Leçon 21

Intervalle de confiance ; Statistique d'ordre

1. Intervalle de confiance : un premier exemple

2. Théorème central limite

3. Échantillon gaussien

4. Estimateur statistique

5. Statistique d'ordre

6. Découverte : Spectre de matrice aléatoire

Exercices

1 Intervalle de confiance : un premier exemple

Soit une pièce de monnaie dont la probabilité p de faire pile est inconnue (il est néanmoins supposé que $p \in]0,1[$) ; afin de tester si elle est équilibrée ($p = \frac{1}{2}$), la pièce est lancée un million de fois ; le résultat est un million de pile. Quelle attitude adopter ?

La probabilité de faire pile un million de fois est égale à p^{10^6}, qui est positive et donc l'événement peut tout à fait avoir lieu. De plus, si $p = \frac{1}{2}$, toute autre suite de pile ou de face de taille un million aura la même probabilité. Maintenant, si ces événements sont équiprobables, il n'en demeure pas moins qu'il y a plus de chances d'espérer une *proportion* à peu près égale (si la pièce est équilibrée) de pile et de face, tout simplement parce qu'il y a plus de façons d'atteindre un telle proportion. Comme dans un jeu avec deux dés, si faire $(5,6)$ et $(3,4)$ ont la même probabilité ($\frac{1}{36}$ si les dés sont distingués, $\frac{1}{18}$ sinon), il y a beaucoup plus de combinaisons possibles fournissant une somme égale à 7 qu'une somme égale à 11.

Suite à une expérience produisant un million de fois pile, tout possible soit-elle, le statisticien, ou l'homme de bon sens (un probabiliste aussi !), se demandera ainsi si la pièce n'est pas biaisée, avec donc p proche de 1 pour favoriser l'apparition de pile. Comme le raisonnement probabiliste ne permet pas d'affirmer qu'il n'est pas possible de produire un million de pile, la conclusion ne peut être que probabiliste elle-même ! À savoir, « il y a de fortes chances » que la pièce soit biaisée, et la probabilité qu'elle le soit peut être quantifiée (à partir de l'intuition découlant de la loi des grands nombres que la proportion de pile ou de face est à peu près la même si la pièce est équilibrée). C'est l'objet de la notion d'intervalle de confiance.

Pour formaliser l'exemple précédent, considérer, sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, un *échantillon* $(X_1, \ldots, X_n)$ de n variables aléatoires indépendantes équidistribuées, de même loi de Bernoulli $\mathcal{B}(p)$ sur $\{0,1\}$ de paramètre de suc-

cès ($= 1$) $p \in]0,1[$ (inconnu donc). Comme $\mathbb{E}(X_1) = p$, d'après la loi des grands nombres, un *estimateur statistique* efficace pour tester le paramètre p est l'estimateur de la moyenne

$$\overline{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k.$$

En effet, comme décrit dans la Leçon 19, l'inégalité de Tchebychev fournit une forme faible de loi des grands nombres concluant à la convergence en probabilité de $\overline{X}_n$ vers la moyenne p. Plus précisément, comme $\mathbb{E}(\overline{X}_n) = p$ et $\mathrm{Var}(\overline{X}_n) = \frac{1}{n}\,p(1-p)$, elle indique que pour tout $\varepsilon > 0$,

$$\mathbb{P}\big(|\overline{X}_n - p| \geq \varepsilon\big) \leq \frac{1}{\varepsilon^2}\,\mathrm{Var}\big(\overline{X}_n\big) = \frac{p(1-p)}{\varepsilon^2 n}$$

(qui tend donc vers 0 quand n tend vers l'infini). En langage statistique, la *moyenne empirique* $\overline{X}_n$ converge vers la *moyenne théorique* $\mathbb{E}(X_1) = p$.

La borne précédente peut être mise à profit pour quantifier la convergence. Dans un premier temps, il peut être utiliser simplement que $p(1-p) \leq \frac{1}{4}$ de sorte que

$$\mathbb{P}\big(|\overline{X}_n - p| \geq \varepsilon\big) \leq \frac{1}{4\varepsilon^2 n}\,.$$

Un *seuil* $\mathrm{s_e}$, *ou niveau de confiance*, est alors fixé, sous la forme (conventionelle) $\mathrm{s_e} = 1-\kappa > 0$, typiquement $\kappa = 5\%$ ou $\kappa = 1\%$ de sorte que $\mathrm{s_e} = 1-\kappa = 95\%$ ou $\mathrm{s_e} = 1-\kappa = 99\%$. Choisir ainsi $\varepsilon > 0$ tel que $\frac{1}{4\varepsilon^2 n} = \kappa$, autrement dit $\varepsilon = \frac{1}{2\sqrt{\kappa n}}$, pour obtenir que $\mathbb{P}(|\overline{X}_n - p| \geq \varepsilon) \leq \kappa$, soit par passage au complémentaire

$$\mathbb{P}\big(|\overline{X}_n - p| < \varepsilon\big) \geq 1-\kappa = \mathrm{s_e}.$$

La notion d'*intervalle de confiance* provient alors du retournement de point de vue dans la description de l'événement $\{|\overline{X}_n - p| < \varepsilon\}$, à savoir

$$\big\{\omega \in \Omega\,;\, \big|\overline{X}_n(\omega) - p\big| < \varepsilon\big\} = \Big\{\omega \in \Omega\,;\, p \in \big]\overline{X}_n(\omega) - \varepsilon\,, \overline{X}_n(\omega) + \varepsilon\big[\Big\}.$$

Ainsi, et puisque $\varepsilon = \frac{1}{2\sqrt{\kappa n}}$, en désignant par I l'intervalle (aléatoire)

$$\left]\overline{X}_n - \frac{1}{2\sqrt{\kappa n}}\,,\overline{X}_n + \frac{1}{2\sqrt{\kappa n}}\right[,$$

il vient

$$\mathbb{P}(p \in I) \;\geq\; 1-\kappa \;=\; \mathrm{s_e}.$$

L'intervalle $I = \left]\overline{X}_n - \frac{1}{2\sqrt{\kappa n}}\,\overline{X}_n + \frac{1}{2\sqrt{\kappa n}}\right[$ est appelé *intervalle de confiance au niveau* $\mathrm{s_e} = 1-\kappa$ *pour l'estimateur de la moyenne* $\overline{X}_n$. Les formules indiquent que l'intervalle de confiance est d'autant plus précis (petit) que la taille n de l'échantillon est grand. Un intervalle de confiance peut être ouvert ou fermé.

