver que ce taux d'éclosion a diminué, il vous faudra prendre $H_0 : \theta = 1/2$ et $H_1 : \theta < 1/2$.

2.5.1 Recettes de construction de test

On verra juste après un test classique, basé sur la vraisemblance. Si vous devez construire "manuellement" un test de niveau α pour les hypothèses Θ_0 et Θ_1 dans le modèle $(\mathcal{X}^n, \mathcal{A}, (P_\theta^{\otimes n})_{\theta\in\Theta}$, on procède généralement comme suit :

1. Choisir une statistique S censée bouger suivant les deux hypothèses (contre ex : $P_\theta = \mathcal{N}(\theta, 1)^{\otimes n}$, $S = \sum_i (X_i - \bar{X}_n)^2$).
2. Choisir a priori la forme d'une *région de rejet* R_α en fonction de l'alternative : $S \in R_\alpha$ correspondra à la valeur 1 du test, c-a-d on posera $T = \mathbb{1}_{S\in R_\alpha}$. (ex dans le même cas : si $H_1 : \theta \geq 10$ et $H_0 : \theta < 10$, $S = \bar{X}_n$, on prendra naturellement R_α du type $[t_\alpha, +\infty[$.
3. Calibration de R_α de manière à avoir

$$\sup_{\theta\in\Theta_0} P_\theta(S(X_{1:n}) \in R_\alpha) \leq \alpha,$$

idéalement avec égalité. Cela se fait souvent à l'aide d'une *quantité pivotale* (comme en IC), c'est à dire en trouvant une transformation g_θ de S dont la loi ne dépend pas de θ : on calibre alors $\sup_{\theta\in\Theta_0} \mathbb{P}(g_\theta(S(X_{1:n})) \in g_\theta(R_\alpha)) \leq \alpha$ (calibration en $g_\theta(R_\alpha)$, la loi de $g_\theta(S(X_{1:n}))$ étant fixe). On peut aussi utiliser de la domination stochastique (hors programme).

Exemple 2.33 *: Oeufs de pingouins, fin.* Dans le modèle d'éclosion de n oeufs donné par $(\{0,1\}^n, \mathcal{P}(\{0,1\}^n), \mathcal{B}(\theta)^{\otimes n}_{\theta\in]0,1[})$.

Point de vue écolo : On cherche à prouver que la fonte des glaces a réduit le taux déclosion des oeufs par rapport à la normale 1/2. On a alors

$$\begin{array}{rcl} H_0 & : & \theta = 1/2, \\ H_1 & : & \theta < 1/2. \end{array}$$

On peut prendre comme statistique de test $S = \sum_{i=1}^n X_i$. Sous H_1 on s'attend à ce que S soit petite, donc on pose $R_\alpha = [0, t_\alpha]$. Comme

$$P_{\theta_0}(S(X_{1:n}) \in R_\alpha) = \mathbb{P}\left(\mathcal{B}(n, 1/2) \leq t_\alpha\right),$$

si on prend comme $t_\alpha + 1$ le quantile d'ordre α d'une $\mathcal{B}(n, 1/2)$, $T = \mathbb{1}_{S\leq t_\alpha}$ est un test de niveau α.

Point de vue pétrolier : On cherche à prouver que le taux d'éclosion des oeufs est resté à la normale. Cela correspond au choix d'hypothèses

$$\begin{array}{rcl} H_0 & : & \theta = 1/2, \\ H_1 & : & \theta > 1/2. \end{array}$$

La statistique de test reste la même, en revanche on prendra plutôt comme région de rejet $R_\alpha = [t_\alpha, n]$. Il faut maintenant calibrer t_α de sorte que

$$P_(1/2)(S(X_{1:n}) \geq t_\alpha) \leq \alpha.$$

En choisissant t_α le quantile d'ordre $1-\alpha$ d'une $\mathcal{B}(n,1/2)$, on a un test de niveau α. On peut remarquer alors que la puissance aussi est majorée par α (c'est souvent le cas lorsque les deux hypothèses sont contigües).

