

Collection
Les cours

Introduction aux probabilités et à la statistique

Exercices et problèmes corrigés

Jean-François Delmas



LES PRESSES DE L'ENSTA

Introduction
au calcul des probabilités
et à la statistique

Exercices, problèmes et corrections

Responsables de publication : Thomas Loiseleux et Sophie Chouaf

En application du Code de la Propriété Intellectuelle et notamment de ses articles L. 122.4, L. 122-5 et L. 335-2, toute représentation ou reproduction intégrale ou partielle faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite. Une telle représentation ou reproduction constituerait un délit de contrefaçon, puni de trois ans d'emprisonnement et de 300 000 euros d'amende.

Ne sont autorisés que les copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective, ainsi que les analyses et courtes citations, sous réserve que soient indiqués clairement le nom de l'auteur et la source.

© *Les Presses de l'ENSTA*, 2012

Imprimé en France

ISBN 978-2-7225-0937-5

www.ensta-paristech.fr

Les cours

Jean-François Delmas

**Introduction
au calcul des probabilités
et à la statistique**

Exercices, problèmes et corrections

PARIS
LES PRESSES DE L'ENSTA
32, boulevard Victor, Paris 15^e
2012

À mes parents et grand-parents.

Préface

Ce livre d'exercices et de problèmes corrigés s'appuie sur le texte de cours *Introduction aux probabilités et à la statistique*. Il reprend, avec leur correction détaillée, les exercices et les problèmes proposés aux élèves de première année à l'École Nationale Supérieure des Techniques Avancées depuis 1999. L'accent est mis sur la modélisation et les applications.

Gardant la structure du texte de cours, les exercices sont regroupés en chapitres, avec en probabilités : les espaces probabilisés I (probabilités discrètes, dénombrement, probabilités conditionnelles et indépendance), les variables aléatoires discrètes II (loi, loi conditionnelle, espérance), les variables aléatoires continues III (loi, loi conditionnelle, espérance, fonction de répartition), les fonctions caractéristiques IV (ou transformée de Fourier), les théorèmes limites V (les différentes convergences, loi forte des grands nombres, théorème central limite), les vecteurs gaussiens VI, la simulation VII (qui correspond à un paragraphe dans le livre de cours) ; et avec en statistique : les estimateurs VIII (estimation paramétrique, estimateur du maximum de vraisemblance, propriétés des estimateurs à horizon fini et asymptotique, comparaison d'estimateurs, amélioration d'estimateurs par conditionnement par rapport à une statistique exhaustive), les tests IX (tests asymptotiques, tests du χ^2 empiriques, tests d'adéquation de loi et tests non paramétriques) et les intervalles et régions de confiance X. Dans chaque chapitre d'exercices, nous nous sommes efforcés de présenter des exercices de manipulation qui permettent d'appréhender les concepts du cours, puis parfois des exercices spécifiques (comme sur les lois ou probabilités conditionnelles, sur le jeu de pile ou face ou sur le test du χ^2), et surtout des exercices de modélisation avec des applications diverses dans plusieurs domaines scientifiques (mathématiques, physique, sciences de l'ingénieur, sciences du vivant, économie...). Citons par exemple : le calcul du nombre de points fixes d'une permutation I.9, les notions de sensibilité et de spécificité des tests de dépistage I.11, l'analyse du jeu "Let's Make a Deal" I.14, une

méthode d'optimisation de coût de dépistage II.18, un calcul (peu efficace) de π avec l'aiguille de Buffon III.16, le calcul de la loi de Maxwell pour la distribution des vitesses des molécules d'un gaz III.18, des calculs de limites déterministes par des méthodes stochastiques V.11 et V.12, le paradoxe de Saint-Petersbourg V.13, l'analyse des sondages V.14 (voir aussi les problèmes XI.10 et XII.8), le théorème de Weierstrass V.17, un modèle de contamination au mercure V.18, la détection de la contamination dans l'agro-alimentaire VIII.10, la détection d'un vaccin déficient IX.7, l'analyse du nombre de garçons dans les familles à 5 enfants IX.18 (voir aussi le problème XII.10), les dés (non-équilibrés) de Weldon IX.19, l'analyse des intervalles de confiance d'une fréquence X.3...

Viennent enfin deux chapitres de problèmes qui utilisent l'ensemble des concepts du cours de probabilités (chapitre XI) et du cours de probabilités et de statistique (chapitre XII). Les problèmes ont une forte composante de modélisation et d'analyse des modèles correspondant à des phénomènes réels. Ils sont révélateurs de la diversité de l'utilisation actuelle des probabilités et des statistiques. Citons par exemple : l'analyse des populations biologiques (leur dynamique avec le processus de Galton-Watson XI.8 et l'estimation de leur taille XII.2), l'analyse génétique (modèle de Hardy-Weinberg XII.1 et modèle de mutation de l'ADN XII.13), l'analyse des temps d'attente (problèmes du collectionneur XI.3 et XI.4, paradoxe du bus XI.5), les modèles en économie (mathématiques financières XI.12, stratégie optimale dans le problème du mariage XI.14, vente de disques XI.11 et XII.9, taille des villes XII.6), la théorie de l'information XI.13, l'analyse des séries temporelles XII.12, la théorie des sondages XI.10 et XII.8, des modèles de physique (loi de Bose-Einstein XI.9, résistance de céramiques XII.7, chaleur latente XII.5), les méthodes statistiques de comparaison d'échantillons XII.11, XII.6, XII.3, XII.4... Le dernier chapitre comprend les corrections détaillées des exercices et des problèmes.

Grâce à la participation active des élèves et surtout des enseignants, ce fut un plaisir de rédiger ces exercices et problèmes. Je souhaite ainsi remercier : Mohamed Ben Alaya, Hermine Biermé, Antoine Chambaz, Jean-Stéphane Dhersin, Roland Donat, Anne Dutfoy, Xavier Epiard, Marie-Pierre Etienne, Josselin Garnier, Julien Guyon, Far Hadda, Olivier Hénard, Patrick Hoscheit, Lauris Joubert, Régis Lebrun, Vincent Lefieux, Jérôme Lelong, Eulalia Nualart, Christian Paroissin, Bénédicte Puig, Victor Rivero, Raphaël Roux, Mohamed Sbaï, Simone Scotti, Michel Sortais, Emmanuel Temam et Mathias Winkel. Je souhaite également remercier Romain Abraham, Didier Chauveau, Benjamin Jourdain et Bernard Lapeyre pour les nombreuses discussions ainsi que Jean-Philippe Chancelier pour son aide précieuse concernant l'utilisation des logiciels Latex et Scilab (<http://www.scilab.org/>). Enfin je remercie les personnels du Cermics, labo-

ratoire de l'École des Ponts et Chaussées, pour l'ambiance de travail agréable et stimulante qu'ils ont su créer et développer.

Avec pour la fin, les remerciements à ceux qui m'ont entouré et soutenu tout au long de la rédaction de cet ouvrage.

Champs-sur-Marne,
Octobre 2011.

Jean-François Delmas

Table des matières

partie I Énoncés

| | | |
|------------|--|-----------|
| I | Espaces probabilisés | 3 |
| I.1 | Dénombrement | 3 |
| I.2 | Formule du crible et applications | 4 |
| I.3 | Probabilités conditionnelles | 6 |
| II | Variables aléatoires discrètes | 9 |
| II.1 | Exercices de manipulation | 9 |
| II.2 | Jeu de pile ou face | 11 |
| II.3 | Lois conditionnelles | 13 |
| II.4 | Modélisation | 14 |
| III | Variables aléatoires continues | 19 |
| III.1 | Calculs de probabilités ou d'espérance | 19 |
| III.2 | Calculs de loi | 20 |
| III.3 | Modélisation | 22 |
| IV | Fonctions caractéristiques | 27 |
| IV.1 | Calculs de loi | 27 |
| IV.2 | Modélisation | 29 |
| V | Théorèmes limites | 31 |
| V.1 | Quelques convergences | 31 |
| V.2 | Calcul de limites | 34 |
| V.3 | Modélisation | 35 |

| | | |
|-------------|--|------------|
| VI | Vecteurs gaussiens | 41 |
| VI.1 | Exemples | 41 |
| VI.2 | Propriétés et applications | 42 |
| VII | Simulation | 45 |
| VIII | Estimateurs | 47 |
| VIII.1 | Modèles paramétriques usuels | 47 |
| VIII.2 | Modélisation | 49 |
| IX | Tests | 55 |
| IX.1 | Mise en œuvre | 55 |
| IX.2 | Modélisation | 58 |
| IX.3 | Tests du χ^2 | 61 |
| X | Intervalles et régions de confiance | 67 |
| XI | Problèmes (probabilités) | 69 |
| XI.1 | Loi uniforme sur la sphère | 69 |
| XI.2 | Le collectionneur | 70 |
| XI.3 | Le paradoxe du bus | 74 |
| XI.4 | La statistique de Mann et Whitney | 79 |
| XI.5 | Le processus de Galton Watson | 82 |
| XI.6 | Loi de Bose-Einstein | 83 |
| XI.7 | Sondages (II) | 86 |
| XI.8 | Loi de Yule (I) | 88 |
| XI.9 | Mathématiques financières | 91 |
| XI.10 | Transmission de message | 94 |
| XI.11 | Mariage d'un prince | 96 |
| XII | Problèmes (probabilités et statistique) | 101 |
| XII.1 | Le modèle de Hardy-Weinberg | 101 |
| XII.2 | Estimation de la taille d'une population | 104 |
| XII.3 | Comparaison de traitements | 106 |
| XII.4 | Ensemencement des nuages | 110 |
| XII.5 | Chaleur latente de fusion | 114 |
| XII.6 | Taille des grandes villes | 119 |
| XII.7 | Résistance d'une céramique | 124 |
| XII.8 | Sondages (II) | 128 |
| XII.9 | Loi de Yule (II) | 131 |
| XII.10 | Sexe des anges | 133 |
| XII.11 | Comparaison d'échantillons appariés | 136 |

| | | |
|--------|--|-----|
| XII.12 | Modèle auto-régressif pour la température..... | 140 |
| XII.13 | Mutation de l'ADN mitochondrial..... | 144 |

partie II Corrections

| | | |
|--------------|---|------------|
| XIII | Corrections | 151 |
| XIII.1 | Espaces probabilisés | 151 |
| XIII.2 | Variables aléatoires discrètes | 161 |
| XIII.3 | Variables aléatoires continues | 179 |
| XIII.4 | Fonctions caractéristiques | 197 |
| XIII.5 | Théorèmes limites | 202 |
| XIII.6 | Vecteurs gaussiens..... | 219 |
| XIII.7 | Simulation | 226 |
| XIII.8 | Estimateurs | 229 |
| XIII.9 | Tests | 243 |
| XIII.10 | Intervalles et régions de confiance | 268 |
| XIII.11 | Problèmes (probabilités) | 273 |
| XIII.12 | Problèmes (probabilités et statistique) | 318 |
| Index | | 383 |

Première partie

Énoncés

Espaces probabilisés

I.1 Dénombrement

Exercice I.1 (Jeu de cartes).

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 (Jeu de dés).

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

△

Exercice I.3 (Anniversaires simultanés).

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.4 (Jeu de pile ou face).

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 ?

△

Exercice I.5 (Tirage avec remise et sans remise).

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 ($p_r \leq r$, $p_b \leq b$ et $p_r + p_b = p$).
3. Calculer, dans les deux cas, les probabilités limites quand $r \rightarrow \infty$, $b \rightarrow \infty$ et $r/(b+r) \rightarrow \theta \in]0, 1[$.

△

I.2 Formule du crible et applications**Exercice I.6** (Formule du crible).

Soit A_1, \dots, 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. Montrer la formule du crible par récurrence.

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

3. Inégalités de Bonferroni. Montrer, par récurrence sur n , que pour $1 \leq m \leq n$,

$$\sum_{p=1}^m (-1)^{p+1} \sum_{1 \leq i_1 < \dots < i_p \leq n} \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_p})$$

est une majoration (resp. minoration) de $\mathbb{P}(\bigcup_{i=1}^n A_i)$ lorsque m est impair (resp. pair).