De façon pratique, si $n = 10^6$ et $\kappa = 5\%$, un résultat de un million de pile, autrement dit $\overline{X}_n = 1$, fournit un intervalle de confiance approché de la forme $]0,9977, 1,0023[$ (il est à noter que le côté droit de l'intervalle n'est pas pertinent et doit être remplacé par 1 puisque $p \in]0,1[$). À l'issue donc d'une suite de lancers produisant un million de fois pile, la seule conclusion raisonnable est d'affirmer qu'il y a 95% de chance que la pièce soit biaisée, et que son paramètre de succès p soit entre $0,9977$ et 1. Au seuil $\mathrm{s_e} = 1-\kappa = 99\%$, le paramètre de succès p serait estimé entre $0,995$ et 1. Cette affirmation ne contredit pas le point de vue probabiliste de la probabilité positive de la réalisation d'un million de pile, mais quantifie la forte possibilité que la pièce ne soit pas équilibrée.

Le caractère aléatoire de l'intervalle de confiance

$$I(\omega) \;=\; \left]\overline{X}_n(\omega) - \frac{1}{2\sqrt{\kappa n}}\,,\overline{X}_n(\omega) + \frac{1}{2\sqrt{\kappa n}}\right[$$

prend son sens dans un sondage par exemple. Dans un sondage ω à réponse binaire (0 ou 1) de n individus, la moyenne (empirique) ou fréquence $\overline{X}_n(\omega)$ représente la proportion d'individus répondant 1, et l'intervalle de confiance décrit une estimation au seuil $\mathrm{s_e} = 1-\kappa$ autour de la fréquence obtenue. Bien entendu, plus le nombre de personnes sondées est grand, plus la précision

est accrue. Il est aussi parfois nécessaire de répéter les sondages afin d'exclure les valeurs aberrantes (comme dans l'exemple précédent). Si A est l'événement $A = \{\omega \in \Omega\,;\, p \in I(\omega)\}$, de probabilité donc $\geq 1-\kappa$, un aléa $\omega \in \Omega$ appartient à A avec probabilité au moins $1-\kappa$ et à son complémentaire avec probabilité au plus κ. Donc, même s'il n'y a qu'une expérience (sondage) ω, c'est ainsi qu'il faut entendre le niveau de confiance $\mathrm{s_e} = 1-\kappa$.

L'analyse précédente sur le modèle de Bernoulli peut être précisée quelque peu en évitant la majoration, qui peut être grossière, $p(1-p) \leq \frac{1}{4}$. Il suffit en fait de conserver l'expression $p(1-p)$ et d'effectuer une analyse sur un polynôme du second degré (en p). En effet, si $\kappa = \frac{p(1-p)}{\varepsilon^2 n}$, l'inégalité $|\overline{X}_n - p| < \varepsilon$ prend la forme

$$\big(\overline{X}_n - p\big)^2 \,<\, \varepsilon^2 \,=\, \frac{p(1-p)}{\kappa n}\,,$$

autrement dit

$$(1+\kappa n)p^2 - \big(2\kappa n\overline{X}_n + 1\big)p + \kappa n\overline{X}_n^2 \,<\, 0.$$

La résolution de cette inéquation du second degré en la variable p indique que $p \in]p_-, p_+[$ avec

$$p_\pm \,=\, \frac{\overline{X}_n + \frac{1}{2\kappa n} \pm \sqrt{\frac{1}{4\kappa^2 n^2} + \frac{1}{\kappa n}\,\overline{X}_n(1-\overline{X}_n)}}{1+\frac{1}{\kappa n}}\,.$$

(Bien entendu, ces valeurs sont à restreindre à $]0,1[$.) L'intervalle de confiance au niveau $\mathrm{s_e} = 1-\kappa$ est ainsi $I =]p_-, p_+[$. Pour l'expérience $n = 10^6$ avec $\overline{X}_n = 1$ au seuil $\kappa = 1\%$, p_- est de l'ordre de $0,9999$.

Il faut observer que quand n devient grand, le terme d'erreur pertinent (autour de $\overline{X}_n$) dans l'expression de $p_\pm$ est $\sqrt{\frac{1}{\kappa n}\,\overline{X}_n(1-\overline{X}_n)}$, donc d'ordre $\frac{1}{\sqrt{n}}$.

L'Exercice 1 présente une étude similaire lorsque l'échantillon $(X_1, \ldots, X_n)$ suit une loi commune exponentielle $\mathcal{E}(\alpha)$ de paramètre $\alpha > 0$ inconnu ou une loi uniforme $\mathcal{U}(0,a)$, $a > 0$.

