Introduction au calcul des probabilités et à la statistique

Exercices, Problèmes et corrections

22 juin 2005

Table des matières

Chapitre I

Espaces probabilisés

I.1 Énoncés

Exercice I.1.
On tire au hasard deux cartes dans un jeu de 52 cartes.

1. Quelle est la probabilité pour que la couleur des deux cartes soit pique ?
2. Quelle est la probabilité pour que les deux cartes ne soient pas de la même couleur (pique, cœur, carreau, trèfle) ?
3. Quelle est la probabilité pour que la première carte soit un pique et la seconde un cœur ?
4. Quelle est la probabilité pour qu'il y ait un pique et un cœur ?
5. Quelle est la probabilité pour qu'il y ait un pique et un as ?

△

Exercice I.2.
On considère une classe de n élèves. On suppose qu'il n'y a pas d'année bissextile.

1. Quelle est la probabilité, p_n, pour que deux élèves au moins aient la même date d'anniversaire ? Trouver le plus petit entier n_1 tel que $p_{n_1} \geq 0.5$. Calculer p_{366}.
2. Quelle est la probabilité, q_n, pour qu'au moins un élève ait la même date d'anniversaire que Socrate ? Calculer q_{n_1} et q_{366}.

△

Exercice I.3.
A possède deux dés à six faces, et B posséde un dé à douze faces. Le joueur qui fait le plus grand score remporte la mise (match nul si égalité). Le jeu est-il équilibré ? On calculera la probabilité que A gagne et la probabilité d'avoir un match nul.

△

Exercice I.4.
Afin de savoir si les élèves travaillent indépendamment ou en groupe, un enseignant donne m exercices à une classe de n élèves. Chaque élève choisit k exercices parmi les m.

1. Calculer la probabilité pour que les élèves aient tous choisi une combinaison fixée de k exercices.
2. Calculer la probabilité pour que tous les élèves aient choisi les k mêmes exercices.
3. Calculer la probabilité pour qu'une combinaison fixée à l'avance, n'ait pas été choisie.
4. Calculer la probabilité pour qu'il existe au moins une combinaison de k exercices qui n'ait pas été choisie. (On utilisera la formule du crible (I.1) cf. exercice I.6)
5. A.N. Donner les résultats pour $n = 20$, $m = 4$, $k = 2$. Comparer les valeurs pour les questions 1 et 2 puis 3 et 4.

△

Exercice I.5.
On utilise dans cet exercice la formule du crible (I.1) (cf exercice I.6).

1. Calculer à l'aide de la formule du crible le nombre de surjections de $\{1, \cdots, n\}$ dans $\{1, \cdots, k\}$.
2. En déduire $\left\{ {n \atop k} \right\}$, le nombre de partitions d'un ensemble à n éléments en k sous-ensembles non vides.
3. Montrer que

$$\left\{ {n \atop k} \right\} = \left\{ {n-1 \atop k-1} \right\} + k \left\{ {n-1 \atop k} \right\}, \quad \left\{ {n \atop 1} \right\} = 1, \quad \left\{ {n \atop k} \right\} = 0 \text{ si } k > n.$$

Les nombres $\left\{ {n \atop k} \right\} = \dfrac{1}{k!} \displaystyle\sum_{j=0}^{k} (-1)^j C_k^j (k-j)^n$ sont appelés les nombres de Stirling de deuxième espèce.

△

Exercice I.6.
La **formule du crible**. Soit $A_1, \cdots, A_n$ des évènements.

1. Montrer que $\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$.
2. Établir une formule analogue pour $\mathbb{P}(A_1 \cup A_2 \cup A_3)$.
3. Montrer la formule du crible par récurrence.

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i \right) = \sum_{p=1}^{n} (-1)^{p+1} \sum_{1 \le i_1 < \cdots < i_p \le n} \mathbb{P}(A_{i_1} \cap \cdots \cap A_{i_p}). \tag{I.1}$$

△

Exercice I.7.
On suppose que l'on a autant de chance d'avoir une fille ou un garçon à la naissance. Votre voisin de palier vous dit qu'il a deux enfants.

1. Quelle est la probabilité qu'il ait au moins un garçon ?

2. Quelle est la probabilité qu'il ait un garçon, sachant que l'aînée est une fille ?
3. Quelle est la probabilité qu'il ait un garçon, sachant qu'il a au moins une fille ?
4. Vous sonnez à la porte de votre voisin. Une fille ouvre la porte. Sachant que l'aîné(e) ouvre la porte avec probabilité p, et ce indépendamment de la répartition de la famille, quelle est la probabilité que votre voisin ait un garçon ?
5. Vous téléphonez à votre voisin. Une fille décroche le téléphone. Vous savez que dans les familles avec un garçon et une fille, la fille décroche le téléphone avec probabilité p, quelle est la probabilité que votre voisin ait un garçon ?

△

Exercice I.8.
Un laboratoire pharmaceutique vend un test avec la notice suivante : si vous êtes malade, alors le test est positif avec probabilité $\alpha = 98\%$ (α est la sensibilité du test), si vous êtes sain, alors le test est positif avec probabilité $\beta = 3\%$ ($1-\beta$ est la spécificité du test) . Sachant qu'en moyenne il y a un malade sur $\tau^{-1} = 1000$ personnes, calculer la probabilité pour que vous soyez un sujet sain alors que votre test est positif. Calculer la probabilité d'être malade alors que le test est négatif. Commentaire.

△

Exercice I.9.
Le gène qui détermine la couleur bleue des yeux est récessif. Pour avoir les yeux bleus, il faut donc avoir le génotype bb. Les génotypes mm et bm donnent des yeux marron. On suppose que les parents transmettent indifféremment un de leurs gènes à leurs enfants. La sœur et la femme d'Adrien ont les yeux bleus, mais ses parents ont les yeux marron.

1. Quelle est la probabilité pour qu'Adrien ait les yeux bleus ?
2. Quelle est la probabilité que le premier enfant d'Adrien ait les yeux bleus sachant que Adrien a les yeux marron ?
3. Quelle est la probabilité pour que le deuxième enfant d'Adrien ait les yeux bleus sachant que le premier a les yeux marron ?
4. Comment expliquez-vous la différence des résultats entre les deux dernières questions ?

△

Exercice I.10.
On considère trois cartes : une avec les deux faces rouges, une avec les deux faces blanches, et une avec une face rouge et une face blanche. On tire une carte au hasard. On expose une face au hasard. Elle est rouge. Parieriez-vous que la face cachée est blanche ? Pour vous aider dans votre choix :

1. Déterminer l'espace de probabilité.
2. Calculer la probabilité que la face cachée est blanche sachant que la face visible est rouge.

△

Exercice I.11.
Le jeu télévisé "la porte de la fortune" comporte trois portes : A, B et C. Derrière l'une d'entre elles se trouve une voiture haut de gamme et rien derrière les deux autres. Vous choisissez au hasard une des trois portes sans l'ouvrir, par exemple la porte A. À ce moment-là, le présentateur, qui sait derrière quelle porte se trouve la voiture, ouvre une porte parmi les deux B et C, derrière laquelle il n'y a évidement rien. On vous propose alors de changer ou non de porte, le but étant d'ouvrir la porte qui cache la voiture afin de gagner. L'objectif de cet exercice est de déterminer votre meilleure stratégie.

1. On suppose que si la voiture est derrière la porte A, alors le présentateur choisit au hasard entre les deux autres portes. Calculer la probabilité pour que la voiture soit derrière la porte B sachant que le présentateur ouvre la porte C. Que faites-vous ?
2. On suppose que si la voiture est derrière la porte A, alors le présentateur choisit systématiquement la porte B. Que faites-vous si le présentateur ouvre la porte B ?, la porte C ?
3. Montrer que quelle que soit la valeur de la probabilité pour que le présentateur ouvre la porte B (respectivement C) sachant que la voiture est en A, vous avez intérêt à changer de porte. En déduire que la meilleure stratégie consiste à changer systématiquement de porte.
4. Une fois que le présentateur a ouvert une porte, et quel que soit le mécanisme de son choix, vous tirez à pile ou face pour choisir si vous changez ou non de porte. Quelle est votre probabilité de gagner la voiture ? Vérifier que cette stratégie est moins bonne que la précédente.

△

Exercice I.12.
On utilise dans cet exercice la formule du crible (I.1) (cf exercice I.6).

1. Pour fêter leur réussite à un concours, n étudiants se donnent rendez-vous dans un chalet. En entrant chaque personne dépose sa veste dans un vestiaire. Au petit matin, quand les esprits ne sont plus clairs, chacun prend au hasard une veste. Quelle est la probabilité pour qu'une personne au moins ait sa propre veste ?
2. En s'inspirant de la question précédente, calculer la probabilité $\pi_n(k)$ pour que k personnes exactement aient leur propre veste.
3. Calculer la limite $\pi(k)$ de $\pi_n(k)$ quand $n \to \infty$. Vérifier que la famille $(\pi(k), k \in \mathbb{N})$ détermine une probabilité sur $\mathbb{N}$. Il s'agit en fait de la loi de Poisson.

△

Exercice I.13.
Une urne contient r boules rouges et b boules bleues.

1. On tire **avec** remise $p \in \mathbb{N}^*$ boules. Calculer la probabilité pour qu'il y ait p_r boules rouges et p_b boules bleues ($p_r + p_b = p$).
2. On tire **sans** remise $p \leq r + b$ boules. Calculer la probabilité pour qu'il y ait p_r boules rouges et p_b boules bleues ($p_r + p_b = p$).

3. Calculer, dans les deux cas, les probabilités limites quand $r \to \infty$, $b \to \infty$ et $r/(b+r) \to \theta \in]0,1[$.

△

Exercice I.14.
Eugène et Diogène ont l'habitude de se retrouver chaque semaine autour d'un verre et de décider à pile ou face qui règle l'addition. Eugène se lamente d'avoir payé les quatre dernières additions et Diogène, pour faire plaisir à son ami, propose de modifier exceptionnellement la règle : "Eugène, tu vas lancer la pièce cinq fois et tu ne paieras que si on observe une suite d'au moins trois piles consécutifs ou d'au moins trois faces consécutives". Eugène se félicite d'avoir un si bon ami. À tort ou à raison ?

△

I.2 Corrections

Exercice I.1.
1) $\frac{1}{17}$; 2) $\frac{13}{17}$; 3) $\frac{13}{52}\frac{13}{51} = \frac{13}{204}$; 4) $\frac{13}{102}$; 5) $\frac{2}{52} + 2\frac{12}{52}\frac{3}{51} = \frac{29}{26x17}$. ▲

Exercice I.2.

Pour répondre à la première question on définit d'abord l'espace de probabilités : $\Omega = \{1, \dots, 365\}^n$ avec $\omega = (\omega_1, \dots, \omega_n)$ où ω_i est la date d'anniversaire de l'élève i. On choisit la probabilité uniforme sur Ω. On a alors :

$$\begin{aligned} p_n &= \mathbb{P}(\text{au moins 2 élèves ont la même date d'anniversaire}) \\ &= 1 - \mathbb{P}(\text{tous les élèves ont des dates d'anniverssaires différentes}) \\ &= 1 - \mathbb{P}(\{\omega; \omega_i \neq \omega_j, \forall i \neq j\}) \\ &= 1 - \frac{\text{Card}\,\{\omega; \omega_i \neq \omega_j, \forall i \neq j\}}{365^n} \\ &= 1 - \frac{\text{Card}\,\{\text{injections de } \{1, \dots, n\} \text{ dans } \{1, \dots, 365\}\}}{365^n} \\ &= \begin{cases} 1 - \dfrac{365!}{(365-n)!365^n} & \text{si } n \leq 365, \\ 1 & \text{si } n \geq 366. \end{cases} \end{aligned}$$

On obtient les valeurs numériques suivantes :

$$p_{22} \simeq 0.476; \qquad p_{23} \simeq 0.507; \qquad p_{366} = 1.$$

En fait, les naissances ne sont pas uniformément réparties sur l'année. Les valeurs statistiques de p_n sont donc plus élevées.

Pour la deuxième question, on a, en notant x la date d'anniversaire de Socrate :

$$\begin{aligned} q_n &= \mathbb{P}(\text{au moins un élève a son anniversaire le jour } x) \\ &= 1 - \mathbb{P}(\text{tous les élèves ont leur date d'anniversaire différente de } x) \\ &= 1 - \mathbb{P}(\{\omega; \omega_i \neq x, \forall i \in \{1, \dots, n\}\}) \\ &= 1 - \frac{\text{Card}\,\{\omega; \omega_i \neq x, \forall i \in \{1, \dots, n\}\}}{365^n} \\ &= 1 - \left(\frac{364}{365}\right)^n. \end{aligned}$$

On obtient les valeurs numériques suivantes :

$$q_{23} \simeq 0.061; \qquad q_{366} \simeq 0.634.$$

Les valeurs p_n et q_n sont très différentes. ▲

Exercice I.3.

L'espace d'état est $\Omega = \{(i, j, k); 1 \leq i, j \leq 6, 1 \leq k \leq 12\}$, où i représente le résultat du premier dé à 6 faces, j celui du second à 6 faces et k celui du dé à 12 faces. On munit $(\Omega, \mathcal{P}(\Omega))$ de la probabilité uniforme $\mathbb{P}$ (dés équilibrés indépendants). Pour calculer $\mathbb{P}$(A gagne), on

compte le nombre de cas favorables divisé par le nombre de cas possibles (Card $\Omega = 6.6.12$). Le nombre de cas favorables est :

$$\begin{aligned}\text{Card}\,\{(i,j,k)\in\Omega; i+j>k\} &= \sum_{1\le i,j\le 6}\sum_{k=1}^{12}\mathbf{1}_{\{i+j>k\}} = \sum_{1\le i,j\le 6}(i+j-1)\\ &= 2.6.\sum_{i=1}^{6} i - 36 = 36.7 - 36 = 36.6.\end{aligned}$$

Donc on a $\mathbb{P}(\text{A gagne}) = 6/12 = 1/2$.

$$\begin{aligned}\mathbb{P}(\text{match nul}) &= \text{Card}\,\{(i,j,k)\in\Omega; i+j=k\}/6.6.12\\ &= \frac{1}{6.6.12}\sum_{1\le i,j\le 6}\sum_{k=1}^{12}\mathbf{1}_{\{k=i+j\}} = 1/12.\end{aligned}$$

Le jeu n'est pas équilibré car $\mathbb{P}(\text{A gagne}) = \frac{1}{2} > \mathbb{P}(\text{A perd}) = \frac{1}{2} - \frac{1}{12}$. ▲

Exercice I.4.

Le nombre de combinaisons de k exercices est $N = C_m^k$.

1. $p_1 = N^{-n}$.
2. $p_2 = N^{-n+1}$.
3. $p_3 = (1 - N^{-1})^n$.
4. Par la formule du crible, on a

$$p_4 = \mathbb{P}\left(\bigcup_{i\in\{1,\dots,N\}}\{\text{combinaison } i \text{ non choisie}\}\right) = \sum_{p=1}^{N}(-1)^{p+1}C_N^p\left(\frac{N-p}{N}\right)^n.$$

5. $N = 6$ et $p_1 \simeq 2,7\ 10^{-16}$; $p_2 \simeq 1,6\ 10^{-15}$; $p_3 \simeq 2,6\%$; $p_4 \simeq 15,2\%$.

▲

Exercice I.5.

1. Soit F l'ensemble des fonctions de $\{1,\dots,n\}$ dans $\{1,\dots,k\}$. On pose $A_i = \{f \in E; f^{-1}(\{i\}) = \emptyset\}$. Le nombre de surjection, N, est donc égal à Card $(F)-$Card $(\cup_{i=1}^k A_i)$. D'après la formule du crible, on a

$$\begin{aligned}\text{Card}\,(\cup_{i=1}^k A_i) &= \sum_{j=1}^{k}(-1)^{j+1}\sum_{1\le i_1\le\cdots\le i_j}\text{Card}\,(A_{i_1}\cap\cdots\cap A_{i_j})\\ &= \sum_{j=1}^{k}(-1)^{j+1}C_k^j(k-j)^k.\end{aligned}$$

2. Comme $k!$ surjections distinctes de $\{1,\dots,n\}$ dans $\{1,\dots,k\}$ définissent la même partition en k sous-ensembles non vides, il vient

$$\left\{{n \atop k}\right\} = \frac{N}{k!} = \frac{1}{k!}\sum_{j=0}^{k}(-1)^j C_k^j(k-j)^n.$$

3. Il est clair que $\left\{ {n \atop 1} \right\} = 1$ et $\left\{ {n \atop k} \right\} = 0$. Qaund on dispose de $n > k > 1$ éléments à répartir en k sous-ensembles non vides, alors :
   – Soit les $n-1$ premiers éléments sont répartis en $k-1$ sous-ensembles non vides, et le dernier élément forme un sous-ensemble à lui tout seul. Ceci représente $\left\{ {n-1 \atop k-1} \right\}$ cas possibles.
   – Soit les $n-1$ premiers éléments sont répartis en k sous-ensembles non vides, et le dernier élément appartient à l'un de ces k sous-ensemble. Ceci représente $k \left\{ {n-1 \atop k} \right\}$ cas possibles.

   On en déduit donc la formule de récurrence.

▲

Exercice I.6.

La correction est élémentaire

▲

Exercice I.7.

On décrit une réalisation $\omega = (\omega_1, \omega_2)$ où ω_1 est le sexe de l'aîné(e) G ou F, et ω_2 le sexe du second. L'espace d'états est donc

$$\Omega = \{(G,G),(G,F),(F,G),(F,F)\}.$$

Les naissances étant équiprobables, on choisit la probabilité uniforme sur Ω.

1. $\mathbb{P}(\exists G) = \dfrac{\text{Card}\,\{(G,F),(F,G),(G,G)\}}{\text{Card}\,\{(F,F),(G,F),(F,G),(G,G)\}} = 3/4.$
2. $\mathbb{P}(\text{cadet} = G | \text{ainée} = F) = 1/2.$
3. $\mathbb{P}(\exists G | \exists F) = \dfrac{\text{Card}\,\{(G,F),(F,G)\}}{\text{Card}\,\{(F,F),(G,F),(F,G)\}} = 2/3.$
4. On a
$$\mathbb{P}(\exists G | \exists F, F \text{ ouvre la porte}) = \frac{\mathbb{P}(\exists G, \exists F, F \text{ ouvre la porte})}{\mathbb{P}(\exists F, F \text{ ouvre la porte})}.$$
Calculons $\mathbb{P}(\exists G, \exists F, F \text{ ouvre la porte})$. On a par la formule de décomposition et par indépendance
$$\begin{aligned}\mathbb{P}(\exists G, \exists F, F \text{ ouvre la porte}) &= \mathbb{P}(\text{aînée ouvre la porte }, (F,G)) \\ &\quad + \mathbb{P}(\text{cadette ouvre la porte }, (G,F)) \\ &= \mathbb{P}(\text{aîné(e) ouvre la porte })\mathbb{P}((F,G)) \\ &\quad + \mathbb{P}(\text{cadet(te) ouvre la porte })\mathbb{P}((G,F)) \\ &= p\,\frac{1}{4} + (1-p)\,\frac{1}{4} = \frac{1}{4}.\end{aligned}$$
De même, on a par indépendance
$$\begin{aligned}\mathbb{P}(\exists F, F \text{ ouvre la porte}) &= \mathbb{P}(\text{aîné(e) ouvre la porte })\mathbb{P}((F,G) \text{ ou } (F,F)) \\ &\quad + \mathbb{P}(\text{cadet(te) ouvre la porte })\mathbb{P}((G,F) \text{ ou } (F,F)) \\ &= p\,\frac{1}{2} + (1-p)\,\frac{1}{2} = \frac{1}{2}.\end{aligned}$$

Ainsi on trouve $\mathbb{P}(\exists G|\exists F \text{ ouvre la porte}) = \frac{1}{4}\frac{1}{\frac{1}{2}} = \frac{1}{2}$.

5. On a $\mathbb{P}(\exists G|\exists F, F \text{ décroche}) = \dfrac{\mathbb{P}(\exists G, \exists F, F \text{ décroche})}{\mathbb{P}(\exists F, F \text{ décroche})}$. On calcule d'abord

$$\mathbb{P}(\exists G, \exists F, F \text{ décroche}) = \mathbb{P}(F \text{ décroche } |(G,F) \text{ ou } (F,G))\mathbb{P}((G,F) \text{ ou } (F,G)) = p/2.$$

Ainsi on obtient

$$\mathbb{P}(\exists G|\exists F, F \text{ décroche}) = \frac{\frac{1}{2}\,p}{\mathbb{P}((F,F)) + \mathbb{P}(\exists G, \exists F, F \text{ décroche})} = \frac{2p}{1+2p} \in [0, 2/3].$$

▲

Exercice I.8.

On note les évènements $S = \{\text{je suis sain}\}$, $M = \{\text{je suis malade}\}$, $+ = \{\text{mon test est positif}\}$ et $- = \{\text{mon test est négatif}\}$. On cherche $\mathbb{P}(S|+)$ et $\mathbb{P}(M|-)$. On connaît les valeurs des probabilités suivantes : $\mathbb{P}(+|M) = \alpha$, $\mathbb{P}(+|S) = \beta$ et $\mathbb{P}(M) = \tau$. On déduit de la formule de Bayes que :

$$\mathbb{P}(S|+) = \frac{\mathbb{P}(+|S)\mathbb{P}(S)}{\mathbb{P}(+|M)\mathbb{P}(M) + \mathbb{P}(+|S)\mathbb{P}(S)} = \frac{\beta(1-\tau)}{\alpha\tau + \beta(1-\tau)}.$$

On déduit également de la formule de Bayes que :

$$\mathbb{P}(M|-) = \frac{\mathbb{P}(-|M)\mathbb{P}(M)}{\mathbb{P}(-|M)\mathbb{P}(M) + \mathbb{P}(-|S)\mathbb{P}(S)} = \frac{(1-\alpha)\tau}{(1-\alpha)\tau + (1-\beta)(1-\tau)}.$$

A.N. $\mathbb{P}(S|+) \simeq 30/31$ (en fait $\mathbb{P}(S|+) \simeq 96.8\%$) et $\mathbb{P}(M|-) \simeq 2.10^{-5}$. ▲

Exercice I.9.

On note M le phénotype "yeux marron", et B "yeux bleus". Le phénotype M provient des génotypes mm et mb, alors que le phénotype B provient du génotype bb.

1. $\mathbb{P}(\text{Adrien} = B) = \dfrac{1}{4}$.
2. $\mathbb{P}(1 = B|\ \text{Adrien} = M) = \dfrac{1}{3}$.
3. En décomposant suivant le génotype d'Adrien, on a

$$\begin{aligned}\mathbb{P}(1 = M, 2 = B) &= \mathbb{P}(1 = M, 2 = B,\ \text{Adrien} = bb) + \mathbb{P}(1 = M, 2 = B,\ \text{Adrien} = mb) \\ &\qquad + \mathbb{P}(1 = M, 2 = B,\ \text{Adrien} = mm) \\ &= \mathbb{P}(1 = M, 2 = B|\ \text{Adrien} = mb)\ \mathbb{P}(\text{Adrien} = mb) \\ &= \frac{1}{2}\frac{1}{2}\frac{1}{2} = \frac{1}{8},\end{aligned}$$

et

$$\mathbb{P}(1 = M) = \mathbb{P}(1 = M,\ \text{Adrien} = mm) + \mathbb{P}(1 = M,\ \text{Adrien} = mb) = \frac{1}{2}.$$

On en déduit donc que $\mathbb{P}(2 = B|1 = M) = \dfrac{\mathbb{P}(1 = M, 2 = B)}{\mathbb{P}(1 = M)} = \dfrac{1}{4}$.

4. Si le premier enfant a les yeux marron, on sait déjà qu'Adrien a les yeux marron. On a donc plus d'information dans la question 3) que dans la question 2).

▲

Exercice I.10.

La description de l'espace d'états doit être faite avec soin. On numérote les faces de la première carte a, b, de la deuxième c, d et de la troisième e, f. Les couleurs des faces sont :

$$a = R,\ b = R,\ c = R,\ d = B,\ e = B,\ f = B.$$

Une carte c'est une face exposée et une face cachée : (E, C). L'espace d'état est donc

$$\Omega = \{(a, b), (b, a), (c, d), (d, c), (e, f), (f, e)\}.$$

On munit $(\Omega, \mathcal{P}(\Omega))$ de la probabilité uniforme. Il faut ici se convaincre que les faces sont discernables et donc que $(b, a) \neq (a, b)$. On pourra raisonner en remarquant que le résultat ne change pas si on numérote effectivement les faces des cartes. On a

$$\mathbb{P}(C = B|E = R) = \mathbb{P}(E = R, C = B)/\mathbb{P}(E = R) = \frac{\text{Card}\,\{(c, d)\}}{\text{Card}\,\{(a, b), (b, a), (c, d)\}} = \frac{1}{3}\,.$$

▲

Exercice I.11.

La probabilité, $p_{B|A;C}$, pour que la voiture soit derrière la porte B sachant que vous avez choisi la porte A et que le présentateur ouvre la porte C est égale à

$$\frac{\mathbb{P}(\text{vous avez choisi la porte } A, \text{ la voiture est en } B, \text{ le présentateur ouvre la porte } C)}{\mathbb{P}(\text{vous avez choisi la porte } A, \text{ le présentateur ouvre la porte } C)}.$$

Si vous avez choisi la porte A, et que la voiture est en B, alors le présentateur ouvre la porte C. Votre choix étant indépendant de la position de la voiture, on en déduit que le numérateur est égal à $1/3$. Pour calculer le dénominateur, on décompose suivant les positions possibles de la voiture (la position C est impossible quand le présentateur ouvre la porte C) : pour la position B, on a obtenu que la probabilité est $1/3$, pour la position A, il vient

$$\begin{aligned}&\mathbb{P}(\text{vous avez choisi la porte } A, \text{ la voiture est en } A, \text{ le présentateur ouvre la porte } C)\\ &= \mathbb{P}(\text{le présentateur ouvre la porte } C|\text{vous avez choisi la porte } A, \text{ la voiture est en } A)\\ &\qquad\qquad \mathbb{P}(\text{vous avez choisi la porte } A, \text{ la voiture est en } A).\end{aligned}$$

En notant $q_{x|y}$, la probabilité que le présentateur ouvre la porte x sachant que vous avez choisi la porte y et que la voiture est en y, on a donc obtenu

$$p_{B|A;C} = \frac{1/3}{q_{C|A}/3 + 1/3} = \frac{1}{q_{C|A} + 1}.$$

1. On modélise "au hasard" par $q_{C|A} = 1/2$. Il vient alors $p_{B|A;C} = 2/3$. On a donc intérêt à changer de porte.

2. On a $q_{C|A} = 0$. Il vient alors $p_{B|A;C} = 1$. Si le présentateur ouvre la porte C, on est certain que la voiture est en B. On note $p_{C|A;B}$ la probabilité pour que la voiture soit derrière la porte C sachant que vous avez choisi la porte A et que le présentateur ouvre la porte B. Des calculs similaires donnent

$$p_{C|A;B} = \frac{1}{q_{B|A} + 1}.$$

On en déduit donc que $p_{C|A;B} = 1/2$. On ne perd rien à changer de porte.

3. Comme $q_{C|A}$ et $q_{B|A}$ sont dans $[0, 1]$, on en déduit donc que $p_{B|A;C}$ et $p_{C|A;B}$ sont dans $[1/2, 1]$. Dans tous les cas, vous avez intérêt à changer de porte !

4. Supposons que vous ayez choisi la porte A et que le présentateur ait ouvert la porte C. Votre probabilité de gagner est $p = \frac{1}{2}\, p_{B|A;C} + \frac{1}{2}\, p_{A|A;C} = \frac{1}{2}$, où $p_{A|A;C} = 1 - p_{B|A;C}$ est la probabilité pour que la voiture soit derrière la porte A sachant que vous avez choisi la porte A et que le présentateur ouvre la porte C. Cette stratégie est moins bonne que celle qui consiste à changer de porte, pour laquelle la probabilité de gagner est comprise entre 1/2 et 1.

▲

Exercice I.12.

1. On pose $A_i = \{i \text{ a sa veste}\}$, on a par la formule du crible

$$\begin{aligned}\mathbb{P}\left(\bigcup_{1\leq i\leq n} A_i\right) &= \sum_{p=1}^{n}(-1)^{p+1} \sum_{1\leq i_1<\ldots<i_p\leq n} \mathbb{P}(A_{i_1}\cap\ldots\cap A_{i_p}) \\ &= \sum_{p=1}^{n}(-1)^{p+1} C_n^p \frac{(n-p)!}{n!} = \sum_{p=1}^{n}(-1)^{p+1}\,\frac{1}{p!}\,.\end{aligned}$$

On note

$$\gamma(n) = \text{Card } \{\text{permutations de } \{1,\ldots,n\} \text{ sans point fixe}\}.$$

On a $1 - \dfrac{\gamma(n)}{n!} = \displaystyle\sum_{p=1}^{n}(-1)^{p+1}\,\frac{1}{p!}$ soit $\gamma(n) = n!\displaystyle\sum_{p=0}^{n}(-1)^{p}\,\frac{1}{p!}$.

On en déduit donc le nombre de permutations de $\{1,\ldots,n\}$ sans point fixe (problème formulé par P. R. de Montmort en 1708)[1]

2. On remarque que

$$\pi_n(k) = \frac{\text{Card } \{\text{permutations de } \{1,\ldots,n\} \text{ ayant } k \text{ points fixes}\}}{\text{Card } \{\text{permutations de } \{1,\ldots,n\}\}}$$

[1] Voir L. Takacs (The problem of coincidences, *Arch. Hist. Exact Sci.* 21 :3 (1980), 229-244) pour une étude historique du problème des coïncidences vu par les probabilistes.

Il existe C_n^k possibilités pour les k points fixes. On en déduit

$$\begin{aligned}\pi_n(k) &= \frac{C_n^k \text{ Card } \{\text{permutations de } \{1,\dots,n-k\} \text{ sans point fixe}\}}{\text{Card } \{\text{permutations de } \{1,\dots,n\}\}} \\ &= \frac{C_n^k \gamma(n-k)}{n!} = \frac{1}{k!}\sum_{p=0}^{n-k}(-1)^p\,\frac{1}{p!}\,.\end{aligned}$$

3. On a $\lim_{n\to\infty}\pi_n(k) = \pi_\infty(k) = \dfrac{1}{k!}\,\mathrm{e}^{-1}$. On retrouve la loi de Poisson de paramètre 1. En fait, on peut montrer le résultat suivant sur la vitesse de convergence des probabilités π_n vers π_∞ : pour tout $B \subset \mathbb{N}$, on a

$$\left|\sum_{k\in B}\pi_n(k) - \sum_{k\in B}\pi_\infty(k)\right| \le \frac{2^n}{(n+1)!}.$$

▲

Exercice I.13.

1. $\mathbb{P}((p_r,p_b)) = \dfrac{C_r^{p_r}C_b^{p_b}}{C_{r+b}^{p_r+p_b}} = \dfrac{C_p^{p_r}C_{r+b-p}^{r-p_r}}{C_{r+b}^r}$.
2. $\mathbb{P}((p_r,p_b)) = C_{p_r+p_b}^{p_r}\left(\dfrac{r}{r+b}\right)^{p_r}\left(\dfrac{b}{r+b}\right)^{p_b}$.
3. Dans les deux cas, on trouve $\mathbb{P}((p_r,p_b)) = C_{p_r+p_b}^{p_r}\theta^{p_r}\,(1-\theta)^{p_b}$.

▲

Exercice I.14.

On jette une pièce n fois. Notons E_n l'évènement "on observe au moins trois piles ou trois faces consécutifs" et p_n sa probabilité. Il est facile de calculer les premières valeurs de la suite $(p_n, n \ge 1)$: $p_1 = p_2 = 0$ et $p_3 = 1/4$. Notons A l'évènement "les deux premiers jets donnent deux résultats différents", B l'évènement "les deux premiers jets donnent deux résultats identiques mais différents du résultat du troisième jet" et enfin C l'évènement "les trois premiers jets donnent trois résultats identiques", de sorte que $\{A,B,C\}$ forme une partition de l'ensemble fondamental Ω. Pour $n \ge 3$, on a donc

$$\begin{aligned}p_n &= \mathbb{P}(A)\mathbb{P}(E_n|A) + \mathbb{P}(B)\mathbb{P}(E_n|B) + \mathbb{P}(C)\mathbb{P}(E_n|C) \\ &= \frac{1}{2}\,p_{n-1} + \frac{1}{4}\,p_{n-2} + \frac{1}{4}\,.\end{aligned}$$

Par conséquent, $p_4 = 3/8$ et $p_5 = 1/2$. Eugène s'est donc réjoui un peu vite...

On peut vérifier, à l'aide de la formule de récurrence que

$$p_n = 1 - \frac{1}{2\sqrt{5}}\left[(3+\sqrt{5})\left(\frac{1+\sqrt{5}}{4}\right)^{n-1} + (3-\sqrt{5})\left(\frac{1-\sqrt{5}}{4}\right)^{n-1}\right].$$

On obtient bien sûr que $\lim_{n\to\infty} p_n = 1$. Par exemple on a $p_{10} \simeq 0.826$. ▲

Chapitre II

Variables aléatoires discrètes

II.1 Énoncés

Exercice II.1.
Calculer les fonctions génératrices des lois usuelles : Bernoulli, binomiale, géométrique et Poisson. En déduire leur moyenne et leur variance.

△

Exercice II.2.
On jette 5 dés. Après le premier lancer, on reprend et on lance les dés qui n'ont pas donné de six, jusqu'à ce qu'on obtienne 5 six. Soit X le nombre de lancers nécessaires.

1. Calculer $\mathbb{P}(X \leq n)$ puis $\mathbb{P}(X = n)$ pour tout $n \in \mathbb{N}^*$.
2. Combien de lancers sont nécessaires en moyenne pour obtenir les 5 six ?

△

Exercice II.3.
Deux joueurs lancent une pièce de monnaie parfaitement équilibrée, n fois chacun. Calculer la probabilité qu'ils obtiennent le même nombre de fois pile.

△

Exercice II.4.
Les boîtes d'allumettes de Banach[1]. Ce problème est dû à H. Steinhaus (1887-1972) qui le dédia à S. Banach (1892-1945), lui aussi grand fumeur.

Un fumeur a dans chacune de ses deux poches une boîte d'allumettes qui contient initialement N allumettes. À chaque fois qu'il veut fumer une cigarette, il choisit au hasard une de ses deux poches et prend une allumette dans la boîte qui s'y trouve.

1. Lorsqu'il ne trouve plus d'allumette dans la boite qu'il a choisi, quelle est la probabilité pour qu'il reste k allumettes dans l'autre boite ?

[1] Feller, *An introduction to probability theory and its applications*, Vol. 1. Third ed. (1968). Wiley & Sons.

2. Le fumeur cherche alors une allumette dans son autre poche. Quelle est la probabilité pour que l'autre boîte soit vide, ce qui suffit à gâcher la journée ? Application numérique : $N = 20$ (la boîte plate), $N = 40$ (la petite boîte).

△

Exercice II.5.
On pose 20 questions à un candidat. Pour chaque question k réponses sont proposées dont une seule est la bonne. Le candidat choisit au hasard une des réponses proposées.

1. On lui attribue un point par bonne réponse. Soit X le nombre de points obtenus. Quelle est la loi de X ?
2. Lorsque le candidat donne une mauvaise réponse, il peut choisir à nouveau une des autres réponses proposées. On lui attribue alors $\frac{1}{2}$ point par bonne réponse. Soit Y le nombre de $\frac{1}{2}$ points obtenus lors de ces seconds choix. Quelle est la loi de Y ?
3. Soit S le nombre total de points obtenus. Déterminer k pour que le candidat obtienne en moyenne une note de 5 sur 20.

△

Exercice II.6.
On considère deux urnes contenant chacune R boules rouges et B boules bleues. On note X le nombre de fois où en retirant une boule de chacune des deux urne, elles ont la même couleur. Les tirages sont sans remise.

1. Calculer la probabilité pour que lors du i-ème tirage, les deux boules aient même couleur. En déduire $\mathbb{E}[X]$.
2. Calculer la loi de X. Est-il facile de calculer $\mathbb{E}[X]$ à partir de la loi de X ?

△

Exercice II.7.
Une urne contient N boules numérotées de 1 à N. On tire n boules une à une avec remise. Soit X et Y le plus petit et le plus grand des nombres obtenus.

1. Calculer $\mathbb{P}(X \geq x)$ pour tout $x \in \{1, \cdots, N\}$. En déduire la loi de X.
2. Calculer $\mathbb{P}(Y \leq y)$ pour tout $y \in \{1, \cdots, N\}$. En déduire la loi de Y.
3. Calculer $\mathbb{P}(X > x, Y \leq y)$ pour tout $(x, y) \in \{1, \cdots, N\}^2$. En déduire la loi du couple (X, Y).

△

Exercice II.8.
Soit X une variable aléatoire de loi de Poisson de paramètre $\lambda > 0$ (*i.e.* $\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}, \quad k \geq 0$). Vérifier que $\dfrac{1}{1+X}$ est une variable aléatoire intégrable. Calculer $\mathbb{E}[\frac{1}{1+X}]$. Calculer $\mathbb{E}[\frac{1}{(1+X)(2+X)}]$ et en déduire $\mathbb{E}[\frac{1}{2+X}]$.

△

Exercice II.9.
Soit X, Y, Z trois v.a. à valeurs dans $\mathbb{N}$, indépendantes, avec
$\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$, $0 < p < 1$.
$\mathbb{P}(Y = k) = (1 - a)a^{k-1}, \quad k \in \mathbb{N}^*, \ \ 0 < a < 1$.
$\mathbb{P}(Z = k) = e^{-\theta}\frac{\theta^k}{k!}, \quad k \in \mathbb{N}, \ \ \theta > 0$.

1. Identifier les lois des v.a. Y et Z et donner leur fonction génératrice.
2. Soit U la v.a. égale à 0 si $X = 0$, égale à Y si $X = 1$. Calculer la fonction génératrice de U, en déduire $\mathbb{E}[U]$ et $\mathbb{E}[U^2]$.
3. Soit V la v.a. égale à Y si $X = 0$, égale à Z si $X = 1$. Calculer la fonction génératrice de V, en déduire $\mathbb{E}[V]$ et $\mathbb{E}[V^2]$.

△

Exercice II.10.
Un gardien de nuit doit ouvrir une porte dans le noir, avec n clefs dont une seule est la bonne.

1. Donner la loi de probabilité du nombre X d'essais nécessaires s'il essaie les clefs une à une sans utiliser deux fois la même. Calculer l'espérance et la variance de X.
2. Lorsque le gardien est ivre, il mélange toutes les clefs à chaque tentative. Identifier la loi de X. Rappeler l'espérance et la variance de X.
3. Le gardien est ivre un jour sur trois. Sachant qu'un jour n tentatives ont été nécessaires pour ouvrir la porte, quelle est la probabilité que le gardien ait été ivre ce jour là ? Calculer sa limite.

△

Exercice II.11.
Soit $(X_1, \ldots, X_n)$ une suite de variables aléatoires indépendantes de loi de Bernoulli de paramètre $p \in]0, 1[$. On note $S_n = \sum_{i=1}^{n} X_i$.

1. Calculer la loi de $(X_1, \ldots, X_n)$ conditionnellement à S_n.
2. Calculer la loi de X_i conditionnellement à S_n (pour $n \geq i$).
3. Les variables X_1 et X_2 sont-elles indépendantes conditionnellement à S_n $(n \geq 2)$?

△

Exercice II.12.
On désire modéliser le temps d'attente d'une panne de machine à l'aide de variables aléatoires sans mémoire : la probabilité pour que la machine tombe en panne après la date $k+n$ sachant qu'elle fonctionne à l'instant n est indépendante de n.

1. Montrer que la loi géométrique de paramètre p est sans mémoire : c'est-à-dire que $\mathbb{P}(X > k + n | X > n)$ est indépendant de n.
2. Caractériser toutes les lois des variables aléatoires X à valeurs dans $\mathbb{N}^*$ qui sont sans mémoire. On pourra calculer $\mathbb{P}(X > 1 + n)$ en fonction de $\mathbb{P}(X > 1)$.
3. Caractériser toutes les lois des variables aléatoires X à valeurs dans $\mathbb{N}$ qui sont sans mémoire.

△

Exercice II.13.
Soient X et Y deux variables aléatoires indépendantes de loi géométrique de même paramètre $p \in]0,1[$.

1. Calculer les fonctions génératrices de X et de Y, en déduire celle de $S = X+Y$. Calculer $\mathbb{P}(S = n)$ pour $n \in \mathbb{N}$.
2. Déterminer la loi de X sachant S. En déduire $\mathbb{E}[X|S]$.
3. Vérifier la formule $\mathbb{E}\left[\mathbb{E}[X|S]\right] = \mathbb{E}[X]$.

△

Exercice II.14.
On considère un jeu de pile ou face biaisé : les variables aléatoires $(X_i, i \in \mathbb{N}^*)$ sont indépendantes et de même loi de Bernoulli de paramètre $p \in]0,1[$:

$$\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p.$$

On note T_k l'instant du k-ième succès. $T_1 = \inf\{n \geq 1; X_n = 1\}$ et pour $k \geq 2$,

$$T_k = \inf\{n \geq T_{k-1} + 1; X_n = 1\}.$$

1. Montrer que T_1 et $T_2 - T_1$ sont indépendants.
2. On pose $T_0 = 0$. Montrer que $T_1 - T_0$, $T_2 - T_1, \cdots, T_{k+1} - T_k$ sont indépendantes et de même loi.
3. Calculer $\mathbb{E}[T_k]$ et $\mathrm{Var}(T_k)$.
4. Déterminer $\mathbb{P}(T_k = n)$ directement. Donner la fonction génératrice de T_k. La loi de T_k est appelée loi binomiale négative de paramètre (k, p).

On possède une autre pièce (paramètre ρ). On note τ l'instant du premier succès de la seconde pièce. On décide de jouer avec la première pièce jusqu'au τ-ième succès (c'est-à-dire T_τ).

5. Déterminer la loi de T_τ à l'aide des fonctions génératrices. Reconnaître la loi de T_τ.
6. Retrouver ce résultat à l'aide d'un raisonnement probabiliste sur les premiers temps de succès.

△

Exercice II.15.
Soit $(X_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires de Bernoulli indépendantes et identiquement distribuées :

$$\forall n \in \mathbb{N}^*, \quad \mathbb{P}(X_n = 1) = 1 - \mathbb{P}(X_n = 0) = p \quad \text{où } p \in]0,1[.$$

Soit $T = \inf\{n \geq 2; X_{n-1} = 1, X_n = 0\}$, avec la convention $\inf \emptyset = +\infty$.

1. Montrer que $\mathbb{P}(T < +\infty) = 1$.

2. Calculer la loi de T, et montrer que T a même loi que $U + V$ où U et V sont des variables aléatoires indépendantes de loi géométrique de paramètre respectif p et $1 - p$.
3. Trouver la fonction génératrice de T.
4. Calculer la moyenne et la variance de T.

△

Exercice II.16.
La France a eu 38 médailles dont 13 d'or aux jeux olympiques de Sydney en 2000, sur 928 médailles dont 301 d'or. On estime la population à 6 milliards dont 60 millions en France. Peut-on dire que les sportifs de haut niveau sont uniformément répartis dans la population ?

△

Exercice II.17.
Les gènes se présentent le plus souvent en paire et sous deux formes d'allèles que nous noterons A et B. Cela donne donc trois génotypes possibles : AA, AB et BB. Chaque individu reçoit au hasard un gène de chacun de ses parents. Chaque allèle composant le gène d'un des parents ayant la probabilité $1/2$ d'être transmis à l'enfant.

On suppose que les génotypes des deux parents sont indépendants et de même loi : soit x la probabilité d'avoir le génotype AA, $2y$ celle d'avoir le génotype AB, z celle d'avoir le génotype BB (remarque évidente : $x + 2y + z = 1$).

1. Calculer la probabilité de chacun des génotypes pour un enfant.
2. Vérifier que, si $x = y = z = 1/4$, alors les génotypes de l'enfant ont les mêmes probabilités que ceux des parents.
3. Calculer la probabilité de chacun des génotypes pour un enfant de la seconde génération. Que constatez-vous ? La loi de répartition des génotypes s'appelle la loi de Hardy-Weinberg.

△

Exercice II.18.
Optimisation de coûts. Le coût de dépistage de la maladie M à l'aide d'un test sanguin est c. La probabilité qu'une personne soit atteinte de la maladie M est p. Chaque personne est malade indépendamment des autres. Pour effectuer un dépistage parmi N personnes, on propose les deux méthodes suivantes :

- Un test par personne.

- On mélange les prélèvements sanguins de n personnes et on effectue le test. Si on détecte la maladie M dans le mélange, alors on refait un test sanguin pour chacune des n personnes.

Calculer le coût de la première stratégie, puis le coût moyen de la seconde stratégie. On supposera $np \ll 1$, et on montrera que $n \simeq p^{-1/2}$ est une taille qui minimise correctement le coût du dépistage. Quelle méthode choisissez-vous ? Illustrer vos conclusions pour le cas où $p = 1\%$.

Cette démarche a été utilisée initialement par R. Dorfman[2], durant la Seconde Guerre mondiale, dans un programme commun avec l'United States Public Health Service et le

[2]The detection of defective numbers of large populations, *Ann. Math. Statist.* **14**, 436-440 (1943).

Selective Service System, afin de réformer les futurs appelés du contingent ayant la syphilis (taux de syphilis en Amérique du nord : de l'ordre de 1 à 2 pour 100 dans les années 1940 et de l'ordre de 5 à 20 pour 100 000 en 1990). De nombreuses généralisations ont ensuite été étudiées.

△

Exercice II.19.
On considère le modèle d'Ehrenfest : N particules sont placées dans deux recipients A et B. À chaque instant on choisit au hasard une particule et on la change de recipient. Soit X_n le nombre de particules dans le recipient A à l'instant n.

1. Montrer que $p_n(i,j) = \mathbb{P}(X_{n+1} = j | X_n = i)$ ne dépend pas de n.
2. On suppose que X_0 est distribuée suivant la loi binomiale $\mathcal{B}(N, 1/2)$. Déterminer la loi de X_1. En déduire la loi de X_n.

Ce modèle a été introduit en 1907 par les physiciens autrichiens Tatiana et Paul Ehrenfest[3] pour décrire la diffusion d'un gaz placé dans deux récipients à températures différentes mis en communication. On peut alors montrer, qu'après un certain temps, il y a très peu de chance de revenir à l'état initial.

△

Exercice II.20.
On souhaite modéliser des phénomènes qui se répètent à des temps alétoires indépendants et identiquement distribués, par exemple : une machine tombe en panne, l'arrivée de particules α à un compteur Geiger, la chute de météorites, etc. On ne peut pas prédire la date où ces phénomènes vont se produire mais on peut estimer la probabilité que ces phénomènes se produisent à un instant n donné. C'est–à–dire,

$$\mathbb{P}(\text{le phénomène } H \text{ a lieu à l'instant } n) = v_n, \qquad n \in \mathbb{N}.$$

L'objetif de cet exercice est d'utiliser les fonctions génératrices pour calculer ces probabilités.

On note T_n l'instant de n–ième occurrence du phénomène H. On pose $X_1 = T_1$, et pour $n \geq 2$,

$$X_n = T_n - T_{n-1}.$$

La variable X_n, pour $n \geq 2$, représente le temps écoulé entre la n–ième et la $(n-1)$–ième occurrence du phénomène H. On suppose que les variables aléatoires $(X_n, n \geq 1)$ sont indépendantes. On suppose de plus que les variables aléatoires $(X_n, n \geq 2)$ sont identiquement distribuées à valeurs dans $\mathbb{N}^*$ et que X_1 est à valeurs dans $\mathbb{N}$ avec une loi qui est éventuellement différente de celle de X_2.

1. On pose $b_k = \mathbb{P}(X_1 = k)$, $k \in \mathbb{N}$. Montrer que

$$v_n = b_n u_0 + b_{n-1} u_1 + b_{n-2} u_2 + \cdots + b_0 u_n,$$

avec $u_0 = 1$ et

$$\begin{aligned} u_j &= \mathbb{P}(\text{le phénomène } H \text{ se produit au temps } j+k | X_1 = k) \\ &= \mathbb{P}(X_2 + \cdot + X_i = j \text{ pour un } i \in \{2, \ldots, j+1\}) \qquad j \geq 1. \end{aligned}$$

[3]The conceptual foundations of the statistical approach in mechanics, *Ithaca, NY, Cornell Univ. Press* (1959)

2. On considère $B(s) = \sum_{n=0}^{\infty} s^n \mathbb{P}(X_1 = n)$ la fonction génératrice de X_1, et on pose

$$U(s) = \sum_{n=0}^{\infty} s^n u_n, \qquad |s| < 1.$$

Déduire de la question 1, que

$$V(s) = \sum_{n=0}^{\infty} s^n v_n = B(s)U(s), \qquad |s| < 1.$$

3. Démontrer que pour tout $n \geq 1$, on a

$$u_n = f_1 u_{n-1} + f_2 u_{n-2} + \cdots + f_n u_0,$$

avec $f_i = \mathbb{P}(X_2 = i)$. En déduire que si $F(s) = \sum_{n=1}^{\infty} s^n f_n$ est la fonction génératrice de X_2, alors

$$U(s) = \frac{1}{1 - F(s)}, \qquad |s| < 1.$$

4. On suppose que X_1 est une variable aléatoire de loi définie par

$$\mathbb{P}(X_1 = k) = (1-p)^k p, \qquad k \in \mathbb{N},\ p \in]0,1[.$$

On suppose que les variables aléatoires $X_i, i \geq 2$ suivent une loi géométrique de paramétre p. Vérifier que

$$B(s) = p\, \frac{1 - F(s)}{1 - s}, \qquad |s| < 1.$$

Déduire des questions précédentes que $v_n = p$. Les probabilités v_n sont stationnaires.

Ce modèle simple sert à modèliser les arrivées des particules α à un compteur Geiger. Les temps $(X_j, j \geq 2)$ suivent une loi géomètrique de paramètre p. À l'instant où l'on commence à observer il est possible qu'il y ait une arrivée d'une particule α avec probabilité p. Il est donc nécessaire de permettre à X_1 d'avoir une distribution différente des autres X_j. Ce qui explique le choix de la distribution de X_1 dans la dernière question. Le résultat obtenu nous confirme que la probabilité d'observer l'arrivée d'une particule α au compteur Geiger ne varie pas avec le temps d'observation.

En 1913, Victor Hess[4] a découvert que la radiation augmente avec l'altitude. Ainsi, le nombre moyen de particules α capté par un compteur Geiger au niveau du sol est de 12 par minute tandis qu'à 36000 pieds il est de 360 par minute. On prend le dixième de seconde comme unité de temps. Alors, la probabilité qu'il y ait une arrivée d'une particule α à l'instant n à un compteur Geiger situé au sol est de 1/50 (une particule arrive en moyenne chaque 50 dixièmes de seconde) et de 30/50 pour un compteur Geiger situé à 36000 pieds d'altitude (30 particules arrivent en moyenne chaque 50 dixièmes de seconde).

△

[4] *Cosmic Radiation and Its Biological Effects,* Victor Hess et Jakob Eugster, 1940, 2d ed. 1949

II.2 Corrections

Exercice II.1.

Loi	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	$\phi_X(z)$
Bernoulli $p \in [0,1]$	p	$p(1-p)$	$1-p+pz$
binomiale $(n,p) \in \mathbb{N} \times [0,1]$	np	$np(1-p)$	$(1-p+pz)^n$
géométrique $p \in]0,1]$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\frac{pz}{1-(1-p)z}$
Poisson $\theta \in]0,\infty[$	θ	θ	$\mathrm{e}^{-\theta(1-z)}$

▲

Exercice II.2.

Soit X_i le nombre de jets nécessaires pour que le $i^{\text{ème}}$ dé amène un six pour la première fois. Les v.a.d. $(X_i)_{1\leq i\leq 5}$ sont indépendantes et suivent une loi géomètrique de paramètre $1/6$. Comme $X = \max_{1\leq i\leq 5} X_i$ on obtient

$$\begin{aligned}
\mathbb{P}(X \leq k) &= \mathbb{P}(X_i \leq k, \forall 1 \leq i \leq 5) \\
&= \prod_{i=1}^{5} \mathbb{P}(X_i \leq k) \quad \text{par indépendance,} \\
&= (\mathbb{P}(X_i \leq k))^5 \quad \text{car les lois sont identiques,} \\
&= (1 - \mathbb{P}(X_i \geq k+1))^5 \\
&= \left(1 - \left(\frac{5}{6}\right)^k\right)^5 .
\end{aligned}$$

Cette formule est valable pour $k = 0$. En particulier $\mathbb{P}(X = 0) \leq \mathbb{P}(X \leq 0) = 0$. On en déduit que pour $k \geq 1$,

$$\begin{aligned}
\mathbb{P}(X = k) &= \mathbb{P}(X \leq k) - \mathbb{P}(X \leq k-1) \\
&= \left(1 - \left(\frac{5}{6}\right)^k\right)^5 - \left(1 - \left(\frac{5}{6}\right)^{k-1}\right)^5 .
\end{aligned}$$

Sous réserve d'existence, $\mathbb{E}[X] = \sum_{k=1}^{+\infty} k\mathbb{P}(X=k)$.

$$\begin{aligned}
\sum_{k=1}^{n} k\mathbb{P}(X=k) &= \sum_{k=1}^{n} k\left[\left(1-\left(\frac{5}{6}\right)^k\right)^5 - \left(1-\left(\frac{5}{6}\right)^{k-1}\right)^5\right] \\
&= \sum_{k=1}^{n}\left[k\left(1-\left(\frac{5}{6}\right)^k\right)^5 - (k-1)\left(1-\left(\frac{5}{6}\right)^{k-1}\right)^5 - \left(1-\left(\frac{5}{6}\right)^{k-1}\right)^5\right] \\
&= n\left(1-\left(\frac{5}{6}\right)^n\right)^5 - \sum_{k=0}^{n-1}\left(1-\left(\frac{5}{6}\right)^k\right)^5 \\
&= n\left[\left(1-\left(\frac{5}{6}\right)^n\right)^5 - 1\right] + \sum_{k=0}^{n-1}\left[1-\left(1-\left(\frac{5}{6}\right)^k\right)^5\right].
\end{aligned}$$

On a $\lim_{n\to+\infty} n\left[\left(1-\left(\frac{5}{6}\right)^n\right)^5 - 1\right] = \lim_{n\to+\infty} -5n\left(\frac{5}{6}\right)^n = 0$. Donc X est intégrable et on a

$$\mathbb{E}[X] = 5\frac{1}{1-\frac{5}{6}} - 10\frac{1}{1-\left(\frac{5}{6}\right)^2} + 10\frac{1}{1-\left(\frac{5}{6}\right)^3} - 5\frac{1}{1-\left(\frac{5}{6}\right)^4} + \frac{1}{1-\left(\frac{5}{6}\right)^5} \simeq 13,02.$$

▲

Exercice II.3.

Soient X et Y les variables aléatoires discrètes qui représentent le nombre de piles obtenus par chacun des joueurs au cours des n lancers. Les v.a.d. X et Y sont **indépendantes** et suivent des loi **binomiales** $\mathcal{B}(n, 1/2)$. On a, en utilisant la formule de décomposition puis l'indépendance de X et Y,

$$\begin{aligned}
\mathbb{P}(X=Y) &= \sum_{k=0}^{n}\mathbb{P}(X=k, Y=k) = \sum_{k=0}^{n}\mathbb{P}(X=k)\mathbb{P}(Y=k) \\
&= \sum_{k=0}^{n} C_n^k \frac{1}{2^n} C_n^k \frac{1}{2^n} = \frac{1}{2^{2n}}\sum_{k=0}^{n}(C_n^k)^2 \\
&= \frac{C_{2n}^n}{2^{2n}}.
\end{aligned}$$

Pour démontrer l'égalité $\sum_{k=0}^{n}(C_n^k)^2 = C_{2n}^n$, on peut calculer de deux manières différentes le coefficient de x^n dans le polynôme $(1+x)^{2n} = (1+x)^n(1+x)^n$. ▲

Exercice II.4.

On suppose que la phrase "il choisit au hasard une de ses deux poches" signifie que le choix de la poche est une réalisation d'une variable aléatoire de Bernoulli de paramètre p (on code par exemple 0 pour la poche de gauche et 1 pour la poche de droite) et que chaque choix est indépendant des choix précédents. En absence d'information supplémentaire, il sera naturel de choisir $p = 1/2$. On note $(X_n, n \geq 1)$ une suite de v.a. de Bernoulli de paramètre $p \in]0,1[$ indépendantes. La variable X_n représente le choix de la poche lors du n-ième tirage (i.e. de la n-ième cigarette).

1. Lorsque le fumeur s'aperçoit que la boîte qu'il a choisie est vide, s'il reste k allumettes dans l'autre boîte, c'est qu'il a déjà fumé $2N-k$ cigarettes, et donc il a cherché $2N+1-k$ fois une allumette. En particulier l'évènement "quand le fumeur ne trouve plus d'allumette dans une boite, il reste k allumettes dans l'autre boite", de probabilité p_k, est donc égal à la réunion des deux évènements exclusifs suivants : "à la $2N-k+1$-ième cigarette, le fumeur a choisi la poche de droite pour la $N+1$-ième fois" et "à la $2N-k+1$-ième cigarette, le fumeur a choisi la poche de gauche pour la $N+1$-ième fois", c'est-à-dire à la réunion de $\{\sum_{i=1}^{2N+1-k} X_i = N+1, X_{2N+1-k}=1\}$ et de $\{\sum_{i=1}^{2N+1-k} X_i = N-k, X_{2N+1-k}=0\}$. On en déduit donc

$$\begin{aligned} p_k &= \mathbb{P}(\sum_{i=1}^{2N+1-k} X_i = N+1, X_{2N+1-k}=1) + \mathbb{P}(\sum_{i=1}^{2N+1-k} X_i = N-k, X_{2N+1-k}=0) \\ &= \mathbb{P}(\sum_{i=1}^{2N-k} X_i = N)\mathbb{P}(X_{2N+1-k}=1) + \mathbb{P}(\sum_{i=1}^{2N-k} X_i = N-k)\mathbb{P}(X_{2N+1-k}=0) \\ &= C_{2N-k}^N [p^{N+1}(1-p)^{N-k} + (1-p)^{N+1}p^{N-k}]. \end{aligned}$$

2. La probabilité cherchée est $p_0 = C_{2N}^N p^N (1-p)^N$. Si on suppose que $p=1/2$, alors $p_0 = C_{2N}^N / 2^{2N}$. Il vient $p_0 \simeq 12.5\%$ pour $N=20$ et $p_0 \simeq 8.9\%$ pour $N=40$.

Comme $\sum_{k=0}^N p_k = 1$, on en déduit, avec $p=1/2$, la relation suivante non triviale sur les coeffcicients binomiaux :

$$\sum_{k=0}^{N} C_{2N-k}^N 2^k = 2^{2N}.$$

▲

Exercice II.5.

1. On note X_i le résultat du candidat à la question i. Les v.a.d. $(X_i, 1 \le i \le 20)$ sont des v.a.d. de Bernoulli **indépendantes** et de même paramètre $\mathbb{P}(X_i=1) = 1/k$. Comme $X = \sum_{i=1}^{20} X_i$, la loi de X est donc la loi binomiale $\mathcal{B}(20, 1/k)$.
2. Si $X_i = 0$, on note Y_i le résultat du candidat à la question i lors du second choix. Si $X_i = 1$, on pose $Y_i = 0$. Ainsi $Y = \sum_{i=1}^{20} Y_i$ représente le nombre de bonnes réponses au deuxième choix. Les v.a.d. $(Y_i, 1 \le i \le 20)$ sont des v.a.d. **indépendantes** et de même loi. Comme elles sont à valeurs dans $\{0,1\}$, il s'agit de v.a. de Bernoulli. On détermine leur paramètre :

$$\mathbb{P}(Y_i=1) = \mathbb{P}(Y_i=1, X_i=0) = \mathbb{P}(Y_i=1|X_i=0)\mathbb{P}(X_i=0) = \frac{1}{k-1}\frac{k-1}{k} = \frac{1}{k}.$$

   Donc Y suit la loi binomiale $\mathcal{B}(20, 1/k)$. On remarquera que les v.a.d. X et Y ne sont pas indépendantes.
3. La note obtenue à la fin est $S = X + \frac{1}{2}Y$. Il vient $\mathbb{E}[S] = \mathbb{E}[X] + \frac{1}{2}\mathbb{E}[Y] = \frac{30}{k}$. Il faut prendre $k=6$.

▲

Exercice II.6.

1. La probabilité pour que lors du i-ème tirage, les deux boules aient même couleur est :
$$p = \frac{R^2}{(R+B)^2} + \frac{B^2}{(R+B)^2}.$$
On définit les variables aléatoires X_i par $X_i = 1$ si lors du i-ème tirage, les deux boules ont même couleur et $X_i = 0$ sinon. On a $X = \sum_{i=1}^{R+B} X_i$. On en déduit par linéarité que
$$\mathbb{E}[X] = \sum_{i=1}^{R+B} \mathbb{E}[X_i] = (R+B)p = \frac{R^2+B^2}{R+B}.$$
2. Pour $k = R+B-2d$ et $d \in \{0, \ldots, \min(R,B)\}$ représentant le nombre de boules rouge (et donc bleu) de l'urne 1 n'ayant pas la même couleur que la boule de l'urne 2 du même tirage, on a
$$\mathbb{P}(X=k) = \mathbb{P}(X = R+B-2d) = \frac{R!^2\ B!^2}{(R+B)!(R-d)!(B-d)!d!^2}.$$

▲

Exercice II.7.

On note par $X_1, X_2, \cdots,$ et X_n les résultats des n boules.

1. Soit $x \in \{1, \cdots, N\}$. L'évènement $\{X \geq x\}$ est réalisé si et seulement si les n nombres, $X_1, \cdots, X_n$, sont supérieurs ou égaux à x. On a $\mathbb{P}(X \geq x) = \mathbb{P}(X_1 \geq x, \cdots, X_n \geq x)$. Comme les variables aléatoires $X_1, \cdots, X_n$ sont indépendantes et identiquement distribuées, on déduit que $\mathbb{P}(X \geq x) = \mathbb{P}(X_1 \geq x)^n = \frac{(N-x+1)^n}{N^n}$. Pour la loi de X, on a pour tout $x \in \{1, \cdots, N-1\}$,
$$\mathbb{P}(X=x) = \mathbb{P}(X \geq x) - \mathbb{P}(X \geq x+1) = \frac{(N-x+1)^n - (N-x)^n}{N^n}.$$
Cette formule est encore valable pour $x = N$, car $\mathbb{P}(X=N) = \frac{1}{N^n}$.
2. Soit $y \in \{1, \cdots, N\}$, on a
$$\mathbb{P}(Y \leq y) = \mathbb{P}\left(X_i \leq y, \forall i \in \{1, \cdots, n\}\right) = \mathbb{P}(X_1 \leq y)^n = \frac{y^n}{N^n}.$$
Pour la loi de Y, on a $\mathbb{P}(Y=y) = \mathbb{P}(Y \leq y) - \mathbb{P}(Y \leq y-1) = \frac{y^n - (y-1)^n}{N^n}$, pour tout $y \in \{2, \cdots, N\}$. Cette formule est encore valable pour $y = 1$ car $\mathbb{P}(Y=1) = \frac{1}{N^n}$.
3. Soit $(x,y) \in \{1, \cdots, N\}^2$. Si $x \geq y$, $\mathbb{P}\left((X > x) \cap (Y \leq y)\right) = \mathbb{P}\left(x < X \leq Y \leq y\right)) = 0$. Si $x < y$, on a
$$\begin{aligned}\mathbb{P}\left(X > x, Y \leq y\right) &= \mathbb{P}\left(x < X \leq Y \leq y\right) \\ &= \mathbb{P}\left(x < X_i \leq y, \forall i \in \{1, \cdots, n\}\right) \\ &= \mathbb{P}\left(x < X_1 \leq y\right)^n = \frac{(y-x)^n}{N^n}.\end{aligned}$$
Cette formule est encore valable pour $x = y$. Pour la loi du couple (X,Y), on a

– si $x > y$, $\mathbb{P}(X = x, Y = y) = 0$ car on a toujours $X \leq Y$.
– si $x = y$, $\mathbb{P}(X = x, Y = y) = \mathbb{P}\left(X_i = x, \forall i \in \{1, \cdots, n\}\right) = \dfrac{1}{N^n}$.
– si $x < y$,

$$\begin{aligned}\mathbb{P}(X = x, Y = y) &= \mathbb{P}(X > x - 1, Y = y) - \mathbb{P}(X > x, Y = y)\\ &= \mathbb{P}(X > x - 1, Y \geq y) - \mathbb{P}(X > x - 1, Y \geq y - 1)\\ &\quad - \mathbb{P}(X > x, Y \geq y) + \mathbb{P}(X > x, Y \geq y - 1)\\ &= \frac{(y - x + 1)^n - 2(y - x)^n + (y - x - 1)^n}{N^n}.\end{aligned}$$

▲

Exercice II.8.

Comme $X \geq 0$ p.s., on en déduit que $\left|\dfrac{1}{1+X}\right| = \dfrac{1}{1+X} \leq 1$ p.s. Donc la v.a. $\dfrac{1}{1+X}$ est **intégrable**. On a

$$\begin{aligned}\mathbb{E}\left[\frac{1}{1+X}\right] &= \sum_{k=0}^{\infty} \frac{1}{1+k} \mathbb{P}\left(X = k\right)\\ &= \sum_{k=0}^{\infty} \frac{1}{1+k} e^{-\lambda} \frac{\lambda^k}{k!}\\ &= \frac{e^{-\lambda}}{\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{(k+1)!}\\ &= \frac{1 - e^{-\lambda}}{\lambda}.\end{aligned}$$

De même $\dfrac{1}{(1+X)(2+X)}$ est intégrable, et on a

$$\begin{aligned}\mathbb{E}\left[\frac{1}{(1+X)(2+X)}\right] &= \sum_{k=0}^{\infty} \frac{1}{(1+k)(2+k)} \mathbb{P}\left(X = k\right)\\ &= \sum_{k=0}^{\infty} \frac{1}{(1+k)(2+k)} e^{-\lambda} \frac{\lambda^k}{k!}\\ &= \frac{e^{-\lambda}}{\lambda^2} \sum_{k=0}^{\infty} \frac{\lambda^{k+2}}{(k+2)!}\\ &= \frac{1 - e^{-\lambda} - \lambda e^{-\lambda}}{\lambda^2}.\end{aligned}$$

Comme $\dfrac{1}{1+X} - \dfrac{1}{2+X} = \dfrac{1}{(1+X)(2+X)}$, on déduit que $\dfrac{1}{2+X}$ est intégrable et que

$$\begin{aligned}\mathbb{E}\left[\frac{1}{2+X}\right] &= \mathbb{E}\left[\frac{1}{1+X}\right] - \mathbb{E}\left[\frac{1}{(1+X)(2+X)}\right]\\ &= \frac{1 - e^{-\lambda}}{\lambda} - \frac{1 - e^{-\lambda} - \lambda e^{-\lambda}}{\lambda^2}\\ &= \frac{\lambda - 1 + e^{-\lambda}}{\lambda^2}.\end{aligned}$$

▲

Exercice II.9.

1. Y suit une loi géomètrique de paramètre $1-a$, $\phi_Y(z) = \frac{(1-a)z}{1-az}$. Z suit une loi de Poisson de paramètre θ, $\phi_Z(z) = e^{-\theta(1-z)}$.
2. On a $U = \mathbf{1}_{\{X=0\}} + Y\mathbf{1}_{\{X=1\}}$, $z^U = \mathbf{1}_{\{X=0\}} + z^Y\mathbf{1}_{\{X=1\}}$.

$$\begin{aligned}\phi_U(z) &= \mathbb{E}(z^U)\\ &= \mathbb{E}(\mathbf{1}_{X=0} + z^Y\mathbf{1}_{X=1})\\ &= 1-p+p\phi_Y(z)\\ &= 1-p+\frac{p(1-a)z}{1-az}\end{aligned}$$

Comme

$$\phi'_U(z) = \frac{p(1-a)}{(1-az)^2} \quad \text{et} \quad \phi''_U(z) = \frac{2p(1-a)a}{(1-az)^3},$$

on déduit

$$\begin{aligned}\mathbb{E}[U] &= \phi'_U(1) = \frac{p}{1-a},\\ \mathbb{E}[U(U-1)] &= \phi''_U(1) = \frac{2pa}{(1-a)^2},\\ \text{et } \mathbb{E}[U^2] &= \frac{2pa}{(1-a)^2} + \frac{p}{1-a} = \frac{p(a+1)}{(1-a)^2}.\end{aligned}$$

3. On a $V = Y\mathbf{1}_{\{X=0\}} + Z\mathbf{1}_{\{X=1\}}$, $z^V = \mathbf{1}_{\{X=0\}}z^Y + \mathbf{1}_{\{X=1\}}z^Z$.

$$\begin{aligned}\phi_V(z) &= \mathbb{E}(z^V)\\ &= \mathbb{E}(\mathbf{1}_{\{X=0\}}z^Y + \mathbf{1}_{\{X=1\}}z^Z)\\ &= (1-p)\phi_Y(z) + p\phi_Z(z)\\ &= \frac{(1-p)(1-a)z}{1-az} + pe^{\theta(z-1)}\end{aligned}$$

Comme

$$\phi'_V(z) = \frac{(1-p)(1-a)}{(1-az)^2} + p\theta e^{\theta(z-1)} \quad \text{et} \quad \phi''_V(z) = \frac{2(1-p)(1-a)a}{(1-az)^3} + p\theta^2 e^{\theta(z-1)},$$

on déduit

$$\begin{aligned}\mathbb{E}[V] &= \phi'_V(1) = \frac{1-p}{1-a} + p\theta,\\ \mathbb{E}[V(V-1)] &= \phi''_V(1) = \frac{2(1-p)a}{(1-a)^2} + p\theta^2,\\ \text{et } \mathbb{E}[V^2] &= \frac{(1-p)(a+1)}{(1-a)^2} + p\theta(1+\theta).\end{aligned}$$

▲

Exercice II.10.

Soit A_k l'événement : "le gardien choisit la bonne clef lors de la k-ième tentative".

1. Comme le gardien essaie les clefs une et une seule fois (c'est un tirage sans remise), les événements $(A_k, k \geq 1)$ ne sont pas indépendants. On a $X \in \{1, \cdots, n\}$ et pour $1 \leq k \leq n$,
$$\{X = k\} = A_1^c \cap A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k.$$
On en déduit que
$$\begin{aligned} \mathbb{P}(X = k) &= \mathbb{P}(A_1^c \cap A_2^c \cap \cdots \cap A_{k-1}^c \cap A_k) \\ &= \mathbb{P}(A_k | A_1^c \cap \cdots \cap A_{k-1}^c)\mathbb{P}(A_{k-1}^c | A_1^c \cap \cdots \cap A_{k-2}^c) \cdots \mathbb{P}(A_2^c | A_1^c)\mathbb{P}(A_1^c) \\ &= \frac{1}{n-(k-1)} \frac{n-(k-1)}{n-(k-2)} \cdots \frac{n-2}{n-1} \frac{n-1}{n} = \frac{1}{n}. \end{aligned}$$
La loi de X est la loi uniforme sur $\{1, \cdots, n\}$. On a
$$\mathbb{E}[X] = \sum_{k=1}^{n} k\mathbb{P}(X = k) = \frac{1}{n}\sum_{k=1}^{n} k = \frac{n+1}{2},$$
$$\mathbb{E}[X^2] = \sum_{k=1}^{n} k^2\mathbb{P}(X = k) = \frac{1}{n}\sum_{k=1}^{n} k^2 = \frac{(n+1)(2n+1)}{6},$$
$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}.$$
2. Maintenant le gardien essaie les clefs éventuellement plusieurs fois chacune (on parle de tirage avec remise), les événements $(A_k, k \geq 1)$ sont indépendants et de probabilité $\frac{1}{n}$. La loi de X est donc la loi géométrique de paramètre $\frac{1}{n}$. On a $X \in \mathbb{N}^*$ et pour $k \geq 1$,
$$\mathbb{P}(X = k) = (1 - \frac{1}{n})^{k-1}\frac{1}{n}.$$
On sait que si une variable aléatoire X suit une loi géométrique de paramètre p alors $\mathbb{E}[X] = \frac{1}{p}$ et $\mathrm{Var}(X) = \frac{1-p}{p^2}$. En remplaçant p par $\frac{1}{n}$, on trouve $\mathbb{E}[X] = n$ et $\mathrm{Var}(X) = n(n-1)$.
3. On note par I l'événement : "le gardien est ivre". Par la formule de Bayes, on obtient
$$\begin{aligned} \mathbb{P}(I | X = n) &= \frac{\mathbb{P}(X = n|I)\mathbb{P}(I)}{\mathbb{P}(X = n|I)\mathbb{P}(I) + \mathbb{P}(X = n|I^c)\mathbb{P}(I^c)} \\ &= \frac{1}{1 + 2(\frac{n}{n-1})^{n-1}} = \frac{1}{1+2\,\mathrm{e}} + O(\frac{1}{n}). \end{aligned}$$

▲

Exercice II.11.

1. On a
$$\mathbb{P}\left(X_1 = k_1, \ldots, X_n = k_n \mid S_n = k\right) = \begin{cases} 0 \quad \text{si} \quad \sum_{i=1}^{n} k_i \neq k, \\ \dfrac{\prod_{i;k_i=1} p \prod_{j;k_j=0}(1-p)}{C_n^k p^k (1-p)^{n-k}} = \dfrac{1}{C_n^k} \quad \text{sinon.} \end{cases}$$
On en déduit que la loi conditionnelle de $(X_1, \ldots, X_n)$ sachant S_n est la loi uniforme sur $\{(k_1, \ldots, k_n) \in \{1,0\}^n; \sum_{i=1}^n k_i = k\}$.
2. La variable X_i prend les valeurs 0 ou 1. Il en est de même quand on conditionne par rapport à S_n. La loi de X_i conditionnellement à S_n est donc une loi de Bernoulli de paramètre $\tilde{p}$. Comme $\mathbb{P}(X_i = 1|S_n = k) = C_{n-1}^{k-1}/C_n^k = k/n$ pour $k \geq 1$, et $\mathbb{P}(X_i = 1|S_n = 0) = 0$, on en déduit que $\tilde{p} = S_n/n$.
3. On a $\mathbb{P}(X_1 = 1, X_2 = 1|S_n) = S_n(S_n - 1)/(n(n-1))$. En particulier $\mathbb{P}(X_1 = 1, X_2 = 1|S_n) \neq \mathbb{P}(X_1 = 1|S_n)\mathbb{P}(X_2 = 1|S_n)$. Conditionnellement à S_n, les variables X_1 et X_2 ne sont pas indépendantes.

▲

Exercice II.12.

1. Soit X une v.a.d géométrique de paramètre $p \in]0,1[$. On a pour $n \in \mathbb{N}^*$,
$$\mathbb{P}(X > n) = \sum_{k \geq n+1} \mathbb{P}(X = k) = \sum_{k \geq n+1} p(1-p)^{k-1} = p(1-p)^n \sum_{k=0}^{\infty} (1-p)^k = (1-p)^n.$$
On en déduit que
$$\mathbb{P}(X > k + n|X > n) = \mathbb{P}(X > k + n)/\mathbb{P}(X > n) = (1-p)^k = \mathbb{P}(X > k).$$
2. Comme $\mathbb{P}(X > 1 + n|X > n)$ est indépendant de n, on a
$$\mathbb{P}(X > 1 + n|X > n) = \mathbb{P}(X > 1|X > 0) = \mathbb{P}(X > 1) = q \quad \text{car } X \in \mathbb{N}^* \text{ p.s.},$$
où $q \in [0,1]$. Si $q = 0$, alors $\mathbb{P}(X > n) = 0$, et les probabilités conditionnelles ne sont pas définies. On a donc $q > 0$. Il vient $\mathbb{P}(X > n + 1) = q\mathbb{P}(X > n)$ i.e $\mathbb{P}(X > n) = q^n \mathbb{P}(X > 0) = q^n$. Cela implique que pour $n \in \mathbb{N}^*$,
$$\mathbb{P}(X = n) = \mathbb{P}(X > n - 1) - \mathbb{P}(X > n) = (1-q)q^{n-1} = p(1-p)^{n-1},$$
où $p = 1 - q$. Comme $\mathbb{P}(X \in \mathbb{N}^*) = 1$, on en déduit que $\sum_{n \in \mathbb{N}^*} \mathbb{P}(X = n) = 1$ i.e. $p > 0$. On reconnaît, pour X, la loi géométrique de paramètre $p \in]0,1[$.
3. On note $\alpha = \mathbb{P}(X > 0)$. On vient de traiter le cas $\alpha = 1$. On suppose $\alpha \in]0,1[$. (Le cas $\alpha = 0$ implique $\mathbb{P}(X > n) = 0$, et donc le caractère sans mémoire n'a pas de sens.) On a $\mathbb{P}(X > 1 + n|X > n) = \mathbb{P}(X > 1|X > 0) = q$. On en déduit que pour $n \geq 0$, $\mathbb{P}(X > n + 1) = q\mathbb{P}(X > n)$ i.e $\mathbb{P}(X > n + 1) = q^{n+1}\mathbb{P}(X > 0)$. Donc, il vient $\mathbb{P}(X = 0) = 1 - \alpha$ et pour $n \geq 1$, $\mathbb{P}(X = n) = \mathbb{P}(X > n-1) - \mathbb{P}(X > n) = \alpha p(1-p)^{n-1}$, où $p = 1 - q = \mathbb{P}(X = 1|X > 0)$ et $p \in]0,1[$. On peut remarquer que X a même loi que YZ, où Y est une v.a. de Bernoulli de paramètre α, et Z est une v.a. indépendante de Y de loi géométrique de paramètre p. En ce qui concerne les temps de panne de machines, soit la machine est en panne (probabilité $1 - \alpha$), soit la machine est en état de marche. Dans ce dernier cas, la loi du temps de panne est une géométrique.

▲

Exercice II.13.

1. On note par ϕ_X la fonction génératrice de X. On a

$$\phi_X(z) = \mathbb{E}[z^X] = \sum_{k=1}^{\infty} z^k \mathbb{P}(X=k) = \sum_{k=1}^{\infty} z^k p(1-p)^{k-1} = \frac{pz}{1-(1-p)z}.$$

On note par ϕ_Y la fonction génératrice de Y et ϕ_S la fonction génératrice de S. Il est clair que $\phi_Y(z) = \frac{pz}{1-(1-p)z}$. Par indépendance, on a $\phi_S(z) = \phi_X(z)\phi_Y(z) = \frac{p^2z^2}{(1-(1-p)z)^2}$. En utilisant le développement $\frac{1}{(1-x)^2} = \sum_{k=0}^{\infty}(k+1)x^k$, on déduit que

$$\phi_S(z) = \sum_{k=0}^{\infty} p^2(k+1)(1-p)^k z^{k+2} = \sum_{k=2}^{\infty} p^2(k-1)(1-p)^{k-2}z^k.$$

Ainsi on a pour $n \geq 2$, $\quad \mathbb{P}(S=n) = p^2(n-1)(1-p)^{n-2}$.

2. Si $k \notin \{1, \cdots, n-1\}$ on a $\mathbb{P}(X=k|S=n) = 0$ et pour $k \in \{1, \cdots, n-1\}$.

$$\begin{aligned}
\mathbb{P}(X=k|S=n) &= \frac{\mathbb{P}(X=k, S=n)}{\mathbb{P}(S=n)} \\
&= \frac{\mathbb{P}(X=k, Y=n-k)}{\mathbb{P}(S=n)} \\
&= \frac{p(1-p)^{k-1}p(1-p)^{n-k-1}}{p^2(n-1)(1-p)^{n-2}} \\
&= \frac{1}{n-1}.
\end{aligned}$$

Conditionnellement à S, X suit la loi uniforme sur $\{1, .., S-1\}$. On note par $h(n) = \mathbb{E}[X|S=n]$. On a $h(n) = \sum_{k=1}^{n-1} k\mathbb{P}(X=k|S=n) = \sum_{k=1}^{n-1} \frac{k}{n-1} = \frac{n}{2}$. On en déduit que $\mathbb{E}[X|S] = \frac{S}{2}$.

3. On a bien $\mathbb{E}\left[\mathbb{E}[X|S]\right] = \frac{\mathbb{E}[S]}{2} = \mathbb{E}[X]$.

▲

Exercice II.14.

1. Soient $m, n \geq 1$. L'évènement $\{T_1 = m, T_2 - T_1 = n\}$ est égal à $\{X_1 = 0, \ldots, X_{m-1} = 0, X_m = 1, X_{m+1} = 0, \ldots, X_{m+n-1} = 0, X_{m+n} = 1\}$. Par *indépendance* des v.a. X_i, on a donc

$$\begin{aligned}
\mathbb{P}(T_1 = m, T_2 - T_1 = n) &= \mathbb{P}(X_1=0)\cdots\mathbb{P}(X_{m-1}=0)\mathbb{P}(X_m=1)\mathbb{P}(X_{m+1}=0)\cdots \\
&\qquad \cdots\mathbb{P}(X_{m+n-1}=0)\mathbb{P}(X_{m+n}=1) \\
&= p^2(1-p)^{m+n-2}.
\end{aligned}$$

Par la formule des lois marginales,

$$\begin{aligned}\mathbb{P}(T_1 = m) &= \sum_{n\geq 1} \mathbb{P}(T_1 = m, T_2 - T_1 = n) = \sum_{n\geq 1} p^2(1-p)^{m+n-2} \\ &= p^2(1-p)^{m-1} \sum_{n\geq 1} (1-p)^{n-1} = p(1-p)^{m-1}.\end{aligned}$$

De même, $\mathbb{P}(T_2 - T_1 = n) = p(1-p)^{n-1}$. Par conséquent, pour tous $m, n \geq 1$, $\mathbb{P}(T_1 = m, T_2 - T_1 = n) = \mathbb{P}(T_1 = m)\mathbb{P}(T_2 - T_1 = n)$, ce qui prouve que les v.a. T_1 et $T_2 - T_1$ sont indépendantes. Noter qu'elles suivent toutes les deux la loi géométrique de paramètre p.

2. Plus généralement, si $n_1, n_2, \ldots, n_{k+1} \geq 1$, notons $I = \{n_1, n_1 + n_2, \ldots, n_1 + \cdots + n_{k+1}\}$ et $J = \{1, \ldots, n_1 + \cdots + n_{k+1}\} \backslash I$. L'évènement $\{T_1 - T_0 = n_1, T_2 - T_1 = n_2, \ldots, T_{k+1} - T_k = n_{k+1}\}$ est alors égal à

$$\bigcap_{i\in I} \{X_i = 1\} \cap \bigcap_{i\in J} \{X_i = 0\}.$$

Par *indépendance* des v.a. X_i, on a donc

$$\begin{aligned}\mathbb{P}(T_1 - T_0 = n_1, T_2 - T_1 = n_2, \ldots, T_{k+1} - T_k = n_{k+1}) &= \prod_{i\in I} \mathbb{P}(X_i = 1) \prod_{i\in J} \mathbb{P}(X_i = 0) \\ &= p^{k+1}(1-p)^{n_1+\cdots+n_{k+1}-k-1}.\end{aligned}$$

Comme ci-dessus, la formule des lois marginales implique que pour tous $j \in \{0, \ldots, k\}$ et $n_{j+1} \geq 1$,

$$\mathbb{P}(T_{j+1} - T_j = n_{j+1}) = p(1-p)^{n_{j+1}-1}. \tag{II.1}$$

Par conséquent,

$$\mathbb{P}(T_1 - T_0 = n_1, T_2 - T_1 = n_2, \ldots, T_{k+1} - T_k = n_{k+1}) = \prod_{j=0}^{k} \mathbb{P}(T_{j+1} - T_j = n_{j+1}),$$

ce qui prouve que les v.a. $T_1 - T_0, T_2 - T_1, \ldots, T_{k+1} - T_k$ sont indépendantes. De plus, d'après (II.1), elles suivent toutes la loi géométrique de paramètre p.

3. Noter que $T_k = \sum_{j=0}^{k-1} (T_{j+1} - T_j)$. Par linéarité de l'espérance,

$$\mathbb{E}[T_k] = \mathbb{E}\left[\sum_{j=0}^{k-1} (T_{j+1} - T_j)\right] = \sum_{j=0}^{k-1} \mathbb{E}[T_{j+1} - T_j] = \sum_{j=0}^{k-1} \frac{1}{p} = \frac{k}{p}.$$

Par *indépendance* des v.a. $T_{j+1} - T_j$,

$$\mathrm{Var}(T_k) = \mathrm{Var}\left(\sum_{j=0}^{k-1} (T_{j+1} - T_j)\right) = \sum_{j=0}^{k-1} \mathrm{Var}(T_{j+1} - T_j) = \sum_{j=0}^{k-1} \frac{1-p}{p^2} = \frac{k(1-p)}{p^2}.$$

4. Soit $n \geq k \geq 1$. On suppose $n \geq 2$. L'évènement $\{T_k = n\}$ est égal à $\{\sum_{i=1}^{n-1} X_i = k-1\} \cap \{X_n = 1\}$. De plus la loi de $\sum_{i=1}^{n-1} X_i$ est la loi binomiale de paramètre $(n-1, p)$. Par *indépendance* des v.a. X_i, on a donc

$$\mathbb{P}(T_k = n) = \mathbb{P}(\sum_{i=1}^{n-1} X_i = k-1)\mathbb{P}(X_n = 1) == C_{n-1}^{k-1} p^k (1-p)^{n-k}.$$

Pour $n = k = 1$, on obtient $\mathbb{P}(T_1 = 1) = p$.

Par *indépendance* des v.a. $T_{j+1} - T_j$, la fonction génératrice ϕ_{T_k} de T_k est le produit des fonctions génératrices des v.a. $T_{j+1} - T_j$ ($j \in \{0, \ldots, k-1\}$). Comme ces dernières suivent la loi géométrique de paramètre p, on a donc

$$\phi_{T_k}(z) = \prod_{j=0}^{k-1} \frac{pz}{1-(1-p)z} = \left(\frac{pz}{1-(1-p)z}\right)^k.$$

5. On décompose sur les évènements $\{\tau = n\}$:

$$\phi_{T_\tau}(z) = \mathbb{E}[z^{T_\tau}] = \sum_{n\geq 1} \mathbb{E}[z^{T_\tau} | \tau = n]\mathbb{P}(\tau = n) = \sum_{n\geq 1} \phi_{T_n}(z)\mathbb{P}(\tau = n).$$

Puisque $\mathbb{P}(\tau = n) = \rho(1-\rho)^{n-1}$, on a donc d'après la question précédente

$$\phi_{T_\tau}(z) = \sum_{n\geq 1} \rho(1-\rho)^{n-1} \left(\frac{pz}{1-(1-p)z}\right)^n = \phi_\tau\left(\frac{pz}{1-(1-p)z}\right) = \frac{\rho p z}{1-(1-\rho p)z},$$

ce qui prouve que T_τ suit la loi géométrique de paramètre ρp.

6. Considérons l'expérience suivante. On jette la première pièce et chaque fois que l'on obtient 1 (pile) on jette la seconde pièce. On note T' le premier instant où la seconde pièce montre pile. D'une part T' et T_τ ont même loi puisque, les jets étant tous indépendants, l'ordre des jets n'importe pas dans le calcul des lois de ces deux v.a.. D'autre part, T' est l'instant de premier succès dans une suite d'expériences indépendantes où la probabilité de succès vaut $p\rho$. En effet, les deux jets étant indépendants, la probabilité que la première pièce montre pile puis que la seconde montre aussi pile est le produit $p\rho$. T_τ et T' suivent donc la loi géométrique de paramètre $p\rho$.

▲

Exercice II.15.

1. On a $T \leq 2T'$ avec $T' = \inf\{n \in \mathbb{N}^*; X_{2n-1} = 1, X_{2n} = 0\}$. La variable aléatoire T' est le premier instant où $Y_n = (X_{2n-1}, X_{2n})$ est égal à $(1,0)$. Les variables aléatoires discrètes $(Y_n, n \in \mathbb{N}^*)$ sont indépendantes et de même loi. T' est un premier instant de succès. La loi de T' est donc la loi géométrique de paramètre $\mathbb{P}(Y_n = (1,0)) = p(1-p)$. En particulier T' est finie p.s. On en déduit que T est fini p.s.
2. On calcule $\mathbb{P}(T = k)$ en décomposant suivant les valeurs de $T_1 = \inf\{n \in \mathbb{N}^*; X_n = 1\}$:

$$\mathbb{P}(T = k) = \sum_{i=1}^{k-1} \mathbb{P}(T = k, T_1 = i) = \sum_{i=1}^{k-1} (1-p)^{i-1} p^{k-i}(1-p).$$

Soit U et V deux variables aléatoires indépendantes de loi géométrique de paramètres respectifs p et $1-p$. La loi de $U+V$ est :

$$\mathbb{P}(U+V=k)=\sum_{i=1}^{k-1}\mathbb{P}(U=i,V=k-i)=\sum_{i=1}^{k-1}(1-p)^{i-1}p^{k-i}(1-p).$$

3. On a $\phi_U(z)=\dfrac{zp}{1-(1-p)z}$, $\phi_V(z)=\dfrac{z(1-p)}{1-pz}$. On en déduit $\phi_T(z)=\phi_{U+V}(z)=\phi_U(z)\phi_V(z)$.
4. On a $\mathbb{E}[T]=\mathbb{E}[U]+\mathbb{E}[V]$ ou bien $\mathbb{E}[T]=\phi_T'(1)=\dfrac{1}{p(1-p)}$. On a également par indépendance $\mathrm{Var}(T)=\mathrm{Var}(U)+\mathrm{Var}(V)$ ou bien $\mathrm{Var}(T)=\phi_T''(1)+\phi_T'(1)-\phi_T'(1)^2=\dfrac{1-3p(1-p)}{p^2(1-p)^2}$.

▲

Exercice II.16.

Si les joueurs de haut niveau sont uniformément répartis dans la population, les médailles le sont aussi. Chaque individu a donc, environ, une probabilité $p_m=928/6\ 10^9$ d'avoir une médaille et $p_o=301/6\ 10^9$ d'avoir une médaille d'or. Le nombre de médailles française suit donc une loi binomiale de paramètre (n,p_m), avec $n=60\ 10^6$. On peut approcher cette loi par la loi de Poisson de paramètre $\theta_m=np_m\simeq 9$. De même, on peut approcher la loi du nombre de médaille d'or par une loi de Poisson de paramètre $\theta_o=np_o\simeq 3$. La probabilité d'avoir plus de 20 médailles est de 4 pour 10 000, et la probabilité d'avoir plus de 10 médailles d'or est de 3 pour 10 000. Ceci est invraisemblable. L'hypothèse selon laquelle les sportifs de haut niveau sont uniformément répartis dans la population est donc invraisemblable.

▲

Exercice II.17.

1. Le tableau ci-dessous donne la probabilité de chacun des croisements et la probabilité des génotypes d'un enfant pour chaque croisement :

		Proba. géno. enfant		
Croisement	Proba. du croisement	AA	AB	BB
$AA\times AA$	x^2	1	0	0
$AA\times AB$	$4xy$	$\frac{1}{2}$	$\frac{1}{2}$	0
$AA\times BB$	$2xz$	0	1	0
$AB\times AB$	$4y^2$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$AB\times BB$	$4yz$	0	$\frac{1}{2}$	$\frac{1}{2}$
$BB\times BB$	z^2	0	0	1

D'où, en notant x' la probabilité d'avoir le génotype AA, $2y'$ celle d'avoir le génotype AB, z' celle d'avoir le génotype BB pour l'enfant, on obtient que :

$$x'=1\times x^2+\frac{1}{2}\times 4xy+\frac{1}{4}\times 4y^2=(x+y)^2\ .$$

De même, on obtient que $y'=(x+y)(z+y)$ et $z'=(y+z)^2$.

2. Supposons que, pour les parents, la probabilité d'avoir respectivement les génotypes (AA, AB, BB) est $(1/4, 1/2, 1/4)$. Alors, les résultats de la question permettent d'en déduire immédiatement qu'un enfant a les mêmes probabilités d'avoir les génotypes (AA, AB, BB) que ses parents.
3. Pas la peine de refaire les calculs ! Il suffit d'utiliser le tableau ci-dessus :

$$\begin{aligned} x'' &= (x'+y')^2 \\ &= ((x+y)^2 + (x+y)(z+y))^2 \\ &= ((x+y)(x+y+z+y))^2 \\ &= (x+y)^2 \\ &= x' . \end{aligned}$$

De la même manière, on obtient que $y'' = y'$ et $z'' = z'$. Donc, dès la seconde génération, les probabilités d'avoir les génotypes (AA, AB, BB) restent inchangées.

▲

Exercice II.18.

Le coût de la première méthode est $c_1 = cN$. Le coût du test sur un groupe de n personnes en mélangeant les prélèvement est C_n, et on a $\mathbb{P}(C_n = c) = (1-p)^n$ et $\mathbb{P}(C_n = c + nc) = 1-(1-p)^n$. En supposant que N soit divisible par n, le coût moyen total du test, en utilisant la deuxième méthode est $c_2 = \frac{N}{n}\mathbb{E}[C_n]$ soit :

$$c_2 = \frac{N}{n}\left[c(1-p)^n + (c+nc)(1-(1-p)^n)\right] = cN\left[1-(1-p)^n + \frac{1}{n}\right].$$

En supposant que $np \ll 1$, il vient en faisant un développement limité :

$$c_2 = cN\left[np + \frac{1}{n}\right] + o(np).$$

Cette quantité est minimale pour $n \simeq 1/\sqrt{p}$. La condition $np \ll 1$ est alors équivalente à $p \ll 1$. On obtient donc $c_2 \simeq 2cN\sqrt{p}$. On choisit donc la deuxième méthode. On peut vérifier que pour $p \ll 1$, en prenant $n = \left[1/\sqrt{p}\right]$, on a $c_2 \leq 2c + 2cN\sqrt{p}$.