**Exercice I.7** (Entente).

Afin de savoir si les élèves travaillent indépendamment ou en groupe, un enseignant donne m exercices à une classe de n élèves et demande à chaque élève de choisir 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. Que peut dire l'enseignant si tous les élèves ont choisi la même combinaison de 2 exercices ?

**Exercice I.8** (Nombre de Stirling).

On utilise dans cet exercice la formule du crible (I.1) (cf exercice I.6). Soit $1 \leq k \leq n$.

1. Calculer à l'aide de la formule du crible le nombre de surjections de $\{1, \dots, n\}$ dans $\{1, \dots, k\}$.
2. En déduire $\left\{ \begin{matrix} n \\ k \end{matrix} \right\}$, le nombre de partitions d'un ensemble à n éléments en k sous-ensembles non vides. Les nombres $\left\{ \begin{matrix} n \\ k \end{matrix} \right\}$ sont appelés les nombres de Stirling de deuxième espèce.
3. Montrer que

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

**Exercice I.9** (Points fixes d'une permutation).

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 déduire le nombre de permutations de $\{1, \dots, n\}$ sans point fixe (problème formulé par P. R. de Montmort en 1708)¹
3. En s'inspirant de la question 1, calculer la probabilité $\pi_n(k)$ pour que k personnes exactement aient leur propre veste.
4. Calculer la limite $\pi(k)$ de $\pi_n(k)$ quand n tend vers l'infini. 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.

△

I.3 Probabilités conditionnelles

Exercice I.10 (Famille à deux enfants).

On suppose que l'on a autant de chance d'avoir une fille ou un garçon à la naissance (et qu'il n'y a pas de jumeau). 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 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 ?
5. 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 ?

La question 3 a une variante² (laissée à la sagacité du lecteur) : Quelle est la probabilité que votre voisin qui a deux enfants, ait un garçon, sachant qu'il a une fille qui est née un mardi ?

△

1. L. Takacs. The problem of coincidences. *Arch. Hist. Exact Sci.*, vol. 21(3), pp. 229-244 (1980).

2. T. Khovanova. Martin Gardner's Mistake. <http://arxiv.org/pdf/1102.0173v1> (2011).

Exercice I.11 (Test médical).

Les laboratoires pharmaceutiques indiquent pour chaque test sa sensibilité α , qui est la probabilité que le test soit positif si le sujet est malade, et sa spécificité β , qui est la probabilité que le test soit négatif si le sujet est sain. Sachant qu'en moyenne il y a un malade sur 1000 personnes, calculer la probabilité pour que vous soyez un sujet sain alors que votre test est positif, avec $\alpha = 98\%$ et $\beta = 97\%$. Calculer la probabilité d'être malade alors que le test est négatif. Commentaire.

△

Exercice I.12 (Couleur des yeux).

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 qu'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.13 (Paradoxe de Bertrand (1889)).

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 soit blanche sachant que la face visible est rouge.

△

Exercice I.14 (Let's Make a Deal).

Le problème qui suit est inspiré du jeu télévisé américain "Let's Make a Deal" (1963-1977) présenté par Monty Hall³. On considère trois portes : A , B et C .

3. Wikipedia : http://en.wikipedia.org/wiki/Monty_Hall_problem.

Derrière l'une d'entre elles se trouve un cadeau 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 le cadeau, ouvre une porte parmi les deux B et C , derrière laquelle il n'y a évidemment rien. On vous propose alors de changer ou non de porte, le but étant d'ouvrir la porte qui cache le cadeau afin de gagner. L'objectif de cet exercice est de déterminer votre meilleure stratégie.

1. On suppose que si le cadeau est derrière la porte A , alors le présentateur choisit au hasard entre les deux autres portes. Calculer la probabilité pour que le cadeau soit derrière la porte B sachant que le présentateur ouvre la porte C . Que faites-vous ?
2. On suppose que si le cadeau est derrière la porte A , alors le présentateur choisit systématiquement la porte B . Que faites-vous si le présentateur ouvre la porte B (respectivement C) ?
3. Montrer que quelle que soit la valeur de la probabilité pour que le présentateur ouvre la porte B (respectivement C) sachant que le cadeau est derrière la porte 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 le cadeau ? Vérifier que cette stratégie est moins bonne que la précédente.

△

II

Variables aléatoires discrètes

II.1 Exercices de manipulation

Exercice II.1 (Lois discrètes usuelles).

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

△

Exercice II.2 (Loi de Poisson).

Soit X une variable aléatoire de loi de Poisson de paramètre $\theta > 0$: $\mathbb{P}(X = k) = e^{-\theta} \frac{\theta^k}{k!}$, $k \in \mathbb{N}$.

1. Vérifier que $\frac{1}{1+X}$ est une variable aléatoire intégrable. Calculer $\mathbb{E} \left[\frac{1}{1+X} \right]$.
2. Calculer $\mathbb{E} \left[\frac{1}{(1+X)(2+X)} \right]$ et en déduire $\mathbb{E} \left[\frac{1}{2+X} \right]$.

△

Exercice II.3 (Ivresse).

Un gardien de nuit doit ouvrir une porte dans le noir, avec n clefs dont une seule est la bonne. On note X le nombre d'essais nécessaires pour trouver la bonne clef.

1. On suppose que le gardien essaie les clefs une à une sans utiliser deux fois la même. Donner la loi de X , son espérance et sa variance.
2. Lorsque le gardien est ivre, il mélange toutes les clefs à chaque tentative. Donner la loi de X , son espérance et sa variance.
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 la limite quand n tend vers l'infini.

△

Exercice II.4 (Jeu de dés).

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 $\mathbb{P}(X \leq k)$ pour $k \in \mathbb{N}$.
2. Soit Y une variable à valeurs dans \mathbb{N} . Montrer que

$$\mathbb{E}[Y] = \sum_{k=1}^{\infty} \mathbb{P}(Y \geq k).$$

3. Combien de lancers sont nécessaires en moyenne pour obtenir les 5 six ?

△

Exercice II.5 (Mélange de lois discrètes).

Soit des variables aléatoires indépendantes X , de loi de Bernoulli de paramètre $p \in]0, 1[$, Y de loi géométrique de paramètre $a \in]0, 1[$ et Z de loi de Poisson de paramètre $\theta > 0$.

1. Donner les fonctions génératrices de Y et Z .
2. Soit U la variable aléatoire 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 variable aléatoire 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.6 (Questionnaire à choix multiples).

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.7** (Deux urnes).

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 urnes, elles n'ont pas la même couleur. Les tirages sont sans remise.

1. Calculer la probabilité pour que lors du i -ème tirage, les deux boules n'aient pas la 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.8** (Urne).

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 $x \in \{1, \dots, N\}$. En déduire la loi de X .
2. Donner la loi de Y .
3. Calculer $\mathbb{P}(X > x, Y \leq y)$ pour tout $(x, y) \in \{0, \dots, N\}^2$. En déduire la loi du couple (X, Y) .

**Exercice II.9** (Dé à 11 faces).

On désire répondre à la question suivante : Peut-on reproduire le résultat d'un lancer d'un dé équilibré à onze faces, numérotées de 2 à 12, comme la somme d'un lancer de deux dés à six faces, numérotées de 1 à 6, éventuellement différemment biaisés ?

1. Soit X de loi uniforme sur $\{2, \dots, 12\}$. Vérifier que la fonction génératrice de X est un polynôme. Quelles sont ses racines réelles ?
2. Étudier les racines de la fonction génératrice associée à la somme d'un lancer de deux dés à six faces. Conclure.

**II.2 Jeu de pile ou face****Exercice II.10** (Égalité de pile).

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.11 (Les boîtes d'allumettes de Banach).

Ce problème¹. 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 boîte qu'il a choisie, quelle est la probabilité pour qu'il reste k allumettes dans l'autre boîte ?
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.12 (Loi binomiale négative).

On considère un jeu de pile ou face biaisé : les variables aléatoires $(X_n, n \in \mathbb{N}^*)$ sont indépendantes et de même loi de Bernoulli de paramètre $p \in]0, 1[$: $\mathbb{P}(X_n = 1) = p$ et $\mathbb{P}(X_n = 0) = 1 - p$. 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 $T_1 - T_0, T_2 - T_1, \dots, T_{k+1} - T_k$ sont indépendantes et de même loi.
3. Calculer $\mathbb{E}[T_k]$ et $\text{Var}(T_k)$.
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 seconde pièce de paramètre $\rho \in]0, 1[$. 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.

△

1. W. Feller. *An introduction to probability theory and its applications*, vol. 1. Wiley, third ed. (1968).

Exercice II.13 (Temps d'apparition d'une séquence).

On considère un jeu de pile ou face biaisé : les variables aléatoires $(X_n, n \in \mathbb{N}^*)$ sont indépendantes et de même loi de Bernoulli de paramètre $p \in]0, 1[$: $\mathbb{P}(X_n = 1) = p$ et $\mathbb{P}(X_n = 0) = 1 - p$. On considère le premier temps d'apparition de la séquence 10 : $T = \inf\{n \geq 2; X_{n-1} = 1, X_n = 0\}$, 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.14 (Loi de Poisson).

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de loi de Bernoulli de paramètre $p \in]0, 1[$, $\mathbb{P}(X_n = 1) = p$ et $\mathbb{P}(X_n = 0) = 1 - p$, et N une variable aléatoire à valeurs dans \mathbb{N} indépendante de $(X_n, n \geq 1)$. On pose $S = 0$ si $N = 0$ et $S = \sum_{k=1}^N X_k$ sinon. On désire déterminer les lois de N telles que les variables S et $N - S$ soient indépendantes. On note ϕ la fonction génératrice de N .

1. On suppose que la loi de N est la loi de Poisson de paramètre $\theta > 0$: $\mathbb{P}(N = n) = \frac{\theta^n}{n!} e^{-\theta}$ pour $n \in \mathbb{N}$. Déterminer la loi de $(S, N - S)$. Reconnaître la loi de S . Vérifier que S et $N - S$ sont indépendants.

On suppose que S et $N - S$ sont indépendants.

2. Montrer que pour tout $z \in [-1, 1]$, $\phi(z) = \phi((1 - p) + pz)\phi(p + (1 - p)z)$. On pose $h(z) = \phi'(z)/\phi(z)$, pour $z \in]0, 1[$. Montrer que $h(z) = ph((1 - p) + pz) + (1 - p)h(p + (1 - p)z)$.
3. On suppose $p = 1/2$. Vérifier que $h(z) = h((1 + z)/2)$. En déduire que $h(z) = \lim_{r \rightarrow 1^-} h(r)$, puis que soit $N = 0$ p.s., soit N suit une loi de Poisson.
4. On suppose $p \in]0, 1/2[$. Montrer, en s'inspirant de la question précédente, que soit $N = 0$ p.s., soit N suit une loi de Poisson.

△

II.3 Lois conditionnelles

Exercice II.15 (Loi de Bernoulli).

Soit (X_1, \dots, 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, \dots, 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 (pour $n \geq 2$) ?

△

Exercice II.16 (Loi de Poisson).

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 $\mathbb{E}[X_1 | X_1 + X_2]$.

△

Exercice II.17 (Somme de variables aléatoires géométriques).

Soit 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 $\mathbb{E}[\mathbb{E}[X | S]] = \mathbb{E}[X]$.

△

II.4 Modélisation

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 stratégies suivantes :

Stratégie 1 : Un test par personne.

Stratégie 2 : On suppose que N/n est entier et on regroupe les N personnes en N/n groupes de n personnes. Par groupe, on mélange les prélèvements sanguins des 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 stratégie 1, puis le coût moyen de la stratégie 2. On supposera $np \ll 1$, et on montrera que $n \simeq p^{-1/2}$ est une taille qui minimise correctement le coût moyen de la stratégie 2. Quelle stratégie choisissez-vous? Illustrer vos conclusions pour le cas où $p = 1\%$.

Cette démarche a été utilisée initialement par R. Dorfman², durant la Seconde Guerre mondiale, dans un programme commun avec l'United States Public Health Service et le 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% 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 (Temps de panne).

On désire modéliser la loi du temps d'attente d'une panne de machine à l'aide d'une loi 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 . Plus précisément, on dit qu'une variable aléatoire X à valeurs dans \mathbb{N} est de loi sans mémoire si $\mathbb{P}(X > k + n | X > n)$ est indépendant de $n \in \mathbb{N}$ pour tout $k \in \mathbb{N}$.