2 Théorème central limite

Même si cet aspect ne sera pas développé, il est intuitivement important de comprendre que les facteurs $\sqrt{n}$ apparaissant dans les intervalles de confiance précédents sont directement liés au théorème central limite (Leçon 20). En effet, d'après ce dernier, et pour l'échantillon de variables de Bernoulli précédent,

$$\sqrt{\frac{n}{p(1-p)}}\,(\overline{X}_n - p)$$

converge en loi quand n tend vers l'infini vers une variable aléatoire G de loi $\mathcal{N}(0,1)$. En particulier, d'après la convergence des fonctions de répartition,

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}\,|\overline{X}_n - p| > t\right) \;=\; \mathbb{P}\big(|G| > t\big)$$

pour tout $t > 0$ (la fonction de répartition d'une loi normale est continue).

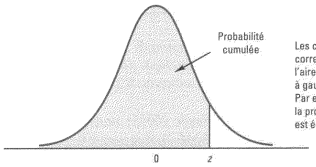
Or la loi $\mathcal{N}(0,1)$ a pour propriété que la probabilité $\mathbb{P}(|G| > t)$ est assez petite même pour des valeurs de $t > 0$ modestes (voir par exemple les encadrements de l'Exercice 1, Leçon 14). Pour s'en rendre compte, par symétrie,

$$\mathbb{P}\big(|G| > t\big) \;=\; \mathbb{P}(G > t) + \mathbb{P}(-G > t) \;=\; 2\big(1 - \Phi(t)\big)$$

où

$$\Phi(t) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}x^2} dx, \quad t \in \mathbb{R},$$

est la fonction de répartition de la loi normale $\mathcal{N}(0,1)$. Des *tables de loi normale*



Les chiffres de la table correspondent à la valeur de l'aire située sous la courbe à gauche de la valeur z. Par exemple, pour $z = 1{,}25$ la probabilité cumulée est égale à 0,8944.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9913
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9986	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

évaluent $\Phi(t)$ pour de nombreuses valeurs de t. Il peut y être lu par exemple que pour t de l'ordre de 2 ($t = 1,96$), $2(1 - \Phi(t))$ est plus petit que 5%.

Si donc la probabilité

$$\mathbb{P}\left(\sqrt{\frac{n}{p(1-p)}}\,\left|\overline{X}_n - p\right| > t\right)$$

est proche de $\mathbb{P}(|G| > t)$, pour t de l'ordre de 2 elle sera plus petite que 5%. Ainsi

$$\mathbb{P}\left(\left|\overline{X}_n - p\right| \leq 2\sqrt{\frac{p(1-p)}{n}}\right) \geq 0,95.$$

Ceci revient, dans les notations du paragraphe précédent, à considérer $\varepsilon = 2\sqrt{\frac{p(1-p)}{n}}$, et donc à formellement remplacer κ par $\frac{1}{4}$ dans les intervalles de confiance (donc des intervalles plus étroits).

Maintenant, tous ces arguments sont sujets à une bonne approximation des probabilités dans le théorème central limite, question chargée de risque et amplement étudiée notamment en statistiques. Le théorème de Berry-Esseen (Leçon 20) est un des outils permettant de quantifier les erreurs.

3 Échantillon gaussien

Ainsi qu'il est connu, une variable aléatoire de loi normale $\mathcal{N}(m, \sigma^2)$ est caractérisée, dans la famille des lois gaussiennes, par ses paramètres de moyenne m et de variance σ^2 (Leçon 14).

Soit alors un échantillon $(X_1, \ldots, X_n)$ de variables aléatoires gaussiennes indépendantes de même variance σ^2 *connue* et de moyenne m *inconnue*. En rappelant $\overline{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$, il peut être observé que

$$\frac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$$

suit une loi $\mathcal{N}(0,1)$, car $X_1 + \cdots + X_n$ a pour loi $\mathcal{N}(nm, n\sigma^2)$ d'après les propriétés d'addition des variables gaussiennes indépendantes, et donc $\overline{X}_n$ a pour loi $\mathcal{N}(m, \frac{\sigma^2}{n})$; l'affirmation s'ensuit après translation et dilatation.

Un intervalle de confiance de la moyenne m au niveau $\mathrm{s_e} = 1 - \kappa > 0$ s'obtient en posant

$$\mathbb{P}\big(|\overline{X}_n - m| \leq \varepsilon\big) \;=\; 1 - \kappa$$

car en renversant les inégalités

$$\mathbb{P}\big(m \;\in\; \big[\,\overline{X}_n - \varepsilon, \overline{X}_n + \varepsilon\big]\big) \;=\; 1 - \kappa.$$

Afin de déterminer $\varepsilon > 0$ en fonction de κ, et donc l'intervalle de confiance au seuil $\mathrm{s_e} = 1 - \kappa$, soit à nouveau, pour faciliter les notations, G une variable aléatoire de loi $\mathcal{N}(0,1)$; en normalisant $\overline{X}_n - m$ par $\frac{\sigma}{\sqrt{n}}$, comme $\frac{\overline{X}_n - m}{\frac{\sigma}{\sqrt{n}}}$ est de loi $\mathcal{N}(0,1)$, l'identité précédente s'exprime comme

$$\mathbb{P}\big(|G| \leq \varepsilon_n\big) \,=\, 1 - \kappa$$

pour $\varepsilon_n = \frac{\varepsilon\sqrt{n}}{\sigma}$. En rappelant $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{1}{2}x^2} dx$, $t \in \mathbb{R}$, la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$, il vient

$$\mathbb{P}\big(|G| \leq \varepsilon_n\big) \,=\, \mathbb{P}(G \leq \varepsilon_n) - \mathbb{P}(G \leq -\varepsilon_n) \,=\, \Phi(\varepsilon_n) - \Phi(-\varepsilon_n).$$