Dans l'exemple précédent, on remarque que à chaque fois H_0 est de la forme $\{\theta_0\}$ (un singleton), ce qui est plus facile pour calibrer le seuil via $P_{\theta_0}(S(X_{1:n} \in R_\alpha) \leq \alpha$. Il aurait été plus intuitif de prendre, par exemple dans le cas pétrolier

$$\begin{array}{rcl} H_0 & : & \theta < 1/2, \\ H_1 & : & \theta \geq 1/2. \end{array}$$

Mais dans ce cas il aurait fallu calculer

$$\sup_{\theta<1/2} P_\theta(S \geq t_\alpha).$$

Dans ce cas précis il n'est pas difficile d'avoir l'intuition que le supremum à gauche est atteint lorsque la probabilité de succès est maximale, c'est à dire

$$\sup_{\theta<1/2} P_\theta(S \geq t_\alpha) = P_{1/2}(S \geq t_\alpha).$$

On peut le montrer par le calcul, ou en utilisant un argument de couplage. Toujours est-il que ce genre de méthode renvoie à la notion de *domination stochastique*, qui permet de relier

$$\sup_{\theta\in\Theta_0} P_\theta(S \in R_\alpha)$$

à $P_{\theta_0}(S \in R_\alpha)$, où θ_0 est au bord de Θ_0.

La notion de domination stochastique dépassant le cadre de ce cours, on se cantonnera (sauf mention contraire) au cas où $\Theta_0 = \{\theta_0\}$ (ce qui revient à assimiler une hypothèse nulle à son bord, la domination stochastique permet de le justifier proprement).

2.5.2 p-valeur

En pratique, lorsque vous lancez un test à partir d'un jeu de données via `R` par exemple, l'issue de la procédure est une p-valeur, à partir de laquelle vous allez décider de rejeter (ou accepter). On peut donner une définition précise de la notion de p-valeur. Dans un cadre appliqué une définition informelle suffit.

Definition 2.34 : p-valeur, définition informelle

> Une p-valeur est une statistique à valeurs dans $[0,1]$, telle que $p(X_{1:n})$ représente "la probabilité de rejeter à tort H_0 en se basant sur $X_{1:n}$".

Une autre manière de l'énoncé peut être : "pour un test donné, la p-valeur correspond au plus petit niveau conduisant au rejet de H_0 en se basant sur $X_{1:n}$". Dès lors, si une p valeur $p(X_{1:n})$ est plus petite que α, pour un α donné, alors un test de niveau α rejetterait H_0 (vu que la p-valeur est "le plus petit niveau qui permet de rejeter").

Ces énoncés peuvent être formalisé un peu plus proprement dans un contexte où les régions de rejet sont croissantes en le niveau.

LEMME 2.35

Si $p(X_{1:n})$ est une p-valeur, alors

$$T = \mathbb{1}_{p(X_{1:n}) \leq \alpha}$$

est un test de niveau α.

"Réciproquement", si $(R_\alpha)_{\alpha \in [0,1]}$ est une famille croissante de régions de rejet associées à une famille de tests de niveau α, alors

$$\hat{\alpha} : X_{1:n} \mapsto \inf\{\alpha > 0 \mid S(X_{1:n}) \in R_\alpha\}$$

est une p-valeur.

Cet énoncé correspond aux deux cas de figure que vous rencontrerez en pratique :

1. Soit vous effectuez un test via un logiciel qui vous ressort une p-valeur p. Dans ce cas, pour décider de l'issue du test, vous comparez cette p-valeur avec votre seuil α défini au préalable :
   - Si $p \leq \alpha$, vous rejetez H_0 (et vous en êtes sûr avec probabilité plus grande que $1 - \alpha$).
   - Si $p > \alpha$, vous ne pouvez qu'accepter H_0, sans garanties.
2. Soit on vous demande de construire une p-valeur associée à un type de test et à des observations $X_{1:n}$. Dans ce cas vous construisez S et R_α comme dans la partie précédente, et vous prenez comme p-valeur le plus petit α tel que $S(X_{1:n}) \in R_\alpha$ (ça sera votre p-valeur).

Exemple 2.36.

Si on reprend l'exemple des oeufs de pingouins, du point de vue écolo, avec $n = 1000$, en observant S, alors

1. $[0, S]$ est la plus petite région de rejet contenant S,
2. elle correspond au niveau de test $F_0(S)$, où F_0 est la fonction de répartition d'une $\mathcal{B}(1000, 1/2)$.