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

△

Exercice II.20 (Médailles).

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 mondiale?

△

Exercice II.21 (Suites typiques).

On s'intéresse aux nombres de suites typiques contenant des 0 ou des 1. Plus précisément, on considère Ω_n l'ensemble des suites $\omega = (\omega_1, \dots, \omega_n) \in \{0, 1\}^n$ muni de la probabilité $\mathbb{P}(\{\omega\}) = p^{\sum_{i=1}^n \omega_i} q^{n - \sum_{i=1}^n \omega_i}$. On définit le sous-ensemble

2. R. Dorfman. The detection of defective numbers of large populations. *Ann. Math. Statist.* vol. 14, pp. 436-440 (1943).

de Ω des suites typiques de longueur n par :

$$C_n = \left\{ \omega \in \Omega_n; \left| \frac{1}{n} \sum_{i=1}^n \omega_i - p \right| \leq \delta_n \right\},$$

où $\lim_{n \rightarrow \infty} \delta_n = 0$ et $\lim_{n \rightarrow \infty} \sqrt{n} \delta_n = +\infty$. Le but de l'exercice est de montrer que l'ensemble C_n est de probabilité proche de 1 et d'estimer son cardinal (qui peut être significativement plus petit que celui de Ω_n qui vaut 2^n).

1. Soit $\alpha \in]0, 1[$. Montrer, à l'aide de l'inégalité de Tchebychev que pour n assez grand, on a $\mathbb{P}(C_n) \geq 1 - \alpha$.

On définit l'entropie de la loi de Bernoulli de paramètre p par $H_p = -p \log(p) - (1 - p) \log(1 - p)$.

2. Pour quelle valeur de p l'entropie est-elle maximale ?
3. Montrer que pour n assez grand, on a pour tout $\omega \in C_n$:

$$e^{-n(H_p + \beta_n)} \leq \mathbb{P}(\{\omega\}) \leq e^{-n(H_p - \beta_n/2)} \leq e^{-n(H_p - \beta_n)},$$

avec $0 \leq \beta_n = c_p \delta_n$, la constante c_p dépendant seulement de p .

4. Montrer que pour tout n assez grand, on a :

$$e^{n(H_p - \beta_n)} \leq \text{Card}(C_n) \leq e^{n(H_p + \beta_n)},$$

5. Quel résultat obtenez-vous si $p = 1/2$, si $p \simeq 0$ ou $p \simeq 1$?

Certaines techniques de compression consistent à trouver les suites typiques pour une longueur n fixée, et à les coder par des séquences courtes (d'où la compression). La contrepartie est que les suites non-typiques sont codées par des séquences plus longues. Comme les suites non-typiques sont très rares, elles apparaissent peu souvent et donc la compression est peu dégradée par les suites non-typiques. La compression est d'autant plus importante que l'ensemble des suites typiques est petit. Dans le cas de séquences aléatoires traitées dans cet exercice, cela correspond aux cas $p \simeq 0$ ou $p \simeq 1$.

△

Exercice II.22 (Renouvellement).

On souhaite modéliser des phénomènes qui se répètent à des intervalles de temps aléatoires indépendants et identiquement distribués (théorie des processus de renouvellement). 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, pour $n \in \mathbb{N}$,

$$\mathbb{P}(\text{le phénomène se produit à l'instant } n) = v_n.$$

L'objectif de cet exercice est d'utiliser les fonctions génératrices pour calculer ces probabilités en fonction des lois des intervalles de temps aléatoires, puis, d'en déduire les lois des intervalles de temps quand les probabilités v_n sont stationnaires, i.e. indépendantes de n .

On note T_n l'instant de n -ième occurrence du phénomène considéré. On pose $X_1 = T_1$, et pour $n \geq 2$, $X_n = T_n - T_{n-1}$, l'intervalle de temps ou le temps écoulé entre la n -ième et la $(n - 1)$ -ième occurrence du phénomène. On suppose que les variables aléatoires $(X_n, n \geq 1)$ sont indépendantes, que X_1 est à valeurs dans \mathbb{N} et que les variables aléatoires $(X_n, n \geq 2)$ sont identiquement distribuées à valeurs dans \mathbb{N}^* . La loi de X_1 est *a priori* différente de celle de X_2 car le choix du temps à partir duquel on commence à compter les occurrences est arbitraire.

Pour $k \in \mathbb{N}$, on pose $b_k = \mathbb{P}(X_1 = k)$ et pour $s \in [-1, 1]$, $B(s) = \sum_{k=0}^{\infty} b_k s^k$ la fonction génératrice de X_1 . Pour $k \in \mathbb{N}$, on pose $f_k = \mathbb{P}(X_2 = k)$ (avec $f_0 = 0$) et pour $s \in [-1, 1]$, $F(s) = \sum_{k=0}^{\infty} f_k s^k$ la fonction génératrice de X_2 .

On définit $u_0 = 1$ et, pour $j \geq 1$,

$$u_j = \mathbb{P} \left(\sum_{i=2}^k X_i = j \text{ pour un } k \in \{2, \dots, j + 1\} \right).$$

Pour $s \in]-1, 1[$, on pose $U(s) = \sum_{n=0}^{\infty} s^n u_n$ ainsi que $V(s) = \sum_{n=0}^{\infty} v_n s^n$.

1. Vérifier que les fonctions U et V sont bien définies pour $s \in]-1, 1[$.
2. Montrer que $v_n = b_n u_0 + b_{n-1} u_1 + b_{n-2} u_2 + \dots + b_0 u_n$. En déduire que pour $s \in]-1, 1[$, on a $V(s) = B(s)U(s)$.
3. Montrer que $u_n = f_1 u_{n-1} + f_2 u_{n-2} + \dots + f_n u_0$. En déduire que

$$V(s) = \frac{B(s)}{1 - F(s)} \quad \text{pour } s \in]-1, 1[. \tag{II.1}$$

On suppose que les probabilités v_n sont stationnaires et non triviales, i.e. il existe $p \in]0, 1[$ et $v_n = p$ pour tout $n \geq 0$.

4. Calculer V . Calculer b_0 . Montrer, en calculant les dérivées successives de F en fonction de celles de B , que $f_n = \frac{b_{n-1} - b_n}{p}$ pour tout $n \geq 1$.
5. On suppose de plus que le choix arbitraire du temps à partir duquel on commence à compter les occurrences ne change pas le processus des occurrences : plus précisément s'il n'y a pas d'occurrence à l'instant initial, alors le temps de la première occurrence a même loi que X_2 . Autrement dit, on suppose que la loi de X_1 sachant $\{X_1 > 0\}$ est la loi de X_2 . Montrer alors que X_2 suit la loi géométrique de paramètre p , i.e. $f_n = p(1 - p)^{n-1}$ pour $n \geq 1$, et que X_1 a même loi que $X_2 - 1$.

Ce modèle simple permet, par exemple, de rendre compte des temps d'arrivées des particules α à un compteur Geiger : les probabilités d'observer l'arrivée d'une particule α au compteur Geiger ne varie pas avec le temps d'observation (régime stationnaire), et le choix du début des mesures ne change pas la loi d'attente de la première mesure. Avec une discrétisation régulière du temps, on peut modéliser les intervalles de temps entre deux arrivées de particules α par des variables aléatoires indépendantes de loi géométrique.

△

III

Variables aléatoires continues

III.1 Calculs de probabilités ou d'espérance

Exercice III.1 (Loi de Rayleigh).

Soit la fonction f définie sur \mathbb{R} par : $f(x) = x e^{-x^2/2} \mathbf{1}_{\{x>0\}}$.

1. Vérifier que f est une densité de probabilité.
2. Soit X une variable aléatoire continue dont la loi a pour densité f . Montrer que $Y = X^2$ est une variable aléatoire continue, dont on précisera la densité. Reconnaître la loi de Y .
3. Calculer l'espérance et la variance de Y

La densité f correspond à la loi de Rayleigh.

△

Exercice III.2 (Galette des rois).

On considère une galette des rois circulaire de rayon R , une fève de rayon r cachée dans la galette ($R > 2r > 0$). Calculer la probabilité de toucher la fève quand on donne le premier coup de couteau dans la galette. (On suppose que le coup de couteau correspond à un rayon de la galette.) Donner un équivalent de cette probabilité pour r petit.

△

Exercice III.3 (La cerise sur le gâteau).

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.

1. Avec quelle probabilité la part contenant la cerise est-elle plus petite que la part ne contenant pas la cerise ?
2. Quelle est la longueur angulaire moyenne de la part contenant la cerise ?

△

Exercice III.4 (Couper un bâton).

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.5 (Somme de variables aléatoires exponentielles).

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

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

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

△

III.2 Calculs de loi

Exercice III.6 (Loi exponentielle symétrique).

Soit Y une variable aléatoire de loi exponentielle $\lambda > 0$ et ε une variable aléatoire discrète indépendante de Y et telle que $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$. Montrer que la loi de la variable aléatoire $Z = \varepsilon Y$ possède une densité et la calculer. Cette loi est appelée loi exponentielle symétrique. △

Exercice III.7 (Lois gamma).

Soit X et Y deux variables aléatoires indépendantes de loi respective $\Gamma(\lambda, a)$ et $\Gamma(\lambda, b)$ avec $a, b, \lambda \in]0, \infty[$.

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.8 (Loi de Cauchy).

Soit X une variable aléatoire 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.9 (Loi du χ^2).

Soit X_1, X_2 des variables aléatoires indépendantes de loi $\mathcal{N}(0, 1)$.

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

△

Exercice III.10 (Loi du min et du max).

Soit $n \geq 2$ et X_1, \dots, 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 puis reconnaître la loi de Y/Z conditionnellement à Z . Retrouver le résultat de la question 3.
5. Montrer que $(1 - Z, 1 - Y)$ a même loi que (Y, Z) .
6. En déduire que $(1 - Z)/(1 - Y)$ est indépendant de Y .

△

Exercice III.11 (Lois conditionnelles).

Soit Y une variable aléatoire de loi $\Gamma(1, 1/2)$ de densité :

$$f_Y(y) = \frac{1}{\sqrt{\pi y}} e^{-y} \mathbf{1}_{\{y>0\}}.$$

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 $\mathbb{E}[Y|X]$.

△

III.3 Modélisation

Exercice III.12 (Durée de vie).

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 : $F(t) = (1 - e^{-t^2/2})\mathbf{1}_{\{t \geq 0\}}$.

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 ?

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 , avec $X_i = 1$ si la durée de vie du circuit i est inférieure à un an et $X_i = 0$ sinon.

3. Quelle est la loi du nombre, N , de circuits de E dont la durée de vie est inférieure à 1 an ?
4. 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 ?
5. 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 ?

△

Exercice III.13 (Stockage).

On utilise des fûts pour stocker un déchet toxique liquide. On suppose que sur la très longue période de stockage les fûts se dégradent. En particulier, par corrosion des perforations aléatoires apparaissent sur les fûts. Le liquide toxique s'écoule alors par ces perforations.

On considère n fûts de hauteur h et on suppose que le nombre de perforations par fût suit une loi de Poisson de paramètre θ et que ces perforations sont uniformément réparties sur la hauteur du fût.

On s'intéresse à un seul fût.

1. On note Z la variable aléatoire correspondant à la hauteur de la perforation la plus basse sur le côté du fût et N la variable aléatoire donnant le nombre de perforations sur ce fût. Donner la loi de Z conditionnellement à N .
2. Quel pourcentage de liquide peut-on espérer conserver connaissant le nombre de perforations du fût ?

On considère l'ensemble des n fûts, avec n grand.

3. Quel pourcentage du liquide toxique peut-on espérer conserver ?

Application numérique : $\theta = 5$.

**Exercice III.14** (Le paradoxe de Bertrand).

Le paradoxe de Bertrand¹ est un exemple classique² 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 note L sa longueur. Pour les trois choix ci-dessous de la corde, donner la loi de L , son espérance, sa variance et $p = \mathbb{P}(L > \sqrt{3}r)$ la probabilité pour que la corde soit plus longue que le côté du triangle équilatéral inscrit (i.e. la distance de la corde au centre du cercle est inférieure à $r/2$).