Exemple numérique : si $p = 1\%$, alors il est optimal de réaliser des tests sur des groupes de $n = 10$ personnes. L'économie réalisée est alors de :

$$\frac{c_1 - c_2}{c_1} = 1 - 2\sqrt{p} = 80\% .$$

▲

Exercice II.19.

1. Nous avons, pour $i = 0, ..., N$,

$$\begin{aligned} p_n(i, i+1) &= (N-i)/N, \\ p_n(i, i-1) &= i/N, \\ p_n(i, j) &= 0 \quad \text{si } |i-j| \neq 1. \end{aligned}$$

2. Si $\mathbb{P}(X_0 = i) = C_N^i/2^N$, on a

$$\begin{aligned}\mathbb{P}(X_1 = j) &= \sum_{i=0}^{N} \mathbb{P}(X_1 = j, X_0 = i) = \sum_{i=0}^{N} \mathbb{P}(X_0 = i)\mathbb{P}(X_1 = j|X_0 = i) \\ &= \mathbb{P}(X_0 = j-1)\mathbb{P}(X_1 = j|X_0 = j-1) + \mathbb{P}(X_0 = j+1)\mathbb{P}(X_1 = j|X_0 = j+1) \\ &= \frac{C_N^{j-1}}{2^N}\frac{N-j+1}{N} + \frac{C_N^{j+1}}{2^N}\frac{j+1}{N} \\ &= C_N^j/2^N,\end{aligned}$$

avec la convention que $C_N^{-1} = 0$ et $C_N^{N+1} = 0$. Les variables aléatoires X_0 et X_1 ont donc la même loi binomiale $\mathcal{B}(N, 1/2)$. Par récurrence, on vérifie alors que la loi de X_n est la même loi binomiale. On dit que la loi $\mathcal{B}(N, 1/2)$ est une loi stationnaire du système. En fait, on peut vérifier que c'est la seule.

▲

Exercice II.20.

1. On note $\mathbf{H}_n = \{\text{le phénomène H se produit à l'instant } n\}$.

$$\begin{aligned}v_n = \mathbb{P}(\mathbf{H}_n) &= \sum_{k=0}^{n} \mathbb{P}(\mathbf{H}_n|X_1 = k)\mathbb{P}(X_1 = k) \\ &= \sum_{k=0}^{n} \mathbb{P}(X_1 + X_2 + \cdots + X_i = n, \text{ pour un } i \geq 2|X_1 = k)\mathbb{P}(X_1 = k) \\ &= \sum_{k=0}^{n} \mathbb{P}(X_2 + \cdots + X_i = n-k, \text{ pour un } i \geq 2)\mathbb{P}(X_1 = k).\end{aligned}$$

La seconde égalité est obtenue grâce à l'indépendance de X_1 et de la suite $(X_k, k \geq 2)$. Comme les variables aléatoires $(X_k, k \geq 2)$ sont à valeurs dans $\mathbb{N}^*$, on en déduit que

$$\begin{aligned}&\mathbb{P}(X_2 + \cdots + X_i = n-k \text{ pour un } i \geq 2) \\ &\qquad = \mathbb{P}(X_2 + \cdots + X_i = n-k \text{ pour un } i \in \{2, \ldots, n-k+1\}) = u_{n-k}.\end{aligned}$$

2. On a démontré que $(v_n, n \geq 0)$ est la convolution des suites $(b_n, n \geq 0)$ avec $(u_n, n \geq 0)$. On a donc $V(s) = B(s)U(s)$.
3. La preuve de la première égalité se traite comme dans la question 1. Pour démontrer la relation entre $U(s)$ et $F(s)$ on multiplie par s^n et on somme sur n. On obtient

$$\sum_{n=1}^{\infty} s^n u_n = \sum_{n=1}^{\infty}\sum_{k=0}^{n-1} s^n u_k f_{n-k}.$$

Il vient $U(s) - 1 = F(s)U(s)$, c'est–à–dire, $U(s) = 1/(1 - F(s))$.

4. Comme $b_k = \mathbb{P}(X_1 = k) = (1-p)^k p,\ k \geq 0$, on a

$$B(s) = \sum_{k=0}^{\infty} s^k b_k = p\sum_{k=0}^{\infty} s^k (1-p)^k = \frac{p}{1 - s(1-p)}.$$

En plus, la fonction génératrice d'une loi géométrique de paramètre p est donnée par $F(s) = ps/(1 - s(1 - p))$, on vérifie facilement la relation

$$B(s) = p\ \frac{1 - F(s)}{1 - s}, \qquad |s| < 1.$$

En utilisant les questions précédentes, on obtient

$$\sum_{n=0}^{\infty} s^n v_n = V(s) = B(s)U(s) = \frac{B(s)}{1 - F(s)} = \frac{p}{1 - s} = \sum_{n=0}^{\infty} p\ s^n.$$

En identifiant le coefficient de s^n on trouve $v_n = p$.

▲

Chapitre III

Variables aléatoires continues

III.1 Énoncés

Exercice III.1.
Soit X une v.a. de densité f définie par $f(x) = 0$ si $x < 0$ et $f(x) = x\mathrm{e}^{-x^2/2}$ sinon.

1. Vérifier que f est une densité de probabilité.
2. Montrer que $Y = X^2$ est une v.a. à densité, dont on précisera la loi.
3. Calculer l'espérance et la variance de Y

△

Exercice III.2.
Soit Y une v.a. de loi exponentielle $\lambda > 0$ et ε une v.a. discrète indépendante de Y et telle que $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$. Quelle est la loi de $Z = \varepsilon Y$? Cette loi est appelée loi exponentielle symétrique.

△

Exercice III.3.
Soit X et Y deux variables aléatoires indépendantes de loi respective $\Gamma(\lambda, a)$ et $\Gamma(\lambda, b)$.

1. Calculer la loi du couple $(X + Y, \frac{X}{X+Y})$.
2. Montrer que $X + Y$ et $\frac{X}{X+Y}$ sont indépendantes et identifier leur loi.

△

Exercice III.4.
Soit X une variable de Cauchy.

1. Déterminer la loi de $1/X$.
2. Montrer que si Y et Z sont deux variables gaussiennes centrées réduites indépendantes, alors Y/Z suit une loi de Cauchy.
3. Retrouver ainsi le résultat de la question 1.

△

Exercice III.5.
Soit X_1, X_2 des v.a. indépendantes de loi $\mathcal{N}(0,1)$.

1. Montrer que X_1^2 suit une loi $\chi^2(1)$.
2. Montrer que $X_1^2 + X_2^2$ suit une loi $\chi^2(2)$.

△

Exercice III.6.
La durée de vie, exprimée en années, d'un circuit électronique est une variable aléatoire T dont la fonction de répartition F est définie par :

$$F(t) = \begin{cases} 0 & \text{si } t < 0 \\ 1 - \exp\left(-\frac{1}{2}t^2\right) & \text{si } t \geq 0 \end{cases}$$

1. Donner la densité de probabilité f de T. Calculer $\mathbb{E}[T]$.
2. Sachant que le circuit a déjà fonctionné durant 1 an, quelle est la probabilité qu'il continue à fonctionner encore durant au moins 2 ans ? La loi est-elle sans mémoire ?
3. Un équipement électronique E est composé de 10 circuits identiques et indépendants. Au circuit i $(1 \leq i \leq 10)$ est associée la variable aléatoire :

$$X_i = \begin{cases} 1 & \text{si la durée de vie du circuit } i \text{ est inférieure à un an} \\ 0 & \text{sinon} \end{cases}$$

   (a) Quelle est la loi de probabilité de la variable aléatoire N égale au nombre de circuits de E dont la durée de vie est inférieure à 1 an ?

   (b) L'équipement E est dit en série si la défaillance de l'un de ses circuits entraîne sa défaillance. Quelle est alors la probabilité qu'il soit défaillant avant 1 an ?

   (c) L'équipement E est dit en parallèle si sa défaillance ne peut se produire que si tous ses circuits sont défaillants. Quelle est alors la probabilité qu'il soit défaillant : avant 1 an ? avant t ans ?

△

Exercice III.7.
Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de loi exponentielle de paramètre rescpectif $\lambda_n > 0$. Montrer que les trois conditions suivantes sont équivalentes.

1. $\sum_{n\geq 1} \lambda_n^{-1} < \infty$.
2. $\mathbb{E}[\sum_{n\geq 1} X_n] < \infty$.
3. $\mathbb{P}(\sum_{n\geq 1} X_n < \infty) > 0$.

Pour $3 \Rightarrow 1$, on pourra considérer $\mathbb{E}[\mathrm{e}^{-\sum_{n\geq 1} X_n}]$.

△

Exercice III.8.
Soit $n \geq 2$. Soit $X_1, \ldots, X_n$ une suite de n variables aléatoires indépendantes de loi uniforme sur $[0,1]$. On note $Y = \min_{1\leq i\leq n} X_i$ et $Z = \max_{1\leq i\leq n} X_i$.

1. Calculer la loi de (Y, Z).
2. En déduire la loi de Y et la loi de Z. Reconnaître ces deux lois.
3. Calculer $\mathbb{E}[Y|Z]$.
4. Calculer $\mathbb{E}[g(Y/Z)|Z]$, pour une fonction g mesurable bornée. En déduire la loi de Y/Z conditionnellement à Z. Reconnaître la loi de Y/Z.
5. Montrer que $\mathcal{L}((1-Z, 1-Y)) = \mathcal{L}((Y, Z))$.
6. En déduire que $(1-Z)/(1-Y)$ est indépendant de Y.

△

Exercice III.9.
Un convoi ferroviaire traverse une région de forts troubles politiques. Passant à proximité d'une zone de combat, ce convoi est atteint par un nombre aléatoire de balles qui perforent les flancs latéraux des containers parallélépipédiques des wagons. Le liquide contenu dans chaque container se met alors à fuir jusqu'au niveau du trou le plus bas dans le flanc du container.

On suppose pour l'exercice que le convoi comporte n containers identiques de hauteur h, que le nombre d'impacts par container suit une loi de Poisson de paramètre θ et que ces impacts sont uniformément répartis sur le flanc latéral des containers.

1. On s'intéresse à un seul container. On note Z la variable aléatoire correspondant à la hauteur du point d'impact le plus bas sur le flanc du container et N la variable aléatoire donnant le nombre d'impacts reçus par le container. Donner la loi de Z conditionnellement à N.
2. Quel pourcentage de liquide peut-on espérer sauver connaissant le nombre d'impacts reçus par le container ?
3. On s'intéresse au convoi dans son entier. On suppose qu'il comporte un grand nombre de wagons. Quel pourcentage du chargement peut-on espérer sauver ?

   Application numérique : $\theta = 5$.

△

Exercice III.10.
On considère un gâteau circulaire avec une cerise sur le bord. On découpe le gâteau en deux parts en coupant suivant deux rayons choisis au hasard. Avec quelle probabilité la part contenant la cerise est-elle plus petite que la part ne contenant pas la cerise ? Quelle est la longueur angulaire moyenne de la part contenant la cerise ? Conclusion.

△

Exercice III.11.
On considère un bâton sur lequel on trace au hasard deux marques. On découpe le bâton suivant les deux marques. Quelle est la probabilité pour que l'on puisse faire un triangle avec les trois morceaux ainsi obtenus ? △

Exercice III.12.
Votre ami choisit deux nombres positifs sans vous faire part de la manière dont il les choisit. Après avoir lancé une pièce équilibrée, il vous donne le plus petit s'il a obtenu face et le plus grand sinon. Votre objectif est de déterminer s'il vous a donné le plus petit ou le plus grand, et de maximiser votre probabilité d'avoir raison.

1. Vous lancez une pièce équilibrée ou non. Si vous obtenez face, vous pariez qu'il vous a donné le plus petit, sinon vous pariez qu'il vous a donné le plus grand. Quelle est la probabilité de gagner votre pari ?
2. Vous simulez une variable aléatoire positive continue Z ayant pour support $\mathbb{R}^+$. Si le nombre donné par votre ami est plus petit que Z, alors vous pariez qu'il vous a donné le plus petit, sinon vous pariez qu'il vous a donné le plus grand. Quelle est la probabilité de gagner votre pari ?
3. On suppose que les deux nombres de votre ami, ont été obtenu par simulation suivant une loi (continue de densité strictement positive sur $]0, \infty[$) donnée et connue de vous. Déterminer votre stratégie optimale (i.e. la loi de Z que l'on ne suppose plus continue). Quelle est alors la probabilité de gagner votre pari ?

△

Exercice III.13.
On désire déterminer la distribution des vitesses des molécules d'un gaz monoatomique parfait à l'équilibre (loi de Maxwell (1859)).

1. Soit (X, Y, Z), un vecteur aléatoire continu à support dans $\mathbb{R}^3$ dont la loi est invariante par rotation autour de l'origine et dont les composantes X, Y, Z sont indépendantes. Caractériser[1] les lois marginales de X, Y et Z dans le cas où la densité de probabilité du couple (X, Y, Z) est une fonction $C^1(\mathbb{R}^3, \mathbb{R})$.
2. On représente la vitesse d'une molécule d'un gaz monoatomique parfait à l'équilibre dans un repère orthonormal par un vecteur aléatoire $V = (V_1, V_2, V_3)$. Le choix du repère étant arbitraire, il est naturel de supposer que la loi de V est invariante par rotation. Il est de plus naturel de supposer que les coordonnées de V sont indépendantes. Si on suppose de plus que la loi de V possède une densité dérivable, on en déduit que le vecteur V vérifie les propriétés de la question 1). Déterminer la densité de probabilité de la vitesse d'une molécule, sachant que l'énergie cinétique moyenne d'un atome du gaz de masse m est $\frac{3}{2}kT$ où k est la constante de Boltzmann et T la température du gaz. (Pour des molécules à plusieurs atomes, l'énergie cinétique moyenne tient compte d'effets complexes comme la rotation, les oscillations... La loi de Maxwell n'est plus vérifiée dans ces cas.)
3. Montrer que si X et Y sont deux v.a. indépendantes de loi respective $\Gamma(\lambda, a)$ et $\Gamma(\lambda, b)$, alors la loi de $X + Y$ est une loi gamma dont on précisera les paramètres.
4. Calculer la loi de V_1^2. En déduire la loi de $|V|^2$ et la loi de $|V|$ dite loi de Maxwell.

△

[1] En fait, on peut, en utilisant les fonctions caractéristiques, caractériser toutes les lois de vecteur qui sont invariantes par rotation autour de l'origine. Hormis la variable nulle, on ne trouve pas d'autres lois que celles obtenues sous les hypothèses de cet exercice.

Exercice III.14.
Le paradoxe de Bertrand[2] est un exemple classique (cf le livre *Calcul des probabilités* de Poincaré en 1912), qui met en évidence la difficulté d'étendre la formule classique des probabilités uniformes :

$$\text{Probabilité d'un évènement} = \frac{\text{nombre de résultats favorables}}{\text{nombre de résultats possibles}}$$

aux espaces d'états non dénombrables.

On choisit au hasard une corde sur un cercle de rayon r et de centre O. On désire calculer la probabilité p pour qu'elle soit plus longue que le côté du triangle équilatéral inscrit (de longueur $\sqrt{3}r$). Ceci est équivalent à calculer la probabilité pour que la distance de la corde, choisie au hasard, au centre du cercle soit inférieure à $r/2$.

1. On considère que la longueur de la corde est déterminée par la donnée de la distance H de son milieu au centre du cercle. On suppose que H suit une loi uniforme sur $[0, r]$. Quelle est la loi de la longueur de la corde ? Quelle est son espérance et sa variance ? Quelle est la probabilité qu'elle soit plus grande que $\sqrt{3}r$?
2. On choisit A et B au hasard sur le cercle et de manière indépendante. La longueur de la corde AB est repérée par l'angle au centre $\alpha = \widehat{AOB}$. Montrer que α suit une loi uniforme sur $[0, 2\pi]$. Quelle est la loi de la longueur de la corde ? Quelle est son espérance et sa variance ? Quelle est la probabilité qu'elle soit plus grande que $\sqrt{3}r$?
3. On choisit le point I milieu de la corde aléatoirement sur le disque. Le point I est donc définit par deux coordonnées (x, y) qui suivent une loi uniforme sur le disque de densité $\frac{1}{\pi r^2}\mathbf{1}_{\{x^2+y^2\leq r^2\}}$. Quelle est la loi de la longueur de la corde ? Quelle est son espérance et sa variance ? Quelle est la probabilité qu'elle soit plus grande que $\sqrt{3}r$?

Quelle est votre conclusion ?

△

Exercice III.15.
Les véhicules spatiaux désirant s'arrimer à la Station Spatiale Internationale (ISS) s'appuient sur le système de guidage GPS pour la phase d'approche de la station. Cependant, à faible distance de l'ISS, les signaux émits par la constellation de satellites qui constituent le système GPS sont fortement perturbés par les phénomènes de réflexions multiples sur la structure métallique de la station. L'onde électromagnétique reçue par le récepteur GPS du véhicule spatial se présente donc comme la superposition de deux ondes en quadrature dont les amplitudes X et Y sont des variables aléatoires de loi $\mathcal{N}(0, \sigma^2)$ supposées indépendantes (pour des raisons d'isotropie). L'étude de l'amplitude $R = \sqrt{X^2 + Y^2}$ de l'onde reçue est de première importance pour assurer un guidage fiable du vaisseau lors de la manœuvre d'arrimage[3].

1. Quelle est la loi du couple (X, Y) ?
2. En faisant le changement de variable $X = R\cos\Theta, Y = R\sin\Theta$, donner la loi du couple (R, Θ). R et Θ sont-elles indépendantes ?

[2] Joseph Bertrand, mathématicien français (1822-1900).

[3] cf. "Effects of Multipath and Signal Blockage on GPS Navigation in the Vicinity of the International Space Station (ISS)", David E. Gaylor, R. Glenn Lightsey, Kevin W. Key, ION GPS/GNSS 2003, Portland, OR.

3. En déduire la loi de R. Cette loi est appelée loi de Rayleigh.
4. Calculer l'espérance et la variance de cette loi.

△

Exercice III.16.
L'aiguille de Buffon (1777). On lance des aiguilles de longueur l sur un parquet dont les lames sont parallèles, toutes identiques et de largeur $d > l$. Les lancers sont indépendants et s'effectuent tous dans les mêmes conditions. On paramètre la position d'une aiguille par rapport aux lames de parquet par l'abscisse de son milieu, X, et l'angle, θ, qu'elle fait par rapport à une droite orthogonale à la direction des lames.

1. Traduire le fait qu'une aiguille coupe une rainure du parquet à l'aide des variables X et θ.
2. Proposer un modèle pour la loi de (X, θ). Calculer la probabilité pour qu'une aiguille coupe une rainure du parquet.
3. On note N_n le nombre d'aiguilles coupant une rainure du parquet au bout de n lancers. Que vaut $\lim_{n\to\infty} \frac{N_n}{n}$?
4. On suppose que $2l = d$. Trouver n pour que la précision sur $1/\pi$ soit de 10^{-2} avec une probabilité d'au moins 95%.
5. On effectue 355 lancers avec $2l = d$, et on obtient 113 intersections. On a ainsi une approximation de $1/\pi$ à 3.10^{-7}. Ce résultat est-il en contradiction avec le résultat de la question précédente ? Que devient la précision de l'approximation avec un lancer de plus ?

△

III.2 Corrections

Exercice III.1.

1. La fonction f est positive, continue sur $\mathbb{R}$. De plus
$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_0^{+\infty} x\,\mathrm{e}^{-x^2/2}\,dx = 1$$
donc f est bien une densité de probabilité.
2. Soit $g : \mathbb{R} \to \mathbb{R}$ mesurable bornée.
$$\mathbb{E}[g(Y)] = \mathbb{E}[g(X^2)] = \int_0^{+\infty} g(x^2)\,x\,\mathrm{e}^{-x^2/2}\,dx = \int_0^{+\infty} g(y)\,\frac{1}{2}\mathrm{e}^{-y/2}\,dy,$$
où l'on a fait le changement de variable $y = x^2$ sur $\mathbb{R}^+$. Donc Y suit une loi exponentielle de paramètre $1/2$.
3. Pour une loi exponentielle de paramètre λ, l'espérance est $1/\lambda$ et la variance $1/\lambda^2$, donc $\mathbb{E}[Y] = 2$ et $\mathrm{Var}(Y) = 4$.

▲

Exercice III.2.

Soit g une fonction mesurable bornée. En utilisant la formule de décomposition, puis l'indépendance, on a

$$\begin{aligned}
\mathbb{E}[g(Z)] &= \mathbb{E}[g(Y)\mathbf{1}_{\{\varepsilon=1\}}] + \mathbb{E}[g(-Y)\mathbf{1}_{\{\varepsilon=-1\}}] \\
&= \mathbb{E}[g(Y)]\mathbb{P}(\varepsilon = 1) + \mathbb{E}[g(-Y)]\mathbb{P}(\varepsilon = -1) \\
&= \frac{1}{2}\int_0^{\infty} \lambda\,\mathrm{e}^{-\lambda y}\,g(y)dy + \frac{1}{2}\int_0^{\infty} \lambda\,\mathrm{e}^{-\lambda y}\,g(-y)dy \\
&= \frac{1}{2}\int_{\mathbb{R}} \lambda\,\mathrm{e}^{-\lambda\,|z|}\,g(z)\;dz
\end{aligned}$$

La densité de la loi de Z est donc $f_Z(z) = \dfrac{1}{2}\,\lambda\,\mathrm{e}^{-\lambda\,|z|}\,\mathbf{1}_{\mathbb{R}}(z)$.

▲

Exercice III.3.

Soit $S = X + Y$ et $T = \frac{X}{X+Y}$. Comme $X + Y > 0$ $\mathbb{P}$-p.s., T est une variable aléatoire réelle définie $\mathbb{P}$-p.s. Soit h une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$, mesurable et bornée. On a

$$\begin{aligned}
\mathbb{E}[h(S,T)] &= \mathbb{E}\left[h\left(X+Y, \frac{X}{X+Y}\right)\right] \\
&= \int_{\mathcal{D}} h\left(x+y, \frac{x}{x+y}\right) \frac{\lambda^{a+b}}{\Gamma(a)\Gamma(b)} e^{-\lambda(x+y)} x^{a-1} y^{b-1}\,dx\,dy\,,
\end{aligned}$$

où $\mathcal{D} = \{(x,y) \in \mathbb{R}^2 \,|\, x > 0\,;\; y > 0\}$. On considère la fonction φ définie sur $\mathcal{D}$:

$$\varphi\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} s \\ t \end{pmatrix} \quad \text{où} \quad s = x + y, \quad t = \frac{x}{x+y}.$$

La fonction φ est une bijection de $\mathcal{D}$ dans $\Delta = \big\{(s,t) \in \mathbb{R}^2 \,|\, s > 0\ ;\ 0 < t < 1\big\}$. De plus la fonction φ est de classe C^1 ainsi que son inverse :

$$\varphi^{-1}\begin{pmatrix} s \\ t \end{pmatrix} = \begin{pmatrix} st \\ s(1-t) \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}.$$

Le jacobien de ce changement de variable est :

$$\mathrm{Jac}[\varphi](x,y) = \begin{vmatrix} 1 & 1 \\ \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \end{vmatrix} = -\frac{1}{x+y}\,.$$

On en déduit que $dtds = |\mathrm{Jac}[\varphi](\mathrm{x},\mathrm{y})|\, dxdy$ i.e. $s\, dsdt = dxdy$. D'où, on obtient que :

$$\mathbb{E}\left[h(S,T)\right] = \int_\Delta h(s,t)\frac{\lambda^{a+b}}{\Gamma(a)\Gamma(b)} e^{-\lambda s} s^{a+b-1} t^{a-1}(1-t)^{b-1}\, ds\, dt.$$

Donc, la densité du couple (S,T) est :

$$\frac{\lambda^{a+b}}{\Gamma(a)\Gamma(b)} e^{-\lambda s} s^{a+b-1} t^{a-1}(1-t)^{b-1} \mathbf{1}_{]0;+\infty[}(s)\mathbf{1}_{]0;1[}(t)\,.$$

C'est le produit d'une fonction de s et d'une fonction de t. Donc, les variables aléatoires S et T sont indépendantes. La densité de S est :

$$\frac{\lambda^{a+b}}{\Gamma(a+b)} e^{-\lambda s} s^{a+b-1} \mathbf{1}_{]0;+\infty[}(s)\,.$$

On reconnaît la densité de la loi gamma de paramètres λ et $a+b$.
La densité de T est :

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1}(1-b)^{b-1} \mathbf{1}_{]0;1[}(t)\,.$$

On reconnaît la densité de la loi béta de paramètres a et b. ▲

Exercice III.4.

1. Commençons par remarquer que la v.a. X est a priori à valeurs dans $\mathbb{R}$ entier, ce qui entraîne un problème de définition de $1/X$ lorsque $X = 0$. Toutefois, la v.a. X étant continue, on a $\mathbb{P}(X = 0) = 0$. La v.a. $1/X$ est donc bien définie $\mathbb{P}$-p.s. Soit $g : \mathbb{R} \to \mathbb{R}$ mesurable bornée. On a

$$\begin{aligned}
\mathbb{E}[g(1/X)] &= \int_{-\infty}^{+\infty} g(1/x)\,\frac{1}{\pi}\frac{1}{1+x^2}\,dx \\
&= \int_{-\infty}^{0} g(1/x)\,\frac{1}{\pi}\frac{1}{1+x^2}\,dx + \int_{0}^{+\infty} g(1/x)\,\frac{1}{\pi}\frac{1}{1+x^2}\,dx \\
&= \int_{-\infty}^{0} g(y)\,\frac{1}{\pi}\frac{1}{1+(1/y)^2}\,\frac{dy}{y^2} + \int_{0}^{+\infty} g(y)\,\frac{1}{\pi}\frac{1}{1+(1/y)^2}\,\frac{dy}{y^2} \\
&= \int_{-\infty}^{+\infty} g(y)\,\frac{1}{\pi}\frac{1}{1+y^2}\,dy,
\end{aligned}$$

où l'on a fait le changement de variable $y = 1/x$ sur $]-\infty, 0[$ et sur $]0, +\infty[$. Donc $1/X$ suit une loi de Cauchy.

2. Comme précédemment, si $X = Y/Z$, X n'est pas définie lorsque $Z = 0$, ce qui est un événement de probabilité nulle. Donc cette fois encore, elle est définie $\mathbb{P}$-p.s. Soit $g : \mathbb{R} \to \mathbb{R}$ mesurable bornée. On a

$$\begin{aligned}\mathbb{E}[g(X)] = \mathbb{E}[g(Y/Z)] &= \int_{-\infty}^{+\infty} dz\, \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \int_{-\infty}^{+\infty} dy \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}\, g(y/z) \\ &= 2\int_{0}^{+\infty} dz\, \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \int_{-\infty}^{+\infty} dy \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}\, g(y/z) \\ &= \frac{1}{\pi}\int_{0}^{+\infty} dz\, \mathrm{e}^{-z^2/2} \int_{-\infty}^{+\infty} du\, z\mathrm{e}^{-u^2z^2/2}\, g(u) \\ &= \frac{1}{\pi}\int_{0}^{+\infty} dz\ z\mathrm{e}^{-(1+u^2)z^2/2} \int_{-\infty}^{+\infty} du\, g(u) \\ &= \int_{-\infty}^{+\infty} \frac{1}{\pi}\frac{1}{1+u^2}\ g(u)\, du,\end{aligned}$$

où l'on a fait le changement de variable $u(y) = y/z$ pour $y \in \mathbb{R}$. Donc $X = Y/Z$ suit une loi de Cauchy.

3. Par symétrie la loi de Z/Y est la loi de Cauchy. On retrouve bien que la loi de $1/X$ est la loi de Cauchy si X est de loi de Cauchy.

▲

Exercice III.5.

Soit $g : \mathbb{R} \to \mathbb{R}$ mesurable bornée.

1. On a

$$\begin{aligned}\mathbb{E}[g(X_1^2)] = \int_{-\infty}^{+\infty} g(x^2)\, \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, dx &= 2\int_{0}^{+\infty} g(x^2)\, \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x^2/2}\, dx \\ &= \int_{0}^{+\infty} g(y)\, \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{y}} \mathrm{e}^{-y/2}\, dy,\end{aligned}$$

où l'on a fait le changement de variable $y = x^2$ sur $[0, +\infty[$. Donc X_1^2 suit une loi $\chi^2(1)$.

2. On a

$$\begin{aligned}\mathbb{E}[g(X_1^2 + X_2^2)] &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x_1^2/2}\, dx_1 \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-x_2^2/2}\, dx_2\ g(x_1^2 + x_2^2) \\ &= \frac{1}{2\pi}\int_{0}^{2\pi} d\theta \int_{0}^{+\infty} r\, dr g(r^2)\mathrm{e}^{-r^2/2} \\ &= \int_{0}^{+\infty} g(y)\, \frac{1}{2}\mathrm{e}^{-y/2}\, dy,\end{aligned}$$

où l'on est passé en coordonnées polaires, puis on a fait le changement de variable $y = r^2$. On en déduit que $X_1^2 + X_2^2$ suit une loi $\chi^2(2)$.

▲

Exercice III.6.

1. La densité de probabilité f de la fonction de répartition est $f(t) = t\exp\left(-\frac{1}{2}t^2\right)\mathbf{1}_{\{t>0\}}$.
   L'espérance vaut :
$$\begin{aligned}\mathbb{E}[T] &= \int_0^{+\infty} t^2\exp\left(-\frac{1}{2}t^2\right)dt \\ &= \left[-t\exp\left(-\frac{1}{2}t^2\right)\right]_0^{+\infty} + \int_0^{+\infty}\exp\left(-\frac{1}{2}t^2\right)dt \\ &= \frac{1}{2}\int_{-\infty}^{+\infty}\exp\left(-\frac{1}{2}t^2\right)dt \\ &= \frac{\sqrt{2\pi}}{2}.\end{aligned}$$

2. La probabilité s'écrit :
$$\mathbb{P}(T\geq 3\,|T\geq 1) = \frac{\mathbb{P}[(T\geq 3)\cap(T\geq 1)]}{\mathbb{P}(T\geq 1)} = \frac{\mathbb{P}(T\geq 3)}{\mathbb{P}(T\geq 1)} = \frac{\mathrm{e}^{-\frac{9}{2}}}{\mathrm{e}^{-\frac{1}{2}}} = \mathrm{e}^{-4}$$
$$\neq \mathbb{P}(T\geq 2) = \mathrm{e}^{-2}.$$
   On n'a pas $\mathbb{P}(T\geq 3\,|T\geq 1) = \mathbb{P}(T\geq 2)$ donc la loi n'est pas sans mémoire.
   (a) Les v.a. X_i sont indépendantes et suivent une loi de Bernoulli de paramètre :
$$\mathbb{P}(T\leq 1) = F(1) = 1-\mathrm{e}^{-\frac{1}{2}}$$
   On en déduit que la loi de N est la loi binomiale de paramètre $(10, 1-\mathrm{e}^{-\frac{1}{2}})$.
   (b) La probabilité que l'équipement en série soit défaillant avant 1 an vaut :
$$\begin{aligned}\mathbb{P}\left(\bigcup_{i=1}^{10}(X_i=1)\right) &= 1-\mathbb{P}\left(\bigcap_{i=1}^{10}(X_i=0)\right) \\ &= 1-\prod_{i=1}^{10}\mathbb{P}(X_i=0) \text{ car les v.a. sont indépendantes} \\ &= 1-\mathrm{e}^{-\frac{10}{2}} \\ &\simeq 9,9.10^{-1}.\end{aligned}$$
   (c) La probabilité que l'équipement en parallèle soit défaillant avant 1 an vaut :
$$\begin{aligned}\mathbb{P}\left(\bigcap_{i=1}^{10}(X_i=1)\right) &= \prod_{i=1}^{10}\mathbb{P}(X_i=1) \text{ car les v.a. sont indépendantes} \\ &= \left(1-\mathrm{e}^{-\frac{1}{2}}\right)^{10} \\ &\simeq 8,9.10^{-5}.\end{aligned}$$
   Soit T_i la durée de vie de l'équipement i. La probabilité que l'équipement en parallèle soit défaillant avant t ans vaut :
$$\begin{aligned}p_t &= \mathbb{P}\left(\bigcap_{i=1}^{10}(T_i\leq t)\right) \\ &= \prod_{i=1}^{10}\mathbb{P}(T_i\leq t) \\ &= \left(1-\mathrm{e}^{-\frac{1}{2}t^2}\right)^{10}.\end{aligned}$$

On obtient :$p_2 \simeq 2,3.10^{-1}$, $p_3 \simeq 8,9.10^{-1}$ et $p_4 \simeq 9,97.10^{-2}$.

▲

Exercice III.7.

On a par linéarité $\mathbb{E}[\sum_{n\geq 1} X_n] = \sum_{n\geq 1} \lambda_n^{-1}$. Donc on a $1 \Leftrightarrow 2$.

On montre ensuite que $2 \Rightarrow 3$. Si on a $\mathbb{E}[\sum_{n\geq 1} X_n] < \infty$, alors la variable aléatoire (positive) $\sum_{n\geq 1} X_n$ est finie p.s., et donc $\mathbb{P}(\sum_{n\geq 1} X_n < \infty) = 1$.

On montre maintenant que $3 \Rightarrow 1$. Si $\mathbb{P}(\sum_{n\geq 1} X_n < \infty) > 0$, alors on a

$$\mathbb{E}[\mathrm{e}^{-\sum_{n\geq 1} X_n}] = \mathbb{E}[\mathrm{e}^{-\sum_{n\geq 1} X_n} \mathbf{1}_{\{\sum_{n\geq 1} X_n < \infty\}}] > 0.$$

D'autre part, par indépendance, on a

$$\mathbb{E}[\mathrm{e}^{-\sum_{n\geq 1} X_n}] = \prod_{n\geq 1} \mathbb{E}[\mathrm{e}^{-X_n}] = \prod_{n\geq 1} \frac{\lambda_n}{1+\lambda_n}.$$

En particulier, comme ce produit est strictement positif, cela implique que $\lim_{n\to\infty} \dfrac{\lambda_n}{1+\lambda_n} = 1$ et donc $\lim_{n\to\infty} \lambda_n = \infty$. Remarquons enfin que

$$\prod_{n\geq 1} \frac{\lambda_n}{1+\lambda_n} = \mathrm{e}^{-\sum_{n\geq 1} \log(1+\lambda_n^{-1})}.$$

Comme le produit est strictement positif, la série $\sum_{n\geq 1} \log(1+\lambda_n^{-1})$ converge. Comme $\lim_{n\to\infty} \lambda_n = \infty$, cela implique que la série $\sum_{n\geq 1} \lambda_n^{-1}$ converge également. Ainsi on a montré que $3 \Rightarrow 1$.

▲

Exercice III.8.

1. Soit $D = \{(x_1,\dots,x_n) \in]0,1[^n ; x_i \neq x_j \quad \text{pour tout} \quad i \neq j\}$. Les ensembles $\Delta_\sigma = \{(x_1,\dots,x_n) \in]0,1[^n ; x_{\sigma(1)} < \cdots < x_{\sigma(n)}\}$, pour $\sigma \in \mathcal{S}_n$, où $\mathcal{S}_n$ est l'ensemble des permutations de $\{1,\dots,n\}$, forment une partition de D.

   Soit g une fonction mesurable bornée. On a

$$\begin{aligned}
\mathbb{E}[g(Y,Z)] &= \mathbb{E}[g(\min_{1\leq i\leq n} X_i, \max_{1\leq i\leq n} X_i)] \\
&= \int g(\min_{1\leq i\leq n} x_i, \max_{1\leq i\leq n} x_i) \mathbf{1}_{[0,1]^n}(x_1,\dots,x_n)\, dx_1 \dots dx_n \\
&= \int g(\min_{1\leq i\leq n} x_i, \max_{1\leq i\leq n} x_i) \mathbf{1}_{D}(x_1,\dots,x_n)\, dx_1 \dots dx_n \\
&= \sum_{\sigma\in\mathcal{S}_n} \int g(x_{\sigma(1)}, x_{\sigma(n)}) \mathbf{1}_{\Delta_\sigma}(x_1,\dots,x_n)\, dx_1 \dots dx_n \\
&= n! \int g(x_1,x_n) \mathbf{1}_{\Delta_{\sigma_0}}(x_1,\dots,x_n)\, dx_1 \dots dx_n,
\end{aligned}$$

où σ_0 est l'identité. La dernière égalité s'obtient par un argument de symétrie. On en déduit donc que

$$\begin{aligned}\mathbb{E}[g(Y,Z)] &= \int g(y,z)n!\mathbf{1}_{\{y<x_2<\ldots<x_{n-1}<z\}}\,dx_2\ldots dx_{n-1}\,\mathbf{1}_{\{0<y<z<1\}}\,dydz\\ &= \int g(y,z)\;n(n-1)(z-y)^{n-2}\mathbf{1}_{\{0<y<z<1\}}\,dydz.\end{aligned}$$

Donc (Y,Z) est une variable aléatoire continue de densité $f_{(Y,Z)}(y,z) = n(n-1)(z-y)^{n-2}\mathbf{1}_{\{0<y<z<1\}}$.

2. Par la formule des lois marginales, on en déduit que Y est une variable aléatoire continue de densité $f_Y(y) = n(1-y)^{n-1}\mathbf{1}_{\{0<y<1\}}$. Il s'agit de la loi $\beta(1,n)$. Par symétrie, on vérifie que la loi de Z est la loi $\beta(n,1)$ de densité $f_Z(z) = nz^{n-1}\mathbf{1}_{\{0<z<1\}}$.
3. Y étant intégrable, l'espérance conditionnelle $\mathbb{E}[Y|Z]$ a un sens. On a pour $z\in]0,1[$,

$$\begin{aligned}\mathbb{E}[Y|Z=z] &= \int y\frac{f_{(Y,Z)}(y,z)}{f_Z(z)}\,dy\\ &= \int_0^z (n-1)y(z-y)^{n-2}z^{1-n}\,dy\\ &= \left[-z^{1-n}(z-y)^{n-1}y\right]_0^z + \int_0^z z^{1-n}(z-y)^{n-1}\,dy\\ &= \frac{z}{n}.\end{aligned}$$

On en déduit donc $\mathbb{E}[Y|Z] = Z/n$.

4. On a pour $z\in]0,1[$,

$$\begin{aligned}\mathbb{E}[g(Y/Z)|Z=z] &= \int g(y/z)\frac{f_{(Y,Z)}(y,z)}{f_Z(z)}\,dy\\ &= \int g(y/z)(n-1)(z-y)^{n-2}z^{1-n}\mathbf{1}_{\{0<y<z\}}\,dy\\ &= \int g(\rho)(n-1)(1-\rho)^{n-2}\mathbf{1}_{\{0<\rho<1\}}\,d\rho,\end{aligned}$$

où l'on a effectué le changement de variable $\rho = y/z$ (à z fixé). On en déduit donc que pour toute fonction g bornée,

$$\mathbb{E}[g(Y/Z)|Z] = \int g(\rho)(n-1)(1-\rho)^{n-2}\mathbf{1}_{\{0<\rho<1\}}\,d\rho.$$

Donc la densité de la loi conditionnelle de Y/Z sachant Z est $(n-1)(1-\rho)^{n-2}\mathbf{1}_{\{0<\rho<1\}}$. On reconnaît une loi $\beta(1,n-1)$. Elle est indépendante de Z. Cela signifie donc que Y/Z est indépendante de Z. On retrouve alors

$$\mathbb{E}[Y|Z] = \mathbb{E}[Z\frac{Y}{Z}|Z] = Z\mathbb{E}[\frac{Y}{Z}] = Z/n.$$

5. Le résultat s'obtient par symétrie. En effet X_i a même loi que $1-X_i$. Donc $(1-X_1,\ldots,1-X_n)$ a même loi que $(X_1,\ldots,X_n)$. Comme $1-Z = \min_{1\le i\le n}(1-X_i)$ et $1-Y = \max_{1\le i\le n}(1-X_i)$, on en déduit donc que $(1-Z,1-Y)$ a même loi que (Y,Z).

6. On déduit de la question précédente que $\dfrac{1-Z}{1-Y}$ est indépendante de $1-Y$ donc de Y.

▲

Exercice III.9.

1. On note Z_k la variable aléatoire réelle donnant la hauteur du k-ième impact. Les variables aléatoires $(Z_k, k \geq 1)$ sont indépendantes de même loi $\mathcal{U}([0,h])$, et on a $Z = \min_{1\leq k\leq N} Z_k$. On en déduit, en utilisant l'indépendance entre $(Z_k, k \geq 1)$ et N, que pour $z \in [0,h]$ et $n > 0$:

$$\begin{aligned}\mathbb{P}(Z > z|N = n) &= \mathbb{P}(Z_1 > z, \ldots, Z_n > z|N = n)\\ &= \mathbb{P}(Z_1 > z, \ldots, Z_n > z)\\ &= \prod_{k=1}^{n} \mathbb{P}(Z_k > z)\\ &= \prod_{k=1}^{n} \left(1 - \frac{z}{h}\right)\\ &= \left(1 - \frac{z}{h}\right)^n.\end{aligned}$$

Donc on a $\mathbb{P}(Z < z|N = n) = 1 - \left(1 - \frac{z}{h}\right)^n$. Comme la fonction de répartition de Z sachant N est de classe C^1 (en z), on en déduit que conditionnellement à N, Z est une variable aléatoire continue de densité

$$f_{Z|N}(z|n) = \frac{n}{h}\left(1 - \frac{z}{h}\right)^{n-1}.$$

2. Le container étant parallélépipédique, le pourcentage τ_N de liquide qu'on peut espérer sauver sachant le nombre d'impacts N, correspond à l'espérance de $\mathbb{E}[Z|N]$ rapportée à la hauteur h du container. On a :

$$\begin{aligned}\frac{1}{h}\,\mathbb{E}[Z|N = n] &= \frac{1}{h}\int_0^h z f_{Z|N}(z|n)\, dz\\ &= \frac{1}{h}\int_0^h z \frac{n}{h}\left(1 - \frac{z}{h}\right)^{n-1} dz\\ &= \frac{1}{1+n}.\end{aligned}$$

On constate que cette formule couvre aussi le cas où $n = 0$. En effet, dans ce cas on sauve la totalité du chargement (le container n'a été touché par aucune balle). On a donc $\tau_N = 1/(N+1)$.

3. Si le convoi comporte un grand nombre de wagons, on peut supposer que le pourcentage moyen τ que l'on peut espérer sauver correspond à la moyenne des τ_N. On a donc :

$$\begin{aligned}\tau &= \mathbb{E}[\tau_N]\\ &= \sum_{n=0}^{+\infty} \frac{1}{1+n}\,\mathrm{e}^{-\theta}\,\frac{\theta^n}{n!}\\ &= \frac{1 - e^{-\theta}}{\theta}.\end{aligned}$$

Pour $\theta = 5$, on obtient $\tau = 19,8\%$.

▲

Exercice III.10.

On note Θ_1 et Θ_2 les angles formés par les deux rayons et le rayon qui passe par la cerise. L'énoncé du problème indique que Θ_1 et Θ_2 sont indépendants et suivent la loi uniforme sur $[0, 2\pi]$. La longueur angulaire de la part contenant la cerise est $2\pi - |\Theta_1 - \Theta_2|$. La longueur moyenne de la part contenant la cerise est donc $2\pi - \mathbb{E}[|\Theta_1 - \Theta_2|]$, et la probabilité pour que la part contenant la cerise soit la plus petite est $\mathbb{P}(2\pi - |\Theta_1 - \Theta_2| < |\Theta_1 - \Theta_2|)$. Comme les angles sont indépendants, la loi du couple est la loi produit. On calcule :

$$\begin{aligned}\mathbb{E}[|\Theta_1 - \Theta_2|] &= \frac{1}{(2\pi)^2} \iint_{[0,2\pi]^2} |\theta_1 - \theta_2| \; d\theta_1 d\theta_2 \\ &= \frac{1}{(2\pi)^2}\, 2 \int_{[0,2\pi]} d\theta_1 \int_{[0,\theta_1]} (\theta_1 - \theta_2)\; d\theta_2 \\ &= \frac{2\pi}{3}\end{aligned}$$

et

$$\begin{aligned}\mathbb{P}(2\pi - |\Theta_1 - \Theta_2| < |\Theta_1 - \Theta_2|) &= \frac{1}{(2\pi)^2} \iint_{[0,2\pi]^2} \mathbf{1}_{\{|\theta_1 - \theta_2| > \pi\}} d\theta_1 d\theta_2 \\ &= \frac{1}{(2\pi)^2}\, 2 \int_{[0,2\pi]} d\theta_1 \int_{[0,\theta_1]} \mathbf{1}_{\{\theta_1 - \theta_2 > \pi\}}\; d\theta_2 \\ &= \frac{1}{4}\,.\end{aligned}$$

La longueur moyenne de la part contenant la cerise est donc $4\pi/3$, et la probabilité pour que la part contenant la cerise soit la plus petite est $1/4$. La part qui contient la cerise est plus grande en moyenne et elle est également plus grande dans 75% des cas.

Pour voir que ce résultat ne contredit pas l'intuition il faut inverser les opérations. On découpe d'abord au hasard deux rayons dans le gâteau, puis on jette au hasard la cerise sur le bord. Celle-ci a intuitivement plus de chance de tomber sur la part la plus grosse ! Il reste à se convaincre que jeter la cerise sur le bord puis couper le gâteau au hasard, ou couper le gâteau au hasard puis jeter la cerise sur le bord donne bien le même résultat.

▲

Exercice III.11.

On suppose que la longueur du bâton est de une unité. On note X et Y les emplacements des deux marques. Par hypothèse X et Y sont des variables aléatoires indépendantes de loi uniforme sur $[0, 1]$. On fait un triangle si est seulement si aucune des longueur des morceaux n'est plus grande que la somme des deux autres. Cela est équivalent aux trois conditions suivantes :

$$\max(X, Y) \geq 1/2, \quad \min(X, Y) \leq 1/2, \quad \max(X, Y) - \min(X, Y) \leq 1/2.$$

On obtient

$$\begin{aligned}\mathbb{P}(\text{triangle}) &= \mathbb{P}(\max(X,Y)\geq 1/2, \min(X,Y)\leq 1/2, \max(X,Y)-\min(X,Y)\leq 1/2)\\ &= \mathbb{E}[\mathbf{1}_{\{\max(X,Y)\geq 1/2,\min(X,Y)\leq 1/2,\max(X,Y)-\min(X,Y)\leq 1/2\}}]\\ &= \int \mathbf{1}_{\{\max(x,y)\geq 1/2,\min(x,y)\leq 1/2,\max(x,y)-\min(x,y)\leq 1/2\}}\mathbf{1}_{[0,1]}(x)\mathbf{1}_{[0,1]}(y)\,dxdy\\ &= 2\int_{1/2}^{1} dx \int_{x-1/2}^{1/2} dy = 1/4\end{aligned}$$

▲

Exercice III.12.

On note $s < t$ les deux nombres choisis par votre ami et X la variable aléatoire de Bernoulli de paramètre $1/2$ qui modélise le lancer de sa pièce : $X = 0$ si il a obtenu face et $X = 1$ sinon. Le nombre donné par votre ami est

$$Y = s\mathbf{1}_{\{X=0\}} + t\mathbf{1}_{\{X=1\}}.$$

On note G l'évènement {gagner le pari}.

1. On modélise le lancer de votre pièce par une variable indépendante de X de loi de Bernoulli de paramètre $p \in [0,1]$, U : $U = 0$ si vous avez obtenu face et $U = 1$ sinon. On a

   $$G = \{U = 0, Y = s\} \cup \{U = 1, Y = t\} = \{U = 0, X = 0\} \cup \{U = 1, X = 1\}.$$

   En utilisant l'indépendance entre X et U, on en déduit que la probabilité de gagner est $\mathbb{P}(G) = (1-p)/2 + p/2 = 1/2$.

2. Par construction X et Z sont indépendant. On a

   $$G = \{Z \geq Y, Y = s\} \cup \{Z \leq Y, Y = t\} = \{Z \geq s, X = 0\} \cup \{Z \leq t, X = 1\}.$$

   Il vient, en utilisant l'indépendance entre X et Z, et $t > s$

   $$\mathbb{P}(G) = \frac{1}{2}[\mathbb{P}(Z \geq s) + \mathbb{P}(Z \leq t)] = \frac{1}{2} + \frac{1}{2}\,\mathbb{P}(Z \in [s,t]).$$

   Comme $\mathbb{P}(Z \in [s,t]) > 0$, on a $\mathbb{P}(G) > 1/2$.

3. On suppose que s et t sont les réalisations de variables aléatoire S et T indépendantes et de même loi de fonction de répartition F. Mais comme on a supposé que $s < t$, cela impose que s est la réalisation de $\min(S,T)$ et t la réalisation de $\max(S,T)$. Dans ce cas, le nombre fourni par votre ami est donc

   $$Y = \min(S,T)\mathbf{1}_{\{X=0\}} + \max(S,T)\mathbf{1}_{\{X=1\}}.$$

   On a $G = \{Z \geq \min(S,T), X = 0\} \cup \{Z \leq \max(S,T), X = 1\}$, et en utilisant l'indépendance de S, T, Z et X,

   $$\begin{aligned}\mathbb{P}(G) &= \frac{1}{2}[\mathbb{P}(Z \geq \min(S,T)) + \mathbb{P}(Z \leq \max(S,T))]\\ &= \frac{1}{2} + \frac{1}{2}\,\mathbb{P}(Z \in [\min(S,T), \max(S,T)]).\end{aligned}$$

Pour maximiser la probabilité de gagner, il faut donc maximiser la probabilité $\mathbb{P}(Z \in [\min(S,T), \max(S,T)])$. Supposons un instant que Z soit une variable continue de densité g. On note f la densité de S et T, il vient

$$\begin{aligned}\mathbb{P}(Z \in [\min(S,T), \max(S,T)]) &= \int \mathbf{1}_{\{z \in [\min(s,t), \max(s,t)]\}} g(z) f(s) f(t)\, dzdsdt \\ &= 2 \int \mathbf{1}_{\{z \in [s,t]\}} \mathbf{1}_{\{s<t\}} g(z) f(s) f(t)\, dzdsdt \\ &= 2 \int g(z) F(z)(1 - F(z))\, dz \\ &= 2\mathbb{E}[F(Z)(1 - F(Z))].\end{aligned}$$

Admettons dans un premier temps que $\mathbb{P}(Z \in [\min(S,T), \max(S,T)]) = 2\mathbb{E}[F(Z)(1 - F(Z))]$ même si Z n'est pas une variable continue. Comme $F(x) \in [0,1]$, la valeur de $F(x)(1-F(x))$ est maximale pour $x = x_{1/2}$, le quantile d'ordre $1/2$ de la loi de F (i.e. comme F est strictement croissante, $x_{1/2}$ est l'unique solution de $F(x) = 1/2$), et donc $2\mathbb{E}[F(Z)(1 - F(Z))] \leq 1/2$ avec égalité si $Z = x_{1/2}$ p.s. On obtient, si $Z = x_{1/2}$ p.s.,

$$\mathbb{P}(G) = \frac{1}{2}(1 + \frac{1}{2}) = \frac{3}{4}.$$

Il reste donc à vérifier que si $\mathbb{P}(Z = x_{1/2}) < 1$ alors $\mathbb{P}(Z \in [\min(S,T), \max(S,T)]) < 1/2$. On suppose il exsite $\varepsilon > 0$ et $\eta > 0$, tels que $\mathbb{P}(\big|Z - x_{1/2}\big| > \eta) > \varepsilon$. On décompose, pour n fixé, suivant $Z \in [\frac{k}{n}, \frac{k+1}{n}[$, pour $k \in \mathbb{N}$. On calcule

$$\begin{aligned}&\mathbb{P}(Z \in [\frac{k}{n}, \frac{k+1}{n}[, Z \in [\min(S,T), \max(S,T)]) \\ &\qquad \leq \mathbb{P}(Z \in [\frac{k}{n}, \frac{k+1}{n}[, \min(S,T) \leq \frac{k+1}{n}, \max(S,T) \geq \frac{k}{n}) \\ &\qquad = 2\mathbb{P}\Big(Z \in [\frac{k}{n}, \frac{k+1}{n}[\Big) F(\frac{k+1}{n})(1 - F(\frac{k}{n})).\end{aligned}$$

En sommant sur k, et en distinguant suivant $\frac{k}{n} > x_{1/2} + \eta$, $\frac{k+1}{n} < x_{1/2} - \eta$ et $x_{1/2} - \eta - \frac{1}{n} \leq \frac{k}{n} \leq x_{1/2} + \eta$, on obtient

$$\begin{aligned}&\mathbb{P}(Z \in [\min(S,T), \max(S,T)]) \\ &\qquad \leq 2\mathbb{P}(Z - x_{1/2} > \eta) \sup_{x > x_{1/2} + \eta} F(x + \frac{1}{n})(1 - F(x)) \\ &\qquad\quad + 2\mathbb{P}(Z - x_{1/2} < -\eta) \sup_{x < x_{1/2} + \eta} F(x + \frac{1}{n})(1 - F(x)) \\ &\qquad\quad + 2\mathbb{P}(\big|Z - x_{1/2}\big| \leq \eta) \sup_{x; \big|x - x_{1/2}\big| \leq \eta} F(x + \frac{1}{n})(1 - F(x)).\end{aligned}$$

En laissant $n \to \infty$, et en utilisant le fait que F est continue, on obtient

$$\mathbb{P}(Z \in [\min(S,T), \max(S,T)]) \leq \mathbb{P}(\big|Z - x_{1/2}\big| > \eta) q + \frac{1}{2}\mathbb{P}(\big|Z - x_{1/2}\big| \leq \eta),$$

où $q = \sup_{x;\,|x-x_{1/2}|>\eta}(F(x)(1-F(x)) = \max(F(x_{1/2}+\eta)(1-F(x_{1/2}+\eta)), F(x_{1/2}-\eta)(1-F(x_{1/2}-\eta))$. Comme F est strictement croissante, on a $q < 1/4$ et donc $\mathbb{P}(Z \in [\min(S,T), \max(S,T)]) < 1/2$. On a donc démontré que la quantité $\mathbb{P}(Z \in [\min(S,T), \max(S,T)])$ est maximale si et seulement si $Z = x_{1/2}$ p.s.

▲

Exercice III.13.

1. Invariante par rotation, la densité du couple (X,Y,Z) est de la forme $f_{(X,Y,Z)}(x,y,z) = \phi(x^2+y^2+z^2)$ avec $\phi : \mathbb{R}^+ \to \mathbb{R}^+$. Les variables X, Y et Z étant indépendantes, on a d'autre part
$$f_{(X,Y,Z)}(x,y,z) = f_X(x) f_Y(y) f_Z(z).$$
La loi de (X,Y,Z) étant invariante par rotation, on en déduit que X, Y et Z ont même loi de densité : $f = f_X = f_Y = f_Z$. L'égalité
$$\forall (x,y,z) \in \mathbb{R}^3, \quad f(x)f(y)f(z) = \phi(x^2+y^2+z^2),$$
implique $f(x)f'(y)f(z) = 2y\phi'(x^2+y^2+z^2)$ et $f'(x)f(y)f(z) = 2x\phi'(x^2+y^2+z^2)$. En faisant le rapport des ces deux égalités, on en déduit qu'il existe une constante c telle que $\forall x \in \mathbb{R}$, $\dfrac{f'(x)}{f(x)} = 2cx$. Par intégration, on en déduit que $f(x) = a\,\mathrm{e}^{cx^2}$. Finalement, les conditions sur f en qualité de densité de probabilité imposent que X, Y et Z sont des variables aléatoires indépendantes et identiquement distribuées de loi gaussienne centrée : $f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-\frac{x^2}{2\sigma^2}}$.
2. L'énergie cinétique moyenne est définie par $\mathrm{E}_c = \frac{1}{2}\,m\mathbb{E}[|V|^2] = \frac{1}{2}\,m\mathbb{E}[V_1^2+V_2^2+V_3^2]$. Comme $\mathbb{E}[V_1^2] = \sigma^2$, on en déduit que $\mathrm{E}_c = \dfrac{3}{2}m\sigma^2$. On obtient $\sigma^2 = \dfrac{kT}{m}$.
3. Voir l'exercice .4. La loi de $X+Y$ est la loi $\Gamma(\lambda, a+b)$.
4. La loi de V_i^2 est une loi gamma de paramètres $(\frac{1}{2\sigma^2}, 1/2)$. On déduit de la question précédente que la loi de $|V|^2$ est une loi gamma de paramètres $(\frac{1}{2\sigma^2}, 3/2)$. Sa densité est $\dfrac{1}{\sqrt{2\pi}\sigma^3}\,\sqrt{z}\,\mathrm{e}^{-z/2\sigma^2}\,\mathbf{1}_{\{z>0\}}$. La loi de Maxwell est la loi de $|V|$ dont la densité est
$$\frac{\sqrt{2}}{\sqrt{\pi}}\left(\frac{m}{kT}\right)^{3/2} v^2\,\mathrm{e}^{-mv^2/2kT}\,\mathbf{1}_{\{v>0\}}.$$

▲

Exercice III.14.

1. Par une application du théorème de Pythagore, on trouve que la corde est de longueur L :
$$L = 2 * \sqrt{r^2 - H^2}.$$
Pour calculer sa loi, on utilise la méthode de la fonction muette. Soit f une fonction continue bornée, comme H est uniforme sur $[0,r]$,
$$\mathbb{E}\left[f(L)\right] = \frac{1}{r}\int_0^r f\left(2 * \sqrt{r^2-h^2}\right) dh.$$

On effectue le changement de variable $u = 2 * \sqrt{r^2 - h^2}$. Cela donne

$$\mathbb{E}\left[f(L)\right] = \int_0^{2r} f(u) \frac{u du}{4r\sqrt{r^2 - \frac{u^2}{4}}}.$$

La variable aléatoire L est continue de densité $\frac{u}{4r\sqrt{r^2-u^2/4}} \mathbf{1}_{[0,2r]}(u)$. L'espérance de L est (on effectue le changement de variable $u = 2r\sin(t)$)

$$\mathbb{E}\left[L\right] = \int_0^{2r} \frac{u^2 du}{4r\sqrt{r^2 - \frac{u^2}{4}}} = \int_0^{\pi/2} 2r\sin^2(t)dt = \int_0^{\pi/2} r(1-\cos(2t))dt = \frac{r\pi}{2}.$$

Pour calculer la variance, on commence par calculer $\mathbb{E}[L^2]$. Utilisant la même méthode que pour la moyenne, on obtient

$$\mathbb{E}[L^2] = \int_0^{2r} \frac{u^3 du}{4r\sqrt{r^2 - \frac{u^2}{4}}} = \int_0^{\pi/2} 4r^2\sin^3(t)dt = 4r^2\int_0^{\pi/2} \sin(t)(1-\cos^2(t))dt = \frac{8r^2}{3}.$$

La variance est donc

$$\mathrm{Var}(L) = \frac{8r^2}{3} - \frac{r^2\pi^2}{4}.$$

La probabilité que la corde soit plus grande qu'un côté du triangle équilatéral inscrit est la probabilité que sa distance au centre soit plus grande que $r/2$. Comme la loi de H est uniforme sur $[0, r]$,

$$p = \mathbb{P}\left[H \geq r/2\right] = 1/2.$$

2. Si A et B sont choisis uniformément sur le cercle, leurs angles au centre (θ_A et θ_B) par rapport à l'axe des x est uniforme sur $]-\pi, \pi[$. On a $\alpha = (\theta_A - \theta_B)$ modulo 2π. La longueur de la corde est alors $L = 2r\left|\sin\left(\alpha/2\right)\right|$. On utilise la méthode de la fonction muette pour calculer la loi de α. On effectue le changement de variable bijectif

$$\varphi\begin{pmatrix}\alpha'\\ \beta\end{pmatrix} = \varphi(\theta_A, \theta_B) = \begin{pmatrix}\theta_A - \theta_B\\ \theta_B\end{pmatrix}$$

de $]0, 2\pi[^2$ dans

$$\Delta = \{(\alpha', \beta); -2\pi < \alpha' < 2\pi, \max(-\pi, \alpha' - \pi) < \beta < \min(\pi, \alpha' + \pi)\}.$$

Le déterminant de la matrice Jacobienne est

$$\mathrm{Jac}[\varphi](\theta_A, \theta_B) = \begin{vmatrix}1 & -1\\ 0 & 1\end{vmatrix} = 1.$$

Soit f une fonction bornée continue,

$$\begin{aligned}\mathbb{E}[f(\alpha')] &= \frac{1}{4\pi^2}\int_{-\pi}^{\pi} d\theta_A \int_{-\pi}^{\pi} d\theta_B\ f(\theta_A - \theta_B)\\ &= \frac{1}{4\pi^2}\int_{\Delta} f(\alpha')\ d\alpha' d\beta\\ &= \frac{1}{4\pi^2}\int_{-2\pi}^{0} (\alpha' + 2\pi) f(\alpha')\ d\alpha' + \frac{1}{4\pi^2}\int_0^{2\pi} (2\pi - \alpha') f(\alpha')\ d\alpha'.\end{aligned}$$

L'angle qui nous intéresse est $\alpha = \alpha'$ modulo 2π. On a

$$\begin{aligned}\mathbb{E}[f(\alpha)] &= \mathbb{E}[f(\alpha'+2\pi)\mathbf{1}_{]-2\pi,0[}(\alpha')] + \mathbb{E}[f(\alpha')\mathbf{1}_{]0,2\pi[}(\alpha')] \\ &= \frac{1}{4\pi^2}\int_{-2\pi}^{0}(\alpha'+2\pi)f(\alpha'+2\pi)\,d\alpha' + \frac{1}{4\pi^2}\int_0^{2\pi}(2\pi-\alpha')f(\alpha')\,d\alpha' \\ &= \frac{1}{2\pi}\int_0^{2\pi} f(\alpha')\,d\alpha'.\end{aligned}$$

On en déduit que la loi de α est la loi uniforme sur $]0, 2\pi[$.
On utilise, encore la méthode de la fonction muette pour déterminer la loi de L.

$$\mathbb{E}\left[f(L)\right] = \frac{1}{2\pi}\int_0^{2\pi} d\alpha\; f\left(2r\left|\sin(\alpha/2)\right|\right) = \frac{2}{\pi}\int_0^{\pi/2} f(2r\sin(\alpha))d\alpha. \tag{III.1}$$

On effectue le changement de variable $u = 2r\sin(\alpha)$,

$$\mathbb{E}[f(L)] = \int_0^{\pi/2} f(2r\sin(\alpha))\frac{2d\alpha}{\pi} = 2\int_0^{2r} f(u)\frac{du}{\pi\sqrt{4r^2-u^2}}.$$

L est donc une variable aléatoire continue de densité $\dfrac{1}{\pi\sqrt{4r^2-u^2}}\mathbf{1}_{[0,2r]}(u)$.
Pour calculer l'espérance et la variance de L, on utilise (III.1) :

$$\mathbb{E}\left[L\right] = \frac{2r}{\pi}\int_0^{\pi}\sin(t)dt = \frac{4r}{\pi},$$

et pour la variance,

$$\mathrm{Var}(L) = \mathbb{E}[L^2] - (\mathbb{E}[L])^2 = \frac{4r^2}{\pi}\int_0^{\pi}\sin^2(t)dt - \frac{16r^2}{\pi^2} = 2r^2 - \frac{16r^2}{\pi^2}.$$

La probabilité p est calculée en exploitant la loi de L, et en effectuant le changement de variable $u = 2r\sin(\alpha)$:

$$p = \mathbb{P}\left(L \geq \sqrt{3}r\right) = 2\int_{\sqrt{3}r}^{2r}\frac{2du}{\pi\sqrt{4r^2-u^2}} = \int_{\pi/3}^{\pi/2}\frac{2d\alpha}{\pi} = 1/3.$$

3. On calcule la longueur de la corde comme a la question 1, à partir de la distance de I au centre du cercle :
$$L = 2\sqrt{r^2-x^2-y^2}.$$
La loi de L est calculée par la méthode de la fonction muette :

$$\mathbb{E}\left[f(L)\right] = \frac{1}{\pi r^2}\int_{-r}^{r}dx\int_{-r}^{r}dy\; f\left(2\sqrt{r^2-x^2-y^2}\right)\mathbf{1}_{\{x^2+y^2\leq r^2\}}.$$

On effectue un changement de coordonnées polaires dans $\mathbb{R}^2$:

$$\mathbb{E}\left[f(L)\right] = \frac{1}{\pi r^2}\int_0^r udu\int_{-\pi}^{\pi}d\theta\; f\left(2\sqrt{r^2-u^2}\right) = \frac{2}{r^2}\int_0^r f\left(2\sqrt{r^2-u^2}\right)u\,du,$$

On conclut par le même changement de variable qu'à la question 1 :

$$\mathbb{E}\left[f(L)\right] = \frac{1}{2r^2}\int_0^{2r} f(v)v dv.$$

La variable L est donc continue de densité $\dfrac{v}{2r^2}\mathbf{1}_{[0,2r]}(v)$.
La moyenne et la variance de cette loi se calcule immédiatement :

$$\mathbb{E}[L] = \frac{1}{2r^2}\int_0^{2r} v^2 dv = \frac{4r}{3}; \qquad \mathrm{Var}(L) = \frac{1}{2r^2}\int_0^{2r} v^3 dv - \frac{16r^2}{9} = 2r^2 - \frac{16r^2}{9} = \frac{2r^2}{9}.$$

Enfin, la probabilité que la corde soit de longueur plus grande que $\sqrt{3}r$ est

$$p = \frac{1}{2r^2}\int_{\sqrt{3}r}^{2r} v dv = \frac{1}{2r^2}\,\frac{4r^2 - 3r^2}{2} = 1/4.$$

On résume dans le tableau ci-dessous, les valeurs numériques (avec $r = 1$) des différentes moyennes et variances et des valeurs de p. On voit bien que la notion de "choisir au hasard" dans un énoncé doit être précisée.

Cas	moyenne	variance	p
1	1.57	0.2	1/2
2	1.27	0.37	1/3
3	1.33	0.22	1/4

▲

Exercice III.15.

1. Les variables X et Y étant indépendantes, la loi du couple (X, Y) est :

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right).$$

2. Soit $\Delta = \{(r,\theta) \in \mathbb{R}^2; r > 0, -\pi < \theta < \pi\}$. On considère la fonction φ définie sur Δ :

$$\varphi\begin{pmatrix} r \\ \theta \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \text{où} \quad x = r\cos\theta, \quad y = r\sin\theta.$$

Cette fonction établit un C^1-difféomorphisme entre Δ et $\mathcal{D} = \mathbb{R}^2\backslash\{\mathbb{R}^- \times \{0\}\}$, dont l'inverse a pour expression :

$$\varphi^{-1}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} r \\ \theta \end{pmatrix} \quad \text{où} \quad r = \sqrt{x^2+y^2}, \quad \theta = \mathrm{sgn}(y)\arccos\left(x/\sqrt{x^2+y^2}\right).$$

En effet, on a :
– φ est injective : si $(x_1, y_1) \in \mathcal{D}$ et $(x_2, y_2) \in \mathcal{D}$ sont tels que $(x_1, y_1) = (x_2, y_2)$, alors $r_1 = \sqrt{(x_1^2 + y_1^2)} = \sqrt{(x_2^2 + y_2^2)} = r_2$ avec $r_1 > 0$, et $\cos\theta_1 = x_1/r_1 = x_2/r_2 = \cos\theta_2$ ainsi que $\sin\theta_1 = y_1/r_1 = y_2/r_2 = \sin\theta_2$.
Comme $\theta_1 \in]-\pi, \pi[$ et $\theta_2 \in]-\pi, \pi[$, l'égalité des cosinus donne que $|\theta_1| = |\theta_2|$ et l'égalité des sinus donne que $\mathrm{sgn}(\theta_1) = \mathrm{sgn}(\theta_2)$, d'où $\theta_1 = \theta_2$.

– φ est surjective : si $(x,y) \in \mathcal{D}$, on a $\sqrt{x^2+y^2} > 0$ et $x/\sqrt{x^2+y^2} \in]-1,1]$, donc

$$\cos\left(\arccos\left(x/\sqrt{x^2+y^2}\right)\right) = x/\sqrt{x^2+y^2}$$

et

$$\sin\left(\arccos\left(x/\sqrt{x^2+y^2}\right)\right) = \sqrt{1-x^2/(x^2+y^2)} = |y|/\sqrt{x^2+y^2}.$$

Si on pose $r = \sqrt{x^2+y^2}$ et $\theta = \mathrm{sgn}(y)\arccos(x/\sqrt{x^2+y^2})$, on a donc bien :

$$\begin{aligned} r\cos\theta &= r\cos\left(\mathrm{sgn}(y)\arccos\left(x/\sqrt{x^2+y^2}\right)\right) \\ &= r\cos\left(\arccos\left(x/\sqrt{x^2+y^2}\right)\right) \\ &= x \end{aligned}$$

et

$$\begin{aligned} r\sin\theta &= r\sin\left(\mathrm{sgn}(y)\arccos\left(x/\sqrt{x^2+y^2}\right)\right) \\ &= \mathrm{sgn}(y)\sin\left(\arccos\left(x/\sqrt{x^2+y^2}\right)\right) \\ &= \mathrm{sgn}(y)|y| \\ &= y. \end{aligned}$$

– φ est C^1 : la matrice jacobienne de φ existe pour tout $(r,\theta) \in \Delta$ et ses éléments sont des fonctions C^1 de (r,θ) :

$$\nabla\varphi(r,\theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}$$

– φ^{-1} est C^1 : la matrice jacobienne de φ^{-1} existe pour tout $(x,y) \in \mathcal{D}\backslash(\mathbb{R}^+ \times \{0\})$ et ses éléments sont des fonctions C^1 de (x,y) :

$$\nabla\varphi^{-1}(x,y) = \begin{pmatrix} x/\sqrt{x^2+y^2} & y/\sqrt{x^2+y^2} \\ -|y|/(x^2+y^2) & x/(x^2+y^2) \end{pmatrix}$$

La vérification du caractère C^1 de φ^{-1} pour les points $(x,0) \in \mathcal{D}$ se fait par une étude directe. Pour $x > 0$, $\varphi^{-1}(x,0) = (x, \mathrm{sgn}(0)\arccos(1)) = (x,0)$ et :

$$\begin{aligned} &\varphi^{-1}(x+h_x, 0+h_y) - \varphi^{-1}(x,0) \\ &\qquad = (\sqrt{(x+h_x)^2+h_y^2} - x, \mathrm{sgn}(h_y)\arccos\left((x+h_x)/\sqrt{(x+h_x)^2+h_y^2}\right)) \\ &\qquad = (h_x, h_y/x) + o(h_x^2, h_y^2), \end{aligned}$$

ce qui donne à la fois la continuité, la différentiabilité et la continuité des dérivées partielles de φ^{-1} en $(x,0)$.

Soit g une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$, mesurable et bornée. On remarque que $(X,Y) \in \mathcal{D}$ $\mathbb{P}$-p.s, donc $(R,\Theta) = \varphi^{-1}(X,Y)$ est défini $\mathbb{P}$-p.s. On a :

$$\begin{aligned}\mathbb{E}[g(R,\Theta)] &= \mathbb{E}[g(\sqrt{X^2+Y^2}, \operatorname{sgn}(Y)\arccos(X/\sqrt{X^2+Y^2})] \\ &= \int_{\mathcal{D}} g(\sqrt{x^2+y^2}, \operatorname{sgn}(y)\arccos(x/\sqrt{x^2+y^2}) \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) dxdy\end{aligned}$$

or $dxdy = |\mathrm{Jac}[\varphi](\mathrm{r},\theta)|\mathrm{drd}\theta = \mathrm{rdrd}\theta$ donc :

$$\begin{aligned}\mathbb{E}[g(R,\Theta)] &= \int_{\Delta} g(r,\theta)\frac{1}{2\pi\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) r dr d\theta \\ &= \int_{\mathbb{R}^2} g(r,\theta)\frac{r}{2\pi\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) \mathbf{1}_{]0,\infty[}(r)\mathbf{1}_{]-\pi,\pi[}(\theta) dr d\theta.\end{aligned}$$

On en déduit que la densité de (R,Θ) est :

$$f_{R,\Theta}(r,\theta) = \frac{r}{2\pi\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)\mathbf{1}_{]0,\infty[}(r)\mathbf{1}_{]-\pi,\pi[}.$$

Comme $f_{R,\Theta}(r,\theta) = f_R(r) f_\Theta(\theta)$ avec

$$f_R(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)\mathbf{1}_{]0,\infty[}(r)$$

et

$$f_\Theta(\theta) = \frac{1}{2\pi}\mathbf{1}_{]-\pi,\pi[}(\theta),$$

on en déduit que R et Θ sont indépendantes.

3. D'après la question précédente, la loi de R a pour densité :

$$f_R(r) = \frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right)\mathbf{1}_{]0,\infty[}(r)$$

4. L'espérance de R vaut :

$$\begin{aligned}\mathbb{E}[R] = \mu &= \int_0^{+\infty} r\frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= \left[-r\exp\left(-\frac{r^2}{2\sigma^2}\right)\right]_0^{+\infty} + \int_0^{+\infty}\exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= \sigma\sqrt{\frac{\pi}{2}}.\end{aligned}$$

La variance de R vaut :

$$\begin{aligned}\mathrm{Var}(R) &= \int_0^{+\infty}(r-\mu)^2\frac{r}{\sigma^2}\exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= \left[-(r-\mu)^2\exp\left(-\frac{r^2}{2\sigma^2}\right)\right]_0^{+\infty} + 2\int_0^{+\infty}(r-\mu)\exp\left(-\frac{r^2}{2\sigma^2}\right) dr \\ &= \mu^2 + 2(\sigma^2-\mu^2) \\ &= (2-\pi/2)\sigma^2.\end{aligned}$$

▲

Exercice III.16.

1. En supposant que la rainure de gauche a pour abscisse $-d/2$ et celle de droite $d/2$, l'aiguille coupe une rainure sous la condition :
$$|X| > \frac{d}{2} - \frac{l}{2}\cos\theta$$
2. Au vu des hypothèses, il est naturel de proposer que X et θ sont indépendantes et de loi uniforme respectivement sur $[-d/2, d/2]$ et $[-\pi/2, \pi/2]$. La probabilité cherchée vaut :
$$\begin{aligned}\mathbb{P}(|X| > \frac{d}{2}\cos\theta - l/2) &= \int \mathbf{1}_{\{|x|>d/2-\frac{l}{2}\cos\theta\}}\frac{1}{\pi d}\mathbf{1}_{[-d/2,d/2]\times[-\pi/2,\pi/2]}(x,\theta)\,dxd\theta \\ &= 2\int_{-\pi/2}^{\pi/2}\left(\int_{d/2-\frac{l}{2}\cos\theta}^{d/2}\frac{1}{\pi d}dx\right)d\theta \\ &= \frac{2}{\pi d}\int_{-\pi/2}^{\pi/2}\frac{l}{2}\cos\theta\,d\theta \\ &= \frac{2l}{\pi d}.\end{aligned}$$
3. On note Y_i le résultat du i-ème lancer : $Y_i = 1$ si l'aiguille coupe la rainure et 0 sinon. La suite $(Y_i, i \in \mathbb{N}^*)$ forme un schéma de Bernoulli de paramètre $p = 2l/\pi d$. D'après la loi faible des grands nombres (Proposition II.33 du cours), la moyenne empirique $\frac{N_n}{n}$ converge en probabilité vers l'espérance de Y_1, c'est-à dire vers $2l/\pi d$. On a ainsi un moyen expérimental de calculer une approximation de $1/\pi$ et donc de π.
4. On veut trouver n tel que $\mathbb{P}(\left|\frac{N_n}{n} - \frac{1}{\pi}\right| > 10^{-2}) \leq 5\%$. On déduit de la démonstration de la loi faible des grands nombres que
$$\mathbb{P}\left(\left|\frac{N_n}{n} - \frac{1}{\pi}\right| > a\right) \leq \frac{\frac{1}{\pi}(1-\frac{1}{\pi})}{na^2}.$$
On trouve $n = 43398$. Mais la précision à 10^{-2} sur $1/\pi$ correspond à une précision de l'ordre de 10^{-1} sur π.
5. Il n'y a pas de contradiction avec le résultat précédent qui quantifiait le nombre de lancer n à effectuer pour que dans 95% des réalisations de ces lancers on obtienne une approximation à 10^{-2} de $1/\pi$. Cela n'impose pas que l'approximation soit systématiquement de l'ordre de 10^{-2} ! Avec 355 lancers, la meilleure approximation de $1/\pi$ que l'on puisse obtenir est précisément 113/355, et cette approximation est de très bonne qualité car 113/355 est une fraction de la suite des approximations de $1/\pi$ en fractions continues. Le caractère artificiel de ce résultat apparaît si l'on effectue un lancer supplémentaire : on obtient 113/356 ou 114/356 comme approximation de $1/\pi$, qui sont des approximations à 10^{-3} et 2.10^{-3} respectivement. Cette brutale perte de précision caractérise la supercherie du choix de 355 lancers et certainement du nombre "aléatoire" de 113 lancers avec intersection.

▲

Chapitre IV

Fonctions caractéristiques

IV.1 Énoncés

Exercice IV.1.
Autour des lois gamma.

1. Soit X une variable aléatoire de loi $\Gamma(\lambda, \alpha)$. Calculer son espérance et sa variance.
2. Soit X_1, X_2 deux variables aléatoires indépendantes et de loi respective $\Gamma(\lambda, \alpha_1)$ et $\Gamma(\lambda, \alpha_2)$. Le paramètre λ est identique. Montrer que la loi de $X_1 + X_2$ est également une loi gamma de paramètre $(\lambda, \alpha_1 + \alpha_2)$.
3. En déduire que si $(X_n, n \in \mathbb{N}^*)$ est une suite de variables aléatoires indépendantes de loi exponentielle de paramètre $\lambda > 0$, alors la loi de la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ est la loi $\Gamma(n\lambda, n)$. Calculer la variance de $\bar{X}_n$ et retrouver la loi faible des grands nombres pour la suite $(X_n, n \in \mathbb{N}^*)$.

△

Exercice IV.2.
Soit Y une v.a.c. de loi exponentielle $\lambda > 0$ et ε une v.a.d. indépendante de Y et telle que $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$. Calculer la densité et la fonction caractéristique de $Z = \varepsilon Y$ (loi exponentielle symétrique). En déduire la fonction caractéristique de la loi de Cauchy.

△

Exercice IV.3.
Soit N une variable aléatoire discrète à valeurs dans $\mathbb{N}$. Soit $(X_k, k \in \mathbb{N})$ une suite de variables aléatoires continues de même loi, indépendantes et indépendantes de N. On pose $S_0 = 0$, et pour $n \geq 1$, $S_n = \sum_{k=1}^n X_k$.

1. Calculer $\mathbb{E}[S_N]$ et $\mathrm{Var}(S_N)$.
2. Calculer la fonction caractéristique de S_N, et retrouver les résultats précédents.

△

Exercice IV.4.
Soit X une variable aléatoire réelle dont la fonction caractéristique est $\psi_X(u)$.

1. Montrer que X est symétrique (i.e. X et $-X$ ont même loi) si et seulement si $\psi_X(u) \in \mathbb{R}$ pour tout $u \in \mathbb{R}$.
2. Montrer que $|\psi_X(u)|^2$ est la fonction caractéristique d'une variable aléatoire réelle. On pourra écrire $|\psi_X(u)|^2$ comme le produit de deux fonctions.
3. Qu'est-ce qu'on peut dire à propos des fonctions caractéristiques des variables aléatoires réelles qui sont symétriques par rapport à $a \neq 0$ (i.e. X et $2a - X$ ont même loi) ?

△

Exercice IV.5.
La *loi de défaut de forme* est utilisée pour la maîtrise statistique des procédés (MSP). Cette loi est décrite dans les normes AFNOR (E60-181) et CNOMO (E 41 32 120 N) et sert à quantifier les défauts géométriques de type planéité, parallélisme, circularité. Il s'agit de la loi de $|X - Y|$ où X et Y sont deux v.a. indépendantes suivant respectivement les lois $\mathcal{N}(\mu_x, \sigma_x)$ et $\mathcal{N}(\mu_y, \sigma_y)$.

1. Calculer la loi de $X - Y$.
2. En déduire la loi de $Z = |X - Y|$.
3. Calculer $\mathbb{E}[Z]$, $\mathbb{E}[Z^2]$ et $\mathrm{Var}(Z)$.

△

IV.2 Corrections

Exercice IV.1.

1. En utilisant $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$, on obtient $\mathbb{E}[X]=\alpha/\lambda$ et $\operatorname{Var} X=\alpha/\lambda^2$.
2. Par indépendance, on a
$$\psi_{X_1+X_2}(u)=\psi_{X_1}(u)\psi_{X_2}(u)=\left(\frac{\lambda}{\lambda-iu}\right)^{\alpha_1}\left(\frac{\lambda}{\lambda-iu}\right)^{\alpha_2}=\left(\frac{\lambda}{\lambda-iu}\right)^{\alpha_1+\alpha_2}.$$
La loi de X_1+X_2 est donc la loi $\Gamma(\lambda,\alpha_1+\alpha_2)$.
3. On en déduit par récurrence que la loi de $\sum_{i=1}^n X_i$ est la loi $\Gamma(\lambda,n)$. En regardant encore les fonctions caractéristiques on montre que si Z a la loi $\Gamma(a,b)$ alors cZ a la loi $\Gamma(a/c,b)$ pour tout $a,b,c>0$:
$$\psi_{cZ}(u)=\left(\frac{a}{a-icu}\right)^b=\left(\frac{a/c}{a/c-iu}\right)^b.$$
On en déduit que $\bar{X}_n$ a pour loi $\Gamma(n\lambda,n)$. On sait que $\mathbb{E}(\bar{X}_n)=\lambda^{-1}$ et que $\operatorname{Var}(\bar{X}_n)=1/n\lambda^2$. L'inégalité de Tchebychev nous donne que pour tout $\varepsilon>0$,
$$\mathbb{P}(|\bar{X}_n-\lambda^{-1}|>\varepsilon)\le\frac{\operatorname{Var}\bar{X}_n}{\varepsilon^2}=\frac{1}{\varepsilon^2 n\lambda^2}.$$
On en déduit que la suite $(\bar{X}_n, n\in\mathbb{N}^*)$ converge en probabilité vers $1/\lambda$.

▲

Exercice IV.2.

On a déjà calculé la densité de la loi de la v.a. continue Z (voir exercice III.2) : $f_Z(z)=\frac{\lambda}{2}\,\mathrm{e}^{-\lambda|z|}$. On utilise la formule de décomposition pour obtenir
$$\begin{aligned}\psi_Z(u)&=\mathbb{E}\left[\mathrm{e}^{iuY}\,\mathbf{1}_{\{\varepsilon=1\}}\right]+\mathbb{E}\left[\mathrm{e}^{-iuY}\,\mathbf{1}_{\{\varepsilon=-1\}}\right]\\&=\frac{1}{2}\left[\frac{\lambda}{\lambda-iu}+\overline{\left(\frac{\lambda}{\lambda-iu}\right)}\right]\\&=\frac{\lambda^2}{\lambda^2+u^2}.\end{aligned}$$
On remarque que, à un coefficient multiplicatif près, la fonction caractéristique de la loi de Z est la densité de la loi de Cauchy de paramètre λ. À l'aide du théorème d'inversion de la transformée de Fourier pour les fonctions intégrables, on a donc
$$f_Z(z)=\int_{\mathbb{R}}\mathrm{e}^{-iuz}\,\psi_Z(u)\,\frac{du}{2\pi},\quad\text{soit}\quad\frac{\lambda}{2}\,\mathrm{e}^{-\lambda|z|}=\int_{\mathbb{R}}\mathrm{e}^{-iuz}\,\frac{\lambda^2}{\lambda^2+u^2}\frac{du}{2\pi}.$$
On en déduit ainsi la fonction caractéristique de la loi de Cauchy de paramètre λ :
$$\int_{\mathbb{R}}\mathrm{e}^{iuz}\,\frac{1}{\pi}\,\frac{\lambda}{\lambda^2+u^2}\,du=\mathrm{e}^{-\lambda|z|}.$$

▲

Exercice IV.3.

1. En utilisant l'espérance conditionnelle, on a pour $n \in \mathbb{N}$,

$$\mathbb{E}[S_N|N=n] = \mathbb{E}[\sum_{k=1}^{n} X_k|N=n] = n\mathbb{E}[X_1|N=n] = n\mathbb{E}[X_1].$$

On en déduit donc $\mathbb{E}[S_N|N] = N\mathbb{E}[X_1]$ et $\mathbb{E}[S_N] = \mathbb{E}[N]\mathbb{E}[X_1]$. Le calcul de la variance est similaire. On a pour $n \in \mathbb{N}$,

$$\mathbb{E}[S_N^2|N=n] = \mathbb{E}[S_n^2] = \mathrm{Var}(S_n) + \mathbb{E}[S_n]^2 = n\,\mathrm{Var}(X_1) + n^2\mathbb{E}[X_1]^2.$$

On en déduit donc que $\mathbb{E}[S_N^2] = \mathbb{E}[N]\,\mathrm{Var}(X_1) + \mathbb{E}[N^2]\mathbb{E}[X_1]^2$. On obtient alors la formule de Wald :

$$\mathrm{Var}(S_N) = \mathbb{E}[N]\,\mathrm{Var}(X_1) + \mathbb{E}[X_1]^2\,\mathrm{Var}(N).$$

On peut également utiliser une formule de décomposition pour donner la réponse.

2. De la même manière, on a

$$\mathbb{E}[\mathrm{e}^{iuS_N}\,|N=n] = \mathbb{E}[\mathrm{e}^{iuS_n}] = \psi_{X_1}(u)^n,$$

où ψ_{X_1} est la fonction caractéristique de X_1. On a utilisé le fait que les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes et de même loi. Il vient

$$\mathbb{E}[\mathrm{e}^{iuS_N}] = \sum_{n=0}^{\infty} \mathbb{E}[\mathrm{e}^{iuS_N}\,|N=n]\mathbb{P}(N=n) = \sum_{n=0}^{\infty} \psi_{X_1}(u)^n\mathbb{P}(N=n) = \phi(\psi_{X_1}(u)),$$

où ϕ est la fonction génératrice de N. Comme $\mathbb{E}[S_N] = -i(\phi\circ\psi_{X_1})'(0)$, il vient

$$\mathbb{E}[S_N] = -i\phi'(\psi_{X_1}(0))\psi'_{X_1}(0) = \phi'(1)(-i\psi'_{X_1}(0)) = \mathbb{E}[N]\mathbb{E}[X_1].$$

Comme $\mathbb{E}[S_N^2] = -(\phi\circ\psi_{X_1})''(0)$, il vient

$$\begin{aligned}\mathbb{E}[S_N^2] &= -(\phi'\circ\psi_{X_1}\psi'_{X_1})'(0)\\ &= -\phi''(1)\psi'_{X_1}(0)^2 - \phi'(1)\psi''_{X_1}(0)\\ &= \mathbb{E}[N(N-1)]\mathbb{E}[X_1]^2 + \mathbb{E}[N]\mathbb{E}[X_1^2]\\ &= \mathbb{E}[N]\,\mathrm{Var}(X_1) + \mathbb{E}[N^2]\mathbb{E}[X_1]^2.\end{aligned}$$

On retrouve ainsi les résultats de la question précédente.

▲

Exercice IV.4.

On note $\Im(z)$ la partie imaginaire de $z \in \mathbb{C}$.

1. Si X est symétrique alors $\Im(\psi_X(u)) = \frac{1}{2}(\psi_X(u) - \bar{\psi}_X(u)) = \frac{1}{2}(\mathbb{E}(\mathrm{e}^{iuX}) - \mathbb{E}(\mathrm{e}^{-iuX})) = 0$. Si $\Im(\psi_X(u)) = 0$ alors le calcul précédent montre que les fonctions caractéristiques de X et $-X$ coïncident, alors X et $-X$ sont égales en loi.

2. Si Y est de même loi et indépendant de X, alors $X - Y$ a la fonction caractéristique $\psi_X(u)\bar{\psi}_X(u) = |\psi_X(u)|^2$.
3. X est symétrique par rapport à $a \in \mathbb{R}$ si et seulement si $e^{-iau}\psi_X(u) \in \mathbb{R}$ pour tout $u \in \mathbb{R}$.

▲

Exercice IV.5.

1. En utilisant les fonctions caractéristiques, il vient par indépendance, pour $u \in \mathbb{R}$,

$$\phi_{X-Y}(u) = \phi_X(u)\phi_{-Y}(u) = \phi_X(u)\phi_Y(-u) = \mathrm{e}^{iu(\mu_X+\mu_Y)-\frac{u^2(\sigma_X^2+\sigma_Y^2)}{2}}.$$

   On en déduit que la loi de $X - Y$ est la loi gaussienne de paramètre $\mu = \mu_X + \mu_Y$ et de variance $\sigma^2 = \sigma_X^2 + \sigma_Y^2$.
2. On utilise la méthode de la fonction muette. Soit g une fonction continue bornée, on a, en notant

$$p(v) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-(v-\mu)^2/2\sigma^2},$$

   la densité de la loi $\mathcal{N}(\mu, \sigma^2)$,

$$\begin{aligned}\mathbb{E}[g(Z)] &= \mathbb{E}[g(|X-Y|)]\\ &= \int_{\mathbb{R}} g(|v|)p(v)dv\\ &= \int_{\mathbb{R}} g(v)p(v)\mathbf{1}_{\{v>0\}}dv + \int_{\mathbb{R}} g(-v)p(v)\mathbf{1}_{\{v<0\}}dv\\ &= \int_{\mathbb{R}} g(v)[p(v)+p(-v)]\mathbf{1}_{\{v>0\}}dv.\end{aligned}$$

   On en déduit que Z est une v.a. continue et la densité de sa loi est donnée par $[p(v) + p(-v)]\mathbf{1}_{\{v>0\}}$.
3. Remarquons que Z est égal en loi à $|\sigma G + \mu|$, où G est une v.a. de loi $\mathcal{N}(0, 1)$. En particulier, on a

$$\mathbb{E}[Z^2] = \mathbb{E}[(\sigma G + \mu)^2] = \sigma^2 + \mu^2,$$

   et

$$\begin{aligned}\mathbb{E}[Z] &= \mathbb{E}[|\sigma G + \mu|]\\ &= \mathbb{E}[(\sigma G + \mu)\mathbf{1}_{\{G>-\mu/\sigma\}}] - \mathbb{E}[(\sigma G + \mu)\mathbf{1}_{\{G<-\mu/\sigma\}}]\\ &= 2\frac{\sigma}{\sqrt{2\pi}}\,\mathrm{e}^{-\mu^2/2\sigma^2} + \mu(\Phi(\mu/\sigma) - \Phi(-\mu/\sigma))\\ &= \sqrt{\frac{2}{\pi}}\sigma\,\mathrm{e}^{-\mu^2/2\sigma^2} + \mu(2\Phi(\mu/\sigma) - 1),\end{aligned}$$

   où Φ est la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Enfin, la variance de Z est égale à $\mathbb{E}[Z^2] - \mathbb{E}[Z]^2$.

▲

Chapitre V

Théorèmes limites

V.1 Énoncés

Exercice V.1.
Soit $(X_n, n \in \mathbb{N}^*)$ une suite de v.a. indépendantes de loi exponentielle de paramètre $\lambda > 0$. Montrer que la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converge dans L^2 vers la moyenne $\frac{1}{\lambda}$. On pose $Y_n = \sqrt{n}\left(\bar{X}_n - \frac{1}{\lambda}\right)$. Montrer que $(Y_n, n \in \mathbb{N}^*)$ converge en loi vers une loi gaussienne dont on calculera les paramètres.

△

Exercice V.2.
Soit $(X_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires de loi exponentielle de paramètre λ_n. Étudier la convergence en loi dans les trois cas suivants :

1. $\lim_{n\to\infty} \lambda_n = \lambda \in]0, \infty[$,
2. $\lim_{n\to\infty} \lambda_n = +\infty$,
3. $\lim_{n\to\infty} \lambda_n = 0$.

△

Exercice V.3.
Soit $(U_n, n \geq 1)$ une suite de variables aléatoires indépendantes de loi uniforme sur l'intervalle $[0, \theta]$, où $\theta > 0$. On pose pour $n \geq 1$, $X_n = \max_{1\leq i\leq n} U_i$.

1. Montrer que $(X_n, n \geq 1)$ converge p.s. et déterminer sa limite. On pourra calculer $\mathbb{P}(|X_n - \theta| > \varepsilon)$ pour $\varepsilon > 0$.
2. Étudier la convergence en loi de la suite $(n(\theta - X_n), n \geq 1)$.

△

Exercice V.4.
Soient X une variable aléatoire intégrable à valeurs dans $\mathbb{N}$ et $(X_n, n \geq 1)$ une suite de variables aléatoires de même loi que X.

1. (a) Montrer que
$$\sum_{k=0}^{+\infty} \mathbb{P}(X > k) = \mathbb{E}[X].$$
   (b) Soit $m \in \mathbb{N}^*$. Montrer que la variable aléatoire $Y_m = \sum_{n=1}^{+\infty} \mathbf{1}_{\{\frac{X_n}{n} > \frac{1}{m}\}}$ est finie p.s.
   (c) En déduire que $\frac{X_n}{n}$ tend p.s. vers 0.
2. Montrer que $\frac{1}{n}\max(X_1, \ldots, X_n)$ tend en probabilité vers 0.

△

Exercice V.5.
Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de loi de Cauchy de paramètre $a > 0$. On note $S_n = \sum_{k=1}^n X_k$. Étudier les convergences en loi et en probabilité des suites :

1. $\left(\frac{S_n}{\sqrt{n}}, n \geq 1\right)$.
2. $\left(\frac{S_n}{n^2}, n \geq 1\right)$.
3. $\left(\frac{S_n}{n}, n \geq 1\right)$. On pourra déterminer la loi de $\frac{S_{2n}}{2n} - \frac{S_n}{n}$, et en déduire que la suite $\left(\frac{S_{2n}}{2n} - \frac{S_n}{n}, n \geq 1\right)$ ne converge pas en probabilité vers 0. On montrera alors que l'on ne peut avoir la convergence en probabilité de la suite $\left(\frac{S_n}{n}, n \geq 1\right)$.