On en déduit que $F_0(S)$ est une p-valeur.

Application : si on observe $S_{obs} = 460$ oeufs éclos, la p-valeur correspondante est de $0,6\%$. Le point de vue écolo semble alors fondé.

2.5.3 Tests dans un cadre asymptotique

Comme pour les intervalles de confiance, il existe des notions de niveau asymptotique et de consistance pour les tests. On se place ici dans le modèle i.i.d. $(\mathcal{X}^n, \mathcal{A}^{\otimes n}, (P_\theta^{\otimes n})_{\theta \in \Theta})$. La notion de consistance est la celle usuelle pour l'estimation de $\mathbb{1}_{\theta \in \Theta_1}$ (c-à-d on veut que $T \xrightarrow{\mathbb{P}} \mathbb{1}_{\theta \in \Theta_1}$ pour tout θ). La notion de niveau asymptotique est un peu similaire à celle du niveau de confiance asymptotique pour les IC.

Definition 2.37

Pour une hypothèse nulle $\Theta_0 \subset \Theta$ et $\alpha \in [0,1]$, un test T est dit de niveau *asymptotique* α si

$$\sup_{\theta\in\Theta_0} \lim \sup_{n\to+\infty} P_\theta^{\otimes n}(T(X_{1:n}) = 1) \leq \alpha.$$

La recette de construction d'un test de niveau asymptotique α est la même que dans le cadre général, avec une étape de convergence en loi la plupart du temps.

Exemple 2.38 *: Oeufs de pingouins encore.*

Dans le modèle d'éclosion des oeufs de pingouins, testé d'un point de vue pétrolier, on cherche toujours un test de la forme $S \geq t_\alpha$, mais on va calibrer le seuil de manière asymptotique.

Pour ce faire, il nous faut **une suite** $t_{\alpha,n}$ telle que

$$\lim \sup_{n\to+\infty} P_{1/2}(S \geq t_{\alpha,n}) \leq \alpha.$$

Or, sous $P_{1/2}$, $\frac{2(S-n/2)}{\sqrt{n}} \rightsquigarrow \mathcal{N}(0,1)$. Si on note $q_{1-\alpha}$ le $1-\alpha$-quantile d'une $\mathcal{N}(0,1)$, on a

$$\lim_{n\to\infty} P_{1/2}\left(\frac{2(S-n/2)}{\sqrt{n}} \geq q_{1-\alpha}\right) = \alpha,$$

et donc, $t_{\alpha,n} = n/2 + \sqrt{n}q_{1-\alpha}/2$ convient, et le test

$$T = \mathbb{1}_{S\geq t_{\alpha,n}}$$

est un test de niveau **asymptotique** α.

Pour montrer la consistance de ce test, on se place sous H_1, c'est à dire pour un $\theta > 1/2$. On a alors

$$\begin{aligned} P_\theta(T(X_{1:n}) = 1) &= P_\theta\left(\frac{2(S-n/2)}{\sqrt{n}} \geq q_{1-\alpha}\right) \\ &= P_\theta\left(\frac{2(S-n\theta)}{\sqrt{n}} + \frac{2n(\theta-1/2)}{\sqrt{n}} \geq q_{1-\alpha}\right). \end{aligned}$$

Or, le TCL montre que $Z_n = 2(S-n\theta)/\sqrt{n}$ converge en loi sous P_θ vers une variable Z sur $\mathbb{R}$. Par ailleurs $2\sqrt{n}(\theta-1/2)$ tend vers $+\infty$. On a alors

$$P_\theta(T(X_{1:n}) = 1) = P_\theta\left(\frac{2(S-n\theta)}{\sqrt{n}} \geq u_n\right),$$

où $u_n \underset{n\to+\infty}{\longrightarrow} -\infty$. Le Théorème Central Limite Uniforme (Proposition **??**) mène alors à

$$\begin{aligned} P_\theta(T(X_{1:n}) = 1) &\geq P_\theta(Z \geq u_n) - \|F_{Z_n} - F_Z\|_\infty \\ &\underset{n\to+\infty}{\longrightarrow} 1 - 0. \end{aligned}$$

Chapitre 3

Deux modèles classiques (et méthodes qui vont avec)

3.1 Modèle lin gaussien

3.2 Modèles multinomiaux