1. On se donne un rayon au hasard et un point J uniformément sur le rayon. La corde choisie est perpendiculaire au rayon et passe par J .
2. On choisit la corde AB où A et B sont choisis indépendamment et uniformément sur le cercle. On pourra faire intervenir l'angle au centre \widehat{AOB} et montrer qu'il est de loi uniforme sur $[0, 2\pi]$.
3. On choisit le milieu de la corde, I , uniformément dans le disque : les coordonnées (X, Y) de I suivent la loi uniforme sur le disque de densité $\frac{1}{\pi r^2} \mathbf{1}_{\{x^2+y^2 \leq r^2\}}$.

Quelle est votre conclusion ?

**Exercice III.15** (GPS et loi de Rayleigh).

Les véhicules spatiaux désirant s'arrimer à la Station Spatiale Internationale (ISS) s'appuient sur le système de guidage GPS (Global Positioning system) pour la phase d'approche de la station. Cependant, à faible distance de l'ISS, les signaux émis 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 gaussienne $\mathcal{N}(0, \sigma^2)$ supposées indépendantes (pour des raisons d'isotropie). L'étude de l'amplitude

1. Joseph Bertrand, mathématicien français (1822-1900).

2. H. Poincaré. *Calcul des probabilités*. Gauthier-Villars (1912).

$R = \sqrt{X^2 + Y^2}$ de l'onde reçue est de première importance pour assurer un guidage fiable du vaisseau lors de la manœuvre d'arrimage³.

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) .
3. Montrer que R et Θ sont indépendants. Reconnaître la loi de Θ . Donner la densité de la loi de R . Cette loi est appelée loi de Rayleigh.
4. Calculer l'espérance et la variance de R .

△

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 $d > \ell$. 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 sa direction, donnée par 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 effectue n lancers et on note N_n le nombre d'aiguilles coupant une rainure du parquet. Que vaut $\lim_{n \rightarrow \infty} \frac{N_n}{n}$?
4. On suppose que $2\ell = d$. 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 $2\ell = d$, 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 ?

△

Exercice III.17 (Devinette).

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

3. D. E. Gaylor, R. Glenn Lightsey and K. W. Key. Effects of Multipath and Signal Blockage on GPS Navigation in the Vicinity of the International Space Station (ISS). *ION GPS/GNSS* Portland, OR (2003).

a obtenu face et le plus grand sinon. Vous devez parier s'il vous a donné le plus petit ou le plus grand. Votre objectif est de maximiser votre probabilité de gagner votre pari.

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}^+ (i.e. $\mathbb{P}(Z \in O) > 0$ pour tout ouvert O non vide de \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 obtenus 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.18 (loi de Maxwell).

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 valeurs dans \mathbb{R}^3 dont la loi est invariante par rotation autour de l'origine et dont les composantes X, Y, Z sont indépendantes. Caractériser⁴ les lois marginales de X, Y et Z dans le cas où les densités des lois marginales sont des fonctions de classe C^1 .
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.)

4. En fait, on peut, en utilisant les fonctions caractéristiques, caractériser toutes les lois des vecteurs aléatoires qui sont invariantes par rotation autour de l'origine et dont les coordonnées sont indépendantes, voir l'exercice IV.6. Hormis la variable nulle, on ne trouve pas d'autres lois que celles obtenues sous les hypothèses de cet exercice.

3. Montrer que si X et Y sont deux variables aléatoires 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 $|V| = \sqrt{V_1^2 + V_2^2 + V_3^2}$ dite loi de Maxwell.

△

Exercice III.19 (Formule de duplication de la fonction Γ).

Soit X et Y deux variables aléatoires indépendantes de loi respective $\Gamma(\lambda, \alpha)$ et $\Gamma(\lambda, \alpha + \frac{1}{2})$ avec $\alpha > 0$ et $\lambda > 0$. On rappelle que la densité de la loi $\Gamma(\lambda, a)$ est $x \mapsto \frac{1}{\Gamma(a)} \lambda^a x^{a-1} e^{-\lambda x} \mathbf{1}_{\{x>0\}}$ où $\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$.

1. Déterminer la loi de (\sqrt{XY}, \sqrt{Y}) .
2. Calculer la loi de \sqrt{XY} . On pourra utiliser l'égalité suivante :

$$\int_0^\infty e^{-\lambda(v^2 + \frac{u^2}{v^2})} dv = \frac{\sqrt{\pi}}{2\sqrt{\lambda}} e^{-2\lambda u}.$$

3. Reconnaître la loi de \sqrt{XY} (identité en loi due à S. Wilks⁵). En déduire, en utilisant $\Gamma(1/2) = \sqrt{\pi}$, la formule de duplication de la fonction Γ : $\Gamma(2\alpha) = 2^{2\alpha-1} \frac{\Gamma(\alpha)\Gamma(\alpha + \frac{1}{2})}{\Gamma(1/2)}$.

△

5. S. Wilks. Certain generalizations in the analysis of variance. *Biometrika*, vol. 24, pp.471-494 (1932).

IV

Fonctions caractéristiques

IV.1 Calculs de loi

Exercice IV.1 (Stabilité de la loi gamma par addition).

Propriétés des lois gamma.

1. Soit X_1, X_2 deux variables aléatoires indépendantes et de loi gamma de paramètres respectifs (λ, α_1) et (λ, α_2) . (Le paramètre λ est identique.) Montrer que la loi de $X_1 + X_2$ est une loi gamma de paramètre $(\lambda, \alpha_1 + \alpha_2)$.
2. Soit $(X_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes de loi exponentielle de paramètre $\lambda > 0$. Donner la loi de la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

△

Exercice IV.2 (Loi exponentielle symétrique et loi de Cauchy).

Soit Y une variable aléatoire de loi exponentielle de paramètre $\lambda > 0$ et ε une variable aléatoire indépendante de Y et telle que $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$.

1. Calculer la densité et la fonction caractéristique de $Z = \varepsilon Y$. La loi de Z est appelée loi exponentielle symétrique.
2. En déduire la fonction caractéristique de la loi de Cauchy.

△

Exercice IV.3 (Matrice aléatoire).

Le but de cet exercice est le calcul de la loi du déterminant d'une matrice aléatoire gaussienne.

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 gaussienne centrée $\mathcal{N}(0, 1)$. On considère la matrice aléatoire $A = \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix}$ et on note $Y = \det(A)$ le déterminant de A . Calculer $\mathbb{E} [e^{iuY} | X_1, X_2]$, puis en déduire la fonction caractéristique de Y .

△

Exercice IV.4 (Formule de Wald).

Soit N une variable aléatoire de carré intégrable et à valeurs dans \mathbb{N} . Soit $(X_k, k \in \mathbb{N})$ une suite de variables aléatoires réelles, de même loi, de carré intégrable, 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 $\text{Var}(S_N)$ (formule de Wald).
2. Calculer la fonction caractéristique de S_N et retrouver les résultats précédents.

△

Exercice IV.5 (Loi symétrique).

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. Que peut-on dire à propos des fonctions caractéristiques des variables aléatoires réelles dont la loi est symétrique par rapport à $a \neq 0$ (i.e. X et $2a - X$ ont même loi) ?

△

Exercice IV.6 (Loi invariante par rotation).

Soit $X = (X_1, \dots, X_d)$ un vecteur aléatoire à valeurs dans \mathbb{R}^d . On suppose que la loi de X est invariante par rotation et que les variables aléatoires X_1, \dots, X_d sont indépendantes. Le but de cet exercice est de déterminer la loi de X . On note ψ_{X_1} la fonction caractéristique de X_1 .

1. On pose $g(x) = \psi_{X_1}(\sqrt{x})$ pour $x \geq 0$. Vérifier que g est réelle et solution de :

$$\prod_{k=1}^d g(v_k) = g\left(\sum_{k=1}^d v_k\right), \quad \text{pour tout } v_1, \dots, v_d \in [0, \infty[. \quad (\text{IV.1})$$

2. En déduire que $\psi_{X_1}(u_1) = e^{-\sigma^2 u_1^2/2}$, pour $\sigma \geq 0$. Montrer que soit $X = 0$ p.s. soit X_1, \dots, X_d sont de même loi gaussienne centrée.

△

IV.2 Modélisation

Exercice IV.7 (Loi de défaut de forme).

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 variables aléatoires indépendantes suivant respectivement les lois gaussiennes $\mathcal{N}(\mu_x, \sigma_x^2)$ et $\mathcal{N}(\mu_y, \sigma_y^2)$.

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

△

Théorèmes limites

V.1 Quelques convergences

Exercice V.1 (Loi exponentielle).

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

△

Exercice V.2 (Convergence de lois géométriques).

Soit $(T_n, n \geq n_0)$ une suite de variables aléatoires de loi géométrique de paramètre $p_n = \frac{\theta}{n}$ avec $n_0 > \theta > 0$. Montrer que la suite $\left(\frac{T_n}{n}, n \geq n_0\right)$ converge en loi et déterminer sa limite.

△

Exercice V.3 (Maximum de variables uniformes).

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 $n \geq 1$, $M_n = \max_{1 \leq i \leq n} U_i$.

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

△

Exercice V.4 (Moyenne empirique de variables de Cauchy).

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 suivantes.

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.5 (Le TCL n'est pas une convergence en probabilité).

Soit $(X_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires réelles de même loi, indépendantes de carré intégrable. On note $\mu = \mathbb{E}[X_1]$, $\sigma^2 = \text{Var}(X_1)$ et $Z_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n (X_k - \mu)$.

1. Rappeler la convergence en loi de la suite $(Z_n, n \in \mathbb{N}^*)$.
2. Établir la convergence de la suite $(Z_{2n} - Z_n, n \in \mathbb{N}^*)$ et donner sa limite.
3. En déduire que la suite $(Z_n, n \in \mathbb{N}^*)$ ne converge pas en probabilité si $\sigma^2 > 0$. △

Exercice V.6 (Approximations pour une jeu de pile ou face).

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

1. 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 $n = 20$?
2. Est-il plus probable que $N = 0$, que $N = 1$ ou que $N \geq 2$? △

Exercice V.7 (Produit de variables aléatoires).

Soit $(U_k, k \in \mathbb{N}^*)$ 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 $Y_n = (X_n e^\alpha)^{\sqrt{n}}$, 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.

△

Exercice V.8 (Loi hypergéométrique).

Soit X_N une variable aléatoire de loi hypergéométrique de paramètre (N, m, n) . On rappelle que X_N représente le nombre de boules blanches obtenues lors d'un tirage sans remise de n boules hors d'une urne contenant m boules blanches et $N - m$ boules noires.

1. Vérifier que $\mathbb{P}(X_N = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} = \frac{\binom{n}{k} \binom{N-n}{m-k}}{\binom{N}{m}}$ pour $n - N + m \leq k \leq m$ et $n \geq k \geq 0$.
2. On suppose que le nombre de boules blanches, m , est fixé, n et N tendent vers $+\infty$ avec $\lim_{N \rightarrow +\infty} n/N = p \in [0, 1]$ (p est la proportion limite du nombre de boules obtenues lors du tirage). Montrer que la suite $(X_N, N \in \mathbb{N}^*)$ converge en loi vers la loi binomiale de paramètre (m, p) .
3. On suppose que le tirage est de taille n fixé et que m et N tendent vers $+\infty$ avec $\lim_{N \rightarrow +\infty} m/N = \theta \in [0, 1]$ (θ est la proportion limite du nombre de boules blanches dans l'urne). Montrer que la suite $(X_N, N \in \mathbb{N}^*)$ converge en loi vers la loi binomiale de paramètre (n, θ) .

△

Exercice V.9 (Majoration du théorème des grandes déviations).

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes et de même loi que X . On suppose que X est une variable aléatoire bornée non constante de moyenne $\mu = \mathbb{E}[X]$. Le but de cet exercice est de trouver une majoration exponentielle de

l'événement rare $\{|\bar{X}_n - \mu| \geq \varepsilon\}$ avec $\varepsilon > 0$, où $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Cette majoration

est un exemple particulier des résultats de la théorie des grandes déviations¹.

On note Φ la transformée de Laplace de X définie par $\Phi(\theta) = \mathbb{E}[e^{\theta X}]$ pour $\theta \in \mathbb{R}$.