△

Exercice V.6.
Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de loi de Poisson de paramètre 1. On note $S_n = \sum_{k=1}^n X_k$.

1. Montrer que
$$\frac{S_n - n}{\sqrt{n}} \xrightarrow[n\to\infty]{\text{en loi}} \mathcal{N}(0,1).$$
2. Montrer que
$$e^{-n}(1 + n + \frac{n^2}{2!} + \cdots + \frac{n^n}{n!}) \underset{n\to\infty}{\longrightarrow} \frac{1}{2}.$$

△

Exercice V.7.
Calcul de limites d'intégrales.

1. En considérant une suite de variables aléatoires indépendantes de loi uniforme sur $[0,1]$. Calculer à l'aide de la loi faible des grands nombres
$$\lim_{n\to\infty} \int_{[0,1]^n} f(\frac{x_1 + \cdots + x_n}{n})\, dx_1 \cdots dx_n,$$
   où f est une application continue bornée de $\mathbb{R}$ dans $\mathbb{R}$.

2. Soit $(Y_k, k \geq 1)$ une suite de variables aléatoires indépendantes de loi de Poisson $\mathcal{P}(\alpha)$, $\alpha > 0$. Déterminer la loi de $Z_n = \sum_{k=1}^{n} Y_k$.
3. Montrer que la suite des moyennes empiriques $(\bar{Y}_n = Z_n/n, n \geq 1)$ converge en loi vers une limite que l'on déterminera. Calculer, en s'inspirant de la question 1,

$$\lim_{n\to\infty} \sum_{k\geq 0} \mathrm{e}^{-\alpha n} \frac{(\alpha n)^k}{k!} f(\frac{k}{n}),$$

où $\alpha > 0$ et f est une application continue bornée de $\mathbb{R}$ dans $\mathbb{R}$.

△

Exercice V.8.
On effectue n séries de 400 tirages de pile ou face avec une pièce équilibrée. On observe les fréquences empiriques de pile $F_1, \ldots, F_n$ dans ces séries.

Quelle est (approximativement) la loi de probabilité du nombre N de ces fréquences $(F_i, 1 \leq i \leq n)$ qui ne vérifient pas la condition $0.45 < F_i < 0.55$, lorsque $n = 20$? Est-il plus probable que $N = 0$, que $N = 1$ ou que $N \geq 2$?

△

Exercice V.9.
Majoration du théorème des grandes déviations pour des v.a. de Bernoulli[1].

Soit $(X_n, n \in \mathbb{N}^*)$ une suite de v.a. i.i.d. de loi de Bernoulli de paramètre $p \in]0,1[$. On pose, pour tout $n \in \mathbb{N}^*$, $\bar{X}_n = \sum_{i=1}^{n} X_i/n$.

1. Calculer la transformée de Laplace de la v.a. X_1, i.e. l'application g_{X_1} de $\mathbb{R}$ dans $[0, +\infty[$ définie, pour tout $t \in \mathbb{R}$, par $g_{X_1}(t) = \mathbb{E}[\mathrm{e}^{tX_1}]$. Exprimer à l'aide de g_{X_1} la transformée de Laplace de $\bar{X}_n$.
2. Soit $\varepsilon > 0$, montrer que pour tout $t \geq 0$,

$$\mathbb{P}(\bar{X}_n \geq p + \varepsilon) \leq \exp\left[-t(p+\varepsilon) + n \ln g_{X_1}(t/n)\right].$$

En déduire

$$\mathbb{P}(\bar{X}_n \geq p + \varepsilon) \leq \exp\left[-n \sup_{s\geq 0}((p+\varepsilon)s - \ln g_{X_1}(s))\right],$$

et montrer que

$$\mathbb{P}(\bar{X}_n \leq p - \varepsilon) \leq \exp\left[-n \sup_{s\leq 0}((p-\varepsilon)s - \ln g_{X_1}(s))\right].$$

3. Vérifier que, pour tout $\varepsilon > 0$, assez petit, $\sup_{s\geq 0}((p+\varepsilon)s - \ln g_{X_1}(s)) \in]0, +\infty[$ et $\sup_{s\leq 0}((p-\varepsilon)s - \ln g_{X_1}(s)) \in]0, +\infty[$.

[1] Les théorèmes des grandes déviations étudient les équivalents logarithmiques des probabilités d'évènements rares. L'exemple typique d'évènement considéré est $\{|\bar{X}_n - \mathbb{E}[X_1]| > \varepsilon\}$, dont l'étude est due à Cramèr (1938). D'autre part, la théorie des grandes déviations est un sujet d'étude qui connaît un essor important depuis les années 1980.

4. Montrer la majoration du théorème des grandes déviations (pour les v.a. de Bernoulli) : Si $(X_n, n \in \mathbb{N}^*)$ est une suite de v.a. i.i.d. de loi de Bernoulli de paramètre $p \in]0,1[$, alors pour tout $\varepsilon > 0$ il existe une constante $C_{\varepsilon,p} > 0$ telle que, pour tout $n \in \mathbb{N}^*$,

$$\mathbb{P}\left(\left|\bar{X}_n - p\right| \geq \varepsilon\right) \leq 2\exp(-nC_{\varepsilon,p}).$$

△

Exercice V.10.
Le paradoxe de Saint-Petersbourg est d'abord un problème imaginé par Nicolas Bernoulli, qui obtint une solution partielle donnée par Daniel Bernoulli (1738) dans les Commentaires de l'Académie des sciences de Saint-Petersbourg (d'où son nom). Aujourd'hui encore, ce problème attire l'attention de certaines personnes en mathématiques et en économie[2].

Un casino propose le jeu suivant qui consiste à lancer plusieurs fois de suite une pièce équilibrée jusqu'à obtenir pile. Le joueur gagne 2^k francs si le premier pile a lieu au k-ième jet. La question est de savoir quel doit être le prix à payer pour participer à ce jeu.

Soit X_n le gain réalisé lors du n-ième jeu et $S_n = X_1 + \cdots + X_n$ le gain obtenu lors de n jeux successifs.

1. Peut-on appliquer la loi forte des grands nombres pour donner un prix équitable ?

Les fonctions d'utilité qui quantifient l'aversion au risque permettent de proposer des prix pour ce jeu. La suite de l'exercice est consacré à l'étude de la convergence de la suite $(S_n, n \geq 1)$ convenablement renormalisée[3]

2. On pose $S'_n = \sum_{k=1}^n X_k^n$, où pour $k \in \{1, \ldots, n\}$,

$$X_k^n = X_k 1_{\{X_k \leq n\log_2 n\}},$$

où $\log_2 n$ est le logarithme en base 2, i.e. $2^{\log_2 n} = n$. Après avoir vérifié que pour tout $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{S'_n}{n\log_2 n} - 1\right| > \varepsilon\right) \leq \mathbb{P}\left(\left|\frac{S'_n - \mathbb{E}[S'_n]}{n\log_2 n}\right| > \varepsilon/2\right) + \mathbb{P}\left(\left|\frac{\mathbb{E}[S'_n]}{n\log_2 n} - 1\right| > \varepsilon/2\right),$$

montrer que la suite $(\frac{S'_n}{n\log_2 n}, n \geq 1)$ converge en probabilité vers 1.

3. Calculer $\mathbb{P}(S_n \neq S'_n)$, et en déduire sa limite quand $n \to \infty$.

4. En déduire que la suite $(\frac{S_n}{n\log_2 n}, n \geq 1)$ converge en probabilité vers 1.

△

Exercice V.11.
Précision des sondages.

[2]Voir par exemple l'article suivant et les références citées : G. Székely and D. Richards, The St. Petersburg paradox and the crash of high-tech stocks in 2000, *Amer. Statist.* **58**, 225–231 (2004).

[3]Feller, *An introduction to probability theory and its applications*, Vol. 1. Third ed. (1968). Wiley & Sons.

1. À quelle précision peut prétendre un sondage sur deux candidats effectué sur un échantillon de 1 000 personnes ? Est-ce que ce résultat dépend de la taille de la population ?
2. En Floride, pour l'élection présidentielle américaine 2000, on compte 6 millions de votants. Sachant qu'il y a environ 4 000 voix d'écart, quel est le nombre de personnes qu'il faudrait interroger dans un sondage pour savoir avec 95% de chance qui est le vainqueur ?

△

Exercice V.12.
On souhaite réaliser un sondage[4] sur le taux d'abstention p au sein d'une population totale de taille N. On note $(y_1, \dots, y_N)$ les choix de la population (0 pour le vote, 1 pour l'abstention) et $p = \frac{1}{N}\sum_{i=1}^N y_i$ le taux d'abstention réel. Pour estimer p, on décide d'interroger n personnes choisies au hasard parmi les N. Soit $(Y_1, \dots, Y_n)$ l'échantillon des réponses obtenues. On définit $\hat{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ l'estimateur de p obtenu au cours du sondage. On approche le tirage sans remise par un tirage avec remise, de sorte que les variables aléatoires $(Y_1, \dots, Y_n)$ sont indépendantes de loi de Bernoulli de paramètre $p \in]0, 1[$.

1. Montrer que $(\hat{Y}_n(1-\hat{Y}_n), n \geq 1)$ converge p.s. vers $\sigma^2 = \mathrm{Var}(Y_1)$.
2. Vérifier que $\sigma^2 = \frac{1}{N}\sum_{i=1}^N (y_i - p)^2 = p(1-p)$.
3. Donner un intervalle de confiance asymptotique de niveau $1-\alpha$ pour p utilisant $\hat{Y}_n$. Vérifier que la largeur de l'intervalle de confiance est de l'ordre de $2\phi_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{Y}_n)}$, où $\phi_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi gaussienne centrée réduite.

Pour améliorer le sondage, on répartit les électeurs en H groupes homogènes (par exemple le niveau d'études). Il s'agit d'une technique dite de stratification. On note N_h le cardinal de la strate h pour $h \in \{1, \dots, H\}$. À l'intérieur de la strate h on réalise le sondage auprès de n_h personnes et on note $\hat{Y}^h_{n_h}$ l'estimateur du taux d'abstention p_h. On note $\sigma_h^2 = p_h(1-p_h)$. On définit l'estimateur de Horvitz Thomsom par

$$\hat{Y}^{st}_n = \sum_{h=1}^H \frac{N_h}{N}\hat{Y}^h_{n_h}.$$

4. Donner l'expression de la variance de l'estimateur de Horvitz Thomsom. (On peut montrer que $[\hat{Y}^{st}_n \pm \phi_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{Y}^{st}_n)}]$ est un intervalle de confiance asymptotique à $1-\alpha$ de p et que l'on peut remplacer $\sqrt{\mathrm{Var}(\hat{Y}^{st}_n)}$ par $\sum_{h=1}^H \frac{N_h}{N}\hat{Y}^h_n(1-\hat{Y}^h_n)$.)
5. Monter que la variance totale σ^2 peut s'écrire $\sigma^2 = \sum_{h=1}^H \frac{N_h}{N}\sigma_h^2 + \sum_{h=1}^H \frac{N_h}{N}(p_h - p)^2$. Le premier terme de la somme représente la variance intra-strates et le second la variance inter-strates.
6. On introduit $\eta^2 = \dfrac{\text{variance inter-strates}}{\sigma^2}$. On suppose que le rapport $\frac{n_h}{N_h}$ est constant. Ce choix de répartition est connu sous le nom d'allocation proportionnelle. Exprimez $\mathrm{Var}(\hat{Y}^{st}_n)$ en fonction σ^2, n et η. À quelle condition a-t-on $\mathrm{Var}(\hat{Y}^{st}_n) \ll \mathrm{Var}(\hat{Y}_n)$? Conclusion ?

[4]Pour un exposé détaillé sur les sondages, on peut consulter la monographie de W. Cochran, *Sampling techniques*. Thrid ed. (1977). Wiley & Sons.

7. Application numérique. Il y a en France $N = 42.\,10^6$ votants et on désire connaître le taux d'abstention. Pour cela on effectue un sondage par strates sur $n = 8023$ personnes. On considère les trois strates suivantes classées par niveau d'études : inférieur au Bac, égal au Bac et Bac +2 ou mieux. En fonction des données a posteriori suivantes (France 2002) calculer le gain dû à la stratification.

niveau d'études	$N_h/10^6$	p_h
inférieur au Bac	25	33%
Bac	6	31%
Bac+2 ou plus	11	21%

△

Exercice V.13.
Répartition des bombes sur Londres lors de la Seconde Guerre Mondiale.

1. Soit $(Y_m, m \in \mathbb{N})$ une suite de variables aléatoires de loi binomiale de paramètres (m, p_m). On suppose que $m \to \infty$ et $\lim_{m\to\infty} mp_m = \theta \in (0, \infty)$. Montrer que la suite $(Y_m, m \in \mathbb{N})$ converge en loi vers la loi de Poisson de paramètre θ.

   Cette approximation est utile quand m est grand car le calcul numérique des coefficients binômiaux C_k^m est peu efficace.

2. Les données suivantes représentent le nombre de bombes qui ont touché le sud de Londres pendant la Seconde Guerre Mondiale[5] [6]. Le sud de Londres a été divisé en $N = 576$ domaines de taille $t = \frac{1}{4}$ km^2 chacun. On a compté les nombres N_k de domaines qui ont été touchés exactement k fois :

k	0	1	2	3	4	5+
N_k	229	211	93	35	7	1

   Faire un modèle simple qui représente cette expérience. Le nombre total d'impacts dans le sud de Londres est $T = \sum_{k\geq 1} kN_k = 537$. Calculer les probabilités théoriques pour qu'un domaine contienne exactement k impacts. Comparer avec les fréquences empiriques N_k/N ci-dessus.

△

Exercice V.14.
L'objectif de cet exercice est de démontrer le théorème suivant du à Borel (1909) : "Tout nombre réel choisi au hasard et uniformémént dans $[0, 1]$ est presque sûrement absolument normal".

Soit $x \in [0, 1]$, et considérons son écriture en base $b \geq 2$:

$$x = \sum_{n=1}^{\infty} \frac{x_n}{b^n},$$

avec $x_n \in \{0, \ldots, b-1\}$. Cette écriture n'est pas unique seulement pour les fractions rationnelles de la forme $x = a/b^n$ et $a \in \{1, \ldots, b^n - 1\}$. En effet, dans ce cas deux représentations sont possibles : l'une telle que $x_k = 0$ pour $k \geq n+1$ et l'autre telle que $x_k = b-1$ pour

[5] Clarke R. D., An application of the Poisson distribution, *J. of Institute of Actuaries* (1946), **72**, p. 481.
[6] Feller, W. *An Introduction to Probability Theory and Applications*, Wiley, 3rd edition, vol. 1, pp. 160-161

$k \geq n+1$. On dit que x est simplement normal en base b si et seulement si pour tout $i \in \{0,\dots,b-1\}$, $\lim_{n\to\infty} \frac{1}{n}\text{Card}\,\{1 \leq k \leq n; x_k = i\}$ existe et vaut $1/b$. Cela revient à dire que les fréquences d'apparition de i dans le développement de x en base b sont uniformes. Remarquons que les fractions rationnelles ne sont pas simplement normales, quelle que soit leur représentation.

On dit que x est normal en base b si et seulement si il est simplement normal en base b^r pour tout $r \in \mathbb{N}^*$. Remarquons qu'un nombre est normal en base b si et seulement si pour tout $r \in \mathbb{N}^*$, la fréquence d'apparition d'une séquence donnée de longueur r, dans le développement de x est uniforme (et vaut donc $1/b^r$) i.e. pour tout $i \in \{0,\dots,b-1\}^r$,

$$\lim_{n\to\infty} \frac{1}{n}\text{Card}\,\{0 \leq k \leq n; (x_{rk+1},\dots,x_{r(k+1)}) = i\} = \frac{1}{b^r}\,.$$

Il est bien connu que le nombre de Champernowne[7] dont la partie décimale est la suite consécutive des entiers (0,12345678910111213...) est normal en base 10.

On dit que x est absolument normal si et seulement si il est normal en toute base $b \geq 2$.

1. Soit X une variable aléatoire de loi uniforme sur $[0,1]$. Quelle est la loi de X_n, le n-ième chiffre du développement de X en base b ?
2. Montrer que les variables aléatoires $X_1,\dots,X_n$ sont indépendantes. En déduire que les variables aléatoires $(X_n, n \geq 1)$ sont indépendantes.
3. En utilisant la loi forte des grands nombres, montrer que X est p.s. simplement normal en base b.
4. Montrer que X est p.s. normal en base b, puis qu'il est p.s. absolument normal.

Bien que presque tous les réels soient absolument normaux, il est très difficile de montrer qu'un réel donné est absolument normal. On ne sait toujours pas si des nombres tels que π, e, $\sqrt{2}$ ou $\ln 2$ sont absolument normaux, ni même normaux en base 10 (cf. *Pour la Science*, janvier 1999).

△

Exercice V.15.
Théorème de Weierstrass (1885) : "Toute fonction continue sur un intervalle borné est limite uniforme d'une suite de polynômes".

Cet exercice s'inspire de la démonstration de Bernstein du théorème de Weierstrass. Soit $(X_k, k \geq 1)$ une suite de variables aléatoires indépendantes de loi de Bernoulli de paramètre $x \in [0,1]$. On considère la moyenne empirique $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$ (de loi binomiale de paramètre (n,x)). Soit $h : [0,1] \to \mathbb{R}$ une fonction continue. On pose $\Delta_n = \{\left|\bar{X}_n - x\right| > \delta\}$.

1. Montrer que $\mathbb{P}(\Delta_n) \leq \delta^{-2}\mathbb{E}[(\bar{X}_n - x)^2]$. Majorer $\mathbb{P}(\Delta_n)$ indépendanment de $x \in [0,1]$.
2. Déterminer $\lim_{n\to\infty} \sup_{x\in[0,1]} \left|h(x) - \mathbb{E}[h(\bar{X}_n)]\right|$, en écrivant

$$\left|h(x) - h(\bar{X}_n)\right| = \left|h(x) - h(\bar{X}_n)\right| \mathbf{1}_{\Delta_n} + \left|h(x) - h(\bar{X}_n)\right| \mathbf{1}_{\Delta_n^c}\,.$$

3. Quelle est la loi de $n\bar{X}_n$?

[7] Champernowne D. G., The construction of decimals normal in the scale of ten, *J. London MAth. Soc.* (1933), **8** pp. 254-260

4. En déduire que

$$\lim_{n\to\infty} \sup_{x\in[0,1]} \left| h(x) - \sum_{k=0}^{n} C_n^k h(k/n) x^k (1-x)^{n-k} \right| = 0.$$

5. Soit $f : \mathbb{R}^+ \to \mathbb{R}$ continue bornée. Montrer, en s'inspirant des questions précédentes, que pour tout $x \in \mathbb{R}^+$,

$$\lim_{n\to\infty} \left| f(x) - \sum_{k=0}^{\infty} \mathrm{e}^{-nx} \frac{(nx)^k}{k!} f(k/n) \right| = 0.$$

Si l'on suppose f uniformément continue, la convergence ci-dessus est-elle uniforme en x ? (Prendre par exemple $f(x) = \cos(x)$ pour $x_n = 2\pi n$.)

△

Exercice V.16.
Contamination au mercure.

1. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de même loi. On suppose qu'il existe deux réels $\alpha > 0, \lambda > 0$ tels qu'au voisinage de $+\infty$,

$$\mathbb{P}(X_1 > x) \sim \frac{\alpha}{x^\lambda}.$$

Montrer que

$$Z_n = n^{-\frac{1}{\lambda}} \max(X_1, \ldots, X_n)$$

converge en loi vers la loi de Fréchet. Si la loi de Y est la loi de Fréchet, sa fonction de répartition est donnée pour $y > 0$ par $\mathbb{P}(Y \leq y) = \exp(-\alpha y^{-\lambda})$.

2. Le mercure, métal lourd, est présent dans peu d'aliments. On le trouve essentiellement dans les produits de la mer. L'Organisation Mondiale de la Santé fixe la dose journalière admissible en mercure à 0.71 μg par jour et par kilo de poids corporel. Des études statistiques[8] donnent la forme de la queue de distribution empirique de la contamination globale annuelle en gramme de mercure pour un individu de 70 kg :

$$\mathbb{P}(X > x) = \frac{\alpha}{x^\lambda} \qquad \text{pour } x \text{ assez grand,}$$

avec $\alpha = 3.54\, 10^{-9}$ et $\lambda = 2.58$.

Seriez-vous étonné(e) qu'au moins une personne soit exposée à ce risque sanitaire en France ? Dans le 15ème arrondissement de Paris ? Dans une promo de l'ENSTA ? À partir de quelle valeur de n pouvez-vous affirmer, avec seulement 5% de chances de vous tromper : "Parmi ces n personnes, au moins une a un niveau de mercure trop élevé" ?

△

[8] *Evaluation des risques d'exposition à un contaminant alimentaire : quelques outils statistiques*, P. Bertail, Laboratoire de statistique, CREST, août 2002, disponible à l'adresse :
`www.crest.fr/doctravail/document/2002-41.pdf`

V.2 Corrections

Exercice V.1.

Remarquons que $\mathbb{E}[\bar{X}_n] = \dfrac{1}{\lambda}$. Donc on a $\mathbb{E}\left[(\bar{X}_n - \frac{1}{\lambda})^2\right] = \mathrm{Var}(\bar{X}_n)$. En utilisant l'indépendance des variables aléatoires, il vient $\mathrm{Var}(\bar{X}_n) = 1/n\lambda^2$. On en déduit donc la convergence dans L^2.

En utilisant l'indépendance des v.a. X_i, on a

$$\psi_{Y_n}(u) = \left(1 - \frac{iu}{\sqrt{n}\,\lambda}\right)^{-n} \mathrm{e}^{-iu\sqrt{n}/\lambda}\,.$$

On note $\rho_n \geq 0$ et $\theta_n \in]-\frac{\pi}{2}, \frac{\pi}{2}[$, le module et l'argument de $1 - \frac{iu}{\sqrt{n}\,\lambda}$. Ainsi on a $\rho_n^2 = 1 + \frac{u^2}{n\lambda^2}$ et $\tan\theta_n = \frac{-u}{\sqrt{n}\,\lambda}$. Donc

$$\psi_{Y_n}(u) = \mathrm{e}^{-\frac{iu\sqrt{n}}{\lambda} - in\theta_n - n\log\rho_n}\,.$$

Remarquons que $\lim_{n\to\infty} n\log\rho_n = u^2/2\lambda^2$ et que $\tan\theta_n = -u/\sqrt{n}\lambda$ implique que $\theta_n = \frac{-u}{\sqrt{n}\,\lambda} + O(n^{-3/2})$. Par passage à la limite, on obtient

$$\psi_{Y_n}(u) = \mathrm{e}^{-\frac{u^2}{2\lambda^2} + O(n^{-1/2})} \underset{n\to\infty}{\longrightarrow} \mathrm{e}^{-\frac{u^2}{2\lambda^2}}\,.$$

On reconnaît pour la limite la fonction caractéristique de la loi gaussienne $\mathcal{N}\left(0, \frac{1}{\lambda^2}\right)$. On en déduit que la suite $(Y_n, n \in \mathbb{N}^*)$ converge en loi vers $\mathcal{N}\left(0, \frac{1}{\lambda^2}\right)$.

▲

Exercice V.2.

1. Soit g continue bornée. On a $\mathbb{E}[g(X_n)] = \int_0^\infty \lambda_n \mathrm{e}^{-\lambda_n x}\, g(x)\, dx$. Il existe $n_0 \in \mathbb{N}^*$, et $0 < \lambda_- < \lambda_+ < \infty$ tels que pour tout $n \geq n_0$, on a $\lambda_n \in [\lambda_-, \lambda_+]$. On a alors $\left|\lambda_n \mathrm{e}^{-\lambda_n x}\, g(x)\right| \leq \|g\|_\infty \lambda_+ \mathrm{e}^{-\lambda_- x} = h(x)$. La fonction h est intégrable sur $[0, \infty[$. Remarquons que l'on a aussi $\lim_{n\to\infty} \lambda_n \mathrm{e}^{-\lambda_n x}\, g(x) = \lambda \mathrm{e}^{-\lambda x}\, g(x)$. On déduit du théorème de convergence dominée que

$$\mathbb{E}[g(X_n)] \underset{n\to\infty}{\longrightarrow} \int_0^\infty \lambda \mathrm{e}^{-\lambda x}\, g(x)\, dx.$$

   Donc la suite $(X_n, n \in \mathbb{N}^*)$ converge en loi vers la loi exponentielle de paramètre λ.

2. Soit g continue bornée. On a $\mathbb{E}[g(X_n)] = \int_0^\infty \mathrm{e}^{-x}\, g(x/\lambda_n)\, dx$. On a également la majoration $|\mathrm{e}^{-x}\, g(x/\lambda_n)| \leq \|g\|_\infty \mathrm{e}^{-x} = h(x)$, et la fonction h est intégrable sur $[0, \infty[$. Comme la fonction g est continue, on a $\lim_{n\to\infty} g(x/\lambda_n) = g(0)$. Par convergence dominée, il vient

$$\mathbb{E}[g(X_n)] \underset{n\to\infty}{\longrightarrow} g(0) = \mathbb{E}[g(X)],$$

   où $X = 0$ p.s. Donc la suite $(X_n, n \in \mathbb{N}^*)$ converge en loi vers 0.

3. Les fonctions $x \mapsto \mathrm{e}^{iux}$ sont continues bornées. Si la suite $(X_n, n \in \mathbb{N}^*)$ convergeait en loi vers une v.a. X, alors les fonctions caractéristiques $\psi_{X_n}(u)$ convergeraient vers $\psi_X(u)$ pour tout $u \in \mathbb{R}$. On a

$$\mathbb{E}\left[\mathrm{e}^{iuX_n}\right] = \frac{\lambda_n}{\lambda_n - iu} \underset{n\to\infty}{\longrightarrow} \mathbf{1}_{\{u=0\}}.$$

Si on a la convergence en loi alors $\psi_X(u) = \mathbf{1}_{\{u=0\}}$. Or la fonction $u \mapsto \mathbf{1}_{\{u=0\}}$ n'est pas continue en 0. Par contraposée, ce n'est donc pas la fonction caractéristique d'une v.a. La suite $(X_n, n \in \mathbb{N}^*)$ ne converge donc pas en loi.

▲

Exercice V.3.

1. La suite $(X_n, n \geq 1)$ est croissante et bornée p.s. par θ. Elle converge donc p.s. vers une limite X. Soit $\varepsilon \in]0, \theta]$, on a

$$\mathbb{P}(|X_n - \theta| > \varepsilon) = \mathbb{P}(X_n > \theta + \varepsilon) + \mathbb{P}(X_n < \theta - \varepsilon) = 0 + (\frac{\theta - \varepsilon}{\theta})^n.$$

Donc on a

$$\lim_{n\to\infty} \mathbb{P}(|X_n - \theta| > \varepsilon) = 0,$$

i.e. la suite $(X_n, n \geq 1)$ converge en probabilité vers θ. Comme la suite converge aussi en probabilité vers X, par unicité de la limite, on en déduit que $X = \theta$ p.s.

2. On étudie la fonction de répartition sur $\mathbb{R}^+$, car $n(\theta - X) \geq 0$. Soit $a > 0$.

$$\mathbb{P}(n(\theta - X_n) \leq a) = \mathbb{P}(X_n > \theta - \frac{a}{n}) = 1 - (1 - \frac{a}{\theta n})^n \xrightarrow[n\to\infty]{} 1 - \mathrm{e}^{-\frac{a}{\theta}}.$$

On reconnaît la fonction de répartition de la loi exponentielle de paramètre $1/\theta$. La suite $(n(\theta - X_n), n \geq 1)$ converge donc en loi vers la loi exponentielle de paramètre $1/\theta$.

▲

Exercice V.4.

1. (a) Il suffit d'appliquer l'espérance à l'égalité $X = \sum_{k=0}^{+\infty} \mathbf{1}_{\{X>k\}}$. L'intervertion de l'espérance et de la somme est licite car $\mathbf{1}_{\{X>k\}}$ est positif pour tout k.

   (b) $\mathbf{1}_{\{\frac{X_n}{n} > \frac{1}{m}\}}$ est positif pour tout n donc on peut intervertir l'espérance et la somme. En appliquant le résultat de la question précédente à la variable aléatoire mX intégrable et à valeurs dans $\mathbb{N}$, il vient

$$\begin{aligned}
\mathbb{E}[Y_m] &= \sum_{n=1}^{+\infty} \mathbb{P}\left(\frac{X_n}{n} > \frac{1}{m}\right) \\
&= \sum_{n=1}^{+\infty} \mathbb{P}\left(mX_n > n\right) \\
&= \sum_{n=1}^{+\infty} \mathbb{P}\left(mX > n\right) \\
&\leq \sum_{n=0}^{+\infty} \mathbb{P}\left(mX > n\right) \\
&= \mathbb{E}[mX] < +\infty.
\end{aligned}$$

La variable aléatoire Y_m est d'espérance finie, elle est donc finie p.s.

(c) Soit $A_m = \{Y_m < +\infty\}$. C'est un évènement de probabilité 1. Par conséquent, l'évènement $A = \bigcap_{m\geq 1} A_m$ est lui aussi de probabilité 1. Montrons que pour tout $\omega \in A$, $\frac{X_n(\omega)}{n}$ tend vers 0.
Soient $\omega \in A$, $\varepsilon > 0$ et $m \in \mathbb{N}^*$ tel que $\frac{1}{m} < \varepsilon$. Comme $\omega \in A_m$, $Y_m(\omega) = \sum_{n=1}^{+\infty} \mathbf{1}_{\{\frac{X_n(\omega)}{n} > \frac{1}{m}\}}$ est fini, i.e.

$$\exists N(\omega) \in \mathbb{N}^*, \quad \forall n \geq N(\omega), \quad 0 \leq \frac{X_n(\omega)}{n} \leq \frac{1}{m} < \varepsilon,$$

ce qu'il fallait démontrer.

2. Soit $\varepsilon > 0$. Comme

$$\left\{\frac{1}{n}\max(X_1,\ldots,X_n) > \varepsilon\right\} = \bigcup_{k=1}^{n} \{X_k > n\varepsilon\},$$

on a

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\max(X_1,\ldots,X_n) > \varepsilon\right) &\leq \sum_{k=1}^{n} \mathbb{P}(X_k > n\varepsilon) \\ &= n\mathbb{P}(X > n\varepsilon) \\ &\leq \frac{\mathbb{E}[X\mathbf{1}_{\{X>n\varepsilon\}}]}{\varepsilon},\end{aligned}$$

la dernière inégalité provenant du fait que

$$\mathbb{E}[X\mathbf{1}_{\{X>n\varepsilon\}}] \geq \mathbb{E}[n\varepsilon\mathbf{1}_{\{X>n\varepsilon\}}] = n\varepsilon\mathbb{P}(X > n\varepsilon).$$

La suite de variables $(X\mathbf{1}_{\{X>n\varepsilon\}}, n \geq 1)$ est dominée par la variable aléatoire intégrable X donc, d'après le théorème de convergence dominée,

$$\lim_{n\to+\infty} \mathbb{E}[X\mathbf{1}_{\{X>n\varepsilon\}}] = \mathbb{E}\left[\lim_{n\to+\infty} X\mathbf{1}_{\{X>n\varepsilon\}}\right] = 0.$$

Ainsi, pour tout $\varepsilon > 0$,

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{1}{n}\max(X_1,\ldots,X_n) > \varepsilon\right) = 0.$$

Autrement dit, la variable aléatoire $\frac{1}{n}\max(X_1,\ldots,X_n)$ tend en probabilité vers 0.

▲

Exercice V.5.

On rappelle que $\psi_{X_n}(u) = \mathrm{e}^{-a|u|}$.

1. On a

$$\psi_{S_n/\sqrt{n}}(u) = \psi_{S_n}(u/\sqrt{n}) = \prod_{k=1}^{n} \psi_{X_k}(u) = (\mathrm{e}^{-a|u|/\sqrt{n}})^n = \mathrm{e}^{-a\sqrt{n}|u|},$$

où l'on a utilisé l'indépendance des variables aléatoires pour la 2-ième égalité. Donc on en déduit que

$$\psi_{S_n/\sqrt{n}}(u) \xrightarrow[n\to\infty]{} \mathbf{1}_{\{u=0\}}.$$

La limite est une fonction discontinue en 0. Ce n'est donc pas une fonction caractéristique. La suite ne converge donc pas en loi. Elle ne converge pas non plus en probabilité.

2. On a
$$\psi_{S_n/n^2}(u) = \psi_{S_n}(u/n^2) = (\mathrm{e}^{-a\,|u|\,/n^2})^n = \mathrm{e}^{-a\,|u|\,/n}\,.$$
Donc on en déduit que
$$\psi_{S_n/n^2}(u) \xrightarrow[n\to\infty]{} 1.$$
La suite converge en loi vers la variable aléatoire constante égale à 0. Pour la convergence en probabilité, remarquons tout d'abord que S_n/n^2 suit la loi de Cauchy de paramètre a/n. Donc on a
$$\begin{aligned}\mathbb{P}(|S_n/n^2| \geq \varepsilon) &= 1 - \mathbb{P}(|S_n/n^2| < \varepsilon) = 1 - \int_{-\varepsilon}^{\varepsilon} \frac{a/n}{\pi}\frac{1}{x^2 + (a/n)^2}dx \\ &= 1 - \frac{2}{\pi}\arctan(\frac{n\varepsilon}{a}).\end{aligned}$$
Donc on a
$$\mathbb{P}(|S_n/n^2| > \varepsilon) \xrightarrow[n\to\infty]{} 0,$$
i.e. la suite $(S_n/n^2, n \geq 1)$ converge en probabilité vers 0.

3. On a
$$\psi_{S_n/n}(u) = \psi_{S_n}(u/n) = (\mathrm{e}^{-a\,|u|\,/n})^n = \mathrm{e}^{-a\,|u|}\,.$$
On reconnaît la fonction caractéristique d'une loi de Cauchy de paramètre a. La suite $(S_n/n, n \geq 1)$ est donc constante en loi. Montrons maintenant que la suite $(\frac{S_n}{n}, n \geq 1)$ ne converge pas en probabilité. On a
$$\frac{S_{2n}}{2n} - \frac{S_n}{n} = \frac{X_{n+1} + \cdots + X_{2n}}{2n} - \frac{X_1 + \cdots + X_n}{2n}.$$
On a donc par indépendance, puis en utilisant le fait que $\frac{X_{n+1}+\cdots+X_{2n}}{2n}$ et $\frac{X_1+\cdots+X_n}{2n}$ ont même loi que $S_n/2n$, que
$$\psi_{(\frac{S_{2n}}{2n} - \frac{S_n}{n})}(u) = \psi_{\frac{X_{n+1}+\cdots+X_{2n}}{2n}}(u)\psi_{\frac{X_1+\cdots+X_n}{2n}}(u) = \psi_{S_n/2n}(u)^2 = \mathrm{e}^{-a\,|u|}\,.$$
Donc pour tout n, $(\frac{S_{2n}}{2n} - \frac{S_n}{n})$ est une variable aléatoire de Cauchy de paramètre a. On en déduit que la suite $(\frac{S_{2n}}{2n} - \frac{S_n}{n}, n \geq 1)$ ne converge pas en probabilité vers 0. Raisonnons par l'absurde. Si $(\frac{S_n}{n}, n \geq 1)$ convergeait en probabilité vers une limite X, on aurait alors
$$\left\{\left|\frac{S_{2n}}{2n} - \frac{S_n}{n}\right| \geq \varepsilon\right\} \subset \left\{\left|\frac{S_{2n}}{2n} - X\right| \geq \varepsilon/2\right\} \cup \left\{\left|\frac{S_n}{n} - X\right| \geq \varepsilon/2\right\}.$$
En particulier on aurait
$$\mathbb{P}\left(\left|\frac{S_{2n}}{2n} - \frac{S_n}{n}\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\left|\frac{S_{2n}}{2n} - X\right| \geq \varepsilon/2\right) + \mathbb{P}\left(\left|\frac{S_n}{n} - X\right| \geq \varepsilon/2\right),$$

et par passage à la limite

$$\mathbb{P}\left(\left|\frac{S_{2n}}{2n}-\frac{S_n}{n}\right|\geq\varepsilon\right)\xrightarrow[n\to\infty]{}0,$$

pour tout $\varepsilon>0$. La suite $(\frac{S_{2n}}{2n}-\frac{S_n}{n}, n\geq 1)$ converge alors en probabilité, et donc en loi, vers 0. Ceci est absurde. La suite $(\frac{S_n}{n}, n\geq 1)$ ne converge donc pas en probabilité.

▲

Exercice V.6.

1. Notons que S_n est une variable aléatoire de loi de Poisson de paramètre n. Les variables aléatoires X_n sont indépendantes, de même loi et de carré intégrable, et $\mathbb{E}[X_n]=1$, $\mathrm{Var}(X_n)=1$. On a donc, par le théorème central limite, que $\frac{S_n-n}{\sqrt{n}}$ converge en loi vers la loi gaussienne $\mathcal{N}(0,1)$.
2. Soit $F_n(x)$ la fonction de répartition de $\frac{S_n-n}{\sqrt{n}}$. Par le résultat de la question précédente on a que la suite $F_n(x)$ converge, quand n tends vers l'infini, vers la fonction de répartition de la loi gaussienne $\mathcal{N}(0,1)$, $F(x)$, pout tout x point de continuité de F. Comme F est une fonction continue, en particulier, on a

$$\mathbb{P}(\frac{S_n-n}{\sqrt{n}}\leq 0)\underset{n\to\infty}{\longrightarrow}F(0)=\frac{1}{2}.$$

D'autre part, le terme de gauche est égal à

$$\mathbb{P}(S_n\leq n)=\sum_{k=0}^{n}\mathbb{P}(S_n=k)=\sum_{k=0}^{n}e^{-n}\frac{n^k}{k!}.$$

▲

Exercice V.7.

1. Soit $(X_n, n\in\mathbb{N}^*)$ une suite de variables aléatoires indépendantes de loi uniforme sur $[0,1]$. On déduit de la loi faible des grands nombres que la moyenne empirique $\bar{X}_n=\frac{1}{n}\sum_{k=1}^{n}X_k$ converge en probabilité vers $\mathbb{E}[X_1]=1/2$. La convergence en probabilité implique la convergence en loi. On a donc, comme f est continue, que

$$\mathbb{E}[f(\frac{1}{n}\sum_{k=1}^{n}X_k)]\xrightarrow[n\to\infty]{}\mathbb{E}[f(1/2)]=f(1/2).$$

D'autre part on a

$$\mathbb{E}[f(\frac{1}{n}\sum_{k=1}^{n}X_k)]=\int_{[0,1]^n}f(\frac{x_1+\cdots+x_n}{n})\,dx_1\cdots dx_n.$$

On en déduit donc

$$\lim_{n\to\infty}\int_{[0,1]^n}f(\frac{x_1+\cdots+x_n}{n})\,dx_1\cdots dx_n=f(1/2).$$

Le résultat reste vrai si on suppose seulement que f est bornée et continue en $1/2$.

2. La loi de $\sum_{k=1}^n Y_k$ est la loi de Poisson $\mathcal{P}(\alpha n)$. Pour s'en convaincre, on regarde par exemple sa fonction génératrice. On a, en utilisant l'indépendance des variables aléatoires $Y_1, \ldots, Y_n$: pour tout $z \in [-1, 1]$,

$$\phi_{\sum_{k=1}^n Y_k}(z) = \prod_{k=1}^n \phi_{Y_k}(z) = \prod_{k=1}^n \mathrm{e}^{-\alpha(1-z)} = \mathrm{e}^{-n\alpha(1-z)} = \phi_{\mathcal{P}(n\alpha)}(z).$$

3. La loi faible des grands nombres assure que la moyenne empirique $\bar{Y}_n = \frac{1}{n}\sum_{k=1}^n Y_k$ converge en probabilité vers $\mathbb{E}[Y_1] = \alpha$. La convergence en probabilité implique la convergence en loi. On a donc, comme f est continue bornée, que

$$\mathbb{E}[f(\frac{1}{n}\sum_{k=1}^n Y_k)] \xrightarrow[n\to\infty]{} \mathbb{E}[f(\alpha)] = f(\alpha).$$

On en déduit que

$$\lim_{n\to\infty} \sum_{l\geq 0} \mathrm{e}^{-\alpha n} \frac{(\alpha n)^l}{l!} f(\frac{l}{n}) = f(\alpha).$$

Le résultat est vrai en fait dès que f est bornée et continue en α.

▲

Exercice V.8.

Soit S_m le nombre de fois où pile est apparu lors de m tirages. Ainsi $S_m = \sum_{i=1}^m X_i$, où $(X_i, i \geq 1)$ est un schéma de Bernouilli de paramètre $1/2$. En particulier $\mathbb{E}[X_i] = 1/2$ et $\mathrm{Var}(X_i) = 1/4$. La loi de S_m est une loi binomiale $\mathcal{B}(m, 1/2)$. Remarquons que si m est grand, alors, d'après le théorème central limite, la loi de $\dfrac{S_m - m\frac{1}{2}}{\frac{1}{2}\sqrt{m}}$ est asymptotiquement la loi gaussienne centrée.

Les fréquences empiriques de pile lors de n séries de m tirages sont égales en loi à S_m/m. Notons que le variables aléatoires F_1,...,F_n sont indepéndantes et de même loi. Par le théorème central limite on a que, pour chaque n,

$$p_m := \mathbb{P}(F_i \notin]0.45, 0.55[) = \mathbb{P}(\frac{S_m}{m} \notin]0.45, 0.55[) = \mathbb{P}\left(\left|\frac{S_m - m\frac{1}{2}}{\frac{1}{2}\sqrt{m}}\right| \geq \sqrt{m}\, 0.1\right).$$

En particulier, lors de 400 tirages et à l'aide des tables de la fonction de répartition de la loi normale $\mathcal{N}(0,1)$ on a que

$$p := \mathbb{P}(F_n \notin]0.45, 0.55[) = \mathbb{P}\left(\left|\frac{S_{400} - 400\frac{1}{2}}{\frac{1}{2}\sqrt{400}}\right| \geq 2\right) \simeq \mathbb{P}(|\mathcal{N}(0,1)| \geq 2) \simeq 0.05.$$

On considère maintenant une suite de variables aléatoires $(\xi_i, i \geq 1)$ telles que

$$\begin{aligned} &\mathbb{P}(\xi_i = 1) = p, \quad \text{si} \quad F_i \notin]0.45, 0.55[\\ &\mathbb{P}(\xi_i = 0) = 1 - p, \quad \text{sinon.} \end{aligned}$$

On a que le nombre des fréquences $(F_i, 1 \leq i \leq n)$ qui ne vérifient pas la condition $0.45 < F_i < 0.55$, lorsque $n = 20$, est égal à $N = \sum_{i=1}^{20} \xi_i$. La loi de N est donc la loi binomiale de

paramètre (n,p). Comme p est petit, la loi binomiale de paramètre (n,p) peut être approchée par la loi de Poisson de paramètre $np \simeq 1$ (on a la convergence en loi de la loi binomiale de paramètre (n,p) vers la loi de Poisson de paramètre θ quand $n \to \infty$, et $np \to \theta > 0$). En particulier, on a

$$\mathbb{P}(N=0) \simeq \mathrm{e}^{-1} \simeq \frac{1}{3}, \quad \mathbb{P}(N=1) \simeq \mathrm{e}^{-1} \simeq \frac{1}{3}, \quad \mathbb{P}(N \geq 2) \simeq 1 - 2\,\mathrm{e}^{-1} \simeq \frac{1}{3}\,.$$

Ainsi les évènements $\{N=0\}$, $\{N=1\}$ et $\{N \geq 2\}$ sont à peu près de même probabilité.

▲

Exercice V.9.

1. Un simple calcul d'espérance donne

$$g_{X_1}(t) = \mathbb{E}[\mathrm{e}^{tX_1}] = 1 - p + p\,\mathrm{e}^t\,.$$

Les variables étant i.i.d., il vient

$$g_{\bar{X}_n}(t) = g_{X_1}(t/n)^n.$$

2. On utilise l'inégalité de Markov, avec $t > 0$:

$$\begin{aligned} \mathbb{P}\left(\bar{X}_n \geq p + \varepsilon\right) &= \mathbb{P}\left(\exp(t(\bar{X}_n - p - \varepsilon)) \geq 1\right) \\ &\leq \mathbb{E}\left[\exp(t(\bar{X}_n - p - \varepsilon))\right]. \end{aligned}$$

La dernière égalité est vraie pour $t = 0$. D'où l'inégalité, pour tout $t \geq 0$

$$\mathbb{P}\left(\bar{X}_n \geq p + \varepsilon\right) \leq \exp\left[-n\left((p+\varepsilon)t/n - \ln g_{X_1}(t/n)\right)\right].$$

En prenant la borne inférieure du second membre sur les $t \geq 0$, cela donne

$$\begin{aligned} \mathbb{P}\left(\bar{X}_n \geq p + \varepsilon\right) &\leq \inf_{t\geq 0} \exp\left[-n\left((p+\varepsilon)t/n - \ln g_{X_1}(t/n)\right)\right] \\ &\leq \exp\left[-n \sup_{t\geq 0}\left((p+\varepsilon)t/n - \ln g_{X_1}(t/n)\right)\right] \\ &\leq \exp\left[-n \sup_{t\geq 0}\left((p+\varepsilon)t - \ln g_{X_1}(t)\right)\right]. \end{aligned}$$

On démontre la dernière inégalité, avec le même raisonnement et $t < 0$.

3. On note pour $\varepsilon > 0$ tel que $0 < p - \varepsilon < p + \varepsilon < 1$,

$$A_{\varepsilon,p} = \sup_{s\geq 0}((p+\varepsilon)s - \ln g_{X_1}(s)) \text{ et } B_{\varepsilon,p} = \sup_{s\leq 0}((p-\varepsilon)s - \ln g_{X_1}(s)).$$

La fonction $f(x) = (p+\varepsilon)x - \ln(1 - p + p\,\mathrm{e}^x)$ définie sur $[0, +\infty[$ est nulle en 0 et sa dérivée en ce point est $f'(0) = \varepsilon > 0$. Comme elle est continue, elle prend donc des valeurs strictement positives sur $]0, +\infty[$ et donc $A_{\varepsilon,p}$ est strictement positif. Comme $\lim_{x\to\infty} f(x) = -\infty$, $A_{\varepsilon,p}$ est fini.
Un raisonnement similaire assure que $B_{\varepsilon,p}$ est strictement positif et fini.

4. On remarque dans un premier temps que

$$\{\left|\bar{X}_n - p\right| \geq \varepsilon\} = \{\bar{X}_n \leq p - \varepsilon\} \cup \{\bar{X}_n \geq p + \varepsilon\}.$$

On obtient donc (avec les notations de la question précédente)

$$\begin{aligned}\mathbb{P}\left(\left|\bar{X}_n - p\right| \geq \varepsilon\right) &\leq \mathbb{P}\left(\bar{X}_n \leq p - \varepsilon\right) + \mathbb{P}\left(\bar{X}_n \geq p + \varepsilon\right) \\ &\leq \exp(-nB_{\varepsilon,p}) + \exp(-nA_{\varepsilon,p}) \\ &\leq 2\exp(-nC_{\varepsilon,p}),\end{aligned}$$

où on a posé $C_{\varepsilon,p} = \min(A_{\varepsilon,p}, B_{\varepsilon,p}) \in]0, +\infty[$.
On retrouve la loi faible des grands nombres en ayant de plus une majoration exponentielle de la vitesse de convergence. On peut aussi obtenir une minoration de la vitesse de convergence.

▲

Exercice V.10.

1. Les variables aléatoires X_n sont i.i.d. de loi

$$\mathbb{P}(X_1 = 2^k) = \frac{1}{2^k},\ k \geq 1.$$

On observe que $\mathbb{E}[X_1] = \infty$. Donc la loi forte des grands nombres n'est pas applicable.

2. On note $[x]$ la partie entière de x. On a

$$\mathbb{P}\left(\left|\frac{S'_n}{n\log_2 n} - 1\right| > \varepsilon\right) \leq \mathbb{P}\left(\left|\frac{S'_n - \mathbb{E}[S'_n]}{n\log_2 n}\right| > \varepsilon/2\right) + \mathbb{P}\left(\left|\frac{\mathbb{E}[S'_n]}{n\log_2 n} - 1\right| > \varepsilon/2\right),$$

où

$$\mathbb{E}[S'_n] = n\mathbb{E}[X_1 1_{\{X_1 \leq n\log_2 n\}}] = n[\log_2(n\log_2 n)].$$

On remarque que pour n suffisament grand, on a

$$\log_2(n) < [\log_2(n\log_2 n)] \leq \log_2(n) + \log_2(\log_2 n).$$

Ce qui implique que

$$\frac{\mathbb{E}[S'_n]}{n\log_2 n} \underset{n\to\infty}{\longrightarrow} 1.$$

Donc pour n assez grand (dépendant de ε), on a $\left|\frac{\mathbb{E}[S'_n]}{n\log_2 n} - 1\right| \leq \varepsilon/2$.
D'autre part, par l'inégalité de Tchebychev, on a

$$\mathbb{P}\left(\left|\frac{S'_n - \mathbb{E}[S'_n]}{n\log_2 n}\right| > \varepsilon/2\right) \leq \frac{4\mathrm{Var}(S'_n)}{(\varepsilon n\log_2 n)^2}.$$

Par l'indépendance des X_n, on a que

$$\mathrm{Var}(S'_n) = n\mathrm{Var}(X_1 1_{\{X_1 \leq n\log_2 n\}}) \leq n\mathbb{E}[X_1^2 1_{\{X_1 \leq n\log_2 n\}}].$$

De plus,

$$\begin{aligned}\mathbb{E}[X_1^2 1_{\{X_1\le n\log_2 n\}}] = \sum_{k:2^k\le n\log_2 n} 2^k &\le \sum_{k\le[\log_2(n\log_2 n)]} 2^k\\ &\le 2^{[\log_2(n\log_2 n)]+1}\\ &\le 2n2^{\log_2\log_2 n}.\end{aligned}$$

D'où, on conclut que

$$\mathbb{P}\left(\left|\frac{S'_n-\mathbb{E}[S'_n]}{n\log_2 n}\right|>\varepsilon/2\right)\le\frac{4\mathrm{Var}(S'_n)}{(\varepsilon\log_2 n)^2}\le\frac{8}{\varepsilon^2\log_2 n}\underset{n\to\infty}{\longrightarrow}0,$$

pout tout $\varepsilon>0$.

On en déduit donc que $\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{S'_n}{n\log_2 n}-1\right|>\varepsilon\right)=0$. Donc la suite $(\frac{S'_n}{n\log_2 n}, n\ge 1)$ converge en probabilité vers 1.

3. On observe que

$$\mathbb{P}(X_1>2^m)=\sum_{n=m+1}^{\infty}2^{-n}=2^{-m}.$$

D'autre part

$$\mathbb{P}(X_1>n\log_2 n)=\sum_{k:2^k>n\log_2 n}2^{-k}\le\sum_{k>[\log_2(n\log_2 n)]}2^{-k}=2^{-[\log_2(n\log_2 n)]},$$

où $[x]$ denote la partie entière de x. Or

$$[\log_2(n\log_2 n)]=[\log_2 n+\log_2\log_2 n]\ge\log_2 n+\log_2\log_2 n-1.$$

D'où, on a

$$\begin{aligned}\mathbb{P}(S_n\neq S'_n)=\mathbb{P}(\cup_{k=1}^n\{X_k\neq X_k^n\})&\le\sum_{k=1}^n\mathbb{P}(X_k\neq X_k^n)\\ &=n\mathbb{P}(X_1>n\log_2 n)\le\frac{2}{\log_2 n}\underset{n\to\infty}{\longrightarrow}0.\end{aligned}$$

4. Soit $\varepsilon>0$. Comme

$$\begin{aligned}\left\{\left|\frac{S_n}{n\log_2 n}-1\right|>\varepsilon\right\}=\left(\left\{\left|\frac{S'_n}{n\log_2 n}-1\right|>\varepsilon\right\}\cap\{S_n=S'_n\}\right)\\ \cup\left(\left\{\left|\frac{S_n}{n\log_2 n}-1\right|>\varepsilon\right\}\cap\{S_n\neq S'_n\}\right),\end{aligned}$$

on a

$$\mathbb{P}\left(\left|\frac{S_n}{n\log_2 n}-1\right|>\varepsilon\right)\le\mathbb{P}\left(\left|\frac{S'_n}{n\log_2 n}-1\right|>\varepsilon\right)+\mathbb{P}(S_n\neq S'_n).$$

On déduit des questions précédentes que la suite $(\frac{S_n}{n\log_2 n}, n\ge 1)$ converge en probabilité vers 1.

On peut en fait montrer que la suite $(2^{-n}S_{2^n} - n, n \geq 1)$ converge en loi[9]. ▲

Exercice V.11.

1. On note X_i la réponse de la i-ème personne interrogée ($X_i = 1$ si il vote pour le candidat A et $X_i = 0$ si il vote pour le candidat B). Les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes de même loi de Bernoulli de paramètre $p \in]0, 1[$. (Les variables X_i sont effectivement indépendantes si l'on a à faire à un tirage avec remise : une personne peut être interrogée plusieurs fois. Dans le cas d'un tirage sans remise (ce qui est souvent le cas d'un sondage), alors les variables ne sont pas indépendantes. Mais on peut montrer que si la taille de la population est grande devant n, le nombre de personnes interrogées, alors tout ce passe comme si le tirage était avec remise). On estime p par la moyenne empirique : $\bar{X}_n = \frac{1}{n}\sum_{k=1}^n X_k$. On déduit du TCL que avec une probabilité asymptotique de 95%,
$$p \in [\bar{X}_n \pm 1.96\frac{\sqrt{p(1-p)}}{\sqrt{n}}] \subset [\bar{X}_n \pm \frac{1.96}{2\sqrt{n}}] \simeq [\bar{X}_n \pm \frac{1}{\sqrt{n}}].$$
En particulier pour $n = 1\ 000$, on obtient une précision asymptotique (par excès) d'environ ± 3 points. La précision ne dépend pas de la taille de la population, pourvu qu'elle soit bien plus grande que n.
2. On a $p = 0.5 + 3.3\ 10^{-4}$. Comme on assure que A est le vainqueur dès que $\bar{X}_n - \frac{1}{\sqrt{n}} \geq 1/2$, et que dans 95% des cas $\bar{X}_n \in [p \pm \frac{1}{\sqrt{n}}]$ ($p \simeq 1/2$), on en déduit que l'on donne A gagnant dès que $p - \dfrac{2}{\sqrt{n}} \geq \dfrac{1}{2}$ avec une certitude d'au moins 95%. On en déduit $\dfrac{2}{\sqrt{n}} = 3.3\ 10^{-4}$ soit $n = 36\ 000\ 000$. Comme n est plus grand que la taille de la population, cela signifie que l'on ne peut pas faire l'approximation de la question précédente (variables indépendantes). Cela signifie aussi que l'on ne peut pas déterminer le gagnant en faisant un sondage, mais qu'il faut interroger toute la population.

▲

Exercice V.12.

1. D'après la loi forte des grands nombres, et la continuité de la fonction $t \mapsto t(1-t)$, on en déduit que la suite $(\hat{Y}_n(1-\hat{Y}_n), n \geq 1)$ converge p.s. vers $\sigma^2 = \mathrm{Var}(Y_1)$.
2. On a $\mathrm{Var}(Y_1) = p(1-p)$ et, comme $y_i^2 = y_i$, il vient
$$\frac{1}{N}\sum_{i=1}^N (y_i - p)^2 = \frac{1}{N}\sum_{i=1}^N y_i^2 - 2p\frac{1}{N}\sum_{i=1}^N y_i + p^2 = p - 2p^2 + p^2 = p(1-p).$$
3. Le théorème central limite assure que la suite $(\frac{\hat{Y}_n - p}{\sigma_n}, n \geq 1)$ converge en loi vers la loi gaussienne centrée réduite. Ainsi avec une probabilité (asymptotique) de $1-\alpha$, $p \in [\hat{Y}_n \pm \dfrac{\phi_{1-\alpha/2}\sigma}{\sqrt{n}}]$. Grâce au théorème de Slutsky, on peut remplacer σ par $\sqrt{\hat{Y}_n(1-\hat{Y}_n)}$. Comme $\hat{Y}_n(1-\hat{Y}_n)/n$ est de l'ordre de $\mathrm{Var}(\hat{Y}_n)$, la largeur de l'intervalle de confiance est de l'ordre de $2\phi_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{Y}_n)}$.

[9] Voir A. Martin-Löf, A limit theorem which clarifies the "Petersburg paradox", *J. Appl. Prob.*, **22**, 634-643 (1985).

4. On pose $s_n^2 = \mathrm{Var}(\hat{Y}_n^{st})$. Par indépendance des variables $(\hat{Y}_{n_h}^h, h \in \{1, \dots, H\})$, on a
$$s_n^2 = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \mathrm{Var}(\hat{Y}_{n_h}^h) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\sigma_h^2}{n_h}.$$

5. On vérifie que l'expression proposée correspond bien à σ^2. Comme $\sum_{h=1}^{H} \frac{N_h}{N} p_h = p$, il vient
$$\begin{aligned}\sum_{h=1}^{H} \frac{N_h}{N}\sigma_h^2 + \sum_{h=1}^{H} \frac{N_h}{N}(p_h - p)^2 =& \sum_{h=1}^{H} \frac{N_h}{N}(-p_h^2 + p - h + (p_h - p)^2) \\ =& \sum_{h=1}^{H} \frac{N_h}{N}(p^2 + p_h(1-2p)) \\ =& p^2 + p(1-2p) \\ =& \sigma^2.\end{aligned}$$

6. En reprenant l'expression de s_n^2, on peut écrire
$$s_n^2 = \frac{1}{N}\sum_{h=1}^{H} \frac{N_h^2}{N}\sigma_h^2 \frac{1}{n_h} = \sum_{h=1}^{H} \frac{N_h}{N}\sigma_h^2 \frac{N_h}{N n_h} = \frac{1}{n}\sigma^2(1-\eta^2).$$
On compare cette quantité à $\sigma_n^2 = \mathrm{Var}(\hat{Y}_n) = \dfrac{\sigma^2}{n}$. Ainsi plus η est proche de 1, plus la diminution de la variance est important. La valeur de η est proche de 1 lorsque les strates sont homogènes : p_h est proche de 0 ou proche de 1. Dans les 2 cas σ_h^2 est proche de 0. Le gain obtenu par la méthode de stratification peut être important.

7. Le taux d'abstention est d'environ 0.3%. Ainsi on trouve $\sigma^2 \simeq 0.21$, la variance inter-strates $\simeq 27.\, 10^{-4}$ et $\eta^2 \simeq 1.2\%$. La largeur de l'intervalle de confiance est donc multipliée par environ $\sqrt{0.988}$. Ainsi le gain apporté est équivalent à celui que l'on aurait obtenu si on avait procédé à un sondage classique et interrogé 8120 personnes ($8023/8120 \simeq 0.988$). Cet exemple numérique réel ne met pas en valeur la méthode de stratification. Il souligne en fait que les strates choisies ne sont pas homogènes pour l'abstention. Le choix des strates n'étant pas pertinent, la méthode de stratification n'est pas efficace.

▲

Exercice V.13.

1. On considère les fonctions caractéristiques :
$$\psi_{Y_m}(u) = (1 - p_m + p_m\, \mathrm{e}^{iu})^m = \left(1 - \frac{m p_m (1 - \mathrm{e}^{iu})}{m}\right)^m.$$
Rappelons que $(1 - \dfrac{x_n}{n})^n$ converge vers e^{-x} si x_n converge dans $\mathbb{C}$ vers x, quand n tend vers l'infini. On en déduit que $\lim_{n\to\infty} \psi_{Y_m}(u) = \mathrm{e}^{-\theta(1-\mathrm{e}^{iu})}$. On reconnaît la fonction caractérisitique de la loi de Poisson de paramètre θ. On en déduit donc $(Y_m, m \in \mathbb{N})$ converge en loi vers la loi de Poisson de paramètre θ.

2. Chaque impact a une probabilité $1/N = 1/576$ d'être dans le domaine i_0 donné. La loi du nombre d'impacts dans le domaine i_0 est donc une variable aléatoire binomiale $B(537, 1/576)$ soit approximativement une loi de Poisson de paramètre $\theta = 0,9323$. On peut alors comparer les probabilités empiriques N_k/N et les probabilités théoriques $p_k = \mathbb{P}(X = k)$, où X est une variable aléatoire de loi de Poisson de paramètre θ. Comparons en fait N_k et $N\,p_k$:

k	0	1	2	3	4	5+
N_k	229	211	93	35	7	1
Np_k	226.74	211.39	98.54	30.62	7.14	1.57

Les valeurs sont très proches. En fait, en faisant un test du χ^2, on peut vérifier qu'il est raisonnable d'accepter le modèle.

▲

Exercice V.14.

Remarquons que l'ensemble des fractions rationnelles $F = \{a/b^n; a \in \mathbb{N}, n \in N^*\} \cap [0,1]$ est dénombrable. En particulier, on a $\mathbb{P}(X \in F) = 0$. Cela implique que p.s. la représentation $X = \sum_{n=1}^{+\infty} \frac{X_n}{b^n}$ est unique.

1. Soit $n \geq 1$ et $a \in \{0, \ldots, b-1\}$. On a $X_n = a$ si et seulement si $b^n X = a + bq + r$ avec $q \in \{0, \ldots, b^{n-1} - 1\}$ et $r \in [0, 1[$. Donc $X_n = a$ si et seulement si $b^n X$ prend une valeur dans un intervalle de la forme $I_q = [a + bq, a + bq + 1[$ avec $q \in \{0, \ldots, b^{n-1} - 1\}$. Il vient :

$$\begin{aligned}
\mathbb{P}(X_n = a) &= \mathbb{P}(b^n X \in \cup_{q=0}^{b^{n-1}-1} I_q) \\
&= \sum_{q=0}^{b^{n-1}-1} \mathbb{P}(X \in b^{-n} I_q) \quad \text{car les intervalles } I_q \text{ sont disjoints} \\
&= b^{n-1} b^{-n} \quad \text{car les intervalles } I_q \text{ sont tous de mesure 1} \\
&= 1/b.
\end{aligned}$$

La variable aléatoire X_n suit donc une loi uniforme sur $\{0, \ldots, b-1\}$ pour $n \geq 1$.

2. On a $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X = \sum_{k=1}^n \frac{x_k}{b^k} + r)$ avec $r \in [0, b^{-n}[$. Il vient $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = b^{-n} = \prod_{k=1}^n \mathbb{P}(X_k = x_k)$. Cela implique que les variables aléatoires $X_1, \ldots, X_n$ sont bien indépendantes. Ceci étant vrai pour tout $n \geq 2$, cela signifie que les variables aléatoires $(X_n, n \geq 1)$ sont indépendantes. Elles ont même loi d'après la question précédente.

3. Soit $a \in \{0, \ldots, b-1\}$. On pose $Y_k = \mathbf{1}_{\{X_k = a\}}$. La quantité $\sum_{k=1}^n Y_k$ compte le nombre d'apparition du chiffre a parmi les n premiers chiffres de l'écriture en base b de X, et $\frac{1}{n}\sum_{k=1}^n Y_k$ est la fréquence correspondante. Les variables $(X_k, k \in \mathbb{N})$ étant indépendantes et de même loi, il en est de même pour les variables $(Y_k, k \in \mathbb{N})$. De plus, Y_k est une variable aléatoire de Bernoulli de paramètre $\mathbb{P}(X_k = a) = 1/b$. D'après la loi forte des grands nombres, on a : p.s.

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n Y_k = \mathbb{E}[Y_1] = 1/b.$$

Ceci signifie que p.s. X est simplement normal en base b. ce qui traduit bien le fait que la proportion de chiffres de l'écriture en base b égaux à a vaut $1/b$ pour presque tous les x de $[0,1]$.

4. Le raisonnement qui précède est vrai pour toute base $b \geq 2$. On note N_b l'ensemble des réels de $[0,1]$ simplement normal en base b. On a pour tout $b \geq 2$, $\mathbb{P}(X \in N_b) = 1$ et $\mathbb{P}(X \notin N_b) = 0$. On en déduit que

$$\mathbb{P}(X \notin \cup_{r\geq 1} N_{b^r}) \leq \sum_{r\geq 1} \mathbb{P}(X \notin N_b) = 0.$$

Cela implique que X est p.s. normal en base b. De même, on a

$$\mathbb{P}(X \notin \cup_{b\geq 2} N_b) \leq \sum_{b\geq 2} \mathbb{P}(X \notin N_b) = 0.$$

On en déduit que X est p.s. absolument normal.

▲

Exercice V.15.

1. On déduit de l'inégalité de Tchebychev que

$$\mathbb{P}(\Delta_n) = \mathbb{E}[\mathbf{1}_{\{|\bar{X}_n - x|>\delta\}}] \leq \frac{\mathbb{E}[(\bar{X}_n - x)^2]}{\delta^2} = \frac{\mathrm{Var}(\bar{X}_n)}{\delta^2} = \frac{x(1-x)}{n\delta^2} \leq \frac{1}{4n\delta^2},$$

car $x \in [0,1]$.

2. Remarquons que h étant continue sur $[0,1]$ compact, h est uniformément continue sur $[0,1]$ (théorème de Heine). Donc pour tout $\varepsilon > 0$, il existe $\delta > 0$, tel que pour tout $x, y \in [0,1]$, tel que $|x-y| \leq \delta$, alors $|h(x) - h(y)| \leq \varepsilon$. D'autre part, h étant continue sur $[0,1]$, elle est donc bornée par une constante $M > 0$. On a

$$\big|h(x) - \mathbb{E}[h(\bar{X}_n)]\big| \leq \mathbb{E}[\big|h(x) - h(\bar{X}_n)\big|] = A + B,$$

avec

$$\begin{aligned} A =& \mathbb{E}[|h(x) - h(\bar{X}_n)|\mathbf{1}_{\{|\bar{X}_n - x|\leq\delta\}}] \\ B =& \mathbb{E}[|h(x) - h(\bar{X}_n)|\mathbf{1}_{\{|\bar{X}_n - x|>\delta\}}]. \end{aligned}$$

D'après les remarques préliminaires on a $A < \varepsilon\mathbb{P}(|\bar{X}_n - x| \leq \delta) \leq \varepsilon$. D'après la question précédente, on a

$$B \leq 2M\mathbb{P}(|\bar{X}_n - x| > \delta) \leq \frac{M}{2n\delta^2}.$$

Pour $n \geq M/2\varepsilon\delta^2$, on a $B \leq \varepsilon$ et $\big|h(x) - \mathbb{E}[h(\bar{X}_n)]\big| \leq 2\varepsilon$ pour tout $x \in [0,1]$. En conclusion, on a

$$\lim_{n\to\infty} \sup_{x\in[0,1]} \big|h(x) - \mathbb{E}[h(\bar{X}_n)]\big| = 0.$$

3. $n\bar{X}_n = \sum_{k=1}^n X_k$ est une variable aléatoire de loi binomiale de paramètres (n, x).

4. Donc on a
$$\mathbb{E}[h(\bar{X}_n)] = \mathbb{E}[h(n\bar{X}_n/n)] = \sum_{k=1}^{n} h(\frac{k}{n}) C_n^k x^k (1-x)^{n-k}.$$
On déduit de la question 2, que la suite de polynômes $(\sum_{k=1}^{n} h(\frac{k}{n}) C_n^k x^k (1-x)^{n-k}, n \geq 1)$, appelés polynômes de Bernstein, converge uniformément vers h sur $[0,1]$.

5. On raisonne cette fois-ci avec des variables aléatoires $(X_k, k \geq 1)$ indépendantes de loi de Poisson de paramètre 1. Et on utilise le fait que $\sum_{k=1}^{n} X_k$ suit une loi de Poisson de paramètre n. La convergence n'est pas uniforme. En effet, pour l'exemple donné, on obtient
$$\lim_{n\to\infty} \left| f(x_n) - \sum_{k=0}^{\infty} \mathrm{e}^{-nx_n} \frac{(nx_n)^k}{k!} f(k/n) \right| = \mathrm{e}^{-\pi}.$$

▲

Exercice V.16.

1. Pour tout $y \in \mathbb{R}$, par indépendance des X_k,
$$\mathbb{P}(Z_n \leq y) = \mathbb{P}\left(X_1 \leq n^{1/\lambda} y\right)^n = (1-\varepsilon_n)^n$$
où $\varepsilon_n = \mathbb{P}\left(X_1 > n^{1/\lambda} y\right)$. Pour $y \leq 0$, on a $\varepsilon_n > \eta > 0$, et donc $\lim_{n\to+\infty} \mathbb{P}(Z_n \leq y) = 0$. Pour $y > 0$, on a $\varepsilon_n \sim \frac{\alpha y^{-\lambda}}{n}$ quand $n \to \infty$. Par conséquent, $\lim_{n\to+\infty} \mathbb{P}(Z_n \leq y) = \exp(-\alpha y^{-\lambda})$. Cela prouve que la fonction de répartition de Z_n converge vers $\exp(-\alpha y^{-\lambda})\mathbf{1}_{\{y>0\}}$ qui est la fonction de répartion de la loi de Fréchet (i.e. de Y). Cela prouve que Z_n converge en loi vers Y.

2. Soient n individus. On note X_k le niveau annuel d'exposition au mercure de l'individu k. On suppose les X_k indépendants et de même loi que X. Pour un individu de 70 kg, la dose annuelle admissible fixée par l'OMS est $s = 18.14\, 10^{-3}$ g. La probabilité qu'un individu au moins ait un niveau de mercure trop élevé est
$$\begin{aligned} p_n = \mathbb{P}(\max(X_1, \ldots, X_n) > s) &= 1 - \mathbb{P}(\max(X_1, \ldots, X_n) \leq s) \\ &= 1 - \mathrm{e}^{n \ln(1-\alpha s^{-\lambda})} \\ &\simeq 1 - \exp(-\alpha s^{-\lambda} n). \end{aligned}$$
Si $n_1 = 225\,362$[10], $n_2 = 100$, on a
$$p_{n_1} \simeq 1, \qquad p_{n_2} \simeq 0.01.$$
Enfin, $1 - \exp(-\alpha s^{-\lambda} n) \geq 0.95$ si et seulement si $n \geq \dfrac{\ln 20}{\alpha s^{-\lambda}}$, i.e. $n \geq 27\,214$.

▲

[10] Population du 15ème arrondissement de Paris, Recensement 1999.

Chapitre VI

Vecteurs Gaussiens

VI.1 Énoncés

Exercice VI.1.
Soit $X = (X_1, X_2, X_3, X_4)$ un vecteur gaussien centré de matrice de covariance :

$$\Gamma = \begin{pmatrix} 2 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 2 \end{pmatrix}$$

1. Que peut-on dire de X_3 et de (X_1, X_2, X_4) ?
2. Donner la loi marginale de (X_1, X_2) et calculer $\mathbb{E}[X_1|X_2]$.
3. Même question pour (X_2, X_4).
4. En déduire deux variables indépendantes de X_2, fonctions respectivement de X_1, X_2 et de X_2, X_4.
5. Trouver une décomposition de X en quatre vecteurs indépendants.

△

Exercice VI.2.
Soit $X = (X_1, X_2, X_3, X_4)^t$ un vecteur gaussien de loi $\mathcal{N}(\mu_X, \Sigma_X)$ avec

$$\mu_X = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \Sigma_X = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

Soit $\alpha \in \mathbb{R}$ et $Y = (Y_1, Y_2, Y_3)^t$ défini par $Y_1 = X_1 + \alpha X_2$, $Y_2 = X_2$ et $Y_3 = X_4$.

1. Déterminer la loi de Y.
2. Quelle condition doit vérifier α pour que les variables aléatoires Y_1, Y_2, Y_3 soient indépendantes ? Calculer $\mathbb{E}\left[Y_1^2 Y_2^2 Y_3^2\right]$ sous ces conditions.

△

Exercice VI.3.
Soient X et Z deux variables aléatoires réelles indépendantes, X étant de loi $\mathcal{N}(0,1)$ et Z de loi définie par $\mathbb{P}(Z=1)=\mathbb{P}(Z=-1)=1/2$. On pose $Y=ZX$.

1. Déterminer la loi de Y.
2. Calculer $\mathrm{Cov}(X,Y)$.
3. Le vecteur (X,Y) est-il gaussien ? Les variables X et Y sont-elles indépendantes ?

△

Exercice VI.4.
Soit $X_1,\dots,X_n$ des variables aléatoires réelles indépendantes de même loi, d'espérance m, de variance σ^2 finie. On pose

$$\bar{X}_n=\frac{1}{n}\sum_{i=1}^{n}X_i,\quad \Sigma_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-m)^2 \text{ et } V_n=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2.$$

On suppose que la loi de X_i est la loi gaussienne $\mathcal{N}(m,\sigma^2)$.

1. Quelle est la loi de $\bar{X}_n$?
2. Quelle est la loi de $n\Sigma_n^2/\sigma^2$?
3. Montrer que $\bar{X}_n$ et V_n sont indépendantes.
4. Montrer que $(n-1)V_n/\sigma^2$ suit la loi $\chi^2(n-1)$.

△

Exercice VI.5.
Soit $X_1,\dots,X_n$ des variables aléatoires réelles indépendantes de même loi, d'espérance m, de variance σ^2 finie. On suppose que la moyenne empirique $\bar{X}_n=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}X_i$ et la variance empirique $V_n=\dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i-\bar{X})^2$ sont des variables aléatoires indépendantes. Le but de cet exercice est de démontrer que la loi de X_i est alors la loi gaussienne $\mathcal{N}(m,\sigma^2)$.

On note ψ la fonction caractéristique de X_i. On suppose $m=0$.

1. Calculer $\mathbb{E}[(n-1)V_n]$ en fonction de σ^2. Montrer que pour tout réel t,

$$\mathbb{E}[(n-1)V_n\,\mathrm{e}^{itn\bar{X}_n}]=(n-1)\psi(t)^n\sigma^2.$$

2. En développant V_n dans l'égalité précédente, vérifier que

$$\mathbb{E}[(n-1)V_n\,\mathrm{e}^{itn\bar{X}_n}]=-(n-1)\psi''(t)\psi(t)^{n-1}-(n-1)\psi'(t)^2\psi(t)^{n-2}.$$

3. En déduire que ψ est solution de l'équation différentielle

$$\begin{cases}\dfrac{\psi''}{\psi}-(\dfrac{\psi'}{\psi})^2=-\sigma^2\\ \psi(0)=1\,,\psi'(0)=0\end{cases}$$

4. En déduire que la loi des variables X_i est la loi gaussienne $\mathcal{N}(0, \sigma^2)$.
5. Que peut on dire si on ne suppose plus $m = 0$?

△

Exercice VI.6.
Soient X et Y deux variables aléatoires réelles gaussiennes indépendantes.

1. Donner une condition nécessaire et suffisante pour que $X+Y$ et $X-Y$ soient indépendantes.
2. On suppose de plus que X et Y sont des gaussiennes centrées réduites. Calculer la fonction caractéristique de $Z_1 = X^2/2$ puis celle de $Z_2 = (X^2 - Y^2)/2$.
3. Montrer que Z_2 peut s'écrire comme le produit de deux variables aléatoires normales indépendantes.

△

Exercice VI.7.
Soit $(X_n, n \geq 1)$, une suite de variables aléatoires indépendantes, de loi normale $\mathcal{N}(\theta, \theta)$, avec $\theta > 0$. L'objectif de cet exercice est de présenter une méthode pour estimer θ, et de donner un (bon) intervalle de confiance pour cette estimation. On note

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k \quad \text{et} \quad V_n = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \bar{X}_n)^2.$$

1. Donner la loi de $\bar{X}_n$, son espérance et sa variance. Déterminer la limite de $(\bar{X}_n, n \geq 1)$.
2. Donner la loi de V_n, son espérance et sa variance. Déterminer la limite de $(V_n, n \geq 2)$.
3. Donner la loi du couple $(\bar{X}_n, V_n)$. Déterminer la limite de $((\bar{X}_n, V_n), n \geq 2)$.
4. On considère la classe des variables aléatoires T_n^λ de la forme :

$$T_n^\lambda = \lambda \bar{X}_n + (1-\lambda)V_n, \quad \lambda \in \mathbb{R}.$$

   Calculer son espérance, sa variance, et montrer la convergence presque sûre de $(T_n^\lambda, n \geq 2)$.
5. Étudier la convergence en loi de $(\sqrt{n}(\bar{X}_n - \theta), n \geq 1)$.
6. Étudier la convergence en loi de $(\sqrt{n}(V_n - \theta), n \geq 2)$.
7. Étudier la convergence en loi de $(\sqrt{n}(\bar{X}_n - \theta, V_n - \theta), n \geq 2)$.
8. Étudier la convergence en loi de $(\sqrt{n}(T_n^\lambda - \theta), n \geq 2)$.
9. On pose $\sigma = \sqrt{\lambda^2\theta + 2(1-\lambda)^2\theta^2}$. Construire, à partir de T_n^λ, σ et n, un intervalle de confiance de θ au niveau asymptotique 95%. Autrement dit trouver un intervalle aléatoire I_n, fonction de T_n^λ, σ et n, qui contient le paramètre θ, avec une probabilité asymptotique de 95%.
10. Comme σ est inconnu on l'estime par $\sigma_n = \sqrt{\lambda^2 T_n^\lambda + 2(1-\lambda)^2 (T_n^\lambda)^2}$ et on le remplace dans l'expression de I_n. Montrer que l'intervalle obtenu est encore un intervalle de confiance de θ au niveau asymptotique de 95%. Donner un tel intervalle pour la réalisation $\lambda = 0.5$, $n = 100$, $\bar{x}_n = 4.18$ et $s_n = 3.84$.

11. Vérifier qu'il existe un unique réel $\lambda^* \in [0,1]$, fonction de θ, qui minimise la longueur de l'intervalle de confiance, I_n. On considère maintenant les variables aléatoires $\lambda_n^* = \dfrac{2V_n}{1+2V_n}$. Montrer que la suite $(\lambda_n^*, n \geq 2)$ converge presque sûrement vers λ^*.
12. Étudier la convergence en loi de la suite $(\sqrt{n}(T_n^{\lambda_n^*} - \theta), n \geq 2)$. En déduire un intervalle de confiance de θ au niveau asymptotique de 95%. Donner un tel intervalle pour les données ci-dessus.

△

VI.2 Corrections

Exercice VI.1.

1. On voit que quelque soit $i = 1, 2, 4$, on a $\mathrm{Cov}(X_3 X_i) = 0$, donc X_3 est indépendant du vecteur gaussien (X_1, X_2, X_4).

2. Le vecteur gaussien (X_1, X_2) est centré, de matrice de covariance $\Gamma_{12} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$. Pour le calcul de l'espérance conditionnelle, on cherche à écrire $X_1 = aX_2 + W$ où W est une variable aléatoire indépendante de X_2. Comme X_1 et X_2 sont centrées, W l'est aussi. On calcule alors

$$\mathrm{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] = a\mathbb{E}[X_2^2] + \mathbb{E}[X_2]\mathbb{E}[W] = a,$$

   car W est indépendante de X_2. On en déduit que $a = 1$ et que $W = X_1 - X_2$. Il vient ensuite $\mathbb{E}[X_1|X_2] = \mathbb{E}[W] + X_2 = X_2$.

3. Le vecteur (X_2, X_4) est également gaussien centré, de matrice de covariance $\Gamma_{24} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. On obtient de même que $\mathbb{E}[X_4|X_2] = X_2$.

4. Si (X, Y) est un vecteur gaussien alors $(X - \mathbb{E}[X|Y], Y)$ est aussi un vecteur gaussien car $\mathbb{E}[X|Y]$ est une fonction linéaire de Y. Or $\mathbb{E}\left[(X - \mathbb{E}[X|Y])Y\right] = 0$, donc ces deux variables sont indépendantes. Au vu des questions 2 et 3, on sait que $X_1 - X_2$ et $X_4 - X_2$ sont deux variables gaussiennes indépendantes de X_2.

5. On choisit comme décomposition de X : $(X_1 - X_2, X_2, X_3, X_4 - X_2)$. Pour montrer que ces vecteurs sont indépendants, il suffit de montrer que $X_1 - X_2$ et $X_4 - X_2$ sont indépendantes. Or le vecteur $(X_1 - X_2, X_4 - X_2)$ est gaussien donc ces variables sont indépendantes si leur covariance est nulle. Or, on a $\mathbb{E}\left[(X_1 - X_2)(X_4 - X_2)\right] = \mathrm{Cov}(X_1, X_4) - \mathrm{Cov}(X_2, X_4) - \mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_2, X_2) = 0$.

▲

Exercice VI.2.

1. On peut écrire $Y = AX$ avec

$$A = \begin{bmatrix} 1 & \alpha & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

   Donc $Y = (Y_1, Y_2, Y_3)$ est un vecteur gaussien (en tant que transformation linéaire d'un vecteur gaussien). Il ne reste donc plus qu'à déterminer l'espérance et la matrice de variance-covariance :

$$\mathbb{E}[Y] = A\mathbb{E}[X] = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\begin{aligned}
\Sigma_Y &= A\Sigma_X A^t \\
&= \begin{bmatrix} 1 & \alpha & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
&= \begin{bmatrix} 1 & \alpha & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2+\alpha & 1 & 0 \\ 1+\alpha & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 2 \end{bmatrix} \\
&= \begin{bmatrix} 2+\alpha+\alpha\,(1+\alpha) & 1+\alpha & 0 \\ 1+\alpha & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.
\end{aligned}$$

2. Le vecteur Y étant gaussien, il y a équivalence entre indépendance des composantes et covariances nulles. Donc les variables Y_1, Y_2, Y_3 sont indépendants si et seulement si $1+\alpha = 0$ soit $\alpha = -1$.
   On a alors
   $$\Sigma_Y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$
   Donc dans ce cas, on a
   $$\mathbb{E}\left[Y_1^2 Y_2^2 Y_3^2\right] = \mathbb{E}\left[Y_1^2\right]\mathbb{E}\left[Y_2^2\right]\mathbb{E}\left[Y_3^2\right] = 2.$$

▲

Exercice VI.3.

1. Soit g une fonction mesurable bornée. On a par indépendance
   $$\begin{aligned}
   \mathbb{E}[g(Y)] &= \mathbb{E}[g(ZX)\mathbf{1}_{\{Z=1\}} + g(ZX)\mathbf{1}_{\{Z=-1\}}] \\
   &= \mathbb{E}[g(X)]\mathbb{P}(Z=1) + \mathbb{E}[g(-X)]\mathbb{P}(Z=-1) \\
   &= \frac{1}{2}\Big(\mathbb{E}[g(X)] + \mathbb{E}[g(-X)]\Big) \\
   &= \mathbb{E}[g(X)],
   \end{aligned}$$
   car la loi de X est symétrique. Donc X et Y ont même loi : $\mathcal{N}(0,1)$.
2. On a
   $$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] = \mathbb{E}[X^2]\mathbb{P}(Z=1) + \mathbb{E}[-X^2]\mathbb{P}(Z=-1) = 0,$$
3. On a $\mathbb{P}(X+Y=0) = \mathbb{P}(Z=-1) = 1/2$ donc $X+Y$ n'est pas une variable aléatoire continue. En particulier ce n'est pas une variable gaussienne. Donc (X,Y) n'est pas un vecteur gaussien. De plus X et Y ne sont pas indépendants, sinon (X,Y) serait un vecteur gaussien. Ainsi deux variables aléatoires gaussiennes de covariance nulle ne sont pas forcément indépendantes.

▲

Exercice VI.4.

1. Le vecteur $X = (X_1, \dots, X_n)$ est un vecteur gaussien $\mathcal{N}(m\mathbf{1}_n, \sigma^2 I_n)$. Donc, $\bar{X}_n$ qui est combinaison linéaire des composantes de X est une variable aléatoire gaussienne. On a $\mathbb{E}[\bar{X}_n] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = m$ et $\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) = \frac{\sigma^2}{n}$. Donc on en déduit que la loi de $\bar{X}$ est $\mathcal{N}(m, \sigma^2/n)$
2. On a $n\Sigma_n^2/\sigma^2 = \sum_{i=1}^n \left(\frac{X_i - m}{\sigma}\right)^2$ où $\frac{X_i - m}{\sigma}$ a pour loi $\mathcal{N}(0,1)$. Il vient donc que $n\Sigma_n^2/\sigma^2$ suit la loi $\chi^2(n)$.
3. Pour montrer que $\bar{X}$ et V_n^2 sont indépendants, il suffit de montrer que $\bar{X}_n$ et $(X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$ le sont. Or le vecteur $Y = (\bar{X}_n, X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$ est gaussien (les coordonnées sont des combinaisons linéaires de celles de X), donc il suffit de montrer que $\mathrm{Cov}(\bar{X}_n, X_i - \bar{X}n_)= 0$, ce qui est le cas :

$$\mathrm{Cov}(\bar{X}_n, X_i) = \mathrm{Cov}(\frac{1}{n}\sum_{j=1}^n X_j, X_i) = \frac{1}{n}\mathrm{Cov}(X_i, X_i) = \frac{1}{n}\sigma^2 = \mathrm{Var}(\bar{X}_n).$$

4. Supposons $m = 0$. En développant, on obtient

$$(n-1)V_n = \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \sum_{i=1}^n X_i^2 - \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n X_i X_j.$$

On en déduit que

$$\sum_{i=1}^n X_i^2 = (n-1)V_n + \left(\frac{\sum_{i=1}^n X_i}{\sqrt{n}}\right)^2.$$

La loi de $\sum_{i=1}^n X_i^2/\sigma^2$ est la loi $\chi^2(n)$. Comme $\frac{\sum_{i=1}^n X_i}{\sqrt{n}\sigma} = \sqrt{n}\bar{X}_n/\sigma$ est une variable aléatoire gaussienne centrée réduite, la loi de $Z = \left(\frac{\sum_{i=1}^n X_i}{\sigma\sqrt{n}}\right)^2$ est donc la loi $\chi^2(1)$. Par indépendance, obtient pour les fonctions caractéristiques

$$\left(\frac{1}{1-2iu}\right)^{n/2} = \psi_{\frac{\sum_{i=1}^n X_i^2}{\sigma^2}}(u) = \psi_{\frac{(n-1)V_n}{\sigma^2}}(u)\psi_Z(u) = \psi_{\frac{(n-1)V_n}{\sigma^2}}(u)\left(\frac{1}{1-2iu}\right)^{1/2}.$$

On en déduit donc que $\psi_{(n-1)V_n/\sigma^2}(u) = \left(\frac{1}{1-2iu}\right)^{(n-1)/2}$. On reconnaît la fonction caractéristique de la loi $\chi^2(n-1)$v. Si $m \neq 0$, alors on considère $X_i - m$ au lieu de X_i dans ce qui précède. Cela ne change pas la définition de V_n et on obtient la même conclusion.

▲

Exercice VI.5.

1. On a

$$\mathbb{E}[(n-1)V_n] = \mathbb{E}[\sum_{j=1}^{n}(X_j^2 - 2\,\bar{X}_n X_j + \bar{X}_n^2)] = n\sigma^2 - 2n\frac{\sigma^2}{n} + n^2\frac{\sigma^2}{n^2} = (n-1)\sigma^2.$$

Comme V_n et $\bar{X}_n$ sont indépendants, on a

$$\mathbb{E}[(n-1)V_n\ \mathrm{e}^{itn\bar{X}_n}] = \mathbb{E}[(n-1)V_n]\,\mathbb{E}[\mathrm{e}^{itn\bar{X}_n}] = (n-1)\sigma^2\psi(t)^n.$$

2. En développant, il vient

$$\begin{aligned}
\mathbb{E}[(n-1)V_n\ \mathrm{e}^{itn\bar{X}_n}] &= \mathbb{E}\big[\big((1-\frac{1}{n})\sum_{j=1}^{n}X_j^2 - \frac{2}{n}\sum_{1\le j<k\le n}X_jX_k\big)\,\mathrm{e}^{it\sum_{j=1}^{n}X_j}\,\big] \\
&= (1-\frac{1}{n})\sum_{j=1}^{n}\mathbb{E}[X_j^2\,\mathrm{e}^{itX_j}]\mathbb{E}[\mathrm{e}^{itX_1}]^{n-1} \\
&\qquad - \frac{2}{n}\sum_{1\le j<k\le n}\mathbb{E}[X_j\,\mathrm{e}^{itX_j}]\mathbb{E}[X_k\,\mathrm{e}^{itX_k}]\mathbb{E}[\mathrm{e}^{itX_1}]^{n-2} \\
&= -(1-\frac{1}{n})n\psi''(t)\psi(t)^{n-1} - \frac{2}{n}\frac{n(n-1)}{2}(-i\psi'(t))^2\psi(t)^{n-2} \\
&= -(n-1)\psi''(t)\psi(t)^{n-1} - (n-1)\psi'(t)^2\psi(t)^{n-2}.
\end{aligned}$$

3. Donc ψ est solution de l'équation différentielle $\begin{cases} \dfrac{\psi''}{\psi} - (\dfrac{\psi'}{\psi})^2 = -\sigma^2 \\ \psi(0) = 1\,, \psi'(0) = 0. \end{cases}$
4. On pose $\varphi = \psi'/\psi$. On obtient que $\varphi'(t) = -\sigma^2$ et $\varphi(0) = 0$. On en déduit que $\varphi(t) = -\sigma^2 t$. Donc $\psi'(t) = -\sigma^2 t\psi(t)$. Une solution est $\psi(t) = \mathrm{e}^{-t^2\sigma^2/2}$. On cherche alors les solutions sous la forme $\psi(t) = \mathrm{e}^{-t^2\sigma^2/2}\ h(t)$ (méthode de la variation de la constante). On en déduit que $h' = 0$. La condition $\psi(0) = 1$ implique que $\mathrm{e}^{-t^2\sigma^2/2}$ est la seule solution de l'équation différentielle considérée. La loi de X_i est donc la loi gaussienne $\mathcal{N}(0, \sigma^2)$.
5. Si $m \neq 0$, on applique ce qui précède à $X_i^* = X_i - m$. On trouve alors que la loi de X_i est la loi gaussienne $\mathcal{N}(m, \sigma^2)$.

▲

Exercice VI.6.

1. Comme X et Y sont indépendantes et gaussiennes, (X, Y) est un vecteur gaussien. Par conséquent, $(X+Y, X-Y)$ est un vecteur gaussien et les variables $X+Y$ et $X-Y$ sont des variables gaussiennes. Ces deux variables sont indépendantes si et seulement si leur covariance est nulle :

$$\mathbb{E}\,[(X+Y)(X-Y)] - \mathbb{E}[X+Y]\mathbb{E}[X-Y] = \mathrm{Var}(X) - \mathrm{Var}(Y).$$

On conclut donc que $X+Y$ et $X-Y$ sont indépendantes si et seulement si $\mathrm{Var}(X) = \mathrm{Var}(Y)$.

2. On sait que X^2 suit une loi $\chi^2(1) = \Gamma(1/2, 1/2)$. On a donc pour Z_1,

$$\psi_{Z_1}(u) = \psi_{X^2}(u/2) = \left(\frac{1/2}{1/2 - iu/2}\right)^{1/2} = \left(\frac{1}{1-iu}\right)^{1/2}.$$

Ainsi la loi de Z_1 est la loi $\Gamma(1, 1/2)$. Pour Z_2 on utilise le fait que X et Y sont indépendantes et donc $\psi_{Z_2}(u) = \psi_{Z_1}(u)\psi_{Z_1}(-u)$. On en déduit que $\psi_{Z_2}(u) = \frac{1}{\sqrt{1+u^2}}$.

3. On voit que $Z_2 = \left(\frac{X-Y}{\sqrt{2}}\right)\left(\frac{X+Y}{\sqrt{2}}\right)$ et vu la question 1, ces variables sont indépendantes. De plus $\mathrm{Var}(\frac{X-Y}{\sqrt{2}}) = \mathrm{Var}(\frac{X+Y}{\sqrt{2}}) = 1$, d'où le résultat demandé.

▲

Exercice VI.7.