Pour des raisons de symétries déjà évoquées dans le paragraphe précédent ($\Phi(-t) = 1 - \Phi(t)$ pour tout $t > 0$), $\mathbb{P}(|G| \leq \varepsilon_n) = 2\Phi(\varepsilon_n) - 1$, de sorte que

$$2\Phi(\varepsilon_n) - 1 \,=\, 1 - \kappa.$$

La fonction Φ étant strictement croissante et continue sur $\mathbb{R}$, donc inversible, il s'ensuit que

$$\frac{\varepsilon\sqrt{n}}{\sigma} \,=\, \varepsilon_n \,=\, \Phi^{-1}\Big(1 - \frac{\kappa}{2}\Big).$$

Comme déjà vu, si par exemple $\kappa = 5\%$, une table de loi normale indique que $\Phi^{-1}\big(1 - \frac{\kappa}{2}\big)$ est de l'ordre de 2. Pour plus de simplicité dans l'exposition, prendre cette valeur exacte 2. Poser alors $\varepsilon = \frac{2\sigma}{\sqrt{n}}$ de sorte que

$$I \,=\, \Big[\overline{X}_n - \frac{2\sigma}{\sqrt{n}}\,, \overline{X}_n + \frac{2\sigma}{\sqrt{n}}\Big]$$

constitue un intervalle de confiance au niveau 95% pour le paramètre de moyenne m de l'échantillon gaussien $(X_1, \ldots, X_n)$.

Si la moyenne m est connue et la variance σ^2 inconnue, l'estimateur correspondant est fourni par

$$\frac{1}{n} \sum_{k=1}^{n} (X_k - m)^2$$

qui suit ce qui est appelée une *loi du chi 2*, χ^2 (à n degrés de liberté). C'est en fait une loi Gamma $\gamma(\frac{n}{2}, \alpha)$ pour une valeur de $\alpha > 0$ dépendant de la variance σ^2 ($\alpha = \frac{1}{2}$ si $\sigma^2 = 1$).

Si m et σ^2 sont toutes deux inconnues, il convient de considérer successivement

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} \left(X_k - \overline{X}_n\right)^2$$

(variance empirique) puis, à l'image de la première situation,

$$\frac{\overline{X}_n - m}{\frac{S}{\sqrt{n}}}$$

qui suit une loi dite de *Student*[1], indépendante de m et σ^2.

4 Estimateur statistique

Ce court paragraphe descriptif introduit la notion d'estimateur statistique.

Définition 1 (Estimateur statistique)**.** *Un estimateur statistique d'un échantillon $(X_1, \ldots, X_n)$ de variables aléatoires indépendantes de même loi est une fonctionnelle*

$$\widehat{\xi} = \mathrm{Est}(X_1, \ldots, X_n)$$

dépendant d'un paramètre (réel) ξ de la loi des variables à estimer.

Typiquement, le paramètre ξ pourra être la moyenne ou la variance, comme dans les exemples de cette leçon.

Un intervalle de confiance de seuil $\mathrm{s_e} = 1 - \kappa$ associé à un estimateur statistique est déterminé à partir d'une estimation probabiliste (inégalité de

1. William Gosset, connu sous le pseudonyme Student, statisticien anglais (1876–1937).

Bienaymé-Tchebychev ou théorème central limite par exemple) du type

$$\mathbb{P}\big(e_1 < \widehat{\xi} < e_2\big) \ \geq \ 1 - \kappa$$

pour des valeurs $e_1 < e_2$ (qui dépendent d'ordinaire de κ). Suivant les cas, les inégalités $e_1 < \widehat{\xi} < e_2$ peuvent se renverser pour fournir un encadrement de ξ, appelé donc intervalle de confiance (qui pourra être ouvert ou fermé).

Les premiers exemples développés dans les paragraphes précédents illustrent ainsi ce mécanisme pour l'estimateur de la moyenne

$$\widehat{\xi} \ = \ \overline{X}_n$$

si $\mathbb{E}(X_1) = \xi$. Cet estimateur est particulièrement pertinent lorsque dans une famille de lois, l'espérance détermine le paramètre (comme par exemple pour les lois de Bernoulli, de Poisson, exponentielle, certaines lois uniformes etc.). En fait, un estimateur est dit *sans biais* si $\mathbb{E}(\widehat{\xi}\,) = \xi$.

La qualité d'un estimateur peut être mesurée en moyenne quadratique $\mathbb{E}([\widehat{\xi} - \xi]^2)$. Comme $\widehat{\xi} = \widehat{\xi}_n$ dépend de la taille de l'échantillon, sa convergence (en probabilité ou presque sûre), quand n tend vers l'infini, vers la vraie valeur ξ peut être analysée, en vue de la *consistance* de l'estimateur.

Différents estimateurs peuvent être considérés. Voici une courte liste incluant certains exemples déjà rencontrés, d'autres seront étudiés dans les exercices :

$$\overline{X}_n \ = \ \frac{1}{n}\sum_{k=1}^{n} X_k \qquad \text{moyenne}$$

$$S^2 \ = \ \frac{1}{n-1}\sum_{k=1}^{n} \big(X_k - \overline{X}_n\big)^2 \qquad \text{variance}$$

$$M_n \ = \ \max_{1\leq k\leq n} X_k \qquad \text{maximum}$$

$$F_n(t) \;=\; \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{]-\infty,t]}(X_k), \quad t \in \mathbb{R} \qquad \text{fonction de répartition empirique}$$

$$\varphi_n(u) \;=\; \frac{1}{n}\sum_{k=1}^{n} e^{iuX_k}, \quad u \in \mathbb{R} \qquad \text{fonction caractéristique empirique}$$

5 Statistique d'ordre

Étant donné un échantillon aléatoire $(X_1, \ldots, X_n)$ de variables aléatoires réelles (définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$), issu par exemple d'une expérience statistique, il est une démarche naturelle consistant à ordonner les variables $X_1, \ldots, X_n$, en ordre croissant par exemple. Cette étape permet souvent de manière pratique d'identifier, voire d'éliminer, des valeurs extrêmes (ou aberrantes).