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

1. Montrer que Φ est de classe C^∞ . Calculer ses dérivées.
2. Montrer que $(\Phi')^2 \leq \Phi\Phi''$. En déduire que la fonction $\lambda = \log(\Phi)$ est convexe. Vérifier que $\lambda(0) = 0$.

On considère la transformée de Legendre de λ , I , définie pour $x \in \mathbb{R}$ par :

$$I(x) = \sup_{\theta \in \mathbb{R}} \{\theta x - \lambda(\theta)\}.$$

3. Montrer que I est convexe, positive et nulle en μ .
4. Montrer que $\mathbb{P}(\bar{X}_n \geq \mu + \varepsilon) \leq e^{-\theta(\mu+\varepsilon) + n\lambda(\theta/n)}$ pour $\theta \geq 0$. En déduire que :

$$\mathbb{P}(\bar{X}_n \geq \mu + \varepsilon) \leq e^{-nI(\mu+\varepsilon)} = e^{-n \inf\{I(x); x \geq \mu + \varepsilon\}}.$$

5. Montrer que :

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq 2e^{-n \inf\{I(x); |x - \mu| \geq \varepsilon\}}. \quad (\text{V.1})$$

6. Calculer explicitement la majoration exponentielle quand X suit la loi de Bernoulli de paramètre $p \in]0, 1[$.

La majoration exponentielle (V.1) est en fait vraie pour des variables aléatoires non bornées. La théorie des grandes déviations permet également d'obtenir un équivalent logarithmique de $\mathbb{P}(\bar{X}_n \leq \mu - \varepsilon)$.

△

V.2 Calcul de limites

Exercice V.10 (Convergence du maximum).

Soit X une variable aléatoire à valeurs dans \mathbb{N} .

1. Montrer que $\sum_{k=1}^{+\infty} \mathbb{P}(X \geq k) = \mathbb{E}[X]$.

On suppose que X est intégrable. Soit $(X_n, n \geq 1)$ une suite de variables aléatoires de même loi que X .

2. Soit $m \in \mathbb{N}^*$. Montrer que la variable aléatoire $Y_m = \sum_{n=1}^{+\infty} \mathbf{1}_{\{\frac{X_n}{n} \geq \frac{1}{m}\}}$ est p.s. finie.
3. En déduire que $\frac{X_n}{n}$ converge p.s. vers 0 quand n tend vers l'infini.
4. Montrer que $\frac{1}{n} \max(X_1, \dots, X_n)$ converge p.s. vers 0 quand n tend vers l'infini.

5. Soit X une variable aléatoire réelle intégrable et $(X_n, n \geq 1)$ une suite de variables aléatoires de même loi que X . Montrer que $\frac{1}{n} \max(X_1, \dots, X_n)$ converge p.s. vers 0 quand n tend vers l'infini.

△

Exercice V.11 (Calcul de limites d'intégrales).

Soit f une application continue bornée de \mathbb{R} dans \mathbb{R} .

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 :

$$\lim_{n \rightarrow \infty} \int_{[0,1]^n} f\left(\frac{x_1 + \dots + x_n}{n}\right) dx_1 \cdots dx_n.$$

2. Soit $(Y_k, k \geq 1)$ une suite de variables aléatoires indépendantes de loi de Poisson de paramètre $\alpha > 0$. 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 vers une limite que l'on déterminera. Calculer, en s'inspirant de la question 1 :

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

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

△

Exercice V.12 (Un équivalent de e^n).

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 $\frac{S_n - n}{\sqrt{n}} \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}(0, 1)$.

2. Déterminer la loi de S_n , puis montrer que $\lim_{n \rightarrow \infty} e^{-n} \left(1 + n + \frac{n^2}{2!} + \dots + \frac{n^n}{n!}\right) = \frac{1}{2}$.

△

V.3 Modélisation

Exercice V.13 (Paradoxe de Saint-Petersbourg).

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 économie².

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 euros 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 $S_n = X_1 + \dots + X_n$ 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 consacrée à l'étude de la convergence de la suite $(S_n, n \geq 1)$ convenablement renormalisée³.

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

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

où $\log_2 x$ est le logarithme en base 2 de $x > 0$: $2^{\log_2 x} = x$. Après avoir vérifié que pour tout $\varepsilon > 0$:

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

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

3. Calculer $\mathbb{P}(S_n \neq S'_n)$, et en déduire sa limite quand n tend vers l'infini.
4. En déduire que la suite $\left(\frac{S_n}{n \log_2 n}, n \geq 1\right)$ converge en probabilité vers 1.

△

Exercice V.14 (Sondage).

Précision des sondages.

2. G. Székely and D. Richards. The St. Petersburg paradox and the crash of high-tech stocks in 2000. *Amer. Statist.* vol. 58, pp. 225-231 (2004).

3. W. Feller. *An introduction to probability theory and its applications*, vol. 1. Wiley, third ed. (1968).

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 en 2000, on compte 6 millions de votants. Sachant qu'il y a eu environ 4 000 voix d'écart, quel est le nombre de personnes qu'il aurait fallu interroger dans un sondage pour savoir avec 95% de chance qui allait être le vainqueur ?

△

Exercice V.15 (Points de chute).

On considère la répartition des points de chute des bombes sur Londres lors de la Seconde Guerre Mondiale⁴.

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 tend vers l'infini et que $\lim_{m \rightarrow \infty} mp_m = \theta \in]0, \infty[$. 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 binomiaux $\binom{k}{m}$ est peu efficace.

2. Les données suivantes représentent le nombre de bombes qui sont tombées dans le sud de Londres pendant la Seconde Guerre Mondiale⁵. Le sud de Londres a été divisé en $N = 576$ domaines de taille 0.25 km^2 chacun. On a recensé dans la table V.1 le nombre 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 |

Table V.1. Nombre N_k de domaines du sud de Londres qui ont été touchés par exactement k bombes.

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.

△

4. W. Feller. *An Introduction to Probability Theory and Applications*, vol. 1, pp. 160-161. Wiley, third ed. (1968).

5. R. D. Clarke. An application of the Poisson distribution. *J. of Institute of Actuaries*, vol. 72, p. 481 (1946).

Exercice V.16 (Nombres absolument normaux).

L'objectif de cet exercice est de démontrer le théorème suivant dû à Borel (1909) : "Tout nombre réel choisi au hasard et uniformément 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 $x_n \in \{0, \dots, b-1\}$. Cette écriture est unique, sauf pour les fractions rationnelles de la forme $x = a/b^n$ et $a \in \{1, \dots, b^n - 1\}$. 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 $k \geq n+1$. On dit que x est simplement normal en base b si et seulement si pour tout $i \in \{0, \dots, b-1\}$, $\lim_{n \rightarrow \infty} \frac{1}{n} \text{Card} \{1 \leq k \leq n; x_k = i\}$ 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.

On dit que x est normal en base b si et seulement si il est simplement normal en base b^r pour tout $r \in \mathbb{N}^*$. Ainsi, 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 $i \in \{0, \dots, b-1\}^r$:

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

Le nombre de Champernowne⁶ dont la partie décimale est la suite consécutive des entiers (0.12345678910111213...) est normal en base 10. Les fractions rationnelles ne sont pas normales, quelle que soit leur représentation.

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

Soit X une variable aléatoire de loi uniforme sur $[0, 1]$.

1. Quelle est la loi de (X_1, \dots, X_n) , des n premiers chiffres du développement de X en base b ?
2. Calculer la loi de X_n . Montrer que les variables aléatoires $(X_n, n \geq 1)$ sont indépendantes et de même loi.
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.

6. D. G. Champernowne. The construction of decimals normal in the scale of ten. *J. London Math. Soc.*, vol. 8, pp. 254-260 (1933).

Bien que presque tous les réels soient absolument normaux, il est très difficile de montrer qu'un réel donné est absolument normal⁷. On ne sait toujours pas si des nombres tels que π , e , $\sqrt{2}$ ou $\log(2)$ sont absolument normaux, ni même normaux en base 10.

△

Exercice V.17 (Théorème de Weierstrass).

Cet exercice s'inspire de la démonstration de Bernstein du théorème de Weierstrass (1885) : "Toute fonction continue sur un intervalle fermé borné est limite uniforme d'une suite de polynômes".

Soit $(X_n, n \geq 1)$ une suite de variables aléatoires indépendantes de loi de Bernoulli de paramètre $x \in [0, 1]$. Pour $n \geq 1$, on considère la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$. Soit $h : [0, 1] \rightarrow \mathbb{R}$ une fonction continue. Soit $\delta > 0$. On pose $\Delta_n = \{|\bar{X}_n - x| > \delta\}$.

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

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

3. Quelle est la loi de $n\bar{X}_n$?
4. En déduire que :

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

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

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

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

△

7. J.-P. Delahaye. Champernowne et quelques autres. *Pour la Science*, pp. 102-106 (décembre 1998).

Exercice V.18 (Contamination au mercure).

Cet exercice présente un modèle pour la 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 l'infini :

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

Montrer que la suite $(Z_n, n \geq 1)$ définie par :

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

converge en loi vers la loi de Fréchet. On rappelle que la fonction de répartition de la loi de Fréchet de paramètre $(\alpha, \lambda) \in]0, \infty[^2$ est $y \mapsto \exp(-\alpha y^{-\lambda}) \mathbf{1}_{\{y > 0\}}$.

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 $0.71 \mu\text{g}$ par jour et par kilo de poids corporel. Des études statistiques⁸ donnent la forme de la queue de distribution empirique de la contamination globale annuelle en gramme de mercure pour un individu de 70 kg :

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

avec $\alpha = 3.54 \cdot 10^{-9}$ 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 Paris⁹ ? Dans une promotion de cent étudiants ? 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. P. Bertail. Evaluation des risques d'exposition à un contaminant alimentaire : quelques outils statistiques. www.crest.fr/doctravail/document/2002-41.pdf (2002).

9. Population du 15ème arrondissement de Paris, Recensement 1999 : 225 362 personnes.

VI

Vecteurs gaussiens

VI.1 Exemples

Exercice VI.1 (Exemple numérique).

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. Vérifier que $X_1 - X_2$ et $X_4 - X_2$ sont indépendants et écrire X comme la somme de quatre vecteurs gaussiens indépendants.

△

Exercice VI.2 (Couple de variables gaussiennes).

Soit X et Z deux variables aléatoires réelles indépendantes, X étant de loi gaussienne centrée réduite $\mathcal{N}(0, 1)$ 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 $\text{Cov}(X, Y)$.
3. Le vecteur (X, Y) est-il gaussien ? Les variables X et Y sont-elles indépendantes ?

△

Exercice VI.3 (Somme et produit de variables gaussiennes).

Soit 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.

△

VI.2 Propriétés et applications

Exercice VI.4 (Théorème de Cochran).

Le but de cet exercice est la démonstration du théorème de Cochran. Soit $n \geq 2$, et X_1, \dots, X_n des variables aléatoires indépendantes de même loi gaussienne centrée réduite $\mathcal{N}(0, 1)$. Soit $e = \{e_1, \dots, e_n\}$ 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 $U^t U = U U^t = I_n$, 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 $E_1 \oplus \dots \oplus E_p = \mathbb{R}^n$: 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 aléatoires 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 nombre de degré de liberté.
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 $T_n = \sum_{i=1}^n |X_i - \bar{X}_n|^2$, et donner la loi de T_n .

△

Exercice VI.5 (Moyenne et variance empirique de variables gaussiennes).

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, pour tout i , 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.6 (Moyenne et variance empirique indépendantes).

Soit X_1, \dots, X_n des variables aléatoires réelles indépendantes de même loi, de carré intégrable, d'espérance m et de variance σ^2 . On suppose que la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ et la variance empirique $V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ 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 :

$$\mathbb{E}[(n-1)V_n e^{itn\bar{X}_n}] = (n-1)\psi(t)^n \sigma^2.$$

2. En développant V_n dans l'égalité précédente, vérifier que :

$$\mathbb{E}[(n-1)V_n e^{itn\bar{X}_n}] = -(n-1)\psi''(t)\psi(t)^{n-1} + (n-1)\psi'(t)^2\psi(t)^{n-2}.$$

3. En déduire que, sur un voisinage ouvert de 0, ψ est solution de l'équation différentielle :

$$\begin{cases} \frac{\psi''}{\psi} - \left(\frac{\psi'}{\psi}\right)^2 = -\sigma^2, \\ \psi(0) = 1, \quad \psi'(0) = 0. \end{cases}$$

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 l'on ne suppose plus $m = 0$?