1. D'après la proposition 11 page 137 du polycopié : $\bar{X}_n$ est une variable aléatoire gaussienne $\mathcal{N}(\theta, \frac{\theta}{n})$, $\mathbb{E}[\bar{X}_n] = \theta$ et $\mathrm{Var}(\bar{X}_n) = \dfrac{\theta}{n}$. Par la loi forte des grands nombres, la suite $(\bar{X}_n, n \geq 1)$ converge presque sûrement vers $\mathbb{E}[X_1] = \theta$.
2. $(n-1)V_n/\theta$ suit la loi $\chi^2(n-1)$, $\mathbb{E}[V_n] = \theta$ et $\mathrm{Var}(V_n) = \dfrac{2\theta^2}{(n-1)}$. Remarquons que $V_n = \frac{n}{n-1}\frac{1}{n}\sum_{k=1}^n X_k^2 - \frac{n}{n-1}\left(\frac{1}{n}\sum_{k=1}^n X_k\right)^2$. Comme les variables aléatoires X_k sont indépendantes, de même loi et de carré intégrable, on déduit de la loi forte des grands nombres que la suite $(\frac{1}{n}\sum_{k=1}^n X_k^2, n \geq 2)$ converge presque sûrement vers $\mathbb{E}[X_1^2]$. On en déduit donc que la suite $(V_n, n \geq 2)$ converge presque sûrement vers $\mathrm{Var}(X_1) = \theta$.
3. Les variables aléatoires $\bar{X}_n$ et V_n sont indépendants. La loi du couple est donc la loi produit. La suite $((\bar{X}_n, V_n), n \geq 2)$ converge p.s. vers (θ, θ).
4. On a

$$\mathbb{E}[T_n^\lambda] = \lambda\mathbb{E}[\bar{X}_n] + (1-\lambda)\mathbb{E}[V_n] = \theta,$$
$$\mathrm{Var}[T_n^\lambda] = \lambda^2\,\mathrm{Var}[\bar{X}_n] + (1-\lambda)^2\,\mathrm{Var}[V_n] + 2\,\mathrm{Cov}(\bar{X}_n, V_n) = \lambda^2\frac{\theta}{n} + (1-\lambda)^2\frac{2\theta^2}{(n-1)},$$

car $\bar{X}_n$ et V_n sont indépendants. Par continuité de l'application $(x, s) \mapsto \lambda x + (1-\lambda)s$, on en déduit que la suite $(T_n^\lambda, n \geq 2)$ converge presque sûrement vers $\lambda\mathbb{E}[X_1] + (1-\lambda)\,\mathrm{Var}(X_1) = \theta$.
5. Les variables aléatoires $(X_k, k \geq 1)$ sont de même loi, indépendantes, et de carré intégrable. De plus $\mathbb{E}[X_k] = \theta$ et $\mathrm{Var}(X_k) = \theta$. On déduit du théorème central limite, que la suite $\left(\sqrt{n}(\bar{X}_n - \theta), n \geq 1\right)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \theta)$.
6. Remarquons que $\frac{n-1}{\theta}V_n$ est distribué suivant la loi $\chi^2(n-1)$. En particulier, $\frac{n-1}{\theta}V_n$ a même loi que $\sum_{k=1}^{n-1} g_k^2$, où $(g_k, k \geq 1)$ est une suite de variables aléatoires indépendantes, identiquement distribuées, de loi $\mathcal{N}(0, 1)$. On en déduit donc

$$\begin{aligned}\sqrt{n}(V_n - \theta) &= \sqrt{\frac{n}{n-1}}\theta\left(\sqrt{n-1}\left(\frac{1}{\theta}V_n - 1\right)\right)\\ &\overset{loi}{=} \sqrt{\frac{n}{n-1}}\theta\left(\sqrt{n-1}(\frac{1}{n-1}\sum_{k=1}^{n-1} g_k^2 - 1)\right).\end{aligned}$$

Comme $\mathbb{E}[g_k^2] = 1$ et $\mathrm{Var}(g_k^2) = 2$, on déduit du théorème central limite, que la suite $(\sqrt{n-1}(\frac{1}{n-1}\sum_{k=1}^{n-1} g_k^2 - 1), n \geq 2)$ converge en loi vers G de loi gaussienne $\mathcal{N}(0,2)$. Remarquons que $\sqrt{\frac{n}{n-1}}\theta$ converge vers θ. On déduit du théorème de Slutsky, la convergence en loi de la suite $((\sqrt{\frac{n}{n-1}}\theta, \sqrt{n-1}(\frac{1}{n-1}\sum_{k=1}^{n-1} g_k^2 - 1), n \geq 1)$ vers (θ, G). Par continuité de la multiplication, on en déduit que la suite $(\sqrt{n}(V_n - \theta), n \geq 2)$ converge en loi vers θG. Soit encore $(\sqrt{n}(V_n - \theta), n \geq 2)$ converge en loi vers $\mathcal{N}(0, 2\theta^2)$.

7. On pose $Y_n = \sqrt{n}(\bar{X}_n - \theta)$ et $W_n = \sqrt{n}(V_n - \theta)$. En utilisant l'indépendance entre Y_n et W_n, et les deux questions précédentes, on obtient

$$\lim_{n\to\infty} \psi_{(Y_n,W_n)}(v,w) = \lim_{n\to\infty} \psi_{Y_n}(v)\psi_{W_n}(w) = \mathrm{e}^{-\theta v^2/2}\,\mathrm{e}^{-2\theta^2 w^2/2}$$

pour tout $v, w \in \mathbb{R}$. On reconnaît la fonction caractéristique du couple (Y, W), où Y et W sont indépendants de loi respective $\mathcal{N}(0,\theta)$ et $\mathcal{N}(0, 2\theta^2)$. On en déduit que la suite $(\sqrt{n}(\bar{X}_n - \theta, V_n - \theta), n \geq 2)$ converge en loi vers le vecteur gaussien (Y, W).

8. Par continuité de l'application $(a,b) \to \lambda a + (1-\lambda)b$, on en déduit que la suite $(\sqrt{n}(T_n^\lambda - \theta) = \lambda\sqrt{n}(\bar{X}_n - \theta) + (1-\lambda)\sqrt{n}(V_n - \theta), n \geq 2)$ converge en loi vers $\lambda Y + (1-\lambda)W$. Autrement dit la suite $(\sqrt{n}(T_n^\lambda - \theta), n \geq 2)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \lambda^2\theta + 2(1-\lambda)^2\theta^2)$.

9. Comme $\sqrt{n}\dfrac{T_n^\lambda - \theta}{\sigma}$ converge en loi vers une loi gaussienne $\mathcal{N}(0,1)$, et que la loi gaussienne est une loi à densité, on a

$$\lim_{n\to\infty} \mathbb{P}\left(\theta \in [T_n^\lambda - a\frac{\sigma}{\sqrt{n}}, T_n^\lambda + a\frac{\sigma}{\sqrt{n}}]\right) = \frac{1}{\sqrt{2\pi}}\int_{-a}^{a} e^{-\frac{x^2}{2}}dx.$$

L'intervalle aléatoire $[T_n^\lambda - a\frac{\sigma}{\sqrt{n}}, T_n^\lambda + a\frac{\sigma}{\sqrt{n}}]$ est un intervalle de confiance de θ de niveau asymptotique $\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-a}^{a} e^{-\frac{x^2}{2}}dx$. Pour le niveau de 95%, on trouve $a = 1.96$. Soit alors

$$I_n = [T_n^\lambda - 1.96\frac{\sigma}{\sqrt{n}}, T_n^\lambda + 1.96\frac{\sigma}{\sqrt{n}}].$$

10. Comme σ_n converge presque sûrement vers σ, en appliquant le théorème de Slutsky, on montre que $(\sqrt{n}\dfrac{T_n^\lambda - \theta}{\sigma_n}, n \geq 2)$ converge en loi vers une loi gaussienne $\mathcal{N}(0,1)$. On a

$$\lim_{n\to\infty} \mathbb{P}\left(\theta \in [T_n^\lambda - a\frac{\sigma_n}{\sqrt{n}}, T_n^\lambda + a\frac{\sigma_n}{\sqrt{n}}]\right) = \frac{1}{\sqrt{2\pi}}\int_{-a}^{a} e^{-\frac{x^2}{2}}dx.$$

L'intervalle aléatoire $[T_n^\lambda - a\frac{\sigma_n}{\sqrt{n}}, T_n^\lambda + a\frac{\sigma_n}{\sqrt{n}}]$ est un intervalle de confiance de θ de niveau asymptotique $\dfrac{1}{\sqrt{2\pi}}\displaystyle\int_{-a}^{a} e^{-\frac{x^2}{2}}dx$. Pour le niveau de 95%, on trouve $a = 1.96$. Soit alors

$$\tilde{I}_n = [T_n^\lambda - 1.96\frac{\sigma_n}{\sqrt{n}}, T_n^\lambda + 1.96\frac{\sigma_n}{\sqrt{n}}].$$

On obtient l'intervalle $\tilde{I}_n = [3.421, 4.599]$.

11. En minimisant la fonction $\lambda \to \lambda^2\theta + 2(1-\lambda)^2\theta^2$, on trouve $\lambda^* = \dfrac{2\theta}{1+2\theta}$. Par la convergence presque sûre de la suite V_n vers θ et la continuité de la fonction $x \to \dfrac{2x}{1+2x}$ sur $[0, +\infty[$, on a la convergence presque sûre de la suite $(\lambda_n^*, n \geq 2)$ vers λ^*.

12. On utilise les notations des questions précédentes. Par le théorème de Slutsky, on a la convergence en loi de $((\lambda_n^*, Y_n, W_n), n \geq 2)$ vers (λ^*, Y, W). Comme la fonction $(\lambda', a, b) \to \lambda' a + (1-\lambda')b$ est continue, on en déduit que $(\sqrt{n}(T_n^{\lambda_n^*} - \theta) = \lambda_n^*\sqrt{n}(\bar{X}_n - \theta) + (1-\lambda_n^*)\sqrt{n}(V_n - \theta), n \geq 2)$ converge en loi vers $\lambda^* Y + (1-\lambda^*)W$. Autrement dit $(\sqrt{n}(T_n^{\lambda_n^*} - \theta), n \geq 2)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \dfrac{2\theta^2}{1+2\theta})$. En remarquant que $T_n^{\lambda_n^*}$ converge presque sûrement vers θ, il résulte du théorème de Slutsky que $(\sqrt{n(1+2T_n^{\lambda_n^*})}\dfrac{T_n^{\lambda_n^*} - \theta}{\sqrt{2}T_n^{\lambda_n^*}}, n \geq 2)$ converge en loi vers une loi gaussienne $\mathcal{N}(0,1)$. Par un raisonnement similaire à celui utilisé dans les questions précédentes, on obtient l'intervalle de confiance de niveau asymtopique 95% :

$$\tilde{I}_n^* = [T_n^{\lambda_n^*} - 1.96\frac{\sqrt{2}T_n^{\lambda_n^*}}{\sqrt{n(1+2T_n^{\lambda_n^*})}}, T_n^{\lambda_n^*} + 1.96\frac{\sqrt{2}T_n^{\lambda_n^*}}{\sqrt{n(1+2T_n^{\lambda_n^*})}}]$$

On a $\lambda_n^* = 0.885$, $T_n^{\lambda_n^*} = 4.015$, $1.96\dfrac{\sqrt{2}T_n^{\lambda_n^*}}{\sqrt{n(1+2T_n^{\lambda_n^*})}} = 0.370$. On obtient l'intervalle $\tilde{I}_n^* = [3.645, 4.385]$.

▲

Chapitre VII

Simulation

VII.1 Énoncés

Exercice VII.1.
Soient X et Y deux variables aléatoires continues de densité continue f et g strictement positive. On considère les fonctions de répartitions $F(x) = \int_{-\infty}^{x} f(t)\,dt$ et $G(x) = \int_{-\infty}^{x} g(t)\,dt$.

1. Calculer la loi de $G(Y)$.
2. Calculer et reconnaître la loi de $F^{-1}(G(Y))$.
3. En déduire une méthode pour simuler la v.a. X (il s'agit de la méthode d'inversion de la fonction de répartition).

△

Exercice VII.2.
Soit $(U_n, n \in \mathbb{N}^*)$ une suite de v.a. indépendantes de loi uniforme sur $[0,1]$. Soit $\theta > 0$.

1. Donner la loi de $X_k = -\log(U_k)/\theta$.
2. Donner la loi de $\sum_{k=1}^{n} X_k$.
3. Calculer la loi de la v.a. N définie par $N = \inf\left\{n; \prod_{k=1}^{n+1} U_k < \mathrm{e}^{-\theta}\right\}$.
4. En déduire une méthode pour simuler des variables aléatoires de Poisson.

△

VII.2 Corrections

Exercice VII.1.

1. Les fonctions G et F sont des bijections de $\mathbb{R}$ dans $]0,1[$. En utilisant la fonction de répartition, on a pour $u \in]0,1[$,
$$\mathbb{P}(G(Y) \leq u) = \mathbb{P}(Y \leq G^{-1}(u)) = G(G^{-1}(u)) = u.$$
On en déduit que la loi de $G(Y)$ est la loi uniforme sur $[0,1]$. On en déduit que si U est une variable aléatoire de loi uniforme sur $[0,1]$, alors $G^{-1}(U)$ a même loi que Y. Cette méthode, dite méthode d'inversion de la fonction de répartition, permet de simuler Y à partir d'un générateur de nombres aléatoires de loi uniforme sur $[0,1]$.
2. En utilisant la fonction de répartition, on a pour $u \in \mathbb{R}$,
$$\mathbb{P}(F^{-1}(G(Y)) \leq u) = \mathbb{P}(Y \leq G^{-1}(F(u))) = G(G^{-1}(F(u))) = F(u).$$
La fonction de répartition de $F^{-1}(G(Y))$ est la fonction de répartition de X. On en déduit que $F^{-1}(G(Y))$ a même loi que X.
3. Ainsi, si U est de loi uniforme sur $[0,1]$, alors $F^{-1}(X)$ a même loi que X. Le générateur de nombres pseudo-aléatoires fournit une réalisation, $x_1, x_2, \ldots$, d'une suite de v.a. indépendantes de loi uniforme sur $[0,1]$. La suite $F^{-1}(x_1), F^{-1}(x_2), \ldots$, est donc la réalisation d'une suite de v.a. indépendantes de même loi que X.

▲

Exercice VII.2.

1. On calcule la fonction de répartition de $X_k = -\log(U_k)/\theta$, qui est une v.a. positive : soit $x > 0$, on a
$$\mathbb{P}(-\log(U_k)/\theta \leq x) = \mathbb{P}(U_k \geq \mathrm{e}^{-\theta x}) = 1 - \mathrm{e}^{-\theta x}\,.$$
On en déduit que la loi de X_k est la loi exponentielle de paramètre θ.
2. En utilisant les fonction caractéristique, on en déduit que la loi de $\sum_{k=1}^n X_k$ est la loi $\Gamma(\theta, n)$.
3. On a pour $n \geq 1$,
$$\begin{aligned}
\mathbb{P}(N = n) &= \mathbb{P}(\prod_{k=1}^{n} U_k \geq \mathrm{e}^{-\theta} > \prod_{k=1}^{n+1} U_k) \\
&= \mathbb{P}(\sum_{k=1}^{n} X_k \leq 1 < \sum_{k=1}^{n+1} X_k) \\
&= \mathbb{P}(1 < \sum_{k=1}^{n+1} X_k) - \mathbb{P}(1 < \sum_{k=1}^{n} X_k) \\
&= \int_1^\infty \frac{n!}{\theta^{n+1}} x^n \,\mathrm{e}^{-\theta x}\; dx - \int_1^\infty \frac{(n-1)!}{\theta^n} x^{n-1} \,\mathrm{e}^{-\theta x}\; dx \\
&= \Big[-\frac{n!}{\theta^n} x^n \,\mathrm{e}^{-\theta x} \Big]_1^\infty \\
&= \frac{n!}{\theta^n} \,\mathrm{e}^{-\theta}\,.
\end{aligned}$$

Pour $n = 0$, on a

$$\mathbb{P}(N = 0) = \mathbb{P}(U_1 \leq \mathrm{e}^{-\theta}) = \mathrm{e}^{-\theta}\,.$$

On en déduit donc que la loi de N est la loi de Poisson de paramètre θ.

4. Le générateur de nombres pseudo-aléatoires fournit une réalisation, $x_1, x_2, \ldots$, d'une suite de v.a. indépendantes de loi uniforme sur $[0, 1]$. On déduit de ce qui précède que $\inf\left\{n; \prod_{k=1}^{n+1} x_k < \mathrm{e}^{-\theta}\right\}$ est la réalisation d'une v.a. de loi de Poisson de paramètre θ.

▲

Chapitre VIII

Estimateurs

VIII.1 Énoncés

Exercice VIII.1.
Soit $X_1, \dots, X_n$ un échantillon indépendant de taille n d'une loi de moyenne θ et de variance finie σ^2. Trouver l'estimateur de θ de variance minimale dans la classe des estimateurs linéaires, $\hat{\theta} = \sum_{k=1}^{n} a_k X_k$, et sans biais.

△

Exercice VIII.2.
On considère le modèle d'échantillonnage $X_1, \dots, X_n$ de taille n associé à la famille de lois exponentielles $\mathcal{P} = \{\mathcal{E}(\lambda), \lambda > 0\}$. On veut estimer λ.

1. À partir de la méthode des moments, construire un estimateur convergent $\hat{\lambda}_n$ de λ.
2. Vérifier qu'il s'agit de l'estimateur du maximum de vraisemblance.
3. Déterminer la loi de $\sum_{i=1}^{n} X_i$. Calculer $\mathbb{E}_\lambda[\hat{\lambda}_n]$. L'estimateur est-il sans biais ?
4. Déterminer un estimateur $\hat{\lambda}_n^*$ sans biais et un estimateur $\hat{\lambda}_n^\circ$ qui minimise le risque quadratique parmi les estimateurs
$$\hat{\lambda}_n^{(c)} = \frac{c}{\sum_{i=1}^{n} X_i}, \quad \text{où} \quad c > 0.$$
5. Calculer le score, l'information de Fisher et la borne FDCR.
6. Les estimateurs étudiés font intervenir la statistique $S_n = \displaystyle\sum_{i=1}^{n} X_i$. Est-elle exhaustive et totale ?
7. Résumé : quelles propriétés $\hat{\lambda}_n^*$ a-t-il parmi les suivantes ?
   (a) Sans biais.
   (b) Optimal.
   (c) Efficace.
   (d) Préférable à $\hat{\lambda}_n$.
   (e) Inadmissible.

(f) Régulier.

(g) Asymptotiquement normal.

△

Exercice VIII.3.
On considère le modèle d'échantillonnage $X_1, \dots, X_n$ de taille n associé à la famille de lois de Poisson $\mathcal{P} = \{\mathcal{P}(\theta), \theta > 0\}$. On cherche à estimer $\mathbb{P}_\theta(X_i = 0)$.

1. Montrer que le modèle est exponentiel. Déterminer la statistique canonique S. Est-elle exhaustive et totale ? Donner sa loi.
2. Calculer $\mathbb{P}_\theta(X_i = 0)$ et montrer que $\mathbf{1}_{\{X_1=0\}}$ en est un estimateur sans biais.
3. Montrer que la loi conditionnelle de X_1 sachant S est une binomiale de paramètres $\left(S, \frac{1}{n}\right)$.
4. En déduire que $\delta_S = \left(1 - \frac{1}{n}\right)^S$ est l'estimateur optimal de $\mathbb{P}_\theta(X_i = 0)$. Est-il convergent ?
5. Calculer le score et l'information de Fisher.
6. En déduire la borne FDCR pour l'estimation de $\mathbb{P}_\theta(X_i = 0)$. Est-elle atteinte par δ_S ?

△

Exercice VIII.4.
On observe la réalisation d'un échantillon $X_1, \dots, X_n$ de taille n de loi P_θ de densité

$$f(x, \theta) = \frac{1}{\theta}(1 - x)^{\frac{1}{\theta} - 1} \mathbf{1}_{]0,1[}(x), \quad \theta \in \mathbb{R}_*^+.$$

1. Donner une statistique exhaustive. Est-elle totale ?
2. Déterminer l'estimateur du maximum de vraisemblance T_n de θ.
3. Montrer que $-\log(1 - X_i)$ suit une loi exponentielle dont on précisera le paramètre.
4. Calculer le biais et le risque quadratique de T_n. Cet estimateur est-il convergent, optimal, efficace ?
5. Étudier la limite en loi de $\sqrt{n}(T_n - \theta)$ quand $n \to \infty$.

△

Exercice VIII.5.
Soient Z et Y deux variables indépendantes suivant des lois exponentielles de paramètres respectifs $\lambda > 0$ et $\mu > 0$. On dispose d'un échantillon de variables aléatoires indépendantes $(Z_1, Y_1), \dots, (Z_n, Y_n)$ de même loi que (Z, Y).

1. Calculer la loi de $\sum_{i=1}^{n} Z_i$.
2. S'agit-il d'un modèle exponentiel ? Si oui, peut-on exhiber une statistique exhaustive ?
3. Calculer l'estimateur du maximum de vraisemblance $(\hat{\lambda}_n, \hat{\mu}_n)$ de (λ, μ).

4. Montrer qu'il est asymptotiquement normal et déterminer sa matrice de covariance asymptotique.

On suppose dorénavant que l'on observe seulement $X_i = \min(Z_i, Y_i)$ pour $i \in \{1, \ldots, n\}$.

5. Calculer la fonction de répartition de la variable X_i.
6. Écrire le modèle statistique correspondant. Le modèle est-il identifiable ? Quelle fonction de (λ, μ) est identifiable ?
7. Quels sont les estimateurs du maximum de vraisemblance de $\gamma = \lambda + \mu$ fondés sur les observations
   (a) de $X_1, \ldots, X_n$,
   (b) de $(Z_1, Y_1), \ldots, (Z_n, Y_n)$?
   Est-il naturel que ces estimateurs soient différents ?
8. Comparer les propriétés asymptotiques de ces estimateurs.

△

Exercice VIII.6.
Une machine produit N micro-chips par jour, N connu. Chacun d'entre eux a un défaut avec la même probabilité θ inconnue. On cherche à estimer la probabilité d'avoir au plus k défauts sur un jour. À ce propos, on teste tous les micro-chips pendant une période de n jours et on retient chaque jour le nombre de défauts.

1. Choisir un modèle. Est-ce un modèle exponentiel ?
2. Déterminer une statistique S exhaustive et totale. Calculer sa loi.
3. Construire un estimateur δ sans biais qui ne fait intervenir que les données du premier jour.
4. En déduire un estimateur optimal δ_S. Qu'est-ce qu'on observe quand on fait varier k ?

△

Exercice VIII.7.
Soit X une variable aléatoire à valeurs dans $\mathbb{N}^*$ définie comme l'instant de premier succès dans un schéma de Bernoulli de paramètre $q \in]0, 1[$.

1. Vérifier que la loi de X est une loi géométrique dont on précisera le paramètre.
2. Vérifier qu'il s'agit d'un modèle exponentiel. Donner une statistique exhaustive.
3. Déterminer $I(q)$, l'information de Fisher sur q d'un échantillon de taille 1.

Soit $X_1, \ldots, X_n$ un échantillon indépendant de taille n de même loi que X.

4. Déterminer $\hat{q}_n$, l'estimateur du maximum de vraisemblance de q.
5. Montrer que l'estimateur du maximum de vraisemblance est asymptotiquement normal.
6. Donner un intervalle de confiance pour q de niveau $1 - \alpha$.

Une société de transport en commun par bus veut estimer le nombre de passagers ne validant pas leur titre de transport sur une ligne de bus déterminée. Elle dispose pour cela, pour un jour de semaine moyen, du nombre n_0 de tickets compostés sur la ligne et des résultats de l'enquête suivante : à chacun des arrêts de bus de la ligne, des contrôleurs comptent le nombre de passagers sortant des bus et ayant validé leur ticket jusqu'à la sortie du premier fraudeur. Celui-ci étant inclus on a les données suivantes :

44	09	11	59	81	44	19	89	10	24
07	21	90	38	01	15	22	29	19	37
26	219	02	57	11	34	69	12	21	28
34	05	07	15	06	129	14	18	02	156

7. Estimer la probabilité de fraude. Donner un intervalle de confiance de niveau 95%. Estimer le nombre de fraudeur n_f si $n_0 = 20\,000$.

△

Exercice VIII.8.
Le total des ventes mensuelles d'un produit dans un magasin $i \in \{1, \dots, n\}$ peut être modélisé par une variable aléatoire de loi normale $\mathcal{N}\left(m_i, \sigma^2\right)$. On suppose les constantes $m_i > 0$ et $\sigma > 0$ connus. Une campagne publicitaire est menée afin de permettre l'augmentation des ventes. On note X_i la vente mensuelle du magasin i après la campagen publicitaire. On suppose que les variables X_i sont indépendantes.

1. On suppose que l'augmentation des ventes se traduit par une augmentation de chacune des moyennes m_i d'une quantité α. Déterminer l'estimateur du maximum de vraisemblance de α. Donner sa loi et ses propriétés.
2. On suppose que l'augmentation des ventes se traduit par une multiplication de chacune des moyennes m_i par une quantité β. On considère l'estimateur de β :

$$\tilde{\beta}_n = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i}{m_i}.$$

   Donner sa loi et ses propriétés à horizon fini.
3. Déterminer l'estimateur du maximum de vraisemblance de β. Donner sa loi et ses propriétés à horizon fini.
4. Application numérique aux données simulées du tableau VIII.1, avec $n = 15$ et $\sigma = 12$:

m_i	1023	981	1034	1007	988	1021	1005	995
x_i	1109	1075	1129	1123	1092	1087	1129	1122
m_i	1020	1013	1030	1046	1003	1039	968	
x_i	1105	1124	1103	1072	1065	1069	1098	

TAB. VIII.1 – Effet (simulé) d'une campagne publicitaire

△

Exercice VIII.9.
La hauteur maximale H de la crue annuelle d'un fleuve est observée car une crue supérieure à 6 mètres serait catastrophique. On a modélisé H comme une variable de Rayleigh, i.e. H a une densité donnée par

$$f_H(x) = \mathbf{1}_{\mathbb{R}_+}(x) \frac{x}{a} \exp\left(-\frac{x^2}{2a}\right),$$

où $a > 0$ est un paramètre inconnu. Durant une période de 8 ans, on a observé les hauteurs de crue suivantes en mètres :

2.5 1.8 2.9 0.9 2.1 1.7 2.2 2.8

1. Donner l'estimateur du maximum de vraisemblance, $\hat{a}_n$, de a.
2. Quelles propriétés $\hat{a}_n$ possède-t-il parmi les suivantes ?
   (a) Sans biais.
   (b) Optimal.
   (c) Efficace.
   (d) Asymptotiquement normal.
3. Une compagnie d'assurance estime qu'une catastrophe n'arrive qu'au plus une fois tous les mille ans. Ceci peut-il être justifié par les observations ?

△

Exercice VIII.10.
Soient $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées de loi de Bernoulli $p \in [0, 1]$. On pose $\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$.

1. Montrer l'inégalité $\mathrm{Var}_p(\hat{p}_n) \leq \frac{1}{4n}$.
2. Un institut de sondage souhaite estimer avec une précision de 3 points (à droite et à gauche) la probabilité qu'un individu vote pour le maire actuel aux prochaines elections. Combien de personnes est-il nécessaire de sonder ?
3. Sur un échantillon représentatif de 1000 personnes, on rapporte les avis favorables pour un homme politique. En novembre, il y avait 38% d'avis favorables, en décembre 36%. Un éditorialiste dans son journal prend très au sérieux cette chute de 2 points d'un futur candidat à la présidentielle ! Confirmer ou infirmer la position du journaliste.

△

Exercice VIII.11.
Dans l'industrie agroalimentaire, on s'intéresse à la détection de la contamination[1] du lait par un micro-organisme : les spores de *clostridia*. Cette bactérie, naturellement présente dans les organismes humains et animaux, peut causer des maladies chez les individus fragiles, qui peuvent même être mortelles. Le lait ne figure pas parmi les premiers aliments à risque mais il est très important de contrôler une éventuelle contamination.

Deux problèmes importants se posent :
– On ne dispose pas de l'observation du nombre de micro-organismes présents mais seulement de l'indication présence-absence.
– Sans connaissance a priori de l'ordre de grandeur du taux de contamination, l'estimation du taux risque de ne donner aucun résultat exploitable.

[1] La méthode présentée, MPN ("most probable number"), est une méthode largement utilisée pour détecter des contaminations dans l'agroalimentaire, mais également en environement (rivière, ...).

L'expérience menée consiste à observer la présence-absence de ces spores dans un tube de 1ml de lait, le but étant au final d'estimer la densité en spores : λ (nombre de spores par unité de volume).

Soit Z_k la variable aléatoire désignant le nombre (non observé) de spores présents dans le tube k et $X_k = \mathbf{1}_{\{Z_k=0\}}$ la variable aléatoire valant 1 s'il n'y a pas de spore dans le tube k et 0 sinon. On suppose que Z_k suit une loi de Poisson de paramètre λ.

On effectue une analyse sur n tubes indépendants et $Y = \sum_{k=1}^{n} X_k$ donne le nombre de tubes stériles (négatifs).

1. Donner les lois de X_k et Y. On notera $\pi = \mathbb{P}(X_k = 1)$.
2. Donner l'estimateur du maximum de vraisemblance de π. En déduire l'estimateur du maximum de vraisemblance de λ.
3. Donner les intervalles de confiance de π et λ.
4. Donner les résultats numériques des deux questions précédentes lorsqu'on observe 6 tubes stériles sur 10 au total. Pour les intervalles de confiance, on se placera au niveau $\alpha = 5\%$.
5. Indiquer quels sont les problèmes lorsque la densité λ est très faible ou très forte.

Des densités extrêmes induisant des problèmes d'estimation, on va utiliser le principe de la dilution :

- Si on craint de n'observer que des tubes positifs, on ajoute des expériences sur des tubes dilués. Dans un tube dilué d fois, la densité en spores est égale à λ/d, et le nombre de spores suit une loi de Poisson de paramètre λ/d.
- Si on craint de n'observer que des tubes négatifs, on effectue l'analyse sur de plus grands volumes. Dans un volume d fois plus grand, le nombre de spores suit une loi de Poisson de paramètre λd.

Pour utiliser cette méthode on doit avoir une idée a priori de l'ordre de grandeur de la densité.

Considérons N échantillons contenant chacun n_i tubes avec un taux de dilution égal à d_i (avec $i \in \{1, ..., N\}$). On note Y_i le nombre de tubes négatifs du i-ème échantillon.

6. Donner la loi de Y_i.
7. Donner l'équation qui détermine l'estimateur du maximum de vraisemblance de λ.
8. Que vaut la variance asymptotique de cet estimateur ?
9. On étudie le cas particulier d'un petit nombre de dilutions. Donner le résultat formel lorsque $N = 2$, $d_1 = 1$, $d_2 = \dfrac{1}{2}$. Donner les résultats numériques, estimation et intervalle de confiance de λ, si on observe $y_1 = 3$ et $y_2 = 6$ (pour $n_1 = n_2 = 10$).

△

VIII.2 Corrections

Exercice VIII.1.

Un estimateur linéaire et sans biais de θ s'écrit sous la forme $\hat{\theta}_n = \sum_{i=1}^n a_i X_i$ avec $\sum_{i=1}^n a_i = 1$. Les variables alátoires étant indépendantes $\mathrm{Var}(\hat{\theta}_n) = \sum_{i=1}^n a_i^2 \sigma^2$. D'après l'inégalité de Cauchy-Schwarz on a $(\sum_{i=1}^n a_i^2)(\sum_{i=1}^n \frac{1}{n^2}) \geq \left(\sum_{i=1}^n \frac{a_i}{n}\right)^2$. Puisque $(\sum_{i=1}^n a_i = 1)$, on déduit $\sum_{i=1}^n a_i^2 \geq \frac{1}{n}$ avec égalité si et seulement si $a_i = \frac{1}{n}$ pour tout $i \in \{1, \ldots, n\}$. L'estimateur de variance minimale dans la classe des estimateurs linéaires et sans biais est donc la moyenne empirique.

▲

Exercice VIII.2.

1. La méthode des moments pour $l(x) = x$ est fondée sur la loi forte des grands nombres qui nous donne un estimateur convergent pour $\mathbb{E}_\lambda[l(X_1)] = 1/\lambda = m(\lambda)$

$$\delta(X) = \frac{1}{n}\sum_{i=1}^n X_i \to \frac{1}{\lambda} \quad \mathbb{P}_\lambda - \text{p.s.}$$

   L'estimateur convergent de λ qu'on en déduit est

$$\hat{\lambda}_n = m^{-1}(\delta(X)) = \frac{n}{\sum_{i=1}^n X_i}.$$

2. La log-vraisemblance est donnée par

$$L_n(x;\lambda) = \sum_{i=1}^n \log\left(\lambda e^{-\lambda x_i}\mathbf{1}_{\{x_i>0\}}\right) = n\log\lambda - \lambda\sum_{i=1}^n x_i + \log\prod_{i=1}^n \mathbf{1}_{\{x_i>0\}}.$$

   Pour calculer l'estimateur du maximum de vraisemblance, on cherche les zéros de la dérivée de la log-vraisemblance L_n :

$$\frac{\partial}{\partial\lambda}L_n(x;\lambda) = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \quad \iff \lambda = \frac{n}{\sum_{i=1}^n x_i}.$$

   Comme $\lim_{\lambda\to 0} L_n(x;\lambda) = \lim_{\lambda\to\infty} L_n(x;\lambda) = -\infty$, on en déduit que la log-vraisemblance est maximale pour $\lambda = \frac{n}{\sum_{i=1}^n x_i}$. L'estimateur du maximum de vraisemblance est donc $\hat{\lambda}_n$.

3. En utilisant les fonctions caractéristiques, on vérifie que la loi de $\sum_{i=1}^n X_i$ est la loi $\Gamma(n,\lambda)$ sous $\mathbb{P}_\lambda$. On obtient pour $n \geq 2$,

$$\begin{aligned}\mathbb{E}_\lambda\left[\frac{1}{\sum_{i=1}^n X_i}\right] &= \int_0^\infty \frac{1}{s}\frac{\lambda^n}{(n-1)!}s^{n-1}e^{-\lambda s}ds \\ &= \frac{\lambda}{n-1}\int_0^\infty \frac{\lambda^{n-1}}{(n-2)!}s^{n-2}e^{-\lambda s}ds \\ &= \frac{\lambda}{n-1},\end{aligned}$$

   où on a identifié la densité de $\Gamma(n-1,\lambda)$ sous l'intégrale. Donc on a $\mathbb{E}_\lambda[\hat{\lambda}_n] = \frac{n}{n-1}\lambda$. L'estimateur $\hat{\lambda}_n$ est donc biaisé pour $n \geq 2$. (Pour $n = 1$, on vérifie que $\mathbb{E}_\lambda[\hat{\lambda}_1] = \infty$.)

4. Les calculs précédents nécessitent de supposer $n \geq 2$ et donnent comme estimateur sans biais alors

$$\hat{\lambda}_n^* = \frac{n-1}{\sum_{i=1}^n X_i}.$$

On calcule le deuxième moment de la même manière que le premier, pour $n \geq 3$

$$\mathbb{E}_\lambda \left[\frac{1}{\left(\sum_{i=1}^n X_i\right)^2} \right] = \frac{\lambda^2}{(n-1)(n-2)} \int_0^\infty \frac{\lambda^{n-2}}{(n-3)!} s^{n-3} e^{-\lambda s} ds = \frac{\lambda^2}{(n-1)(n-2)}.$$

Donc pour tout $c > 0$, on a

$$\begin{aligned} R(\hat{\lambda}_n^{(c)}, \lambda) &= \mathbb{E}_\lambda \left(\left(\frac{c}{\sum_{i=0}^n X_i} - \lambda \right)^2 \right) \\ &= \frac{\lambda^2}{(n-1)(n-2)} \left(c^2 - 2(n-2)c + (n-1)(n-2) \right). \end{aligned}$$

Le risque quadratique est minimal pour $2c - 2(n-2) = 0$ soit $c = n-2$. Parmi les estimateurs $\hat{\lambda}^{(c)}$, c'est donc $\hat{\lambda}^\circ = \hat{\lambda}^{(n-2)}$ qui minimise le risque quadratique. Pour $n \leq 2$ le risque quadratique est infini.

5. Notons d'abord que le modèle est régulier. En effet, le support de toutes les lois exponentielles est $(0, \infty)$ et est donc indépendant du paramètre λ ; la fonction de vraisemblance (densité) est de classe C^2 (en λ) ; on peut vérifier des formules d'interversion entre l'intégrale en x et la differentiation en λ ; on verra dans la suite que l'information de Fisher existe. Par définition on a

$$L_1(x_1; \lambda) = \log(\lambda) - \lambda x_1 + \log(\mathbf{1}_{\{x_1 > 0\}}).$$

Il vient

$$\frac{\partial}{\partial \lambda} L_1(x_1; \lambda) = \frac{1}{\lambda} - x_1 \quad \text{et} \quad \frac{\partial^2}{\partial \lambda \partial \lambda} L_1(x_1; \lambda) = -\frac{1}{\lambda^2}.$$

On a donc

$$I(\lambda) = -\mathbb{E}_\lambda \left[\frac{\partial^2}{\partial \lambda \partial \lambda} L_1(X_1; \lambda) \right] = \frac{1}{\lambda^2} \quad \text{et} \quad FDCR(\lambda) = \frac{1}{nI(\lambda)} = \frac{\lambda^2}{n}.$$

6. On vérifie que le modèle est exponentiel en regardant la densité jointe de $(X_1, \ldots, X_n)$

$$p_n(\lambda, x) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i \right) = C(\lambda) h(x) e^{Q(\lambda) S(x)}$$

où $C(\lambda) = \lambda^n$, $h(x) = \mathbf{1}_{\{x>0\}}$, $Q(\lambda) = -\lambda$ et $S(x) = \sum_{i=1}^n x_i$. On en déduit que $S_n = S(X) = \sum_{i=1}^n X_i$ est la statistique canonique. Elle est donc exhaustive et totale.

7. On suppose $n \geq 3$.
   (a) $\hat{\lambda}_n^*$ a été choisi sans biais.
   (b) $\hat{\lambda}_n^*$ est optimal car il est fonction de la statistique exhaustive et totale $S(X)$ (Théorème de Lehman-Sheffé).

(c) L'estimateur sans biais $\hat{\lambda}_n^*$ a comme risque quadratique

$$\mathrm{Var}_\lambda(\hat{\lambda}_n^*) = R(\hat{\lambda}_n^*, \lambda) = \frac{\lambda^2}{(n-1)(n-2)}\left((n-1)^2 - (n-1)(n-2)\right) = \frac{\lambda^2}{n-2}.$$

Il n'atteint donc pas la borne FDCR. Il n'est donc pas efficace. Il n'existe pas d'estimateur efficace, parceque tout estimateur efficace serait optimal et alors (p.s.) égal à $\hat{\lambda}^*$ par l'unicité (p.s.) de l'estimateur optimal.

(d) $\hat{\lambda}_n^*$ est préférable à $\hat{\lambda}_n$ car $R(\hat{\lambda}_n^*, \lambda) < R(\hat{\lambda}_n, \lambda)$ pour tout $\lambda > 0$.

(e) $\hat{\lambda}_n^*$ est inadmissible car $R(\hat{\lambda}_n^\circ, \lambda) < R(\hat{\lambda}_n^*, \lambda)$ pour tout $\lambda > 0$.

(f) $\hat{\lambda}_n^*$ est régulier parce qu'il est de carré intégrable et le modèle est régulier (on peut vérifier que $\hat{\lambda}_n^*$ satisfait les propriétés d'interversion de l'intégrale en x et de la differentiation en λ).

(g) On est dans un modèle régulier et identifiable. L'estimateur du maximum de vraisemblance $\hat{\lambda}_n$ est asymptotiquement efficace, i.e. $\sqrt{n}\left(\hat{\lambda}_n - \lambda\right)$ converge en loi vers une loi normale centrée de variance $1/I(\lambda)$. En remarquant que $\sqrt{n}(\hat{\lambda}_n^* - \hat{\lambda}_n)$ tend vers 0 p.s., le Théorème de Slutsky entraîne que $\hat{\lambda}_n^*$ aussi est asymptotiquement normal de variance $1/I(\lambda)$.

▲

Exercice VIII.3.

1. Écrivons

$$\mathbb{P}_\theta(X_1 = k_1, \ldots, X_n = k_n) = e^{-n\theta}\left(\prod_{i=1}^n \frac{1}{k_i!}\right)\exp\left(\log(\theta)\sum_{i=1}^n k_i\right)$$

et identifions la statistique canonique $S(X) = \sum_{i=1}^n X_i$. La statistique canonique d'un modèle exponentiel est toujours exhaustive et totale. La variable aléatoire $S(X)$ suit une loi de Poisson de paramètre $n\theta$ sous $\mathbb{P}_\theta$.

2. On a $\mathbb{P}_\theta(X_i = 0) = e^{-\theta}$ et $\mathbb{E}_\theta[\mathbf{1}_{\{X_1=0\}}] = \mathbb{P}_\theta(X_1 = 0) = e^{-\theta}$.

3. On calcule pour tout $k, s \in \mathbb{N}$ les probabilités conditionnelles élémentaires

$$\begin{aligned}
\mathbb{P}_\theta(X_1 = k | S = s) &= \frac{\mathbb{P}_\theta(X_1 = k, S - X_1 = s - k)}{\mathbb{P}_\theta(S = s)} \\
&= \frac{e^{-\theta}\theta^k/k! e^{-(n-1)\theta}((n-1)\theta)^{s-k}/(s-k)!}{e^{-n\theta}(n\theta)^s/s!} \\
&= C_n^k\left(\frac{n-1}{n}\right)^{s-k}\left(\frac{1}{n}\right)^k
\end{aligned}$$

et on identifie les probabilités binomiales de paramètres $\left(s, \frac{1}{n}\right)$. La loi de X_1 conditionnellement à S est donc la loi binomiale de paramètre $(S, 1/n)$.

4. On applique le Théorème de Lehman-Sheffé pour $\delta = \mathbf{1}_{\{X_1=0\}}$ et on calcule l'estimateur optimal δ_S par la méthode de Rao-Blackwell :

$$\delta_S(X) = \mathbb{E}[\delta(X)|S(X)] = \mathbb{E}[\mathbf{1}_{\{X_1=0\}}|S(X)] = \mathbb{P}(X_1 = 0|S(X)) = \left(1 - \frac{1}{n}\right)^{S(X)}.$$

Par la loi forte des grands nombres on a $S_n/n \to \theta$, en particulier $S_n \to \infty$, et alors

$$\delta_{S_n} = \left(1 - \frac{1}{n}\right)^{S_n} = \left(1 - \frac{S_n/n}{S_n}\right)^{S_n} \to e^{-\theta}$$

en utilisant $(1 - x_m/m)^m \to e^{-\lim_m x_m}$. Ainsi, δ_S est un estimateur convergent de $\mathrm{e}^{-\theta}$.

5. On vérifie la régularité du modèle (support, dérivabilité, interchangeabilité et existence de $I(\theta)$), et on calcule

$$\begin{aligned} V_\theta(k) &= \frac{\partial}{\partial\theta} \log\left(e^{-\theta}\theta^k/k!\right) = \frac{\partial}{\partial\theta}\left(k\log(\theta)\right) - \theta) = \frac{k}{\theta} - 1 \\ I(\theta) &= -\mathbb{E}_\theta\left[\frac{\partial}{\partial\theta} V_\theta(X_1)\right] = \frac{\mathbb{E}_\theta[X_1]}{\theta^2} = \frac{1}{\theta} \end{aligned}$$

6. On en déduit

$$FDCR(e^{-\theta}) = \left(\frac{\partial}{\partial\theta} e^{-\theta}\right)^2 \frac{1}{nI(\theta)} = \frac{\theta e^{-2\theta}}{n}.$$

Pour calculer la variance de δ_S on se souvient de la fonction génératrice $\mathbb{E}(z^Y) = e^{-\lambda(1-z)}$ d'une variable Poisson Y de paramètre λ :

$$\begin{aligned} \mathrm{Var}_\theta(\delta_S) &= \mathbb{E}_\theta\left[\left(\left(1 - \frac{1}{n}\right)^2\right)^S\right] - e^{-2\theta} \\ &= \exp\left(-n\theta\left(1 - \left(1 - \frac{1}{n}\right)^2\right)\right) - e^{-2\theta} = e^{-2\theta}\left(e^{\theta/n} - 1\right). \end{aligned}$$

La borne FDCR n'est donc pas atteinte. L'estimateur δ_S est donc optimal mais pas efficace.

▲

Exercice VIII.4 .

1. La densité peut s'écrire

$$p(x,\theta) = \frac{1}{\theta}(1-x)^{-1}\exp(\frac{1}{\theta}\log(1-x))\mathbf{1}_{[0,1]}(x), \quad \theta \in \mathbb{R}_*^+.$$

Le modèle $\mathcal{P} = \{P_\theta, \theta > 0\}$ forme un modèle exponentiel. La variable aléatoire $S_n = \sum_{i=1}^n \log(1 - X_i)$ est la statistique canonique. Elle est exhaustive et totale.

2. La log-vraisemblance du modèle est donnée par

$$L_n(x_1,\dots,x_n;\theta) = (\frac{1}{\theta}-1)\sum_{i=1}^{n}\log(1-x_i) - n\log(\theta) + \prod_{i=1}^{n}\log(\mathbf{1}_{[0,1]}(x_i)).$$

La log-vraisemblance vaut $-\infty$ sur la frontière de $]0,1[$. Son maximum est donc atteint au point θ qui annule sa dérivée. On obtient l'estimateur du maximun de vraisemblance : $T_n = \frac{1}{n}\sum_{i=1}^{n} -\log(1-X_i)$.

3. En utilisant la méthode de la fonction muette et le changement de variable $y = -\log(1-x)$, on obtient pour une fonction h mesurable bornée

$$\mathbb{E}_\theta[h(-\log(1-X_i))] = \int_0^1 h(-\log(1-x))\frac{1}{\theta}(1-x)^{\frac{1}{\theta}-1}dx = \int_0^\infty h(y)\frac{1}{\theta}\exp(-\frac{y}{\theta})dy.$$

On en déduit donc que $-\log(1-X_i)$ suit une loi exponentielle de paramètre $\frac{1}{\theta}$

4. (a) L'estimateur est sans biais car $\mathbb{E}_\theta[T_n] = \mathbb{E}_\theta[-\log(1-X_1)] = \theta$.

   (b) Le risque quadratique est

$$R(T_n,\theta) = \mathbb{E}_\theta[T_n-\theta]^2 = \mathrm{Var}_\theta(T_n) = \frac{\theta^2}{n}.$$

   (c) Les variables aléatoires $(-\log(1-X_i), i \geq 1)$ sont indépendantes, de même loi et intégrables. La loi forte des grands nombres assure que la suite $(T_n, n > 1)$ converge presque sûrement vers θ.

   (d) La statistique T_n est une fonction de la statistique S_n qui est exhaustive et totale. En appliquant les théorèmes de Rao-Blackwell et de Lehman-Shefé, on en déduit que T_n est un estimateur optimal de θ.

   (e) On calcule l'information de Fisher

$$I_1(\theta) = -\mathbb{E}[\frac{\partial^2}{\partial\theta^2}L_1(X_1,\theta)] = -\frac{2}{\theta^3}\mathbb{E}[\log(1-X_1)] - \frac{1}{\theta^2} = \frac{1}{\theta^2}.$$

   On a donc $I_n(\theta) = nI_1(\theta) = n/\theta^2$ et $R(T_n,\theta) = 1/nI_1(\theta)$. L'estimateur T_n est donc efficace.

5. Remarquons que les variables aléatoires $(-\log(1-X_i), i \geq 1)$ sont également de carré intégrable. Par le théorème central limite, la suite $(\sqrt{n}(T_n-\theta), n \geq 1)$ est asymptotiquement normale de variance asymptotique θ^2. On peut aussi dire que le modèle est régulier et identifiable, donc l'estimateur du maximum de vraisemblance est asymptotiquement normal de variance l'inverse de l'information de Fisher, $1/I_1(\theta) = \theta^2$.

▲

Exercice VIII.5.

1. En utilisant les fonctions caractéristiques, on obtient que $\sum_{i=1}^{n} Z_i$ suit une loi Gamma $\Gamma(\lambda, n)$.

2. La densité du couple (Z_1, Y_1) est $p_1(z_1, y_1; \lambda, \mu) = \lambda\mu\, \mathrm{e}^{(-\lambda z_1 - \mu y_1)}\, \mathbf{1}_{\{z_1>0, y_1>0\}}$. Il s'agit d'un modèle exponentiel. La vraisemblance de l'échantillon de taille n vaut pour $z = (z_1, \ldots, z_n)$ et $y = (y_1, \ldots, y_n)$

$$p_n(z, y; \lambda\mu) = \lambda^n \mu^n\, \mathrm{e}^{\left(-\lambda \sum_{i=1}^n z_i - \mu \sum_{i=1}^n y_i\right)} \left(\prod_{i=1}^n \mathbf{1}_{\{z_i>0, y_i>0\}}\right).$$

Une statistique exhaustive (et totale) est $T = \left(\sum_{i=1}^n Z_i, \sum_{i=1}^n Y_i\right)$.

3. La log-vraisemblance vaut

$$L_n(z, y; \lambda, \mu) = n\log(\lambda) + n\log(\mu) - \lambda \sum_{i=1}^n z_i - \mu \sum_{i=1}^n y_i + \log\left(\prod_{i=1}^n \mathbf{1}_{\{z_i>0, y_i>0\}}\right).$$

Si on dérive cette log-vraisemblance par rapport à λ on obtient que $\partial_\lambda L_n(z, y; \lambda, \mu) = 0$ implique $\lambda = n/\sum_{i=1}^n z_i$. Comme cette fonction est concave, le maximum de vraisemblance est bien obtenu pour $\lambda = n/\sum_{i=1}^n z_i$. L'estimateur du maximum de vraisemblance est donc $\hat{\lambda}_n = n/\sum_{i=1}^n Z_i$. En utilisant les mêmes arguments on trouve $\hat{\mu}_n = n/\sum_{i=1}^n Y_i$.

4. L'estimateur $(\hat{\lambda}_n, \hat{\mu}_n)$ est asymptotiquement normal car il s'agit de l'estimateur du maximum de vraisemblance dans un modèle exponentiel (modèle régulier et identifiable). De plus la loi normale asymptotique est de moyenne nulle et de variance l'inverse de l'information de Fisher. Un rapide calcul donne

$$\frac{\partial^2 L_n(z, y; \lambda, \mu)}{\partial\lambda^2} = -\frac{n}{\lambda^2}, \quad \frac{\partial^2 L_n(z, y; \lambda, \mu)}{\partial\lambda\partial\mu} = 0 \quad \text{et} \quad \frac{\partial^2 L_n(z, y; \lambda, \mu)}{\partial\mu^2} = -\frac{n}{\mu^2}.$$

L'information de Fisher vaut donc $\begin{pmatrix} 1/\lambda^2 & 0 \\ 0 & 1/\mu^2 \end{pmatrix}$. On en déduit donc que

$$\sqrt{n}\begin{pmatrix} \hat{\lambda}_n - \lambda \\ \hat{\mu}_n - \mu \end{pmatrix} \xrightarrow[n\to\infty]{\text{Loi}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda^2 & 0 \\ 0 & \mu^2 \end{pmatrix}\right).$$

5. Par indépendance, on calcule

$$\mathbb{P}(X_i > t) = \mathbb{P}(Z_i > t, Y_i > t) = \mathbb{P}(Z_i > t)\mathbb{P}(Y_i > t) = \exp(-(\lambda+\mu)t),$$

si $t > 0$ et $\mathbb{P}(X_i > t) = 1$ sinon. Donc la fonction de répartition de X_i est $F(t) = (1 - \exp(-(\lambda+\mu)t))\mathbf{1}_{\{t>0\}}$. La loi de X_i est la loi exponentielle de paramètre $\gamma = \lambda + \mu$.

6. Le modèle statistique est $\mathcal{P} = \{\mathcal{E}(\lambda+\mu), \lambda > 0, \mu > 0\}$. Il n'est pas identifiable car si $\lambda + \mu = \lambda' + \mu'$, alors la loi de X_i est la même. En revanche le modèle $\mathcal{P} = \{\mathcal{E}(\gamma), \gamma > 0\}$ est identifiable en $\gamma = \lambda + \mu$: la vraisemblance de l'échantillon est

$$\prod_{i=1}^n p(x_i; \gamma) = \prod_{i=1}^n \left[\gamma \exp(-\gamma x_i)\mathbf{1}_{\{x_i>0\}}\right].$$

7. Le premier est l'estimateur du maximum de vraisemblance du modèle ci-dessus et par analogie à la question 3, on trouve $\hat{\gamma}_n = n/\sum_{i=1}^{n} X_i$. Pour le deuxième, on trouve

$$\hat{\lambda}_n + \hat{\mu}_n = \frac{n}{\sum_{i=1}^n Z_i} + \frac{n}{\sum_{i=1}^n Y_i}.$$

Les estimateurs sont différents car fondés sur des observations différentes. Ces dernières sont plus riches que les premières.

8. Comme pour la question 4, on a

$$\sqrt{n}(\hat{\gamma}_n - \lambda - \mu) \xrightarrow[n\to\infty]{\text{Loi}} \mathcal{N}(0, (\lambda+\mu)^2).$$

On déduit de la question 4 que

$$\sqrt{n}(\hat{\lambda}_n + \hat{\mu}_n - \lambda - \mu) \xrightarrow[n\to\infty]{\text{Loi}} \mathcal{N}(0, \lambda^2+\mu^2).$$

Comme les deux paramètres sont strictement positifs, on a $(\lambda+\mu)^2 > \lambda^2+\mu^2$. L'estimateur $\hat{\lambda}_n + \hat{\mu}_n$ est toujours préférable à $\hat{\gamma}_n$. Cela correspond bien à l'intuition car l'estimateur $\hat{\lambda}_n + \hat{\mu}_n$ repose sur des observations plus détaillées.

▲

Exercice VIII.6.

1. Chaque micro-chip produit peut être représenté par une variable aléatoire de Bernoulli de paramètre θ. Mais, en groupant par jour, on ne retient que les sommes quotidiennes, ce qui est le nombre de défauts X_i de chaque jour $i = 1, \ldots, n$. Les X_i sont alors indépendantes et identiquement distribuées selon une loi binomiale de mêmes paramètres (N, θ), N connu, $\theta \in (0,1)$ inconnu. On peut écrire

$$\mathbb{P}(X_i = k) = C_N^k \theta^k (1-\theta)^{N-k} = (1-\theta)^N C_N^k \exp\left(\log\left(\frac{\theta}{1-\theta}\right) k\right)$$

et conclure qu'il s'agit d'un modèle exponentiel.

2. On identifie la statistique canonique $S(X) = \sum_{i=1}^n X_i$ du modèle qui est toujours exhaustive et totale. Elle suit une loi binomiale de paramètres (Nn, θ).
3. L'estimateur $\delta = \mathbf{1}_{\{X_1 \le k\}}$ est un estimateur sans biais de $\mathbb{P}_\theta(X_i \le k)$.
4. On utilise l'amélioration de Rao-Blackwell. D'abord on a

$$\begin{aligned}
\mathbb{P}_\theta(X_1 = k | S = s) &= \frac{\mathbb{P}_\theta(X_1 = k, S - X_1 = s-k)}{\mathbb{P}_\theta(S = s)} \\
&= \frac{C_N^k \theta^k (1-\theta)^{N-k} C_{N(n-1)}^{s-k} \theta^{s-k} (1-\theta)^{N(n-1)-(s-k)}}{C_{Nn}^s \theta^s (1-\theta)^{Nn-s}} \\
&= \frac{C_N^k C_{Nn-N}^{s-k}}{C_{Nn}^s}
\end{aligned}$$

On appelle cette loi la loi hypergéométrique de paramètres (Nn, s, N). On peut l'interprêter dans le cadre d'un modèle d'urne de la manière suivante : Imaginons une urne qui contient m boules dont b blanches et $m-b$ noires. On tire n fois sans remise. Soit Y le nombre de boules blanches tirées. Alors, Y est une variable hypergéométrique de paramètres (m, b, n). A vrai dire, on ne fait rien d'autre dans notre exercice : supposons, que le nombre total de défauts est fixé à $b = s$ sur les $m = Nn$ micro-chips produits, on demande la probabilité que parmi N micro-chips choisis au hasard (sans remise) il y a k défauts.
Donc, on obtient

$$\delta_S = \mathbb{E}(\delta|S) = \sum_{j=0}^{k} \mathbb{P}(X_1 = j|S) = \sum_{j=0}^{k} \frac{C_N^j C_{Nn-N}^{S-j}}{C_{Nn}^S}$$

et, d'après le théoreme de Lehman-Sheffé, δ_S est optimal.
Lorsque k varie, $\delta_S(X, k)$ est la valeur en k de la fonction de répartition d'une loi hypergéométrique de paramètres (Nn, S, N).

▲

Exercice VIII.7.

1. Soit $x \in \mathbb{N}^*$, $\mathbb{P}_q(X = x) = (1-q)^{x-1}q$. La loi de X est la loi géométrique de paramètre q.
2. La vraisemblance est $p(x; q) = (1-q)^{x-1}q$. Il s'agit d'un modèle exponentiel avec $T(X) = X$ comme statistique canonique. La statistique T est exhaustive et totale.
3. On a $\dfrac{\partial}{\partial q} \log p(x; q) = \dfrac{1}{q} - \dfrac{x-1}{1-q}$ et $\dfrac{\partial^2}{\partial^2 q} \log p(x; q) = -\dfrac{1}{q^2} - \dfrac{(x-1)}{(1-q)^2}$. On en déduit

$$I(q) = \frac{1}{q^2} + \frac{1}{(1-q)^2}(\mathbb{E}_q[X] - 1) = \frac{1}{q^2} + \frac{1}{(1-q)q} = \frac{1}{q^2(1-q)}.$$

4. La vraisemblance du n échantillon est

$$p_n(x_1, \ldots, x_n; q) = q^n (1-q)^{\sum_{i=1}^n x_i - n}.$$

On regarde les zéros de la dérivées en q de la log-vraisemblance $L_n = \log p_n$, et on vérifie que $\hat{q}_n = \dfrac{n}{\sum_{i=1}^n X_i} = \dfrac{1}{\bar{X}_n}$ est l'estimateur du maximum de vraisemblance de q.
5. On peut appliquer le T.C.L. Comme $\mathrm{Var}_q(X) = (1-q)/q^2$, il vient

$$\sqrt{n}(\bar{X}_n - \frac{1}{q}) \overset{Loi}{\to} \mathcal{N}(0, \frac{1-q}{q^2}).$$

Comme la fonction $h(u) = \frac{1}{u}$ est continue pour $u > 0$, on a la convergence en loi suivante

$$\sqrt{n}(\frac{1}{\bar{X}_n} - q) \overset{Loi}{\to} \mathcal{N}(0, \frac{1-q}{q^2} q^4).$$

Le modèle étant régulier, on a ainsi directement les propriétés asymptotiques de l'EMV. Dans tous les cas, nous concluons

$$\sqrt{n}(\hat{q}_n - q) \overset{Loi}{\to} \mathcal{N}(0, q^2(1-q)).$$

6. La détermination de l'intervalle de confiance est réalisée à partir de l'approximation asymptotique valable pour des grands échantillons. Nous avons $\hat{q}_n\sqrt{1-\hat{q}_n}$ qui converge presque sûrement vers $q\sqrt{1-q}$ d'une part, et $\frac{\sqrt{n}}{q\sqrt{1-q}}(\hat{q}_n - q)$ qui converge en loi vers une gaussienne centrée réduite $\mathcal{N}(0,1)$ d'autre part. Appliquant le théorème de Slutsky, on conclut $\dfrac{\sqrt{n}}{\hat{q}_n\sqrt{1-\hat{q}_n}}(\hat{q}_n - q)$ converge en loi vers G de loi $\mathcal{N}(0,1)$. Désignant par u_α le fractile tel que $P(|G| \leq u_\alpha) = 1-\alpha$, un intervale de confiance pour q de niveau $1-\alpha$ est :
$$[\hat{q}_n - u_\alpha \frac{\hat{q}_n\sqrt{1-\hat{q}_n}}{\sqrt{n}}; \hat{q}_n + u_\alpha \frac{\hat{q}_n\sqrt{1-\hat{q}_n}}{\sqrt{n}}].$$

7. Les paramètres de la modélisation sont $n = 40$ et nous avons $\sum_{i=1}^n x_i = 1779$. Nous déduisons comme estimation $\hat{q}_n \simeq 0.02249$ et comme intervalle de confiance de niveau 95%, $(u_\alpha = 1,96)$ l'intervalle $[0,016; 0,028]$. Le nombre de fraudeur estimé est $n_f \simeq n_0\hat{q}_n \in [320; 560]$.

▲

Exercice VIII.8.

1. La loi de X_i est la loi $\mathcal{N}\left(m_i + \alpha, \sigma^2\right)$. La log-vraisemblance s'écrit :
$$L_n\left(x_1, ..., x_n; \alpha\right) = -\frac{n}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(x_i - m_i - \alpha\right)^2$$
On obtient :
$$\frac{\partial}{\partial\alpha}L_n\left(x_1, ..., x_n; \alpha\right) = \frac{1}{\sigma^2}\sum_{i=1}^n \left(x_i - m_i - \alpha\right),$$
et donc
$$\frac{\partial}{\partial\alpha}L_n\left(x_1, ..., x_n; \alpha\right) = 0 \Leftrightarrow \alpha = \frac{1}{n}\sum_{i=1}^n (x_i - m_i).$$
L'étude du signe la dérivée de la log-vraisemblance, montre que la log-vraisemblance atteint son maximum en $\dfrac{1}{n}\sum_{i=1}^n (x_i - m_i)$. L'estimateur du maximum de vraisemblance est donc
$$\hat{\alpha}_n = \frac{1}{n}\sum_{i=1}^n (X_i - m_i).$$
La loi de $\hat{\alpha}_n$ est la loi $\mathcal{N}(\alpha, \sigma^2/n)$. En particulier, cet estimateur est sans biais. On vérifie s'il est efficace. On a également
$$\frac{\partial^2}{\partial^2\alpha}L_n\left(x_1, ..., x_n; \alpha\right) = -\frac{n}{\sigma^2},$$
et, on en déduit que
$$I_n = \mathbb{E}_\alpha[-\frac{\partial^2}{\partial^2\alpha}L_n\left(X_1, ..., X_n; \alpha\right)] = \frac{n}{\sigma^2}.$$

Comme $\mathrm{Var}_\alpha(\hat{\alpha}_n) = \dfrac{\sigma^2}{n} = \dfrac{1}{I_n}$, l'estimateur est efficace.
De plus les variables $(X_i - m_i)$ sont indépendantes et de même loi gaussienne. Par la loi forte des grands nombre, l'estimateur est convergent. Comme la loi de $\sqrt{n}(\hat{\alpha}_n - \alpha)$ est la loi $\mathcal{N}(0, \sigma^2)$, l'estimateur du maximum de vraisemblance est donc asymptotiquement normal (et asymptotiquement efficace).

2. La loi de X_i est la loi $\mathcal{N}(\beta m_i, \sigma^2)$. En particulier, la loi de $\tilde{\beta}_n$ est la loi $\mathcal{N}(\beta, \dfrac{\sigma^2}{n^2}\sum_{i=1}^n \dfrac{1}{m_i^2})$.
Ainsi $\tilde{\beta}_n$ est un estimateur sans biais de β_n. On vérifie s'il est efficace. La log-vraisemblance s'écrit :
$$L_n(x_1, ..., x_n; \beta) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \beta m_i)^2 .$$
On a
$$\frac{\partial^2}{\partial^2\beta} L_n(x_1, ..., x_n; \beta) = -\frac{1}{\sigma^2}\sum_{i=1}^n m_i^2,$$
et donc
$$I_n = \mathbb{E}_\beta\left[-\frac{\partial^2}{\partial^2 b} L_n(X_1, ..., X_n; \beta)\right] = \frac{1}{\sigma^2}\sum_{i=1}^n m_i^2.$$
D'autre part on a
$$\mathrm{Var}_\beta(\tilde{\beta}_n) = \frac{\sigma^2}{n^2}\sum_{i=1}^n \frac{1}{m_i^2} .$$
Par Cauchy-Schwarz, on a $\dfrac{1}{n^2}\sum_{i=1}^n \dfrac{1}{m_i^2} \geq \dfrac{1}{\sum_{i=1}^n m_i^2}$, et l'inégalité est stricte dès qu'il existe $m_i \neq m_j$. En particulier $\mathrm{Var}_\beta(\tilde{\beta}_n) > \dfrac{1}{I_n}$, s'il existe $m_i \neq m_j$, et l'estimateur n'est alors pas efficace.

3. On a
$$\frac{\partial}{\partial\beta} L_n(x_1, ..., x_n; \beta) = \frac{1}{\sigma^2}\sum_{i=1}^n m_i (x_i - \beta m_i) .$$
En étudiant le signe de cette dérivée, on en déduit que l'estimateur du maximum de vraisemblance de β est
$$\hat{\beta}_n = \frac{\sum_{i=1}^n m_i X_i}{\sum_{i=1}^n m_i^2}.$$
La loi de $\hat{\beta}_n$ est la loi $\mathcal{N}(\beta, \dfrac{\sigma^2}{\sum_{i=1}^n m_i^2})$. En particulier, cet estimateur est sans biais et il est efficace. Il est préférable à $\tilde{\beta}_n$.

4. On obtient : $\hat{\alpha}_n \simeq 88.6$, $\tilde{\beta}_n = \simeq 1.088$ et $\hat{\beta}_n \simeq 1.087$.

▲

Exercice VIII.9.

1. Notons $x_1, \ldots, x_n$ les $n = 8$ observations. La vraisemblance s'écrit

$$p_n(x_1, \ldots, x_n; a) = \frac{1}{a^n} \left(\prod_{i=1}^{n} x_i \right) \exp\left(-\frac{1}{2a} \sum_{i=1}^{n} x_i^2 \right).$$

   Son maximum est atteint pour

$$\hat{a}_n = \frac{1}{2n} \sum_{i=1}^{n} x_i^2,$$

   ce qui nous donne pour a la valeur estimée $\tilde{a} = 2.42$.

2. On vérifie que l'estimateur $\hat{a}_n$ est :
   (a) Sans biais.
   (b) Optimal, car fonction de la statistique exhaustive et totale $S_n = \frac{1}{2} \sum_{i=1}^{n} X_i^2$ (modèle exponentiel).
   (c) Efficace, car le modèle exponentiel sous sa forme naturelle donne $\lambda = -1/a$ et $\varphi(\lambda) = \log(a) = \log(-1/\lambda)$. Et dans un modèle exponentiel, sous sa forme naturelle, S_n est un estimateur efficace de $\varphi'(\lambda) = a$.
   (d) Asymptotiquement normal, par le théorème central limite.

3. La probabilité qu'une catastrophe se produise durant une année donnée s'écrit

$$p = \int_6^{+\infty} \frac{x}{\tilde{a}} \exp\left(-\frac{x^2}{2\tilde{a}} \right) \simeq 5.8\, 10^{-4}.$$

   Par indépendance, la probabilité d'avoir strictement plus d'une catastrophe en mille ans vaut

$$1 - \left((1-p)^{1000} + 1000 p (1-p)^{999} \right) \simeq 0.11,$$

   ce qui n'est pas négligeable ! Notons qu'en moyenne une catastrophe se produit tous les $1/p \simeq 1710$ ans.

▲

Exercice VIII.10.

1. On utilise l'indépendance des variables X_i :

$$\mathrm{Var}_p(\hat{p}_n) = \mathrm{Var}_p\left(\frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}_p(X_i) = \frac{p(1-p)}{n}.$$

   Comme $p \in [0,1]$, on a donc $p(1-p) \leq \frac{1}{4}$. Le maximum est atteint pour $p = \frac{1}{2}$. Finalement : $\mathrm{Var}_p(\hat{p}_n) \leq \frac{1}{4n}$.

2. Modélisons le vote d'un électeur par une loi de Bernoulli : la variable X_i, désignant le vote de l'individu i, vaut 1 s'il vote pour l'ancien maire et 0 sinon. On utilise l'approximation gaussienne donnée par le TCL :

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow[n\to\infty]{\textbf{Loi}} \mathcal{N}(0, p(1-p)).$$

Un intervalle de confiance asymptotique de niveau 5% est donné par :

$$\left[\hat{p}_n - u_{\frac{\alpha}{2}}\sqrt{\mathrm{Var}_p\left(\hat{p}_n\right)}, \hat{p}_n + u_{\frac{\alpha}{2}}\sqrt{\mathrm{Var}_p\left(\hat{p}_n\right)}\right].$$

On a $u_{\frac{\alpha}{2}}\sqrt{\mathrm{Var}_p\left(\hat{p}_n\right)} \leq u_{\frac{\alpha}{2}}\sqrt{\frac{1}{4n}}$. On veut que $u_{\frac{\alpha}{2}}\sqrt{\frac{1}{4n}} \leq 0.03$. On obtient donc : $n \geq \left(\frac{u_{\frac{\alpha}{2}}}{2 \times 0.03}\right)^2$ soit $n \geq 1068$.

3. On teste : $H_0 = \{p = q\}$ contre $H_1 = \{p \neq q\}$ où p et q désignent respectivement le nombre d'avis favorables pour le maire actuel lors du premier et du second sondage. Le test est équivalent à $H_0 = \{p - q = 0\}$ contre $H_1 = \{p - q \neq 0\}$. On peut utiliser le formalisme du test de Wald ou du test de Hausman. On peut aussi faire le raisonnement suivant, qui permet d'obtenir le même résultat. Il est naturel de considérer la statistique de test $\hat{p}_n - \hat{q}_n$ et la zone de rejet est $\{|\hat{p}_n - \hat{q}_n| > c\}$: on rejettera d'autant plus facilement H_0 que $\hat{p}_n$ sera éloigné de $\hat{q}_n$.
On utilise l'approximation gaussienne donnée par le TCL : comme les variables $\hat{p}_n$ et $\hat{q}_n$ sont indépendantes, on en déduit que

$$\sqrt{n}(\hat{p}_n - p, \hat{q}_n - q) \xrightarrow[n\to\infty]{\textbf{Loi}} \mathcal{N}(0, \Sigma),$$

où la matrice de covariance Σ est diagonale : $\Sigma = \begin{pmatrix} p(1-p) & 0 \\ 0 & q(1-q) \end{pmatrix}$. Sous H_0, on en déduit donc que

$$\sqrt{n}(\hat{p}_n - \hat{q}_n) \xrightarrow[n\to\infty]{\textbf{Loi}} \mathcal{N}(0, p(1-p) + q(1-q)).$$

On utilise enfin un estimateur convergent de la variance $\hat{p}_n(1-\hat{p}_n) + \hat{q}_n(1-\hat{q}_n)$, pour déduire du théorème de Slutsky que

$$\frac{\sqrt{n}(\hat{p}_n - \hat{q}_n)}{\sqrt{\hat{p}_n(1-\hat{p}_n) + \hat{q}_n(1-\hat{q}_n)}} \xrightarrow[n\to\infty]{\textbf{Loi}} \mathcal{N}(0,1).$$

À la vue de ce résultat, il est plus naturel de considérer la statistique de test

$$T_n = \frac{\sqrt{n}(\hat{p}_n - \hat{q}_n)}{\sqrt{\hat{p}_n(1-\hat{p}_n) + \hat{q}_n(1-\hat{q}_n)}}.$$

Remarquons que sous H_1, la statistique T_n diverge vers $+\infty$ ou $-\infty$. On choisit donc la région critique

$$W_n = \{|T_n| > c\}.$$

On détermine c le plus petit possible (région critique la plus grande possible) pour que le test soit de niveau asymptotique α. On obtient, avec l'approximation gaussienne

$$\mathbb{P}_{p,p}(W_n) = \mathbb{P}_{p,p}(|T_n| > c) \xrightarrow[n\to\infty]{} \mathbb{P}(|G| > c),$$

où G est une variable aléatoire de loi $\mathcal{N}(0,1)$. Pour obtenir un test de niveau asymptotique α, on choisit donc $c = u_{\frac{\alpha}{2}}$, le quantile d'ordre $\left(1 - \frac{\alpha}{2}\right)$ de la loi normale centrée réduite.

On rejette donc H_0 avec un risque de première espèce α si : $|T_n| > u_{\frac{\alpha}{2}}$.
Application numérique. Avec $\alpha = 0,05$: $c = 1,96$ et $t_n = -0.93$. On ne peut pas rejeter H_0 au niveau 5%, infirmant ainsi la position du journaliste. La p-valeur de ce test est : $\mathbb{P}(|G| > t_n) = 0.177$.

▲

Exercice VIII.11.

1. On a : $\pi = \mathbb{P}(X_k = 1) = \mathbb{P}(Z_k = 0) = \mathrm{e}^{-\lambda}$, et $\mathbb{P}(X_k = 0) = 1 - \mathrm{e}^{-\lambda}$. La loi de X_k est la loi de Bernoulli de paramètre π. La loi de Y, comme somme de Bernoulli indépendantes et de même paramètre, suit une loi binomiale de paramètre (n, π).
2. La vraisemblance s'écrit : $p(x;\pi) = C_n^y \pi^y (1-\pi)^{n-y}$, avec $y = \sum_{i=1}^n x_i$. La log-vraisemblance est donnée par

$$L(x;\pi) = \log p(y;\pi) = \log C_n^y + y \log \pi + (n-y) \log(1-\pi).$$

On a $\dfrac{\partial}{\partial \pi} L(x;\pi) = \dfrac{y}{\pi} - \dfrac{n-y}{1-\pi}$. En résolvant : $\dfrac{\partial}{\partial \pi} L(y;\pi) = 0$, on vérifie que l'estimateur du maximum de vraisemblance de π est $\widehat{\pi} = \dfrac{Y}{n}$. On en déduit l'estimateur du maximum de vraisemblance de λ : $\widehat{\lambda} = -\log(\widehat{\pi}) = -\log(Y/n)$.
3. En utilisant l'approximation gaussienne, on a :

$$IC_{1-\alpha}(\pi) = \left[\widehat{\pi} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}}, \widehat{\pi} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}} \right],$$

où $u_{1-\frac{\alpha}{2}}$ désigne le quantile d'ordre $\left(1 - \dfrac{\alpha}{2}\right)$ de la loi normale centrée réduite. On déduit de l'intervalle de confiance de λ à partir de celui de π, en utilisant la transformation logarithmique :

$$IC_{1-\alpha}(\lambda) = \left[-\log\left(\widehat{\pi} + u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}} \right), -\log\left(\widehat{\pi} - u_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{\pi}(1-\widehat{\pi})}{n}} \right) \right].$$

4. On trouve : $\widehat{\pi} = 0.6$, $\sqrt{\dfrac{\widehat{\pi}(1-\widehat{\pi})}{n}} \simeq 0.155$, $\widehat{\lambda} \simeq 0.51$, $IC_{95\%}(\pi) = [0.3; 0.9]$ et $IC_{95\%}(\lambda) = [0.11; 1.2]$ (non centré sur $\widehat{\lambda}$).
5. Si la densité λ est très forte alors π est très proche de 0, ce qui implique que la probabilité d'avoir des tubes négatifs est très proche de 0. On ne parviendra pas à estimer λ. Remarquons que l'estimateur du maximum de vraisemblance de λ n'est pas défini pour $\pi = 0$.

   Si la densité λ est très faible alors π est très proche de 1, ce qui implique que la probabilité d'avoir des tubes négatifs est très proche de 1. On ne parviendra pas non plus à estimer λ dans ce cas.

6. La loi de Y_i est la loi binomiale $\mathcal{B}(n_i, \pi_i)$ où $\pi_i = e^{-\lambda d_i}$. La vraisemblance s'écrit :

$$p(y_1, \ldots, y_N; \lambda) = \prod_{i=1}^{N} C_{n_i}^{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i} = \prod_{i=1}^{N} C_{n_i}^{y_i} \mathrm{e}^{-\lambda d_i y_i} \left(1 - \mathrm{e}^{-\lambda d_i}\right)^{n_i - y_i},$$

et la log-vraisemblance

$$\begin{aligned} L(y_1, \ldots, y_N; \lambda) &= \log p(y_1, \ldots, y_N; \lambda) \\ &= \sum_{i=1}^{N} \log C_{n_i}^{y_i} + \sum_{i=1}^{N} y_i \log \pi_i + \sum_{i=1}^{N} (n_i - y_i) \log(1 - \pi_i) \\ &= \sum_{i=1}^{N} \log C_{n_i}^{y_i} - \sum_{i=1}^{N} \lambda y_i d_i + \sum_{i=1}^{N} (n_i - y_i) \log\left(1 - \mathrm{e}^{-\lambda d_i}\right). \end{aligned}$$

La dérivée de la log-vraisemblance s'crit donc

$$\frac{\partial L(y_1, \ldots, y_N; \lambda)}{\partial \lambda} = -\sum_{i=1}^{N} y_i d_i + \sum_{i=1}^{N} (n_i - y_i) d_i \frac{\mathrm{e}^{-\lambda d_i}}{(1 - \mathrm{e}^{-\lambda d_i})}.$$

Elle s'annule pour λ tel que :

$$\sum_{i=1}^{N} y_i d_i = \sum_{i=1}^{N} (n_i - y_i) d_i \left(\frac{\mathrm{e}^{-\lambda d_i}}{1 - \mathrm{e}^{-\lambda d_i}}\right).$$

Le menbre de droite est une fonction strictement décroissante de λ prenant les valeurs limite $+\infty$ en 0 et 0 en $+\infty$. En particulier, l'équation ci-dessus possède une seule racine. La résolution de cette équation n'est pas toujours possible formellement. L'estimateur obtenu s'appelle MPN : most probable number.

7. On a :

$$I(\lambda) = -\mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2} L(Y_1, \ldots, Y_N; \lambda)\right] = \sum_{i=1}^{N} n_i d_i^2 \frac{\mathrm{e}^{-\lambda d_i}}{1 - \mathrm{e}^{-\lambda d_i}}.$$

Comme on a un modèle régulier, la variance asymptotique de l'estimateur du maximum de vraisemblance est donc

$$V(\lambda) = \frac{1}{I(\lambda)} = \left(\sum_{i=1}^{N} n_i d_i^2 \frac{\mathrm{e}^{-\lambda d_i}}{1 - \mathrm{e}^{-\lambda d_i}}\right)^{-1}$$

8. Dans ce cas :

$$\frac{\partial}{\partial \lambda} L(y_1, y_2; \lambda) = -y_1 + (n - y_1) \frac{\mathrm{e}^{-\lambda}}{1 - \mathrm{e}^{-\lambda}} - \frac{y_2}{2} + \frac{n - y_2}{2} \frac{\mathrm{e}^{-\frac{\lambda}{2}}}{1 - \mathrm{e}^{-\frac{\lambda}{2}}}.$$

Après avoir résolu une équation du second degré, on trouve

$$\widehat{\lambda} = -2 \log \left[\frac{-(n - Y_2) + \sqrt{(n - Y_2)^2 + 12n(2Y_1 + Y_2)}}{6n}\right].$$

L'intervalle de confiance asymptotique de niveau $1-\alpha$ de λ est donc

$$IC_{1-\alpha}(\lambda) = \left[\hat{\lambda} \pm u_{1-\frac{\alpha}{2}}\sqrt{V\left(\hat{\lambda}\right)}\right].$$

Numériquement, on trouve : $\widehat{\pi} \simeq 0.569$, $\widehat{\lambda} \simeq 1,23$ et $IC_{95\%}(\lambda) \simeq [0.44, 1.81]$.

▲

Chapitre IX

Tests

IX.1 Énoncés

Exercice IX.1.
Soit $X_1, \ldots, X_n$ un n-échantillon de loi exponentielle de paramètre $1/\theta > 0$.

1. Construire le test de niveau α $H_0 = \{\theta = \theta_0\}$ contre $H_1 = \{\theta > \theta_0\}$.
2. Construire le test de niveau α $H_0 = \{\theta = \theta_0\}$ contre $H_1 = \{\theta \neq \theta_0\}$.

△

Exercice IX.2.
Des plaignants[1] ont poursuivi en justice le Ministère israëlien de la Santé suite à une campagne de vaccination menée sur des enfants et ayant entraîné des dommages fonctionnels irréversibles pour certains d'entre eux. Ce vaccin était connu pour entraîner ce type de dommages en de très rares circonstances. Des études antérieures menées dans d'autres pays ont montré que ce risque était d'un cas sur 310 000 vaccinations. Les plaignants avaient été informés de ce risque et l'avaient accepté. Les doses de vaccin ayant provoqué les dommages objet de la plainte provenaient d'un lot ayant servi à vacciner un groupe de 300 533 enfants. Dans ce groupe, quatre cas de dommages ont été détectés.

1. On modélise l'événement "le vaccin provoque des dommages fonctionnels irréversibles sur l'enfant i" par une variable aléatoire de Bernoulli, X_i, de paramètre p. Calculer la valeur p_0 correspondant aux résultats des études antérieures.
2. Justifier qu'on peut modéliser la loi du nombre N de cas de dommages par une loi de Poisson de paramètre θ. Calculer la valeur θ_0 attendue si le vaccin est conforme aux études antérieures.
3. L'hypothèse $H_0 = \{p = p_0\}$ correspond au risque que les plaignants avaient accepté, l'hypothèse alternative étant $H_1 = \{p > p_0\}$. Construire un test de niveau α à partir de la variable N. Accepte-t'on H_0 au seuil de 5% ? Donner la p-valeur de ce test.

△

[1] cf. Murray Aitkin, *Evidence and the Posterior Bayes Factor*, 17 Math. Scientist 15 (1992)

Exercice IX.3.
Une agence de voyage souhaite cibler sa clientèle. Elle sait que les coordonnées du lieu de vie d'un client (X, Y) rapportées au lieu de naissance $(0, 0)$ sont une information significative pour connaître le goût de ce client. Elle distingue :

– La population 1 (Hypothèse H_0) dont la loi de répartition a pour densité :

$$p_1(x, y) = \frac{1}{\sqrt{4\pi^2}} e^{-\frac{x^2+y^2}{2}} dxdy.$$

– La population 2 (Hypothèse H_1) dont la loi de répartition a pour densité :

$$p_2(x, y) = \frac{1}{16} \mathbf{1}_{[-2;2]}(x) \mathbf{1}_{[-2;2]}(y) dxdy.$$

L'agence souhaite tester l'hypothèse qu'un nouveau client vivant en (x, y) appartient à la population 1 plutôt qu'à la population 2.

1. Proposer un test de niveau inférieur à $\alpha = 5\%$ et de puissance maximale, construit à partir du rapport de vraisemblance.
2. Donner une statistique de test et caractériser graphiquement la région critique dans $\mathbb{R}^2$.

△

Exercice IX.4.
On considère un échantillon gaussien $(X_1, Y_1), \ldots, (X_n, Y_n)$ de variables aléatoires indépendantes de loi $\mathcal{N}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right)$, où σ_1 et σ_2 sont inconnus. Le paramètre est $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$.

1. Décrire le test de Wald pour tester si $\mu_1 = \mu_2$ et donner une région critique de niveau asymptotique 5%.
2. On suppose $\sigma_1 = \sigma_2$. Donner une région critique de niveau exact 5%, construite à l'aide de la même statistique de test que celle utilisée dans la question précédente. Faire l'application numérique pour $n = 15$.

△

Exercice IX.5.
Soit $(X_n, n \geq 1)$, une suite de variables aléatoires indépendantes, de loi normale $\mathcal{N}(\theta, \theta)$, avec $\theta > 0$. L'objectif de cet exercice est de présenter deux tests pour déterminer si pour une valeur déterminée $\theta_0 > 0$, on a $\theta = \theta_0$ (hypothèse H_0) ou $\theta > \theta_0$ (hypothèse H_1). On note

$$\bar{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \quad \text{et} \quad V_n = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \bar{X}_n)^2 = \frac{1}{n-1} \sum_{k=1}^{n} X_k^2 - \frac{n}{n-1} \bar{X}_n^2.$$

1. Déterminer $\hat{\theta}_n$, l'estimateur du maximum de vraisemblance de θ. Montrer directement qu'il est convergent, asymptotiquement normal et donner sa variance asymptotique. Est-il asymptotiquement efficace ?
2. Construire un test asymptotique convergent à l'aide de l'estimateur du maximum de vraisemblance.

On considère la classe des estimateurs T_n^λ de la forme : $T_n^\lambda = \lambda \bar{X}_n + (1-\lambda)V_n, \quad \lambda \in \mathbb{R}$.

3. Montrer que la suite d'estimateurs $(T_n^\lambda, n \geq 2)$ est convergente, sans biais. Donner la variance de T_n^λ.
4. Étudier la convergence en loi de $(Z_n, n \geq 1)$, avec $Z_n = \sqrt{n}(\bar{X}_n - \theta, n^{-1}\sum_{k=1}^{n} X_k^2 - \theta - \theta^2)$.
5. Étudier la convergence en loi de $(\sqrt{n}(T_n^\lambda - \theta), n \geq 2)$. En déduire que la suite d'estimateurs $(T_n^\lambda, n \geq 2)$ est asymptotiquement normale. Calculer la variance asymptotique.
6. On considère pour $n \geq 2$, la statistique de test

$$\zeta_n^\lambda = \frac{\sqrt{n}(T_n^\lambda - \theta_0)}{\sqrt{\lambda^2\theta_0 + 2(1-\lambda)^2\theta_0^2}}.$$

   Construire un test asymptotique convergent à partir de cette statistique de test. On donne les valeurs numériques suivantes : $\theta_0 = 3.8$, $\lambda = 0.5$, $n = 100$, $\bar{x}_n = 4.18$ et $v_n = 3.84$. Calculer la p-valeur du test. Quelle est votre décision ?
7. On considère maintenant la valeur λ^* qui minimise $\lambda^2\theta_0 + 2(1-\lambda)^2\theta_0^2$. Comparer les variances asymptotique de $T_n^{\lambda^*}$ et $\hat{\theta}_n$. Reprendre la question précédente avec $\lambda = \lambda^*$ et comparer les résultats à ceux déjà obtenus avec l'estimateur du maximum de vraisemblance.

△

Exercice IX.6.
L'information dans une direction de l'espace prise par un radar de surveillance aérienne se présente sous la forme d'un n-échantillon $X = (X_1, \ldots, X_n)$ de variables aléatoires indépendantes de même loi gaussienne de moyenne θ, paramètre inconnu, et de variance σ^2 connu. On notera $f(x, \theta)$, $x \in \mathbb{R}^n$ la densité du vecteur aléatoire.

En l'absence de tout Objet Volant (Hypothèse H_0), $\theta = \theta_0 \in \mathbb{R}^+$, sinon (Hypothèse H_1), $\theta = \theta_1$, avec $\theta_1 > \theta_0$.

1. Montrer comment le lemme de Neyman-Pearson permet la construction d'un test de l'hypothèse $\theta = \theta_0$ contre l'hypothèse $\theta = \theta_1$ de niveau α et de puissance maximale.
2. Quelle est la plus petite valeur de n permettant de construire un test de niveau $\alpha, \alpha \in [0;1]$ et d'erreur de deuxième espèce inférieure ou égale à $\beta, \beta \in [0;1]$, avec $\alpha < \beta$?
3. Supposons maintenant qu'en presence d'objet volant l'information fournie par le radar est un n-échantillon $X = (X_1, \ldots, X_n)$ de variables aléatoires indépendates de même loi gaussienne de moyenne $\theta \neq \theta_0, \theta \in \mathbb{R}$ et de variance σ^2. Peut-on construire un test de l'hypothèse $\theta = \theta_0$ contre l'hypothèse $\theta \neq \theta_0$ de niveau α donné uniformément plus puissant ?
4. On se propose de trouver un test de niveau α uniformement plus puissant sans biais parmi les tests purs pour tester l'hypothèse $\theta = \theta_0$ contre l'hypothèse $\theta \neq \theta_0$.
   (a) Soit $\theta_1 \neq \theta_0, \theta_1 \in \mathbb{R}$. Prouver l'existence de deux constantes c et c^* définissant un ensemble :

$$\mathcal{X} = \{x \in \mathbb{R}^n / f(x, \theta_1) \geq c f(x, \theta_0) + c^* \frac{\partial f(x, \theta_0)}{\partial \theta}\},$$

   avec $\mathbb{P}_{\theta_0}(X \in \mathcal{X}) = \alpha$ et $\int_{\mathcal{X}} \frac{\partial f(x,\theta)}{\partial \theta}_{|\theta_0} dx = 0$. Cet ensemble dépend-il de θ_1 ?

(b) Montrer que le test pur de niveau α dont la région critique est $\mathcal{X}$ est uniformément plus puissant dans la classe des tests purs sans biais.

△

Exercice IX.7.
Une machine outil fabrique des ailettes de réacteur avec les caractéristiques suivantes : la longueur d'une ailette suit une loi $\mathcal{N}(\bar{L}_0, \sigma_0)$ avec $\bar{L}_0 = 785\,\text{mm}$ et $\sigma_0 = 2\,\text{mm}$.

Une trop grande dispersion dans les caractéristiques de la machine peut avoir deux conséquences :

- La première est de produire des ailettes trop longues, qui alors ne sont pas montables.
- La seconde est de produire des ailettes trop courtes, ce qui nuit aux performances du réacteur.

On a donc particulièrement étudié la machine afin de maîtriser au mieux le paramètre σ_0, qui peut être considéré comme connu et invariable. Par contre, la longueur moyenne a tendance à varier au fil de la production. On désire vérifier que les caractéristiques de la machine n'ont pas trop dérivé, à savoir que $\bar{L} \in [\bar{L}_0 \pm \delta L]$ avec $\delta L = 1.5\,\text{mm}$. On procède à l'analyse de 200 ailettes, ce qui conduit à une moyenne empirique $\dfrac{1}{200}\displaystyle\sum_{i=1}^{200} L_i = 788.3\,\text{mm}$.

1. Vérifier que le modèle est exponentiel. Construire un estimateur de $\bar{L}$.
2. On souhaite tester l'hypothèse $H_0 = \bar{L} \in [\bar{L}_0 - \delta L, \bar{L}_0 + \delta L]$ contre $H_1 = \bar{L} \notin [\bar{L}_0 - \delta L, \bar{L}_0 + \delta L]$. Construire un test bilatéral UPPS au seuil α pour tester H_0 contre H_1.
3. Faire l'application numérique pour $\alpha = 5\%$. Conclusion ?
4. Calculer un intervalle de confiance centré à 95% sur $\bar{L}$ et le comparer à $\bar{L}_0 + \delta L$.

△

Exercice IX.8.
Dans l'outillage de votre usine vous utilisez une grande quantité de pièces d'un certain modèle. Dans les conditions usuelles d'emploi, vous avez observé que la durée de vie de ces pièces est une variable aléatoire normale dont l'espérance mathématique est $\mu_0 = 120$ heures, et l'écart-type est $\sigma = 19,4$ heures.

Le représentant d'un fournisseur vous propose un nouveau modèle, en promettant un gain de performance en moyenne de 5%, pour une dispersion identique σ.

Vous décidez de tester le nouveau modèle sur un échantillon de $n = 64$ unités. On note $(X_i, i \in \{1, \ldots, 64\})$ la durée de vie des pièces testées.

1. Quelle loi proposez vous pour les variables aléatoires $(X_i, i \in \{1, \ldots, 64\})$?
2. Soit μ l'espérance mathématique du nouveau modèle. Donner un estimateur sans biais de μ. Identifier la loi de cet estimateur.
3. Vous ne voulez pas changer de modèle si le nouveau n'est pas plus performant que l'ancien (hypothèse H_0). Plus précisément, vous voulez que la probabilité d'adopter à tort le nouveau modèle ne dépasse pas le seuil de 0.05. Quelle est alors la procédure de décision construite à partir de l'estimateur de μ ?

4. Évaluez le risque que cette procédure vous fasse rejeter le nouveau modèle si l'annonce du représentant est exacte. Les 64 pièces testées ont eu une durée de vie moyenne égale à 123.5 heures. Que concluez-vous ?

Le représentant conteste cette procédure, prétextant qu'il vaut mieux partir de l'hypothèse H_0', selon laquelle le gain de performance moyen est réellement de 5%. Il souhaite que la probabilité de rejeter à tort le nouveau modèle ne dépasse pas le seuil de 0.05.

5. Quelle est alors la procédure de décision ? Quel est le risque de l'acheteur ? Quel est le résultat de cette procédure au vu des observations faites. Commentez.
6. Quelle procédure peut-on proposer pour égaliser les risques de l'acheteur et du vendeur ? Quel est alors ce risque ?

△

Exercice IX.9.
Suite de l'exercice IX.8. Un représentant d'une autre société se présente et déclare avoir un produit moins cher et équivalent à celui des questions précédentes (de moyenne $\nu = 1.05\mu_0$ et de variance σ). L'acheteur le teste sur un échantillon de m pièces. Le résultat obtenu est une moyenne de 124.8. On veut tester si les deux modèles sont de performances équivalentes. On note $p(x, y; \mu, \nu)$ la densité du modèle.

7. Expliciter l'estimateur $\hat{\theta}$ du maximum de vraisemblance sachant que $\mu = \nu$. Expliciter $\hat{\mu}$ et $\hat{\nu}$ les estimateurs de vraisemblance dans le cas général.
8. Expliciter la forme de la région critique. Que peut-on dire des performances relatives des deux types de pièces si $m = 64$?

△

Exercice IX.10.
On souhaite vérifier la qualité du générateur de nombres aléatoires d'une calculatrice scientifique. Pour cela, on procède à 250 tirages dans l'ensemble $\{0, \ldots, 9\}$ et on obtient les résultats suivants :

x	0	1	2	3	4	5	6	7	8	9
$N(x)$	28	32	23	26	23	31	18	19	19	31

À l'aide du test du χ^2, vérifier si le générateur produit des entiers indépendants et uniformément répartis sur $\{1, \ldots, 9\}$.

△

Exercice IX.11.
Test d'égalité des lois marginales (Test de McNemar). On considère deux juges qui évaluent les mêmes évènements, répondant chaque fois par l'affirmative (+) ou négative (−). On note $p_+^{(i)}$ la probabilité que le juge i réponde par l'affirmative et $p_-^{(i)}$ la probabilité qu'il réponde par la négative. On désire savoir si les lois des réponses des deux juges sont les mêmes (hypothèse H_0) ou non (hypothèse H_1).

1. Vérifier que sous H_0 la loi du couple de réponse ne dépend que de deux paramètres, i.e. qu'il existe α et β tels que

	$p_+^{(2)}$	$p_-^{(2)}$	
$p_+^{(1)}$	$\beta-\alpha$	α	β
$p_-^{(1)}$	α	$1-\alpha-\beta$	$1-\beta$
	β	$1-\beta$	1

2. Calculer l'estimateur du maximum de vraisemblance de (α, β).
3. On désire réaliser un test sur un échantillon de taille n. Donner la statistique de test ζ_n du χ^2 et vérifier que
$$\zeta_n = \frac{(N_{+-} - N_{-+})^2}{N_{+-} + N_{-+}},$$
où N_{+-} est le nombre de fois où le juge 1 a répondu + et le juge 2 a répondu −, et N_{-+} est le nombre de fois où le juge 1 a répondu − et le juge 2 a répondu +. Donner la région critique à 5%.
4. Pour $n = 200$, on observe $N_{+-} = 15$ et $N_{-+} = 5$, calculer la $p-$valeur du test.

△

Exercice IX.12.
L'examen de 320 familles ayant cinq enfants donne pour résultat le tableau IX.1.