L'échantillon réordonné (en ordre croissant), désigné par $(X_1^*, \ldots, X_n^*)$, est défini de la manière suivante : pour tout $\omega \in \Omega$, en l'absence de cas d'égalité, soit τ la permutation de $\{1, \ldots, n\}$ (dépendant de ω !) telle que

$$X_{\tau(1)}(\omega) \;<\; \cdots \;<\; X_{\tau(n)}(\omega).$$

Poser alors $X_k^*(\omega) = X_{\tau(k)}(\omega)$, $k = 1, \ldots, n$.

Ce paragraphe décrit brièvement la loi d'un échantillon ainsi réordonné, appelé *statistique d'ordre*, dans un cas particulier.

Soient $U_1, \ldots, U_n$ des variables aléatoires indépendantes de même loi uniforme $\mathcal{U}(0,1)$ sur $]0,1[$. Pour assurer un ordre sans ambiguïté, il est commode d'observer que

$$\mathbb{P}(\exists\, k \neq \ell\,; U_k = U_\ell) \;=\; 0.$$

En effet, par sous-additivité, cette probabilité est inférieure ou égale à

$$\sum_{k\neq\ell} \mathbb{P}(U_k = U_\ell).$$

Comme U_k et U_ℓ $(k \neq \ell)$ sont indépendantes de même loi $\mathcal{U}(0,1)$,

$$\mathbb{P}(U_k = U_\ell) \;=\; \int_{]0,1[} \mathbb{P}(U_k = x) d\lambda(x) \;=\; 0.$$

Il pourra donc être supposé dans la suite que pour tout ω d'un événement de probabilité 1, $U_1(\omega), \ldots, U_n(\omega)$ sont distincts.

Considérer alors la partie Δ de $\mathbb{R}^n$

$$\Delta \;=\; \big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \,;\, 0 < x_1 < x_2 < \cdots < x_n < 1\big\}$$

dont la mesure de Lebesgue est $\frac{1}{n!}$ (par récurrence par exemple, ou comme conséquence du raisonnement à suivre). Si $\phi : \mathbb{R}^n \to \mathbb{R}$ est borélienne, positive ou bornée, par définition du réarrangement,

$$\mathbb{E}\big(\phi(U_1^*, \ldots, U_n^*)\big) \;=\; \sum_\tau \int_{\{U_{\tau(1)} < \cdots < U_{\tau(n)}\}} \phi(U_{\tau(1)}, \ldots, U_{\tau(n)}) d\,\mathbb{P},$$

la somme portant sur toutes les permutations τ de $\{1, \ldots, n\}$. Comme les variables aléatoires $U_1, \ldots, U_n$ sont indépendantes et équisdistribuées, la loi de chaque vecteur $(U_{\tau(1)}, \ldots, U_{\tau(n)})$ est la même que celle de $(U_1, \ldots, U_n)$, qui est simplement la loi uniforme sur le cube $]0,1[^n$ (mesure de Lebesgue produit). Ainsi, toutes les intégrales dans la somme ci-dessus sur les permutations τ sont égales à

$$\int_{\{U_1 < \cdots < U_n\}} \phi(U_1, \ldots, U_n) d\,\mathbb{P} \;=\; \int_\Delta \phi \, d\lambda^n,$$

et il y en a $n!$. Autrement dit

$$\mathbb{E}\big(\phi(U_1^*, \ldots, U_n^*)\big) \;=\; n! \int_\Delta \phi \, d\lambda^n,$$

exprimant que la loi de $(U_1^*, \ldots, U_n^*)$ est uniforme sur Δ (de densité $n!\mathbb{1}_\Delta$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$). L'énoncé suivant rassemble la conclusion obtenue.

Proposition 2. *Si $U_1, \ldots, U_n$ sont des variables aléatoires indépendantes de même loi uniforme $\mathcal{U}(0,1)$ sur l'intervalle $]0,1[$, la loi de l'échantillon réordonné (en ordre croissant) $(U_1^*, \ldots, U_n^*)$ a pour densité $n!\mathbb{1}_\Delta$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ où*

$$\Delta = \big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \,;\, 0 < x_1 < x_2 < \cdots < x_n < 1\big\}.$$

La proposition suivante décrit une représentation très pratique de la loi de cette statistique d'ordre $(U_1^*, \ldots, U_n^*)$ à partir de variables exponentielles.

Proposition 3. *Soient $X_1, \ldots, X_n, X_{n+1}$ des variables aléatoires indépendantes de même loi exponentielle $\mathcal{E}(1)$ de paramètre 1 ; poser $S_k = X_1 + \cdots + X_k$, $k = 1, \ldots, n+1$. Alors, le vecteur aléatoire*

$$\left(\frac{S_1}{S_{n+1}}, \frac{S_2}{S_{n+1}}, \ldots, \frac{S_n}{S_{n+1}}\right)$$

a même loi que $(U_1^*, \ldots, U_n^*)$.