△

Exercice VI.7 (Estimation optimisée).

Soit $(X_n, n \geq 1)$, une suite de variables aléatoires indépendantes, de loi gaussienne $\mathcal{N}(\theta, \theta)$, 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 leur espérance, leur 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 de 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 de niveau asymptotique de 95%. Donner un tel intervalle pour la réalisation $\lambda = 0.5$, $n = 100$, $\bar{x}_n = 4.18$ et $v_n = 3.84$.
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 $\lambda_n^* = \frac{2V_n}{1 + 2V_n}$. 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 de niveau asymptotique de 95%. Donner un tel intervalle pour les valeurs numériques présentées à la question 10.

△

VII

Simulation

Exercice VII.1 (Loi exponentielle et loi géométrique).

Soit U une variable aléatoire de loi uniforme sur $[0, 1]$. Soit $\lambda > 0$.

1. Montrer que $T = -\log(U)/\lambda$ est une variable aléatoire exponentielle de paramètre λ .
2. Montrer que $X = [T] + 1$, où $[x]$ représente la partie entière de x est une variable aléatoire géométrique dont on déterminera le coefficient.

△

Exercice VII.2 (Loi de Poisson).

Soit $(U_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires 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 N défini par $N = \inf \left\{ n \in \mathbb{N}; \prod_{k=1}^{n+1} U_k < e^{-\theta} \right\}$.
4. En déduire une méthode pour simuler des variables aléatoires de Poisson.

△

Exercice VII.3 (Loi gaussienne).

Soit X, Y des variables indépendantes de loi $\mathcal{N}(0, 1)$.

1. On pose $R = \sqrt{X^2 + Y^2}$ et $\Theta \in [0, 2\pi[$, définis par $R \cos(\Theta) = X$ et $R \sin(\Theta) = Y$. Calculer la loi de (R, Θ) . En déduire que R et Θ sont indépendants.
2. Reconnaître la loi $e^{-R^2/2}$.
3. Soit U_1, U_2 des variables aléatoires indépendantes de loi uniforme sur $[0, 1]$. Déduire des questions précédentes la loi du couple (X', Y') défini par :

$$X' = \cos(2\pi U_1) \sqrt{2 |\log U_2|} \quad \text{et} \quad Y' = \sin(2\pi U_1) \sqrt{2 |\log U_2|}.$$

Cette transformation dite de Box-Muller permet de simuler simultanément deux variables aléatoires gaussiennes centrées réduites indépendantes.

△

Exercice VII.4 (Méthode du rejet).

Le but de cet exercice est de présenter la méthode du rejet pour la simulation d'une variable aléatoire de densité h donnée.

Soit X une variable aléatoire à valeurs dans \mathbb{R}^d et soit $A \subset \mathbb{R}^d$ un ensemble mesurable tel que $\mathbb{P}(X \in A) > 0$. Soit $(X_n, n \in \mathbb{N}^*)$ des variables aléatoires indépendantes de même loi que X . On pose $T = \inf\{n \in \mathbb{N}^*; X_n \in A\}$, avec la convention $\inf \emptyset = +\infty$, et $Y = X_T$ si $T < +\infty$ et $Y = 0$ si $T = +\infty$.

1. Montrer que les variables aléatoires Y et T sont indépendantes.
2. Montrer que la loi de T est la loi géométrique de paramètre $\mathbb{P}(X \in A)$.
3. Montrer que la loi de Y est la loi conditionnelle de X sachant $\{X \in A\}$: pour tout borélien $B \subset \mathbb{R}^d$, $\mathbb{P}(Y \in B) = \mathbb{P}(X \in B | X \in A)$.

Soit h la densité d'une variable aléatoire à valeurs dans \mathbb{R} . On suppose qu'il existe une densité g et une constante $c > 0$ telles que $ch \leq g$ (et que l'on sait simuler des variables aléatoires indépendantes de densité g).

Soit $(Z_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes de même loi de densité g . Soit $(U_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires de loi uniforme sur $[0, 1]$, indépendantes et indépendantes de $(Z_n, n \in \mathbb{N}^*)$. On pose $T' = \inf\{n \in \mathbb{N}^*; U_n \leq ch(Z_n)/g(Z_n)\}$ et $A' = \{(z, u); g(z) > 0 \text{ et } u \leq ch(z)/g(z)\}$.

4. Calculer $\mathbb{P}((Z_1, U_1) \in A')$.
5. Montrer que la variable aléatoire $Z_{T'}$ (que l'on peut donc simuler) a pour densité h .

△

VIII

Estimateurs

VIII.1 Modèles paramétriques usuels

Exercice VIII.1 (Estimation linéaire de la moyenne).

Soit X_1, \dots, X_n n variables aléatoires indépendantes de même loi et de carré intégrable. Trouver l'estimateur de la moyenne, $\theta = \mathbb{E}[X_1]$, qui soit de variance minimale dans la classe des estimateurs linéaires, $\hat{\theta}_n = \sum_{k=1}^n a_k X_k$, et sans biais. \triangle

Exercice VIII.2 (Loi exponentielle).

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 :

$$\hat{\lambda}_n^{(c)} = \frac{c}{\sum_{i=1}^n X_i}, \quad \text{où } c > 0.$$

5. Calculer le score, l'information de Fisher et la borne FDCR.
6. Les estimateurs étudiés font intervenir la statistique $S_n = \sum_{i=1}^n X_i$. 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 (Loi de Poisson).

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_n . 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_n est une loi binomiale de paramètres $(S_n, \frac{1}{n})$.
4. En déduire que $\delta_{S_n} = (1 - \frac{1}{n})^{S_n}$ 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 (Loi bêta).

On observe la réalisation d'un échantillon X_1, \dots, X_n de taille n de loi bêta $\beta(1, 1/\theta)$, $\theta > 0$, de densité :

$$p(x; \theta) = \frac{1}{\theta} (1-x)^{\frac{1}{\theta}-1} \mathbf{1}_{]0,1[}(x).$$

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 tend vers l'infini.

△

Exercice VIII.5 (Loi exponentielle et censure).

Soit Z et Y deux variables indépendantes suivant des lois exponentielles de paramètres respectifs $\lambda > 0$ et $\mu > 0$. On dispose d'un échantillon de n variables aléatoires indépendantes $(Z_1, Y_1), \dots, (Z_n, Y_n)$ de même loi que (Z, Y) .

1. S'agit-il d'un modèle exponentiel? Si oui, peut-on exhiber une statistique exhaustive?
2. Calculer l'estimateur du maximum de vraisemblance $(\hat{\lambda}_n, \hat{\mu}_n)$ de (λ, μ) .
3. Montrer qu'il est asymptotiquement normal et déterminer sa matrice de covariance asymptotique.

On suppose dorénavant que les observations sont censurées et que l'on observe seulement $X_i = \min(Z_i, Y_i)$ pour $i \in \{1, \dots, n\}$.

4. Calculer la fonction de répartition de la variable X_i .
5. Écrire le modèle statistique correspondant. Le modèle est-il identifiable? Quelle fonction de (λ, μ) est identifiable?
6. Donner les estimateurs du maximum de vraisemblance de $\gamma = \lambda + \mu$ fondés sur les observations :
 - a) de X_1, \dots, X_n ;
 - b) de $(Z_1, Y_1), \dots, (Z_n, Y_n)$.
 Est-il naturel que ces estimateurs soient différents?
7. Comparer les propriétés asymptotiques de ces estimateurs.

△

VIII.2 Modélisation

Exercice VIII.6 (Estimation de la qualité).

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 . Vérifier que δ_S est la fonction de répartition de la loi hypergéométrique¹ de paramètre (Nn, S, N) évaluée en k .

△

Exercice VIII.7 (Loi géométrique).

On considère la famille paramétrique des lois géométriques. Soit X une variable aléatoire à valeurs dans \mathbb{N}^* de loi géométrique de paramètre $\theta \in]0, 1[$.

1. Vérifier que le modèle est exponentiel. Donner une statistique exhaustive.
2. Déterminer $I(\theta)$, l'information de Fisher sur θ d'un échantillon de taille 1.

Soit X_1, \dots, X_n un échantillon de taille n de variables aléatoires indépendantes de même loi que X .

3. Déterminer $\hat{\theta}_n$, l'estimateur du maximum de vraisemblance de θ construit à partir de l'échantillon de taille n .
4. Montrer directement que l'estimateur du maximum de vraisemblance est asymptotiquement normal.
5. Donner un intervalle de confiance pour θ 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 inclus. Le tableau VIII.1 regroupe les données (simulées).

| | | | | | | | | | |
|----|-----|----|----|----|-----|----|----|----|-----|
| 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 |

Table VIII.1. Nombres (simulés) de passagers jusqu'au premier fraudeur inclus

1. On peut décrire la loi hypergéométrique dans un modèle d'urne. Soit une urne qui contient m boules dont b blanches et $m - b$ noires. Soit Y le nombre de boules blanches obtenues lors de n tirages sans remise. La loi de Y est la loi hypergéométrique de paramètres (m, b, n) . Pour $k \in \{\max(0, n + b - m), \dots, \min(n, b)\}$, on a :

$$\mathbb{P}(Y = k) = \frac{\binom{n}{k} \binom{m-n}{b-k}}{\binom{m}{b}} = \frac{\binom{b}{k} \binom{m-b}{n-k}}{\binom{m}{n}}.$$

6. Estimer la probabilité de fraude et donner un intervalle de confiance de niveau asymptotique 95%. Estimer le nombre de fraudeurs n_f si $n_0 = 20\,000$.

△

Exercice VIII.8 (Estimation d'évènements rares).

La hauteur maximale H de la crue annuelle d'un fleuve est modélisée par une variable aléatoire de Rayleigh, H , qui a pour densité :

$$f_H(x) = \mathbf{1}_{\mathbb{R}_+}(x) \frac{x}{a} \exp\left(-\frac{x^2}{2a}\right),$$

où $a > 0$ est un paramètre inconnu. Le tableau VIII.2 donne les hauteurs maximales annuelles de crue (simulées) sur 8 ans.

| | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|
| 2.5 | 1.8 | 2.9 | 0.9 | 2.1 | 1.7 | 2.2 | 2.8 |
|-----|-----|-----|-----|-----|-----|-----|-----|

Table VIII.2. Hauteurs maximales annuelles de crue (simulées)

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 et donner la variance asymptotique.
3. Une compagnie d'assurance estime qu'une catastrophe, qui correspond à une crue de plus de 6m, n'arrive qu'au plus une fois tous les mille ans. Ceci peut-il être justifié par les observations ? (En général les modèles paramétriques ne sont pas adaptés pour estimer des probabilités d'évènements rares. La théorie des valeurs extrêmes² a été développée pour répondre à ce type de question.)

△

Exercice VIII.9 (Modèle additif et modèle multiplicatif).

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 $\mathcal{N}(m_i, \sigma^2)$. On suppose les constantes $m_i > 0$ et $\sigma > 0$ connus. Une campagne publicitaire est menée

² L. de Haan. Fighting the arch-enemy with mathematics. *Stat. Neerl.*, vol. 44(2), pp. 45-68 (1990).

afin de permettre l'augmentation des ventes. On note X_i la vente mensuelle du magasin i après la campagne 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.3, 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 | |

Table VIII.3. Effet (simulé) d'une campagne publicitaire

△

Exercice VIII.10 (Contamination).

Dans l'industrie agro-alimentaire, on s'intéresse à la détection de la contamination³ 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 :

3. La méthode présentée, MPN ("most probable number"), est une méthode largement utilisée pour détecter des contaminations dans l'agro-alimentaire, mais également en environnement (rivière, ...).

- 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.