Nombres de garçons et de filles	(5,0)	(4,1)	(3,2)	(2,3)	(1,4)	(0,5)	Total
Nombre de familles	18	52	110	88	35	17	320

TAB. IX.1 – Observation de 320 familles

On veut savoir si ce résultat est compatible avec l'hypothèse que la naissance d'un garçon et la naissance d'une fille sont des événements équiprobables. Soit r la probabilité d'avoir un garçon.

1. Calculer la proportion de garçons. Appliquez le test du χ^2 de niveau $\alpha = 0.01$, basé sur cette proportion, pour déterminer si $r = 1/2$. Donner la p-valeur de ce test.
2. Donner un intervalle de confiance pour r de niveau α. Remarquer que le test du χ^2 est équivalent à l'intervalle de confiance.
3. Appliquez le test du χ^2 de niveau $\alpha = 0.01$, directement à partir des données du tableau IX.1. Donner approximativement la p-valeur de ce test. Conclusion ?
4. Appliquer le test du χ^2 pour vérifier si les données du tableau IX.1 suivent une loi binomiale de paramètre 5 et r, avec r inconnu. Donner approximativement la p-valeur de ce test. Conclusion ?

△

Exercice IX.13.
En 1986, à Boston, le docteur Spock, militant contre la guerre du Vietnam, fut jugé pour incitation publique à la désertion. Le juge chargé de l'affaire était soupçonné de ne pas être équitable dans la sélection des jurés[2]. En effet, il y avait 15% de femmes parmi les 700 jurés qu'il avait nommés dans ses procès précédents, alors qu'il y avait 29% de femmes éligibles sur l'ensemble de la ville.

1. Tester si le juge est impartial dans la selection des jurés. Quelle est la p-valeur de ce test ?
2. Que conclure si étant plus jeune, le juge n'a pas nommé 700, mais 40 jurés ?

△

Exercice IX.14.
On se propose de comparer les réactions produites par deux vaccins B.C.G. désignés par A et B[3]. Un groupe de 348 enfants a été divisé par tirage au sort en deux séries qui ont été vaccinées, l'une par A, l'autre par B. La réaction a été ensuite lue par une personne ignorant le vaccin utilisé. Les résultats figurent dans le tableau suivant :

Vaccin	Réaction légère	Réaction moyenne	Ulcération	Abcès	Total
A	12	156	8	1	177
B	29	135	6	1	171
Total	41	291	14	2	348

1. En ce qui concerne les réactions, peut-on dire que les vaccins A et B sont identiques ? On utilisera un test du χ^2 en supposant dans un premier temps que le tirage au sort est équitable.
2. Que se passe-t-il si on ne suppose plus que le tirage au sort est équitable ?

△

Exercice IX.15.
On désire étudier la prédominance visuelle de l'oeil et l'habilité de la main. Un expérimentateur établit la table de contingence IX.15. Á l'aide d'un test du χ^2 dire s'il existe une relation entre la prédominance visuelle et l'habilité des mains (avec un seuil de 0.25) ?

Mobilité manuelle \ Vue	Gauche	Deux yeux	Droit	Total
Gauche	9	15	7	31
Deux mains	8	7	4	19
Droite	15	26	9	50
Total	32	48	20	100

TAB. IX.2 – Résultats du test mains-yeux

△

[2]T.H Wonnacott & R.J Wonnacot, statistique, Economica, 1991
[3]D. Schwartz et P. Lazar, *Éléments de statistique médicale et biologique*, Flammarion, Paris (1964).

Exercice IX.16.
Pour déterminer si les merles vivent en communauté ou en solitaire, on procède à l'expérience suivante[4] : on dispose un filet dans la zone d'habitat des merles, et on vient relever le nombre de captures pendant 89 jours. On obtient les résultats suivants :

Nombre de captures	0	1	2	3	4	5	6
Nombre de jours	56	22	9	1	0	1	0

1. On suppose qu'une loi de Poisson est représentative de l'expérience. Construire un estimateur du paramètre de cette loi.
2. Vérifier à l'aide d'un test du χ^2 l'adéquation du modèle aux données. Faire l'application numérique au niveau $\alpha = 5\%$.
3. Reprendre l'exercice en groupant les catégories Nombre de captures = 2, 3, 4, 5 et 6 en Nombre de captures ≥ 2.

△

Exercice IX.17.
Le tableau IX.3 donne, sur une période de vingt ans, le nombre de décès par an et par régiment dans la cavalerie prussienne causés par un coup de sabot de cheval. On dispose de 200 observations. Appliquer le test du χ^2 pour vérifier si les données suivent une loi de Poisson (dont on estimera le paramètre).

Nombre de décès par an et par régiment	0	1	2	3	4
Nombre d'observations	109	65	22	3	1

TAB. IX.3 – Décés par an et par régiment

△

Exercice IX.18.
Les dés de Weldon. Weldon a effectué $n = 26306$ lancers de douze dés à six faces[5]. On note X_i le nombre de faces indiquant cinq ou six lors du i-ième lancer. Les fréquences empiriques observées sont notées :

$$\hat{f}_j = \frac{N_j}{n} ,$$

où N_j est le nombre de fois où l'on a observé j faces indiquant cinq ou six, sur les douze lancers $N_j = \sum_{i=1}^{n} \mathbf{1}_{\{X_i=j\}}$. Les observations sont données dans les tableaux IX.4 et IX.5.

Si les dés sont non biaisés, la probabilité d'observer les faces cinq ou six dans un lancer de dés est de 1/3. Les variables aléatoires $(X_i, 1 \leq i \leq n)$ suivent donc la loi binomiale de paramètres 12 et 1/3. Les fréquences théoriques sont données dans le tableau IX.6.

[4]Revue du CEMAGREF (Clermond Ferrand) juin 1996
[5]W. Feller, *An introduction to probability theory and its applications*, volume 1, third ed., p. 148.

$N_0 = 185$	$N_1 = 1149$	$N_2 = 3265$	$N_3 = 5475$
$N_4 = 6114$	$N_5 = 5194$	$N_6 = 3067$	$N_7 = 1331$
$N_8 = 403$	$N_9 = 105$	$N_{10} = 14$	$N_{11} = 4$
$N_{12} = 0$			

TAB. IX.4 – Observations

$\hat{f}_0 = 0.007033$	$\hat{f}_1 = 0.043678$	$\hat{f}_2 = 0.124116$	$\hat{f}_3 = 0.208127$
$\hat{f}_4 = 0.232418$	$\hat{f}_5 = 0.197445$	$\hat{f}_6 = 0.116589$	$\hat{f}_7 = 0.050597$
$\hat{f}_8 = 0.015320$	$\hat{f}_9 = 0.003991$	$\hat{f}_{10} = 0.000532$	$\hat{f}_{11} = 0.000152$
$\hat{f}_{12} = 0.000000$			

TAB. IX.5 – Fréquences empiriques observées

$f_0 = 0.007707$	$f_1 = 0.046244$	$f_2 = 0.127171$	$f_3 = 0.211952$
$f_4 = 0.238446$	$f_5 = 0.190757$	$f_6 = 0.111275$	$f_7 = 0.047689$
$f_8 = 0.014903$	$f_9 = 0.003312$	$f_{10} = 0.000497$	$f_{11} = 0.000045$
$f_{12} = 0.000002$			

TAB. IX.6 – Fréquences théoriques

1. Donner la statistique du test du χ^2 et la p-valeur. En déduire que l'on rejette l'hypothèse des dés non biaisés.
2. Rejette-t-on également l'hypothèse selon laquelle les variables sont distribuées suivant une loi binomiale de même paramètre $(12, r)$, r étant inconnu ?

△

IX.2 Corrections

Exercice IX.1.

1. Soit $\theta' > \theta$. Le rapport de vraisemblance est :

$$\frac{p_n(x_1,\dots,x_n,\theta')}{p_n(x_1,\dots,x_n,\theta)} = \left(\frac{\theta}{\theta'}\right)^n \exp\left\{-\left(\frac{1}{\theta'}-\frac{1}{\theta}\right)\sum_{i=1}^n x_i\right\} .$$

Donc, il existe un test U.P.P. de niveau α donné par la région critique :

$$W_\alpha = \left\{(x_1,\dots,x_n) \mid \sum_{i=1}^n x_i > k_\alpha\right\} ,$$

où k_α est tel que $\alpha = \mathbb{P}_{\theta_0}(W)$. La statistique de test $\sum_{i=1}^n X_i$ suit la loi Gamma de paramètres n et $\frac{1}{\theta_0}$, sous l'hypothèse nulle. Donc, $\frac{2}{\theta_0}\sum_{i=1}^n X_i$ suit la loi Gamma de paramètres n et $\frac{1}{2}$, i.e. la loi du χ^2 à $2n$ degrés de liberté (toujours sous l'hypothèse nulle). Donc, on peut déterminer k_α :

$$\alpha = \mathbb{P}_{\theta_0}(W) = \alpha = \mathbb{P}_{\theta_0}\left(\sum_{i=1}^n X_i > k_\alpha\right) = \mathbb{P}_{\theta_0}\left(\frac{2}{\theta_0}\sum_{i=1}^n X_i > \frac{2}{\theta_0}k_\alpha\right) .$$

Donc, $\frac{2}{\theta_0}k_\alpha = z_\alpha(2n)$. On obtient donc la région critique :

$$W_\alpha = \left\{(x_1,\dots,x_n) \mid \sum_{i=1}^n x_i > \frac{\theta_0}{2}z_\alpha(2n)\right\} ,$$

où $z_\alpha(2n)$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $2n$ degrés de liberté.

2. La région critique du test U.P.P. de niveau α est donné par :

$$W_\alpha = \left\{(x_1,\dots,x_n) \mid \sum_{i=1}^n x_i < k_1 \quad \text{et} \quad \sum_{i=1}^n x_i > k_2\right\} ,$$

avec k_1 et k_2 tels que $\alpha = \mathbb{P}_{\theta_0}(W)$:

$$\begin{aligned} 1-\alpha &= \mathbb{P}_{\theta_0}\left(k_1 < \sum_{i=1}^n X_i < k_2\right) \\ &= \mathbb{P}_{\theta_0}\left(\frac{2}{\theta_0}k_1 < \frac{2}{\theta_0}\sum_{i=1}^n X_i < \frac{2}{\theta_0}k_2\right) . \end{aligned}$$

Or, sous H_0, $\frac{2}{\theta_0}\sum_{i=1}^n X_i$ suit la loi du χ^2 à $2n$ degrés de liberté. On choisit donc $\frac{2}{\theta_0}k_1 = z_{\alpha_1}(2n)$ et $\frac{2}{\theta_0}k_2 = z_{1-\alpha_2}(2n)$ avec $\alpha_1+\alpha_2 = \alpha$. Un choix classique consiste à prendre $\alpha_1 = \alpha_2 = \alpha/2$ (on accorde alors autant d'importance de se tromper d'un côté que de l'autre). Cela correspond en fait au test bilatère UPPS dans un modèle exponentiel. Ce test est toutefois dégénéré car l'intervalle $[\theta_0,\theta_0]$ est réduit à un singleton.

▲

Exercice IX.2.

1. D'après les données de l'énoncé, $p_0 = 1/310\ 000 \simeq 3.23\,10^{-6}$.
2. En supposant les X_i indépendants, la variable aléatoire N suit une loi binomiale de paramètres $n = 300\ 533 \gg 1$ et p de l'ordre de $p_0 = 3.23\,10^{-6} \ll 1$. On peut donc approcher cette loi par une loi de Poisson de paramètre $\theta = np$. Dans le cas des études antérieures, on a $\theta_0 = 300\ 533\ p_0 \simeq 0.969$.
3. On construit un test de la forme $\varphi(N) = \mathbf{1}_{\{N \geq N_\alpha\}}$. Sous l'hypothèse H_0, N suit une loi de Poisson de paramètre θ_0. On a alors :

$$\alpha = \mathbb{E}[\varphi(N)] = \mathbb{P}(N \geq N_\alpha) = 1 - \sum_{k<N_\alpha} \frac{\theta^k}{k!}\,\mathrm{e}^{-\theta}\,.$$

On obtient pour $N_\alpha = 3$, $\alpha \simeq 7.47\%$ et pour $N_\alpha = 4$, $\alpha \simeq 1.71\%$. Comme on a observé $N = 4$ cas de dommages, on rejette l'hypothèse H_0 au seuil de 5%. La p-valeur de ce test est de $\alpha \simeq 1.71\%$.

▲

Exercice IX.3.

1. Cet exemple rentre dans le cadre du Lemme de Neyman-Pearson. On définit la région critique du test :

$$A_\alpha = \{(x,y) \in \mathbb{R}^2; \frac{p_2(x,y)}{p_1(x,y)} > k_\alpha\}.$$

On a aussi

$$A_\alpha = \{(x,y) \in \mathbb{R}^2 \cap ([-2;2] \times [-2;2]); x^2 + y^2 \geq K_\alpha\}.$$

Sous l'hypothèse H_0, $X^2 + Y^2$ suit une loi du $\chi^2(2)$. On a $5\% = \mathbb{P}_1(\chi^2(2) \geq 5.99)$. Soit α tel que $K_\alpha = 5.99$. On en déduit donc que $\mathbb{P}_1(A_\alpha) = \alpha \leq 5\%$. C'est le test le plus puissant parmi les tests de même niveau.

2. Une statistique de test est $X^2 + Y^2$. La région critique est l'intersection de l'extérieur du cercle de rayon $\sqrt{5.99}$ avec le carré $[-2;2] \times [-2;2]$.

▲

Exercice IX.4.

1. Le vecteur des paramètres est $\theta' = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. On veut tester $H_0 : \{\mu_1 = \mu_2\}$ contre $H_1 : \{\mu_1 \neq \mu_2\}$. On applique le test de Wald avec la fonction $g(\mu_1, \mu_2, \sigma_1, \sigma_2) = \mu_1 - \mu_2$. L'estimateur du maximum de vraisemblance de θ

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n X_i,\ \ \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^n Y_i,\ \ \hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\mu}_1)^2,\ \ \hat{\sigma}_2^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\mu}_2)^2.$$

La log-vraisemblance associée à une observation est :

$$\log p(x,y,\theta) = -\log(2\pi) - \frac{1}{2}\log(\sigma_1^2\sigma_2^2) - \frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(y-\mu_2)^2}{2\sigma_2^2}\,.$$

On calcule ensuite la matrice d'information de Fisher :

$$I(\theta) = \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & 0 \\ 0 & 1/\sigma_2^2 & 0 & 0 \\ 0 & 0 & 4/\sigma_1^2 & 0 \\ 0 & 0 & 0 & 4/\sigma_2^2 \end{pmatrix}.$$

Puis on calcule la variance asymptotique :

$$\Sigma(\theta) = \frac{\partial g'}{\partial \theta}(\theta) I^{-1}(\theta) \frac{\partial g}{\partial \theta'}(\theta) = \sigma_1^2 + \sigma_2^2.$$

La statistique du test de Wald est donc :

$$\zeta_n = \frac{n(\hat{\mu}_1 - \hat{\mu}_2)^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}\ .$$

Sous H_0, ζ_n converge en loi vers un χ^2 à 1 degré de liberté. La région critique de niveau asymptotique 5% est donnée par $\{\zeta_n > z\}$, où z est le quantile d'ordre 95% de la loi $\chi^2(1)$ soit $z = 3.84$.

2. On note σ^2 la valeur commune de σ_1^2 et σ_2^2. Pour déterminer une région critique de niveau exact, il faut déterminer la loi de ζ_n sous H_0. Remarquons que $n\hat{\sigma}_1^2/\sigma^2$ et $n\hat{\sigma}_2^2/\sigma^2$ sont indépendants et de même loi $\chi^2(n-1)$. On en déduit, en utilisant les fonctions caractéristiques que $n(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/\sigma^2$ suit la loi $\chi^2(2n-2)$. On remarque aussi que, sous H_0, $n(\hat{\mu}_1 - \hat{\mu}_2) = \sum_{i=1}^n (X_i - Y_i)$ a pour loi $\mathcal{N}(0, 2n\sigma^2)$. Il en découle que $\dfrac{n(\hat{\mu}_1 - \hat{\mu}_2)^2}{2\sigma^2}$ a pour loi $\chi^2(1)$. Ainsi, on obtient finalement que :

$$\zeta_n \overset{\text{Loi}}{=} 2n \frac{Z_1}{Z_{2n-2}},$$

avec Z_1 et Z_{2n-2} de $\chi^2(2n-2)$. De l'indépendance de $\hat{\mu}_1$ avec $\hat{\sigma}_1^2$ et de $\hat{\mu}_2$ avec $\hat{\sigma}_2^2$, il résulte que Z_1 et Z_{2n-2} sont indépendantes. Ainsi la loi de $\dfrac{n-1}{n}\zeta_n$ paramètre $(1, 2n-2)$. La région critique de niveau exact 5% est donnée par $\{\zeta_n > \dfrac{n}{n-1} z'\}$, où z' est le quantile d'ordre 95% de la loi $(1, 2n-2)$ soit si $n = 15$, $z' = 4.20$
Remarquons que le facteur $n/(n-1)$ disparaît si l'on considère les estimateurs sans biais $\tilde{\sigma}_1^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \hat{\mu}_1)^2$ et $\tilde{\sigma}_2^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \hat{\mu}_2)^2$ de σ_1^2 et σ_2^2. Remarquons enfin que $(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2$ est bien l'estimateur du maximum de vraisemblance de σ^2 dans le modèle où $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

▲

Exercice IX.5.

1. La vraisemblance du modèle est pour $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$,

$$p_n(x;\theta) = \prod_{i=1}^n \frac{\mathrm{e}^{-(x_i-\theta)^2/2\theta}}{\sqrt{2\pi\theta}},$$

et la log-vraisemblance est donnée par

$$L_n(x;\theta) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\theta) - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\theta}.$$

On obtient

$$\begin{aligned}\frac{\partial}{\partial\theta}L_n(x;\theta) &= -\frac{n}{2\theta} + \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\theta^2} + \sum_{i=1}^{n}\frac{(x_i-\theta)}{\theta}\\ &= -\frac{n}{2\theta^2}(\theta^2+\theta-\frac{1}{n}\sum_{i=1}^{n}x_i^2).\end{aligned}$$

Les conditions $\frac{\partial}{\partial\theta}L_n(x;\theta)=0$ et $\theta>0$ impliquent $\theta' = -\frac{1}{2} + \sqrt{\frac{1}{4}+\frac{1}{n}\sum_{i=1}^{n}x_i^2}$. Comme on a $\lim_{\theta\to 0} L_n(x;\theta) = \lim_{\theta\to\infty} L_n(x;\theta) = -\infty$, la log-vraisemblance est maximale en θ'. L'estimateur du maximum de vraisemblance est donc

$$\hat\theta_n = -\frac{1}{2} + \sqrt{\frac{1}{4}+\frac{1}{n}\sum_{i=1}^{n}X_i^2}.$$

Les variables $(X_n^2, n\geq 1)$ sont indépendantes et intégrables, la fonction $f: x\mapsto -\frac{1}{2}+\sqrt{\frac{1}{4}+x}$ est continue sur $\mathbb{R}^+$, on en déduit que la suite d'estimateur $(\hat\theta_n, n\geq 1)$ converge p.s. vers $f(\mathbb{E}_\theta[X_1^2]) = f(\theta+\theta^2) = \theta$. La suite d'estimateur est donc convergente. Les variables $(X_n^2, n\geq 1)$ sont indépendantes et de carré intégrable, et la fonction f est de classe C^1 sur $\mathbb{R}^+$. On déduit donc du théorème central limite, que la suite d'estimateur $(\hat\theta_n, n\geq 1)$ est asymptotiquement normale de variance asymptotique

$$\sigma^2 = f'(\mathbb{E}_\theta[X_1^2])^2\,\mathrm{Var}_\theta(X_1^2) = \frac{2\theta^2}{1+2\theta}.$$

De plus, on a

$$I(\theta) = -\mathbb{E}_\theta[\frac{\partial}{\partial\theta}L_1(X_1;\theta)] = -\frac{1}{2\theta^2} + \frac{1}{\theta^3}\,\mathbb{E}_\theta[X_1^2] = \frac{1+2\theta}{2\theta^2}.$$

Comme $\sigma^2 = 1/I(\theta)$, la suite d'estimateurs $(\hat\theta_n, n\geq 1)$ est asymptotiquement efficace.

2. On considère la statistique de test

$$\zeta_n' = \sqrt{n}(\hat\theta_n-\theta_0)\frac{\sqrt{1+2\theta_0}}{\theta_0\sqrt{2}}.$$

On déduit de la question précédente que sous H_0, la statistique de test converge en loi vers une gaussienne $\mathcal{N}(0,1)$. Sous H_1, la suite $(\hat\theta_n, n\geq 1)$ converge p.s. vers $\theta>\theta_0$ En particulier $(\zeta_n', n\geq 1)$ converge p.s. vers $+\infty$. On en déduit que les régions critiques

$$W_n' = \{\zeta_n' > c\}$$

définissent un test asymptotique convergent de niveau $\alpha = \mathbb{P}(\mathcal{N}(0,1)>c)$. Comme $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n (x_i-\bar x_n)^2 + n\bar x_n^2$, on obtient $\hat\theta_n = 4.17$, $\zeta_n' = 2.00$ et une p-valeur de 2.3%. On rejette donc H_0 au niveau de 5%.

3. Par la loi forte des grands nombres, la suite $(\bar{X}_n, n \geq 1)$ converge presque sûrement vers $\mathbb{E}_\theta[X_1] = \theta$. Comme les variables aléatoires X_k sont indépendantes, de même loi et de carré intégrable, on déduit de la loi forte des grands nombres que la suite $(\frac{1}{n}\sum_{k=1}^n X_k^2, n \geq 2)$ converge presque sûrement vers $\mathbb{E}_\theta[X_1^2]$. Comme

$$V_n = \frac{n}{n-1}\frac{1}{n}\sum_{k=1}^n X_k^2 - \frac{n}{n-1}\bar{X}_n^2,$$

on en déduit donc que la suite $(V_n, n \geq 2)$ converge presque sûrement vers $\mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2 = \mathrm{Var}_\theta(X_1) = \theta$. On a

$$\mathbb{E}_\theta[T_n^\lambda] = \lambda\mathbb{E}_\theta[\bar{X}_n] + (1-\lambda)\mathbb{E}_\theta[V_n] = \theta,$$

$$\mathrm{Var}_\theta[T_n^\lambda] = \lambda^2\,\mathrm{Var}_\theta[\bar{X}_n] + (1-\lambda)^2\,\mathrm{Var}_\theta[V_n] + 2\,\mathrm{Cov}_\theta(\bar{X}_n, V_n) = \lambda^2\frac{\theta}{n} + (1-\lambda)^2\frac{2\theta^2}{(n-1)},$$

car $\bar{X}_n$ et V_n sont indépendantes et car $\dfrac{n-1}{\theta}V_n$ suit la loi $\chi^2(n-1)$ de moyenne $n-1$ et de variance $2(n-1)$. La suite d'estimateurs est sans biais. Par continuité de l'application $(x, v) \mapsto \lambda x + (1-\lambda)v$, on en déduit que la suite $(T_n^\lambda, n \geq 2)$ converge presque sûrement vers $\lambda\mathbb{E}_\theta[X_1] + (1-\lambda)\,\mathrm{Var}_\theta(X_1) = \theta$. La suite d'estimateurs est donc convergente.

4. Les variables aléatoires $((X_k, X_k^2), k \geq 1)$ sont de même loi, indépendantes, et de carré intégrable. De plus $\mathbb{E}_\theta[X_k] = \theta$ et $\mathbb{E}_\theta[X_k^2] = \mathrm{Var}_\theta(X_k) + \mathbb{E}_\theta[X_k]^2 = \theta + \theta^2$. On déduit du théorème central limite, que la suite $(Z_n, n \geq 1)$ converge en loi vers un vecteur gaussien centré de matrice de covariance Σ. La matrice Σ est la matrice de covariance de (X_k, X_k^2). On a

$$\begin{aligned}
&\mathrm{Var}_\theta(X_k) = \theta,\\
&\mathrm{Cov}_\theta(X_k, X_k^2) = \mathbb{E}[X_k^3] - \mathbb{E}_\theta[X_k]\mathbb{E}_\theta[X_k^2] = \theta^3 + 3\theta^2 - \theta(\theta+\theta^2) = 2\theta^2,\\
&\mathrm{Var}_\theta(X_k^2) = \mathbb{E}[X_k^4] - \mathbb{E}_\theta[X_k^2]^2 = \theta^4 + 6\theta^3 + 3\theta^2 - (\theta+\theta^2)^2 = 4\theta^3 + 2\theta^2.
\end{aligned}$$

5. La suite $\left(\sqrt{n}(\bar{X}_n - \theta), n \geq 1\right)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \theta)$. (En fait la suite est constante en loi, égale en loi à $\mathcal{N}(0,\theta)$.)
Remarquons que $\frac{n-1}{n}V_n$ est l'image de Z_n par la fonction h, de classe C^1 : $h(x, y) = y - x^2$. En particulier, la suite $(\sqrt{n}(\dfrac{n-1}{n}V_n - \theta), n \geq 2)$ converge en loi vers la loi gaussienne centrée de variance

$$(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y})(\theta, \theta+\theta^2)\;\Sigma\;(\frac{\partial h}{\partial x}, \frac{\partial h}{\partial y})^t(\theta, \theta+\theta^2) = 2\theta^2.$$

Enfin remarquons que $\sqrt{n}V_n - \sqrt{n}\dfrac{n-1}{n}V_n = \dfrac{V_n}{\sqrt{n}}$. En particulier, la suite $(\sqrt{n}V_n - \sqrt{n}\dfrac{n-1}{n}V_n, n \geq 2)$ converge p.s. vers 0. On déduit du théorème de Slutsky et de la continuité de l'addition que la suite $(\sqrt{n}(V_n - \theta), n \geq 2)$ converge en loi vers la loi gaussienne centrée de variance $2\theta^2$. On pose $Y_n = \sqrt{n}(\bar{X}_n - \theta)$ et $W_n = \sqrt{n}(V_n - \theta)$. En utilisant l'indépendance entre Y_n et W_n, on obtient

$$\lim_{n\to\infty} \psi_{(Y_n, W_n)}(y, w) = \lim_{n\to\infty} \psi_{Y_n}(y)\psi_{W_n}(w) = \mathrm{e}^{-\theta y^2/2}\,\mathrm{e}^{-2\theta^2 w^2/2}$$

pour tout $y, w \in \mathbb{R}$. On reconnaît la fonction caractéristique du couple (Y, W), où Y et W sont indépendants de loi respective $\mathcal{N}(0, \theta)$ et $\mathcal{N}(0, 2\theta^2)$. On en déduit que la suite $(\sqrt{n}(\bar{X}_n - \theta, V_n - \theta), n \geq 2)$ converge en loi vers le vecteur gaussien (Y, W). Par continuité de l'application $(a, b) \to \lambda a + (1-\lambda)b$, on en déduit que la suite $(\sqrt{n}(T_n^\lambda - \theta) = \lambda\sqrt{n}(\bar{X}_n - \theta) + (1-\lambda)\sqrt{n}(V_n - \theta), n \geq 2)$ converge en loi vers $\lambda Y + (1-\lambda)W$. Autrement dit, la suite $(\sqrt{n}(T_n^\lambda - \theta), n \geq 2)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \lambda^2\theta + 2(1-\lambda)^2\theta^2)$. Ainsi la suite d'estimateur est asymptotiquement normale de variance asymptotique $\sigma_\lambda^2 = \lambda^2\theta + 2(1-\lambda)^2\theta^2$.

6. On déduit de la question précédente, que sous H_0, la statistique de test converge en loi vers une gaussienne $\mathcal{N}(0, 1)$. Sous H_1, la suite $(T_n^\lambda, n \geq 2)$ converge p.s. vers $\theta > \theta_0$. En particulier $(\zeta_n^\lambda, n \geq 2)$ converge p.s. vers $+\infty$. On déduit du théorème de convergence dominée que les régions critiques

$$W_n = \{\zeta_n^\lambda > c\}$$

définissent un test asymptotique convergent. Ces tests sont de niveau asymptotique $\alpha = \mathbb{P}(\mathcal{N}(0,1) > c)$. On obtient $\zeta_n^\lambda = 0.73$ et une p-valeur de 0.23. On accepte H_0 au niveau de 5%.

7. On trouve $\lambda^* = 2\theta_0/(1+2\theta_0)$ et $\sigma_{\lambda^*}^2 = \dfrac{2\theta_0^2}{1+2\theta_0}$. L'application numérique donne $\zeta_n^{\lambda^*} = 1.86$ et une p-valeur de 3.2%. On rejette H_0 au niveau de 5%.
Les suites d'estimateurs $(\hat{\theta}_n, n \geq 1)$ et $(T_n^{\lambda^*}, n \geq 2)$ ont même variance asymptotique. Remarquons que

$$\hat{\theta}_n = -\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{n-1}{n}V_n + \bar{X}_n^2} = -\frac{1}{2} + \sqrt{(\theta + \frac{1}{2})^2 + (\frac{n-1}{n}V_n - \theta) + (\bar{X}_n^2 - \theta^2)}.$$

En effectuant un développement limité, car p.s. $\lim_{n\to\infty} z_n = 0$, avec $z_n = (\frac{n-1}{n}V_n - \theta) + (\bar{X}_n^2 - \theta^2)$, on obtient

$$\hat{\theta}_n = \theta + \frac{z_n}{1+2\theta} + g(z_n^2),$$

où $|g(z)| \leq z^2/4(1+2\theta)$. On en déduit que sous H_0,

$$\hat{\theta}_n - T_n^{\lambda^*} = \frac{1}{1+2\theta_0}\,\frac{1}{n-1}\,V_n + (\bar{X}_n - \theta_0)^2 + g(z_n).$$

En particulier, pour tout $\varepsilon \in]0, 1/2[$, on a $\lim_{n\to\infty} n^{1-\varepsilon}(\hat{\theta}_n - T_n^{\lambda^*}) = 0$ en probabilité. Ainsi les deux statistiques de test ζ_n' et $\zeta_n^{\lambda^*}$ définissent asymtotiquement les mêmes régions critiques.

On peut aussi remarquer que le modèle considéré est un modèle exponentiel de statistique canonique $\sum_{k=1}^n X_k^2$. De plus la fonction en facteur de la statistisque de test dans la vraisemblance, $Q(\theta) = -1/\theta$, est monotone. En particulier, le test (pur) UPP pour tester H_0 contre H_1 est de région critique $\{\sum_{k=1}^n X_k^2 > c\}$, c'est exactement le test construit à partir de l'estimateur du maximum de vraisemblance. Ainsi le test construit à l'aide de l'estimateur du maximum de vraisemblance a de bonnes propriétés asymptotique, mais il est même UPP à horizon fini.

▲

Exercice IX.6.

1. θ_0 est fixé. Ici $\theta_1 > \theta_0$. Les hypothèses du lemme de Neymann-Pearson sont verifiés. On pose
$$f(x,\theta) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\sum_{i=1}^n -\frac{(x_i-\theta)^2}{2\sigma^2}}.$$
Il vient
$$\frac{f(x,\theta_1)}{f(x,\theta_0)} = \mathrm{e}^{-\sum_{i=1}^n -\frac{((x_i-\theta_1)^2-(x_i-\theta_0)^2)}{2\sigma^2}} = \mathrm{e}^{M\lambda-\frac{1}{2}M^2},$$
avec $M = \frac{\sqrt{n}(\theta_1-\theta_0)}{\sigma}$ et $\lambda(X) = \frac{\sqrt{n}(\bar{X}-\theta_0)}{\sigma}$. On notera λ la variable aléatoire à valeur dans $\mathbb{R}$. On considère la région critique du test $A_\alpha = \{x \in \mathbb{R}^n, \frac{f(x,\theta_1)}{f(x,\theta_0)} \geq c_\alpha(\theta_1)\}$. A_α est aussi l'ensemble $A_\alpha = \{x \in \mathbb{R}^n, \lambda(x) \geq \lambda_\alpha(\theta_1)\}$. Ce test associé à la région critique A_α est le test de niveau α le plus puissant pour tester l'hypothèse $\{\theta = \theta_0\}$ contre $\{\theta = \theta_1\}$. Reste à déterminer λ_α. Sous θ_0, λ est une variable aléatoire de loi gaussienne $\mathcal{N}(0,1)$ Il s'agit donc de choisir $\lambda_\alpha(\theta_1)$ tel que : $\int_{\lambda_\alpha(\theta_1)}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \alpha$. On en déduit que $\lambda_\alpha(\theta_1) = \lambda_\alpha$ est indépendant de θ_1. La région critique définie est indépendante de $\theta_1 > \theta_0$. Ce test est donc aussi un test de niveau α uniformément plus puissant pour tester $\{\theta = \theta_0\}$ contre $\{\theta > \theta_0\}$.

2. Si $\theta = \theta_0$, λ suit une loi gaussienne $\mathcal{N}(0,1)$. Si $\theta = \theta_1$, $\lambda - \frac{\sqrt{n}(\theta_1-\theta_0)}{\sigma}$ suit une loi gaussienne $\mathcal{N}(0,1)$. On souhaite definir n tel que $\mathbb{P}_{\theta_0}(A_\alpha) = \alpha$ et $\mathbb{P}_{\theta_1}(\bar{A_\alpha}) = \beta$, c'est à dire
$$\int_{\lambda_\alpha}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{u^2}{2}} du = \alpha \quad \text{et} \quad \int_{-\infty}^{\lambda_\alpha - \frac{\sqrt{n}(\theta_1-\theta_0)}{\sigma}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{u^2}{2}} du \leq \beta = \int_{-\infty}^{\lambda_\beta} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{u^2}{2}} du.$$
Cette procédure nous permet ainsi de choisir n tel que $\lambda_\beta \geq \lambda_\alpha - \frac{\sqrt{n}(\theta_1-\theta_0)}{\sigma}$. On choisit donc $n \geq \sigma^2(\lambda_\beta - \lambda_\alpha)^2/(\theta_1-\theta_0)^2$.

3. Les régions critiques définies pour le test de niveau α uniformément plus puissant de $\theta = \theta_0$ contre $\theta > \theta_0$ et le test de niveau α uniformement plus puissant de $\theta = \theta_0$ contre $\theta < \theta_0$ sont distinctes.

4. (a) L'ensemble $\mathcal{X}$ peut aussi s'écrire :
$$\mathcal{X} = \{x \in \mathbb{R}^n / e^{M\lambda(x)-\frac{1}{2}M^2} \geq c + c^* \frac{\sqrt{n}}{\sigma}\lambda(x)\}.$$
Pour tout $\theta \in \mathbb{R}$, sous $\mathbb{P}_\theta$, λ admet comme distribution la loi $\mathcal{N}(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma}, 1)$. Soit $\lambda_{\frac{\alpha}{2}}$ tle que $\mathbb{P}_{\theta_0}(|\lambda| \geq \lambda_{\frac{\alpha}{2}}) = 2\int_{\lambda_{\frac{\alpha}{2}}}^{\infty} \frac{1}{2\pi} e^{-\frac{u^2}{2}} du = \alpha$, on détermine uniquement les valuers de $c_{\frac{\alpha}{2}}(\theta_1)$, $c^*_{\frac{\alpha}{2}}(\theta_1)$, telles que $\mathcal{X} = \{x \in \mathbb{R}^n / |\lambda(x)| \geq \lambda_{\frac{\alpha}{2}}\}$ (ensemble indépendant de θ_1). Et ceci pour tout $\theta_1 \neq \theta_0$. On a
$$\frac{\partial \mathbb{P}(X \in \mathcal{X}]}{\partial \theta}_{|\theta_0} = \frac{\partial \int \mathbf{1}_{|u| \geq \lambda_{\frac{\alpha}{2}}} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(u-\frac{\sqrt{n}(\theta-\theta_0)}{\sigma})^2}{2}} du}{\partial \theta}_{|\theta_0} = 0.$$

(b) Soit $\mathcal{S}$ une autre région critique de niveau α, définissant un test sans biais, $\forall\theta$, $\mathbb{P}_\theta(\mathcal{S}) \geq \alpha$, et $\mathbb{P}_{\theta_0}(\mathcal{S}) = \alpha$. Nos hypothèses (fonctions régulières et dérivation sous le signe somme) nous permettent d'affirmer : $\frac{\partial P_\theta(\mathcal{S})}{\partial\theta}_{|\theta_0} = 0$. On a donc

$$\frac{\partial \mathbb{P}_\theta(\mathcal{X} - \mathcal{S}\cap\mathcal{X})}{\partial\theta}_{|\theta_0} = \frac{\partial \mathbb{P}_\theta(\mathcal{S} - \mathcal{S}\cap\mathcal{X})}{\partial\theta}_{|\theta_0} \quad \text{et} \quad \mathbb{P}_{\theta_0}(\mathcal{X} - \mathcal{S}\cap\mathcal{X}) = \mathbb{P}_{\theta_0}(\mathcal{S} - \mathcal{S}\cap\mathcal{X}).$$

Soit $A \subset \mathcal{X}$. Alors pour tout $x \in A$, on

$$f(x,\theta_1) \geq cf(x,\theta_0) + c^* \frac{\partial f(x,\theta)}{\partial\theta}_{|\theta_0}.$$

En intégrant sur A, et en permutant l'intégrale en x et la dérivation en θ, il vient

$$\int_A f(x,\theta_1)\,dx \geq c\int_A f(x,\theta_0)\,dx + c^* \frac{\partial \int_A f(x,\theta)\,dx}{\partial\theta}_{|\theta_0}.$$

Si $A \subset \mathcal{X}^\rfloor$, alors on obtient

$$c\int_A f(x,\theta_0)\,dx + c^* \frac{\partial \int_A f(x,\theta)\,dx}{\partial\theta}_{|\theta_0} \geq \int_A f(x,\theta_1)\,dx.$$

En utilisant ces deux inégalités et les deux égalités précédentes, il vient

$$\begin{aligned}
\mathbb{P}_{\theta_1}(\mathcal{X} - \mathcal{S}\cap\mathcal{X}) &\geq c\mathbb{P}_{\theta_0}(\mathcal{X} - \mathcal{S}\cap\mathcal{X}) + c^* \frac{\partial \mathbb{P}_\theta(\mathcal{X} - \mathcal{S}\cap\mathcal{X})}{\partial\theta}_{|\theta_0} \\
&\geq c\mathbb{P}_{\theta_0}(\mathcal{S} - \mathcal{S}\cap\mathcal{X}) + c^* \frac{\partial \mathbb{P}_\theta(\mathcal{S} - \mathcal{S}\cap\mathcal{X})}{\partial\theta}_{|\theta_0} \\
&\geq \mathbb{P}_{\theta_1}(\mathcal{S} - \mathcal{S}\cap\mathcal{X}).
\end{aligned}$$

Finalement, on en déduit

$$\mathbb{P}_{\theta_1}(\mathcal{X}) \geq \mathbb{P}_{\theta_1}(\mathcal{S}).$$

Le test pur de région critique $\mathcal{X}$ est plus puissant que n'importe quel autre test pur sans biais.

▲

Exercice IX.7.

1. La fonction de vraisemblance est :

$$p(l_1;\bar{L}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(l_1-\bar{L})^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{l_1^2+\bar{L}^2}{2\sigma^2}} e^{\frac{\bar{L}}{\sigma^2}l_1}.$$

Il s'agit donc bien d'un modèle exponentiel de statistique canonique $S = \sum_{i=1}^n L_i$. L'estimateur du maximum de vraisemblance de $\bar{L}$ est :

$$\hat{\bar{L}}_n = \frac{1}{n}\sum_{i=1}^n L_i.$$

2. On peut donc définir un test UPPS au seuil α pour tester H_0 contre H_1, à l'aide de la statistique de test S/n : pour $l = (l_1, \ldots, l_n)$,

$$\begin{array}{lll} \varphi(l) = 0 & si & S(l)/n \in [c_1, c_2] \\ \varphi(l) = \gamma_i & si & S(l)/n = c_i \\ \varphi(l) = 1 & si & S(l)/n \notin [c_1, c_2] \end{array}$$

les constantes γ_i et c_i étant déterminées par les conditions $\mathbb{E}_{\bar{L}_0 - \delta L}[\varphi(L_1, \ldots, L_n)] = \alpha = \mathbb{E}_{\bar{L}_0 + \delta L}[\varphi(L_1, \ldots, L_n)]$. Ici, S est la somme de variables aléatoires gaussiennes indépendantes, donc c'est une variable aléatoire gaussienne donc continue : on peut prendre les γ_i nuls. Le test UPPS est donc un test pur de région critique :

$$W = \{l; S(l)/n \notin [c_1, c_2]\}.$$

On détermine c_1 et c_2 par les relations :

$$\mathbb{P}_{(\bar{L}_0 - \delta L)}(W) = 1 - \int_{\sqrt{n}(c_1 - (\bar{L}_0 - \delta L))/\sigma_0}^{\sqrt{n}(c_2 - (\bar{L}_0 - \delta L))/\sigma_0} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = \alpha$$

et

$$\mathbb{P}_{(\bar{L}_0 + \delta L)}(W) = 1 - \int_{\sqrt{n}(c_1 - (\bar{L}_0 + \delta L))/\sigma_0}^{\sqrt{n}(c_2 - (\bar{L}_0 + \delta L))/\sigma_0} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = \alpha.$$

On cherche c_1 sous la forme $c_1 = \bar{L}_0 - \dfrac{\varepsilon_1 \sigma_0}{\sqrt{n}}$ et c_2 sous la forme $c_2 = \bar{L}_0 + \dfrac{\varepsilon_2 \sigma_0}{\sqrt{n}}$. On a donc à résoudre :

$$\int_{\delta L \sqrt{n}/\sigma_0 - \varepsilon_1}^{\delta L \sqrt{n}/\sigma_0 + \varepsilon_2} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = 1 - \alpha \quad \text{et} \quad \int_{-\delta L \sqrt{n}/\sigma_0 - \varepsilon_1}^{-\delta L \sqrt{n}/\sigma_0 + \varepsilon_2} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = 1 - \alpha.$$

Comme $\delta L \neq 0$, on déduit du changement de variable $u = -y$ dans la première intégrale que $\varepsilon_1 = \varepsilon_2 = \varepsilon$ et que :

$$\int_{-\delta L \sqrt{n}/\sigma_0 - \varepsilon}^{-\delta L \sqrt{n}/\sigma_0 + \varepsilon} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = 1 - \alpha.$$

3. L'application numérique donne :

$$\int_{-10.61 - \varepsilon}^{-10.61 + \varepsilon} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = 0.95.$$

On en déduit que $\varepsilon > 10.61$ car $\int_{-\infty}^{0} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = 0.5 < 0.95$, donc la borne inférieure de l'intégrale est < -21, ce qui revient numériquement à la prendre infinie. On cherche donc ε tel que :

$$\int_{-\infty}^{-10.61 + \varepsilon} e^{-y^2/2} \frac{dy}{\sqrt{2\pi}} = 0.95$$

ce qui donne $-10.61 + \varepsilon = 1.64$, c'est-à dire $\varepsilon = 12.25$. On a donc $c_1 = 783.3$ et $c_2 = 786.7$. Comme $\hat{\bar{L}}_n = 788.3 \notin [c_1, c_2]$, on rejette l'hypothèse H_0. La machine nécessite donc d'être révisée. La p-valeur associée à ce test est de l'ordre de 10^{-37}.

4. On sait que l'estimateur $\hat{\bar{L}}$ de $\bar{L}$ est sans biais et efficace, et que sa loi est une loi gaussienne $\mathcal{N}\left(0, \sigma_0^2/n\right)$. On a donc $\mathcal{L}\left(\frac{\sqrt{n}(\hat{\bar{L}}-\bar{L})}{\sigma_0}\right) = \mathcal{N}(0,1)$, ce qui donne :

$$\mathbb{P}(\sqrt{n}(\hat{\bar{L}} - \bar{L})/\sigma_0 \in [-a, a]) = \int_{-a}^{a} e^{-\frac{y^2}{2}} \frac{dy}{\sqrt{2\pi}} = 1 - \alpha.$$

On a $a = 1.96$ pour $1 - \alpha = 95\%$. L'intervalle de confiance de $\bar{L}$ est donc $[\hat{\bar{L}} \pm \frac{1.96\sigma_0}{\sqrt{n}}] = [788.0, 788.6]$. Comme $[\bar{L}_0 - \delta L, \bar{L}_0 + \delta L] \cap [788.0, 788.6] = \emptyset$, on confirme le fait que la machine doit être révisée.

▲

Exercice IX.8.

1. On suppose que les variables sont indépendantes et de même loi.
2. On considère l'estimateur de la moyenne empirique $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Comme les $(X_i, i \in \{1 \dots, 64\})$ sont des variables aléatoires indépendantes de même loi gaussiennes, $\hat{\mu}$ suit une loi normale de moyenne μ, et de variance $\frac{\sigma^2}{n}$.
3. On teste l'hypothèse nulle $H_0 :\ \mu \leq \mu_0$ (le nouveau modèle n'est pas plus performant que l'ancien) contre $H_1 :\ \mu > \mu_0$. Dans ce cas l'erreur de première espèce est d'adopter le nouveau modèle alors qu'il n'est pas plus performant que l'ancien. Il s'agit d'un test unilatéral UPP dans un modèle exponentiel. On choisit alors une région critique de la forme $W = \{\hat{\mu} > k\}$, où k est tel que l'erreur de première espèce est de 5%. On a donc $\sup_{\mu \in H_0} \mathbb{E}_\mu(\mathbf{1}_W) = \sup_{\mu \leq \mu_0} \mathbb{P}_\mu(W) = 5\%$. Or, comme $\hat{\mu}$ est une variable normale, il vient

$$\mathbb{P}_\mu(W) = \mathbb{P}_\mu\left(\hat{\mu} > k\right) = \mathbb{P}_\mu\left(\sqrt{n}\left(\frac{\hat{\mu} - \mu}{\sigma}\right) > \sqrt{n}\frac{k - \mu}{\sigma}\right).$$

Comme $\sqrt{n}\left(\frac{\hat{\mu}-\mu}{\sigma}\right)$ suit une loi $\mathcal{N}(0,1)$, on a

$$\mathbb{P}_\mu(W) = 1 - N\left(\sqrt{n}\left(\frac{k - \mu}{\sigma}\right)\right),$$

où N est la fonction de répartition de la loi normale. On remarque également qu'il s'agit d'une fonction croissante de μ donc $\sup_{\mu \leq \mu_0} \mathbb{P}_\mu(W) = \mathbb{P}_{\mu_0}(W)$. Pour un seuil de 5%, k est donc déterminé par l'équation :

$$1 - N\left(\sqrt{n}\left(\frac{k - \mu_0}{\sigma}\right)\right) = 0.05 \Leftrightarrow \sqrt{n}\left(\frac{k - \mu_0}{\sigma}\right) = u_{95\%} \Leftrightarrow k = \frac{u_{95\%}}{\sqrt{n}}\sigma + \mu_0.$$

Pour l'application numérique, on sait que $u_{95\%} = N^{-1}(95\%) = 1.64$. On a donc $k = 123.9 > \hat{\mu} = 123.5$. On accepte donc l'hypothèse H_0, au vu des observations le nouveau modèle n'est pas plus performant que l'ancien. La p-valeur de ce test est $p = 0.07$. Ceci confirme que l'on rejette H_0 au niveau de 5%, mais cela montre également que le choix de 5% est presque critique.

4. On demande également d'évaluer l'erreur de deuxième espèce pour $\mu = \mu_0$, c'est à dire la probabilité de rejeter le nouveau modèle sachant qu'il est plus performant (i.e. l'annonce du représentant est exacte). Il s'agit donc de

$$1 - \mathbb{P}_{\mu=1.05\mu_0}(\hat{\mu} > k) = N\left(\sqrt{n}\left(\frac{k - 1.05\mu_0}{\sigma}\right)\right).$$

Utilisant la valeur de k trouvée précédemment, on a $\mathbb{P}_{\mu=1.05\mu_0}(\hat{\mu} < k) = N(-0,86) = 0,195 \simeq 20\%$. Il s'agit du risque du vendeur, i.e. le risque que sa machine ne soit pas achetée alors qu'elle est 5% meilleure que l'ancienne.

5. L'hypothèse à tester est $H_0' :\ \mu = 1.05\mu_0$ contre $H_1 :\ \mu \leq \mu_0$. Dans ce cas la région critique est de la forme $W = \{\hat{\mu} < k\}$. On cherche k tel que l'erreur de première espèce soit 0.05 ce qui donne $\mathbb{P}_{1.05\mu_0}(W) = 0.05$. En raisonnant comme précédemment (on rappelle que $\sqrt{n}\frac{\hat{\mu}-1.05\mu_0}{\sigma}$ est une gaussienne centrée réduite sous H_0'), on aboutit à

$$k = 1.05\mu_0 - \frac{\sigma N^{-1}(0.05)}{\sqrt{n}} \simeq 122.02.$$

On rejette H_0' si $\hat{\mu} < 122.02$. Au vue des données, on accepte l'hypothèse $\mu = 1.05\mu_0$: le nouveau modèle est donc plus performant que l'ancien. La p-valeur de ce test est $p = 0.19$. Elle confirme qu'il n'est pas raisonnable de rejeter H_0.
Le risque de deuxième espèce est alors $\sup_{\mu\leq\mu_0} \mathbb{P}_\mu(\hat{\mu} > k) = \mathbb{P}_{\mu_0}(\hat{\mu} > k)$. En utilisant la valeur $k = 122.02$, on obtient $\mathbb{P}_{\mu_0}(\hat{\mu} > k) = 1 - N\left(\sqrt{n}\left(\frac{k-\mu_0}{\sigma}\right)\right) \simeq 20\%$. L'acheteur a alors au plus 20% de chance d'accepter l'annonce du représentant alors que celle-ci est fausse.

6. Dans la question 3, le risque de l'acheteur est $\mathbb{P}_{\mu_0}(\hat{\mu} > k)$ et dans la question 4, celui du vendeur est $\mathbb{P}_{1.05\mu_0}(\hat{\mu} < k)$. On peut donc chercher k tel que ces deux risques soient égaux. Or on sait que $\mathbb{P}_{\mu_0}(\hat{\mu} > k) = 1 - N\left(\sqrt{n}\frac{k-\mu_0}{\sigma}\right)$ et $\mathbb{P}_{1.05\mu_0}(\hat{\mu} < k) = N\left(\sqrt{n}\frac{k-1.05\mu_0}{\sigma}\right)$. L'égalité des deux probabilités donne $k - \mu_0 = 1.05\mu_0 - k$ et donc $k = 123$. Dans ce cas le risque vaut $N((123 - 126) * \sqrt{64}/19.4) = 11\%$.

▲

Exercice IX.9.

7. On dispose donc d'un deuxième échantillon $Y_1, ..., Y_m$ de moyenne ν et on désire tester $H_0 :\ \mu = \nu$ contre $\mu \neq \nu$. L'échantillon global est donc $(X_1, ..., X_n, Y_1, ..., Y_m)$ et la densité du modèle est

$$p(x, y; \mu, \nu) = \frac{1}{(2\pi)^{n+m}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x - \mu)^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y - \nu)^2\right).$$

Pour $\theta = (\mu, \nu)$, on définit $\Theta_0 = \{\theta :\ \mu = \nu\}$ et $\Theta_1 = \mathbb{R}^2 - \Theta_0$.
L'estimateur de vraisemblance dans le cas $\mu = \nu$ est $\hat{\theta} = \frac{1}{m+n}\left(\sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i\right)$ et les estimateurs dans le cas général sont $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ et $\hat{\nu} = \frac{1}{n}\sum_{i=1}^{m} Y_i$. On remarque que $(m + n)\hat{\theta} = n\hat{\mu} + m\hat{\nu}$.

8. Pour construire la région critique on procède par la méthode du rapport de vraisemblance : on recherche une région critique de la forme

$$W = \{(x,y)|\ p_n(x,y,\hat{\mu},\hat{\nu}) > kp_n(x,y,\hat{\theta},\hat{\theta})\}.$$

En calculant le quotient des deux densités on obtient

$$W = \{(x,y)|\ \sum_{i=1}^{n}(x_i-\hat{\mu})^2 - \sum_{i=1}^{n}(x_i-\hat{\theta})^2 + \sum_{i=1}^{m}(y_i-\hat{\nu})^2 - \sum_{i=1}^{m}(y_i-\hat{\theta})^2 < k'\}.$$

En utilisant les identités

$$\sum_{i=1}^{n}(x_i-\hat{\theta})^2 = \sum_{i=1}^{n}(x_i-\hat{\mu})^2 + n(\hat{\mu}-\hat{\theta})^2$$
$$\sum_{i=1}^{m}(y_i-\hat{\theta})^2 = \sum_{i=1}^{m}(y_i-\hat{\nu})^2 + m(\hat{\nu}-\hat{\theta})^2$$

et le fait que $(m+n)(\hat{\theta}-\hat{\mu}) = m(\hat{\nu}-\hat{\mu})$ et $(m+n)(\hat{\theta}-\hat{\nu}) = m(\hat{\mu}-\hat{\nu})$ on obtient que la région critique est de la forme

$$W = \left\{(x,y)|\ \sqrt{\frac{mn}{m+n}}|\hat{\nu}-\hat{\mu}| > c\sigma\right\}.$$

Or sous l'hypothèse H_0, $\hat{\nu}-\hat{\mu}$ suit une loi normale de moyenne nulle et de variance $\frac{m+n}{mn}\sigma^2$, donc pour un risque de 0.05, on obtient que $\mathbb{P}_{H_0}(W) = 0.05$ implique $c = 1 - N^{-1}(0.025) = 1.96$. Avec les valeurs données dans l'énoncé, on trouve que $\frac{1}{\sigma}\sqrt{\frac{mn}{m+n}}|\hat{\nu}-\hat{\mu}| = 0.36$ et donc on accepte l'hypothèse nulle : les deux machines sont équivalentes. La p-valeur de ce test est $p = 0.70$, ceci confirme que l'on en peut pas rejeter H_0.

▲

Exercice IX.10.

Les statistiques

$$\xi_n^{(1)} = n\sum_{j=0}^{j=9}\frac{(\hat{p}_j-p_j)^2}{\hat{p}_j} \quad \text{et} \quad \xi_n^{(2)} = n\sum_{j=0}^{j=9}\frac{(\hat{p}_j-p_j)^2}{p_j},$$

où $(\hat{p}_j, j\in\{0,\dots,9\})$ sont les fréquences empiriques observées et $p_j = 1/10$ sont les fréquences théoriques, convergent en loi vers un χ^2 à $10-1=9$ degrés de liberté. Les tests définis par les régions critiques :

$$W_i = \{\xi_n^{(i)} \geq z_\alpha(9)\}, \quad i\in\{1,2\}$$

sont convergents de niveau asymptotique α, avec $z_\alpha(9)$ défini par :

$$\mathbb{P}(\chi^2(9) \geq z_\alpha(9)) = \alpha.$$

L'application numérique donne $\xi_n^{(1)} = 11.1$ et $\xi_n^{(2)} = 10.4$. Au seuil $\alpha = 5\%$, on lit dans la table des quantiles de la loi du χ^2 : $z_\alpha(9) = 16.92$. Comme $z_\alpha(9) > \xi_n^{(i)}$ pour $i\in\{1,2\}$, on accepte l'hypothèse de distribution uniforme pour chacun des tests. Les p-valeurs de ces deux tests sont $p^{(1)} = 0.27$ et $p^{(2)} = 0.32$. Ces valeurs confirme qu'il n'est pas raisonnable de rejeter H_0.

▲

Exercice IX.11.

1. Puisque l'on est sous H_0, la loi des réponses des deux juges est la même. Notons β la probabilité de répondre par l'affirmative (notée $p_+ = p_+^{(1)} = p_+^{(2)}$), alors $p_- = p_-^{(1)} = p_-^{(2)} = 1-\beta$. Notons $p_{+-} = \alpha$ et remarquons ensuite que $p_+ = p_+^{(1)} = p_{++} + p_{+-}$. Ainsi $p_{++} = \beta - \alpha$. De même $p_{--} = 1-\beta-\alpha$, puis finalement $p_{-+} = p_+ - p_{++} = \beta - \beta - \alpha = \alpha$.
2. La vraisemblance de l'échantillon $x = (x_1, \dots, x_n)$ où $x_i \in \{-,+\}^2$ est donnée par

   $$p_n(x;\alpha,\beta) = (\beta-\alpha)^{\sum_{i=1}^n \mathbf{1}_{\{x_i=(+,+)\}}} \alpha^{\sum_{i=1}^n \mathbf{1}_{\{x_i=(-,+)\}} + \mathbf{1}_{\{x_i=(+,-)\}}} (1-\beta-\alpha)^{\sum_{i=1}^n \mathbf{1}_{\{x_i=(-,-)\}}}.$$

   On remarque que $N_{++} = \sum_{i=1}^n \mathbf{1}_{\{x_i=(+,+)\}}$ et $N_{--} = \sum_{i=1}^n \mathbf{1}_{\{x_i=(-,-)\}}$. De plus le couple (N_{++}, N_{--}) est la statistisque canonique du modèle. La log-vraisemblance $L_n(x;\alpha,\beta)$ peut alors s'écrire

   $$L_n(x,\alpha,\beta) = N_{++}\log(\beta-\alpha) + (n - N_{--} - N_{++})\log\alpha + N_{++}\log(1-\beta-\alpha).$$

   On cherche à annuler les deux dérivées partielles

   $$\begin{cases} \dfrac{-N_{++}}{\beta-\alpha} + \dfrac{n-N_{--}-N_{++}}{\alpha} - \dfrac{N_{--}}{1-\beta-\alpha} &= 0, \\ \dfrac{N_{++}}{\beta-\alpha} - \dfrac{N_{--}}{1-\beta-\alpha} &= 0. \end{cases}$$

   Ce système est équivalent à

   $$\begin{cases} \dfrac{n-N_{--}-N_{++}}{\alpha} - \dfrac{2N_{--}}{1-\beta-\alpha} &= 0, \\ -\dfrac{2N_{++}}{\beta-\alpha} + \dfrac{n-N_{--}-N_{++}}{\alpha} &= 0. \end{cases}$$

   Ce système se résoud alors facilement et on trouve

   $$\begin{cases} \alpha &= \dfrac{n-N_{--}-N_{++}}{2n}, \\ \beta &= \dfrac{n+N_{++}-N_{--}}{2n}. \end{cases}$$

3. L'hypothèse H_0 est équivalente à l'hypothèse : $p_{-+} = p_{+-}$. On vient de voir que sous H_0, l'estimateur du maximum de vraisemblance de p_{-+} est $\dfrac{n-N_{--}-N_{++}}{2n}$ que l'on peut réécrire $\dfrac{N_{+-}+N_{-+}}{2n}$. On peut alors calculer la statistique de test ζ_n du χ^2.

   $$\begin{aligned} \zeta_n &= n\left(\frac{\left(\frac{N_{+-}}{n} - \frac{N_{+-}+N_{-+}}{2n}\right)^2}{\frac{N_{-+}+N_{+-}}{2n}} + \frac{\left(\frac{N_{-+}}{n} - \frac{N_{+-}+N_{-+}}{2n}\right)^2}{\frac{N_{-+}+N_{+-}}{2n}} \right) \\ &= \frac{(N_{+-}-N_{-+})^2}{N_{-+}-N_{+-}}. \end{aligned}$$

   Sous H_0, ζ_n tend en loi vers χ^2 à 4-1-2 degré de liberté (on a retiré 2 degrés de liberté pour l'estimation de α et β). Considérons le test asymptotique de niveau 0.05 et $a = 3.84$ le quantile d'ordre 0.95 de la loi $\chi^2(1)$ (i.e. $\mathbb{P}(\chi^2(1) \le a) = 0.95$). La région critique est $[a, \infty[$.

4. la $p-$valeur du test est donnée par $\mathbb{P}(Z \geq \zeta_n^{\text{obs}})$ où Z suit une loi du $X^2(1)$, et ζ_n^{obs} est la statistique calculée sur les données observées. On a $\zeta_n^{\text{obs}} = 5$. On rejette donc l'hypothèse nulle au seuil de 5%. La p-valeur du test est d'environ 2.5%.

▲

Exercice IX.12.

1. Le nombre total d'enfants observés est $n = 5 \times 320 = 1600$. Le nombre total de garçons observés est $n_g = 18 \times 5 + 52 \times 4 + 110 \times 3 + 88 \times 2 + 35 \times 1 = 839$. Donc, la proportion de garçons observés est $\hat{r} = n_g/n = 839/1600 = 0.524375$. La statistique du χ^2 empirique est

$$\xi_n^{(1)} = n \sum_{i=1}^{2} \frac{(\hat{p}_i - p_i)^2}{\hat{p}_i},$$

où $\hat{p}_1 = \hat{r}$ est la fréquence empirique des garçons, $\hat{p}_2 = 1 - \hat{r}$, la fréquence empirique des filles, et $p_1 = p_2 = 1/2$. Or, on sait que la statistique $(\xi_n^{(1)}, n \geq 1)$ converge en loi vers un χ^2 avec $2 - 1 = 1$ degré de liberté. Donc, $W = \{\xi_n^{(1)} > z_\alpha(1)\}$ est la région critique du test du χ^2.

Application numérique : $p_1 = p_2 = 1/2$, $\hat{p}_1 = \hat{r} = 0.524$, $\hat{p}_2 = 1 - \hat{r} = 0.476$, $\xi_{1600}^{(1)} = 3.81$. Pour $\alpha = 0.01$, $z_\alpha(1) = 6.63$. Donc, on accepte l'hypothèse $r = 1/2$: il y a autant de garçons que de filles à la naissance. La p-valeur de ce test (la plus grande valeur de α qui permette d'accepter ce test) est $p = 0.051$.

2. $\hat{r}$ est un estimateur de r. On déduit du théorème central limite et du théorème de Slutsky que $(\sqrt{n}(\hat{r}_n - r)/\sqrt{\hat{r}_n(1 - \hat{r}_n)}, n \geq 1)$ converge en loi vers vers la loi gaussienne $\mathcal{N}(0, 1)$. Un intervalle de confiance pour r de niveau asymptotique α est :

$$IC_\alpha = \left[\hat{r} - a_\alpha \sqrt{\frac{\hat{r}(1-\hat{r})}{n}}; \hat{r} + a_\alpha \sqrt{\frac{\hat{r}(1-\hat{r})}{n}} \right] .$$

Application numérique : pour $\alpha = 0.01$, $a_\alpha = 2.5758$. On a déjà calculé $\hat{r}$: $\hat{r} = 0.524$. On obtient donc :

$$IC_{0.01} = [0.492; 0.557] .$$

On s'aperçoit que $1/2 \in IC_{0.01}$. Vérifions que, pour cette question, il y a équivalence entre le test du χ^2 et l'intervalle de confiance :

$$\begin{aligned} \xi_n^{(1)} &= n \sum_{i=1}^{2} \frac{(\hat{p}_i - p_i)^2}{\hat{p}_i} \\ &= n \left(\frac{(\hat{r} - 1/2)^2}{\hat{r}} + \frac{(\hat{r} - 1/2)^2}{1 - \hat{r}} \right) \\ &= \left(\sqrt{n} \frac{\hat{r} - 1/2}{\sqrt{\hat{r}(1-\hat{r})}} \right)^2 . \end{aligned}$$

Comme $a_\alpha^2 = z_\alpha(1)$, nous voyons bien qu'il y a équivalence entre les deux approches.

Répartition (G,F)	(5,0)	(4,1)	(3,2)	(2,3)	(1,4)	(0,5)	Total
Nbre de familles	18	52	110	88	35	17	320
Prop. observées	0.05625	0.1625	0.34375	0.27500	0.10938	0.05312	1
Prop. théoriques	1/32	5/32	10/32	10/32	5/32	1/32	1

TAB. IX.7 – Proportions observées et théoriques

3. On calcule d'abord les probabilités théoriques : *cf.* tableau IX.7.
   La statistique du χ^2 empirique est :

$$\begin{aligned}\xi_n^{(2)} &= n\sum_{i=1}^{6}\frac{(\hat{p}_i - p_i)^2}{p_i} \\ &= 320\left(\frac{(0.056250-1/32)^2}{1/32}+\frac{(0.1625-5/32)^2}{5/32}+\frac{(0.34375-10/32)^2}{1/32}\right. \\ &\qquad \left.+\frac{(0.275-10/32)^2}{10/32}+\frac{(0.10938-5/32)^2}{5/32}+\frac{(0.05312-1/32)^2}{1/32}\right)\end{aligned}$$

   Or, on sait que la statistique $(\xi_n^{(2)}, n \geq 1)$ converge en loi vers un χ^2 avec $6-1=5$ degrés de liberté. Donc, $W = \{\xi_n^{(2)} > z_\alpha(5)\}$ est la région critique du test du χ^2.
   Application numérique : $\xi_{320}^{(2)} = 18.3$. Pour $\alpha = 0.01$, $z_\alpha(5) = 15.09$. La p-valeur de ce test est $p2 = 0.003$. On peut également calculer $\xi_{320}^{(1)} = 14.5$. La p-valeur associée est $p1 = 0.008$. Donc, on rejette l'hypothèse $r = 1/2$ au niveau $\alpha = 1\%$.
   Conclusion : on obtient un test plus fin en gardant les cases, c'est-à-dire en tenant compte de toute l'information contenue dans les données. Ce principe est utilisé pour les sondages d'opinion.

4. Nous souhaitons tester si les données suivent la loi binomiale de paramètres 5 et r. Les probabilités théoriques sont alors $p_1(r) = r^5$, $p_2(r) = (1-r)r^4C_5^1, \ldots, p_6(r) = (1-r)^5$. L'estimateur du maximum de vraisemblance de r est $\hat{r}$. La statistique du χ^2

$$\xi_n^{(2)} = n\sum_{i=1}^{6}\frac{(\hat{p}_i - p_i(\hat{r}))^2}{p_i(\hat{r})},$$

   converge en loi vers la loi du χ^2 à $6-1-1=4$ degrés de liberté. On a retiré un degré de liberté pour l'estimation du paramètre r de dimension 1.
   Application numérique : $\xi_{320}^{(2)} = 15.9$ et pour $\alpha = 0.01$, $z_\alpha(4) = 13.28$. Donc, on rejette l'hypothèse selon laquelle les naissances des garçons suit une loi binomiale. La p-valeur de ce test est $p2 = 0.003$. On peut émettre l'hypothèse qu'il existe des familles à garçons et des familles à filles.
   On peut également calculer $\xi_{320}^{(1)} = 9.7$. La p-valeur associée est $p1 = 0.045$. On devrait alors accepter H_0 au niveau $\alpha = 1\%$. La différence entre les deux statistiques provient de l'écart important entre le nombre escompté de famille à 5 garçons, 13, contre 18 observées réellement et entre le nombre escompté de famille à 5 filles, 8, contre 17 observées réellement. Les valeurs $\hat{p}_5$ et $\hat{p}_0$ sont trés largement supérieures à $p_5(\hat{r})$ et $p_0(\hat{r})$.

En particulier comme ces quantitées interviennent au dénominateur des statistiques, cela implique que $\xi^{(1)}_{320}$ est bien plus faible que $\xi^{(2)}_{320}$. Ainsi la statistique $\xi^{(1)}_{320}$ sous estime l'écart entre $\hat{p}$ et $p(\hat{r})$. Si par exemple, on regroupe les cases '5 garçons' et '5 filles', on obtient les p-valeurs suivantes : pour la statistique $\xi^{(2)}_{320} = 13.1$, $p2 = 0.004$ et pour la statistique $\xi^{(1)}_{320} = 9.2$, $p1 = 0.027$ (le nombre de degrés de liberté est ici $5 - 1 - 1 = 3$). On remarque que la p-valeur $p2$ est peu modifiée par cette transformation alors que la p-valeur $p1$ varie, en valeur absolue, de manière importante. Cela signifie que la statistique $\xi^{(1)}_{320}$ est peu fiable, et il convient dans ce cas précis de conserver les résultats donnés par la statistique $\xi^{(2)}_{320}$: on rejette donc H_0 au niveau $\alpha = 1\%$.

▲

Exercice IX.13.

1. Soit p la proportion de femmes du jury. On teste : $H_0 = \{p = p_0\}$ avec $p_0 = 0.29$ contre $H_1 = \{p \neq p_0\}$ au niveau α. Il s'agit d'un test bilatéral. Soit X la variable aléatoire donnant, pour tous les jurys, le nombre de femmes qu'il contient. Sous H_0 : X suit une loi binomiale de paramètres $\mathcal{B}(n, p_0) = \mathcal{B}(700, 0.29)$. On considère la variable $Z_n = \dfrac{X - np_0}{\sqrt{np_0(1-p_0)}}$. En utilisant l'approximation gaussienne, on en déduit que sous H_0, la loi de Z_n est approximativement la loi gaussienne $\mathcal{N}(0,1)$. Sous H_1, en revanche on vérifie aisément que p.s. $\lim_{n\to\infty} |Z_n| = +\infty$, et donc $\mathbb{P}_p(|Z_n| \leq a)$ tend vers 0 quand n tend vers l'infini. On peut donc contruire un test asymptotique convergent de région critique $W = \{|z| > c\}$. Le test est de niveau saymptotique α pour c égal à $u_{1-\frac{\alpha}{2}}$, le quantile d'ordre $\left(1 - \frac{\alpha}{2}\right)$ de la loi normale centrée réduite.
   Application numérique. On observe $z_n = -8.16$, et la p-valeur est donnée par $\mathbb{P}(|G| > z_n) = 3\ 10^{-16}$, où G est une variable gaussienne cantrée réduite. On rejette donc l'hypothèse d'impartialité du juge.
2. Dans ce cas, on a $z_n = -1.95$. Pour $\alpha = 5\%$, on a $u_{1-\frac{\alpha}{2}} = 1.96$, on ne peut donc pas rejetter l'hypothèse d'impartialité du juge au niveau 5%.

▲

Exercice IX.14.

Dans un premier temps on regroupe les cases afin d'avoir au moins 5 individus par case. On considère donc le tableau suivant :

Vaccin	Réaction légère	Réaction moyenne	Ulcération ou Abcès	Total
A	12	156	9	177
B	29	135	7	171
Total	41	291	16	348

On note l, m et u les réactions suivantes : légères, moyennes et ulcération ou abcès. Le paramètre du modèle est donc le vecteur des fréquences $p = (p_{l,A}, p_{m,A}, p_{u,A}, p_{l,B}, p_{m,B}, p_{u,B})$ où $p_{i,*}$ est la probabilité d'avoir le vaccin $* \in \{A, B\}$ et la réaction $i \in \{l, m, u\}$. On désire savoir si les vaccins sont égaux (hypothèse H_0) c'est-à-dire si $p_{l|A} = p_{l|B}$, $p_{m|A} = p_{m|B}$ et $p_{u|A} = p_{u|B}$ où $p_{i|*}$ est la probabilité d'avoir la réaction $i \in \{l, m, u\}$ sachant que l'on a le vaccin $* \in \{A, B\}$.

1. On désire écrire l'hypothèse H_0 comme une hypothèse explicite. Comme le tirage au sort est équitable, cela implique que $p_A = p_B = 1/2$, où p_* est la probabilité d'avoir le vaccin $* \in \{A, B\}$. On a $p_{i,*} = p_* p_{i|*} = p_{i|*}/2$. En particulier, si $p_{i|A} = p_{i|B}$, il vient

$$p_i = p_{i,A} + p_{i,B} = \frac{1}{2}(p_{i|A} + p_{i|B}) = p_{i|*} \quad \text{et} \quad p_{i,*} = \frac{1}{2}\, p_{i|*} = \frac{1}{2}\, p_i.$$

Comme $p_l + p_m + p_u = 1$, on en déduit que $\gamma = (p_l, p_m)$ varie dans un ouvert de $\mathbb{R}^q$ avec $q = 2$. L'hypothèse nulle H_0 est une hypothèse explicite qui s'écrit :

$$p = h(\gamma) = h(p_l, p_m) = \frac{1}{2}(p_l, p_m, 1 - p_l - p_m, p_l, p_m, 1 - p_l - p_m).$$

Il est facile de vérifier que le Jacobien de h est de rang $q = 2$. Les statistiques de tests

$$\zeta_n^{(1)} = n \sum_{i \in \{l,m,u\}, * \in \{A,B\}} \frac{(\hat{p}_{i,*} - p_{i,*}(\hat{\gamma}))^2}{\hat{p}_{i,*}} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i \in \{l,m,u\}, * \in \{A,B\}} \frac{(\hat{p}_{i,*} - p_{i,*}(\hat{\gamma}))^2}{p_{i,*}(\hat{\gamma})}$$

convergent (quand $n \to \infty$) sous H_0 vers un χ^2 à $6-1-q = 3$ degrés de liberté. Sous H_1 $\zeta_n^{(1)}$ et $\zeta_n^{(2)}$ divergent. Le test de région critique $W^{(j)} = \{\zeta_n^{(j)} > z_\alpha\}$ est un test convergent de niveau asymptotique α, où z_α est défini par $\mathbb{P}(\chi^2(3) \geq z_\alpha) = \alpha$. L'estimateur du maximum de vraisemblance d'une fréquence est la fréquence empirique. On obtient donc

$$\hat{p}_{l,A} = \frac{12}{348}, \ldots, \hat{p}_{u,B} = \frac{7}{348} \quad \text{et} \quad \hat{p}_l = \frac{41}{348}, \quad \hat{p}_m = \frac{291}{348}.$$

Il vient pour $n = 348$: $\zeta_n^{(1)} = 10,3$ et $\zeta_n^{(2)} = 8,8$. On lit dans les tables $z_\alpha = 7,81$ pour le seuil usuel $\alpha = 5\%$. On rejette donc H_0. Les p-valeurs de ces tests sont $p^{(1)} = 0.016$ et $p^{(2)} = 0.032$, elles confirment que l'on peut rejeter H_0 au niveau de 5%.

2. On ne suppose plus que $p_A = p_B = 1/2$. Il faut donc estimer $a = p_A = 1 - p_B$. Si de plus $p_{i|A} = p_{i|B}$, il vient

$$p_i = p_{i,A} + p_{i,B} = a p_{i|A} + (1-a) p_{i|B} = p_{i|*} \quad \text{et} \quad p_{i,A} = a p_{i|*} = 1 - p_{i,B}.$$

Le paramètre $\gamma = (a, p_l, p_m)$ varie dans un ouvert de $\mathbb{R}^q$ avec $q = 3$. L'hypothèse nulle H_0 est une hypothèse explicite qui s'écrit :

$$\begin{aligned} p = h(\gamma) &= h(a, p_l, p_m) \\ &= (a p_l, a p_m, a(1 - p_l - p_m), (1-a) p_l, (1-a) p_m, (1-a)(1 - p_l - p_m)). \end{aligned}$$

Il est facile de vérifier que le Jacobien de h est de rang $q = 3$. Les statistiques de tests

$$\zeta_n^{(1)} = n \sum_{i \in \{l,m,u\}, * \in \{A,B\}} \frac{(\hat{p}_{i,*} - p_{i,*}(\hat{\gamma}))^2}{\hat{p}_{i,*}} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i \in \{l,m,u\}, * \in \{A,B\}} \frac{(\hat{p}_{i,*} - p_{i,*}(\hat{\gamma}))^2}{p_{i,*}(\hat{\gamma})}$$

convergent (quand $n \to \infty$) sous H_0 vers un χ^2 à $6-1-q = 2$ degrés de liberté. Sous H_1 $\zeta_n^{(1)}$ et $\zeta_n^{(2)}$ divergent. Le test de région critique $W^{(j)} = \{\zeta_n^{(j)} > z_\alpha\}$ est un test convergent de niveau asymptotique α, où z_α est défini par $\mathbb{P}(\chi^2(2) \geq z_\alpha) = \alpha$. L'estimateur du

maximum de vraisemblance d'une fréquence est la fréquence empirique. On obtient donc

$$\hat{p}_{l,A} = \frac{12}{348}, \ldots, \hat{p}_{u,B} = \frac{7}{348} \quad \text{et} \quad \hat{a} = \frac{177}{348}, \quad \hat{p}_l = \frac{41}{348}, \quad \hat{p}_m = \frac{291}{348}.$$

Il vient pour $n = 348$: $\zeta_n^{(1)} = 10,3$ et $\zeta_n^{(2)} = 8,7$. On lit dans les tables $z_\alpha = 5,99$ pour le seuil usuel $\alpha = 5\%$. On rejette donc H_0. Les p-valeurs de ces tests sont $p^{(1)} = 0.006$ et $p^{(2)} = 0.013$, elles confirment que l'on peut rejeter H_0 au niveau de 5%.

▲

Exercice IX.15. Table des fréquences estimées :

Mobilité manuelle \ Vue	Gauche	Deux yeux	Droit
Gauche	9.92	14.88	6.20
Deux mains	6.08	9.12	3.90
Droite	16.0	24.0	10.0

Il s'agit d'un test d'indépendance. La statistique du χ^2 vaut environ 1.63. Le nombre de degrés de liberté de cette statistique est égal à $(3-1) \times (4-1) = 4$. La valeur critique pour $\alpha = 0.25$ vaut 5.39. Donc on ne peut pas dire qu'il y a une relation entre la prédominance visuelle et l'habilité des mains, même en prenant un risque élevé (ici 25%).

▲

Exercice IX.16.

1. La vraisemblance du modèle de Poisson est :

$$p(x_1; \lambda) = \mathrm{e}^{-\lambda} \frac{\lambda^{x_1}}{x_1!} = \mathrm{e}^{-\lambda} \frac{\mathrm{e}^{\log(\lambda) x_1}}{x_1!}$$

Il s'agit d'un modèle exponentiel. La densité de l'échantillon de taille n est :

$$p(x_1, \ldots, x_n; \lambda) = \mathrm{e}^{-n\lambda} \frac{\mathrm{e}^{\log(\lambda) \sum_{i=1}^n x_i}}{x_1! \ldots x_n!}.$$

L'estimateur du maximum de vraisemblance est $\hat{\lambda}_n = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i$.

2. On pose

$$\xi_n^{(1)} = n \sum_{j=0}^{j=6} \frac{\left(\hat{p}_j - p_j(\hat{\lambda}_n)\right)^2}{\hat{p}_j} \quad \text{et} \quad \xi_n^{(2)} = n \sum_{j=0}^{j=6} \frac{\left(\hat{p}_j - p_j(\hat{\lambda}_n)\right)^2}{p_j(\hat{\lambda}_n)}$$

où $(\hat{p}_j)$ sont les fréquences empiriques observées et $(p_j(\hat{\lambda}_n) = e^{-\hat{\lambda}_n} \frac{\hat{\lambda}_n^j}{j!})$ sont les fréquences théoriques de la loi de Poisson de paramètre $\hat{\lambda}_n$. On sait que les statistiques ci-dessus convergent en loi vers un χ^2 à $7 - 1 - 1 = 5$ degrés de liberté (on a retranché un degré de liberté supplémentaire du fait de l'estimation de λ). Les tests définis par les régions critiques :

$$W_i = \{\xi_n^{(i)} \geq z_\alpha(5)\}, \quad i \in \{1, 2\}$$

sont convergent de niveau asymptotique α, avec $z_\alpha(5)$ définit par :

$$\mathbb{P}(\chi^2(5) \geq z_\alpha(5)) = \alpha.$$

3. Les fréquences empiriques observées, $\hat{p}_j$, et les fréquences théoriques, $p_j(\hat{\lambda}_n)$, sont données dans le tableau suivant :

j	0	1	2	3	4	5	6
$\hat{p}_j$	0,629	0,247	0,101	0,0112	0	0,0112	0
$p_j(\hat{\lambda}_n)$	0,583	0,315	0,0848	0,0152	$2,06\,10^{-3}$	$2,22\,10^{-4}$	$2,16\,10^{-5}$

TAB. IX.8 – Fréquences estimées et théoriques

Application numérique. On obtient $\hat{\lambda}_n = 0.539$, $\xi_n^{(1)} = +\infty$ (certaines fréquences empiriques sont nulles) et $\xi_n^{(2)} = 50.9$. Au seuil $\alpha = 5\%$, on a $z_\alpha(5) = 11.07 < \xi_n^{(i)} \quad i \in \{1, 2\}$. Donc on rejette l'hypothèse de distribution selon une loi de Poisson avec les deux tests. La p-valeur associée à $\xi_n^{(2)}$ est $p \simeq 10^{-9}$, cela confirme que l'on rejette H_0.

4. On regroupe les résultats des classes peu représentées en une seule classe d'effectif $9+1+0+1+0 = 11$ et correspondant au nombre de jours où l'on a capturé au moins 2 merles. En reprenant les statistiques du χ^2 empirique, on trouve :

$$\xi'^{(1)}_n = n \sum_{j=0}^{j=2} \frac{\left(\hat{p}_j - p_j(\hat{\lambda}_n)\right)^2}{\hat{p}_j} \quad \text{et} \quad \xi'^{(2)}_n = n \sum_{j=0}^{j=2} \frac{\left(\hat{p}_j - p_j(\hat{\lambda}_n)\right)^2}{p_j(\hat{\lambda}_n)},$$

avec $p_2(\hat{\lambda}_n) = \sum_{k\geq 2} e^{-\hat{\lambda}_n} \frac{\hat{\lambda}_n^k}{k!} = 1 - e^{-\hat{\lambda}_n}(1 + \hat{\lambda}_n)$. Ces statistiques convergent en loi vers un χ^2 à $3 - 1 - 1 = 1$ degré de liberté. Les tests définis par les régions critiques :

$$W_i = \{\xi'^{(i)}_n \geq z_\alpha(1)\}, \quad i \in \{1, 2\}$$

sont convergent de niveau asymptotique α, avec $z_\alpha(1)$ définit par :

$$\mathbb{P}(\chi^2(1) \geq z_\alpha(1)) = \alpha.$$

On trouve numériquement $\xi'^{(1)}_n = 2.26$ et $\xi'^{(2)}_n = 2.00$. Au seuil $\alpha = 5\%$, on a $z_\alpha(1) = 3.84 > \xi'^{(i)}_n, \quad i \in \{1, 2\}$. Donc on accepte l'hypothèse de distribution selon une loi de Poisson ! Les p-valeurs de ces deux tests sont $p^{(1)} = 0.13$ et $p^{(2)} = 0.16$, elles confirment qu'il n'est pas raisonnable de rejeter H_0.

Ce paradoxe est dû au fait que l'approximation asymptotique dans le test du χ^2 est mauvaise s'il existe des indices i tels que $np_i \leq 5$. Il faut alors regrouper les cases de sorte que cette dernière condition soit satisfaite.

▲

Exercice IX.17.

On doit d'abord estimer le paramètre de la loi de Poisson. L'estimateur du maximum de vraisemblance est :

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} X_i \, .$$

Application numérique : $\hat{\lambda} = 0.61$.

On veut donc tester si les données suivent la loi de Poisson de paramètre $\hat{\lambda}$: sous l'hypothèse poissonnienne, la statistique

$$\xi_n^{(2)} = n \sum_{j=0}^{4} \frac{\left(\hat{p}_j - p_j(\hat{\lambda})\right)^2}{p_j(\hat{\lambda})} ,$$

où $(\hat{p}_j)$ sont les fréquences empiriques et $(p_j(\hat{\lambda}))$ les fréquences théoriques, converge en loi vers un χ^2 à $q = 5 - 1 - 1 = 3$ degrés de liberté. On a retiré un degré de liberté pour l'estimation de λ (paramètre de dimension 1). Alors, le test défini par la région critique :

$$W = \{\xi_n^{(2)} \geq z_\alpha(3)\} ,$$

est asymptotiquement convergent de niveau α, avec $z_\alpha(3)$ défini par :

$$\mathbb{P}(\chi^2(3) \geq z_\alpha(3)) = \alpha .$$

Application numérique : $\alpha = 0.05$. On lit dans la table statistique des quantiles du χ^2 : $z_{0.05}(3) = 7.81$. On calcule la statistique du test : $\xi_n^{(2)} = 0.60$. La p-valeur de ce test est $p = 0.90$. Donc, on accepte l'hypothèse nulle : les données suivent une loi de Poisson.

▲

Exercice IX.18.

1. La statistique du test du χ^2 est :

$$\xi_n^{(2)} = n \sum_{j=0}^{12} \frac{\left(\hat{f}_j - f_j\right)^2}{f_j} .$$

   Sous l'hypothèse de dés non biaisés, la statistique $(\xi_n^{(2)}, n \geq 1)$ converge en loi vers un χ^2 à $13 - 1 = 12$ degrés de liberté. Le test défini par la région critique :

$$W = \{\xi_n^{(2)} \geq z_\alpha(12)\},$$

   est convergent de niveau asymptotiquement α, avec $z_\alpha(12)$ défini par :

$$\mathbb{P}(\chi^2(12) \geq z_\alpha(12)) = \alpha.$$

   Application numérique : $\alpha = 0.05$. On lit dans la table statistique des quantile de la loi du χ^2 : $z_{0.05}(12) = 21.03$. On calcule la statistique du test : $\xi_n^{(2)} = 41.31$. La p-valeur de ce test est $p = 4/100\,000$. Donc, on rejette l'hypothèse nulle. Les dés sont donc biaisés. (Il faudrait en fait regrouper les cases 10, 11 et 12 afin d'avoir au moins 5 observations (réelles ou théoriques) dans toutes les cases. Dans ce cas, on obtient une p-valeur de l'ordre de $p = 1/10\,000$. On rejette aussi l'hypothèse nulle.)

2. L'estimateur du maximum de vraisemblance de r, la probabilité d'obtenir un 5 ou un 6 lors d'un lancer de 12 dés, est donné par la fréquence empirique :

$$\hat{r} = \frac{\sum_{i=0}^{12} i N_i}{12 \sum_{i=0}^{12} N_i}.$$

Les fréquences $f_j(r)$ sont données sous H_0 par $f_j(r) = C^j - 12r^j(1-r)^{12-j}$. La statistique du test du χ^2 est :

$$\xi_n^{(2)} = n \sum_{j=0}^{12} \frac{\left(\hat{f}_j - f_j(\hat{r})\right)^2}{f_j(\hat{r})} .$$

Sous l'hypothèse de dés ayant un même biais, la statistique $(\xi_n^{(2)}, n \geq 1)$ converge en loi vers un χ^2 à $13 - 1 - 1 = 11$ degrés de liberté (on a retiré un degré de liberté supplémentaire pour l'estimation de r). Le test défini par la région critique :

$$W = \{\xi_n^{(2)} \geq z_\alpha(11)\},$$

est convergent de niveau asymptotiquement α, avec $z_\alpha(11)$ défini par :

$$\mathbb{P}(\chi^2(11) \geq z_\alpha(11)) = \alpha.$$

Application numérique : On trouve $\hat{r} = 0.338$. Rappelons $\alpha = 0.05$. On lit dans la table statistique des quantile de la loi du χ^2 : $z_{0.05}(11) = 19.68$. On calcule la statistique du test : $\xi_n^{(2)} = 13.16$. La p-valeur de ce test est $p = 0.28$. On accepte donc le fait que les 12 dés ont même biais. (Il faudrait en fait regrouper les cases 10, 11 et 12 afin d'avoir au moins 5 observations (réelles ou théoriques) dans toutes les cases. Dans ce cas, on obtient une p-valeur de l'ordre de $p = 0.52$. On accepte aussi l'hypothèse nulle.)

▲

Chapitre X

Intervalles et régions de confiance

X.1 Énoncés

Exercice X.1.
Soit $X_1, \ldots, X_n$ un échantillon de taille n de variables aléatoires indépendantes de loi uniforme sur $[0, \theta]$ où θ est un paramètre strictement positif. Soit $\theta_0 > 0$. On veut tester $H_0 : \theta \leq \theta_0$ contre $H_1 : \theta > \theta_0$ au seuil α.

1. Trouver une zone de rejet W associée à une constante z_α vérifiant $0 < z_\alpha \leq \theta_0$.
2. Calculer la puissance du test $\pi(\theta)$.
3. On prend $\theta_0 = \frac{1}{2}$. Calculer z_α en fonction de n pour que le test soit de niveau 5%.
4. Avec $\theta_0 = \frac{1}{2}$ et $\alpha = 5\%$, calculer n pour que la puissance du test soit d'au moins 98% en $\theta = \frac{3}{4}$. Que vaut la puissance du test en $\theta = \frac{3}{4}$ si $n = 2$?
5. On examine $H_0' : \theta = \theta_0$ contre $H_1' : \theta > \theta_0$. Que proposez-vous ?
6. Soit $\alpha' > 0$ donné. Calculer $k_n \geq 1$ tel que $\mathbb{P}_\theta(\theta < k_n \max_{1 \leq i \leq n} X_i)$ soit égale à $1 - \alpha'$. En déduire un intervalle de confiance pour θ au niveau $1 - \alpha'$.
7. Montrer que la suite $(n(\theta - \max_{1 \leq i \leq n} X_i), n \in \mathbb{N}^*)$ converge en loi vers une loi exponentielle. En déduire un intervalle de confiance pour θ de niveau asymptotique $1 - \alpha'$. Le comparer avec l'intervalle de confiance de la question précédente.

△

Exercice X.2.
Soit X une v.a. de densité :

$$f(x) = \frac{1}{2} \mathrm{e}^{-|x-\theta|}, \quad \forall x \in \mathbb{R},$$

où θ est un paramètre réel inconnu.

1. Calculer $\mathbb{E}_\theta[X]$ et $\mathrm{Var}_\theta(X)$. En déduire un estimateur T_n de θ.
2. Construire un intervalle de confiance de niveau asymptotique 95% pour θ dans le cas où $n = 200$.

△

X.2 Corrections

Exercice X.1.

1. Soit $\theta' > \theta$. Le rapport de vraisemblance est :

$$\frac{p_n(x_1,\dots,x_n;\theta')}{p_n(x_1,\dots,x_n;\theta)} = \left(\frac{\theta}{\theta'}\right)^n \frac{\mathbf{1}_{\{\max_{1\le i\le n} x_i \le \theta',\, \inf_{1\le i\le n} x_i \ge 0\}}}{\mathbf{1}_{\{\max_{1\le i\le n} x_i \le \theta,\, \inf_{1\le i\le n} x_i \ge 0\}}} .$$

Si $0 \le \max_{1\le i\le n} x_i \le \theta$, alors on a

$$\frac{p_n(x_1,\dots,x_n;\theta')}{p_n(x_1,\dots,x_n;\theta)} = \left(\frac{\theta}{\theta'}\right)^n ,$$

et si $\max_{1\le i\le n} x_i \ge \theta$, on pose :

$$\frac{p_n(x_1,\dots,x_n;\theta')}{p_n(x_1,\dots,x_n;\theta)} = +\infty .$$

Le rapport de vraisemblance est donc une fonction croissante de la statistique $\max\limits_{1\le i\le n} x_i$. Donc il existe un test U.P.P donné par la région critique :

$$W_\alpha = \{(x_1,\dots,x_n);\quad \max_{1\le i\le n} x_i > z_\alpha\},$$

où z_α est déterminé par $\alpha = \sup_{\theta\le\theta_0} \mathbb{P}_\theta(W_\alpha)$. Il reste encore à déterminer z_α. Calculons la loi de $\max_{1\le i\le n} X_i$:

$$\mathbb{P}_\theta(\max_{1\le i\le n} X_i \le x) = \mathbb{P}_\theta(X_1 \le x)^n = \begin{cases} 0 & \text{si} \quad x < 0, \\ \left(\frac{x}{\theta}\right)^n & \text{si} \quad 0 \le x \le \theta, \\ 1 & \text{si} \quad x > \theta. \end{cases}$$

On en déduit que : $\alpha = \sup\limits_{\theta\le\theta_0} \left(1 - \left(\frac{z_\alpha}{\theta}\right)^n\right)$. Le maximum est atteint pour la plus grande valeur de θ. Ainsi on a :

$$\alpha = \left(1 - \left(\frac{z_\alpha}{\theta_0}\right)^n\right), \quad \text{soit} \quad z_\alpha = \theta_0\,(1-\alpha)^{1/n}.$$

En conclusion, on obtient la région critique suivante :

$$W_\alpha = \{(x_1,\dots,x_n);\quad \max_{1\le i\le n} x_i > \theta_0(1-\alpha)^{1/n}\}.$$

2. Calculons la puissance du test pour $\theta \in H_1$. Comme $0 < z_\alpha \le \theta$, on a

$$\begin{aligned} \pi(\theta) &= \mathbb{P}_\theta(W_\alpha), \\ &= \mathbb{P}_\theta(\max_{1\le i\le n} X_i > z_\alpha), \\ &= 1 - \mathbb{P}_\theta(\max_{1\le i\le n} X_i \le z_\alpha), \\ &= 1 - \left(\frac{z_\alpha}{\theta}\right)^n. \end{aligned}$$

3. Pour $\theta_0 = \frac{1}{2}$ et $\alpha = 5\%$, on obtient $z_\alpha = \frac{1}{2}(0,95)^{1/n}$.
4. Pour $\theta_0 = \frac{1}{2}$ et $\alpha = 5\%$, il faut choisir n tel que :

$$1 - \left(\frac{(0.95)^{1/n} \frac{1}{2}}{3/4} \right)^n \geq 0.98 \quad \text{soit} \quad n \geq 9.5.$$

On trouve donc $n = 10$.
Si $n = 2$, on obtient $\pi(\frac{3}{4}) = 0.578$. On remarquera que plus n grand, plus $\pi(\theta)$ est proche de 1. L'erreur de deuxième espèce pour θ est proche de 0 quand n est grand. En effet le test associé à la région critique W est convergent.

5. C'est un cas particulier du test précédent. La région critique

$$W_\alpha = \{(x_1, \ldots, x_n); \quad \max_{1 \leq i \leq n} x_i > \theta_0 (1-\alpha)^{1/n}\}$$

définit un test UPP de niveau α.