Démonstration. Par construction (les variables $X_1, \ldots, X_n, X_{n+1}$ prennent leurs valeurs dans $]0,\infty[$ presque sûrement), le vecteur aléatoire considéré dans l'énoncé, noté Z dans la suite, est bien à valeurs (presque sûrement) dans Δ. La loi de $(X_1, \ldots, X_{n+1})$ étant le produit de $n+1$ lois exponentielles $\mathcal{E}(1)$, par le théorème de transport (dans $\mathbb{R}^{n+1}$), pour toute fonction $\phi : \mathbb{R}^n \to \mathbb{R}$ borélienne, positive ou bornée,

$$\begin{aligned}&\mathbb{E}\big(\phi(Z)\big)\\ &= \int_{]0,\infty[^{n+1}} \phi\left(\frac{x_1}{x_1 + \cdots + x_{n+1}}, \ldots, \frac{x_1 + \cdots + x_n}{x_1 + \cdots + x_{n+1}}\right) e^{-\sum_{k=1}^{n+1} x_k} d\lambda^{n+1}.\end{aligned}$$

Il est naturel de considérer le changement de variables $y = h(x)$ de $]0,\infty[^{n+1}$ dans $\Delta\times\,]0,\infty[$ défini par

$$y_k \;=\; \frac{x_1+\cdots+x_k}{x_1+\cdots+x_{n+1}}\,,\;\; k=1,\ldots,n,\quad y_{n+1}=x_1+\cdots+x_{n+1}.$$

Le calcul du jacobien de h se trouve simplifié en analysant celui de la transformation inverse $x = h^{-1}(y)$,

$$x_k \;=\; (y_k-y_{k-1})y_{n+1},\;\; k=1,\ldots,n,\quad x_{n+1} \;=\; (1-y_n)y_{n+1}$$

(avec la convention $y_0 = 0$) qui est égal à

$$\text{dét}\left(\Big(\frac{\partial h_k^{-1}}{\partial y_\ell}(y)\Big)_{k,\ell}\right) \;=\; \begin{vmatrix} y_{n+1} & -y_{n+1} & 0 & \cdots & 0 & 0 \\ 0 & y_{n+1} & -y_{n+1} & \cdots & \vdots & \vdots \\ 0 & 0 & y_{n+1} & \cdots & 0 & 0 \\ \vdots & \vdots & 0 & \ddots & -y_{n+1} & 0 \\ 0 & 0 & \vdots & \ddots & y_{n+1} & -y_{n+1} \\ y_1 & y_2-y_1 & y_3-y_2 & \cdots & y_n-y_{n-1} & 1-y_n \end{vmatrix}.$$

En remplaçant la deuxième colonne par sa somme avec la première, puis la troisième par sa somme avec la deuxième (nouvellement formée) et ainsi de suite, il vient

$$\text{dét}\left(\Big(\frac{\partial h_k^{-1}}{\partial y_\ell}(y)\Big)_{k,\ell}\right) \;=\; \begin{vmatrix} y_{n+1} & 0 & 0 & \cdots & 0 \\ 0 & y_{n+1} & 0 & \vdots & 0 \\ \vdots & \ddots & y_{n+1} & \ddots & \vdots \\ 0 & \cdots & 0 & \ddots & 0 \\ y_1 & y_2 & y_3 & \cdots & 1 \end{vmatrix} \;=\; y_{n+1}^n.$$

Il en résulte que

$$\text{dét}\big(J_h(x)\big) \;=\; \frac{1}{\text{dét}(J_{h^{-1}}(y))} \;=\; \frac{1}{y_{n+1}^n} \;=\; \frac{1}{(x_1+\cdots+x_{n+1})^n}\,.$$

L'application de la formule du changement de variable (Leçon 5), puis le théorème de Fubini-Tonelli, fournissent ainsi

$$\mathbb{E}\big(\phi(Z)\big) \;=\; \int_{\Delta\times\,]0,\infty[} \phi(y_1,\ldots,y_n)\, y_{n+1}^n e^{-y_{n+1}} d\lambda^n(y_1,\ldots,y_n) d\lambda^1(y_{n+1}).$$

Comme $\int_{]0,\infty[} y_{n+1}^n e^{-y_{n+1}} d\lambda^1(y_{n+1}) = \Gamma(n+1) = n!$, la loi de Z a pour densité $n!\mathbb{1}_\Delta$ par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$, ce qui conclut au résultat. □

La démonstration montre de la même façon que Z est indépendant de S_{n+1}. L'Exercice 7 prolonge cette proposition en déterminant la loi de $\frac{S_k}{S_{n+1}}$ pour chaque $k = 1,\ldots,n$.

La construction précédente à travers des variables aléatoires indépendantes de loi exponentielle a d'autres vertus, comme par exemple le résultat décrit dans l'Exercice 8 qui s'établit essentiellement suivant le même schéma.

6 Découverte : Spectre de matrice aléatoire

L'Exercice 6 montre comment, pour un échantillon $(X_1, \ldots, X_n)$ de variables aléatoires (réelles) indépendantes de même loi sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, la mesure empirique

$$\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k}$$

approche la loi commune des éléments de l'échantillon. Cette mesure aléatoire associe en effet à chaque $\omega \in \Omega$ la mesure de probabilité (discrète) $\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k(\omega)}$ sur les boréliens de $\mathbb{R}$, et la fonction de répartition $F_n(\cdot)(\omega)$ de cette dernière est donnée par

$$F_n(t)(\omega) \;=\; \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k(\omega)}\big(]-\infty, t]\big) \;=\; \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{]-\infty,t]}\big(X_k(\omega)\big), \quad t \in \mathbb{R}.$$

D'après la loi des grands nombres, pour chaque $t \in \mathbb{R}$, $F_n(t)$ converge presque sûrement vers $\mathbb{E}(\mathbb{1}_{]-\infty,t]})(X_1) = \mathbb{P}(X_1 \leq t)$, c'est-à-dire la fonction de répartition de la loi commune, traduisant bien l'approximation en loi.

La mesure empirique $\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k}$ peut clairement être considérée pour des variables aléatoires à valeurs dans des espaces plus généraux.

La recherche récente a étudié des propriétés du même type pour des échantillons de valeurs propres de matrices aléatoires. Voici une illustration simple.