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 (non observée) désignant le nombre de spores présents dans le tube k et $X_k = \mathbf{1}_{\{Z_k=0\}}$ la variable aléatoire (observée) 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 négatifs, on effectue l'analyse sur de plus grands volumes. Dans un volume d fois ($d \geq 1$) plus grand, le nombre de spores suit une loi de Poisson de paramètre λd .
- Si on craint de n'observer que des tubes positifs, on ajoute des expériences sur des tubes dilués. Dans un tube dilué $1/d$ fois ($d \leq 1$), la densité en spores est égale à λd , et 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 une dilution égale à d_i , avec $i \in \{1, \dots, N\}$. 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. Donner un équivalent de la variance asymptotique de cet estimateur.
9. On étudie le cas particulier d'un petit nombre de dilutions. Donner le résultat littéral pour $N = 2$, $d_1 = 1$, $d_2 = \frac{1}{2}$. Donner les résultats numériques, estimation et intervalle de confiance de λ , si on observe $y_1 = 3$ et $y_2 = 6$ pour $n_1 = n_2 = 10$.

△

IX

Tests

IX.1 Mise en œuvre

Exercice IX.1 (Test de Neyman).

On distingue deux types de plantes à graines : les plantes qui disséminent leurs graines localement (type 1) et les plantes qui disséminent leurs graines très localement avec une grande probabilité et loin avec une faible probabilité (type 2). La loi de la position relative (X, Y) d'une graine par rapport à la plante est :

– pour le type 1, la loi uniforme sur $[-2, 2]^2$ de densité :

$$p_1(x, y) = \frac{1}{16} \mathbf{1}_{[-2;2]^2}(x, y);$$

– pour le type 2, la loi gaussienne centrée réduite de densité :

$$p_2(x, y) = \frac{1}{\sqrt{4\pi^2}} e^{-(x^2+y^2)/2}.$$

La graine d'une plante est observée à la position relative (x, y) par rapport à la plante. On souhaite savoir s'il s'agit d'une plante de type 2.

1. Décrire le test de Neyman de niveau $\alpha = 5\%$ et sa région critique.
2. Décrire le test de Neyman de niveau $\alpha = 1\%$ et sa région critique.

△

Exercice IX.2 (Loi exponentielle et modèle exponentiel).

Soit $X = (X_1, \dots, X_n)$ un échantillon de taille n de variables aléatoires indépendantes de même loi exponentielle de paramètre $1/\theta > 0$. Soit $\theta_0 > 0$ et $\alpha \in]0, 1[$.

1. Construire le test UPP de niveau α pour l'hypothèse nulle $H_0 = \{\theta = \theta_0\}$ contre l'hypothèse alternative $H_1 = \{\theta > \theta_0\}$.

2. Construire le test UPPS de niveau α pour l'hypothèse nulle $H_0 = \{\theta = \theta_0\}$ contre l'hypothèse alternative $H_1 = \{\theta \neq \theta_0\}$.

△

Exercice IX.3 (Test d'égalité des moyennes dans un modèle gaussien).

On considère un échantillon gaussien $(X_1, Y_1), \dots, (X_n, Y_n)$ 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.

1. Décrire le test de Wald pour tester l'hypothèse nulle $H_0 = \{\mu_1 = \mu_2\}$ contre l'hypothèse alternative $H_1 = \{\mu_1 \neq \mu_2\}$.
2. On suppose $\sigma_1 = \sigma_2$. Donner une région critique de niveau exact $\alpha \in]0, 1[$, construite à l'aide de la même statistique de test que celle utilisée dans la question précédente. Faire l'application numérique pour $n = 15$ et $\alpha = 5\%$.

△

Exercice IX.4 (Tests asymptotiques dans un modèle gaussien).

Soit $(X_n, n \geq 1)$, une suite de variables aléatoires indépendantes, de loi gaussienne $\mathcal{N}(\theta, \theta)$, avec $\theta > 0$. (La loi gaussienne $\mathcal{N}(\theta, \theta)$ apparaît comme approximation de la loi de Poisson de paramètre θ pour θ grand.) Soit $\theta_0 > 0$ fixé. L'objectif de cet exercice est de présenter deux tests pour l'hypothèse nulle $H_0 = \{\theta = \theta_0\}$ contre l'hypothèse alternative $H_1 = \{\theta > \theta_0\}$. 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 : $T_n^\lambda = \lambda \bar{X}_n + (1-\lambda)V_n$, $\lambda \in \mathbb{R}$.

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 la suite $(\sqrt{n}(T_n^\lambda - \theta), n \geq 2)$. 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.

7. On considère maintenant la valeur λ^* qui minimise $\lambda^2\theta_0 + 2(1-\lambda)^2\theta_0^2$. Comparer les variances asymptotiques de $T_n^{\lambda^*}$ et $\hat{\theta}_n$. Reprendre la question précédente avec $\lambda = \lambda^*$. Comparer asymptotiquement le test construit à partir de $T_n^{\lambda^*}$ et celui construit à partir de l'estimateur du maximum de vraisemblance.
8. On donne les valeurs numériques suivantes : $\theta_0 = 3.8$, $n = 100$, $\bar{x}_n = 4.18$ et $v_n = 3.84$. Calculer la p -valeur des deux tests. Quelle est votre décision ?

△

Exercice IX.5 (Test UPPS).

Soit un échantillon de taille n , $X = (X_1, \dots, X_n)$, de variables aléatoires indépendantes de même loi gaussienne de moyenne $\theta \in \mathbb{R}$ (inconnu) et de variance σ^2 connue. On note $p_n(x, \theta)$, pour $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ la densité du vecteur aléatoire X . On considère l'hypothèse nulle $H_0 = \{\theta = \theta_0\}$ et l'hypothèse alternative $H_1 = \{\theta = \theta_1\}$ avec $\theta_0 < \theta_1$ fixés. Soit $\alpha \in]0, 1[$.

1. Donner un test UPP de niveau α .
2. Quelle est la plus petite valeur de n permettant de construire un test UPP de niveau $\alpha \in]0, 1[$ et d'erreur de deuxième espèce inférieure ou égale à $\beta \in]0, 1[$?

On considère l'hypothèse alternative bilatère suivante $H'_1 = \{\theta \neq \theta_0\}$.

3. Peut-on construire un test UPP de H_0 contre H'_1 ?

On se propose de trouver un test de niveau α uniformément plus puissant sans biais parmi les tests purs pour tester l'hypothèse H_0 contre l'hypothèse H'_1 .

4. Soit $\theta_1 \in \mathbb{R}$ distinct de θ_0 . Montrer qu'il existe deux constantes $c(\theta_1)$ et $c^*(\theta_1)$ telles que l'ensemble défini par :

$$W'_n = \left\{ x; p_n(x, \theta_1) \geq c(\theta_1)p_n(x, \theta_0) + c^*(\theta_1)\frac{\partial p_n}{\partial \theta}(x, \theta_0) \right\},$$

est indépendant de θ_1 et vérifie :

$$\mathbb{P}_{\theta_0}(W'_n) = \mathbb{P}_{\theta_0}(X \in W'_n) = \alpha \quad \text{et} \quad \frac{\partial}{\partial \theta} \mathbb{P}_\theta(W'_n)|_{\theta=\theta_0} = 0.$$

5. Montrer que le test pur de région critique W'_n est UPP dans la classe des tests purs sans biais, *i.e.* vérifiant $\mathbb{P}_{\theta_0}(W''_n) = \alpha$ et $\mathbb{P}_{\theta}(W''_n) \geq \alpha$ pour tout $\theta \in \mathbb{R}$.

△

IX.2 Modélisation

Exercice IX.6 (Impartialité).

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 jurés¹. 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 sélection 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 seulement 40 jurés ?

△

Exercice IX.7 (Vaccin).

Des plaignants² ont poursuivi en justice le Ministère israélien 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 en moyenne 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.

1. M. O. Finkelstein and B. Levin. *Statistics for lawyers*. Springer, 2nd edition (2010).

2. M. Aitkin. Evidence and the Posterior Bayes Factor. *Math. Scientist*, vol. 17, pp. 15-25 (1992).

3. L'hypothèse nulle $H_0 = \{p = p_0\}$ correspond au risque que les plaignants aient accepté, l'hypothèse alternative étant $H_1 = \{p > p_0\}$. Construire un test à partir de la variable N . Donner la p -valeur de ce test. Accepte-t-on H_0 au seuil de 5% ?

△

Exercice IX.8 (Sondage).

L'objectif de cet exercice est d'étudier la variabilité des sondages.

1. Soient X_1, \dots, X_n des variables aléatoires indépendantes et identiquement distribuées de loi de Bernoulli $p \in [0, 1]$. On pose $\hat{p}_n = \frac{1}{n} \sum_{i=1}^n X_i$. Montrer l'inégalité $\text{Var}_p(\hat{p}_n) \leq 1/(4n)$.
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 à la prochaine élection municipale. 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, et 36% en décembre. Un éditorialiste dans son journal prend très au sérieux cette chute de 2 points. Confirmer ou infirmer la position du journaliste.

△

Exercice IX.9 (Une question de point de vue).

Dans les conditions usuelles d'emploi, la durée de vie des pièces produites par une machine sont modélisées par une loi gaussienne $\mathcal{N}(\mu_0, \sigma^2)$, avec $\mu_0 = 120$ heures et $\sigma = 19.4$ heures.

Le représentant d'un fournisseur propose un nouveau modèle de machine, en promettant un gain sur la durée de vie des pièces produites de 5% en moyenne, pour un écart-type identique σ .

On décide de tester le nouveau modèle de machine sur un échantillon de $n = 64$ pièces produites. On note $(X_i, i \in \{1, \dots, n\})$ les durées de vie des n pièces produites par le nouveau modèle de machine.

1. Quelle est la loi de $(X_i, i \in \{1, \dots, n\})$?
2. Soit μ la durée de vie moyenne des pièces produites par le nouveau modèle de machine. Donner un estimateur sans biais de μ . Identifier la loi de cet estimateur.
3. On ne souhaite pas changer de modèle si le nouveau n'est pas plus performant que l'ancien. Plus précisément, on souhaite que la probabilité d'adopter à tort le nouveau modèle ne dépasse pas le seuil de $\alpha = 0.05$. Quelle est alors la

procédure de décision construite à partir de l'estimateur de μ ? Les 64 pièces testées ont eu une durée de vie moyenne égale à 123.5 heures. Conclusion.

- Évaluez le risque que cette procédure vous fasse rejeter le nouveau modèle si l'annonce du représentant est exacte.

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%, tout en conservant le même seuil α pour ce test.

- 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. Conclusion.
- Quelle procédure peut-on proposer pour égaliser les risques de l'acheteur et du vendeur? Quel est alors ce risque?

△

Exercice IX.10 (Concurrence).

On reprend la problématique de l'exercice IX.9. Un représentant d'une autre société se présente et déclare avoir un produit moins cher et équivalent à celui des questions précédentes (de moyenne ν et de variance σ). L'acheteur le teste sur un échantillon de m pièces. Le résultat obtenu est une moyenne de 124.8 avec $m = 64$. On veut tester si les deux modèles sont de performances équivalentes *i.e.* si $\mu = \nu$.

- Expliciter $\hat{\mu}_n$ et $\hat{\nu}_m$ les estimateurs du maximum de vraisemblance de μ et ν .
- Construire un test pur à l'aide de la statistique de test $\hat{\mu}_n - \hat{\nu}_m$. Que peut-on dire des performances relatives des deux machines?

△

Exercice IX.11 (Production d'ailettes).

Une machine outil fabrique des ailettes de réacteur avec les caractéristiques suivantes : la longueur d'une ailette suit une loi gaussienne $\mathcal{N}(\ell_0, \sigma_0^2)$ avec $\ell_0 = 785$ mm et $\sigma_0 = 2$ mm. Une trop grande dispersion dans les caractéristiques de la machine peut induire une production d'ailettes trop longues, qui ne sont pas utilisables, ou d'ailettes trop courtes, qui nuisent aux performances du réacteur. On a donc calibré la machine afin de maîtriser au mieux le paramètre σ_0 , qui est donc considéré comme connu et invariable. En revanche, 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 la longueur moyenne ℓ reste dans l'intervalle de tolérance $[\ell_0 \pm \delta\ell_0]$ avec $\delta\ell_0 = 1.5$ mm. On procède à l'analyse de

$n = 200$ ailettes, ce qui conduit à une moyenne empirique $\bar{L}_n = \frac{1}{n} \sum_{i=1}^n L_i$ observée

$$\bar{L}_n^{\text{obs}} = 788.3 \text{ mm.}$$

1. Vérifier que le modèle est exponentiel. Construire un estimateur de ℓ .
2. On souhaite tester l'hypothèse $H_0 = \{\ell \in [\ell_0 \pm \delta\ell_0]\}$ contre $H_1 = \{\ell \notin [\ell_0 \pm \delta\ell_0]\}$. Construire un test bilatéral UPPS au seuil α pour tester H_0 contre H_1 .
3. Faire l'application numérique pour $\alpha = 5\%$ et conclure.
4. Calculer un intervalle de confiance de niveau 95% pour ℓ et le comparer à $[\ell_0 \pm \delta\ell_0]$.