6. Soit $\alpha' > 0$. D'après ce que nous avons vu précédemment, nous avons :

$$1 - \alpha' = \mathbb{P}_\theta \left(\max_{1 \leq i \leq n} X_i > \frac{\theta}{k_n} \right) = 1 - \mathbb{P}_\theta \left(\max_{1 \leq i \leq n} X_i \leq \frac{\theta}{k_n} \right).$$

On suppose $k_n \geq 1$. Ceci entraîne que :

$$1 - \alpha' = 1 - \left(\frac{1}{k_n} \right)^n \quad \text{soit} \quad k_n = \left(\frac{1}{\alpha'} \right)^{1/n}.$$

Comme $\theta \geq \max_{1 \leq i \leq n} X_i$, on obtient l'intervalle de confiance de niveau $1 - \alpha'$ pour θ :

$$IC_{\alpha'}(\theta) = \left] \max_{1 \leq i \leq n} X_i ; \left(\frac{1}{\alpha'} \right)^{1/n} \max_{1 \leq i \leq n} X_i \right[.$$

7. La fonction de répartition F_n de $n(\theta - \max_{1 \leq i \leq n} X_i)$ est pour $x \geq 0$,

$$\begin{aligned} F_n(x) &= 1 - \mathbb{P}_\theta(n(\theta - \max_{1 \leq i \leq n} X_i) > x) \\ &= 1 - \mathbb{P}_\theta(\max_{1 \leq i \leq n} X_i < \theta - \frac{x}{n}) \\ &= 1 - (1 - \frac{x}{n\theta})^n. \end{aligned}$$

En particulier la suite de fonctions de répartition $(F_n, n \in \mathbb{N}^*)$ converge vers la fonction F définie par

$$F(x) = 1 - \mathrm{e}^{-x/\theta} \quad \text{si } x \geq 0, \text{ et} \quad F(x) = 1 \quad \text{si } x \leq 0.$$

Ainsi la suite $(n(\theta - \max_{1 \leq i \leq n} X_i), n \in \mathbb{N}^*)$ converge en loi vers une loi exponentielle de paramètre $1/\theta$. En particulier la suite $(n(1 - \frac{1}{\theta} \max_{1 \leq i \leq n} X_i), n \in \mathbb{N}^*)$ converge en loi vers une variable aléatoire Y de loi exponentielle de paramètre 1. Comme les variables aléatoires exponentielles sont continues, on en déduit que sous $\mathbb{P}_\theta$, on a

$$\mathbb{P}_\theta(0 \leq n(1 - \frac{1}{\theta} \max_{1 \leq i \leq n} X_i) \leq a) \xrightarrow[n \to \infty]{} \mathbb{P}(0 \leq Y \leq a) = 1 - \mathrm{e}^{-a}.$$

$1-\alpha' = 1-\mathrm{e}^{-a}$ implique $a = -\log(\alpha')$. On en déduit donc l'intervalle de confiance de niveau asymptotique $1-\alpha'$ est $IC_{\alpha'}^{\text{asymp}}(\theta) = \left]\max_{1\le i\le n} X_i; +\infty\right[$ si $n \le -\log(\alpha')$ et pour $n > -\log(\alpha')$:

$$IC_{\alpha'}^{\text{asymp}}(\theta) = \left]\max_{1\le i\le n} X_i; \frac{1}{1+\log(\alpha')/n}\max_{1\le i\le n} X_i\right[.$$

Il est facile de vérifier que $\left(\frac{1}{\alpha'}\right)^{1/n} \le \frac{1}{1+\log(\alpha')/n}$ pour tout α', n tels que $n > -\log(\alpha')$. En particulier l'intervalle de confiance asymptotique $IC_{\alpha'}^{\text{asymp}}$ est un intervalle de confiance par excès.

▲

Exercice X.2.

1. Calcul de l'espérance :

$$\begin{aligned}\mathbb{E}_\theta[X] &= \frac{1}{2}\int_{+\infty}^{-\infty} x e^{-|x-\theta|}dx \\ &= \frac{1}{2}\int_{+\infty}^{-\infty} (u+\theta) e^{-|u|}dx \\ &= \frac{\theta}{2}\int_{+\infty}^{-\infty} e^{-|u|}dx \\ &= \theta\,.\end{aligned}$$

De même, on calcule le moment d'ordre 2 : $\mathbb{E}_\theta[X^2] = 2+\theta^2$; puis on en déduit la variance de X : $\mathrm{Var}_\theta(X) = 2$. D'après la méthode des moments, on en déduit que la moyenne empirique $\bar{X}_n$ est un estimateur de θ. De plus, c'est un estimateur convergent par la loi forte des grands nombres.

2. D'après le Théorème Central Limite, on obtient l'intervalle de confiance asymptotique suivant :

$$IC_\alpha(\theta) = \left]\bar{X}_n \pm u_\alpha\sqrt{\frac{2}{n}}\right[,$$

où u_α est le quantile d'ordre $1-\alpha/2$ de la loi gaussienne $\mathcal{N}(0,1)$: $\mathbb{P}(|Y| > u_\alpha) = \alpha$, où Y est de loi $\mathcal{N}(0,1)$. Pour un niveau de confiance de 95%, on obtient $u_\alpha = 1,96$ et avec $n = 200$, on a

$$IC_\alpha(\theta) = \left]\bar{X}_n \pm 0.196\right[.$$

▲

Chapitre XI

Contrôles à mi-cours

XI.1 1999-2000

XI.1.1 Exercices

Exercice XI.1.
Soit $(T_n, n \geq n_0)$ une suite de variables aléatoires de loi géométrique de paramètre $p_n = \dfrac{\theta}{n}$ avec $n_0 > \theta > 0$. Montrer que la suite $\left(\dfrac{T_n}{n}, n \geq n_0\right)$ converge en loi et déterminer sa limite.

△

Exercice XI.2.
Déterminant d'une matrice à coefficients gaussiens.

1. Soit V et W deux variables aléatoires réelles ou vectorielles continues indépendantes. Montrer que si φ est une fonction bornée, alors $\mathbb{E}[\varphi(V,W) \mid W] = h(W)$, où la fonction h est définie par $h(w) = \mathbb{E}[\varphi(V,w)]$.
2. Soit (X_1, X_2, X_3, X_4) des variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$. On considère la matrice aléatoire $A = \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix}$ et on note $Y = \det A$. Calculer $\mathbb{E}\left[\mathrm{e}^{iuY} \,|X_1, X_2\right]$, puis en déduire la fonction caractéristique de Y.

△

XI.1.2 Le collectionneur (I)

Exercice XI.3.
Votre petit frère collectionne les images des joueurs de la coupe du monde que l'on trouve dans les tablettes de chocolat. On suppose qu'il existe n images différentes et qu'elles sont équitablement réparties, à raison de une par tablette. On note $X_i \in \{1, \cdots, n\}$ le numéro de l'image contenue dans la i-ème tablette. On note N_k le nombre de tablettes achetées pour obtenir k images différentes : $N_k = \inf\{j \geq 1; \text{Card } \{X_i, i \leq j\} = k\}$. Enfin, $T_k = N_k - N_{k-1}$, avec la convention $T_1 = 1$, représente le nombre de tablettes achetées pour obtenir une nouvelle image alors que l'on en possède déjà $k-1$.

1. Quelle loi proposez-vous pour la suite de variables aléatoires $(X_i, i \in \mathbb{N}^*)$?
2. Soit T une variable aléatoire géométrique de paramètre $p \in]0,1[$. Montrer que $\mathbb{E}[T] = p^{-1}$ et $\mathrm{Var}(T) = (1-p)p^{-2}$.
3. Calculer $\mathbb{P}(T_2 = l)$. En déduire que T_2 suit une loi géométrique dont on précisera le paramètre.
4. Montrer que

$$\begin{aligned}&\mathbb{P}(T_2 = l_2, T_3 = l_3)\\ &= \sum_{j_1, j_2, j_3 \text{ distincts}} \mathbb{P}(X_1 = j_1, \cdots, X_{l_2} = j_1, X_{l_2+1} = j_2, \cdots, X_{l_2+l_3} \in \{j_1, j_2\}, X_{l_2+l_3+1} = j_3).\end{aligned}$$

5. En déduire que T_3 suit une loi géométrique dont on précisera le paramètre.
6. Vérifier que T_2 et T_3 sont indépendants.
7. Décrire T_k comme premier instant de succès et en déduire sa loi. *On admet dorénavant que les variables aléatoires $T_1, T_2, \cdots, T_n$ sont indépendantes.*
8. Calculer $\mathbb{E}[N_n]$ et vérifier que $\mathbb{E}[N_n] = n\,[\log(n) + O(1)]$ où $O(1)$ désigne une fonction $g(n)$ telle que $\sup_{n\geq 1} |g(n)| \leq M < \infty$.
9. Calculer $\mathrm{Var}(N_n)$ et en donner un équivalent quand $n \to \infty$.
10. Vérifier que $\mathbf{1}_{\{x^2 > \varepsilon^2\}} \leq x^2 \varepsilon^{-2}$ pour tout $x \in \mathbb{R}$ et $\varepsilon > 0$. Majorer $\mathbb{P}\left(\left|\dfrac{N_n}{\mathbb{E}[N_n]} - 1\right| > \varepsilon\right)$.
11. Montrer que la suite $\left(\dfrac{N_n}{n \log n}, n \in \mathbb{N}^*\right)$ converge en probabilité vers 1.

△

XI.2 2000-2001

XI.2.1 Exercice

Exercice XI.4.
Soit X_1, X_2 des variables aléatoires indépendantes de loi de Poisson de paramètres respectifs $\theta_1 > 0$ et $\theta_2 > 0$.

1. Calculer la loi de $X_1 + X_2$.
2. Calculer la loi de X_1 sachant $X_1 + X_2$. Reconnaître cette loi.
3. Calculer $\mathbb{E}[X_1 | X_1 + X_2]$.

△

XI.2.2 Le collectionneur (II)

Exercice XI.5.
Soit $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes de loi exponentielle de paramètre $\lambda > 0$. La variable aléatoire X_i représente le temps de panne de la machine i. On

suppose qu'une fois en panne les machines ne sont pas réparées. On note $X_{(1)} \leq \cdots \leq X_{(n)}$ le réordonnement croissant de $X_1, \ldots, X_n$. Ainsi $X_{(i)}$ représente le temps de la i-ème panne quand on considère l'ensemble des n machines. On pose

$$Y_1 = X_{(1)} = \min_{1 \leq i \leq n} X_i,$$

le temps de la première panne, $Y_2 = X_{(2)} - X_{(1)}$ le temps entre la première et la deuxième panne, et plus généralement pour $k \in \{2, \ldots, n\}$,

$$Y_k = X_{(k)} - X_{(k-1)}.$$

Le but de ce problème est dans un premier temps d'étudier le comportement de l'instant où la dernière machine tombe en panne, $X_{(n)} = \sum_{i=1}^n Y_i$, quand $n \to \infty$. Dans un deuxième temps nous étudierons la loi du vecteur $(Y_1, \ldots, Y_n)$. Enfin nous donnerons une application de ces deux résultats dans une troisième partie.

I Comportement asymptotique de $X_{(n)} = \sum_{i=1}^n Y_i$.

1. Calculer la fonction de répartition de $X_{(n)} = \max_{1 \leq i \leq n} X_i$.
2. Montrer que la suite $(X_{(n)} - \lambda^{-1} \log n, n \in \mathbb{N}^*)$ converge en loi vers une variable aléatoire Z dont on déterminera la fonction de répartition.
3. Pour $\lambda = 1$, en déduire la densité f de la loi de Z. Déterminer, à la première décimale près, a et b tels que $\int_{-\infty}^a f(z)dz = 2,5\%$ et $\int_b^{+\infty} f(z)dz = 2,5\%$.

II Loi du vecteur $(Y_1, \ldots, Y_n)$.

1. Soit $i \neq j$. Calculer $\mathbb{P}(X_i = X_j)$.
2. En déduire que $\mathbb{P}(\exists i \neq j; X_i = X_j) = 0$. Remarquer que presque sûrement le réordonnement croissant est unique, c'est-à-dire $X_{(1)} < \cdots < X_{(n)}$. Ainsi p.s. aucune machine ne tombe en panne au même instant.
3. On suppose dans **cette question et la suivante seulement** que $n = 2$. Soit g_1 et g_2 des fonctions bornées mesurables. En distinguant $\{X_1 < X_2\}$ et $\{X_2 < X_1\}$, montrer que
$$\mathbb{E}[g_1(Y_1)g_2(Y_2)] = \int g_1(y_1)g_2(y_2)\, 2\lambda^2\, \mathrm{e}^{-2\lambda y_1 - \lambda y_2}\, \mathbf{1}_{\{y_1>0, y_2>0\}}\, dy_1 dy_2.$$
En déduire une expression de $\mathbb{E}[g_1(Y_1)]$ puis la loi de Y_1.
4. Déduire la loi de Y_2 et vérifier que Y_1 et Y_2 sont indépendants. Donner la loi du vecteur (Y_1, Y_2).
5. En décomposant suivant les évènements $\{X_{\sigma(1)} < \cdots < X_{\sigma(n)}\}$, où σ parcourt l'ensemble $\mathcal{S}_n$ des permutations de $\{1, \ldots, n\}$, montrer que pour $n \geq 3$,
$$\mathbb{E}[g_1(Y_1) \cdots g_n(Y_n)] = n!\, \mathbb{E}[g_1(X_1)g_2(X_2 - X_1) \cdots g_n(X_n - X_{n-1})\mathbf{1}_{\{X_1 < \cdots < X_n\}}],$$
où les fonctions $g_1, \ldots g_n$ sont mesurables bornées. On pourra utiliser, sans le justifier, le fait que les vecteurs $(X_{\sigma(1)}, \ldots, X_{\sigma(n)})$ et $(X_1, \ldots, X_n)$ ont même loi.
6. Calculer la loi de Y_i, $i \in \{1, \ldots, n\}$. Montrer que les variables aléatoires $Y_1, \ldots, Y_n$ sont indépendantes. Donner la loi du vecteur $(Y_1, \ldots, Y_n)$.
7. En déduire la fonction caractéristique de $X_{(n)}$.

III Application (Facultatif)

Votre petit frère collectionne les images de Pokémons que l'on trouve dans les plaquettes de chocolat. On suppose qu'il existe n images différentes et qu'elles sont réparties au hasard dans les plaquettes. On note $T_{k,n}$ le nombre de plaquettes qu'il faut acheter pour avoir une nouvelle image, alors que l'on en possède déjà $k-1$. On a donc $T_{1,n}=1$. Pour avoir toutes les images, il faut donc acheter $T_{1,n}+\cdots+T_{n,n}=N_n$ plaquettes. On admet (voir le contrôle de 1999, exercice XI.3) que les variables aléatoires $T_{1,n},\ldots,T_{n,n}$ sont indépendantes et que la loi de $T_{k,n}$ est la loi géométrique de paramètre $1-\dfrac{k-1}{n}$. On admet de plus que $\mathbb{E}[N_n]\sim n\log n$ et que $\dfrac{N_n}{n\log n}$ converge en probabilité vers 1 quand $n\to+\infty$. Le but de cette partie est de déterminer à quelle vitesse a lieu cette convergence et d'en déduire un intervalle de confiance pour le nombre de plaquettes de chocolat que votre petit frère doit acheter pour avoir sa collection complète.

Soit $\psi_{k,n}$ la fonction caractéristique de $\dfrac{1}{n}T_{n-k+1,n}$, où $k\in\{1,\ldots,n\}$ et ψ_k la fonction caractéristique de la loi exponentielle de paramètre k.

1. Montrer que la suite $\left(\dfrac{1}{n}T_{n-k+1,n}, n\geq k\right)$ converge en loi vers la loi exponentielle de paramètre k.
   Une étude plus précise permet en fait de montrer que pour tout $u\in\mathbb{R}$, il existe $n(u)$ et C, tels que pour tout $n\geq n(u)$ et $k\in\{1,\ldots,n\}$, on a
   $$|\psi_{k,n}(u)-\psi_k(u)|\leq\frac{1}{kn}C.$$
2. En utilisant l'inégalité $\left|\prod_{k=1}^{n}a_k-\prod_{k=1}^{n}b_k\right|\leq\sum_{k=1}^{n}|a_k-b_k|$ valable pour des complexes a_k,b_k de modules inférieur à 1 ($|a_k|\leq 1$ et $|b_k|\leq 1$), montrer que
   $$\left|\psi_{N_n/n}(u)-\psi_{X_{(n)}}(u)\right|\leq C\frac{\log n}{n},$$
   où $X_{(n)}$ est définie au paragraphe I, avec $\lambda=1$. De manière formelle, on écrira que pour n grand,
   $$\frac{1}{n}\,\text{Géom}(1)+\frac{1}{n}\,\text{Géom}(\frac{n-1}{n})+\cdots+\frac{1}{n}\,\text{Géom}(\frac{1}{n})\overset{\text{Loi}}{\simeq}\mathcal{E}(n)+\mathcal{E}(n-1)+\cdots+\mathcal{E}(1).$$
3. En déduire que la suite $(n^{-1}(N_n-n\log n), n\in\mathbb{N}^*)$ converge en loi vers la variable aléatoire Z définie au paragraphe I.
4. Donner un intervalle de confiance, I_n, de niveau asymptotique $\alpha=95\%$ pour N_n : $\mathbb{P}(N_n\in I_n)\simeq 95\%)$ pour n grand.
5. Application numérique : $n=151$. Quel est le nombre moyen de plaquettes de chocolat que votre petit frère doit acheter pour avoir une collection de Pokémons complète. Donner un intervalle de confiance à 95% du nombre de plaquettes que votre petit frère risque d'acheter pour avoir une collection complète.
   Refaire l'application numérique pour $n=250$, qui correspond au nombre de Pokémons au Japon.

△

XI.3 2001-2002

XI.3.1 Exercice

Exercice XI.6.
Soit $(U_i, i \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes de loi uniforme sur $[0,1]$. Soit $\alpha > 0$.

1. Pour $n \in \mathbb{N}^*$, on pose $X_n = (U_1 \cdots U_n)^{\alpha/n}$. Montrer que la suite $(X_n, n \in \mathbb{N}^*)$ converge presque sûrement et donner sa limite. On pourra considérer dans un premier temps la suite $(\log(X_n), n \in \mathbb{N}^*)$.
2. Montrer que la suite $(Y_n, n \in \mathbb{N}^*)$, définie par $Y_n = [X_n \, \mathrm{e}^{\alpha}]^{\sqrt{n}}$, converge en loi et déterminer la loi limite. On calculera la densité de la loi limite s'il s'agit d'une variable aléatoire continue.

△

XI.3.2 Le paradoxe du bus

Exercice XI.7.
À l'arrêt de bus, il est indiqué que le temps moyen entre les passages de bus est d'environ 10 minutes. Or, lorsqu'un client arrive à l'instant t à l'arrêt de bus, il attend en moyenne une dizaine de minutes. On en déduit naïvement, par symétrie, que le temps moyen entre deux passages de bus est de 20 minutes. Le but de ce problème est de déterminer à l'aide d'un modèle simple qui, de la société de bus ou du client a raison.

On note T_1 le temps de passage du premier bus et T_{i+1} le temps entre le passage du $i^{\text{ème}}$ et du $i+1^{\text{ème}}$ bus. En particulier $V_n = \sum_{i=1}^{n} T_i$ est le temps de passage du $n^{\text{ème}}$ bus, avec la convention que $V_0 = \sum_{i=1}^{0} T_i = 0$. À cause des conditions aléatoires du trafic, on suppose que les variables aléatoires $(T_i, i \in \mathbb{N}^*)$ sont indépendantes et de loi exponentielle de paramètre $\lambda > 0$.

I Préliminaires

1. Quel est, d'après ce modèle, le temps moyen entre les passages des bus annoncé par la société de bus ?
2. Déterminer et reconnaître la loi de V_n pour $n \geq 1$.

On note N_t le nombre de bus qui sont passés avant l'instant t :

$$N_t = \sup\{n \geq 0; \sum_{i=1}^{n} T_i \leq t\}.$$

3. Quelle est la valeur de $\mathbb{P}(N_t = 0)$?
4. Écrire l'évènement $\{N_t = n\}$ en fonction de V_n et T_{n+1}.
5. Pour $n \geq 1$, calculer $\mathbb{P}(N_t = n)$, et vérifier que la loi de N_t est une loi de Poisson dont on précisera le paramètre.

II Les temps moyens

On note $R_t = t - \sum_{i=1}^{N_t} T_i$, le temps écoulé entre le passage du dernier bus avant t et t, avec la convention que $R_t = t$ si $N_t = 0$. Et on note $S_t = \sum_{i=1}^{N_t+1} T_i - t$ le temps d'attente du prochain bus lorsqu'on arrive à l'instant t. Soit $r, s \in [0, \infty[$.

1. Que vaut $\mathbb{P}(R_t \leq t)$? Calculer $\mathbb{P}(R_t \leq r, S_t \leq s, N_t = 0)$.
2. Vérifier que pour $n \geq 1$, on a

$$\mathbb{P}(R_t \leq r, S_t \leq s, N_t = n) = \mathrm{e}^{-\lambda t}(1 - \mathrm{e}^{-\lambda s})\frac{\lambda^n}{n!}[t^n - (t-r)_+^n],$$

   où $z_+ = z$ si $z \geq 0$ et $z_+ = 0$ si $z < 0$.
3. Calculer $\mathbb{P}(R_t \leq r, S_t \leq s)$ la fonction de répartition du couple (R_t, S_t). On distinguera les cas $r < t$ et $r \geq t$.
4. Déterminer et reconnaître la loi de S_t.
5. Montrer que R_t a même loi que $\min(T_1, t)$.
6. En déduire le temps moyen d'attente d'un bus lorsqu'on arrive à l'instant t ainsi que l'écart moyen entre le dernier bus juste avant l'instant t et le premier bus juste après l'instant t. Pouvez vous répondre à la question initiale du problème ? Quel est le paradoxe ?

III Loi du temps entre deux passages de bus

On désire maintenant étudier la loi du temps entre le dernier bus juste avant l'instant t et le premier bus juste après l'instant t : $U_t = R_t + S_t$.

1. Vérifier que les évènements $\{R_t \leq r\}$ et $\{S_t \leq s\}$ sont indépendants pour tout $(r, s) \in \mathbb{R}^2$. On rappelle que cela implique que les variables R_t et S_t sont indépendantes.
2. Vérifier que $(U_t, t > 0)$ converge en loi vers une variable aléatoire U. Déterminer et reconnaître la loi limite.
3. Calculer la loi de $U_t = R_t + S_t$.

△

XI.4 2002-2003

XI.4.1 La statistique de Mann et Whitney

Exercice XI.8.
L'objectif est d'étudier le comportement asymptotique d'une suite de variables aléatoires appelées statistiques de Mann et Whitney.

I Calculs préliminaires

Soit X, Y deux variables aléatoires réelles indépendantes de densité respective f et g. On introduit les fonctions de répartitions

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)\, du \qquad \text{et} \qquad G(y) = \mathbb{P}(Y \leq y) = \int_{-\infty}^{y} g(u)\, du.$$

On suppose que $p = \mathbb{P}(Y \leq X) \in]0,1[$.

1. Quelle est la loi de $\mathbf{1}_{\{Y \leq X\}}$? Donner $\mathrm{Var}(\mathbf{1}_{\{Y \leq X\}})$ en fonction de p.
2. Déterminer p comme une intégrale en fonction de G et f (ou en fonction de F et g). Vérifier que, si X et Y ont même loi (i.e. $f = g$), alors $p = 1/2$.
3. On pose $S = \mathbb{E}\left[\mathbf{1}_{\{Y \leq X\}} | X\right]$ et $T = \mathbb{E}\left[\mathbf{1}_{\{Y \leq X\}} | Y\right]$. Déterminer S et T. Donner $\mathbb{E}[S]$ et $\mathbb{E}[T]$.
4. On pose $\alpha = \mathrm{Var}(S)$ et $\beta = \mathrm{Var}(T)$. Calculer α (respectivement β) en fonction de p, G et f (respectivement p, F et g). On admet que $p \in]0,1[$ implique que $(\alpha, \beta) \neq (0,0)$.
5. Montrer que, si X et Y ont même loi, alors $\alpha = \beta$. Donner alors leur valeur.
6. Calculer $\mathrm{Cov}(S, \mathbf{1}_{\{Y \leq X\}})$ et $\mathrm{Cov}(T, \mathbf{1}_{\{Y \leq X\}})$.

II Étude de la projection de Hajek de la statistique de Mann et Withney

Soit $(X_i, i \geq 1)$ et $(Y_j, j \geq 1)$ deux suites indépendantes de variables aléatoires indépendantes. On suppose de plus que X_i a même loi que X pour tout $i \geq 1$, et Y_j a même loi que Y pour tout $j \geq 1$. La statistique de Mann et Whitney (1947) est la variable définie pour $m \geq 1$, $n \geq 1$, par

$$U_{m,n} = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbf{1}_{\{Y_j \leq X_i\}}.$$

On pose $U^*_{m,n} = U_{m,n} - \mathbb{E}[U_{m,n}] = \displaystyle\sum_{i=1}^{m} \sum_{j=1}^{n} \Big(\mathbf{1}_{\{Y_j \leq X_i\}} - p\Big)$. La projection de Hajek (1968) de $U^*_{m,n}$ est définie par

$$H_{m,n} = \sum_{i=1}^{m} \mathbb{E}[U^*_{m,n} | X_i] + \sum_{j=1}^{n} \mathbb{E}[U^*_{m,n} | Y_j].$$

On pose $S_i = G(X_i)$ et $T_j = 1 - F(Y_j)$.

1. Vérifier que $H_{m,n} = n \sum_{i=1}^{m} (S_i - p) + m \sum_{j=1}^{n} (T_j - p)$.
2. Calculer $\mathrm{Var}(H_{m,n})$ en fonction de α et β.
3. Déterminer la limite en loi des suites

$$\left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} (S_i - p), m \geq 1 \right) \qquad \text{et} \qquad \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{n} (T_j - p), n \geq 1 \right).$$

4. On considère la suite d'indices $((m_k, n_k), k \geq 1)$ telle que

$$m_k \xrightarrow[k\to\infty]{} \infty, \quad n_k \xrightarrow[k\to\infty]{} \infty, \quad \text{et} \quad \frac{m_k}{m_k + n_k} \xrightarrow[k\to\infty]{} \theta \in]0,1[.$$

Montrer en utilisant les fonctions caractéristiques, la convergence en loi du vecteur

$$\left((V_k, W_k) = \left(\frac{1}{\sqrt{m_k}} \sum_{i=1}^{m_k} (S_i - p), \frac{1}{\sqrt{n_k}} \sum_{j=1}^{n_k} (T_j - p) \right), k \geq 1 \right).$$

5. Montrer que $\left(H_{m_k,n_k} / \sqrt{\mathrm{Var}(H_{m_k,n_k})}, k \geq 1 \right)$ converge en loi vers une loi gaussienne dont on déterminera les paramètres. On pourra écrire $H_{m_k,n_k} = a_k V_k + b_k W_k$ avec $a_k = n_k \sqrt{m_k} / \sqrt{m_k n_k^2 \alpha + n_k m_k^2 \beta}$ et $b_k = m_k \sqrt{n_k} / \sqrt{m_k n_k^2 \alpha + n_k m_k^2 \beta}$.
6. On admet la formule suivante (voir la partie IV pour une démonstration) :

$$\mathrm{Var}(U_{m,n}) = mnp(1-p) + n(n-1)m\alpha + m(m-1)n\beta. \tag{XI.1}$$

Déterminer la limite en loi de la suite $\left(H_{m_k,n_k} / \sqrt{\mathrm{Var}(U_{m_k,n_k})}, k \geq 1 \right)$.

III Convergence de la statistique de Mann et Whitney (Facultatif)

1. Montrer que $\mathrm{Cov}(H_{m,n}, U^*_{m,n}) = mn^2 \ \mathrm{Cov}(S, \mathbf{1}_{\{Y \leq X\}}) + nm^2 \ \mathrm{Cov}(T, \mathbf{1}_{\{Y \leq X\}})$.
2. En déduire $\mathrm{Var}(H_{m,n} - U^*_{m,n})$.
3. Calculer la limite de $\mathrm{Var}(H_{m_k,n_k} - U^*_{m_k,n_k}) / \mathrm{Var}(U_{m_k,n_k})$ quand $k \to \infty$.
4. En déduire que la suite $\left(\dfrac{H_{m_k,n_k} - U^*_{m_k,n_k}}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}}, k \geq 1 \right)$ converge en probabilité vers 0.
5. Montrer que la suite

$$\left(\frac{U_{m_k,n_k} - m_k n_k p}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}}, k \geq 1 \right)$$

converge en loi. Déterminer la loi limite.

IV Calcul de la variance de la statistique de Mann et Whitney (Facultatif)

Soit X' et Y' des variables aléatoires de même loi que X et Y. On suppose de plus que les variables aléatoires X, X', Y et Y' sont indépendantes.

1. On considère la partition de $\Delta = \{(i,i',j,j') \in (\mathbb{N}^*)^4; i \leq m, i' \leq m, j \leq n, j' \leq n\}$ en quatre sous-ensembles :

$$\begin{aligned}
\Delta_1 &= \{(i,i,j,j) \in \Delta\} \\
\Delta_2 &= \{(i,i',j,j) \in \Delta; i \neq i'\} \\
\Delta_3 &= \{(i,i,j,j') \in \Delta; j \neq j'\} \\
\Delta_4 &= \{(i,i',j,j') \in \Delta; i \neq i', j \neq j'\}.
\end{aligned}$$

Calculer le cardinal des quatre sous-ensembles.

2. Vérifier que $\mathrm{Cov}(\mathbf{1}_{\{Y\leq X\}}, \mathbf{1}_{\{Y'\leq X\}}) = \alpha$ et $\mathrm{Cov}(\mathbf{1}_{\{Y\leq X\}}, \mathbf{1}_{\{Y\leq X'\}}) = \beta$.
3. Calculer la variance de $U_{m,n}$ et vérifier ainsi la formule (XI.1).
4. Donner $\mathrm{Var}(U_{m,n})$ dans le cas où les variables aléatoires $(X_i, i \geq 1)$ et $(Y_j, j \geq 1)$ ont toutes même loi.

△

XI.5 2003-2004

XI.5.1 Le processus de Galton Watson

Exercice XI.9.
En 1873, Galton publie un problème concernant le calcul de la probabilité d'extinction des noms de familles. N'obtenant pas de réponse satisfaisante, il contacte Watson qui fournit une réponse partielle. Ce n'est qu'à partir de 1930 que ce problème attire à nouveau l'attention et obtient alors une réponse détaillée[1].

Le but du problème qui suit est, à partir d'un modèle élémentaire d'évolution de population, appelé modèle de Galton-Watson, de déterminer cette probabilité d'extinction.

On considère un individu masculin à l'instant 0, et on note Z_n le nombre de descendants masculin de cet individu à la n-ième génération ($Z_0 = 1$ par convention). On suppose que les nombres de garçons de chaque individu sont indépendants et de même loi qu'une variable aléatoire, ξ, à valeurs entières. Plus précisément, soit $(\xi_{i,n}, i \geq 1, n \geq 0)$ une suite doublement indicée de variables aléatoires indépendantes de même loi que ξ. Le nombre d'individus de la $n+1$-ième génération est la somme des garçons des individus de la n-ième génération : pour $n \geq 0$,

$$Z_{n+1} = \sum_{i=1}^{Z_n} \xi_{i,n},$$

avec la convention que $Z_{n+1} = 0$ si $Z_n = 0$. On note

$$\eta = \mathbb{P}(\text{il existe } n \geq 0 \text{ tel que } Z_n = 0$$

la probabilité d'extinction de la population. Pour $k \in \mathbb{N}$, on note $p_k = \mathbb{P}(\xi = k)$, et l'on suppose que $\boxed{p_0 > 0}$ (sinon presque sûrement la population ne s'éteint pas).

I Calcul de la probabilité d'extinction

1. Montrer que η est la limite croissante de la suite $(\mathbb{P}(Z_n = 0), n \geq 0)$.

On suppose que ξ est intégrable, et on pose $m = \mathbb{E}[\xi]$.

2. Calculer $\mathbb{E}[Z_{n+1}|Z_n]$. En déduire que $\mathbb{E}[Z_n] = m^n$.
3. Montrer que si $m < 1$, alors $\eta = 1$, i.e. la population s'éteint presque sûrement.

On note ϕ la fonction génératrice de ξ, et ϕ_n la fonction génératrice de Z_n (et $\phi_0(z) = z$ pour $z \in [0, 1]$).

[1] Voir l'article de D. Kendall. Branching processes since 1873, *J. London Math. Soc.* (1966), **41** pp. 385-406.

4. Calculer $\mathbb{E}[z^{Z_{n+1}}|Z_n]$ pour $z \in [-1,1]$. En déduire que $\phi_{n+1} = \phi_n \circ \phi$, puis que $\phi_{n+1} = \phi \circ \phi_n$.
5. Montrer que $\mathbb{P}(Z_{n+1} = 0) = \phi(\mathbb{P}(Z_n = 0))$. En déduire que η est solution de l'équation

$$\phi(x) = x. \tag{XI.2}$$

6. Calculer $\phi'(1)$. Vérifier que si $m \geq 1$, alors ϕ est strictement convexe sur $[0,1]$. Tracer le graphe $z \mapsto \phi(z)$ pour $z \in [0,1]$.
7. En déduire que si $m = 1$, alors $\eta = 1$.

On suppose dorénavant que $m > 1$.

8. Montrer que (XI.2) possède une unique solution $x_0 \in]0,1[$.
9. Montrer que $\mathbb{P}(Z_n = 0) \leq x_0$ pour tout $n \geq 0$. En déduire que $\eta = x_0$.

II Comportement asymptotique sur un exemple

Les données concernant les U.S.A. en 1920 pour la population masculine (cf la référence (1) en bas de page 171) sont telles que l'on peut modéliser la loi de ξ sous la forme

$$p_0 = \alpha, \quad \text{et pour } k \geq 1, \quad p_k = (1-\alpha)(1-\beta)\beta^{k-1}, \tag{XI.3}$$

avec $0 < \alpha < \beta < 1$. On suppose dorénavant que la loi de ξ est donnée par (XI.3).

1. Calculer $m = \mathbb{E}[\xi]$, vérifier que $m > 1$ et calculer η, l'unique solution de (XI.2) dans $]0,1[$, où ϕ est la fonction génératrice de ξ. Application numérique (cf la note (1) en bas de page 171) : $\alpha = 0.4813$ et $\beta = 0.5586$.
2. Vérifier que $\dfrac{\phi(z)-1}{\phi(z)-\eta} = m\dfrac{z-1}{z-\eta}$. En déduire $\phi_n(z)$, où $\phi_1 = \phi$ et pour $n \geq 2$, $\phi_n = \phi \circ \phi_{n-1}$.
3. Calculer la fonction caractéristique de XY, où X et Y sont indépendants, X est une variable aléatoire de Bernoulli de paramètre $p \in [0,1]$, et Y une variable aléatoire exponentielle de paramètre $\lambda > 0$.
4. Montrer que la suite $(m^{-n}Z_n, n \geq 1)$, où Z_n est une variable aléatoire de fonction génératrice ϕ_n, converge en loi vers une variable aléatoire Z dont on reconnaîtra la loi.

△

XI.6 2004-2005

XI.6.1 Exercice

Exercice XI.10.
Le but de cet exercice est la démonstration du théorème de Cochran. Soit $n \geq 2$, et $X_1, \ldots, X_n$ des variables aléatoires indépendantes de même loi gaussienne $\mathcal{N}(0,1)$. Soit $e = \{e_1, \ldots, e_n\}$ la base canonique de $\mathbb{R}^n$ et $X = \sum_{i=1}^n X_i e_i$ le vecteur aléatoire de $\mathbb{R}^n$.

1. Soit $f = \{f_1, \dots, f_n\}$ une base orthonormée de $\mathbb{R}^n$ et $Y = (Y_1, \dots, Y_n)$ les coordonnées de X dans la base f. Montrer que les variables $Y_1, \dots, Y_n$ sont indépendantes de loi $\mathcal{N}(0,1)$. (On rappelle qu'il existe une matrice U de taille $n \times n$ telle que si $x = (x_1, \dots, x_n)$ sont les coordonnées d'un vecteur dans la base e, alors ses coordonnées dans la base f sont données par $y = Ux$. De plus on a $U^tU = UU^t = I_n$, où I_n est la matrice identité.)
2. Soit $E_1, \dots, E_p$ une famille de $p \geq 2$ sous-espaces vectoriels de $\mathbb{R}^n$ orthogonaux deux à deux tels que $E_1 \oplus \cdots \oplus E_p = \mathbb{R}^n$ (i.e. si $f^{(i)} = \{f_1^{(i)}, \dots, f_{n_i}^{(i)}\}$, où $n_i = \dim(E_i)$, est une base orthonormée de E_i, alors $f = \cup_{1 \leq i \leq p} f^{(i)}$ est une base orthonormée de $\mathbb{R}^n$). On note X_{E_i} la projection orthogonale de X sur E_i. Montrer que les variables $X_{E_1}, \dots, X_{E_p}$ sont indépendantes et que la loi de $\|X_{E_i}\|^2$ est une loi du χ^2 dont on déterminera le paramètre.
3. On note Δ la droite vectorielle engendrée par le vecteur unitaire $f_1 = n^{-1/2} \sum_{i=1}^n e_i$ et H le sous-espace vectoriel orthogonal (en particulier $\Delta \oplus H = \mathbb{R}^n$). Calculer X_Δ et $\|X_H\|^2 = \|X - X_\Delta\|^2$. Retrouver ainsi que la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est indépendante de $T_n = \sum_{i=1}^n \left|X_i - \bar{X}_n\right|^2$, et donner la loi de T_n.

△

XI.6.2 Loi de Bose-Einstein

Exercice XI.11.
L'énergie d'une particule est quantifiée, c'est-à-dire que les valeurs possibles de l'énergie forment un ensemble discret. Mais, pour un niveau d'énergie donné, une particule peut-être dans différents sous-états, que l'on peut décrire à l'aide du moment cinétique (nombre quantique secondaire), du moment magnétique (nombre quantique magnétique) et de la rotation propre des électrons de l'atome (spin). Il existe deux types de particules :

- Les fermions (électron, proton, neutron, etc.) ont un spin demi-entier et obéissent à la statistique de Fermi-Dirac : au plus une particule par sous-état.
- Les bosons (photon, phonon, etc.) obéissent à la statistique de Bose-Einstein : plusieurs particules peuvent occuper le même état, et les particules sont indiscernables.

Le but de ce problème est, après avoir établi la loi de Bose-Einstein, d'évaluer plusieurs quantités naturelles associées à cette loi.

I Convergence en loi pour les variables aléatoires discrètes

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires à valeurs dans $\mathbb{N}$. On pose $p_n(k) = \mathbb{P}(X_n = k)$ pour $k \in \mathbb{N}$, $n \geq 1$.

1. On suppose que, pour tout $k \in \mathbb{N}$, la suite $(p_n(k), n \geq 1)$ converge vers une limite, notée $p(k)$, et que $\sum_{k=0}^\infty p(k) = 1$. Soit g une fonction bornée mesurable. Montrer que pour $n_0 \in \mathbb{N}$ fixé, on a

$$\left|\mathbb{E}[g(X_n)] - \sum_{k=0}^\infty p(k)g(k)\right| \leq \|g\| \left[\sum_{k=0}^{n_0} |p_n(k) - p(k)| + 2 - \sum_{k=0}^{n_0} (p_n(k) + p(k))\right],$$

où $\|g\| = \sup\{|g(x)|; x \in \mathbb{R}\}$. En déduire que la suite $(X_n, n \geq 1)$ converge en loi vers la loi d'une variable aléatoire discrète X à valeurs dans $\mathbb{N}$, où $p(k) = \mathbb{P}(X = k)$ pour $k \in \mathbb{N}$.

2. (Facultatif) Montrer que si la suite $(X_n, n \geq 1)$ converge en loi (vers la loi d'une variable aléatoire X), alors pour tout $k \in \mathbb{N}$, les suites $(p_n(k), n \geq 1)$ convergent vers une limite. De plus, la variable aléatoire X est discrète à valeurs dans $\mathbb{N}$, et si on note $p(k) = \lim_{n\to\infty} p_n(k)$, alors on a $p(k) = \mathbb{P}(X = k)$ et $\sum_{k=0}^{\infty} p(k) = 1$.

II La loi de Bose-Einstein

On suppose que l'on dispose de r particules indiscernables pouvant occuper n sous-états (appelés aussi boîtes) du même niveau d'énergie. Dire que les particules sont indiscernables revient à dire que toutes les configurations sont équiprobables. Ainsi pour $r = n = 2$, on dispose des 3 configurations différentes

$$|**|\,| \quad , \quad |*|*| \quad \text{et} \quad ||**| \quad ,$$

où les étoiles représentent les particules et les barres verticales les bords des boîtes. Chaque configuration est donc de probabilité 1/3.

1. Montrer qu'il existe $\dfrac{(n+r-1)!}{r!(n-1)!}$ configurations différentes. En déduire la probabilité d'une configuration (loi de Bose-Einstein).

On suppose $n \geq 2$. L'état du système est décrit par $X_{n,r} = (X_{n,r}^{(1)}, \ldots, X_{n,r}^{(n)})$, où $X_{n,r}^{(i)}$ est le nombre de particules dans la boîte i.

2. Remarquer que si k particules sont dans la première boîte, alors il reste $r-k$ particules dans les $n-1$ autres boîtes. En déduire la loi de $X_{n,r}^{(1)}$.
3. On suppose que $r \to \infty$, $n \to \infty$ et $r/n \to \theta \in]0, \infty[$. Montrer que sous ces hypothèses la suite $(X_{n,r}^{(1)}, n \in \mathbb{N}^*, r \in \mathbb{N})$ converge en loi vers la loi d'une variable aléatoire entière X. Donner la loi de X, vérifier que la loi de $X+1$ est la loi géométrique dont on précisera le paramètre.
4. Donner la loi de $X_{n,r}^{(i)}$, pour $i \in \{1, \ldots, n\}$. Calculer $\sum_{i=1}^{n} X_{n,r}^{(i)}$. En déduire $\mathbb{E}[X_{n,r}^{(1)}]$.
5. Vérifier que si $r \to \infty$, $n \to \infty$ et $r/n \to \theta \in]0, \infty[$, alors $\mathbb{E}[X_{n,r}^{(1)}]$ converge vers $\mathbb{E}[X]$, où la variable X est définie à la question II.3.
6. (Facultatif) On suppose $r \geq 1$. Donner une relation simple entre $\mathbb{P}(X_{n+1,r-1}^{(1)} = k)$ et $\mathbb{P}(X_{n,r}^{(1)} = k)$. En déduire $\mathbb{E}[r - X_{n,r}^{(1)}]$, puis retrouver $\mathbb{E}[X_{n,r}^{(1)}]$.
7. (Facultatif) En s'inspirant de la question précédente calculer également $\mathbb{E}[(X_{n,r}^{(1)})^2]$ pour $r \geq 2$. Vérifier que si $r \to \infty$, $n \to \infty$ et $r/n \to \theta \in]0, \infty[$, alors $\mathbb{E}[(X_{n,r}^{(1)})^2]$ converge vers $\mathbb{E}[X^2]$, où la variable X est définie à la question II.3.

III Quand on augmente le nombre de particules

On suppose que l'on dispose de n boîtes et de r particules disposées dans ces boîtes suivant la loi de Bose-Einstein. Conditionnellement à l'état du système, $X_{n,r}$, quand on ajoute une

particule, elle est mise dans la boîte i avec une probabilité proportionnelle à $X_{n,r}^{(i)}+1$. Ainsi la nouvelle particule a plus de chance d'être mise dans une boîte contenant déjà beaucoup de particules. On note $X_{n,r+1}=(X_{n,r+1}^{(1)},\dots,X_{n,r+1}^{(n)})$ le nouvel état du sytème.

1. Calculer la loi de $X_{n,r+1}$ sachant $X_{n,r}$.
2. En déduire la loi de $X_{n,r+1}$, et reconnaître cette loi.

△

XI.7 Corrections

Exercice XI.1.

La fonction caractéristique de T_n est $\psi_{T_n}(u)=\dfrac{p_n\,\mathrm{e}^{iu}}{1-(1-p_n)\,\mathrm{e}^{iu}}$. On a

$$\psi_{\frac{T_n}{n}}(u)=\psi_{T_n}\left(\frac{u}{n}\right)=\frac{\theta\,\mathrm{e}^{iu/n}}{n-(n-\theta)\,\mathrm{e}^{iu/n}}.$$

Rappelons que $\mathrm{e}^{iu/n}=1+\frac{iu}{n}+o\left(\frac{1}{n}\right)$. On en déduit que

$$\psi_{\frac{T_n}{n}}(u)=\frac{\theta+o(1)}{\theta-iu+o(1)}\underset{n\to\infty}{\longrightarrow}\frac{\theta}{\theta-iu}.$$

On reconnaît dans le terme de droite la fonction caractéristique de la loi exponentielle de paramètre $\theta>0$. Donc la suite $(T_n/n, n\geq n_0)$ converge en loi vers la loi exponentielle de paramètre $\theta>0$.

▲

Exercice XI.2.

1. Par définition $\mathbb{E}[\varphi(V,W)\mid W]=h(W)$ où

$$h(w)=\int\varphi(v,w)f_{V|W}(v|w)\,dv=\int\varphi(v,w)\frac{f_{V,W}(v,w)}{f_W(w)}\,dv.$$

Comme les v.a. V et W sont indépendantes, on a $f_{V,W}(v,w)=f_V(v)f_W(w)$. Il vient

$$h(w)=\int\varphi(v,w)f_V(v)\,dv=\mathbb{E}[\varphi(V,w)].$$

2. On a par la question précédente $\mathbb{E}\left[\mathrm{e}^{iuX_1X_4-iuX_2X_3}\mid X_1,X_2\right]=h(X_1,X_2)$, où

$$\begin{aligned}h(x_1,x_2)&=\mathbb{E}[\mathrm{e}^{iux_1X_4-iux_2X_3}]\\&=\mathbb{E}[\mathrm{e}^{iux_1X_4}]\mathbb{E}[\mathrm{e}^{-iux_2X_3}]\quad\text{par indépendance}\\&=\mathrm{e}^{-\frac{u^2x_1^2}{2}-\frac{u^2x_2^2}{2}}\quad\text{en utilisant la fonction caractéristique de la loi }\mathcal{N}(0,1).\end{aligned}$$

Ensuite on a

$$\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{iuX_1X_4-iuX_2X_3}\right] &= \mathbb{E}\left[\mathbb{E}\left[\mathrm{e}^{iuX_1X_4-iuX_2X_3} \mid X_1, X_2\right]\right] \\
&= \mathbb{E}\left[h(X_1, X_2)\right] \\
&= \mathbb{E}\left[\mathrm{e}^{-\frac{u^2X_1^2}{2}-\frac{u^2X_2^2}{2}}\right] \\
&= \mathbb{E}\left[\mathrm{e}^{-\frac{u^2X_1^2}{2}}\right]^2 \quad \text{par indépendance} \\
&= \left[\int_{\mathbb{R}} \mathrm{e}^{-\frac{u^2x^2}{2}-\frac{x^2}{2}} \ \frac{dx}{\sqrt{2\pi}}\right]^2 \\
&= \left[\frac{1}{\sqrt{u^2+1}}\right]^2 \\
&= \frac{1}{u^2+1}\,.
\end{aligned}$$

Remarque : on pouvait reconnaître la fonction caractéristique de la loi exponentielle symétrique (cf exercice IV.2).

▲

Exercice XI.3.

1. **Indépendantes** et de même loi **uniforme** sur $\{1, \cdots, n\}$.
2. La fonction génératrice de T est $\Phi(z) = \frac{pz}{1-(1-p)z}$. On a $\mathbb{E}[T] = \Phi'(1) = p^{-1}$ et $\mathrm{Var}(T) = \Phi''(1) + \Phi'(1) - \Phi'(1)^2 = (1-p)p^{-2}$. On peut bien sûr faire un calcul direct.
3. On décompose suivant la valeur de la première image :

$$\begin{aligned}
\mathbb{P}(T_2 = l) &= \sum_{1\le j\le n} \mathbb{P}(X_1 = j; \cdots, X_l = j; X_{l+1} \neq j) \\
&= \sum_{1\le j\le n} \mathbb{P}(X_1 = j) \cdots \mathbb{P}(X_{l+1} \neq j) \quad \text{par } \textbf{indépendance} \\
&= \sum_{1\le j\le n} \left(\frac{1}{n}\right)^l \left(1-\frac{1}{n}\right) \\
&= \left(1-\frac{1}{n}\right)\left(\frac{1}{n}\right)^{l-1}.
\end{aligned}$$

   T_2 suit une loi géométrique de paramètre $\left(1-\frac{1}{n}\right)$.
4. On décompose suivant les cas possibles pour les images.
5. On déduit de la formule précédente en utilisant l'**indépendance** que

$$\begin{aligned}
\mathbb{P}(T_2 = l_2, T_3 = l_3) &= n(n-1)(n-2)\left(\frac{1}{n}\right)^{l_2}\frac{1}{n}\left(\frac{2}{n}\right)^{l_3-1}\frac{1}{n} \\
&= \left(1-\frac{1}{n}\right)\left(\frac{1}{n}\right)^{l_2-1}\left(1-\frac{2}{n}\right)\left(\frac{2}{n}\right)^{l_3-1}.
\end{aligned}$$

En utilisant la **formule des lois marginales**, on obtient :

$$\begin{aligned}\mathbb{P}(T_3 = l_3) &= \sum_{l_2 \geq 1} \mathbb{P}(T_2 = l_2, T_3 = l_3) \\ &= \left(1 - \frac{2}{n}\right)\left(\frac{2}{n}\right)^{l_3 - 1}\end{aligned}$$

T_3 suit une loi géométrique de paramètre $\left(1 - \frac{2}{n}\right)$.

6. On a pour tout $l_2 \geq 1, l_3 \geq 1$ que $\mathbb{P}(T_2 = l_2, T_3 = l_3) = \mathbb{P}(T_2 = l_2)\mathbb{P}(T_3 = l_3)$. Donc T_2 et T_3 sont indépendants.
7. T_k est le premier instant où l'on obtient une nouvelle image alors que l'on en possède déjà $k-1$. C'est donc une v.a. géométrique de paramètre p, où p est la probabilité de succès : obtenir une nouvelle carte alors que l'on en possède déjà $k-1$: $p = 1 - \dfrac{k-1}{n}$.
8. On a par **linéarité** de l'espérance :

$$\begin{aligned}\mathbb{E}[N_n] = \mathbb{E}\left[\sum_{k=1}^{n} T_k\right] &= \sum_{k=1}^{n} \mathbb{E}[T_k] \\ &= \sum_{k=1}^{n} \frac{n}{n-k+1} = n \sum_{j=1}^{n} \frac{1}{j}\,.\end{aligned}$$

On en déduit que $\mathbb{E}[N_n] = n\,[\log(n) + O(1)]$.

9. On a par **indépendance** :

$$\begin{aligned}\mathrm{Var}(N_n) = \mathrm{Var}\left(\sum_{k=1}^{n} T_k\right) &= \sum_{k=1}^{n} \mathrm{Var}(T_k) \\ &= \sum_{k=1}^{n} \frac{n(k-1)}{(n-k+1)^2} = \sum_{k=1}^{n} \frac{n^2}{(n-k+1)^2} - \frac{n}{(n-k+1)} \\ &= n^2 \sum_{j=1}^{n} \frac{1}{j^2} - n \sum_{j=1}^{n} \frac{1}{j} = n^2\left(\frac{\pi^2}{6} + o(1)\right) - n\,[\log(n) + O(1)]\,.\end{aligned}$$

On en déduit que $\mathrm{Var}(N_n) = n^2 \frac{\pi^2}{6} + o(n^2)$.

10. On a $\mathbf{1}_{\{x^2 > \varepsilon^2\}} \leq x^2 \varepsilon^{-2}$ pour tout $x \in \mathbb{R}$ et $\varepsilon > 0$. Par monotonie de l'espérance, il vient

$$\mathbb{P}\left(\left|\frac{N_n}{\mathbb{E}[N_n]} - 1\right| > \varepsilon\right) = \mathbb{E}\left[\mathbf{1}_{\left\{\left|\frac{N_n}{\mathbb{E}[N_n]} - 1\right|^2 > \varepsilon^2\right\}}\right] \leq \varepsilon^{-2}\mathbb{E}\left[\left(\frac{N_n}{\mathbb{E}[N_n]} - 1\right)^2\right].$$

11. On a

$$\begin{aligned}\mathbb{E}\left[\left(\frac{N_n}{\mathbb{E}[N_n]} - 1\right)^2\right] &= \frac{1}{\mathbb{E}[N_n]^2}\,\mathrm{Var}(N_n) \\ &= O\left(\frac{n^2}{n^2(\log n)^2}\right) = O\left((\log n)^{-2}\right).\end{aligned}$$

On en déduit que pour tout $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{N_n}{\mathbb{E}[N_n]} - 1\right| > \varepsilon\right) = 0$. Donc la suite $\left(\frac{N_n}{\mathbb{E}[N_n]}, n \in \mathbb{N}^*\right)$ converge en probabilité vers 1. Comme $\lim_{n\to\infty} \frac{\mathbb{E}[N_n]}{n\log n} = 1$, on en déduit que la suite $\left(\frac{N_n}{n\log n}, n \in \mathbb{N}^*\right)$ converge en probabilité vers 1.

De plus, on peut montrer que la suite $\left(\frac{N_n}{n} - \log n, n \in \mathbb{N}^*\right)$ converge en loi vers une variable aléatoire continue dont la densité est $f(x) = \mathrm{e}^{-x}\,\mathrm{e}^{-\,\mathrm{e}^{-x}}$, $x \in \mathbb{R}$ appelée loi de Gumbel (voir le contrôle ci-dessous). ▲

Exercice XI.4.

1. On utilise les fonctions **génératrices**. Par **indépendance**, on a pour $z \in [-1, 1]$,

$$\phi_{X_1+X_2}(z) = \phi_{X_1}(z)\phi_{X_2}(z) = \mathrm{e}^{-\theta_1(1-z)-\theta_2(1-z)} = \mathrm{e}^{-(\theta_1+\theta_2)(1-z)}\,.$$

   On reconnaît la fonction génératrice de la loi de Poisson de paramètre $\theta_1 + \theta_2$. La loi de $X_1 + X_2$ est donc la loi de Poisson de paramètre $\theta_1 + \theta_2$.

2. On calcule pour $0 \le k \le n$,

$$\begin{aligned}
\mathbb{P}(X_1 = k | X_1 + X_2 = n) &= \frac{\mathbb{P}(X_1 = k, X_2 = n-k)}{\mathbb{P}(X_1 + X_2 = n)} \\
&= \frac{\mathbb{P}(X_1 = k)\mathbb{P}(X_2 = n-k)}{\mathbb{P}(X_1 + X_2 = n)} \quad \text{par } \textbf{indépendance,} \\
&= \mathrm{e}^{-\theta_1}\frac{\theta_1^k}{k!}\,\mathrm{e}^{-\theta_2}\frac{\theta_2^{n-k}}{(n-k)!}\,\mathrm{e}^{(\theta_1+\theta_2)}\frac{n!}{(\theta_1+\theta_2)^n} \\
&= C_n^k \left(\frac{\theta_1}{\theta_1+\theta_2}\right)^k \left(\frac{\theta_2}{\theta_1+\theta_2}\right)^{n-k}.
\end{aligned}$$

   La loi de X_1 sachant $X_1 + X_2 = n$ est une loi binomiale de paramètres n et p, où $p = \frac{\theta_1}{\theta_1+\theta_2}$. On en déduit donc que la loi conditionnelle de X_1 sachant $X_1 + X_2$ est la loi binomiale $\mathcal{B}\left(X_1 + X_2, \frac{\theta_1}{\theta_1+\theta_2}\right)$.

3. Soit Z de loi $\mathcal{B}(n, p)$, on a $\mathbb{E}[Z] = np$. Comme $\mathcal{L}(X_1|X_1+X_2) = \mathcal{B}\left(X_1 + X_2, \frac{\theta_1}{\theta_1+\theta_2}\right)$, on en déduit que

$$\mathbb{E}[X_1|X_1 + X_2] = (X_1 + X_2)\frac{\theta_1}{\theta_1+\theta_2}.$$

▲

Exercice XI.5.

I Comportement asymptotique de $X_{(n)} = \sum_{i=1}^n Y_i$.

1. On a pour $x \geq 0$,

$$\begin{aligned}\mathbb{P}(X_{(n)} \leq x) &= \mathbb{P}(\max_{1\leq i\leq n} X_i \leq x)\\ &= \mathbb{P}(X_1 \leq x, \ldots, X_n \leq x)\\ &= \mathbb{P}(X_1 \leq x)\cdots\mathbb{P}(X_n \leq x) \quad \text{par } \textbf{indépendance},\\ &= (1-\mathrm{e}^{-\lambda x})^n.\end{aligned}$$

2. On utilise la méthode de la **fonction de répartition**. On a pour $x + \lambda^{-1}\log n \geq 0$,

$$\mathbb{P}(X_{(n)} - \lambda^{-1}\log n \leq x) = \mathbb{P}(X_{(n)} \leq x + \lambda^{-1}\log n) = \left(1 - \frac{\mathrm{e}^{-\lambda x}}{n}\right)^n.$$

   Cette quantité converge vers $F(x) = \mathrm{e}^{-\,\mathrm{e}^{-\lambda x}}$ pour tout $x \in \mathbb{R}$. Comme $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$, on en déduit que la suite $(X_{(n)} - \lambda^{-1}\log n, n \in \mathbb{N}^*)$ converge en loi vers une variable aléatoire Z de fonction de répartition F.
3. On a $f(x) = \mathrm{e}^{-x}\,\mathrm{e}^{-\,\mathrm{e}^{-x}}$ pour $x \in \mathbb{R}$. On obtient $\mathrm{e}^{-\,\mathrm{e}^{-a}} = 0,025$ soit $a \simeq -1,3$ et $\mathrm{e}^{-\,\mathrm{e}^{-b}} = 1 - 0,025$ soit $b \simeq 3,7$. La loi de densité f porte le nom de loi de Gumbel.

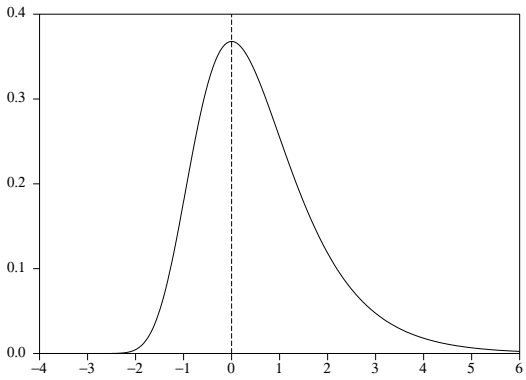

FIG. XI.1 – Densité de la loi de Gumbel

II Loi du vecteur $(Y_1, \ldots, Y_n)$.

1. Comme les lois possèdent des densités et que les v.a. sont indépendantes, $\mathbb{P}(X_i = X_j) = 0$ si $i \neq j$. En effet :

$$\mathbb{P}(X_i = X_j) = \mathbb{E}[\mathbf{1}_{\{X_i=X_j\}}] = \int_{\mathbb{R}_+^2} \mathbf{1}_{\{x=y\}}\lambda^2\,\mathrm{e}^{-\lambda x-\lambda y}\; dxdy = 0.$$

2. On a :

$$\mathbb{P}(\exists i \neq j; X_i = X_j) \leq \sum_{i\neq j} \mathbb{P}(X_i = X_j) = 0.$$

Pour presque tout $\omega \in \Omega$, les réels $X_1(\omega), \dots, X_n(\omega)$ sont distincts deux à deux. Il existe donc un unique réordonnement croissant.

3. Par la **formule de décomposition**, on a

$$\begin{aligned}\mathbb{E}[g_1(Y_1)g_2(Y_2)] &= \mathbb{E}[g_1(X_1)g_2(X_2-X_1)\mathbf{1}_{\{X_1<X_2\}}] + \mathbb{E}[g_1(X_2)g_2(X_1-X_2)\mathbf{1}_{\{X_2<X_1\}}] \\ &= 2\int g_1(x_1)g_2(x_2-x_1)\mathbf{1}_{\{0<x_1<x_2\}}\lambda^2\,\mathrm{e}^{-\lambda x_1-\lambda x_2}\;dx_1dx_2 \\ &= 2\int g_1(x_1)g_2(y_2)\mathbf{1}_{\{0<x_1\}}\mathbf{1}_{\{0<y_2\}}\lambda^2\,\mathrm{e}^{-\lambda x_1-\lambda(y_2+x_1)}\;dx_1dy_2,\end{aligned}$$

où on a posé $y_2 = x_2 - x_1$, à x_1 fixé. Il vient

$$\mathbb{E}[g_1(Y_1)g_2(Y_2)] = \int g_1(y_1)g_2(y_2)\;2\lambda^2\,\mathrm{e}^{-2\lambda y_1-\lambda y_2}\,\mathbf{1}_{\{y_1>0,y_2>0\}}\;dy_1dy_2,$$

où on a posé $y_1 = x_1$. En faisant $g_2 = \mathbf{1}$, on obtient

$$\begin{aligned}\mathbb{E}[g_1(Y_1)] &= \int g_1(y_1)\;2\lambda^2\,\mathrm{e}^{-2\lambda y_1-\lambda y_2}\,\mathbf{1}_{\{y_1>0,y_2>0\}}\;dy_1dy_2 \\ &= \int g_1(y_1)\;2\lambda\,\mathrm{e}^{-2\lambda y_1}\,\mathbf{1}_{\{y_1>0\}}\;dy_1.\end{aligned}$$

On en déduit que la loi de Y_1 a pour densité $f_{Y_1}(y_1) = 2\lambda\,\mathrm{e}^{-2\lambda y_1}\,\mathbf{1}_{\{y_1>0\}}$. On reconnaît la loi exponentielle de paramètre 2λ.

4. Un calcul similaire montre que la loi de Y_2 est la loi exponentielle de paramètre λ : $f_{Y_2}(y_2) = \lambda\,\mathrm{e}^{-\lambda y_2}\,\mathbf{1}_{\{y_2>0\}}$. On remarque enfin que pour toutes fonctions g_1, g_2 mesurables bornées, on a :

$$\mathbb{E}[g_1(Y_1)g_2(Y_2)] = \mathbb{E}[g_1(Y_1)]\mathbb{E}[g_2(Y_2)].$$

Cela implique que les v.a. Y_1 et Y_2 sont **indépendantes**. La densité de la loi du couple est donc la densité **produit** : $f_{(Y_1,Y_2)}(y_1,y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$.

5. En **décomposant** suivant les évènements $\{X_{\sigma(1)} < \cdots < X_{\sigma(n)}\}$, et en utilisant le fait que $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$ et $(X_1, \dots, X_n)$ ont même loi, on a :

$$\begin{aligned}\mathbb{E}[g_1(Y_1)\cdots g_n(Y_n)] &= \sum_{\sigma\in\mathcal{S}_n}\mathbb{E}[g_1(Y_1)\cdots g_n(Y_n)\mathbf{1}_{\{X_{\sigma(1)}<\cdots<X_{\sigma(n)}\}}] \\ &= \sum_{\sigma\in\mathcal{S}_n}\mathbb{E}[g_1(X_{\sigma(1)})\cdots g_n(X_{\sigma(n)}-X_{\sigma(n-1)})\mathbf{1}_{\{X_{\sigma(1)}<\cdots<X_{\sigma(n)}\}}]. \\ &= \sum_{\sigma\in\mathcal{S}_n}\mathbb{E}[g_1(X_1)\cdots g_n(X_n-X_{n-1})\mathbf{1}_{\{X_1<\cdots<X_n\}}]. \\ &= n!\,\mathbb{E}[g_1(X_1)\cdots g_n(X_n-X_{n-1})\mathbf{1}_{\{X_1<\cdots<X_n\}}] \\ &= n!\int g_1(x_1)\cdots g_n(x_n-x_{n-1}) \\ &\qquad\qquad \mathbf{1}_{\{0<x_1<\cdots<x_n\}}\,\lambda^n\,\mathrm{e}^{-\lambda\sum_{i=1}^n x_i}\;dx_1\cdots dx_n.\end{aligned}$$

Pour vérifier que $(X_{\sigma(1)},\dots,X_{\sigma(n)})$ et $(X_1,\dots,X_n)$ ont même loi, on remarque que pour $u=(u_1,\dots,u_n)\in\mathbb{R}^n$, on a

$$\begin{aligned}\psi_{(X_{\sigma(1)},\dots,X_{\sigma(n)})}(u) &= \mathbb{E}\left[\mathrm{e}^{i\sum_{k=1}^n X_{\sigma(k)}u_k}\right]\\ &= \mathbb{E}\left[\mathrm{e}^{i\sum_{j=1}^n X_j u_{\sigma^{-1}(j)}}\right], \quad \text{où } u_{\sigma^{-1}}=(u_{\sigma^{-1}(1)},\dots,u_{\sigma^{-1}(n)})\\ &= \prod_{j=1}^n \mathbb{E}\left[\mathrm{e}^{iX_j u_{\sigma^{-1}(j)}}\right], \quad \text{par indépendance}\\ &= \prod_{k=1}^n \mathbb{E}\left[\mathrm{e}^{iX_1 u_k}\right], \quad \text{car les v.a. ont même loi}\\ &= \psi_{(X_1,\dots,X_n)}(u).\end{aligned}$$

On en déduit donc que les vecteurs $(X_{\sigma(1)},\dots,X_{\sigma(n)})$ et $(X_1,\dots,X_n)$ ont même loi.

6. On considère l'application φ définie sur $\mathbb{R}^n$ par $\varphi(x_1,\dots,x_n)=(y_1,\dots,y_n)$, où $y_1=x_1$ et pour $i>1$, $y_i=x_i-x_{i-1}$. L'application φ est un C^1 **difféomorphisme** de l'**ouvert** $\{(x_1,\dots,x_n);0<x_1<\cdots<x_n\}$ dans l'**ouvert** $\{(y_1,\dots,y_n)\in]0,+\infty[^n\}$. Le **Jacobien** de l'application φ est

$$\mathrm{Jac}[\varphi](\mathrm{x}) = \begin{pmatrix} 1 & 0 & \dots & \dots & 0\\ -1 & 1 & 0 & \dots & 0\\ 0 & -1 & 1 & \dots & 0\\ \vdots & \ddots & \ddots & \ddots & \vdots\\ 0 & \dots & 0 & -1 & 1\end{pmatrix}$$

Son déterminant est $\det(\mathrm{Jac}[\varphi](\mathrm{x}))=1$. Remarquons enfin que $x_i=\sum_{j=1}^i y_j$, et donc

$$\sum_{i=1}^n x_i = \sum_{i=1}^n\sum_{j=1}^i y_j = \sum_{j=1}^n y_j\sum_{i=j}^n 1 = \sum_{j=1}^n (n-j+1)y_j.$$

On en déduit en faisant le changement de variables $(y_1,\dots,y_n)=\varphi(x_1,\dots,x_n)$, que

$$\begin{aligned}&\mathbb{E}[g_1(Y_1)\cdots g_n(Y_n)]\\ &\qquad = n!\int g_1(x_1)\cdots g_n(x_n-x_{n-1})\mathbf{1}_{\{x_1<\cdots<x_n\}}\,\lambda^n\,\mathrm{e}^{-\lambda\sum_{i=1}^n x_i}\,dx_1\cdots dx_n\\ &\qquad = n!\int g_1(y_1)\cdots g_n(y_n)\mathbf{1}_{\{0<y_1,\dots,0<y_n\}}\,\lambda^n\,\mathrm{e}^{-\lambda\sum_{j=1}^n (n-j+1)y_j}\,dy_1\cdots dy_n.\end{aligned}$$

En posant $g_j=\mathbf{1}$ si $j\neq i$, on obtient

$$\begin{aligned}\mathbb{E}[g_i(Y_i)] &= n!\int g_i(y_i)\mathbf{1}_{\{0<y_1,\dots,0<y_n\}}\,\lambda^n\,\mathrm{e}^{-\lambda\sum_{j=1}^n (n-j+1)y_j}\,dy_1\cdots dy_n\\ &= \int g_i(y_i)\,\mathbf{1}_{\{0<y_i\}}(n-i+1)\lambda\,\mathrm{e}^{-\lambda(n-i+1)y_i}\,dy_i.\end{aligned}$$

La loi de Y_i est donc une loi exponentielle de paramètre $n-i+1$. Remarquons enfin que pour toutes fonctions $g_1, \dots g_n$, mesurables bornées, on a

$$\mathbb{E}[g_1(Y_1)\cdots g_n(Y_n)] = \mathbb{E}[g_1(Y_1)]\cdots\mathbb{E}[g_n(Y_n)].$$

Les variables aléatoires $Y_1, \dots, Y_n$ sont donc **indépendantes**. La densité de la loi du vecteur $(Y_1, \dots, Y_n)$ est donc le **produit des densités des lois marginales**.

7. Comme $X_{(n)} = \sum_{i=1}^n Y_i$, on a pour $u \in \mathbb{R}$, en utilisant l'indépendance des variables aléatoires $Y_1, \dots, Y_n$,

$$\psi_{X_{(n)}}(u) = \psi_{\sum_{i=1}^n Y_i}(u) = \prod_{i=1}^n \psi_{Y_i}(u) = \prod_{i=1}^n \frac{(n-i+1)\lambda}{(n-i+1)\lambda - iu} = \prod_{k=1}^n \frac{k\lambda}{k\lambda - iu}.$$

III Application

1. La loi de $T_{n-k+1,n}$ est la loi géométrique de paramètre k/n. La fonction caractéristique de la loi de $\dfrac{1}{n}T_{n-k+1,n}$ est pour $u \in \mathbb{R}$,

$$\psi_{k,n}(u) = \frac{k}{n}\frac{\mathrm{e}^{iu/n}}{1-\left(1-\frac{k}{n}\right)\mathrm{e}^{iu/n}} = k\frac{1+o(1)}{n-(n-k)\left(1+\frac{iu}{n}+o(1/n)\right)} = \frac{k}{k-iu}(1+o(1)).$$

La suite $(\psi_{k,n}, n > k)$ converge vers la fonction ψ_k, où $\psi_k(u) = \dfrac{k}{k-iu}$ est la fonction caractéristique de la loi exponentielle de paramètre k. On en déduit que la suite $\left(\dfrac{1}{n}T_{n-k+1,n}, n \geq k\right)$ converge en loi vers la loi exponentielle de paramètre k.
Nous regardons maintenant en détail $\psi_{k,n} - \psi_k$. On a

$$\begin{aligned}\psi_{k,n}(u) - \psi_k(u) &= \frac{k}{n}\frac{\mathrm{e}^{iu/n}}{1-\left(1-\frac{k}{n}\right)\mathrm{e}^{iu/n}} - \frac{k}{k-iu}\\ &= \frac{k}{k-n(1-\mathrm{e}^{-iu/n})} - \frac{k}{k-iu}\\ &= k\frac{-iu+n(1-\mathrm{e}^{-iu/n})}{(k-iu)(k-n(1-\mathrm{e}^{-iu/n}))}.\end{aligned}$$

Remarquons que $\left|\dfrac{k}{k-iu}\right| \leq 1$ et $\left|-iu+n(1-\mathrm{e}^{-iu/n})\right| \leq C\dfrac{u^2}{n}$, où C est une constante indépendante de n et u. Enfin, on a $\left|k-n(1-\mathrm{e}^{-iu/n})\right| \geq |k-n(1-\cos(u/n))|$. Comme $|1-\cos(x)| \leq x^2/2$, on en déduit que pour u fixé, et n assez grand, on a $n(1-\cos(u/n)) \leq u^2/2n < 1/2$. Pour tout $k \in \{1, \dots, n\}$, on a alors $\left|k-n(1-\mathrm{e}^{-iu/n})\right| \geq k-1/2 \geq k/2$. On en déduit donc que pour tout $u \in \mathbb{R}$, il existe $n(u)$ fini, tel que pour tout $n \geq n(u)$ et $k \in \{1, \dots, n\}$,

$$|\psi_{k,n}(u) - \psi_k(u)| \leq \frac{1}{kn}\,2Cu^2,$$

Quitte à remplacer C par $2Cu^2$, où u est fixé, on obtient bien la majoration annoncée.

2. Comme les v.a. $T_{1,n},\dots,T_{n,n}$ sont indépendantes, on a $\psi_{N_n/n}(u)=\prod_{k=1}^{n}\psi_{k,n}(u)$. On déduit de la dernière question de la partie II que

$$\left|\psi_{N_n/n}(u)-\psi_{X_{(n)}}(u)\right|=\left|\prod_{k=1}^{n}\psi_{k,n}(u)-\prod_{k=1}^{n}\psi_k(u)\right|.$$

Toute fonction caractéristique est majorée en module par 1. On peut donc utiliser le lemme V.25 page 89. On obtient

$$\left|\psi_{N_n/n}(u)-\psi_{X_{(n)}}(u)\right|\le\sum_{k=1}^{n}|\psi_{k,n}(u)-\psi_k(u)|\le C\frac{\log n}{n}.$$

3. On utilise les fonctions caractéristiques. On a

$$\begin{aligned}&\left|\psi_{(N_n-n\log n)/n}(u)-\psi_Z(u)\right|\\&\qquad\le\left|\psi_{(N_n-n\log n)/n}(u)-\psi_{(X_{(n)}-n\log n)}(u)\right|+\left|\psi_{(X_{(n)}-n\log n)}(u)-\psi_Z(u)\right|.\end{aligned}$$

Après avoir remarqué que

$$\psi_{(N_n-n\log n)/n}(u)-\psi_{(X_{(n)}-n\log n)}(u)=\mathrm{e}^{-iu\,n\log n}\left(\psi_{N_n/n}(u)-\psi_{X_{(n)}}(u)\right),$$

On en déduit que

$$\left|\psi_{(N_n-n\log n)/n}(u)-\psi_Z(u)\right|\le\left|\psi_{N_n/n}(u)-\psi_{X_{(n)}}(u)\right|+\left|\psi_{(X_{(n)}-n\log n)}(u)-\psi_Z(u)\right|.$$

Soit $u\in\mathbb{R}$ fixé. Le premier terme du membre de droite converge vers 0 quand $n\to\infty$. On déduit de la convergence en loi de la suite $(X_{(n)}-n\log n, n\in\mathbb{N}^*)$ vers Z, que le deuxième terme converge également vers 0 quand $n\to\infty$. Par conséquent, on a :

$$\lim_{n\to\infty}\psi_{(N_n-n\log n)/n}(u)=\psi_Z(u).$$

La suite $(n^{-1}(N_n-n\log n), n\in\mathbb{N}^*)$ converge en loi vers la variable aléatoire Z.

4. Les points de discontinuité de la fonction $\mathbf{1}_{[a,b]}(x)$ sont $\{a,b\}$. Comme $\mathbb{P}(Z\in\{a,b\})=0$, on déduit de la question précédente que

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}\left((N_n-n\log n)/n\in[a,b]\right)&=\mathbb{P}(Z\in[a,b])\\&=1-\int_{-\infty}^{a}f(x)\,dx-\int_{b}^{+\infty}f(x)\,dx.\end{aligned}$$

Comme $\mathbb{P}\left((N_n-n\log n)/n\in[a,b]\right)=\mathbb{P}\left(N_n\in[an+n\log n,bn+n\log n]\right)$, on en déduit que $I_n=[an+n\log n,bn+n\log n]$ est un intervalle de confiance de niveau asymptotique $\int_a^b f(x)\,dx$ pour N_n. On choisit par exemple a et b tels que

$$\int_{-\infty}^{a}f(x)\,dx=\int_{b}^{+\infty}f(x)\,dx=2,5\%,$$

soit $a=-1,31$ et $b=3,68$.

5. Pour n, on note $r_n = n \log n$ le nombre moyen de plaquettes à acheter pour avoir une collection complète, et I_n l'intervalle de confiance, de niveau 95%, pour le nombre de plaquettes que l'on risque d'acheter afin d'avoir une collection complète.

n	r_n	I_n
151	758	$[560, 1313]$
250	1380	$[1054, 2299]$

▲

Exercice XI.6.

1. On a $\mathbb{P}(\exists i \in \{1, \dots, n\}; U_i \leq 0) \leq \sum_{i=1}^{n} \mathbb{P}(U_i \leq 0) = 0$. La variable aléatoire X_n est donc strictement positive p.s. On peut donc considérer $V_n = \log(X_n) = \frac{1}{n} \sum_{i=1}^{n} \alpha \log(U_i)$. Les variables aléatoires $(\alpha \log(U_i), i \in \mathbb{N}^*)$ sont indépendantes et de même loi. Elles sont également intégrables :

$$\mathbb{E}[|\log(U_i)|] = \int |\log(u)| \, \mathbf{1}_{]0,1[}(u) \, du = -\int_{]0,1[} \log(u) \, du = -[u \log(u) - u]_0^1 = 1.$$

On déduit de la **loi forte des grands nombres** que la suite $(V_n, n \in \mathbb{N}^*)$ converge presque sûrement vers $\mathbb{E}[\alpha \log(U_1)] = -\alpha$. Comme la fonction exponentielle est continue, on en déduit que la suite $(X_n, n \in \mathbb{N}^*)$ converge presque sûrement vers $\mathrm{e}^{-\alpha}$.

2. On a

$$\log(Y_n) = \sqrt{n}(\log(X_n) + \alpha) = \sqrt{n}\left(\frac{1}{n} \sum_{i=1}^{n} \alpha \log(U_i) - \mathbb{E}[\alpha \log(U_i)] \right).$$

Vérifions que les variables aléatoires $\alpha \log(U_i)$ sont de carré intégrables. On a

$$\mathbb{E}[\log(U_i)^2] = \int \log(u)^2 \mathbf{1}_{]0,1[}(u) \, du = [u \log(u)^2 - 2u \log(u) + 2u]_0^1 = 2.$$

On déduit du **théorème de la limite centrale** que la suite $(\log(Y_n), n \in \mathbb{N}^*)$ converge en loi vers Z de loi gaussienne $\mathcal{N}(0, \sigma^2)$, où $\sigma^2 = \mathrm{Var}(\alpha \log(U_i)) = \alpha^2(2-1) = \alpha^2$. Comme la fonction exponentielle est continue, on en déduit que la suite $(Y_n, n \in \mathbb{N}^*)$ converge en loi vers e^Z. Il reste donc à déterminer la loi de $Y = \mathrm{e}^Z$. On utilise la méthode de la fonction muette. Soit g une fonction mesurable bornée. On a

$$\begin{aligned} \mathbb{E}[g(\mathrm{e}^Z)] &= \int g(\mathrm{e}^z) f_Z(z) dz = \int g(\mathrm{e}^z) \, \mathrm{e}^{-z^2/2\alpha^2} \, \frac{dz}{\sqrt{2\pi\alpha^2}} \\ &= \int g(y) \, \frac{1}{y\sqrt{2\pi\alpha^2}} \, \mathrm{e}^{-\log(y)^2/2\alpha^2} \, \mathbf{1}_{]0,\infty[}(y) \, dy, \end{aligned}$$

où l'on a effectué le changement de variable $y = \mathrm{e}^z$ de $\mathbb{R}$ dans $]0,\infty[$. On en déduit donc que $Y = \mathrm{e}^Z$ est une variable aléatoire continue de densité $\dfrac{1}{y\sqrt{2\pi\alpha^2}}\,\mathrm{e}^{-\log(y)^2/2\alpha^2}\,\mathbf{1}_{]0,\infty[}(y)$. La loi de Y est appelée loi log-normale.

▲

Exercice XI.7.

I Préliminaires

1. Le temps moyen est $\mathbb{E}[T_1] = 1/\lambda$.
2. On considère les fonctions caractéristiques : pour $n \geq 1$, $u \in \mathbb{R}$,
$$\begin{aligned}\psi_{V_n}(u) &= \psi_{\sum_{i=1}^n T_i}(u)\\ &= \prod \psi_{T_i}(u) \quad \text{par indépendance}\\ &= \left(\frac{\lambda}{\lambda - iu}\right)^n.\end{aligned}$$
On reconnaît la fonction caractéristique de la loi gamma de paramètre (λ, n).
3. $\mathbb{P}(N_t = 0) = \mathbb{P}(T_1 > t) = \mathrm{e}^{-\lambda t}$.
4. On a $\{N_t = n\} = \{V_n \leq t < V_n + T_{n+1}\}$.
5. Soit $n \geq 1$. Comme V_n et T_{n+1} sont indépendantes, on a
$$\begin{aligned}\mathbb{P}(N_t = n) &= \mathbb{P}(V_n \leq t < V_n + T_{n+1})\\ &= \int \mathbf{1}_{\{v\leq t<v+w\}}\,\frac{1}{(n-1)!}\,\lambda^n v^{n-1}\,\mathrm{e}^{-\lambda v}\,\mathbf{1}_{\{v>0\}}\lambda\,\mathrm{e}^{-\lambda w}\,\mathbf{1}_{\{w>0\}}\,dvdw\\ &= \int \frac{1}{(n-1)!}\,\lambda^n v^{n-1}\,\mathrm{e}^{-\lambda v}\,\lambda\,\mathrm{e}^{-\lambda w}\,\mathbf{1}_{\{0<v\leq t<w+v\}}\,dvdw\\ &= \int \frac{1}{(n-1)!}\,\lambda^n v^{n-1}\,\mathrm{e}^{-\lambda v}\,\mathrm{e}^{-\lambda(t-v)}\,\mathbf{1}_{\{0<v\leq t\}}\,dv\\ &= \frac{1}{(n-1)!}\,\lambda^n\,\mathrm{e}^{-\lambda t}\int v^{n-1}\mathbf{1}_{\{0<v\leq t\}}\,dv\\ &= \mathrm{e}^{-\lambda t}\,\frac{(\lambda t)^n}{n!}.\end{aligned}$$
On reconnaît la loi de Poisson de paramètre λt.

II Les temps moyens

1. Comme $\sum_{i=1}^{N_t} T_i \geq 0$, on en déduit que $R_t \leq t$ p.s. Donc $\mathbb{P}(R_t \leq t) = 1$. On a $\{N_t = 0\} = \{R_t = t\}$. Donc si $r < t$, on a
$$\mathbb{P}(R_t \leq r, S_t \leq s, N_t = 0) = 0.$$
Si $r \geq t$, alors
$$\mathbb{P}(R_t \leq r, S_t \leq s, N_t = 0) = \mathbb{P}(t < T_1 \leq t + s) = \mathrm{e}^{-\lambda t}(1 - \mathrm{e}^{-\lambda s}).$$

2. Soit $n \geq 1$. On pose $I = \mathbb{P}(R_t \leq r, S_t \leq s, N_t = n)$. On a

$$\begin{aligned} I &= \mathbb{P}(t - V_n \leq r, V_n + T_{n+1} \leq t + s, V_n \leq t < V_n + T_{n+1}) \\ &= \mathbb{P}((t-r) \leq V_n \leq t < V_n + T_{n+1} \leq t + s) \\ &= \int \mathbf{1}_{\{t-r \leq v \leq t < v+w \leq t+s\}} \frac{1}{(n-1)!} \lambda^n v^{n-1} \mathrm{e}^{-\lambda v} \mathbf{1}_{\{0<v\}} \lambda \mathrm{e}^{-\lambda w} \mathbf{1}_{\{0<w\}} \, dvdw, \end{aligned}$$

car les variables V_n et T_{n+1} sont indépendantes. Il vient

$$\begin{aligned} I &= \int \mathbf{1}_{\{(t-r)_+ \leq v \leq t\}} \frac{1}{(n-1)!} \lambda^n v^{n-1} \mathrm{e}^{-\lambda v} \mathbf{1}_{\{t-v<w\leq t+s-v\}} \lambda \mathrm{e}^{-\lambda w} \, dvdw \\ &= \int \mathbf{1}_{\{(t-r)_+ \leq v \leq t\}} \frac{1}{(n-1)!} \lambda^n v^{n-1} \mathrm{e}^{-\lambda v} \mathrm{e}^{-\lambda(t-v)} (1 - \mathrm{e}^{-\lambda s}) \, dv \\ &= \frac{1}{(n-1)!} \lambda^n \mathrm{e}^{-\lambda t} (1 - \mathrm{e}^{-\lambda s}) \int \mathbf{1}_{\{(t-r)_+ \leq v \leq t\}} v^{n-1} \, dv \\ &= \mathrm{e}^{-\lambda t} (1 - \mathrm{e}^{-\lambda s}) \frac{\lambda^n}{n!} [t^n - (t-r)_+^n]. \end{aligned}$$

3. En décomposant suivant les valeurs possibles de N_t, on obtient

$$\begin{aligned} \mathbb{P}(R_t \leq r, S_t \leq s) &= \sum_{n=0}^{\infty} \mathbb{P}(R_t \leq r, S_t \leq s, N_t = n) \\ &= \mathrm{e}^{-\lambda t} (1 - \mathrm{e}^{-\lambda s}) \left\{ \mathbf{1}_{\{r \geq t\}} + \sum_{n=1}^{\infty} \frac{\lambda^n}{n!} [t^n - (t-r)_+^n] \right\} \\ &= \mathrm{e}^{-\lambda t} (1 - \mathrm{e}^{-\lambda s}) \left\{ \mathrm{e}^{\lambda t} - \mathrm{e}^{\lambda(t-r)} \mathbf{1}_{\{r<t\}} \right\} \\ &= (1 - \mathrm{e}^{-\lambda s})(1 - \mathrm{e}^{-\lambda r} \mathbf{1}_{\{r<t\}}). \end{aligned}$$

4. Comme $R_t \leq t$ p.s., on a $F(s) = \mathbb{P}(S_t \leq s) = \mathbb{P}(R_t \leq t, S_t \leq s) = (1 - \mathrm{e}^{-\lambda s})$. Comme S_t est positif p.s., on en déduit que sa fonction de répartition est $F(s) = (1 - \mathrm{e}^{-\lambda s}) \mathbf{1}_{\{s>0\}}$. La fonction de répartition F de S_t est une fonction dérivable. On en déduit que S_t est une variable continue de densité $F'(s) = \lambda \mathrm{e}^{-\lambda s} \mathbf{1}_{\{s>0\}}$. On en déduit que la loi de S_t est la loi exponentielle de paramètre λ.

5. La variable aléatoire S_t est fine p.s. On en déduit que

$$\mathbb{P}(R_t \leq r) = \lim_{s \to \infty} \mathbb{P}(R_t \leq r, S_t \leq s) = (1 - \mathrm{e}^{-\lambda r} \mathbf{1}_{\{r<t\}}).$$

Calculons la fonction de répartition de $\min(T_1, t)$. On a

$$\mathbb{P}(\min(T_1, t) \leq r) = \mathbf{1}_{\{r \geq t\}} + \mathbf{1}_{\{r<t\}} \mathbb{P}(T_1 \leq r) = (1 - \mathrm{e}^{-\lambda r} \mathbf{1}_{\{r<t\}}).$$

Comme les fonctions de répartitions caractérisent la loi, on en déduit que R_t et $\min(T_1, t)$ ont même loi.

6. Le temps moyen d'attente d'un bus quand on arrive à l'instant t est $\mathbb{E}[S_t] = 1/\lambda$. L'écart moyen entre le dernier bus juste avant l'instant t et le premier bus juste après l'instant

t est $\mathbb{E}[S_t + R_t] = \mathbb{E}[S_t] + \mathbb{E}[R_t]$. On a

$$\begin{aligned}\mathbb{E}[R_t] = \mathbb{E}[\min(T_1, t)] &= \int \min(w,t)\lambda\,\mathrm{e}^{-\lambda w}\,\mathbf{1}_{\{w>0\}}\,dw \\ &= \int_0^t w\lambda\,\mathrm{e}^{-\lambda w}\,dw + \int_t^\infty t\lambda\,\mathrm{e}^{-\lambda w}\,dw = [-w\,\mathrm{e}^{-\lambda w}]_0^t + \int_0^t \mathrm{e}^{-\lambda w}\,dw + t\,\mathrm{e}^{-\lambda t} \\ &= \frac{1}{\lambda}(1-\mathrm{e}^{-\lambda t}).\end{aligned}$$

On en déduit donc que $\mathbb{E}[S_t + R_t] = \dfrac{2}{\lambda} - \dfrac{1}{\lambda}\,\mathrm{e}^{-\lambda t}$. L'écart moyen entre le dernier bus juste avant l'instant t et le premier bus juste après l'instant t est donc différent de l'écart moyen entre deux bus ! Le client et la société de bus ont tous les deux raison.

III Loi du temps entre deux passages de bus

1. On a $\mathbb{P}(R_t \le r, S_t \le s) = \mathbb{P}(R_t \le r)\mathbb{P}(S_t \le s)$ pour tout $r, s \in \mathbb{R}$. On déduit que les variables R_t et S_t sont indépendantes.
2. Remarquons que U_t a même loi que $W_t = \min(T_1, t) + T_2$. Comme T_1 est une variable aléatoire finie presque sûrement, on en déduit que $\lim_{t\to\infty} W_t = T_1 + T_2$ pour la convergence p.s. En particulier la suite $(W_t, t > 0)$ converge en loi vers $T_1 + T_2$. Comme U_t a même loi que W_t, on en déduit que la suite $(U_t, t > 0)$ converge en loi vers $T_1 + T_2$. De plus la loi de $T_1 + T_2$ est la loi gamma de paramètre $(\lambda, 2)$.
3. On remarque que (R_t, S_t) a même loi que $(\min(T_1, t), T_2)$. Soit g une fonction bornée mesurable. Pour tout $t > 0$, on a

$$\begin{aligned}\mathbb{E}[g(U_t)] = \mathbb{E}[g(R_t + S_t)] &= \mathbb{E}[g(\min(T_1,t) + T_2)] \\ &= \int g(\min(t_1,t) + t_2)\lambda^2\,\mathrm{e}^{-\lambda(t_1+t_2)}\,\mathbf{1}_{\{0<t_1,0<t_2\}}\,dt_1dt_2 \\ &= \mathrm{e}^{-\lambda t}\int g(t+t_2)\lambda\,\mathrm{e}^{-\lambda t_2}\,\mathbf{1}_{\{0<t_2\}}\,dt_2 \\ &\qquad + \int g(t_1+t_2)\lambda^2\,\mathrm{e}^{-\lambda(t_1+t_2)}\,\mathbf{1}_{\{0<t_1<t,0<t_2\}}\,dt_1dt_2 \\ &= \int g(u)\lambda\,\mathrm{e}^{-\lambda u}\,\mathbf{1}_{\{t<u\}}\,du \\ &\qquad + \int g(u)\lambda^2\,\mathrm{e}^{-\lambda u}\,\mathbf{1}_{\{t_2<u<t+t_2,0<t_2\}}\,dudt_2 \\ &= \int g(u)\lambda\,\mathrm{e}^{-\lambda u}[\mathbf{1}_{\{t<u\}} + \lambda(u-(u-t)_+)\mathbf{1}_{\{0<u\}}]\,du \\ &= \int g(u)\lambda\,\mathrm{e}^{-\lambda u}[\lambda u\mathbf{1}_{\{0<u\le t\}} + (\lambda t+1)\mathbf{1}_{\{t<u\}}]\,du.\end{aligned}$$

On en déduit que U_t est une variable aléatoire continue de densité $\lambda\,\mathrm{e}^{-\lambda u}[\lambda u\mathbf{1}_{\{0<u\le t\}} + (\lambda t+1)\mathbf{1}_{\{t<u\}}]$.

▲

Exercice XI.8.

I Calculs préliminaires

1. La variable aléatoire $\mathbf{1}_{\{Y\leq X\}}$ prend les valeurs 0 ou 1. Il s'agit d'une variable aléatoire de Bernoulli. Son paramètre est $p = \mathbb{P}(\mathbf{1}_{\{Y\leq X\}} = 1) = \mathbb{E}\left[\mathbf{1}_{\{Y\leq X\}}\right] = \mathbb{P}(Y \leq X)$. Sa variance est $\mathrm{Var}(\mathbf{1}_{\{Y\leq X\}}) = p(1-p)$.
2. On a
$$p = \mathbb{E}\left[\mathbf{1}_{\{Y\leq X\}}\right] = \int \mathbf{1}_{\{y\leq x\}}\, f(x)g(y)\, dxdy = \int \left(\int_{-\infty}^{x} g(y)\, dy\right) f(x)\, dx$$
$$= \int G(x)f(x)\, dx.$$
En intégrant d'abord en x puis en y, on obtient également
$$p = \int (1 - F(y))g(y)\, dy.$$
Si X et Y ont même loi, alors on a
$$p = \int F(x)f(x)\, dx = \left[F(x)^2/2\right]_{-\infty}^{+\infty} = 1/2.$$
3. Comme X et Y sont indépendants, la densité conditionnelle de Y sachant X est la densité de Y. On a donc
$$\mathbb{E}\left[\mathbf{1}_{\{Y\leq X\}}|X = x\right] = \int \mathbf{1}_{\{y\leq x\}} g(y)\, dy = [G(y)]_{-\infty}^{x} = G(x).$$
Cela implique $S = \mathbb{E}\left[\mathbf{1}_{\{Y\leq X\}}|X\right] = G(X)$. On a
$$\mathbb{E}[S] = \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{Y\leq X\}}|X\right]\right] = \mathbb{E}\left[\mathbf{1}_{\{Y\leq X\}}\right] = p.$$
Des calculs similaires donnent $T = 1 - F(Y)$ et $\mathbb{E}[T] = p$.
4. On a
$$\alpha = \mathrm{Var}(S) = \mathbb{E}[G(X)^2] - p^2 = \int G(x)^2 f(x)\, dx - p^2.$$
De façon similaire, on obtient $\beta = \mathrm{Var}(T) = \int (1 - F(y))^2\, g(y)dy - p^2$.
Vérifions par l'absurde que $(\alpha, \beta) \neq (0, 0)$. Si $\alpha = 0$, alors $\mathrm{Var}(S) = 0$ et donc $S = p$ p.s., c'est-à-dire $G(X) = p$ p.s. Et donc $X \in G^{-1}(\{p\}) = [a, b]$. En particulier, cela implique que $F(x) = 0$ si $x < a$ et $F(x) = 1$ si $x \geq b$. Comme G est constant sur $[a, b]$, cela implique que $Y \notin [a, b]$ p.s. Donc $F(Y)$ est une variable aléatoire de Bernoulli de paramètre $q = \mathbb{P}(Y \geq b)$. Sa variance est $q(1 - q)$. Par définition, elle est égale à β. Si $\beta = 0$, alors $F(Y) = 1$ p.s. ou $F(Y) = 0$ p.s. En prenant l'espérance, cela implique donc que $\mathbb{P}(Y \leq X) = 0$ ou $\mathbb{P}(Y \leq X) = 1$. Ce qui contredit l'hypothèse $p \in]0, 1[$.
5. Si X et Y ont même loi, alors on a
$$\alpha = \int F(x)^2 f(x)\, dx - p^2 = \left[\frac{F(x)^3}{3}\right]_{-\infty}^{\infty} - \frac{1}{4} = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}\,.$$
Si X et Y ont même loi, alors par symétrie, S et T on même loi. En particulier $\beta = \alpha = 1/12$.