Soit $X_{k,\ell}$, $k, \ell \in \mathbb{N}$, une famille de variables aléatoires réelles, indépendantes de même loi ; pour chaque $n \in \mathbb{N}$, former la matrice carrée $M_n = (X_{k,\ell})_{1 \leq k,\ell \leq n}$, qui est donc une « matrice aléatoire ». Cette matrice admet des valeurs propres, complexes, aléatoires, $\rho_1, \ldots, \rho_n$, au nombre de n, éventuellement avec répétition. Il est intéressant de comprendre le comportement asymptotique de la matrice M_n, lorsque sa taille n tend vers l'infini, à travers le comportement de son

spectre, sous la forme globale de sa *mesure spectrale* $\frac{1}{n}\sum_{k=1}^{n}\delta_{\rho_k}$, qui est donc une mesure (sur $\mathbb{C}$) aléatoire (analogue à la mesure empirique précédente).

Un résultat important de la théorie établit que, sous les seules conditions de moments (normalisés) $\mathbb{E}(X_{1,1}) = 0$, $\mathbb{E}(X_{1,1}^2) = 1$, presque sûrement, pour tout rectangle $R =]-\infty, s] \times]-\infty, t]$, $s, t \in \mathbb{R}$, de $\mathbb{R}^2$,

$$\frac{1}{n}\sum_{k=1}^{n}\delta_{\frac{\rho_k}{\sqrt{n}}}(R) \;=\; \frac{1}{n}\sum_{k=1}^{n}\mathbb{1}_R\Big(\frac{\rho_k}{\sqrt{n}}\Big)$$

converge, quand n tend vers l'infini, vers $\frac{1}{\pi}\lambda_{\mathcal{D}}(R)$ où $\lambda_{\mathcal{D}}$ est la mesure de Lebesgue de $\mathbb{C}$ ($\mathbb{R}^2$) restreinte au disque unité $\mathcal{D}$. La mesure spectrale $\frac{1}{n}\sum_{k=1}^{n}\delta_{\rho_k}$ approche donc en loi la mesure uniforme sur le disque.

Exercices

(*Une étoile * désignera une question de difficulté supérieure.*)

Exercice 1. Soit $(X_1, \ldots, X_n)$ un échantillon de variables aléatoires sur $(\Omega, \mathcal{A}, \mathbb{P})$, indépendantes de même loi exponentielle $\mathcal{E}(\frac{1}{\xi})$ de paramètre $\frac{1}{\xi} > 0$. Pour tout $n \geq 1$, poser comme dans la leçon, $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$.

a) Rappeler la moyenne et la variance de X_1.

b) Démontrer que pour tout $t > 0$,

$$\mathbb{P}\big(|\overline{X}_n - \xi| \geq t\big) \leq \frac{\xi^2}{t^2 n}\,.$$

c) Calculer un intervalle de confiance au seuil $\mathrm{s_e} = 1 - \kappa$ de ξ pour l'estimateur de la moyenne lorsque $\kappa n > 1$.

d) Mêmes questions si les variables $X_1, \ldots, X_n$ sont indépendantes de même loi uniforme $\mathcal{U}(0, a)$ sur $[0, a]$, où $a > 0$ est un paramètre inconnu à estimer.

Exercice 2. Un vol Toulouse-Paris est assuré par un avion de $N = 150$ places. Pour un vol de ce type, des estimations ont montré que la probabilité pour qu'une personne confirme son billet est $p = 0,75$. La compagnie vend n billets, $n > 150$. Soit X la variable aléatoire « nombre de personnes parmi les n possibles ayant confirmé leur réservation pour ce vol ».

a) Quelle est la loi exacte suivie par X ?

b) Quel est le nombre maximum de places que la compagnie peut vendre pour que, à au moins 95%, elle soit sûre que tout le monde puisse monter dans l'avion (c'est-à-dire trouver n tel que $\mathbb{P}(X > 150) \leq 0,05$) ? (*Indication* : utiliser l'inégalité de Tchebychev ; proposer une estimation plus précise à partir du théorème central limite.)

c) Discuter suivant les valeurs de N, p et n.

Exercice 3 (*Processus de Poisson*). Soit X_n, $n \in \mathbb{N}$, une suite de variables aléatoires indépendantes de même loi exponentielle $\mathcal{E}(\alpha)$ de paramètre $\alpha > 0$. Poser $S_n = X_1 + \cdots + X_n$, $n \geq 1$, et par convention $S_0 = 0$. Pour tout $t > 0$, soit $N_t = \sup\{n \geq 0\,;\, S_n < t\}$.

a) Quelle est la loi de S_n ?

b) Établir que, pour tout $n \in \mathbb{N}$,

$$\mathbb{P}(N_t = n) \;=\; \mathbb{P}(S_n < t) - \mathbb{P}(S_{n+1} < t).$$

En déduire la loi de N_t.

Exercice 4 (*Loi de Student*). Soient X et Y deux variables indépendantes, X de loi $\mathcal{N}(0,1)$ et Y de loi Gamma $\gamma(\frac{n}{2}, \frac{1}{2})$ de paramètres $\frac{n}{2}$ et $\frac{1}{2}$, aussi appelée loi du χ^2 à $n \geq 1$ degrés de liberté. La loi de

$$T_n \;=\; \frac{X}{\sqrt{\frac{Y}{n}}}$$

est appelée loi de Student à n degrés de liberté. Montrer que cette loi a une densité, et la décrire (*ainsi que son histoire !*).

Exercice 5. Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même loi normale centrée réduite $\mathcal{N}(0,1)$ et soit X le vecteur $(X_1, \ldots, X_n)$. Soit O une matrice $n \times n$ orthogonale, et soit $Y = OX$; rappeler pourquoi $\sum_{k=1}^n X_k^2 = \sum_{k=1}^n Y_k^2$.

a) Démontrer que Y a même loi que X.