6. On a

$$\mathbb{E}\left[S\mathbf{1}_{\{Y\leq X\}}\right] = \mathbb{E}\left[G(X)\mathbf{1}_{\{Y\leq X\}}\right] = \int G(x)\mathbf{1}_{\{y\leq x\}}g(y)f(x)\,dxdy = \int G(x)^2 f(x)\,dx.$$

On en déduit donc que

$$\mathrm{Cov}(S,\mathbf{1}_{\{Y\leq X\}}) = \int G(x)^2 f(x)\,dx - \mathbb{E}[S]\mathbb{E}\left[\mathbf{1}_{\{Y\leq X\}}\right] = \alpha.$$

Des calculs similaires donnent $\mathrm{Cov}(T,\mathbf{1}_{\{Y\leq X\}}) = \beta$.

II Étude de la projection de Hajek de la statistique de Mann et Withney

1. Par indépendance, pour $l \neq i$, on a

$$\mathbb{E}\left[\mathbf{1}_{\{Y_j\leq X_l\}}|X_i\right] = \mathbb{E}\left[\mathbf{1}_{\{Y_j\leq X_l\}}\right] = p.$$

On déduit de ce calcul et de la partie précédente que

$$\mathbb{E}[U^*_{m,n}|X_i] = \sum_{j=1}^{n}\mathbb{E}\left[\mathbf{1}_{\{Y_j\leq X_i\}} - p|X_i\right] = n(G(X_i)-p) = n(S_i-p).$$

Un raisonnement similaire donne

$$\mathbb{E}[U^*_{m,n}|Y_j] = \sum_{i=1}^{m}\mathbb{E}\left[\mathbf{1}_{\{Y_j\leq X_i\}} - p|Y_j\right] = m(1-F(Y_j)-p) = m(T_j-p).$$

On obtient donc

$$H_{m,n} = n\sum_{i=1}^{m}(S_i-p) + m\sum_{j=1}^{n}(T_j-p).$$

2. Les variables aléatoires $(n(S_i-p), i\geq 1)$ et $(m(T_j-p), j\geq 1)$ sont indépendantes. Cela implique

$$\begin{aligned}
\mathrm{Var}(H_{m,n}) &= \mathrm{Var}\left(\sum_{i=1}^{m} n(S_i-p) + \sum_{j=1}^{n} m(T_j-p)\right)\\
&= \sum_{i=1}^{m}\mathrm{Var}(n(S_i-p)) + \sum_{j=1}^{n}\mathrm{Var}(m(T_j-p))\\
&= \sum_{i=1}^{m} n^2\,\mathrm{Var}(S_i) + \sum_{j=1}^{n} m^2\,\mathrm{Var}(T_j)\\
&= mn^2\alpha + nm^2\beta.
\end{aligned}$$

3. Les variables aléatoires $((S_i-p), i\geq 1)$ sont indépendantes, de même loi, et de carré intégrable. Remarquons que $\mathbb{E}[S_i-p] = 0$ et $\mathrm{Var}(S_i-p) = \alpha$. On déduit du théorème

central limite, que la suite $\left(\frac{1}{\sqrt{m}}\sum_{i=1}^{m}(S_i-p), m\geq 1\right)$ converge en loi vers une loi gaussienne $\mathcal{N}(0,\alpha)$. Un raisonnement similaire assure que la suite $\left(\frac{1}{\sqrt{n}}\sum_{j=1}^{n}(T_j-p), n\geq 1\right)$ converge en loi vers la loi gaussienne $\mathcal{N}(0,\beta)$.

4. On a pour tout $v,w\in\mathbb{R}$, en utilisant l'indépendance entre V_k et W_k,

$$\psi_{(V_k,W_k)}(v,w)=\psi_{V_k}(v)\psi_{W_k}(w).$$

On déduit de la question précédente que

$$\lim_{k\to\infty}\psi_{V_k}(v)=\mathrm{e}^{-\alpha v^2/2}\quad\text{et}\quad\lim_{k\to\infty}\psi_{W_k}(w)=\mathrm{e}^{-\beta w^2/2}.$$

On en déduit donc que

$$\lim_{k\to\infty}\psi_{(V_k,W_k)}(v,w)=\mathrm{e}^{-(\alpha v^2+\beta w^2)/2}.$$

On reconnaît la fonction caractéristique du couple (V,W), où V et W sont indépendants de loi respective $\mathcal{N}(0,\alpha)$ et $\mathcal{N}(0,\beta)$. On en déduit que le vecteur

$$\left(\left(\frac{1}{\sqrt{m_k}}\sum_{i=1}^{m_k}(S_i-p),\frac{1}{\sqrt{n_k}}\sum_{j=1}^{n_k}(T_j-p)\right),k\geq 1\right).$$

converge en loi vers le vecteur gaussien (V,W).

5. On pose

$$\begin{pmatrix}a_k\\b_k\end{pmatrix}=\frac{1}{\sqrt{m_kn_k^2\alpha+n_km_k^2\beta}}\begin{pmatrix}n_k\sqrt{m_k}\\m_k\sqrt{n_k}\end{pmatrix}.$$

Remarquons que

$$\lim_{k\to\infty}\begin{pmatrix}a_k\\b_k\end{pmatrix}=\begin{pmatrix}a\\b\end{pmatrix},$$

où

$$\begin{pmatrix}a\\b\end{pmatrix}=\frac{1}{\sqrt{(1-\theta)\alpha+\theta\beta}}\begin{pmatrix}\sqrt{1-\theta}\\\sqrt{\theta}\end{pmatrix}.$$

On déduit du théorème de Slutsky, la convergence en loi de la suite $((a_k,b_k,V_k,W_k),k\geq 1)$ vers (a,b,V,W). Comme l'application $(a',b',V',W')\mapsto a'V'+b'W'$ est continue, on en déduit que la suite $(a_kV_k+b_kW_k,k\geq 1)$ converge en loi vers $aV+bW$. Comme (V,W) est un vecteur gaussien centré, on obtient que la loi de $aV+bW$ est la loi gaussienne $\mathcal{N}(0,\sigma^2)$, avec $\sigma^2=\mathrm{Var}(aV+bW)$. Comme V et W sont des variables aléatoires indépendantes, on a

$$\sigma^2=\mathrm{Var}(aV)+\mathrm{Var}(bW)=a^2\alpha+b^2\beta=\frac{1-\theta}{(1-\theta)\alpha+\theta\beta}\alpha+\frac{\theta}{(1-\theta)\alpha+\theta\beta}\beta=1.$$

Ainsi la suite $\left(H_{m_k,n_k}/\sqrt{\mathrm{Var}(H_{m_k,n_k})},k\geq 1\right)$ converge en loi vers Z de loi $\mathcal{N}(0,1)$.

6. On pose $Z_k = H_{m_k,n_k}/\sqrt{\mathrm{Var}(H_{m_k,n_k})}$ et $c_k = \sqrt{\mathrm{Var}(H_{m_k,n_k})/\mathrm{Var}(U_{m_k,n_k})}$. Comme
$$\lim_{k\to\infty} c_k = 1,$$
on déduit du théorème de Slutsky que la suite $((c_k, Z_k), k \geq 1)$ converge en loi vers $(1, Z)$. Comme la fonction $(c', Z') \mapsto c'Z'$ est continue, on en déduit la convergence en loi de la suite $(c_k Z_k, k \geq 1)$ vers Z. C'est-à-dire la suite $\left(H_{m_k,n_k}/\sqrt{\mathrm{Var}(U_{m_k,n_k})}, k \geq 1\right)$ converge en loi vers la loi gaussienne centrée réduite $\mathcal{N}(0,1)$.

III Convergence de la statistique de Mann et Whitney

1. On a
$$\begin{aligned}
\mathrm{Cov}(H_{m,n}, U^*_{m,n}) &= \mathrm{Cov}(H_{m,n}, U_{m,n}) \\
&= \mathrm{Cov}(n\sum_{i=1}^{m} S_i, U_{m,n}) + \mathrm{Cov}(m\sum_{j=1}^{n} T_j, U_{m,n}) \\
&= n\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{Cov}(S_i, \mathbf{1}_{\{Y_j \leq X_i\}}) + m\sum_{i=1}^{m}\sum_{j=1}^{n} \mathrm{Cov}(T_j, \mathbf{1}_{\{Y_j \leq X_i\}}) \\
&= mn^2\ \mathrm{Cov}(S, \mathbf{1}_{\{Y \leq X\}}) + nm^2\ \mathrm{Cov}(T, \mathbf{1}_{\{Y \leq X\}}).
\end{aligned}$$

2. On déduit de ce qui précède que
$$\begin{aligned}
&\mathrm{Var}(H_{m,n} - U^*_{m,n}) \\
&= \mathrm{Var}(H_{m,n}) + \mathrm{Var}(U^*_{m,n}) - 2\,\mathrm{Cov}(H_{m,n}, U^*_{m,n}) \\
&= mn^2\alpha + nm^2\beta + mnp(1-p) + n(n-1)m\alpha + m(m-1)n\beta - 2mn^2\alpha - 2nm^2\beta \\
&= mnp(1-p) - mn\alpha - mn\beta \\
&= mn(p - p^2 - \alpha - \beta).
\end{aligned}$$

3. En particulier, on a
$$\lim_{k\to\infty} \frac{\mathrm{Var}(H_{m_k,n_k} - U^*_{m_k,n_k})}{\mathrm{Var}(U_{m_k,n_k})} = 0.$$

4. On pose $Q_k = \dfrac{H_{m_k,n_k} - U^*_{m_k,n_k}}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}}$. En utilisant l'inégalité de Tchebychev, on a pour tout $\varepsilon > 0$,
$$\mathbb{P}(|Q_k| > \varepsilon) \leq \mathbb{E}[Q_k^2]/\varepsilon^2.$$
Comme $\lim_{k\to\infty} \mathbb{E}[Q_k^2] = 0$, on déduit que la suite $\left(\dfrac{H_{m_k,n_k} - U^*_{m_k,n_k}}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}}, k \geq 1\right)$ converge en probabilité vers 0.

5. On a
$$\frac{U_{m_k,n_k} - m_k n_k p}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}} = c_k Z_k - Q_k.$$

On déduit du théorème de Slutsky que la suite $((c_k Z_k, -Q_k), k \leq 1)$ converge en loi vers $(Z, 0)$. Par continuité de l'addition, on en déduit que la suite $(c_k Z_k - Q_k, k \geq 1)$ converge en loi vers Z. On a donc montré que la suite

$$\left(\frac{U_{m_k,n_k} - m_k n_k p}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}}, k \geq 1 \right)$$

converge en loi vers la loi gaussienne centrée $\mathcal{N}(0,1)$.

IV Calcul de la variance de la statistique de Mann et Whitney

1. On a

$$\begin{aligned}
\mathrm{Card}\,(\Delta_1) &= mn \\
\mathrm{Card}\,(\Delta_2) &= m(m-1)n \\
\mathrm{Card}\,(\Delta_3) &= n(n-1)m \\
\mathrm{Card}\,(\Delta_4) &= m(m-1)n(n-1).
\end{aligned}$$

2. On a

$$\begin{aligned}
\mathrm{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y' \leq X\}}) &= \mathbb{E}\left[\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y' \leq X\}}\right] - \mathbb{E}\left[\mathbf{1}_{\{Y \leq X\}}\right] \mathbb{E}\left[\mathbf{1}_{\{Y' \leq X\}}\right] \\
&= \int \mathbf{1}_{\{y \leq x\}}, \mathbf{1}_{\{y' \leq x\}} g(y) g(y') f(x)\, dy dy' dx - p^2 \\
&= \int G(x)^2 f(x)\, dx - p^2 \\
&= \alpha.
\end{aligned}$$

Un calcul similaire donne $\mathrm{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y \leq X'\}}) = \beta$.

3. On a

$$\mathrm{Var}(U_{m,n}) = \mathbb{E}\left[(U^*_{m,n})^2\right] = \sum_{(i,i',j,j') \in \Delta} \mathbb{E}\left[(\mathbf{1}_{\{Y_j \leq X_i\}} - p)(\mathbf{1}_{\{Y_{j'} \leq X_{i'}\}} - p)\right].$$

On décompose la somme sur les ensembles d'indices $\Delta_1, \Delta_2, \Delta_3$ et Δ_4. Il vient

$$\begin{aligned}
\sum_{(i,i',j,j') \in \Delta_1} \mathbb{E}\left[(\mathbf{1}_{\{Y_j \leq X_i\}} - p)(\mathbf{1}_{\{Y_{j'} \leq X_{i'}\}} - p)\right] &= \mathrm{Card}\,(\Delta_1)\, \mathbb{E}\left[(\mathbf{1}_{\{Y \leq X\}} - p)^2\right] \\
&= \mathrm{Card}\,(\Delta_1)\ \mathrm{Var}(\mathbf{1}_{\{Y \leq X\}}) \\
&= mnp(1-p),
\end{aligned}$$

$$\begin{aligned}
\sum_{(i,i',j,j') \in \Delta_2} \mathbb{E}\left[(\mathbf{1}_{\{Y_j \leq X_i\}} - p)(\mathbf{1}_{\{Y_{j'} \leq X_{i'}\}} - p)\right] &= \mathrm{Card}\,(\Delta_2)\, \mathbb{E}\left[(\mathbf{1}_{\{Y \leq X\}} - p)(\mathbf{1}_{\{Y \leq X'\}} - p)\right] \\
&= \mathrm{Card}\,(\Delta_2)\ \mathrm{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y \leq X'\}}) \\
&= m(m-1)n\beta,
\end{aligned}$$

$$\sum_{(i,i',j,j')\in\Delta_3} \mathbb{E}\left[(\mathbf{1}_{\{Y_j\le X_i\}}-p)(\mathbf{1}_{\{Y_{j'}\le X_{i'}\}}-p)\right] = \text{Card }(\Delta_3)\ \mathbb{E}\left[(\mathbf{1}_{\{Y\le X\}}-p)(\mathbf{1}_{\{Y'\le X\}}-p)\right]$$
$$= \text{Card }(\Delta_3)\ \text{Cov}(\mathbf{1}_{\{Y\le X\}},\mathbf{1}_{\{Y'\le X\}})$$
$$= n(n-1)m\alpha,$$

$$\sum_{(i,i',j,j')\in\Delta_4} \mathbb{E}\left[(\mathbf{1}_{\{Y_j\le X_i\}}-p)(\mathbf{1}_{\{Y_{j'}\le X_{i'}\}}-p)\right] = \text{Card }(\Delta_4)\ \mathbb{E}\left[(\mathbf{1}_{\{Y\le X\}}-p)(\mathbf{1}_{\{Y'\le X'\}}-p)\right]$$
$$= 0.$$

La dernière égalité s'obtient en utilisant l'indépendance entre (X,Y) et (X',Y'). On obtient donc

$$\text{Var}(U_{m,n}) = mnp(1-p) + n(n-1)m\alpha + m(m-1)n\beta.$$

4. Dans le cas où les variables aléatoires $(X_i, i\ge 1)$ et $(Y_j, j\ge 1)$ ont même loi, on a

$$\text{Var}(U_{m,n}) = mn\,\frac{1}{2}(1-\frac{1}{2}) + \frac{n(n-1)m}{12} + \frac{m(m-1)n}{12}$$
$$= \frac{mn(m+n+1)}{12}.$$

▲

Exercice XI.9.

I Calcul de la probabilité d'extinction

1. On a $\{Z_n = 0\} \subset \{Z_{n+1} = 0\}$ pour tout $n \ge 0$. Ceci implique que $\{Z_n = 0\} = \cup_{k=1}^n \{Z_k = 0\}$. On en déduit que $\{\text{il existe } n \ge 0 \text{ tel que } Z_n = 0\} = \cup_{k\ge 1}\{Z_k = 0\}$ est la limite croissante de la suite d'évènements $(\{Z_n = 0\}, n \ge 0)$. On déduit donc du théorème de convergence monotone que η est la limite croissante de la suite $(\mathbb{P}(Z_n = 0), n \ge 0)$.
2. On a pour $k \ge 1$,

$$\mathbb{E}[Z_{n+1}|Z_n = k] = \mathbb{E}[\sum_{i=1}^{Z_n} \xi_{i,n}|Z_n = k] = \mathbb{E}[\sum_{i=1}^{k} \xi_{i,n}|Z_n = k] = \mathbb{E}[\sum_{i=1}^{k} \xi_{i,n}] = km,$$

où l'on a utilisé pour la 3-ième égalité le fait que les variables $(\xi_{i,n}, i \ge 1)$ sont indépendantes de $(\xi_{i,l}, i \ge 1, 0 \le l < n)$ et donc indépendantes de Z_n. Le résultat est trivialement vrai pour $k = 0$. On en déduit donc que $\mathbb{E}[Z_{n+1}|Z_n] = mZ_n$, et que $\mathbb{E}[Z_{n+1}] = m\mathbb{E}[Z_n]$. En itérant, on obtient que $\mathbb{E}[Z_n] = m^n$.
3. Remarquons que $\mathbb{P}(Z_n > 0) \le \mathbb{E}[Z_n]$. Si $m < 1$, alors on a $\lim_{n\to\infty} \mathbb{P}(Z_n > 0) = 0$, et donc $\eta = 1 - \lim_{n\to\infty} \mathbb{P}(Z_n > 0) = 1$.

4. On a pour $k \geq 1$,

$$\mathbb{E}[z^{Z_{n+1}}|Z_n = k] = \mathbb{E}[z^{\sum_{i=1}^{Z_n}\xi_{i,n}}|Z_n = k] = \mathbb{E}[z^{\sum_{i=1}^{k}\xi_{i,n}}|Z_n = k] = \mathbb{E}[z^{\sum_{i=1}^{k}\xi_{i,n}}] = \phi(z)^k,$$

où l'on a utilisé pour la 3-ième égalité le fait que les variables $(\xi_{i,n}, i \geq 1)$ sont indépendantes de $(\xi_{i,l}, i \geq 1, 0 \leq l < n)$ et donc indépendantes de Z_n. Le résultat est trivialement vrai pour $k = 0$. On en déduit donc que $\mathbb{E}[z^{Z_{n+1}}|Z_n] = \phi(z)^{Z_n}$, et donc comme $\phi(z) \in [-1, 1]$, on a

$$\phi_{n+1}(z) = \mathbb{E}[z^{Z_{n+1}}] = \mathbb{E}[\phi(z)^{Z_n}] = \phi_n(\phi(z)).$$

On en déduit donc que ϕ_n est la fonction ϕ composée avec elle-même n fois. En particulier on a $\phi_{n+1} = \phi \circ \phi_n$.

5. On a

$$\mathbb{P}(Z_{n+1} = 0) = \phi_{n+1}(0) = \phi(\phi_n(0)) = \phi(\mathbb{P}(Z_n = 0)).$$

Comme la fonction ϕ est continue sur $[-1, 1]$, on en déduit par passage à la limite, quand $n \to \infty$, que η est solution de l'équation (1).

6. On a $\phi'(1) = \mathbb{E}[\xi] = m$. Si $p_0 + p_1 = 1$ alors on a $m < 1$ car $p_0 > 0$. Par contraposée, on en déduit que si $m \geq 1$, alors $p_0 + p_1 < 1$. En particulier, il existe $k \geq 2$ tel que $p_k > 0$. Cela implique que pour $z \in]0, 1[$, alors $\phi''(z) > 0$. Et donc la fonction ϕ est strictement convexe sur $[0, 1]$.

7. On a $\phi(1) = 1$. Si $m = 1$, alors $\phi'(1) = 1$, et comme la fonction ϕ est strictement convexe, on en déduit que $\phi(x) > x$ sur $[0, 1[$. En particulier, la seule solution de (1) sur $[0, 1]$ est donc 1. Ainsi on a $\eta = 1$.

8. On a $\phi(1) = 1$. Si $m > 1$, alors la fonction $\phi(x) - x$ est strictement convexe sur $[0, 1]$, strictement positive en 0, nulle en 1 et de dérivée positive en 1. En particulier elle possède un unique zéro, x_0 sur $]0, 1[$.

9. On démontre la propriété par récurrence. Remarquons que $\mathbb{P}(Z_0 = 0) = 0 \leq x_0$. Supposons que $\mathbb{P}(Z_n = 0) \leq x_0$. Comme la fonction ϕ est croissante, on en déduit que $\phi(\mathbb{P}(Z_n = 0)) \leq \phi(x_0) = x_0$. Comme $\phi(\mathbb{P}(Z_n = 0)) = \mathbb{P}(Z_{n+1} = 0)$, on en déduit que $\mathbb{P}(Z_{n+1} = 0) \leq x_0$. Ce qui démontre que $\mathbb{P}(Z_n = 0) \leq x_0$ pour tout $n \geq 0$. Par passage à la limite, on a $\eta = \lim_{n\to\infty} \mathbb{P}(Z_n = 0) \leq x_0$. Comme $\eta \in \{x_0, 1\}$, on en déduit que $\eta = x_0$.

II Comportement asymptotique sur un exemple

1. On obtient

$$\phi(z) = \alpha + \frac{(1-\alpha)(1-\beta)z}{1-\beta z}. \tag{XI.4}$$

On a $m = \phi'(1) = \dfrac{1-\alpha}{1-\beta} > 1$. Enfin si $\phi(x) = x$ alors on a $\beta x^2 - (\alpha + \beta)x + \alpha = 0$. Comme 1 est racine de cette équation, on en déduit que l'autre racine est $x_0 = \alpha/\beta$. Comme on a supposé $m > 1$, on a $\eta = \alpha/\beta$. On obtient les valeurs numériques suivantes : $m \simeq 1.175$ et $\eta \simeq 0.8616$.

2. Il est facile de vérifier que
$$\frac{\phi(z)-1}{\phi(z)-\eta} = m\frac{z-1}{z-\eta}.$$
On en déduit donc par itération que
$$\frac{\phi_n(z)-1}{\phi_n(z)-\eta} = m^n\frac{z-1}{z-\eta},$$
soit
$$\phi_n(z) = \frac{z-\eta-m^n\eta z+m^n\eta}{z-\eta-m^n z+m^n}.$$

3. On a
$$\psi_{XY}(u) = \mathbb{E}[\mathrm{e}^{iuXY}] = \mathbb{E}[\mathbf{1}_{\{X=0\}}] + \mathbb{E}[\mathbf{1}_{\{X=1\}}\,\mathrm{e}^{iuY}] = (1-p)+p\frac{\lambda}{\lambda-iu}.$$

4. La fonction caractéristique, ψ, de ξ se déduit de (XI.4), en remarquant que $\psi(u)=\phi(\mathrm{e}^{iu})$ pour $u\in\mathbb{R}$. En reproduisant les calculs qui permettent de déterminer ϕ_n, on obtient $\psi_{Z_n}(u)=\phi_n(\mathrm{e}^{iu})$. Il vient
$$\begin{aligned}\psi_{m^{-n}Z_n}(u) &= \psi_{Z_n}(m^{-n}u)\\ &= \phi_n(\mathrm{e}^{im^{-n}u})\\ &= \frac{1-\eta-m^n\eta(1+im^{-n}u)+m^n\eta+o(1)}{1-\eta-m^n(1+im^{-n}u)+m^n+o(1)}\\ &= \frac{1-\eta-iu\eta+o(1)}{1-\eta-iu+o(1)}.\end{aligned}$$
On en déduit donc que
$$\lim_{n\to\infty}\psi_{m^{-n}Z_n}(u) = \frac{1-\eta-iu\eta}{1-\eta-iu} = \eta+(1-\eta)\frac{1-\eta}{1-\eta-iu}.$$
On en déduit donc que la suite $(m^{-n}Z_n, n\geq 1)$ converge en loi vers XY, où X et Y sont indépendants, X de loi de Bernoulli de paramètre $1-\eta$, et Y de loi exponentielle de paramètre $1-\eta$.
En fait on peut montrer que la suite $(m^{-n}Z_n, n\geq 1)$ converge p.s. dans cet exemple, et même dans un cadre plus général.

▲

Exercice XI.10.

1. En utilisant la matrice de changement de base, on a $Y=UX$ où $X=(X_1,\ldots,X_n)$. Comme X est un vecteur gaussien de moyenne nulle et de matrice de covariance la matrice identité, I_n, on en déduit que Y est un vecteur gaussien de moyenne $U\mathbb{E}[X]=0$ et de matrice de covariance $UI_nU^t=I_n$. Ainsi X et Y ont même loi.

2. On note $Y = (Y_1, \dots, Y_n)$ les coordonnées du vecteur X dans la base f. Ainsi on a $X_{E_i} = \sum_{j=1}^{n_i} Y_{m_i+j} f_{m_i+j}$, où $m_i = 0$ si $i = 1$ et $m_i = \sum_{k=1}^{i-1} n_i$ sinon. D'après la question précédente, les variables $Y_1, \dots, Y_n$ sont indépendantes de loi $\mathcal{N}(0,1)$. On en déduit donc que les variables $X_{E_1}, \dots, X_{E_p}$ sont indépendantes. On a également

$$\|X_{E_i}\|^2 = \sum_{j=1}^{n_i} Y_{m_i+j}^2.$$

On en déduit que $\|X_{E_i}\|^2$ est la somme de n_i carré de gaussiennes centrées réduites indépendantes. Sa loi est donc la loi du χ^2 à n_i degrés de liberté.

3. On a $X_\Delta = (X, f_1) f_1 = \bar{X}_n \sum_{i=1}^n e_i$ et

$$\|X_H\|^2 = \|X - X_\Delta\|^2 = \sum_{i=1}^n \left|X_i - \bar{X}_n\right|^2 = T_n.$$

D'après la question précédente X_Δ et X_H sont indépendants. En particulier $\bar{X}_n = (X_\Delta, f_1)/\sqrt{n}$ et $T_n = \|X_H\|^2$ sont indépendants. De plus, comme H est de dimension $n-1$, la loi de T_n est la loi du χ^2 à $n-1$ degrés de liberté.

▲

Exercice XI.11.

I Convergence en loi pour les variables aléatoires discrètes

1. Comme $\sum_{k=0}^\infty p(k) = 1$, on en déduit qu'il existe une variable aléatoire, X, à valeurs dans $\mathbb{N}$ telle que $\mathbb{P}(X = k) = p(k)$ pour tout $k \in \mathbb{N}$.
Soit g une fonction bornée mesurable. On a

$$\left|\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]\right| = \left|\sum_{k=0}^\infty p_n(k)g(k) - \sum_{k=0}^\infty p(k)g(k)\right| \leq \|g\| \sum_{k=0}^\infty |p_n(k) - p(k)|.$$

On a de plus

$$\sum_{k=n_0+1}^\infty |p_n(k) - p(k)| \leq \sum_{k=n_0+1}^\infty (p_n(k) + p(k)) = 2 - \sum_{k=0}^{n_0} (p_n(k) + p(k)),$$

car $\sum_{k=0}^\infty p_n(k) = \sum_{k=0}^\infty p(k) = 1$. On en déduit donc

$$\left|\mathbb{E}[g(X_n)] - \sum_{k=0}^\infty p(k)g(k)\right| \leq \|g\| \left[\sum_{k=0}^{n_0} |p_n(k) - p(k)| + 2 - \sum_{k=0}^{n_0} (p_n(k) + p(k))\right].$$

Soit $\varepsilon > 0$. Il existe $n_0 \in \mathbb{N}$ tel que $\sum_{k=n_0+1}^\infty p(k) \leq \varepsilon$ (i.e. $1 - \sum_{k=0}^{n_0} p(k) \leq \varepsilon$). Comme $\lim_{n\to\infty} p_n(k) = p(k)$ pour tout $k \in \mathbb{N}$, il existe $N \geq 1$ tel que pour tout $n \geq N$, $\sum_{k=0}^{n_0} |p_n(k) - p(k)| \leq \varepsilon$. Enfin, on remarque que

$$\left|\sum_{k=0}^{n_0} p_n(k) - \sum_{k=0}^{n_0} p(k)\right| \leq \sum_{k=0}^{n_0} |p_n(k) - p(k)| \leq \varepsilon,$$

et donc

$$-\sum_{k=0}^{n_0} p_n(k) \leq -\sum_{k=0}^{n_0} p(k) + \varepsilon.$$

On en déduit que

$$\left|\mathbb{E}[g(X_n)] - \mathbb{E}[g(X)]\right| \leq \|g\| \left[2\varepsilon + 2 - 2\sum_{k=0}^{n_0} p(k)\right] \leq 4\,\|g\|\,\varepsilon.$$

Ceci implique que pour toute fonction g continue bornée $\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$. Ainsi la suite $(X_n, n \geq 1)$ converge en loi vers la loi de X.

2. Soit $g(x) = \max(1 - 2\,|x|, 0)$, pour $x \in \mathbb{R}$. Pour $k \in \mathbb{N}$, on pose $g_k(\cdot) = g(\cdot - k)$. Ainsi, on a $g_k(X_n) = \mathbf{1}_{\{X_n = k\}}$ et donc $\mathbb{E}[g_k(X_n)] = p_n(k)$. La fonction g_k étant continue bornée, la convergence en loi de la suite $(X_n, n \geq 1)$ implique la convergence de la suite $(p_n(k), n \geq 1)$ (on note $p(k) = \mathbb{E}[g_k(X)]$ la limite), et ce pour tout $k \in \mathbb{N}$. La fonction $G(x) = \sum_{k=0}^{\infty} g_k(x)$ est continue bornée. On a $\mathbb{E}[G(X_n)] = \sum_{k=0}^{\infty} p_n(k) = 1$. Comme

$$\lim_{n\to\infty} \mathbb{E}[G(X_n)] = \mathbb{E}[G(X)] = \sum_{k=0}^{\infty} \mathbb{E}[g_k(X)] = \sum_{k=0}^{\infty} p(k),$$

on en déduit donc que $\sum_{k=0}^{\infty} p(k) = 1$. D'après la question précédente, on en déduit que la suite $(X_n, n \geq 1)$ converge en loi vers la loi de Y, variable aléatoire discrète à valeurs dans $\mathbb{N}$ telle que $\mathbb{P}(Y = k) = p(k)$, pour tout $k \in \mathbb{N}$. Par unicité de la loi limite, cela implique que X et Y ont même loi. Donc X est une variable aléatoire discrète à valeurs dans $\mathbb{N}$ telle que $\mathbb{P}(X = k) = p(k)$, pour tout $k \in \mathbb{N}$.

II La loi de Bose-Einstein

1. On considère une urne contenant $n - 1$ boules marquées "|" et r boules marquées "$*$". Et on effectue un tirage sans remise de toutes les boules. Quitte à rajouter une boule "|" au début de la séquence et une à la fin, on remarque qu'un tirage complet correspond à une (et une seule) configuration possible. Et une configuration possible est également représentée par un tirage complet. Il existe $C_{n+r-1}^{n-1} = \dfrac{(n+r-1)!}{r!(n-1)!}$ tirages possibles. Les configurations étant équiprobables, on en déduit que la probabilité d'obtenir une configuration donnée est $\dfrac{r!(n-1)!}{(n+r-1)!}$.

2. Si $X_{n,r}^{(1)} = k$, alors il faut répartir $r - k$ particules indiscernables dans $n - 1$ boîtes. Il existe donc $C_{n+r-2-k}^{n-2} = \dfrac{(n+r-2-k)!}{(r-k)!(n-2)!}$ configurations possibles. On en déduit donc que pour $k \in \{0, \ldots, r\}$,

$$\mathbb{P}(X_{n,r}^{(1)} = k) = (n-1)\frac{r!(n+r-2-k)!}{(r-k)!(n+r-1)!} = \frac{(n-1)}{n+r-1}\frac{r!}{(r-k)!}\frac{(n+r-2-k)!}{(n+r-2)!}.$$

3. Pour $k \in \mathbb{N}$, on pose $p_{n,r}(k) = \mathbb{P}(X_{n,r}^{(1)} = k)$ pour $k \leq r$ et $p_{n,r}(k) = 0$ si $k \geq r+1$. Sous les hypothèses de la question, on obtient

$$p(0) = \lim\ p_{n,r}(0) = \lim\ \frac{1-\frac{1}{n}}{1+\frac{r}{n}-\frac{1}{n}} = \frac{1}{1+\theta},$$

et pour $k \geq 1$,

$$p(k) = \lim\ p_{n,r}(k) = \lim\ \frac{1-\frac{1}{n}}{1+\frac{r}{n}-\frac{1}{n}}\ \frac{\frac{r}{n}}{1+\frac{r}{n}-\frac{2}{n}} \cdots \frac{\frac{r}{n}-\frac{k-1}{n}}{1+\frac{r}{n}-\frac{2}{n}-\frac{k-1}{n}} = \frac{\theta^k}{(1+\theta)^{k+1}}.$$

On remarque que $\sum_{k=0}^{\infty} p(k) = 1$. On déduit alors de la partie I, que la suite $(X_{n,r}, n \in \mathbb{N}^*, r \in \mathbb{N})$ converge en loi vers la loi de X, où pour $k \in \mathbb{N}$, $\mathbb{P}(X = k) = \dfrac{\theta^k}{(1+\theta)^{k+1}}$.

On pose $Y = X + 1$. Pour $k \in \mathbb{N}^*$, on a

$$\mathbb{P}(Y = k) = \mathbb{P}(X = k-1) = \frac{\theta^{k-1}}{(1+\theta)^k} = \frac{1}{1+\theta}\left(1 - \frac{1}{1+\theta}\right)^{k-1}.$$

La loi de Y est la loi géométrique de paramètre $\rho = 1/(1+\theta)$.

4. Par symétrie la loi de $X_{n,r}^{(i)}$ est la loi de $X_{n,r}^{(1)}$. (Si $X_{n,r}^{(i)} = k$, alors il faut répartir $r-k$ particules indiscernables dans les $n-1$ autres boîtes, cf question II.2.) Le nombre total de particules est r. On en déduit donc que $\sum_{i=1}^{n} X_{n,r}^{(i)} = r$. Par linéarité de l'espérance, puis de l'égalité en loi, on déduit

$$r = \mathbb{E}\Big[\sum_{i=1}^{n} X_{n,r}^{(i)}\Big] = \sum_{i=1}^{n} \mathbb{E}[X_{n,r}^{(i)}] = n\mathbb{E}[X_{n,r}^{(1)}].$$

Ainsi, on a obtenu que $\mathbb{E}[X_{n,r}^{(1)}] = r/n$.

5. On a $\lim \mathbb{E}[X_{n,r}^{(1)}] = \theta$, et $\mathbb{E}[X] = \mathbb{E}[Y-1] = \dfrac{1}{\rho} - 1 = \theta$.

6. On a pour $k \in \{0, \ldots, r-1\}$,

$$\mathbb{P}(X_{n+1,r-1}^{(1)} = k) = n\frac{(r-1)!(n+r-2-k)!}{(r-1-k)!(n+r-1)!} = \frac{n}{n-1}\frac{r-k}{r}\mathbb{P}(X_{n,r}^{(1)} = k),$$

soit

$$(r-k)\mathbb{P}(X_{n,r}^{(1)} = k) = r\frac{n-1}{n}\mathbb{P}(X_{n+1,r-1}^{(1)} = k).$$

Cette égalité est trivialement vérifiée pour $k \geq r$. Il vient donc

$$\mathbb{E}[r - X_{n,r}^{(1)}] = \sum_{k=0}^{\infty}(r-k)\mathbb{P}(X_{n,r}^{(1)} = k) = r\frac{n-1}{n}\sum_{k=0}^{\infty}\mathbb{P}(X_{n+1,r-1}^{(1)} = k) = r\frac{n-1}{n}.$$

On retrouve ainsi que $\mathbb{E}[X_{n,r}^{(1)}] = r - r\dfrac{n-1}{n} = \dfrac{r}{n}$.

7. Pour $r \geq 2$, $k \in \mathbb{N}$, on a

$$\begin{aligned}(r-1-k)(r-k)\mathbb{P}(X_{n,r}^{(1)}=k) &= (r-1-k)r\frac{n-1}{n}\mathbb{P}(X_{n+1,r-1}^{(1)}=k)\\ &= (r-1)r\frac{n-1}{n+1}\mathbb{P}(X_{n+2,r-2}^{(1)}=k).\end{aligned}$$

On en déduit donc que

$$\mathbb{E}[(r-X_{n,r}^{(1)})(r-1-X_{n,r}^{(1)})] = r(r-1)\frac{n-1}{n+1},$$

puis que

$$\mathbb{E}[(X_{n,r}^{(1)})^2] = r(r-1)\frac{n-1}{n+1} - (r^2+\mathbb{E}[X_{n,r}^{(1)}](1-2r)-r) = \frac{2r^2}{n(n+1)} + \frac{r(n-1)}{n(n+1)}.$$

En particulier, on a

$$\lim\ \mathbb{E}[(X_{n,r}^{(1)})^2] = 2\theta^2+\theta.$$

On a également $\mathbb{E}[X^2] = \mathbb{E}[Y^2] - 2\mathbb{E}[Y] + 1 = \mathrm{Var}(Y) + \mathbb{E}[Y]^2 - 2\mathbb{E}[Y] + 1$. Comme $\mathbb{E}[Y] = \dfrac{1}{\rho} = 1+\theta$ et $\mathrm{Var}(Y) = \dfrac{1-\rho}{\rho^2} = \theta(1+\theta)$, il vient

$$\mathbb{E}[X^2] = \theta(1+\theta) + (1+\theta)^2 - 2(1+\theta) + 1 = 2\theta^2+\theta.$$

On vérifie ainsi que $\lim\ \mathbb{E}[(X_{n,r}^{(1)})^2] = \mathbb{E}[X^2]$.

III Quand on augmente le nombre de particules

1. Soit $a = (a_1,\ldots,a_n) \in \mathbb{N}^n$, $b = (b_1,\ldots,b_n) \in \mathbb{N}^n$ avec $\sum_{i=1}^n a_i = r$ et $\sum_{i=1}^n b_i = r+1$. Par construction, on a

$$\mathbb{P}(X_{n,r+1} = b | X_{n,r} = a) = 0$$

si $\sum_{i=1}^n |a_i - b_i| \neq 1$ et sinon, il existe un unique indice $j \in \{1,\ldots,n\}$ tel que $b_j = a_j+1$, et on a

$$\mathbb{P}(X_{n,r+1} = b | X_{n,r} = a) = \frac{a_j+1}{n+r} = \frac{b_j}{n+r}.$$

2. En utilisant la définition des probabilités conditionnelles, il vient

$$\begin{aligned}\mathbb{P}(X_{n,r+1}=b) &= \sum_{a=(a_1,\ldots,a_n)\in\mathbb{N}^n,\ \sum_{i=1}^n a_i = r} \mathbb{P}(X_{n,r+1}=b|X_{n,r}=a)\mathbb{P}(X_{n,r}=a)\\ &= \frac{r!(n-1)!}{(n+r-1)!}\sum_{a=(a_1,\ldots,a_n)\in\mathbb{N}^n,\ \sum_{i=1}^n a_i = r} \mathbb{P}(X_{n,r+1}=b|X_{n,r}=a)\\ &= \frac{r!(n-1)!}{(n+r-1)!}\sum_{j=1}^n \frac{b_j}{n+r}\\ &= \frac{r!(n-1)!}{(n+r-1)!}\frac{r+1}{n+r}\\ &= \frac{(r+1)!(n-1)!}{(n+r)!}.\end{aligned}$$

La loi de $X_{n,r+1}$ est la loi de Bose-Einstein pour $r+1$ particules et n boîtes.

▲

Chapitre XII

Contrôles en fin de cours

XII.1 1999-2000

XII.1.1 Exercices

Exercice XII.1.
Soit Y une variable aléatoire de loi $\Gamma(1, 1/2)$. On rappelle la densité de cette loi :

$$f_Y(y) = \frac{1}{\sqrt{\pi y}} \mathrm{e}^{-y} \mathbf{1}_{\{y>0\}}.$$

On suppose que la loi conditionnelle de X sachant Y est une loi gaussienne $\mathcal{N}\left(0, \frac{1}{2Y}\right)$.

1. Calculer la loi du couple (X, Y).
2. Calculer et reconnaître la loi conditionnelle de Y sachant X.
3. Calculer $\mathbb{E}[Y|X]$.

△

XII.1.2 Le modèle de Hardy-Weinberg

Exercice XII.2.
Le but de cet exercice est l'étude simplifiée de la répartition d'un génotype dans la population humaine.

On considère un gène possédant deux caractères a et A. Le génotype d'un individu est donc soit aa, Aa ou AA. On note 1 le génotype aa ; 2 le génotype Aa et 3 le génotype AA. On s'intéresse à la proportion de la population possédant le génotype $j \in \{1, 2, 3\}$.

I Distribution du génotype : le modèle de Hardy-Weinberg

Le but de cette partie est d'établir que la répartition du génotype de la population est stable à partir de la première génération.

On considère la génération 0 d'une population de grande taille dont les proportions sont les suivantes :

$$\begin{cases} \text{proportion de } aa : & u_1; \\ \text{proportion de } aA : & u_2; \\ \text{proportion de } AA : & u_3. \end{cases}$$

On suppose les mariages aléatoires et la transmission du gène a ou A uniforme. On note M le génotype de la mère, P le génotype du père et E celui de l'enfant.

1. Montrer que $\mathbb{P}(E = aa|M = aA, P = aA) = \frac{1}{4}$, $\mathbb{P}(E = aa|M = aa, P = aA) = \frac{1}{2}$, $\mathbb{P}(E = aa|M = aA, P = aa) = \frac{1}{2}$, et $\mathbb{P}(E = aa|M = aa, P = aa) = 1$.
2. Montrer, en précisant les hypothèses, que $\mathbb{P}(E = aa) = u_1^2 + u_1u_2 + \frac{1}{4}u_2^2 = \left(u_1 + \frac{u_2}{2}\right)^2$.
3. Montrer sans calcul que $\mathbb{P}(E = AA) = \left(u_3 + \frac{u_2}{2}\right)^2$.
4. On pose donc $\theta = u_1 + \dfrac{u_2}{2}$. Montrer, sans calcul, que la répartition du génotype à la première génération est

$$\begin{cases} \text{proportion de } aa : & q_1 = \theta^2; \\ \text{proportion de } aA : & q_2 = 2\theta(1-\theta); \\ \text{proportion de } AA : & q_3 = (1-\theta)^2. \end{cases} \tag{XII.1}$$

5. Calculer la répartition du génotype à la seconde génération. En déduire que la répartition (XII.1) est stationnaire au cours du temps.

II Modèle probabiliste

On suppose que la taille de la population est grande et que la répartition du génotype suit le modèle de Hardy-Weinberg (XII.1). On dispose d'un échantillon de n personnes. On note $X_i \in \{1, 2, 3\}$ le génotype de la i-ème personne. On a donc $\mathbb{P}(X_i = j) = q_j$. On suppose de plus que les variables aléatoires $(X_i; i \in \{1, \dots, n\})$ sont indépendantes. On note

$$N_j = N_j[n] = \sum_{i=1}^{n} \mathbf{1}_{\{X_i=j\}}, \tag{XII.2}$$

le nombre de personnes de l'échantillon possédant le génotype j.

1. Donner la loi $\mathbf{1}_{X_i=j}$. En déduire que la loi de N_j est une binomiale dont on précisera les paramètres.
2. Donner $\mathbb{E}[N_j]$ et $\mathrm{Var}(N_j)$.
3. Déduire des questions précédentes un estimateur sans biais de q_j. Est-il convergent ?
4. Montrer qu'il est asymptotiquement normal et donner sa variance asymptotique.
5. On rappelle que $n_1 + n_2 + n_3 = n$. Montrer que

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \frac{n!}{n_1!n_2!n_3!} q_1^{n_1} q_2^{n_2} q_3^{n_3}. \tag{XII.3}$$

On pourra utiliser que le nombre de partitions d'un ensemble à n éléments en trois sous-ensembles de n_1, n_2 et n_3 éléments est $\dfrac{n!}{n_1!n_2!n_3!}$.

6. Calculer la matrice de covariance du vecteur $(\mathbf{1}_{X_1=1}, \mathbf{1}_{X_1=2})$. En déduire $\mathrm{Cov}(N_1, N_2)$.
7. Donner la limite en loi du couple $\left(\frac{N_1[n]-nq_1}{\sqrt{n}}, \frac{N_2[n]-nq_2}{\sqrt{n}}\right)$ quand $n \to \infty$.

III Estimation de θ à l'aide du génotype

On note P_θ la loi de (N_1, N_2, N_3) définie par l'équation (XII.3) de paramètre $\theta \in]0,1[$.

1. Vérifier que la log vraisemblance $L_n(n_1, n_2, n_3; \theta)$ de l'échantillon de loi P_θ est
$$L_n(n_1, n_2, n_3; \theta) = c + 2n_1 \log\theta + n_2 \log\theta + n_2 \log(1-\theta) + 2n_3 \log(1-\theta),$$
où c est une constante indépendante de θ.
2. Calculer le score de l'échantillon.
3. Calculer la dérivée du score en θ. En déduire que l'information de Fisher de l'échantillon de taille n est
$$I_n(\theta) = \frac{2n}{\theta(1-\theta)}.$$
4. On rappelle que $N_1 + N_2 + N_3 = n$. Montrer que $\hat{\theta}_n = \frac{N_1}{n} + \frac{N_2}{2n}$ est l'estimateur du maximum de vraisemblance de θ.
5. $\hat{\theta}_n$ est-il sans biais ?
6. Est-il efficace ?
7. Montrer que l'estimateur $\hat{\theta}_n$ est convergent. Montrer qu'il est asymptotiquement normal et donner sa variance asymptotique. On pourra soit utiliser directement (XII.2) soit utiliser le résultat de la question **II**.7.

IV Tests asymptotiques sur le modèle de Hardy-Weinberg

On désire savoir si le modèle de Hardy-Weinberg est valide pour certaines maladies génétiques. On teste donc l'hypothèse H_0 : la proportion (q_1, q_2, q_3) des génotypes aa, aA, AA satisfait l'équation (XII.1).

1. Donner l'estimateur du maximum de vraisemblance de (q_1, q_2, q_3) noté $(\hat{q}_1, \hat{q}_2, \hat{q}_3)$.
2. En déduire que $n \sum_{j=1}^{3} \frac{(\hat{q}_j - q_j(\hat{\theta}_n))^2}{\hat{q}_j}$ converge sous H_0. Préciser le mode de convergence et la limite.
3. En déduire un test asymptotique convergent.
4. On étudie la maladie génétique CF. Le génotype aa correspond aux cas pathologiques. Les génotypes aA et AA sont sains. Les valeurs numériques suivantes sont inspirées de valeurs réelles. Nombre de naissances aux USA en 1999 : $n = 3,88$ millions, nombre de nouveau-nés atteints de la maladie CF : $n_1 = 1580$, nombre de nouveau-nés portant le gène a : $n_1 + n_2 = 152480$. Rejetez-vous le modèle au seuil de 5% ?
5. On dispose des données suivantes sur l'hémophilie. Nombre de nouveau-nés atteints d'hémophilie : $n_1 = 388$, nombre de nouveau-nés portant le gène a : $n_1 + n_2 = 1164$. Rejetez vous le modèle au seuil de 5% ? En fait on sait que l'hémophilie concerne essentiellement la population masculine. Commentaire.

△

XII.2 2000-2001

XII.2.1 Exercices

Exercice XII.3.
On désire étudier la répartition des naissances suivant le type du jour de semaine (jours ouvrables ou week-end) et suivant le mode d'accouchement (naturel ou par césarienne). Les données proviennent du "National Vital Statistics Report" et concernent les naissances aux USA en 1997. (On a omis 35 240 naissances pour lesquelles le mode d'accouchement n'a pas été retranscrit.)

Naissances	Naturelles	César.	Total
J.O.	2331536	663540	2995076
W.E.	715085	135493	850578
Total	3046621	799033	3845654

Naissances	Naturelles	César.	Total
J.O.	60.6 %	17.3 %	77.9%
W.E.	18.6 %	3.5 %	22.1%
Total	79.2 %	20.8 %	100.0%

On note $p_{J,N}$ la probabilité qu'un bébé naisse un jour ouvrable et sans césarienne, $p_{W,N}$ la probabilité qu'un bébé naisse un week-end et sans césarienne, $p_{J,C}$ la probabilité qu'un bébé naisse un jour ouvrable et par césarienne, $p_{W,C}$ la probabilité qu'un bébé naisse un week-end et par césarienne.

1. Rappeler l'estimateur du maximum de vraisemblance de $p = (p_{J,N}, p_{W,N}, p_{J,C}, p_{W,C})$.
2. À l'aide d'un test du χ^2, pouvez-vous accepter ou rejeter l'hypothèse d'indépendance entre le type du jour de naissance (jour ouvrable ou week-end) et le mode d'accouchement (naturel ou césarienne) ?
3. On désire savoir s'il existe une évolution significative dans la répartition des naissances par rapport à 1996. À l'aide d'un test du χ^2, pouvez-vous accepter ou rejeter l'hypothèse $p = p_0$, où p_0 correspond aux données de 1996 ? On donne les valeurs suivantes pour p_0 :

Naissances	Naturelles	Césariennes
J.O.	60.5 %	17.0 %
W.E.	18.9 %	3.6 %

△

XII.2.2 Estimation de la taille d'une population

Exercice XII.4.
On désire estimer le nombre inconnu N de chevreuils vivant dans une forêt. Dans une première étape, on capture à l'aide de pièges n_0 chevreuils que l'on marque à l'aide de colliers, puis on les relâche. Dans une deuxième étape, des observateurs se rendent en plusieurs points de la forêt et comptent le nombre de chevreuils qu'ils voient. Un chevreuil peut donc être compté plusieurs fois. Parmi les n chevreuils comptabilisés, m portent un collier. On suppose qu'entre la première et la deuxième étape, la population de chevreuils n'a pas évolué, et que les chevreuils avec ou sans collier ont autant de chance d'être vus. Enfin, on suppose que les chevreuils ne peuvent pas perdre les colliers. Le but de ce problème est de construire un estimateur de N.

I Modélisation

On note $X_i = 1$ si le i$^{\text{ème}}$ chevreuil vu porte un collier et $X_i = 0$ sinon. On dispose donc de l'échantillon $X_1, \ldots, X_n$.

1. Au vu des hypothèses du modèle, quelle est la loi des variables aléatoires $X_1, \ldots, X_n$?
2. Vérifier que la densité de la loi de X_1 est $p(x_1; p) = p^{x_1}(1-p)^{1-x_1}$, avec $x_1 \in \{0, 1\}$. On a $p = \dfrac{n_0}{N} \in]0, 1[$. Comme n_0 est connu, chercher un estimateur de N revient à chercher un estimateur de $1/p$.
3. Quelle est la densité de l'échantillon ?
4. Montrer que $S_n = \sum_{i=1}^{n} X_i$ est une statistique exhaustive. Quelle est la loi de S_n ? On admet que la statistique S_n est totale.
5. Soit h une fonction définie sur $\{0, \ldots, n\}$. Vérifier que $\lim\limits_{p \to 0} \mathbb{E}_p[h(S_n)] = h(0)$. En déduire qu'il n'existe pas d'estimateur de $1/p$, fonction de S_n, qui soit sans biais.
6. Montrer alors qu'il n'existe pas d'estimateur de $1/p$ sans biais.

II Estimation asymptotique

1. Montrer que $\left(\dfrac{n+1}{S_n+1}, n \in \mathbb{N}^*\right)$ est une suite d'estimateurs convergents de $1/p$.
2. Quel est le biais de l'estimateur $\dfrac{n+1}{S_n+1}$ défini par $\mathbb{E}\left[\dfrac{n+1}{S_n+1}\right] - \dfrac{1}{p}$?
3. Déterminer le mode de convergence et la limite de la suite $\left(\sqrt{n}\left(\dfrac{S_n}{n} - p\right), n \in \mathbb{N}^*\right)$.
4. Vérifier que la suite $\left(\sqrt{n}\left(\dfrac{S_n+1}{n+1} - \dfrac{S_n}{n}\right), n \in \mathbb{N}^*\right)$ converge presque sûrement vers 0. En déduire que $\left(\dfrac{S_n+1}{n+1}, n \in \mathbb{N}^*\right)$ est une suite d'estimateurs convergente de p asymptotiquement normale.
5. En déduire que la suite d'estimateurs de substitution $\left(\dfrac{n+1}{S_n+1}, n \in \mathbb{N}^*\right)$ est asymptotiquement normale de variance asymptotique $s^2 = \dfrac{(1-p)}{p^3}$.
6. Calculer le score et l'information de Fisher de l'échantillon de taille 1. La suite d'estimateurs $\left(\dfrac{n+1}{S_n+1}, n \in \mathbb{N}^*\right)$ est-elle asymptotiquement efficace ?

III Intervalle de confiance

1. Donner un estimateur convergent de s^2. En déduire un intervalle de confiance asymptotique à 95% de $1/p$.
2. Les données numériques suivantes proviennent d'un site d'études au Danemark dans les années 1960 (cf le livre de Seber : The estimation of animal abundance, page 110). On a $n_0 = 74$, $n = 462$ et $m = 340$. Donner une estimation de $1/p$ puis de N.
3. Donner un intervalle de confiance asymptotique de niveau 95% de $1/p$ puis de N.

IV Tests

1. On considère l'hypothèse nulle $H_0 = \{p = p_0\}$, où $p_0 = n_0/N_0$, et l'hypothèse alternative $H_1 = \{p = p_1\}$, où $p_1 = n_0/N_1$. On suppose $N_1 > N_0$. Calculer le rapport de vraisemblance $Z(x)$, où $x \in \{0,1\}^n$. Déterminer la statistique de test.
2. Décrire le test UPP de niveau α à l'aide de S_n.
3. Ce test est-il UPP pour tester $H_0 = \{N = N_0\}$ contre $H_1 = \{N > N_0\}$?
4. On reprend les données numériques du paragraphe précédent. On considère l'hypothèse nulle $H_0 = \{N = 96\}$ contre son alternative $H_1 = \{N \geq 97\}$. Rejetez-vous H_0 au niveau $\alpha = 5\%$?
   On utilisera les valeurs suivantes ($p_0 = 74/96$) :

c	338	339	340	341	342
$\mathbb{P}_{p_0}(S_n \leq c)$	0.0270918	0.0344991	0.0435125	0.0543594	0.0672678

△

XII.3 2001-2002

XII.3.1 Comparaison de traitements

Exercice XII.5.
On considère une population de souris en présence de produits toxiques. Le temps d'apparition, en jour, des effets dus aux produits toxiques est modélisé par une variable aléatoire T qui suit une loi de type Weibull :

$$\bar{F}_\theta(t) = \mathbb{P}_\theta(T > t) = \mathrm{e}^{-\theta(t-w)^k} \quad \text{pour } t \geq w, \text{ et } \quad \bar{F}_\theta(t) = 1 \quad \text{pour } t < w.$$

Le paramètre $w \geq 0$ correspond à la période de latence des produits toxiques. Le paramètre $k > 0$ est lié à l'évolution du taux de mortalité des souris en fonction de leur âge. Enfin le paramètre $\theta > 0$ correspond plus directement à la sensibilité des souris aux produits toxiques. En particulier, on pense qu'un pré-traitement de la population des souris permet de modifier le paramètre θ. On considère que les paramètres w et k sont connus, et que seul le paramètre θ est inconnu.

I L'estimateur du maximum de vraisemblance de θ

1. Calculer la densité f_θ de la loi de T.
2. Soit $T_1, \ldots, T_n$ un échantillon de taille de n de variables aléatoires indépendantes de même loi que T. Donner la densité $p_n^*(t;\theta)$ de l'échantillon où $t = (t_1, \ldots, t_n) \in]w, +\infty[^n$.
3. Calculer la log-vraisemblance de l'échantillon, puis $\hat{\theta}_n^*$, l'estimateur du maximum de vraisemblance de θ.
4. Calculer l'information de Fisher $I^*(\theta)$ de l'échantillon de taille 1 associée à θ.
5. Calculer $\mathbb{E}_\theta[(T-w)^k]$ et $\mathbb{E}_\theta[(T-w)^{2k}]$. On pourra utiliser la fonction Γ définie par

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\,\mathrm{e}^{-x}\ dx, \quad \alpha > 0.$$

   On rappelle que si α est entier, alors $\Gamma(\alpha) = (\alpha-1)!$.

6. En déduire que l'estimateur du maximum de vraisemblance est un estimateur convergent et asymptotiquement normal de θ.
7. Calculer la variance asymptotique de $\hat{\theta}_n^*$. Et vérifier que l'estimateur du maximum de vraisemblance est asymptotiquement efficace.
8. On a les $n = 17$ observations suivantes :

   143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 246, 265, 304.

   On suppose $w = 100$ et $k = 3$. Donner une estimation de θ et un intervalle de confiance asymptotique à 95%. On donne la valeur de $\sum_{i=1}^n (t_i - 100)^3 = 33\,175\,533$.

II Données censurées

En fait parmi les souris, certaines sont retirées de la population observée pour diverses raisons (maladie, blessure, analyse) qui ne sont pas liées à l'apparition des effets dus aux produits toxiques. Pour une souris donnée, on n'observe pas T le temps d'apparition des premiers effets mais $R = T \wedge S$ (on utilise la notation $t \wedge s = \min(t, s)$), où S est le temps aléatoire où la souris est retirée de la population étudiée. On observe également la variable X définie par $X = 0$ si $T > S$ (la souris est retirée de la population étudiée avant que les effets dus aux produits toxiques soient observés sur cette souris) et $X = 1$ si $T \leq S$ (les effets apparaissent sur la souris avant que la souris soit retirée de la population). On suppose que les variables T et S sont indépendantes. On suppose aussi que S est une variable aléatoire continue de densité g, et $g(x) > 0$ si $x > 0$. On considère la fonction

$$\bar{G}(s) = \mathbb{P}(S > s) = \int_s^\infty g(u)\, du, \quad s \in \mathbb{R}.$$

Les fonctions $\bar{G}$ et g sont inconnues et on ne cherche pas à les estimer. En revanche on désire toujours estimer le paramètre inconnu θ.

1. Quelle est la loi de X ? On note $p = \mathbb{P}_\theta(X = 1)$.
2. Calculer $\mathbb{P}_\theta(R > r, X = x)$, où $x \in \{0, 1\}$ et $r \in \mathbb{R}$. La densité des données censurées $p(r, x; \theta)$ est la densité de la loi de (R, X). Elle est définie par

   $$\text{pour tout } r \in \mathbb{R}, \quad \int_r^\infty p(u, x; \theta)\, du = \mathbb{P}_\theta(R > r, X = x).$$

   Vérifier que $p(r, x; \theta) = \theta^x \, \mathrm{e}^{[-\theta(r-w)_+^k]}\, c(r, x)$, où $z_+ = z\mathbf{1}_{\{z>0\}}$ désigne la partie positive de z et où la fonction c est indépendante de θ.
3. Soit $(R_1, X_1), \ldots, (R_n, X_n)$ un échantillon de taille n de variables aléatoires indépendantes de même loi que (R, X). Cet échantillon modélise les données censurées. On note $N_n = \sum_{i=1}^n X_i$. Que représente N_n ?
4. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de θ. Quelle est la différence avec l'estimateur $\hat{\theta}_n^*$ défini précédemment ?
5. Montrer que $\hat{p}_n = N_n/n$ est un estimateur sans biais convergent de p.
6. Déterminer pour les données censurées $I(\theta)$, l'information de Fisher de θ pour l'échantillon de taille 1.

7. On admet dorénavant que l'estimateur du maximum de vraisemblance est un estimateur convergent asymptotiquement efficace (voir la partie IV). Déduire des questions précédentes un estimateur convergent de $I(\theta)$.
8. En fait dans la partie I on n'a pas tenu compte des données censurées correspondant à $X_i = 0$, c'est-à-dire à l'observation de S_i et non de T_i. Il s'agit des 2 valeurs supplémentaires suivantes : 216 et 244. On suppose toujours $w = 100$ et $k = 3$. Donner une estimation de θ et un intervalle de confiance asymptotique à 95%. On donne la valeur de $\sum_{i=n+1}^{n+2}(s_i - 100)^3 = 4\ 546\ 880$. Comparer le résultat obtenu à celui de la partie précédente (centre et largeur de l'intervalle de confiance)
9. Vérifier que dans ce modèle, si l'on tient compte de souris supplémentaires $i \in \{n + 1, \ldots, n + m\}$ qui sont retirées avant l'instant w, alors la densité de l'échantillon est multipliée par une quantité qui est indépendante de θ. Vérifier que $\hat{\theta}_{n+m} = \hat{\theta}_n$ et que l'estimation de l'information de Fisher de l'échantillon de taille n, $nI(\theta)$ et de l'échantillon de taille $n + m$, $(n + m)I(\theta)$ sont égales. En particulier l'intervalle de confiance asymptotique de θ reste inchangé.

III Comparaison de traitements

On désire comparer l'influence de deux pré-traitements A et B effectués avant l'exposition des souris aux produits toxiques. On note θ^A le paramètre de la loi de T correspondant au pré-traitement A et θ^B celui correspondant au pré-traitement B. On désire donc établir une procédure pour comparer θ^A et θ^B. On note $(R_1^A, X_1^A), \ldots, (R_{n^A}^A, X_{n^A}^A)$ les données censurées de la population de n^A souris ayant subi le pré-traitement A et $(R_1^B, X_1^B), \ldots, (R_{n^B}^B, X_{n^B}^B)$ les données censurées de la population de n^B souris ayant subi le pré-traitement B. On suppose de plus que les variables (R_i^j, X_i^j), où $1 \leq i \leq n$ et $j \in \{A, B\}$, sont indépendantes. On suppose dans un premier temps que les deux populations ont le même nombre d'individus $n = n^A = n^B$. On note $Z_i = (R_i^A, X_i^A, R_i^B, X_i^B)$ pour $i \in \{1, \ldots, n\}$.

1. Donner la densité de Z_1 et de l'échantillon $Z_1, \ldots, Z_n$.
2. Donner l'estimateur du maximum de vraisemblance $(\hat{\theta}_n^A, \hat{\theta}_n^B)$ du paramètre (θ^A, θ^B).
3. On note
$$p^A = \mathbb{P}_{(\theta^A, \theta^B)}(X_1^A = 1) \quad \text{et} \quad p^B = \mathbb{P}_{(\theta^A, \theta^B)}(X_1^B = 1).$$
Donner l'information de Fisher $I(\theta^A, \theta^B)$ pour l'échantillon de taille 1. Vérifier que la matrice $I(\theta^A, \theta^B)$ est diagonale.
4. Sous l'hypothèse $H_0 = \{\theta^A = \theta^B\}$ donner l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de $\theta = \theta^A = \theta^B$.
5. Donner deux tests asymptotiques convergents pour tester l'hypothèse nulle H_0 contre l'hypothèse alternative $H_1 = \{\theta^A \neq \theta^B\}$ construits à l'aide du test de Hausman.
6. Donner un estimateur convergent de p^j construit à partir de $N_n^j = \sum_{i=1}^n X_i^j$, pour $j \in \{A, B\}$. En déduire un estimateur, $\hat{I}_n$, convergent de la fonction $I : (\theta^A, \theta^B) \mapsto I(\theta^A, \theta^B)$.
7. Vérifier que si l'on rajoute dans chaque population m_A et m_B souris virtuelles qui auraient été retirées avant l'instant w, alors cela multiplie la densité de l'échantillon par une constante indépendante du paramètre (θ^A, θ^B). Vérifier que les estimateurs de

θ^A, θ^B, θ et de la fonction I restent inchangés. On admet que l'on peut remplacer dans les tests précédents la fonction I par la fonction $\hat{I}_n$ et que l'on peut ajouter des souris virtuelles sans changer les propriétés des tests (convergence, région critique et niveau asymptotique).

8. Dans le cas où la taille n^A de la population A est plus petite que la taille n^B de la population B, on complète la population A par $n^B - n^A$ souris virtuelles qui auraient été retirées avant l'instant w. Les données de l'échantillon A correspondent à celles de la partie précédente. Les données de l'échantillon B sont les suivantes. La population comporte $n^B = 21$ souris dont
   – 19 souris pour lesquelles les effets des produits toxiques ont été observés aux instants t_i :

     142, 156, 163, 198, 205, 232, 232, 233, 233, 233, 233,
     239, 240, 261, 280, 280, 296, 296, 323,

     avec $\sum_{i=1}^{19}(t_i - 100)^3 = 64\ 024\ 591$,
   – 2 souris qui ont été retirées de l'expérience aux instants s_i : 204 et 344, avec $\sum_{i=1}^{2}(s_i - 100)^3 = 15\ 651\ 648$.

   Calculer $\hat{\theta}_n^A$, $\hat{\theta}_n^B$, et $\hat{\theta}_n$. Les pré-traitements A et B sont-ils différents ?
9. Donner des intervalles de confiance $J^A_{n^A}$ pour θ^A et $J^B_{n^B}$ pour θ^B tels que la confiance asymptotique sur les deux intervalles soit d'au moins 95% (i.e. $\mathbb{P}_{(\theta^A,\theta^B)}(\theta^A \in J^A, \theta^B \in J^B) \geq 95\%$). Retrouvez-vous la conclusion de la question précédente ?

IV Propriétés asymptotiques de l'estimateur $\hat{\theta}_n$ (Facultatif)

1. Exprimer p comme une intégrale fonction de f_θ et $\bar{G}$.
2. Calculer $\mathbb{P}_\theta(R > r)$ en fonction de $\bar{F}_\theta$ et $\bar{G}$. En déduire la densité de la loi de R.
3. Montrer à l'aide d'une intégration par partie, que $\mathbb{E}_\theta[(R-w)_+^k] = p/\theta$.
4. Montrer directement que $\hat{\theta}_n$ est un estimateur convergent de θ.
5. On pose $\beta = \mathbb{E}_\theta[\mathbf{1}_{\{X=1\}}(R-w)_+^k]$. Vérifier à l'aide d'une intégration par partie que $\mathbb{E}_\theta[(R-w)_+^{2k}] = 2\beta/\theta$.
6. Donner la matrice de covariance du couple $((R-w)_+^k, X)$.
7. Montrer que l'estimateur $\hat{\theta}_n$ est asymptotiquement normal et donner sa variance asymptotique.
8. Vérifier que l'estimateur $\hat{\theta}_n$ est asymptotiquement efficace.

Les données numériques proviennent de l'article de M. Pike, "A method of analysis of a certain class of experiments in carcinogenesis" (*Biometrics*, Vol. 22, p. 142-161 (1966)).

Remarque sur la loi de T. Dans l'exemple considéré, on suppose que chaque cellule de l'organe sensible aux produits toxiques se comporte de manière indépendante. Le temps T apparaît donc comme le minimum des temps $\tilde{T}_k$ d'apparition des effets dans la cellule k. La loi de $\tilde{T}_k$ est inconnue a priori. Mais la théorie des valeurs extrêmes permet de caractériser

la loi limite de $\min_{1\leq k\leq K} \tilde{T}_k$, éventuellement renormalisée, quand $K \to \infty$. En particulier les lois de type Weibull apparaissent comme loi limite.

△

XII.4 2002-2003

XII.4.1 Ensemencement des nuages.

Exercice XII.6.
Il existe deux types de nuages qui donnent lieu à des précipitations : les nuages chauds et les nuages froids. Ces derniers possèdent une température maximale de l'ordre de -10°C à -25°C. Ils sont composés de cristaux de glace et de gouttelettes d'eau. Ces gouttelettes d'eau subsistent alors que la température ambiante est inférieure à la température de fusion. On parle d'eau surfondue. Leur état est instable. De fait, quand une particule de glace rencontre une gouttelette d'eau, elles s'aggrègent pour ne former qu'une seule particule de glace. Les particules de glace, plus lourdes que les gouttelettes, tombent sous l'action de la gravité. Enfin si les températures des couches d'air inférieures sont suffisamment élevées, les particules de glace fondent au cours de leur chute formant ainsi de la pluie.

En l'absence d'un nombre suffisant de cristaux de glace pour initier le phénomène décrit ci-dessus, on peut injecter dans le nuage froid des particules qui ont une structure cristalline proche de la glace, par exemple de l'iodure d'argent (environ 100 à 1000 grammes par nuage). Autour de ces particules, on observe la formation de cristaux de glace, ce qui permet, on l'espère, de déclencher ou d'augmenter les précipitations. Il s'agit de l'ensemencement des nuages. Signalons que cette méthode est également utilisée pour limiter le risque de grêle.

Il est évident que la possibilité de modifier ainsi les précipitations présente un grand intérêt pour l'agriculture. De nombreuses études ont été et sont encore consacrées à l'étude de l'efficacité de l'ensemencement des nuages dans divers pays. L'étude de cette efficacité est cruciale et délicate. Le débat est encore d'actualité.

L'objectif du problème qui suit est d'établir à l'aide des données concernant l'ensemencement des nuages en Floride (1975)[1] si l'ensemencement par iodure d'argent est efficace ou non.

On dispose des données des volumes de pluie déversées en 10^7 m^3 (cf. les deux tableaux ci-dessous) concernant 23 jours similaires dont $m = 11$ jours avec ensemencement correspondant aux réalisations des variables aléatoires $X_1, \ldots, X_m$ et $n = 12$ jours sans ensemencement correspondant aux réalisations des variables aléatoires $Y_1, \ldots, Y_n$. On suppose que les variables aléatoires $X_1, \ldots, X_m$ ont même loi, que les variables aléatoires $Y_1, \ldots, Y_n$ ont même loi, et que les variables $X_1, \ldots, X_m, Y_1, \ldots, Y_n$ sont toutes indépendantes.

i	1	2	3	4	5	6	7	8	9	10	11
X_i	7.45	4.70	7.30	4.05	4.46	9.70	15.10	8.51	8.13	2.20	2.16

TAB. XII.1 – Volume de pluie en 10^7 m^3 déversée avec ensemencement

[1] William L. Woodley, Joanne Simpson, Ronald Biondini, Joyce Berkeley, *Rainfall results, 1970-1975 : Florida Area Cumuls Experiment*, Science, Vol 195, pp. 735-742, February 1977.

j	1	2	3	4	5	6	7	8	9	10	11	12
Y_j	15.53	10.39	4.50	3.44	5.70	8.24	6.73	6.21	7.58	4.17	1.09	3.50

TAB. XII.2 – Volume de pluie en 10^7 m^3 déversée sans ensemencement

On considérera divers modèles pour la quantité d'eau déversée par les nuages. Et l'on construira des statistiques de tests et des régions critiques pour tester l'hypothèse nulle

$$H_0 = \{\text{l'ensemencement n'accroît pas de manière sensible la quantité d'eau déversée}\},$$

contre l'hypothèse alternative

$$H_1 = \{\text{l'ensemencement accroît de manière sensible la quantité d'eau déversée}\}.$$

I Modèle gaussien à variance connue

On suppose que Y_j suit une loi gaussienne $\mathcal{N}(\nu, \sigma_0^2)$, où $\nu \in \mathbb{R}$ est inconnu et $\sigma_0^2 = 13$.

1. Calculer $\hat{\nu}_n$ l'estimateur du maximum de vraisemblance de ν. Est-il biaisé ?
2. Donner la loi de $\hat{\nu}_n$. En déduire un intervalle de confiance exact de ν de niveau $1-\alpha$. Application numérique avec $\alpha = 5\%$.
3. Montrer directement que $\hat{\nu}_n$ est un estimateur convergent de ν.

On suppose que X_i suit une loi gaussienne $\mathcal{N}(\mu, \sigma_0^2)$, où μ est inconnu. L'hypothèse nulle s'écrit dans ce modèle $H_0 = \{\mu = \nu\}$ et l'hypothèse alternative $H_1 = \{\mu > \nu\}$.

On considère le modèle complet formé des deux échantillons indépendants $X_1, \ldots, X_m$ et $Y_1, \ldots, Y_n$.

4. Écrire la vraisemblance et la log-vraisemblance du modèle complet.
5. Calculer $\hat{\mu}_m$, l'estimateur du maximum de vraisemblance de μ. Vérifier que l'estimateur du maximum de vraisemblance de ν est bien $\hat{\nu}_n$.
6. Donner la loi de $(\hat{\mu}_m, \hat{\nu}_n)$.
7. On considère la statistique de test

$$T_{m,n}^{(1)} = \sqrt{\frac{mn}{m+n}}\frac{\hat{\mu}_m - \hat{\nu}_n}{\sigma_0}.$$

   Donner la loi de $T_{m,n}^{(1)}$. En particulier, quelle est la loi de $T_{m,n}^{(1)}$ sous H_0 ?
8. Montrer que pour tout $\mu > \nu$, presque sûrement, on a

$$\lim_{\min(m,n)\to\infty} T_{m,n}^{(1)} = +\infty.$$

9. Déduire des questions précédentes un test convergent de niveau α. Déterminer la région critique exacte.
10. On choisit $\alpha = 5\%$. Rejetez vous l'hypothèse nulle ? On appelle p-valeur, la plus grande valeur p telle que pour tout $\alpha < p$, alors on accepte H_0 au niveau α. C'est aussi la plus petite valeur p telle que pour tout $\alpha > p$, alors on rejette H_0 au niveau α. Calculer p_1 la p-valeur du modèle.

II Modèle non paramétrique de décalage

On note F la fonction de répartition de X_i et G la fonction de répartition de Y_j. On suppose que $F(x+\rho) = G(x)$, où ρ est le paramètre de décalage inconnu. En particulier, X_i a même loi que $Y_j + \rho$. On suppose de plus que X_i (et donc Y_j) possède une densité f. L'hypothèse nulle dans ce modèle est donc $H_0 = \{\rho = 0\}$ et l'hypothèse alternative $H_1 = \{\rho > 0\}$. On considère la statistique de Mann et Whithney :

$$U_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{1}_{\{Y_j \leq X_i\}}.$$

On pose $p = \mathbb{E}[\mathbf{1}_{\{Y_j \leq X_i\}}]$, et il vient $\mathbb{E}[U_{m,n}] = mnp$. On rappelle que

$$\mathrm{Var}(U_{m,n}) = mnp(1-p) + n(n-1)m\alpha' + m(m-1)n\beta',$$

avec $\alpha' \geq 0$ et $\beta' \geq 0$. De plus si $p \notin \{0,1\}$ (cas non dégénéré), alors au moins l'un des deux termes α' ou β' est strictement positif. On suppose $p \in]0,1[$. On pose

$$Z_{m,n} = \frac{U_{m,n} - mnp}{\sqrt{\mathrm{Var}(U_{m,n})}}.$$

On rappelle que si la suite d'indices $((m_k, n_k), k \geq 1)$ est telle que

$$m_k \xrightarrow[k\to\infty]{} \infty, \quad n_k \xrightarrow[k\to\infty]{} \infty, \quad \text{et} \quad \frac{m_k}{m_k + n_k} \xrightarrow[k\to\infty]{} \theta \in]0,1[,$$

alors la suite $(Z_{m_k,n_k}, k \geq 1)$ converge en loi vers la loi gaussienne centrée réduite $\mathcal{N}(0,1)$.

On rapelle que sous H_0 (i.e. X_i et Y_j ont même loi), on a

$$p = 1/2 \quad \text{et} \quad \mathrm{Var}(U_{m,n}) = \frac{mn(m+n+1)}{12}.$$

On admet que sous H_1, on a $p > 1/2$ (Facultatif : le démontrer). Le cas $p = 1$ étant dégénéré, on supposera donc que sous H_1, on a $p \in]1/2, 1[$.

On considère la statistique de test

$$T^{(2)}_{m,n} = \frac{U_{m,n} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}.$$

1. Montrer que
$$\{T^{(2)}_{m,n} > a\} = \{Z_{m,n} > b_{m,n}\},$$
où l'on déterminera $b_{m,n}$. Vérifier que sous H_1, on a $\lim_{k\to\infty} b_{m_k,n_k} = -\infty$.
2. En déduire, que sous H_1, on a pour tout $a > 0$, $\lim_{k\to\infty} \mathbb{P}(T^{(2)}_{m_k,n_k} > a) = 1$.
3. Déduire des questions précédentes un test asymptotique convergent de niveau α.
4. On choisit $\alpha = 5\%$. Rejetez vous l'hypothèse nulle ? Calculer p_2 la p-valeur du modèle.

III Modèle non paramétrique général

On note F la fonction de répartition de X_i et G la fonction de répartition de Y_j. On suppose que F et G sont des fonctions continues. On désire tester l'hypothèse nulle $H_0 = \{F = G\}$ contre son alternative $H_1 = \{F \neq G\}$. On considère la statistique de test

$$T_{m,n}^{(3)} = \sqrt{\frac{mn}{m+n}} \sup_{x \in \mathbb{R}} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{X_i \leq x\}} - \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{Y_j \leq x\}} \right|.$$

1. Donner un test asymptotique convergent de niveau α construit à partir de $T_{m,n}^{(3)}$.
2. On choisit $\alpha = 5\%$. On observe la réalisation de $T_{m,n}^{(3)}$: $t_{m,n}^{(3)} = 0.5082$. Rejetez vous l'hypothèse nulle ? Calculer p_3 la p-valeur du modèle. On donne quelques valeurs de la fonction de répartition K de la loi limite de $T_{m,n}^{(3)}$ quand $\min(m,n) \to \infty$ sous H_0 :

x	0.519	0.571	0.827	1.223	1.358
$K(x)$	0.05	0.10	0.50	0.90	0.95

IV Conclusion

L'expérience d'ensemencement des nuages pratiquée en Floride est-elle concluante ?

△

XII.5 2003-2004

XII.5.1 Comparaison de densité osseuse.

Exercice XII.7.
Des médecins[2] de l'université de Caroline du Nord ont étudié l'incidence de l'activité physique sur les fractures osseuses chez les femmes agées de 55 à 75 ans en comparant la densité osseuse de deux groupes de femmes. Un groupe est constitué de femmes physiquement actives (groupe 1 de $n = 25$ personnes), et l'autre de femmes sédentaires (groupe 2 de $m = 31$ personnes). Les mesures pour chaque groupe sont présentées ci-dessous.

Groupe 1

213	207	208	209	232
202	217	184	203	216
219	211	214	204	210
207	212	212	245	226
227	230	214	210	221

Groupe 2

201	173	208	216	204	210
205	217	199	185	201	211
201	187	209	162	207	213
208	202	209	192	202	219
205	214	203	176	206	194
213					

On souhaite répondre à la question suivante : La densité osseuse chez les femmes du groupe 1 est-elle significativement supérieure à celle des femmes du groupe 2 ?

[2] M. Paulsson, N. Hironori, *Age- and gender-related changes in the cellularity of human bone.* Journal of Orthopedic Research 1985 ; Vol. 3 (2) ; pp. 779-784.

I Le modèle

On note X_i (resp. Y_i) la densité osseuse de la i-ème femme du groupe 1 (resp. du groupe 2) et x_i (resp. y_i) l'observation correspondant à une réalisation de X_i (resp. Y_i). On suppose que les variables $(X_i, i \geq 1)$ sont indépendantes de loi gaussienne de moyenne μ_1 et de variance σ_1^2, que les variables $(Y_j, j \geq 1)$ sont indépendantes de loi gaussienne de moyenne μ_2 et de variance σ_2^2, et enfin que les groupes sont indépendants i.e. les variables $(X_i, i \geq 1)$ et $(Y_j, j \geq 1)$ sont indépendantes. On note n la taille du groupe 1 et m celle du groupe 2. On note $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}_+^*$ le paramètre du modèle.

1. Décrire simplement, pour ce modèle, les hypothèses nulle et alternative correspondant à la question posée.

On pose

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad V_n = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2\right),$$

et

$$\bar{Y}_m = \frac{1}{m}\sum_{j=1}^{m} Y_j, \quad W_m = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2 = \frac{m}{m-1}\left(\frac{1}{m}\sum_{j=1}^{m} Y_j^2 - \bar{Y}_m^2\right).$$

2. Donner la loi de $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$.
3. Rappeler (sans démonstration) la loi de $\left(\bar{X}_n, \frac{n-1}{\sigma_1^2}V_n\right)$.
4. Déduire des deux questions précédentes la loi du couple $\left(\frac{n-1}{\sigma_1^2}V_n, \frac{m-1}{\sigma_2^2}W_m\right)$. Calculer la loi de
$$\frac{n-1}{\sigma_1^2}V_n + \frac{m-1}{\sigma_2^2}W_m.$$
5. Rappeler $\mathbb{E}_\theta[\bar{X}_n]$. En déduire un estimateur sans biais de μ_1, puis un estimateur sans biais de μ_2. Ces estimateurs sont-ils convergents ?
6. Rappeler $\mathbb{E}_\theta[V_n]$. En déduire un estimateur sans biais de σ_1^2, puis un estimateur sans biais de σ_2^2. Ces estimateurs sont-ils convergents ?

On donne les valeurs numériques (à 10^{-3} près) correspondant aux observations :

$$n = 25, \quad \bar{x}_n = 214.120, \quad v_n = 142.277, \tag{XII.4}$$
$$m = 31, \quad \bar{y}_m = 201.677, \quad w_m = 175.959. \tag{XII.5}$$

II Simplification du modèle

Afin de simplifier le problème, on désire savoir si les variances σ_1^2 et σ_2^2 sont significativement différentes. Pour cela, on construit dans cette partie un test associé à l'hypothèse nulle $H_0 = \{\sigma_1 = \sigma_2, \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}\}$ et l'hypothèse alternative $H_1 = \{\sigma_1 \neq \sigma_2, \mu_1 \in \mathbb{R}, \mu_2 \in \mathbb{R}\}$.

On considère la statistique de test

$$Z_{n,m} = \frac{V_n}{W_m}.$$

1. Vérifier que la loi de $Z_{n,m}$ sous H_0 est une loi de Fisher-Snedecor dont on précisera les paramètres.
2. Quelles sont les limites de $(Z_{n,m}, n \geq 2, m \geq 2)$ sous H_0 et sous H_1 quand $\min(n,m) \to \infty$?

Soit α_1, $\alpha_2 \in]0, 1/2[$, et a_{n,m,α_1} (resp. b_{n,m,α_2}) le quantile d'ordre α_1 (resp. $1-\alpha_2$) de $F_{n,m}$ de loi de Fisher-Snedecor de paramètre (n,m) i.e.

$$\mathbb{P}(F_{n,m} \leq a_{n,m,\alpha_1}) = \alpha_1 \quad (\text{resp. } \mathbb{P}(F_{n,m} \geq b_{n,m,\alpha_2}) = \alpha_2).$$

On admet les convergences suivantes

$$\lim_{\min(n,m)\to\infty} a_{n,m,\alpha_1} = \lim_{\min(n,m)\to\infty} b_{n,m,\alpha_2} = 1. \tag{XII.6}$$

3. Montrer que le test pur de région critique $\{Z_{n,m} \notin]a_{n-1,m-1,\alpha_1}, b_{n-1,m-1,\alpha_2}[\}$ est un test convergent, quand $\min(n,m) \to \infty$, pour tester H_0 contre H_1. Déterminer son niveau, et vérifier qu'il ne dépend pas de n et m.
4. On choisit usuellement $\alpha_1 = \alpha_2$. Déterminer, en utilisant les tables en fin de problème, la région critique de niveau $\alpha = 5\%$. Conclure en utilisant les valeurs numériques du paragraphe précédent.
5. Calculer la p-valeur du test (la p-valeur est l'infimum des valeurs de α qui permettent de rejeter H_0, c'est aussi le supremum des valeurs de α qui permettent d'accepter H_0). Affiner la conclusion.
6. (Facultatif) Établir les convergences (XII.6).

III Comparaison de moyenne

On suppose dorénavant que $\sigma_1 = \sigma_2$, et on note σ la valeur commune (inconnue). On note $H_0 = \{\mu_1 = \mu_2, \sigma > 0\}$ et $H_1 = \{\mu_1 > \mu_2, \sigma > 0\}$. On pose

$$S_{n,m} = \frac{(n-1)V_n + (m-1)W_m}{n+m-2},$$

et on considère la statistique de test

$$T_{n,m} = \sqrt{\frac{nm}{n+m}} \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{S_{n,m}}}.$$

1. Déduire de la partie I la loi de $\dfrac{n+m-2}{\sigma^2} S_{n,m}$, et la limite de la suite $(S_{n,m}, n \geq 2, m \geq 2)$ quand $\min(n,m) \to \infty$.
2. Déduire de la partie I la loi de $\dfrac{1}{\sigma}\sqrt{\dfrac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m)$.
3. Vérifier que sous H_0, la loi de $T_{n,m}$ est une loi de Student dont on précisera les paramètres.
4. Quelle est la limite sous H_1 de la suite $(\bar{X}_n - \bar{Y}_m, n \geq 1, m \geq 1)$ quand $\min(n,m) \to \infty$.
5. En déduire sous H_1 la valeur de $\lim\limits_{\min(n,m)\to\infty} T_{n,m}$.

6. Construire un test pur convergent quand $\min(n,m) \to \infty$ de niveau α pour tester H_0 contre H_1.
7. On choisit $\alpha = 5\%$. Quelle est la conclusion avec les données numériques de la fin de la partie I ?
8. Donner, en utilisant les tables à la fin de ce problème, une valeur approchée de la p-valeur associée à ce test, et affiner la conclusion.

IV (Facultatif) Variante sur les hypothèses du test

On reprend les notations de la partie III.

1. On ne présuppose pas que $\mu_1 \geq \mu_2$. On considère donc, au lieu de H_1, l'hypothèse alternative $H_1' = \{\mu_1 \neq \mu_2, \sigma > 0\}$. Quelles sont sous H_1' les valeurs de $\lim_{\min(n,m)\to\infty} T_{n,m}$?
2. En s'inspirant des questions de la partie III, construire un test pur convergent pour tester H_0 contre H_1' et donner une valeur approchée de sa p-valeur. Conclusion.
3. On considère l'hypothèse nulle $H_0' = \{\mu_1 \leq \mu_2, \sigma > 0\}$ et l'hypothèse alternative H_1. Vérifier que

$$\mathbb{P}_{(\mu_1,\sigma,\mu_2,\sigma)}\left(T_{n,m} > c\right) = \mathbb{P}_{(0,\sigma,0,\sigma)}\left(T_{n,m} > c - \sqrt{\frac{nm}{n+m}}\frac{\mu_1-\mu_2}{\sqrt{S_{n,m}}}\right)$$

En déduire que le test pur de la question III.6 est un test convergent de niveau α pour tester H_0' contre H_1. Conclusion.

V Quantiles

On fournit quelques quantiles pour les applications numériques.

Quantiles de la loi de Student. Soit T_k une variable aléatoire de loi de Student de paramètre k. On pose $\mathbb{P}(T_k \geq t) = \alpha$. La table fournit les valeurs de t en fonction de k et α. Par exemple $\boxed{\mathbb{P}(T_{54} \geq 1.674) \simeq 0.05}$.

n \ α	0.05000	0.02500	0.01000	0.00500	0.00250	0.00100	0.00050	0.00025	0.00010
54	1.674	2.005	2.397	2.670	2.927	3.248	3.480	3.704	3.991
55	1.673	2.004	2.396	2.668	2.925	3.245	3.476	3.700	3.986
56	1.673	2.003	2.395	2.667	2.923	3.242	3.473	3.696	3.981

Quantiles de la loi de Fisher-Snedecor. Soit $F_{k,l}$ une variable aléatoire de loi de Fisher-Snedecor de paramètre (k,l). On pose $\mathbb{P}(F_{k,l} \geq f) = \alpha$. La table fournit les valeurs de f en fonction de (k,l) et α. Par exemple $\boxed{\mathbb{P}(F_{24,30} \geq 1.098) \simeq 0.4}$.

(k,l) \ α	0.400	0.300	0.200	0.100	0.050	0.025
$(24,30)$	1.098	1.220	1.380	1.638	1.887	2.136
$(24,31)$	1.096	1.217	1.375	1.630	1.875	2.119
$(25,30)$	1.098	1.219	1.377	1.632	1.878	2.124
$(25,31)$	1.096	1.216	1.372	1.623	1.866	2.107
$(30,24)$	1.111	1.236	1.401	1.672	1.939	2.209
$(30,25)$	1.108	1.231	1.394	1.659	1.919	2.182
$(31,24)$	1.111	1.235	1.399	1.668	1.933	2.201
$(31,25)$	1.108	1.230	1.392	1.655	1.913	2.174

Quantiles de la loi de Fisher-Snedecor. Soit $F_{k,l}$ une variable aléatoire de loi de Fisher-Snedecor de paramètre (k,l). On pose $\mathbb{P}(F_{k,l} \leq f) = \alpha$. La table fournit les valeurs de f en fonction de (k,l) et α. Par exemple $\boxed{\mathbb{P}(F_{24,30} \leq 0.453) \simeq 0.025}$.

(k,l) \ α	0.025	0.050	0.100	0.200	0.300	0.400
$(24,30)$	0.453	0.516	0.598	0.714	0.809	0.900
$(24,31)$	0.454	0.517	0.599	0.715	0.810	0.900
$(25,30)$	0.458	0.521	0.603	0.717	0.812	0.902
$(25,31)$	0.460	0.523	0.604	0.718	0.813	0.902
$(30,24)$	0.468	0.530	0.611	0.725	0.820	0.911
$(30,25)$	0.471	0.532	0.613	0.726	0.821	0.911
$(31,24)$	0.472	0.533	0.614	0.727	0.822	0.912
$(31,25)$	0.475	0.536	0.616	0.729	0.823	0.912

△

XII.6 Corrections

Exercice XII.1.

1. La densité conditionnelle est $f_{X|Y}(x|y) = f_{X,Y}(x,y)/f_Y(y)$. Donc la densité de la loi de (X,Y) est
$$f_{X,Y}(x,y) = f_Y(y) f_{X|Y}(x|y) = \frac{1}{\pi} \mathrm{e}^{-y(1+x^2)} \mathbf{1}_{\{y>0\}}.$$
2. Par la formule des lois marginales, on a $f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\, dy = \frac{1}{\pi} \frac{1}{1+x^2}$. On peut remarquer que $\mathcal{L}(X)$ est la loi de Cauchy. Par définition, on a
$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x) = (1+x^2) \mathrm{e}^{-y(1+x^2)} \mathbf{1}_{\{y>0\}}.$$
La loi de Y sachant X est la loi exponentielle de paramètre $1+X^2$.
3. L'espérance d'une variable aléatoire de loi exponentielle de paramètre λ est $1/\lambda$. On en déduit donc que
$$\mathbb{E}[Y|X] = \frac{1}{1+X^2}\,.$$

▲

Exercice XII.2.

Étude simplifiée de la répartition d'un génotype dans la population humaine

I Distribution du génotype : le modèle de Hardy-Weinberg

1. On note (E_1, E_2) les gènes de l'enfant. E_1 représente le gène transmis par la mère et E_2 celui transmis par le père. L'espace d'états est

$$\Omega = \{(E_1, E_2); E_1 \in \{M_1, M_2\} \text{ et } E_2 \in \{P_1, P_2\}\},$$

où M_1M_2 est le génotype de la mère et P_1P_2 celui du père. Remarquons que le génotype de l'enfant est soit E_1E_2 soit E_2E_1 : on ne distingue pas la provenance des gènes. On suppose les transmissions des gènes de la mère et du père indépendants et indépendants des gènes a ou A. On choisit sur Ω la probabilité **uniforme**. Donc on a

$$\begin{aligned}\mathbb{P}(E = aa|M = aA, P = aA) &= \frac{\text{Card } \{(E_1, E_2) = (a, a); E_1 \in \{A, a\} \text{ et } E_2 \in \{a, A\}\}}{\text{Card } \{(E_1, E_2); E_1 \in \{A, a\} \text{ et } E_2 \in \{a, A\}\}} \\ &= \frac{1}{4}, \\ \mathbb{P}(E = aa|M = aa, P = aA) &= \frac{\text{Card } \{(E_1, E_2) = (a, a); E_1 \in \{a\} \text{ et } E_2 \in \{a, A\}\}}{\text{Card } \{(E_1, E_2); E_1 \in \{a\} \text{ et } E_2 \in \{a, A\}\}} \\ &= \frac{1}{2},\end{aligned}$$

et par symétrie

$$\begin{aligned}\mathbb{P}(E = aa|M = aA, P = aa) &= \frac{1}{2}, \\ \mathbb{P}(E = aa|M = aa, P = aa) &= 1.\end{aligned}$$

2. En distinguant tous les génotypes possibles pour les parents qui peuvent donner lieu au génotype aa pour l'enfant, il vient en utilisant la définition des **probabilités conditionnelles** :

$$\begin{aligned}\mathbb{P}(E = aa) =& \mathbb{P}(E = aa|M = aA, P = aA)\mathbb{P}(M = aA, P = aA) \\ &+ \mathbb{P}(E = aa|M = aa, P = aA)\mathbb{P}(M = aa, P = aA) \\ &+ \mathbb{P}(E = aa|M = aA, P = aa)\mathbb{P}(M = aA, P = aa) \\ &+ \mathbb{P}(E = aa|M = aa, P = aa)\mathbb{P}(M = aa, P = aa).\end{aligned}$$

On suppose les mariages aléatoires. En supposant le génotype de la mère et du père indépendant et de même répartition, on a $\mathbb{P}(M = i, P = j) = \mathbb{P}(M = i)\mathbb{P}(P = j) = u_iu_j$. On obtient $\mathbb{P}(E = aa) = \left(u_1 + \frac{u_2}{2}\right)^2$.

3. Par symétrie on a $\mathbb{P}(E = AA) = \left(u_3 + \frac{u_2}{2}\right)^2$.
4. En passant au **complément** on a $\mathbb{P}(E = aA) = 1 - \mathbb{P}(E \neq aA) = 1 - \mathbb{P}(E = aa) - \mathbb{P}(E = AA)$ car les évènements $\{E = aa\}$ et $\{E = AA\}$ sont **disjoints**. Donc on a $\mathbb{P}(E = aA) = 1 - \theta^2 - (1 - \theta)^2 = 2\theta(1 - \theta)$.
5. Le paramètre à la seconde génération est $\theta_2 = q_1 + \frac{q_2}{2} = \theta$. La répartition est identique à celle de la première génération. La répartition est donc stationnaire au cours du temps.

II Modèle probabiliste

1. Les v.a. $\left(\mathbf{1}_{\{X_i=j\}}, i \in \{1,\dots,n\}\right)$ sont des v.a. **indépendantes** de même loi : la loi de **Bernoulli** de paramètre q_j. La loi de N_j est donc une **binomiale** de paramètre (n, q_j).
2. $\mathbb{E}[N_j] = nq_j$ et $\mathrm{Var}(N_j) = nq_j(1-q_j)$.
3. N_j/n est un estimateur sans biais de q_j. Comme $\mathbf{1}_{X_i=j}$ est intégrable, on déduit de la **loi forte des grands nombres** que N_j/n est un estimateur convergent.
4. Comme $\mathbf{1}_{X_i=j}$ est de carré intégrable avec $\mathrm{Var}\,\mathbf{1}_{X_1=j} = q_j(1-q_j)$, on déduit du **théorème de la limite centrale** que $\sqrt{n}\left(\dfrac{N_j}{n} - q_j\right)$ converge en loi vers $\mathcal{N}(0, q_j(1-q_j))$. L'estimateur $\dfrac{N_j}{n}$ est donc un estimateur asymptotiquement normal de variance asymptotique $q_j(1-q_j)$.
5. $$\mathbb{P}(N_1=n_1, N_2=n_2, N_3=n_3) = \sum_{\substack{(x_1,\dots,x_n)\in\{1,2,3\}^n \\ \mathrm{Card}\,\{i; x_i=1\}=n_1 \\ \mathrm{Card}\,\{i; x_i=2\}=n_2}} \mathbb{P}(X_1=x_1,\dots,X_n=x_n)$$
   par **indépendance**, et comme les X_i ont **même loi**
   $$= \sum_{\substack{(x_1,\dots,x_n)\in\{1,2,3\}^n \\ \mathrm{Card}\,\{i; x_i=1\}=n_1 \\ \mathrm{Card}\,\{i; x_i=2\}=n_2}} q_1^{n_1} q_2^{n_2} q_3^{n-n_1-n_2}$$
   $$= \frac{n!}{n_1!n_2!n_3!} q_1^{n_1} q_2^{n_2} q_3^{n_3}.$$
6. La matrice de covariance de $Z_1 = \left(\mathbf{1}_{\{X_i=1\}}, \mathbf{1}_{\{X_i=2\}}\right)'$ est
   $$\Sigma = \begin{pmatrix} \mathrm{Var}\,\mathbf{1}_{\{X_1=1\}} & \mathrm{Cov}(\mathbf{1}_{\{X_1=1\}}, \mathbf{1}_{\{X_1=2\}}) \\ \mathrm{Cov}(\mathbf{1}_{\{X_1=1\}}, \mathbf{1}_{\{X_1=2\}}) & \mathrm{Var}\,\mathbf{1}_{\{X_1=2\}} \end{pmatrix} = \begin{pmatrix} q_1(1-q_1) & -q_1q_2 \\ -q_1q_2 & q_2(1-q_2) \end{pmatrix}.$$
   Par **indépendance**, on a $\mathrm{Cov}(N_1, N_2) = n\,\mathrm{Cov}(\mathbf{1}_{\{X_1=1\}}, \mathbf{1}_{\{X_1=2\}}) = -nq_1q_2$. Les variables aléatoires N_1 et N_2 ne sont pas indépendantes.
7. Les variables aléatoires vectorielles $Z_i = (\mathbf{1}_{X_i=1}, \mathbf{1}_{X_i=2})'$ sont **indépendantes, de même loi** et de **carrés intégrables**. Le **théorème de la limite centrale** vectoriel assure que $\sqrt{n}\left(\displaystyle\sum_{i=1}^n Z_i - n\mathbb{E}[Z_i]\right) = \left(\dfrac{N_1[n]-nq_1}{\sqrt{n}}, \dfrac{N_2[n]-nq_2}{\sqrt{n}}\right)'$ converge en loi quand $n \to \infty$ vers la loi gaussienne $\mathcal{N}(0, \Sigma)$.

III Estimation de θ à l'aide du génotype

1. $c = \log \dfrac{n!}{n_1!n_2!n_3!} + n_2 \log 2$.
2. $V_n = \dfrac{\partial L_n(N_1, N_2, N_3; \theta)}{\partial \theta} = \dfrac{2N_1}{\theta} + \dfrac{N_2}{\theta} - \dfrac{N_2}{1-\theta} - \dfrac{2N_3}{1-\theta}$.
3. $$\begin{aligned} I_n(\theta) &= -\mathbb{E}_\theta \left[\frac{\partial^2 L_n(N_1, N_2 N_3; \theta)}{\partial \theta^2} \right] \\ &= \mathbb{E}_\theta \left[\frac{2N_1}{\theta^2} + \frac{N_2}{\theta^2} + \frac{N_2}{(1-\theta)^2} + \frac{2N_3}{(1-\theta)^2} \right] = \frac{2n}{\theta(1-\theta)}. \end{aligned}$$
4. Si $(n_1, n_2) \neq (0,0)$ et $(n_2, n_3) \neq (0,0)$, alors
$$\lim_{\theta \downarrow 0} L_n(n_1, n_2, n_3; \theta) = \lim_{\theta \uparrow 1} L_n(n_1, n_2, n_3; \theta) = -\infty.$$
Pour trouver les maximums on regarde les zéros de
$$\frac{\partial L_n(n_1, n_2, n_3; \theta)}{\partial \theta} = \frac{2n_1}{\theta} + \frac{n_2}{\theta} - \frac{n_2}{1-\theta} - \frac{2n_3}{1-\theta}.$$
On trouve un seul zéro : $\theta = \dfrac{n_1}{n} + \dfrac{n_2}{2n}$. Si $(n_1, n_2) = (0,0)$, alors $L_n(n_1, n_2, n_3; \theta) = c + n \log \theta$. Le maximum est atteint pour $\theta = 0 = \dfrac{n_1}{n} + \dfrac{n_2}{2n}$. Si $(n_2, n_3) = (0,0)$, alors $L_n(n_1, n_2, n_3; \theta) = c + n \log(1-\theta)$. Le maximum est atteint pour $\theta = 1 = \dfrac{n_1}{n} + \dfrac{n_2}{2n}$. Dans tous les cas le maximum de la vraisemblance est atteint en un point **unique** $\theta = \dfrac{n_1}{n} + \dfrac{n_2}{2n}$. Donc $\hat{\theta}_n = \dfrac{N_1}{n} + \dfrac{N_2}{2n}$ est l'estimateur du maximum de vraisemblance de θ.
5. Oui. Par **linéarité** de l'espérance, on déduit de **II**.2. que $\mathbb{E}_\theta \left[\hat{\theta}_n \right] = \theta^2 + \theta(1-\theta) = \theta$.
6. On a
$$\operatorname{Var} \hat{\theta}_n = \frac{1}{4n^2} \operatorname{Var}(2N_1 + N_2) = \frac{1}{4n^2} \left[4 \operatorname{Var}(N_1) + \operatorname{Var}(N_2) + 4 \operatorname{Cov}(N_1, N_2) \right]$$
grâce aux questions **II**.2 et **II**.6,
$$= \frac{1}{4n} \left[4q_1(1-q_1) + q_2(1-q_2) - 4q_1 q_2 \right] = \frac{\theta(1-\theta)}{2n}.$$
La **borne FCDR** est $I_n(\theta)^{-1} = \dfrac{\theta(1-\theta)}{2n}$. L'estimateur est donc **efficace**. On peut remarquer sur l'équation (2) qu'il s'agit d'un **modèle exponentiel**.
7. En utilisant les résultats de **II**.3, on a que $\hat{\theta}_n$ est un estimateur convergent de θ. En remarquant que $\hat{\theta}_n = \sum_{i=1}^n \left(1, \frac{1}{2}\right) \cdot Z_i$, on déduit du **II**.7 que $\hat{\theta}_n$ est asymptotiquement normal de variance asymptotique $\lim\limits_{n \to \infty} n \operatorname{Var} \hat{\theta}_n = \dfrac{\theta(1-\theta)}{2}$.

IV Tests asymptotiques sur le modèle de Hardy-Weinberg

1. L'estimateur du maximum de vraisemblance de $q = (\hat{q}_1, \hat{q}_2, \hat{q}_3)$ est le vecteur des **fréquences empiriques** $\left(\frac{N_1}{n}, \frac{N_2}{n}, \frac{N_3}{n}\right)$.
2. Par le corollaire sur le test de Hausmann, on sait que $\zeta_n^{(1)} = n\sum_{j=1}^{3} \frac{(\hat{q}_j - q_j(\hat{\theta}_n))^2}{\hat{q}_j}$ converge sous H_0 en **loi** vers un χ^2 à $3-1-1=1$ degré de liberté. Rappelons que l'on enlève un degré de liberté pour l'estimation de $\theta \in]0,1[\subset \mathbb{R}$.
3. $W_n = \left\{\zeta_n^{(1)} \geq z\right\}$ est la **région critique** d'un test asymptotique convergent de niveau $\alpha = \mathbb{P}(\chi^2(1) \geq z)$.
4. On a

$n = 3,88\ 10^6$		$\hat{\theta}_n = 1,9853\ 10^{-2}$
$n_1 = 1580$	$\hat{q}_1 = 4,0722\ 10^{-4}$	$q_1(\hat{\theta}_n) = 3,9415\ 10^{-4}$
$n_2 = 150900$	$\hat{q}_2 = 3,8892\ 10^{-2}$	$q_2(\hat{\theta}_n) = 3,8918\ 10^{-2}$
$n_3 = 1580$	$\hat{q}_3 = 9,6070\ 10^{-1}$	$q_3(\hat{\theta}_n) = 9,6068\ 10^{-1}$.

   On obtient $\zeta_n^{(1)} = 1.69$. Pour $5\% = \mathbb{P}(\chi^2(1) \geq z)$, on lit dans la table $z = 3,84$. Comme $\zeta_n^{(1)} \leq 3,84$, on accepte donc H_0 au seuil de 5%.
5. On obtient $\zeta_n^{(1)} = 1163$. Comme $\zeta_n^{(1)} \geq 3,84$, on rejette donc H_0 au seuil de 5%. En fait le gène responsable de l'hémophilie est porté par le chromosome sexuel X. Le génotype pour la population féminine est donc aa, aA ou AA. En revanche le génotype pour la population masculine est a ou A. Le modèle de Hardy-Weinberg n'est plus adapté.

▲

Exercice XII.3.

1. L'estimateur du maximum de vraisemblance, $\hat{p}$, de p est le vecteur des **fréquences empiriques**. On a donc $\hat{p} = (0,606; 0,186; 0,173; 0,035)$.
2. La dimension du vecteur p est 4. Il faut tenir compte de la contrainte $p_{J,N}+p_{W,N}+p_{J,C}+p_{W,C} = 1$. Enfin l'hypothèse d'indépendance revient à dire que $p = h(p_J, p_N)$, où p_J est la probabilité de naître un jour ouvrable et p_N la probabilité pour que l'accouchement soit sans césarienne. En particulier, on a $p_{J,N} = p_J p_N$, $p_{W,N} = (1-p_J)p_N$, $p_{J,C} = p_J(1-p_N)$ et $p_{W,C} = (1-p_J)(1-p_N)$. Il faut tenir compte des deux estimations : celle de p_J et celle de p_N. Le nombre de degrés de liberté du test du χ^2 est donc q**=4-1-2=1**. L'estimateur du maximum de vraisemblance $\hat{p}_J$, de p_J, et $\hat{p}_N$, de p_N, est celui des fréquences empiriques. On a donc $\hat{p}_J = 0.779$ et $\hat{p}_N = 0.792$. La statistique du χ^2 est

$$\zeta_n^{(1)} = n\frac{(\hat{p}_{J,N} - \hat{p}_J\hat{p}_N)^2}{\hat{p}_{J,N}} + n\frac{(\hat{p}_{W,N} - (1-\hat{p}_J)\hat{p}_N)^2}{\hat{p}_{W,N}} + n\frac{(\hat{p}_{J,C} - \hat{p}_J(1-\hat{p}_N))^2}{\hat{p}_{J,C}} + n\frac{(\hat{p}_{W,C} - (1-\hat{p}_J)(1-\hat{p}_N))^2}{\hat{p}_{W,C}}.$$

   On obtient $\zeta_n^{(1)} \simeq 18\ 219$. On lit dans la table du χ^2 que $\mathbb{P}(X > 11) \leq 0,1\%$, où la loi de X est $\chi^2(1)$. On rejette donc l'hypothèse d'indépendance au niveau de 99,9%. (On aurait également pu utiliser la statistique $\zeta_n^{(2)}$, avec dans notre cas particulier $\zeta_n^{(2)} \simeq 15\ 594$.)

3. La dimension du vecteur p est 4. Il faut tenir compte de la contrainte $p_{J,N} + p_{W,N} + p_{J,C} + p_{W,C} = 1$. Enfin on teste $p = p^0$, avec $p^0 = (0,605; 0,189; 0,17; 0,036)$. Il n'y a pas de paramètre à estimer. Le nombre de degrés de liberté du test du χ^2 est donc q=**4-1=3**. La statistique du χ^2 est

$$\zeta_n^{(1)} = n\frac{(\hat{p}_{J,N} - p^0_{J,N})^2}{\hat{p}_{J,N}} + n\frac{(\hat{p}_{J,N} - p^0_{J,N})^2}{\hat{p}_{W,N}} + n\frac{(\hat{p}_{J,C} - p^0_{J,C})^2}{\hat{p}_{J,C}} + n\frac{(\hat{p}_{W,C} - p^0_{W,C})^2}{\hat{p}_{W,C}}.$$

On obtient $\zeta_n^{(1)} \simeq 412$. On lit dans la table du χ^2 que $\mathbb{P}(X > 17) \leq 0,1\%$, où la loi de X est $\chi^2(3)$. On rejette donc l'hypothèse au niveau de 99,9%. Il y a donc une évolution entre 1996 et 1997. (On aurait également pu utiliser la statistique $\zeta_n^{(2)}$, avec dans notre cas particulier $\zeta_n^{(2)} \simeq 409$.)

▲

Exercice XII.4.

I Modélisation

1. Les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes de loi de Bernoulli de paramètre $p = n_0/N$.
2. La variable aléatoire est discrète. Sa densité est donnée par $p(x_1; p) = \mathbb{P}_p(X_1 = x_1)$, avec $x_1 \in \{0, 1\}$.
3. La densité de l'échantillon est

$$p_n(x; p) = \left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i} (1-p)^n, \quad \text{où} \quad x = (x_1, \ldots, x_n) \in \{0,1\}^n.$$

4. Remarquons que $p_n(x; p) = \psi(S_n(x), p)$, où $\psi(y, p) = p^y(1-p)^{n-y}$ et $S_n(x) = \sum_{i=1}^n x_i$. Par le théorème de factorisation, on déduit que la statistique $S_n = \sum_{i=1}^n X_i$ est exhaustive. La statistique S_n est en fait totale, car il s'agit d'un modèle exponentiel. On peut démontrer directement que la statistique est totale.
5. Pour que $h(S_n)$ soit intégrable, il faut et il suffit que h, définie sur $\{0, \ldots, n\}$, soit bornée. On a alors

$$\mathbb{E}_p[h(S_n)] = \sum_{k=0}^n C_n^k h(k) p^k (1-p)^{n-k}.$$

En prenant la limite de l'expression ci-dessus pour $p \to 0$, il vient $\lim_{p\to 0} \mathbb{E}_p[h(S_n)] = h(0)$. Soit $\delta = h(S_n)$ un estimateur intégrable sans biais de $1/p$. On a donc $\mathbb{E}_p[h(S_n)] = 1/p$. En regardant $p \to 0$, on déduit de ce qui précède que $h(0) = +\infty$. Or $h(S_n)$ est intégrable, implique que h est bornée. Il y a donc contradiction. Il n'existe pas d'estimateur sans biais de $1/p$ fonction de S_n.
6. Soit δ un estimateur (intégrable) sans biais de $1/p$. Alors l'estimateur $\mathbb{E}_p[\delta|S_n] = h(S_n)$ est un estimateur sans biais de $1/p$, qui est fonction de S_n seulement. Cela est absurde d'après la question précédente. Il n'existe donc pas d'estimateur sans biais de $1/p$.

II Estimation asymptotique

1. Les variables aléatoires $(X_i, i \in \mathbb{N}^*)$ sont indépendantes, de même loi et intégrables. On déduit de la **loi forte des grands nombres** que la suite $(S_n/n, n \in \mathbb{N}^*)$ converge $\mathbb{P}_p$-p.s. vers p. Comme $\lim_{n\to\infty} \frac{n}{n+1} = 1$ et $\lim_{n\to\infty} \frac{1}{n+1} = 0$, on en déduit $\frac{n}{n+1} S_n + \frac{1}{n+1} \overset{\mathbb{P}_p\text{-p.s.}}{\longrightarrow} p$ quand $n \to \infty$. Comme la fonction $g(x) = 1/x$ est continue en p, on en déduit que
$$\frac{n+1}{S_n+1} \overset{\mathbb{P}_p\text{-p.s.}}{\longrightarrow} 1/p \quad \text{quand} \quad n \to \infty.$$
L'estimateur $n_0 \frac{n+1}{S_n+1}$ est l'estimateur de Bailey de N. L'estimateur de Petersen $n_0 \frac{n}{S_n}$ est également un estimateur convergent de N, mais il n'est pas intégrable.
2. On a
$$\begin{aligned}
\mathbb{E}_p\left[\frac{n+1}{S_n+1}\right] &= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \frac{n+1}{k+1} p^k (1-p)^{n-k} \\
&= \frac{1}{p} \sum_{k=0}^{n} \frac{(n+1)!}{(k+1)!(n-k)!} p^{k+1} (1-p)^{n-k} \\
&= \frac{1}{p} \sum_{j=0}^{n+1} \frac{(n+1)!}{j!(n+1-j)!} p^j (1-p)^{n+1-j} - \frac{1}{p}(1-p)^{n+1} \\
&= \frac{1}{p}\left[1 - (1-p)^{n+1}\right].
\end{aligned}$$
Le biais est $b_n = -(1-p)^{n+1}/p$. Remarquons que $-pb_n = \left(1 - \frac{n_0}{N}\right)^{n+1} \simeq \mathrm{e}^{-nn_0/N}$ si $n_0 \ll N$.
3. Les v.a. $(X_i, i \in \mathbb{N}^*)$ sont de carré intégrable. On déduit du **théorème central limite** que la suite $\left(\sqrt{n}\left(\frac{S_n}{n} - p\right), n \in \mathbb{N}^*\right)$ converge en loi vers la loi gaussienne $\mathcal{N}(0, \sigma^2)$, où $\sigma^2 = \mathrm{Var}_p(X_1) = p(1-p)$.
4. On a
$$0 \le \sqrt{n}\left(\frac{S_n+1}{n+1} - \frac{S_n}{n}\right) = \frac{n - S_n}{\sqrt{n}(n+1)} \le \frac{\sqrt{n}}{n+1} \le \frac{1}{\sqrt{n}}.$$
La suite converge donc $\mathbb{P}_p$-p.s. vers 0. On déduit du théorème de Slutsky que la suite $\left(\left(\sqrt{n}\left(\frac{S_n}{n} - p\right), \sqrt{n}\left(\frac{S_n+1}{n+1} - p\right)\right), n \in \mathbb{N}^*\right)$ converge en loi vers $(X, 0)$, où $\mathcal{L}(X) = \mathcal{N}(0, \sigma^2)$. Par continuité, on en déduit que
$$\sqrt{n}\left(\frac{S_n+1}{n+1} - p\right) = \sqrt{n}\left(\frac{S_n}{n} - p\right) + \sqrt{n}\left(\frac{S_n+1}{n+1} - \frac{S_n}{n}\right) \overset{\text{en loi}}{\longrightarrow} X + 0 = X,$$
quand $n \to \infty$. La suite $\left(\frac{S_n+1}{n+1}, n \in \mathbb{N}^*\right)$ est une suite d'estimateurs de p convergente et asymptotiquement normale de variance asymptotique $\sigma^2 = \mathrm{Var}(X)$.

5. La fonction $g(x) = 1/x$ est de classe C^1 au point $p \in]0,1[$. On déduit du théorème de convergence des estimateurs de substitution que la suite de variables aléatoires $\left(\sqrt{n}\left(\frac{n+1}{S_n+1} - \frac{1}{p}\right), n \in \mathbb{N}^*\right)$ converge en loi vers la loi gaussienne $\mathcal{N}(0, s^2)$, où $s^2 = \sigma^2\left(\frac{\mathrm{d}g(p)}{\mathrm{d}p}\right)^2 = \frac{p(1-p)}{p^4} = \frac{1-p}{p^3}$. La suite $\left(\frac{n+1}{S_n+1}, n \in \mathbb{N}^*\right)$ est une suite d'estimateurs de $1/p$ convergente et asymptotiquement normale de variance asymptotique s^2.

6. La log-vraisemblance est $L_1(x_1; p) = x_1 \log\left(\frac{p}{1-p}\right) + \log(1-p)$. Le score est

$$V_1 = \frac{\partial L_1(X_1; p)}{\partial p} = X_1\left(\frac{1}{p} + \frac{1}{1-p}\right) - \frac{1}{1-p} = \frac{X_1}{p(1-p)} - \frac{1}{1-p}\,.$$

L'information de Fisher est

$$I(p) = \mathrm{Var}_p(V_1) = \mathrm{Var}_p(X_1/p(1-p)) = \frac{1}{p^2(1-p)^2}\,p(1-p) = \frac{1}{p(1-p)}\,.$$

Soit la fonction $g(x) = 1/x$. La borne FDCR de l'échantillon de taille 1, pour l'estimation de $g(p) = 1/p$ est

$$\left(\frac{\mathrm{d}g(p)}{\mathrm{d}p}\right)^2 \frac{1}{I(p)} = \frac{1}{p^4}\,p(1-p) = \frac{1-p}{p^3}.$$

La variance asymptotique de l'estimateur $\dfrac{n+1}{S_n+1}$ est égale à la borne FDCR, de l'échantillon de taille 1, pour l'estimation de $1/p$. L'estimateur est donc asymptotiquement efficace.

III Intervalle de confiance

1. Comme la suite $\left(\frac{S_n}{n}, n \in \mathbb{N}^*\right)$ converge $\mathbb{P}_p$-p.s. vers p, on en déduit que

$$Y_n = \sqrt{\left(\frac{S_n}{n}\right)^3 \frac{n}{n - S_n}} \underset{n\to\infty}{\longrightarrow} p^{3/2}(1-p)^{-1/2} \quad \mathbb{P}_p\text{-p.s.}$$

On déduit du théorème de Slutsky que la suite $\left(\left(Y_n, \sqrt{n}\left(\frac{n+1}{S_n+1} - \frac{1}{p}\right)\right), n \in \mathbb{N}^*\right)$ converge en loi vers $\left(p^{3/2}(1-p)^{-1/2}, X\right)$, où la loi de X est la loi gaussienne $\mathcal{N}(0, (1-p)/p^3)$. Par continuité, on en déduit que

$$Y_n\sqrt{n}\left(\frac{n+1}{S_n+1} - \frac{1}{p}\right) \overset{\text{en loi}}{\longrightarrow} \mathcal{N}(0,1), \quad \text{quand} \quad n \to \infty.$$

Soit I un intervalle de $\mathbb{R}$, on a en particulier

$$\mathbb{P}_p\left(Y_n\sqrt{n}\left(\frac{n+1}{S_n+1} - \frac{1}{p}\right) \in I\right) \underset{n\to\infty}{\longrightarrow} \mathbb{P}(T \in I),$$

où la loi de T est la loi gaussienne centrée réduite $\mathcal{N}(0,1)$. Pour $I=[-1,96;1,96]$, on a $\mathbb{P}(T\in I)=95\%$. On en déduit que pour n grand, on a

$$\mathbb{P}_p\left(Y_n\sqrt{n}\left(\frac{n+1}{S_n+1}-\frac{1}{p}\right)\in[-1,96;1,96]\right)=\mathbb{P}_p\left(\frac{1}{p}\in\left[\frac{n+1}{S_n+1}\pm\frac{1,96}{Y_n\sqrt{n}}\right]\right)\simeq 95\%.$$

Donc $\left[\frac{n+1}{S_n+1}-\frac{1,96}{Y_n\sqrt{n}};\frac{n+1}{S_n+1}+\frac{1,96}{Y_n\sqrt{n}}\right]$ est un intervalle de confiance asymptotique à 95%.

2. L'estimateur de $1/p$ est $\frac{n+1}{S_n+1}=463/341\simeq 1,36$. L'estimateur de N est $\frac{n_0}{p}=\frac{74*463}{341}\simeq 100$.

3. On a $\frac{n+1}{S_n+1}=463/341\simeq 1,36$, et $Y_n=\sqrt{\left(\frac{S_n}{n}\right)^3\frac{n}{n-S_n}}\simeq 1,23$. L'intervalle de confiance asymptotique à 95% de $1/p$ est $[1,28;1,43]$. L'intervalle de confiance asymptotique à 95% de $N=n_0/p$ est $[95,106]$.

IV Tests

1. Le rapport de vraisemblance est $Z(x)=\left(\frac{p_1(1-p_0)}{(1-p_1)p_0}\right)^{\sum_{i=1}^n x_i}\left(\frac{1-p_1}{1-p_0}\right)^n$, où $x=(x_1,\dots,x_n)\in\{0,1\}^n$. Le rapport de vraisemblance est une fonction de $S_n(x)=\sum_{i=1}^n x_i$. La statistique de test est $S_n=\sum_{i=1}^n X_i$.

2. Comme $N_1>N_0$, on a $p_0>p_1$. La condition $Z(x)>k$ est équivalente à la condition $S_n(x)<c$, pour une constante c qui ne dépend pas de x. Le test UPP φ de niveau α est alors défini par :

$$\begin{aligned}\varphi(x)&=1 \quad \text{si} \quad S_n(x)<c,\\ \varphi(x)&=\gamma \quad \text{si} \quad S_n(x)=c,\\ \varphi(x)&=0 \quad \text{si} \quad S_n(x)>c,\end{aligned}$$

Les constantes c et γ sont définies par la condition $\mathbb{E}_{p_0}[\varphi]=\alpha$.

3. Comme les constantes c et γ sont définies par la condition $\mathbb{E}_{p_0}[\varphi]=\alpha$, elles sont indépendantes de la valeur de N_1. Le test φ est UPP de niveau α pour tester $H_0=\{N=N_0\}$ contre $H_1=\{N=N_1\}$, et ce pour toutes valeurs de N_1 ($>N_0$). En particulier il est UPP de niveau α pour tester $H_0=\{N=N_0\}$ contre $H_1=\{N>N_0\}$.

4. Pour déterminer le test φ, il faut calculer les constantes c et γ. Elles sont définies par $\mathbb{P}_{p_0}(S_n<c)+\gamma\mathbb{P}_{p_0}(S_n=c)=\alpha$, où $\alpha=5\%$. La constante c est également déterminée par la condition $\mathbb{P}_{p_0}(S_n<c)\leq\alpha<\mathbb{P}_{p_0}(S_n<c+1)$. À la vue du tableau, on en déduit que $c=341$, et $\gamma=\frac{\alpha-\mathbb{P}_{p_0}(S_n<c)}{\mathbb{P}_{p_0}(S_n<c+1)-\mathbb{P}_{p_0}(S_n<c)}$, soit $\gamma\simeq 0,6$. Comme $m<341$, on rejette l'hypothèse H_0 au niveau 5%.

Remarquons que si on a $m=341$, alors pour décider, on tire un nombre u uniforme sur $[0,1]$. Si $u<0,6$, alors on rejette H_0, sinon on accepte H_0.

Remarquons que nous rejetons l'hypothèse $N = 96$ (le même test permet en fait de rejeter l'hypothèse $N \leq 96$, il s'agit d'un test UPP unilatéral dans le cadre du modèle exponentiel). Le test UPP est ici plus précis que l'intervalle de confiance de la question précédente. Il est également plus simple à calculer, dans la mesure, où l'on ne doit pas calculer de variance asymptotique σ^2, ni d'estimation de σ^2.

▲

Exercice XII.5.

I L'estimateur du maximum de vraisemblance de θ

1. On a $f_\theta(t) = -\dfrac{\partial \bar{F}_\theta(t)}{\partial t} = \theta k (t-w)^{k-1}\, \mathrm{e}^{-\theta(t-w)^k}\, \mathbf{1}_{\{t>w\}}$.
2. Comme les variables sont **indépendantes**, la densité de l'échantillon est
$$p_n^*(t_1,\dots,t_n;\theta) = \theta^n k^n \left(\prod_{i=1}^n (t_i-w)^{k-1} \mathbf{1}_{\{t_i>w\}} \right) \mathrm{e}^{-\theta \sum_{i=1}^n (t_i-w)^k}\,.$$
3. La log-vraisemblance est définie pour $t = (t_1,\dots,t_n) \in]w,+\infty[^n$, par
$$L_n^*(t;\theta) = n\log(\theta) - \theta \sum_{i=1}^n (t_i-w)^k + c_n^*(t),$$
où la fonction c_n^* est indépendante de θ. Pour maximiser la log-vraisemblance, on cherche les zéros de sa dérivée. Il vient
$$0 = \frac{\partial L_n^*(t;\theta)}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^n (t_i-w)^k \quad \text{soit} \quad \theta = \frac{n}{\sum_{i=1}^n (t_i-w)^k}.$$
Comme de plus $\lim_{\theta\to 0} L_n^*(t;\theta) = \lim_{\theta\to+\infty} L_n^*(t;\theta) = -\infty$, on en déduit que la log-vraisemblance est maximale pour $\theta = n/\sum_{i=1}^n (t_i-w)^k$. L'estimateur du maximum de vraisemblance est donc
$$\hat{\theta}_n^* = \frac{n}{\sum_{i=1}^n (T_i-w)^k}.$$
4. L'information de Fisher est définie par
$$I^*(\theta) = \mathbb{E}_\theta\left[-\frac{\partial^2 L_1^*(T;\theta)}{\partial \theta^2} \right] = \frac{1}{\theta^2}\,.$$
5. Comme p.s. $T > w$, la variable $(T-w)^{\alpha k}$ est positive. On peut donc calculer son espérance. On a pour $\alpha > 0$,
$$\begin{aligned} \mathbb{E}_\theta[(T-w)^{\alpha k}] &= \int (t-w)^{\alpha k} f_\theta(t)\, dt \\ &= \int_w^\infty \theta k (t-w)^{\alpha k + k - 1}\, \mathrm{e}^{-\theta(t-w)^k}\, dt. \end{aligned}$$
En posant $y = \theta(t-w)^k$, il vient
$$\mathbb{E}_\theta[(T-w)^{\alpha k}] = \theta^{-\alpha} \int_0^\infty y^\alpha\, \mathrm{e}^{-y}\, dy = \theta^{-\alpha}\Gamma(\alpha+1).$$
En particulier, on a $\mathbb{E}_\theta[(T-w)^k] = \theta^{-1}$ et $\mathbb{E}_\theta[(T-w)^{2k}] = 2\theta^{-2}$.

6. Comme les variables aléatoires $(T_1 - w)^k, \dots (T_n - w)^k$ sont indépendantes de même loi et intégrables, on déduit de la loi forte des grands nombres que la suite de terme général $Z_n = \frac{1}{n}\sum_{i=1}^n (T_i - w)^k$ converge presque sûrement vers $\mathbb{E}_\theta[(T-w)^k] = \theta^{-1}$. Comme la fonction $h : x \mapsto 1/x$ est continue en $\theta^{-1} > 0$, on en déduit que la suite $(\hat{\theta}_n^* = h(Z_n), n \geq 1)$ converge presque sûrement vers $h(\theta^{-1}) = \theta$. L'estimateur du maximum de vraisemblance est donc **convergent**. Comme de plus les variables $(T_1 - w)^k, \dots (T_n - w)^k$ sont de carré intégrable, on déduit du théorème de la limite centrale que la suite $(\sqrt{n}(Z_n - \theta^{-1}), n \geq 1)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \sigma^2)$, où $\sigma^2 = \mathrm{Var}((T-w)^k) = \theta^{-2}$. Comme la fonction h est de classe C^1 en $\theta^{-1} > 0$, on en déduit que la suite $(\sqrt{n}(\hat{\theta}_n^* - \theta) = \sqrt{n}(h(Z_n) - h(\theta^{-1})), n \geq 1)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \Sigma^2)$, où $\Sigma^2 = \sigma^2 h'(\theta^{-1})^2 = \theta^2$. L'estimateur du maximum de vraisemblance est donc **asymptotiquement normal**.
7. Comme $\Sigma^2 = I(\theta)^{-1}$, cela signifie que l'estimateur du maximum de vraisemblance est **asymptotiquement efficace**.
8. Remarquons que $(\hat{\theta}_n^*)^2$ est un estimateur convergent de Σ^2. On déduit donc du théorème de Slutsky que la suite $(\sqrt{n}(\hat{\theta}_n^* - \theta)/\hat{\theta}_n^*, n \geq 1)$ converge en loi vers la loi gaussienne $\mathcal{N}(0,1)$. En particulier $J_n^* = \left[\hat{\theta}_n^* \pm \dfrac{1,96\ \hat{\theta}_n^*}{\sqrt{n}}\right]$ est un intervalle de confiance asymptotique à 95% de θ. On a $\hat{\theta}_n^* = 17/33\ 175\ 533 = 5,124.\ 10^{-7}$. On obtient donc l'intervalle de confiance à 95% :

$$\left[5,124.\ 10^{-7} \pm 2,46.\ 10^{-7}\right] = [2,7.\ 10^{-7}; 7,6.\ 10^{-7}].$$

II Données censurées

1. La loi de X est une loi de Bernoulli de paramètre $p = \mathbb{P}_\theta(X = 1)$.
2. On a pour $x = 1$

$$\mathbb{P}_\theta(R > r, X = 1) = \mathbb{P}_\theta(S \geq T > r) = \int \mathbf{1}_{\{s \geq t > r\}} f_\theta(t) g(s)\ dtds = \int_r^\infty f_\theta(t) \bar{G}(t)\ dt,$$

et pour $x = 0$,

$$\mathbb{P}_\theta(R > r, X = 0) = \mathbb{P}_\theta(T > S > r) = \int \mathbf{1}_{\{t > s > r\}} f_\theta(t) g(s)\ dtds = \int_r^\infty g(t) \bar{F}_\theta(t)\ dt.$$

On en déduit donc par dérivation que

$$p(r, 1; \theta) = f_\theta(r)\bar{G}(r) = \theta\, \mathrm{e}^{-\theta(r-w)_+^k}\, c(r,1) \quad \text{avec} \quad c(r,1) = k(w-r)_+^{k-1}\bar{G}(r),$$

et

$$p(r, 0; \theta) = g(r)\bar{F}_\theta(r) = \mathrm{e}^{-\theta(r-w)_+^k}\, c(r,0) \quad \text{avec} \quad c(r,0) = g(r).$$

La fonction c est indépendante de θ.
3. N_n représente le nombre de souris pour lesquelles on a observé les effets des produits toxiques.

4. La densité de l'échantillon de taille n est pour $r = (r_1, \ldots, r_n) \in \mathbb{R}^n$, $x = (x_1, \ldots, x_n) \in \{0,1\}^n$,
$$p_n(r, x; \theta) = \prod_{i=1}^{n} p(r_i, x_i; \theta) = \theta^{\sum_{i=1}^n x_i} \, \mathrm{e}^{-\theta \sum_{i=1}^n (r_i - w)_+^k} \, c_n(r, x),$$
où la fonction $c_n(r,x) = \prod_{i=1}^n c(r_i, x_i)$ est indépendante de θ. La log-vraisemblance de l'échantillon est donc
$$L_n(r, x; \theta) = \log(\theta) \sum_{i=1}^{n} x_i - \theta \sum_{i=1}^{n} (r_i - w)_+^k + \log(c_n(r,x)).$$
En raisonnant comme dans la partie précédente, on déduit que la log-vraisemblance atteint son maximum pour $\theta = \sum_{i=1}^{n} x_i / \sum_{i=1}^{n} (r_i - w)_+^k$. Cela reste vrai même si $\sum_{i=1}^n x_i = 0$. L'estimateur du maximum de vraisemblance est donc
$$\hat{\theta}_n = \frac{N_n}{\sum_{i=1}^n (R_i - w)_+^k}.$$
Contrairement à $\hat{\theta}_n^*$, l'estimateur $\hat{\theta}_n$ tient compte même des données censurées au dénominateur. En revanche le numérateur représente toujours le nombre de souris pour lesquelles on observe l'apparition des effets dus aux produits toxiques. Remarquons également que la loi de S n'intervient pas directement dans l'estimateur $\hat{\theta}_n$. En particulier, il n'est pas nécessaire de connaître explicitement la fonction g ou $\bar{G}$.

5. Les variables aléatoires $X_1, \ldots, X_n$ sont indépendantes intégrables et de même loi de Bernoulli de paramètre p. Par la loi forte des grands nombres, on en déduit que N_n/n est un estimateur convergent de p. Comme $\mathbb{E}_\theta[N_n/n] = \mathbb{E}_\theta[X] = p$, on en déduit que N_n/n est un estimateur sans biais de p.

6. L'information de Fisher pour l'échantillon de taille 1 est définie par
$$I(\theta) = \mathbb{E}_\theta \left[-\frac{\partial^2 L_1^*(R_1, X_1; \theta)}{\partial \theta^2} \right] = \frac{\mathbb{E}_\theta[X_1]}{\theta^2} = \frac{p}{\theta^2}.$$

7. Comme la fonction $h : (y,z) \mapsto y/z^2$ est continue sur $]0, \infty[^2$, on en déduit que la suite $(h(N_n/n, \hat{\theta}_n), n \geq 1)$ converge presque sûrement vers $h(p, \theta) = p/\theta^2 = I(\theta)$. L'estimateur $\dfrac{N_n}{n} \dfrac{1}{\hat{\theta}_n^2}$ est donc un estimateur convergent de $I(\theta)$.

8. L'estimateur du maximum de vraisemblance est asymptotiquement efficace. Donc la suite $(\sqrt{n}(\hat{\theta}_n - \theta), n \geq 1)$ converge en loi vers la loi gaussienne $\mathcal{N}(0, I(\theta)^{(-1)})$. Comme $N_n/(n\hat{\theta}_n^2)$ est un estimateur convergent de $I(\theta)$, on déduit du théorème de Slutsky que la suite $(\sqrt{N_n}(\hat{\theta}_n - \theta)/\hat{\theta}_n, n \geq 1)$ converge en loi vers la loi gaussienne $\mathcal{N}(0,1)$. En particulier $J_n = \left[\hat{\theta}_n \pm \dfrac{1,96 \, \hat{\theta}_n}{\sqrt{N_n}} \right]$ est un intervalle de confiance asymptotique à 95% de θ. On a $\hat{\theta}_n = 17/(33\,175\,533 + 4\,546\,880) = 4,507.\ 10^{-7}$. On obtient donc l'intervalle de confiance à 95% :
$$\left[4,507.\ 10^{-7} \pm 2,142.\ 10^{-7}\right] = [2,4.\ 10^{-7}; 6,6.\ 10^{-7}].$$

L'intervalle de confiance J_n est plus étroit que J_n^*. Mais surtout l'estimation à l'aide de $\hat{\theta}_n^*$ donne des résultats pour θ surévalués. De plus l'estimateur $\hat{\theta}_n^*$ n'est pas convergent en présence de données censurées.

9. Pour les souris $i \in \{n+1, \dots, n+m\}$ supplémentaires retirées avant l'instant, on a $x_i = 0$ et $(r_i - w)_+^k = 0$ aussi. En particulier la densité de l'échantillon $n+m$ est pour $r = (r_1, \dots, r_{n+m})$ et $x = (x_1, \dots, x_{n+m})$,

$$p_{n+m}(r, x; \theta) = p_n((r_1, \dots, r_n), (x_1, \dots, x_n); \theta) \prod_{i=n+1}^{n+m} \theta^0 \, \mathrm{e}^{-0} \, c(r_i, 0) = \theta^{\sum_{i=1}^n x_i} \, \mathrm{e}^{-\theta \sum_{i=1}^n (r_i - w)^k} \, c_{n+m}(r, x).$$

En particulier cela ne modifie pas l'estimateur du maximum de vraisemblance $\hat{\theta}_{n+m} = \sum_{i=1}^n (r_i - w)_+^k = \hat{\theta}_n$, ni l'estimation de $I_n(\theta) = nI(\theta)$ par $\sum_{i=1}^n x_i / \hat{\theta}_n^2$. En particulier l'intervalle de confiance asymptotique de θ reste inchangé.

III Comparaison de traitements

1. La densité de Z_1 est le produit de la densité de (R_1^A, X_1^A) et de la densité de (R_1^B, X_1^B) car les variables aléatoires sont indépendantes. La densité de l'échantillon $Z_1, \dots, Z_n$ est pour $r^A = (r_1^A, \dots, r_n^A)$, $r^B = (r_1^B, \dots, r_n^B) \in \mathbb{R}^n$, et $x^A = (x_1^A, \dots, x_n^A)$, $x^B = (x_1^B, \dots, x_n^B) \in \{0,1\}^n$

$$p_n(r^A, x^A, r^B, x^B; \theta^A, \theta^B) = (\theta^A)^{\sum_{i=1}^n x_i^A} (\theta^B)^{\sum_{i=1}^n x_i^B} \mathrm{e}^{-\theta^A \sum_{i=1}^n (r_i^A - w)_+^k - \theta^B \sum_{i=1}^n (r_i^B - w)_+^k} \, c_n(r^A, x^A) c_n(r^B, x^B).$$

2. On déduit de la partie précédente que l'estimateur du maximum de vraisemblance de (θ^A, θ^B) est $(\hat{\theta}_n^A, \hat{\theta}_n^B)$, où $\hat{\theta}_n^j = \dfrac{N_n^j}{\sum_{i=1}^n (R_i^j - w)_+^k}$ et $N_n^j = \sum_{i=1}^n X_i^j$ pour $j \in \{A, B\}$.
3. La log-vraisemblance pour l'échantillon de taille 1 est

$$L(r_1^A, x_1^A, r_1^B, x_1^B; \theta^A, \theta^B) = x_1^A \log(\theta^A) + x_1^B \log(\theta^B) - \theta^A (r_1^A - w)_+^k - \theta^B (r_1^B - w)_+^k + \log c_1(r_1^A, x_1^A) + \log c_1(r_1^B, x_1^B).$$

La matrice des dérivées secondes de la log-vraisemblance est définie par

$$L^{(2)}(r_1^A, x_1^A, r_1^B, x_1^B) = \begin{pmatrix} \frac{\partial^2 L}{\partial \theta^A \partial \theta^A} & \frac{\partial^2 L}{\partial \theta^A \partial \theta^B} \\ \frac{\partial^2 L}{\partial \theta^A \partial \theta^B} & \frac{\partial^2 L}{\partial \theta^B \partial \theta^B} \end{pmatrix} = \begin{pmatrix} -x_1^A / (\theta^A)^2 & 0 \\ 0 & -x_1^B / (\theta^B)^2 \end{pmatrix}.$$

L'information de Fisher est donc

$$I(\theta^A, \theta^B) = -\mathbb{E}_{(\theta^A, \theta^B)}[L^{(2)}(R_1^A, X_1^A, R_1^B, X_1^B)] = \begin{pmatrix} p_A / (\theta^A)^2 & 0 \\ 0 & p_B / (\theta^B)^2 \end{pmatrix}.$$

Il s'agit bien d'une matrice diagonale.

4. La densité de l'échantillon de taille n est

$$p(r^A, x^A, r^B, x^B; \theta, \theta) = \theta^{\sum_{i=1}^n x_i^A + \sum_{i=1}^n x_i^B} \, \mathrm{e}^{-\theta(\sum_{i=1}^n (r_i^A - w)_+^k + \sum_{i=1}^n (r_i^B - w)_+^k)} \\ c_n(r^A, x^A) c_n(r^B, x^B).$$

On en déduit que l'estimateur du maximum de vraisemblance de θ est donc

$$\hat{\theta}_n = \frac{N_n^A + N_n^B}{\sum_{i=1}^n (R_i^A - w)_+^k + \sum_{i=1}^n (R_i^B - w)_+^k}.$$

5. On utilise les résultats concernant les tests de **Hausman**. Le paramètre (θ_A, θ_B) est de dimension $q = 2$, le paramètre θ est de dimension $q' = 1$. Comme la matrice de l'information de Fisher est diagonale, on obtient

$$\zeta_n^{(1)} = n(\hat{\theta}_n^A - \hat{\theta}_n)^2 p_A (\hat{\theta}_n^A)^{-2} + n(\hat{\theta}_n^B - \hat{\theta}_n)^2 p_B (\hat{\theta}_n^B)^{-2}, \\ \zeta_n^{(2)} = n \left[(\hat{\theta}_n^A - \hat{\theta}_n)^2 p_A + (\hat{\theta}_n^B - \hat{\theta}_n)^2 p_B \right] \hat{\theta}_n^{-2}.$$

Sous H_0 les variables aléatoires $\zeta_n^{(1)}$ et $\zeta_n^{(2)}$ convergent en loi vers un χ^2 à $q - q' = 1$ degré de liberté. Le test de Hausman est défini par la région critique $W_n^k = \{\zeta_n^{(k)} \geq u\}$ pour $k \in \{1, 2\}$. Ce test est convergent de niveau asymptotique $\mathbb{P}(\chi^2(1) \geq u)$.

6. Soit $j \in \{A, B\}$. Comme les variables $X_1^j, \ldots, X_n^j$ sont indépendantes intégrables et de même loi, on déduit de la loi forte des grands nombres que la suite $(N_n^j/n, n \geq 1)$ converge presque sûrement vers $\mathbb{E}_{\theta^A, \theta^B}[X_1^j] = p^j$. Comme la matrice de Fisher est une fonction continue des paramètres p^A, p^B et θ^A, θ^B, on déduit des questions précédentes que la fonction

$$\hat{I}_n : (a, b) \mapsto \hat{I}_n(a, b) = \begin{pmatrix} N_n^A/na^2 & 0 \\ 0 & N_n^B/nb^2 \end{pmatrix}$$

est un estimateur convergent de la fonction $I : (a, b) \mapsto I(a, b)$.

7. On refait ici le même raisonnement que pour la dernière question de la partie précédente.

8. On a déjà calculé $\hat{\theta}_n^A = 4,507.\ 10^{-7}$. On calcule

$$\hat{\theta}_n^B = \frac{19}{64\,024\,591 + 15\,651\,648} = 2,385.\ 10^{-7} \quad \text{et} \quad \hat{\theta}_n = 3,066.\ 10^{-7}.$$

Il vient

$$\zeta_n^{(1)} = (4,507 - 3,066)^2 17/4,507^2 + (2,385 - 3,066)^2 19/2,385^2 = 3,3 \\ \zeta_n^{(2)} = (4,507 - 3,066)^2 17/3,066^2 + (2,385 - 3,066)^2 19/3,066^2 = 4,7.$$

Le quantile à 95% du χ^2 à 1 degré de liberté est $u = 3,84$. Donc $\zeta_n^{(2)}$ est dans la région critique mais pas $\zeta_n^{(1)}$. La différence est due à l'approximation de $I(\theta^A, \theta^B)$ par $\hat{I}_n(\hat{\theta}_n^A, \hat{\theta}_n^B)$ ou par $\hat{I}_n(\hat{\theta}_n, \hat{\theta}_n)$. On ne peut pas rejeter H_0 au niveau de confiance de 95%. Les pré-traitements A et B ne sont pas significativement différents.

9. Pour $j \in \{A, B\}$, on déduit de la partie précédente que l'intervalle de confiance asymptotique $1-\alpha^j$ pour θ^j est $J^j_{n^j} = \left[\hat{\theta}^j_{n^j} \pm \dfrac{z_\alpha\, \hat{\theta}^j_{n^j}}{\sqrt{N^j_{n^j}}}\right]$, avec z_α le quantile d'ordre $1-(\alpha/2)$ de la loi gaussienne $\mathcal{N}(0,1)$. Les variables aléatoires $(\hat{\theta}^A_{n^A}, N_{n^A})$ et $(\hat{\theta}^B_{n^B}, N_{n^B})$ sont indépendantes. On en déduit donc que

$$\mathbb{P}_{(\theta^A,\theta^B)}(\theta^A \in J^A_{n^A}, \theta^B \in J^B_{n^B}) = \mathbb{P}_{\theta^A}(\theta^A \in J^A_{n^A})\mathbb{P}_{\theta^B}(\theta^B \in J^B_{n^B}) = (1-\alpha^A)(1-\alpha^B).$$

Pour tout couple (α^A, α^B) tel que $(1-\alpha^A)(1-\alpha^B) = 95\%$, on obtient des intervalles de confiance $J^A_{n^A}$ pour θ^A et $J^B_{n^B}$ pour θ^B tels que la confiance asymptotique sur les deux intervalles soit de 95%. Si on choisit $\alpha = \alpha^A = \alpha^B = 1 - \sqrt{0,95} \simeq 2,5\%$, on a $z_\alpha \simeq 2,24$ et

$$J^A_{n^A} = \left[4,507 \pm \frac{2,24.\ 4,507}{\sqrt{17}}\right]\ 10^{-7} = [2.\ 10^{-7}; 7.\ 10^{-7}],$$
$$J^B_{n^B} = \left[2,385 \pm \frac{2,24.\ 2,385}{\sqrt{19}}\right]\ 10^{-7} = [1,2.\ 10^{-7}; 3,6.\ 10^{-7}].$$

Les deux intervalles $J^A_{n^A}$ et $J^B_{n^B}$ ne sont pas disjoints. On ne peut donc pas rejeter H_0. On retrouve le résultat de la question précédente. Ce test est toutefois moins puissant que les tests précédents.

IV Propriétés asymptotiques de l'estimateur $\hat{\theta}_n$

1. On a $p = \mathbb{P}_\theta(T \leq S) = \int \mathbf{1}_{\{t\leq s\}} f_\theta(t) g(s)\ dtds = \int f_\theta(t)\bar{G}(t)\ dt$.
2. Il vient par indépendance

$$\mathbb{P}_\theta(R > r) = \mathbb{P}_\theta(T > r, S > r) = \mathbb{P}_\theta(T > r)\mathbb{P}_\theta(S > r) = \bar{F}_\theta(r)\bar{G}(r).$$

Par dérivation de $\mathbb{P}_\theta(R \leq r) = 1 - \mathbb{P}_\theta(R > r)$, on en déduit la densité h de la loi de R :

$$h(r) = f_\theta(r)\bar{G}(r) + g(r)\bar{F}_\theta(r).$$

3. On a

$$\begin{aligned}\mathbb{E}_\theta[(R-w)^k_+] &= \int (r-w)^k_+ h(r)\ dr \\ &= \int (r-w)^k_+ f_\theta(r)\bar{G}(r)\ dr + \int (r-w)^k_+ g(r)\bar{F}_\theta(r)\ dr.\end{aligned}$$

À l'aide d'une intégration sur partie, il vient

$$\begin{aligned}\int (r-w)^k_+ g(r)\bar{F}_\theta(r)\ dr &= \Big[-\bar{G}(r)(r-w)^k_+ F_\theta(r)\Big]^{+\infty}_{-\infty} + \int k(r-w)^{k-1}_+ \bar{F}_\theta(r)\bar{G}(r)\ dr \\ &\qquad - \int (r-w)^k_+ f_\theta(r)\bar{G}(r)\ dr \\ &= \theta^{-1}\int f_\theta(r)\bar{G}(r)\ dr - \int (r-w)^k_+ f_\theta(r)\bar{G}(r)\ dr,\end{aligned}$$

où l'on a utilisé la relation $f_\theta(r) = \theta k(r-w)_+^{k-1}\bar{F}_\theta(r)$. On en déduit donc que

$$\mathbb{E}_\theta[(R-w)_+^k] = \theta^{-1}\int f_\theta(r)\bar{G}(r)\,dr = \theta^{-1}p.$$

4. Les variables aléatoires $(R_1-w)_+^k, \ldots, (R_n-w)_+^k$ sont indépendantes intégrables et de même loi. Par la loi forte des grands nombres, on en déduit que la suite $(Z_n = \frac{1}{n}\sum_{i=1}^n (R_i-w)_+^k, n \geq 1)$ converge presque sûrement vers $\mathbb{E}_\theta[(R-w)_+^k] = p/\theta$. De plus la suite $(N_n/n, n \geq 1)$ converge presque sûrement vers p. Comme la fonction $h(x,z) \mapsto x/z$ est continue sur $]0,+\infty[^2$, on en déduit que la suite $(\hat{\theta}_n = h(N_n/n, Z_n), n \geq 1)$ converge presque sûrement vers $h(p, p/\theta) = \theta$. L'estimateur $\hat{\theta}_n$ est donc un estimateur convergent.
5. En procédant comme pour le calcul de $\mathbb{E}_\theta[(R-w)_+^k]$, on a

$$\mathbb{E}_\theta[(R-w)_+^{2k}] = \int (r-w)_+^{2k} f_\theta(r)\bar{G}(r)\,dr + \int (r-w)_+^{2k} g(r)\bar{F}_\theta(r)\,dr.$$

À l'aide d'une intégration sur partie, il vient

$$\begin{aligned}\int (r-w)_+^{2k} g(r)\bar{F}_\theta(r)\,dr &= \left[-\bar{G}(r)(r-w)_+^{2k}F_\theta(r)\right]_{-\infty}^{+\infty} + \int 2k(r-w)_+^{2k-1}\bar{F}_\theta(r)\bar{G}(r)\,dr \\ &\qquad - \int (r-w)_+^{2k} f_\theta(r)\bar{G}(r)\,dr \\ &= 2\theta^{-1}\int (r-w)_+^k f_\theta(r)\bar{G}(r)\,dr - \int (r-w)_+^{2k} f_\theta(r)\bar{G}(r)\,dr.\end{aligned}$$

Comme

$$\beta = \mathbb{E}_\theta[\mathbf{1}_{\{X=1\}}(R-w)_+^k] = \int \mathbf{1}_{\{t\leq s\}}(t-w)_+^k f_\theta(t)g(s)\,dsdt = \int (t-w)_+^k f_\theta(t)\bar{G}(t)\,dt,$$

on déduit donc que

$$\mathbb{E}_\theta[(R-w)_+^{2k}] = 2\theta^{-1}\int (r-w)_+^k f_\theta(r)\bar{G}(r)\,dr = 2\beta/\theta.$$

6. La matrice de covariance du couple est

$$\Delta = \begin{pmatrix} p(1-p) & \beta - \frac{p^2}{\theta} \\ \beta - \frac{p^2}{\theta} & \frac{2\beta}{\theta} - \frac{p^2}{\theta^2}\end{pmatrix}.$$

7. Les variables $((R_1-w)_+^k, X_1), \ldots, ((R_n-w)_+^k, X_n)$ sont indépendantes de même loi et de carré intégrable. On en déduit donc que la suite $(\sqrt{n}(Z_n - p/\theta, N_n/n - p), n \geq 1)$ converge en loi vers une variable aléatoire gaussienne $\mathcal{N}(0,\Delta)$. La fonction $h(x,z) \mapsto x/z$ est de classe C^1 sur $]0,+\infty[^2$. On en déduit que la suite $(\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}(h(N_n/n, Z_n) - h(p, p/\theta^2)), n \geq 1)$ converge en loi vers une variable aléatoire $\mathcal{N}(0,\sigma^2)$, où

$$\begin{aligned}\sigma^2 &= (\frac{\partial h}{\partial x}(p,p/\theta), \frac{\partial h}{\partial z}(p,p/\theta))\Delta(\frac{\partial h}{\partial x}(p,p/\theta), \frac{\partial h}{\partial z}(p,p/\theta))' \\ &= (\theta/p, -\theta^2/p)\begin{pmatrix} p(1-p) & \beta - \frac{p^2}{\theta} \\ \beta - \frac{p^2}{\theta} & \frac{2\beta}{\theta} - \frac{p^2}{\theta^2}\end{pmatrix}(\theta/p, -\theta^2/p)' \\ &= \theta^2/p.\end{aligned}$$

L'estimateur $\hat{\theta}_n$ est donc un estimateur asymptotiquement normal de θ de variance asymptotique $\sigma^2 = \theta^2/p$. On remarque que la variance de l'estimateur des données censurées est supérieure à la variance de l'estimateur des données non censurées.

8. Comme $I(\theta)^{-1} = \sigma^2$, on en déduit que l'estimateur $\hat{\theta}_n$ est asymptotiquement efficace.

▲

Exercice XII.6.

I Modèle gaussien à variance connue

1. Dans un modèle gaussien avec variance connue, l'estimateur du maximum de vraisemblance de la moyenne est la moyenne empirique :

$$\hat{\nu}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

La moyenne empirique est un estimateur sans biais de la moyenne.

2. Le vecteur $(Y_1, \ldots, Y_n)$ est un vecteur gaussien car les variables aléatoires $Y_1, \ldots, Y_n$ sont gaussiennes et indépendantes. Comme $\hat{\nu}_n$ est une transformation linéaire du vecteur $(Y_1, \ldots, Y_n)$, sa loi est une loi gaussienne. On a $\mathbb{E}[\hat{\nu}_n] = \nu$ et $\mathrm{Var}(\hat{\nu}_n) = \dfrac{\sigma_0^2}{n}$. La loi de $\hat{\nu}_n$ est donc la loi $\mathcal{N}(\nu, \frac{\sigma_0^2}{n})$. En particulier la loi de $\sqrt{n}\dfrac{\hat{\nu}_n - \nu}{\sigma_0}$ est la loi $\mathcal{N}(0,1)$. On en déduit que

$$\mathbb{P}(\sqrt{n}\frac{\hat{\nu}_n - \nu}{\sigma_0} \in [-z_\alpha, z_\alpha]) = 1 - \alpha,$$

où z_α est le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi gaussienne centrée réduite. Donc

$$I_\alpha = [\hat{\nu}_n - z_\alpha \frac{\sigma_0}{\sqrt{n}}, \hat{\nu}_n + z_\alpha \frac{\sigma_0}{\sqrt{n}}]$$

est un intervalle de confiance exact de ν de niveau $1 - \alpha$.

Application numérique. On trouve $\hat{\nu}_n \simeq 6.42$, $z_\alpha \simeq 1.96$ et $I_\alpha \simeq [4.38,\ 8.46]$.

3. Les variables aléatoires $(Y_j, j \geq 1)$ sont indépendantes, de même loi et intégrable. On déduit de la loi forte des grands nombres que p.s. la suite $(\hat{\nu}_n, n \geq 1)$ converge vers ν. On en déduit que $\hat{\nu}_n$ est un estimateur convergent de ν.

4. Comme les variables aléatoires sont indépendantes, la vraisemblance et la log-vraisemblance du modèle complet s'écrivent :

$$p(x_1, \ldots, x_m, y_1, \ldots, y_n; \mu, \nu) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma_0^2}} \mathrm{e}^{-(x_i-\mu)^2/2\sigma_0^2} \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_0^2}} \mathrm{e}^{-(y_j-\nu)^2/2\sigma_0^2}$$

$$L(x_1, \ldots, x_m, y_1, \ldots, y_n; \mu, \nu) = -\frac{m+n}{2}\log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2}\sum_{i=1}^{m}(x_i - \mu)^2 - \frac{1}{2\sigma_0^2}\sum_{j=1}^{n}(y_j - \nu)^2.$$

5. La log-vraisemblance est somme d'une fonction de μ et d'une fonction de ν. Elle atteint son maximum quand chacune des deux fonctions atteint son maximum. Ces deux fonctions sont quadratiques négatives. Leur maximum est atteint pour les valeurs de μ et ν qui annulent leur dérivée, à savoir $\hat{\mu}_m(x_1,\dots,x_m) = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m} x_i$ et $\hat{\nu}_n(y_1,\dots,y_n) = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} y_j$. On en déduit les estimateurs du maximum de vraisemblance de (μ,ν) :
$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} X_i \quad \text{et} \quad \hat{\nu}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$
On retrouve bien le même estimateur pour ν.

6. Comme les variables aléatoires $X_1,\dots,X_m,Y_1,\dots,Y_n$ sont indépendantes, on en déduit que $\hat{\mu}_m$ et $\hat{\nu}_n$ sont indépendants. Les mêmes arguments que ceux de la question 2, assurent que la loi de $\hat{\mu}_m$ est la loi $\mathcal{N}(\mu, \frac{\sigma_0^2}{m})$. Le vecteur $(\hat{\mu}_m,\hat{\nu}_n)$ est donc un vecteur gaussien de moyenne (μ,ν) et de matrice de covariance $\sigma_0^2\begin{pmatrix} 1/m & 0 \\ 0 & 1/n \end{pmatrix}$.

7. La variable aléatoire $T^{(1)}_{m,n}$ est une transformation linéaire du vecteur gaussien $(\hat{\mu}_m,\hat{\nu}_n)$. Il s'agit donc d'une variable aléatoire gaussienne. On calcule
$$\begin{aligned}\mathbb{E}[T^{(1)}_{m,n}] &= \sqrt{\frac{mn}{m+n}}\frac{\mu-\nu}{\sigma_0},\\ \mathrm{Var}(T^{(1)}_{m,n}) &= \frac{mn}{m+n}\frac{\mathrm{Var}(\hat{\mu}_m)+\mathrm{Var}(\hat{\nu}_n)}{\sigma_0^2}\\ &= 1.\end{aligned}$$
On a utilisé l'indépendance de $\hat{\mu}_m$ et $\hat{\nu}_n$ pour écrire $\mathrm{Var}(\hat{\mu}_m-\hat{\nu}_n) = \mathrm{Var}(\hat{\mu}_m)+\mathrm{Var}(\hat{\nu}_n)$. En particulier, sous H_0, la loi de $T^{(1)}_{m,n}$ est la loi gaussienne centrée réduite $\mathcal{N}(0,1)$.

8. De même qu'à la question 3, l'estimateur $\hat{\mu}_m$ est un estimateur convergent de μ. Remarquons enfin que
$$\frac{mn}{m+n} = \min(m,n)\,\frac{1}{1+\frac{\min(m,n)}{\max(m,n)}} \geq \frac{\min(m,n)}{2}.$$
En particulier
$$\lim_{\min(m,n)\to\infty} \frac{mn}{m+n} = +\infty.$$
On en déduit donc que la suite de vecteurs $((\hat{\mu}_m,\hat{\nu}_n,\sqrt{\frac{mn}{m+n}}), m\geq 1, n\geq 1)$ converge presque sûrement vers $(\mu,\nu,+\infty)$ quand $\min(m,n)\to\infty$. Si $\mu>\nu$, alors presque sûrement, on a
$$\lim_{\min(m,n)\to\infty} T^{(1)}_{m,n} = +\infty.$$

9. On définit la fonction sur $\mathbb{R}^{m+n}$ (cf la question 5) :

$$t^{(1)}_{m,n}(x_1,\dots,x_m,y_1,\dots,y_n) = \sqrt{\frac{mn}{m+n}}\frac{\frac{1}{m}\sum_{i=1}^m x_i - \frac{1}{n}\sum_{j=1}^n y_j}{\sigma_0}$$
$$= \sqrt{\frac{mn}{m+n}}\frac{\hat\mu_m(x_1,\dots,x_m) - \hat\nu_n(y_1,\dots,y_n)}{\sigma_0}.$$

En particulier, on a $t^{(1)}_{m,n}(X_1,\dots,X_m,Y_1,\dots,Y_n) = T^{(1)}_{m,n}$. On considère le test pur de région critique

$$W_{m,n} = \{(x_1,\dots,x_m,y_1,\dots,y_n)\in\mathbb{R}^{m+n}; t^{(1)}_{m,n}(x_1,\dots,x_m,y_1,\dots,y_n) > z_\alpha\},$$

avec $z_\alpha\in\mathbb{R}$. Déterminons z_α pour que le test soit de niveau exact α. Comme la loi de $T^{(1)}_{m,n}$ est la loi $\mathcal{N}(0,1)$ sous H_0, l'erreur de première espèce est donc :

$$\mathbb{P}(W_{m,n}) = \mathbb{P}((X_1,\dots,X_m,Y_1,\dots,Y_n)\in W_{m,n}) = \mathbb{P}(T^{(1)}_{m,n} > z_\alpha) = \int_{z_\alpha}^\infty \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\,du.$$

Pour que le test soit de niveau exact α, il faut que $\int_{z_\alpha}^\infty \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\,du = \alpha$ et donc on choisit pour z_α le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$.
Vérifions que le test est convergent. D'après la question précédente, on a sous H_1

$$\lim_{\min(m,n)\to\infty} \mathbf{1}_{\{T^{(1)}_{m,n}>z_\alpha\}} = 1.$$

Par convergence dominée, on en déduit que sous H_1,

$$\lim_{\min(m,n)\to\infty}\mathbb{P}(W_{m,n}) = \lim_{\min(m,n)\to\infty}\mathbb{P}(T^{(1)}_{m,n}>z_\alpha) = \lim_{\min(m,n)\to\infty}\mathbb{E}[\mathbf{1}_{\{T^{(1)}_{m,n}>z_\alpha\}}] = 1.$$

Le test est donc convergent.

10. *Application numérique.* On trouve $\hat\mu_m\simeq 6.71$, $\hat\nu_n\simeq 6.42$, $t^{(1)}_{m,n}\simeq 0.187$ et $z_\alpha\simeq 1.64$. Comme $t^{(1)}_{m,n}\le z_\alpha$, on accepte H_0. On rejette H_0 en dès que $z_\alpha < t^{(1)}_{m,n}$. Comme $\alpha = \int_{z_\alpha}^\infty \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\,du$, on en déduit que le maximum des valeurs de α qui permette d'accepter H_0 est

$$p_1 = \int_{t^{(1)}_{m,n}}^\infty \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\,du.$$

On trouve $p_1\simeq 0.426$.

II Modèle non paramétrique de décalage

Remarquons que sous H_0, on a $p = \int \mathbf{1}_{\{y\le x\}} f(x)f(y)\,dxdy = 1/2$.

Vérifions que sous H_1, on a $p > 1/2$. Soit $\rho>0$. Comme la densité f est non nulle, il existe $\rho/2>\varepsilon>0$ et $x_0\in\mathbb{R}$ tel que $\int_{[x_0-\varepsilon,x_0]} f(x)\,dx > 0$ et $\int_{[x_0,x_0+\varepsilon]} f(x)\,dx > 0$. On en déduit alors que pour tout $x\in[x_0-\varepsilon,x_0]$,

$$F(x+\rho) = F(x) + \int_{[x,x+\rho]} f(x')\,dx' \ge F(x) + \int_{[x_0,x_0+\varepsilon]} f(x')\,dx' > F(x).$$

Comme $\int_{[x_0-\varepsilon,x_0]} f(x)\,dx > 0$, on en déduit que

$$\int_{[x_0-\varepsilon,x_0]} F(x+\rho)f(x)\,dx > \int_{[x_0-\varepsilon,x_0]} F(x)f(x)\,dx.$$

Comme de plus $F(x+\rho) \geq F(x)$ pour tout $x \in \mathbb{R}$, on en déduit que $\int F(x+\rho)f(x)\,dx > \int F(x)f(x)\,dx$. Donc sous H_1, on a

$$\begin{aligned} p &= \int \mathbf{1}_{\{y\leq x\}} f(x)g(y)\,dxdy \\ &= \int \mathbf{1}_{\{y\leq x\}} f(x)f(y+\rho)\,dxdy \\ &= \int F(x+\rho)f(x)\,dx \\ &> \int F(x)f(x)\,dx \\ &= \frac{1}{2}. \end{aligned}$$

On en déduit donc que les valeurs possibles de p sous H_1 sont $]\frac{1}{2}, 1]$.

Enfin, nous renvoyons à la correction du problème concernant l'étude de la statistique de Mann et Withney, pour vérifier que α' et β' sont positifs, et si $p \notin \{0,1\}$, alors parmi α' et β', au moins un des deux termes est strictement positif.

1. On a

$$\begin{aligned} \{T^{(2)}_{m,n} > a\} &= \{\frac{U_{m,n} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} > a\} \\ &= \{\frac{U_{m,n} - mnp}{\sqrt{\mathrm{Var}(U_{m,n})}} > \sqrt{\frac{mn(m+n+1)}{12\,\mathrm{Var}(U_{m,n})}}a - \frac{mn(p-\frac{1}{2})}{\sqrt{\mathrm{Var}(U_{m,n})}}\}. \end{aligned}$$

On pose

$$b_{m,n} = \sqrt{\frac{mn(m+n+1)}{12\,\mathrm{Var}(U_{m,n})}}a - \frac{mn(p-\frac{1}{2})}{\sqrt{\mathrm{Var}(U_{m,n})}}.$$

Comme $p > 1/2$, on en déduit que $\lim_{k\to\infty} \frac{m_k n_k (p-\frac{1}{2})}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}} = +\infty$. De plus comme au moins l'un des deux terme α' ou β' est strictement positif, cela implique la convergence de la suite $(\frac{m_k n_k(m_k+n_k+1)}{12\,\mathrm{Var}(U_{m_k,n_k})}, k \geq 1)$ vers une limite finie. On a donc

$$\lim_{k\to\infty} b_{m_k,n_k} = -\infty.$$

2. De la question précédente, on en déduit que pour tout réel $M > 0$, il existe $k_0 \geq 1$ tel que pour tout $k \geq k_0$, $b_{m_k,n_k} < -M$. Cela implique que pour tout $k \geq k_0$,

$$\mathbb{P}(T^{(2)}_{m_k,n_k} > a) \geq \mathbb{P}(\frac{U_{m_k,n_k} - m_k n_k p}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}} > -M).$$

Grâce à la convergence en loi de $(\frac{U_{m_k,n_k} - m_k n_k p}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}}, k \geq 1)$ vers la loi continue $\mathcal{N}(0,1)$, on en déduit que

$$\lim_{k\to\infty} \mathbb{P}(\frac{U_{m_k,n_k} - m_k n_k p}{\sqrt{\mathrm{Var}(U_{m_k,n_k})}} > -M) = \int_{-M}^{\infty} \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\, du.$$

En particulier on a pour tout M, $\liminf_{k\to\infty} \mathbb{P}(T^{(2)}_{m_k,n_k} > a) \geq \int_{-M}^{\infty} \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\, du$. Le choix de M étant arbitraire, cela implique donc que $\lim_{k\to\infty} \mathbb{P}(T^{(2)}_{m_k,n_k} > a) = 1$.

3. On définit les fonctions

$$u_{m,n}(x_1,\ldots,x_m,y_1,\ldots,y_n) = \sum_{i=1}^{m}\sum_{j=1}^{n} \mathbf{1}_{\{y_j \leq x_i\}},$$
$$t^{(3)}_{m,n}(x_1,\ldots,x_m,y_1,\ldots,y_n) = \frac{u_{m,n}(x_1,\ldots,x_m,y_1,\ldots,y_n) - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}.$$

En particulier, on a $t^{(2)}_{m,n}(X_1,\ldots,X_m,Y_1,\ldots,Y_n) = T^{(2)}_{m,n}$. On considère le test pur de région critique

$$W_{m,n} = \{(x_1,\ldots,x_m,y_1,\ldots,y_n) \in \mathbb{R}^{m+n}; t^{(2)}_{m,n}(x_1,\ldots,x_m,y_1,\ldots,y_n) > z_\alpha\},$$

avec $z_\alpha \in \mathbb{R}$. Déterminons z_α pour que le test soit de niveau asymptotique α. L'erreur de première espèce est :

$$\mathbb{P}(W_{m,n}) = \mathbb{P}((X_1,\ldots,X_m,Y_1,\ldots,Y_n) \in W_{m,n}) = \mathbb{P}(T^{(2)}_{m,n} > z_\alpha).$$

Comme la loi asymptotique de la statistique $T^{(2)}_{m,n}$ est la loi $\mathcal{N}(0,1)$ sous H_0, l'erreur de première espèce converge vers $\int_{z_\alpha}^{\infty} \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\, du$. Pour que le test soit de niveau asymptotique α, on choisit pour z_α le quantile d'ordre $1-\alpha$ de la loi $\mathcal{N}(0,1)$. Le test est convergent d'après la question précédente.

4. *Application numérique.* On a $u_{m,n} = 70$ et $t^{(2)}_{m,n} \simeq 0.25$. et $z_\alpha \simeq 1.64$. Comme $t^{(2)}_{m,n} \leq z_\alpha$, on accepte H_0. On rejette H_0 dès que $z_\alpha < t^{(2)}_{m,n}$. Comme $\alpha = \int_{z_\alpha}^{\infty} \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\, du$, on en déduit que le maximum des valeurs de α qui permette d'accepter H_0 est

$$p_2 = \int_{t^{(2)}_{m,n}}^{\infty} \frac{\mathrm{e}^{-u^2/2}}{\sqrt{2\pi}}\, du.$$

On trouve $p_2 \simeq 0.403$.

Remarquons que les résultats sont identiques si l'on considère la même hypothèse nulle H_0 contre l'hypothèse alternative $H_1 = \{\mathbb{P}(Y \leq X) > 1/2\}$, car les réponses aux questions 1 et 2, et par conséquent les réponses aux questions 3 et 4, de cette partie sont identiques.

IV Modèle non paramétrique général

1. Il s'agit du test de Kolmogorov Smirnov à deux échantillons. C'est un test pur de région critique

$$W_{m,n} = \{(x_1, \dots, x_m, y_1, \dots, y_n) \in \mathbb{R}^{m+n}; t^{(3)}_{m,n}(x_1, \dots, x_m, y_1, \dots, y_n) > z_\alpha\},$$

avec

$$t^{(3)}_{m,n}(x_1, \dots, x_m, y_1, \dots, y_n) = \sqrt{\frac{mn}{m+n}} \sup_{x\in\mathbb{R}} \left| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\{x_i \le x\}} - \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{\{y_j \le x\}} \right|$$

et $z_\alpha \ge 0$.

2. Comme le test est de niveau asymptotique α, on en déduit que z_α est tel que $K(z_\alpha) = 1 - \alpha$, où la fonction K est définie par

$$K(z) = \sum_{k=-\infty}^{+\infty} (-1)^k \, \mathrm{e}^{-2k^2 z^2}.$$

On trouve $t^{(3)}_{m,n} \simeq 0.508$ et $z_\alpha \simeq 1.358$. Comme $t^{(3)}_{m,n} \le z_\alpha$, on accepte H_0. On rejette H_0 dès que $z_\alpha < t^{(3)}_{m,n}$. Comme $\alpha = 1 - K(z_\alpha)$, on en déduit que le maximum des valeurs de α qui permette d'accepter H_0 est

$$p_3 = 1 - K(t^{(3)}_{m,n}).$$

On trouve $p_3 \simeq 0.958$. On ne peut vraiment pas rejeter H_0.

Conclusion

On a $p_1 \simeq p_2 < p_3$. On remarque que pour le modèle le plus général (modèle 3) la p-valeur est très élevées. En revanche, plus on fait d'hypothèse et plus la p-valeur est faible. Dans tous les cas on ne peut pas rejeter H_0 sans commettre une erreur de première espèce supérieure à 40%. Cela signifie que les données de l'expérience pratiquée en Floride reproduitent partiellement ici ne permettent pas de conclure à l'efficacité de l'ensemencement des nuages par iodure d'argent. ▲

Exercice XII.7.

I Le modèle

1. L'hypothèse nulle est $H_0 = \{$*La densité osseuse chez les femmes du groupe 1 n'est pas significativement supérieure à celle des femmes du groupe 2*$\}$ et l'hypothèse alternative $H_1 = \{$*La densité osseuse chez les femmes du groupe 1 est significativement supérieure à celle des femmes du groupe 2*$\}$. On espère rejetter H_0, en contrôlant l'erreur de première espèce.
2. Les variables aléatoires $X_1, \dots, X_n, Y_1, \dots, Y_m$ sont indépendantes de loi gaussienne. On en déduit que $(X_1, \dots, X_n, Y_1, \dots, Y_m)$ est un vecteur gaussien.

3. La loi de $\bar{X}_n$ est la loi gaussienne $\mathcal{N}(\mu_1, \sigma_1^2/n)$, la loi de $\dfrac{n-1}{\sigma_1^2}V_n$ est la loi $\chi^2(n-1)$, enfin ces deux variables sont indépendantes. Cela détermine complètement la loi du couple $\left(\bar{X}_n, \dfrac{n-1}{\sigma_1^2}V_n\right)$.
4. Comme les variables $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ sont indépendantes, on en déduit que les variables $\dfrac{n-1}{\sigma_1^2}V_n$ et $\dfrac{m-1}{\sigma_2^2}W_m$ sont indépendantes et suivent les lois du χ^2 de degrés de liberté $n-1$ et $m-1$. En considérant les fonctions caractéristiques, on obtient en utilisant l'indépendance,

$$\begin{aligned}\psi_{\frac{n-1}{\sigma_1^2}V_n+\frac{m-1}{\sigma_2^2}W_m}(u) &= \psi_{\frac{n-1}{\sigma_1^2}V_n}(u)\ \psi_{\frac{m-1}{\sigma_2^2}W_m}(u)\\ &= \frac{1}{(1-2iu)^{(n-1)/2}}\ \frac{1}{(1-2iu)^{(m-1)/2}}\\ &= \frac{1}{(1-2iu)^{(n+m-2)/2}}\,.\end{aligned}$$

   On en déduit donc que la loi de $\dfrac{n-1}{\sigma_1^2}V_n + \dfrac{m-1}{\sigma_2^2}W_m$ est la loi du χ^2 à $n+m-2$ degrés de liberté.
5. Comme $\mathbb{E}_\theta[\bar{X}_n] = \mu_1$ (resp. $\mathbb{E}_\theta[\bar{Y}_m] = \mu_2$), on en déduit que $\bar{X}_n$ (resp. $\bar{Y}_m$) est un estimateur sans biais de μ_1 (resp. μ_2). Par la loi forte des grands nombres, les estimateurs $(\bar{X}_n, n \geq 1)$ et $(\bar{Y}_m, m \geq 1)$ sont convergents.
6. Comme $\mathbb{E}_\theta[V_n] = \sigma_1^2$ (resp. $\mathbb{E}_\theta[W_m] = \sigma_2^2$), on en déduit que V_n (resp. W_m) est un estimateur sans biais de σ_1^2 (resp. σ_2^2). Par la loi forte des grands nombres, les suites $(n^{-1}\sum_{i=1}^n X_i^2, n \geq 1)$ et $(\bar{X}_n, n \geq 1)$ converge $\mathbb{P}_\theta$-p.s. vers $\mathbb{E}_\theta[X_1^2]$ et $\mathbb{E}_\theta[X_1]$. On en déduit donc que la suite $(V_n = \frac{n}{n-1}(n^{-1}\sum_{i=1}^n X_i^2 - \bar{X}_n^2), n \geq 2)$ converge $\mathbb{P}_\theta$-p.s. vers $\mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2 = \mathrm{Var}_\theta(X_1) = \sigma_1^2$. La suite d'estimateurs $(V_n, n \geq 2)$ est donc convergente. De même, on vérifie que la suite d'estimateurs $(W_m, m \geq 2)$ est convergente.

II Simplification du modèle

1. On a sous H_0,
$$Z_{n,m} = \frac{(n-1)V_n/\sigma_1^2}{n-1}\frac{m-1}{(m-1)W_m/\sigma_2^2}.$$
   On déduit donc de I.4, que la loi de $Z_{n,m}$ sous H_0 est la loi de Fisher-Snedecor de paramètre $(n-1, m-1)$.
2. Comme $(V_n, n \geq 2)$ et $(W_m, m \geq 2)$ sont des estimateurs convergents de σ_1^2 et σ_2^2, la suite $(Z_{n,m}, n \geq 1, m \geq 1)$ converge $\mathbb{P}_\theta$-p.s. vers σ_1^2/σ_2^2 quand $\min(n,m) \to \infty$. Sous H_0 cette limite vaut 1. Sous H_1 cette limite est différente de 1.
3. Sous H_1, on a $\mathbb{P}_\theta$-p.s. $\lim_{\min(n,m)\to\infty} Z_{n,m} \neq 1$. En particulier, on en déduit que sous H_1, $\mathbb{P}_\theta$-p.s.
$$\lim_{\min(n,m)\to\infty} \mathbf{1}_{\{a_{n-1,m-1,\alpha_1} < Z_{n,m} < b_{n-1,m-1,\alpha_2}\}} = 0.$$
   On note la région critique
$$\tilde{W}_{n,m} = \{Z_{n,m} \notin]a_{n-1,m-1,\alpha_1}, b_{n-1,m-1,\alpha_2}[\}.$$

Par convergence dominée, on en déduit que sous H_1, $\lim_{\min(n,m)\to\infty} \mathbb{P}_\theta(\tilde{W}_{n,m}) = 1$. Le test est donc convergent. L'erreur de première espèce associée à la région critique $\tilde{W}_{n,m}$ est pour $\theta \in H_0$,

$$\begin{aligned}\mathbb{P}_\theta(\tilde{W}_{n,m}) &= \mathbb{P}_\theta(Z_{n,m} \leq a_{n-1,m-1,\alpha_1}) + \mathbb{P}_\theta(Z_{n,m} \geq b_{n-1,m-1,\alpha_2}) \\ &= \mathbb{P}(F_{n-1,m-1} \leq a_{n-1,m-1,\alpha_1}) + \mathbb{P}(F_{n-1,m-1} \geq b_{n-1,m-1,\alpha_2}) \\ &= \alpha_1 + \alpha_2,\end{aligned}$$

où l'on a utilisé le fait que $a_{n-1,m-1,\alpha_1} \leq b_{n-1,m-1,\alpha_2}$ pour la première égalité, le fait que sous H_0, $Z_{n,m}$ a même loi que $F_{n-1,m-1}$ pour la deuxième égalité, et la définition de $a_{n-1,m-1,\alpha_1}$ et $b_{n-1,m-1,\alpha_2}$ pour la troisième égalité. En particulier le niveau du test est donc $\alpha = \alpha_1 + \alpha_2$. Il est indépendant de n, m et de $\theta \in H_0$.

4. On a donc $\alpha_1 = \alpha_2 = 2.5\%$. On en déduit donc, avec $n = 25$ et $m = 31$, que $a_{n-1,m-1,\alpha_1} = 0.453$ et $b_{n-1,m-1,\alpha_2} = 2.136$. La région critique est donc

$$[0, 0.453] \cup [2.136, +\infty[.$$

La valeur observée de $Z_{n,m}$ est 0.809 (à 10^{-3} près), elle n'appartient pas à la région critique. On accepte donc H_0.

5. La p-valeur du test est la plus grande valeur de α pour laquelle on accepte H_0. On accepte H_0 si $a_{n-1,m-1,\alpha_1} < 0.809$, donc si $\alpha_1 < 0.3$ et donc si $\alpha = 2\alpha_1 < 0.6$. La p-valeur est donc de 0.6. Il est tout à fait raisonnable d'accepter H_0. La p-valeur est très supérieure aux valeurs critiques classiques (telles que 0.1, 0.05 ou 0.01).

6. Soit $\varepsilon \in]0, 1[$. Comme $F_{n,m}$ a même loi que $Z_{n+1,m+1}$ sous H_0, on a

$$\mathbb{P}(F_{n,m} \leq 1 - \varepsilon) = \mathbb{P}(Z_{n+1,m+1} \leq 1 - \varepsilon).$$

On a vu que sous H_0, $\lim_{\min(n,m)\to\infty} Z_{n,m} = 1$. Par convergence dominée, on en déduit que

$$\lim_{\min(n,m)\to\infty} \mathbb{P}(Z_{n,m} \leq 1 - \varepsilon) = 0.$$

On a donc $\lim_{\min(n,m)\to\infty} \mathbb{P}(F_{n,m} \leq 1 - \varepsilon) = 0$. En particulier, pour $\min(n, m)$ assez grand on a $\mathbb{P}(F_{n,m} \leq 1-\varepsilon) < \alpha_1$, ce qui implique que $a_{n,m,\alpha_1} > 1-\varepsilon$. Un raisonnement similaire assure que $\lim_{\min(n,m)\to\infty} \mathbb{P}(F_{n,m} \geq 1 + \varepsilon) = 0$, et donc $b_{n,m,\alpha_2} < 1 + \varepsilon$ pour $\min(n, m)$ assez grand. Comme $a_{n,m,\alpha_1} < b_{n,m,\alpha_2}$, on en déduit que

$$1 - \varepsilon < \liminf_{\min(n,m)\to\infty} a_{n,m,\alpha_1} \leq \limsup_{\min(n,m)\to\infty} b_{n,m,\alpha_2} < 1 + \varepsilon.$$

Comme $\varepsilon \in]0, 1[$ est arbitraire, on en déduit que

$$\lim_{\min(n,m)\to\infty} a_{n,m,\alpha_1} = \lim_{\min(n,m)\to\infty} b_{n,m,\alpha_2} = 1.$$

III Comparaison de moyenne

1. La loi de $\frac{n+m-2}{\sigma^2}S_{n,m}$ est d'après I.4 la loi du χ^2 à $n+m-2$ degrès de liberté. On a $\mathbb{P}_\theta$-p.s.
$$\lim_{\min(n,m)\to\infty} V_n = \lim_{\min(n,m)\to\infty} W_m = \sigma^2.$$
On en déduit donc que
$$\lim_{\min(n,m)\to\infty} S_{n,m} = \sigma^2.$$

2. On déduit de I.2 et de I.3 que les variables $\bar{X}_n$ et $\bar{Y}_m$ sont indépendantes de loi gaussienne respective $\mathcal{N}(\mu_1, \sigma^2/n)$ et $\mathcal{N}(\mu_2, \sigma^2/m)$. En particulier $(\bar{X}_n, \bar{Y}_m)$ forme un vecteur gaussien. Ainsi $\frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m)$, transformation linéaire du vecteur gaussien $(\bar{X}_n, \bar{Y}_m)$, suit une loi gaussienne de moyenne
$$\mathbb{E}_\theta\left[\frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m)\right] = \frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}(\mu_1 - \mu_2),$$
et de variance, en utilisant l'indépendance entre $\bar{X}_n$ et $\bar{Y}_m$,
$$\mathrm{Var}_\theta\left(\frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m)\right) = \frac{1}{\sigma^2}\frac{nm}{n+m}(\frac{\sigma^2}{n} + \frac{\sigma^2}{m}) = 1.$$

3. D'après I.3, les variables aléatoires $S_{n,m}$ et $\frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m)$ sont indépendantes. La loi de $(n+m-2)S_{n,m}/\sigma^2$ est la loi du χ^2 à $n+m-2$ degrès de liberté. De plus, sous H_0, $\frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m)$ est une gaussienne $\mathcal{N}(0,1)$. On en déduit que sous H_0,
$$T_{n,m} = \frac{1}{\sigma}\sqrt{\frac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m)\sqrt{\frac{n+m-2}{(n+m-2)S_{n,m}/\sigma^2}}$$
suit une loi de Student de paramètre $n+m-2$.

4. Sous H_1, la limite de la suite $(\bar{X}_n - \bar{Y}_m, n \geq 1, m \geq 1)$ quand $\min(n,m) \to \infty$ est $\mu_1 - \mu_2$ $\mathbb{P}_\theta$-p.s. d'après I.5.

5. On déduit de ce qui précède, que sous H_1, la suite $((\bar{X}_n - \bar{Y}_m)/\sqrt{S_{n,m}}, n \geq 2, m \geq 2)$ converge quand $\min(n,m) \to \infty$ vers $(\mu_1 - \mu_2)/\sigma$ $\mathbb{P}_\theta$-p.s. De plus, on a
$$\lim_{\min(n,m)\to\infty}\sqrt{\frac{nm}{n+m}} = \lim_{\min(n,m)\to\infty}\frac{1}{\sqrt{\frac{1}{n}+\frac{1}{m}}} = +\infty.$$
On en déduit que sous H_1, $\mathbb{P}_\theta$-p.s., la suite $(T_{n,m}, n \geq 2, m \geq 2)$ converge quand $\min(n,m) \to \infty$ vers $+\infty$ car $\mu_1 > \mu_2$.

6. On considère la région critique
$$W_{n,m} = \{T_{n,m} > c_{n+m-2,\alpha}\},$$

où $c_{k,\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi de Student de paramètre k. Comme la suite $(S_{n,m}, n \geq 2, m \geq 2)$ converge p.s. vers σ^2 quand $\min(n,m) \to \infty$, on déduit du théorème de Slutsky que la suite $(T_{n,m}, n \geq 2, m \geq 2)$ converge en loi, sous H_0, vers G de loi gaussienne centrée réduite quand $\min(n,m) \to \infty$. Comme G est une variable continue, on en déduit que pour tout $c \in \mathbb{R}$,

$$\lim_{\min(n,m)\to\infty} \mathbb{P}(T_{n,m} > c) = \mathbb{P}(G > c).$$

En particulier, cela implique que la suite $(c_{n+m-2,\alpha}, n \geq 2, m \geq 2)$ converge vers c_α le quantile d'ordre $1-\alpha$ de G. Cette suite est donc majorée par une constant, disons c. Par convergence dominée, on déduit de la réponse à la question précédente que sous H_1,

$$\liminf_{\min(n,m)\to\infty} \mathbb{P}_\theta(T_{n,m} > c_{n+m-2,\alpha}) \geq \liminf_{\min(n,m)\to\infty} \mathbb{P}_\theta(T_{n,m} > c) = 1.$$

Donc on a

$$\lim_{\min(n,m)\to\infty} \mathbb{P}_\theta(W_{n,m}) = 1,$$

et le test est convergent. De plus l'erreur de première espèce associée est, pour $\theta \in H_0$,

$$\mathbb{P}_\theta(W_{n,m}) = \mathbb{P}_\theta(T_{n,m} > c_{n+m-2,\alpha}) = \alpha.$$

Le niveau de ce test est donc α.

7. On obtient pour $\alpha = 5\%$, $c_{n+m-2,\alpha} = 1.674$. La région critique est donc $[1.674, \infty[$. La valeur observée de $T_{n,m}$ est 3.648 (à 10^{-3} près), elle appartient à la région critique. On rejette donc H_0.
8. La p-valeur du test est la plus grande valeur de α pour laquelle on accepte H_0. On accepte H_0 si $c_{n+m-2,\alpha} > 3.648$, ce qui correspond à α compris entre 2.5/10 000 et 5/10 000 (en fait la p-valeur est égale à 0.0003). Il est tout à fait raisonnable de rejeter H_0. La p-valeur est très inférieure aux valeurs critiques classiques (telles que 0.1, 0.05 ou 0.01).

IV Variante sur les hypothèses du test

1. On déduit de III.5 que sous H_1', $\mathbb{P}_\theta$-p.s., la suite $(T_{n,m}, n \geq 2, m \geq 2)$ converge quand $\min(n,m) \to \infty$ vers $+\infty$ si $\mu_1 > \mu_2$, et vers $-\infty$ si $\mu_1 < \mu_2$.
2. On considère la région critique

$$W'_{n,m} = \{|T_{n,m}| > c_{n+m-2,\alpha/2}\},$$

où $c_{k,\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student de paramètre k. Par convergence dominée, on déduit de la réponse à la question précédente que sous H_1',

$$\lim_{\min(n,m)\to\infty} \mathbb{P}_\theta(W'_{n,m}) = 1.$$

Le test est donc convergent. De plus l'erreur de première espèce associée est pour $\theta \in H_0$

$$\mathbb{P}_\theta(W'_{n,m}) = \mathbb{P}_\theta(|T_{n,m}| > c_{n+m-2,\alpha/2}) = \alpha.$$

Le niveau de ce test est donc α. Pour $\alpha = 5\%$, on obtient $c_{n+m-2,\alpha/2} = 2.005$. La p-valeur du test est la plus grande valeur de α pour laquelle on accepte H_0. On accepte H_0 si $c_{n+m-2,\alpha/2} > 3.648$, ce qui correspond à α compris entre 5/10 000 et 10/10 000 (en fait la p-valeur est égale à 0.0006). Il est tout à fait raisonnable de rejeter H_0. La p-valeur est très inférieure aux valeurs critiques classiques (telles que 0.1, 0.05 ou 0.01).

3. Sous $\mathbb{P}_{(\mu_1,\sigma,\mu_2,\sigma)}$, $\bar{X}_n - \bar{Y}_m$ a même loi que $\bar{X}_n - \bar{Y}_m + \mu_1 - \mu_2$ sous $\mathbb{P}_{(0,\sigma,0,\sigma)}$. On en déduit donc que sous H_0',

$$\begin{aligned}
\mathbb{P}_{(\mu_1,\sigma,\mu_2,\sigma)}\left(T_{n,m} > c\right) &= \mathbb{P}_{(0,\sigma,0,\sigma)}\left(T_{n,m} + \sqrt{\frac{nm}{n+m}}\frac{\mu_1-\mu_2}{\sqrt{S_{n,m}}} > c\right) \\
&= \mathbb{P}_{(0,\sigma,0,\sigma)}\left(T_{n,m} > c - \sqrt{\frac{nm}{n+m}}\frac{\mu_1-\mu_2}{\sqrt{S_{n,m}}}\right) \\
&\leq \mathbb{P}_{(0,\sigma,0,\sigma)}\left(T_{n,m} > c\right)
\end{aligned}$$

car $\mu_1 \leq \mu_2$ pour la dernière égalité. En particulier, pour $c = c_{n+m-2,\alpha}$, l'erreur de première espèce du test pur de région critique $W_{n,m} = \{T_{n,m} > c\}$ est donc majorée par $\mathbb{P}_{(0,\sigma,0,\sigma)}\left(T_{n,m} > c_{n+m-2,\alpha}\right)$. Or cette dernière quantité est égale à α d'après III.6. De plus ce test est convergent d'après III.6. En conclusion, le test pur de III.6 est un test convergent de niveau α pour tester H_0' contre H_1. En particulier, on rejette H_0', avec une p-valeur égale à 3/10 000.

▲

Index