Poser

$$\overline{X} \;=\; \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad S^2 \;=\; \frac{1}{n-1}\sum_{k=1}^{n} \big(X_k - \overline{X}\big)^2.$$

b) En choisissant O dont tous les termes de la dernière ligne sont égaux à $\frac{1}{\sqrt{n}}$,

vérifier que

$$(n-1)S^2 \;=\; \sum_{k=1}^{n} X_k^2 - n\overline{X}^2 \;=\; \sum_{k=1}^{n} Y_k^2 - Y_n^2 \;=\; \sum_{k=1}^{n-1} Y_k^2.$$

c) Déduire de ce qui précède que $\overline{X}$ et S^2 sont indépendantes et que $(n-1)S^2$ a la même loi que $\sum_{k=1}^{n-1} X_k^2$.

Exercice 6 (*Théorème de Glivenko*[2]*-Cantelli*[3]). Soit X_n, $n \in \mathbb{N}$, une suite de variables aléatoires réelles définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, indépendantes de même loi P sur les boréliens de $\mathbb{R}$ de fonction de répartition F.

Pour tout $n \geq 1$, et tout $t \in \mathbb{R}$, considèrer la variable aléatoire, dite *fonction de répartition empirique*, $F_n(t) : (\Omega, \mathcal{A}, \mathbb{P}) \to [0,1]$ définie pour tout $\omega \in \Omega$ par

$$F_n(t)(\omega) \;=\; \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{]-\infty,t]}\big(X_k(\omega)\big).$$

C'est la fonction de répartition de la mesure (discrète) $\frac{1}{n}\sum_{k=1}^{n} \delta_{X_k(\omega)}$ sur $\mathcal{B}(\mathbb{R})$.

a) Démontrer que pour tout $t \in \mathbb{R}$, $\lim_{n\to\infty} F_n(t) = F(t)$ presque sûrement.

b) Soit $L \geq 1$ un entier et soient $t_0 < t_1 < \cdots < t_L$ des réels ; démontrer que, presque sûrement, $\lim_{n\to\infty} \max_{0\leq \ell \leq L} |F_n(t_\ell) - F(t_\ell)| = 0$.

Il est supposé dans la suite du problème que P est la loi uniforme sur $[0,1]$.

c) Démontrer que, pour tout $L \geq 1$ et tout $n \geq 1$,

$$\sup_{t\in[0,1]} \big|F_n(t) - t\big| \;\leq\; \max_{0\leq \ell \leq L} \big|F_n\big(\tfrac{\ell}{L}\big) - \tfrac{\ell}{L}\big| + \tfrac{1}{L}\,.$$

2. Valery Glivenko, mathématicien ukrainien et soviétique (1896–1940).
3. Francesco Paolo Cantelli, mathématicien italien (1875–1966).

d) Déduire de ce qui précède que l'ensemble

$$A = \left\{\omega \in \Omega\,;\, \lim_{n\to\infty} \sup_{t\in[0,1]} \left|F_n(t)(\omega) - t\right| = 0\right\}$$

contient un ensemble mesurable de probabilité égale à 1.

Exercice 7 (*Lois Gamma et Beta*). Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même loi exponentielle $\mathcal{E}(\alpha)$ de paramètre $\alpha > 0$.

a) Démontrer que la loi de $X_1 + \cdots + X_n$, est une loi Gamma $\gamma(n, \alpha)$ (Exercice 8, Leçon 10), de densité $x \mapsto \frac{1}{\Gamma(n)}\, \alpha^n x^{n-1} e^{-\alpha x}$ par rapport à la mesure de Lebesgue sur $]0, \infty[$ (rappeler que $\Gamma(n) = (n-1)!$).

Pour plus de simplicité dans la suite, prendre $\alpha = 1$ (le cas général s'en déduisant compte tenu que αX_1 suit une loi exponentielle de paramètre 1 si X_1 est de paramètre α – à vérifier).

b) Démontrer que $\frac{X_1}{X_1+X_2}$ est indépendant de $X_1 + X_2$, et suit la loi uniforme sur $]0, 1[$.

c*) Généraliser la question précédente en déterminant, pour $k = 1, \ldots, n-1$, la loi de

$$\frac{X_1 + \cdots + X_k}{X_1 + \cdots + X_n},$$

dite Beta et notée $\beta(k, n-k)$. En déduire la valeur de $\int_0^1 x^{k-1}(1-x)^{n-k-1} dx$. (*Indication* : écrire $X_1 + \cdots + X_n = (X_1 + \cdots + X_k) + (X_{k+1} + \cdots + X_n)$ et utiliser la question a).)

À l'image des lois $\gamma(n, \alpha)$, les lois Beta $\beta(k, \ell)$ peuvent être extrapolées pour $k, \ell \in]0, \infty[$.

Exercice 8*. Soient $X_1, \ldots, X_n, X_{n+1}$ des variables aléatoires indépendantes de même loi exponentielle $\mathcal{E}(1)$ de paramètre 1 ; poser $S_k = X_1 + \cdots + X_k$, $k = 1, \ldots, n+1$.

a) Démontrer que la loi du vecteur aléatoire

$$\left(\frac{X_1}{S_{n+1}}, \frac{X_2}{S_{n+1}}, \ldots, \frac{X_n}{S_{n+1}} \right)$$

est uniforme sur (le simplexe) $\Delta = \left\{ y \in]0, \infty[^n \, ; \, \sum_{k=1}^n y_k < 1 \right\}$.

b) Vérifier que

$$\left(\frac{X_1}{S_{n+1}}, \frac{X_2}{S_{n+1}}, \ldots, \frac{X_{n+1}}{S_{n+1}} \right)$$

est indépendant de S_{n+1} (et donc aussi $\left(\frac{X_1}{S_{n+1}}, \frac{X_2}{S_{n+1}}, \ldots, \frac{X_n}{S_{n+1}} \right)$).