△

IX.3 Tests du χ^2

Exercice IX.12 (Générateur de nombres aléatoires).

On souhaite vérifier la qualité du générateur de nombres aléatoires d'une calculatrice scientifique. Pour cela, on procède à 250 simulations à valeurs dans $\{0, \dots, 9\}$, voir les résultats du tableau IX.1. À l'aide du test du χ^2 , vérifier si le générateur donne des réalisations uniformément réparties sur $\{1, \dots, 9\}$.

| | | | | | | | | | | |
|--------|----|----|----|----|----|----|----|----|----|----|
| x | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| $N(x)$ | 28 | 32 | 23 | 26 | 23 | 31 | 18 | 19 | 19 | 31 |

Table IX.1. Résultats de 250 simulations d'un générateur de nombres aléatoires : $N(x)$ représente le nombre d'occurrences de x .

△

Exercice IX.13 (Césarienne).

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 du tableau IX.2 proviennent du "National Vital Statistics Report"³ 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 | Naissances | Naturelles | César. | Total |
|------------|------------|---------|-----------|------------|------------|--------|--------|
| J.O. | 2 331 536 | 663 540 | 2 995 076 | J.O. | 60.6 % | 17.3 % | 77.9% |
| W.E. | 715 085 | 135 493 | 850 578 | W.E. | 18.6 % | 3.5 % | 22.1% |
| Total | 3 046 621 | 799 033 | 3 845 654 | Total | 79.2 % | 20.8 % | 100.0% |

Table IX.2. Naissances aux U.S.A en 1997.

3. <http://www.cdc.gov/nchs/products/nvsr.htm>

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 des probabilités $p = (p_{J,N}, p_{W,N}, p_{J,C}, p_{W,C})$.
2. À l'aide d'un test du χ^2 , peut-on 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 , peut-on accepter ou rejeter l'hypothèse $p = p_0$, où p_0 correspond aux données de 1996, voir le tableau IX.3 ?

| Naissances | Naturelles | Césariennes |
|------------|------------|-------------|
| J.O. | 60.5 % | 17.0 % |
| W.E. | 18.9 % | 3.6 % |

Table IX.3. Naissances aux U.S.A en 1996.

△

Exercice IX.14 (Old Faithful).

On considère les données⁴ sur le geyser "Old Faithful" du parc national de Yellowstone aux États Unis d'Amérique. Elles sont constituées de 299 couples correspondant à la durée d'une éruption, et au temps d'attente correspondant au temps écoulé entre le début de cette éruption et la suivante. Les données sont collectées continûment du 1er au 15 août 1985.

Le temps d'attente (en minutes) moyen est de 72.3, l'écart-type de 13.9, et la médiane de 76. Un temps d'attente est dit court (C) s'il est inférieur à 76, et long (L) s'il est supérieur ou égal à 76. (On constate qu'un temps d'attente inférieur à 72 minutes est toujours suivi par une longue éruption et qu'un temps d'attente supérieur à 72 minutes est suivi en égale proportion par une éruption soit longue soit courte. La durée d'une éruption varie entre 2 et 5 minutes.)

On note $w_k \in \{C, L\}$ la durée qualitative du temps d'attente entre le début de la k -ième éruption et la suivante. On considère les couples $(X_k, Y_k) = (w_{2k-1}, w_{2k})$ pour $1 \leq k \leq K = 149$, qui représentent les durées qualitatives de deux temps

4. A. Azzalini and A. W. Bownman. A look at some data on Old Faithful. *Applied Statistics*, vol. 39, pp. 357-365 (1990).

d'attente successifs. On note $N_{i,j}$ pour $i, j \in \{C, L\}$ le nombre d'occurrences de (i, j) pour la suite $((X_k, Y_k), 1 \leq k \leq K)$. On observe les valeurs suivantes :

$$N_{C,C} = 9, \quad N_{C,L} = 70, \quad N_{L,C} = 55 \quad \text{et} \quad N_{L,L} = 15.$$

On suppose que les variables aléatoires $((X_k, Y_k), 1 \leq k \leq K)$ sont indépendantes et de même loi.

1. Tester si deux temps d'attente successifs sont indépendants (i.e. X_k est indépendant de Y_k). Pourquoi n'utilise-t-on pas les occurrences des couples (w_k, w_{k+1}) pour ce test ?
2. Tester si deux temps d'attente successifs sont indépendants et de même loi (i.e. X_k est indépendant de Y_k et X_k a même loi que Y_k). Êtes-vous étonnés de votre conclusion vu la réponse à la question précédente ?

△

Exercice IX.15 (Test de McNemar ou d'égalité des lois marginales).

On considère deux juges qui évaluent les mêmes situations et répondent chaque fois par l'affirmative (+) ou la 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 nulle $H_0 = \{p_+^{(1)} = p_+^{(2)}\}$, ou non, hypothèse alternative $H_1 = \{p_+^{(1)} \neq p_+^{(2)}\}$. On suppose que toutes les réponses possibles pour les deux juges (soit quatre cas possibles) ont une probabilité strictement positive.

1. Vérifier que, sous H_0 , la loi du couple de réponses ne dépend que de deux paramètres α et β tels que $0 < \alpha < \beta < 1 - \alpha$, $\beta = p_+^{(i)}$ et le tableau IX.4 donne les probabilités des réponses possibles.

| | | |
|-----------------|------------------|----------------------|
| | rép. juge 2 = + | rép. juge 2 = - |
| rép. juge 1 = + | $\beta - \alpha$ | α |
| rép. juge 1 = - | α | $1 - \alpha - \beta$ |

Table IX.4. Probabilité d'observer les réponses des juges 1 et 2.

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 de niveau asymptotique 5%.

4. Pour $n = 200$, on observe $N_{+-} = 15$ et $N_{-+} = 5$, calculer la p -valeur asymptotique du test et conclure.

△

Exercice IX.16 (Réactions à des vaccins).

On se propose de comparer les réactions produites par deux vaccins B.C.G. conçus par A et B ⁵. Un groupe de 348 enfants a été divisé par tirage au sort en deux séries qui ont été vaccinées, l'une par le vaccin de A , l'autre par le vaccin de B . La réaction a été ensuite lue par une personne ignorant le vaccin utilisé. Les résultats figurent dans le tableau IX.5. En ce qui concerne les réactions, peut-on dire que les vaccins de A et de B sont identiques?

| 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 |

Table IX.5. Réactions à deux vaccins de B.C.G..

△

Exercice IX.17 (Accidents de cavalerie).

Le tableau IX.6 donne, sur une période de vingt ans (1875-1894), le nombre de décès par an et par régiment dans la cavalerie prussienne causés par un coup de sabot de cheval⁶. On dispose de 280 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 | 144 | 91 | 32 | 11 | 2 |

Table IX.6. Décès par an et par régiment.

△

5. D. Schwartz et P. Lazar. *Éléments de statistique médicale et biologique*. Flammarion (1964).

6. A. Gulberg. Les fonctions de fréquence discontinues et les séries statistiques. *Annales de l'I. H. P.*, vol. 3, pp.229-278 (1933).

Exercice IX.18 (Familles à garçons ou familles à filles?).

La répartition du nombre de garçons et de filles dans 5128 familles⁷ de cinq enfants est donnée dans le tableau IX.7. On veut savoir si ce résultat est compatible avec l'hypothèse d'équiprobabilité de la naissance d'un garçon et de la naissance d'une fille. On note r la probabilité d'avoir un garçon.

| Nombres de garçons et de filles | (5,0) | (4,1) | (3,2) | (2,3) | (1,4) | (0,5) | Total |
|---------------------------------|-------|-------|-------|-------|-------|-------|-------|
| Nombre de familles | 204 | 841 | 1585 | 1544 | 810 | 144 | 5128 |

Table IX.7. Nombres de garçons et de filles dans 5128 familles de cinq enfants.

1. Calculer la proportion de garçons. Utiliser le test du χ^2 de niveau asymptotique $\alpha = 0.01$, basé sur cette proportion, pour déterminer si $r = 1/2$. Donner la p -valeur asymptotique de ce test.
2. Donner un intervalle de confiance pour r de niveau asymptotique α . Remarquer que le test du χ^2 est équivalent à l'intervalle de confiance.
3. Utiliser le test du χ^2 de niveau asymptotique $\alpha = 0.01$, directement à partir des données du tableau IX.7. Donner approximativement la p -valeur asymptotique de ce test. Conclusion? (En général, on obtient un test plus précis en gardant la répartition par famille, c'est-à-dire en tenant compte de toute l'information contenue dans les données. La répartition des données en sous-ensembles est appelée méthode de stratification.)
4. Utiliser le test du χ^2 pour vérifier si les données du tableau IX.7 suivent une loi binomiale de paramètre $(5, r)$, avec r inconnu. Donner approximativement la p -valeur asymptotique de ce test. Conclusion?

△

Exercice IX.19 (Les dés de Weldon).

Weldon a effectué $n = 26\ 306$ lancers de douze dés à six faces⁸. On note X_i le nombre de faces indiquant cinq ou six lors du i -ième lancer. Les fréquences empiriques observées sont notées $\hat{p}_j = N_j/n$, 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.8 et IX.9.

7. M.-P. Schützenberger. Résultats d'une enquête sur la distribution du sexe dans les familles nombreuses. *Semaine des Hôpitaux de Paris*, vol. 25(61), pp. 2579–2582 (1949).

8. W. Feller. *An introduction to probability theory and its applications*, vol. 1, p. 148. Wiley, third ed. (1968).

| | | | |
|--------------|--------------|---------------|--------------|
| $N_0 = 185$ | $N_1 = 1149$ | $N_2 = 3265$ | $N_3 = 5475$ |
| $N_4 = 6114$ | $N_5 = 5194$ | $N_6 = 3067$ | $N_7 = 1331$ |
| $N_8 = 403$ | $N_9 = 105$ | $N_{10} = 14$ | $N_{11} = 4$ |
| $N_{12} = 0$ | | | |

Table IX.8. Observations.

| | | | |
|---------------------------|------------------------|---------------------------|---------------------------|
| $\hat{p}_0 = 0.007033$ | $\hat{p}_1 = 0.043678$ | $\hat{p}_2 = 0.124116$ | $\hat{p}_3 = 0.208127$ |
| $\hat{p}_4 = 0.232418$ | $\hat{p}_5 = 0.197445$ | $\hat{p}_6 = 0.116589$ | $\hat{p}_7 = 0.050597$ |
| $\hat{p}_8 = 0.015320$ | $\hat{p}_9 = 0.003991$ | $\hat{p}_{10} = 0.000532$ | $\hat{p}_{11} = 0.000152$ |
| $\hat{p}_{12} = 0.000000$ | | | |

Table IX.9. Fréquences empiriques observées.

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.10.

| | | | |
|---------------------|------------------|---------------------|---------------------|
| $p_0 = 0.007707$ | $p_1 = 0.046244$ | $p_2 = 0.127171$ | $p_3 = 0.211952$ |
| $p_4 = 0.238446$ | $p_5 = 0.190757$ | $p_6 = 0.111275$ | $p_7 = 0.047689$ |
| $p_8 = 0.014903$ | $p_9 = 0.003312$ | $p_{10} = 0.000497$ | $p_{11} = 0.000045$ |
| $p_{12} = 0.000002$ | | | |

Table IX.10. 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 ?

△

X

Intervalles et régions de confiance

Exercice X.1 (Loi exponentielle symétrique décalée).

Soit X une variable aléatoire exponentielle symétrique décalée de densité :

$$f(x) = \frac{1}{2} e^{-|x-\theta|}, \quad \forall x \in \mathbb{R},$$

où $\theta \in \mathbb{R}$ est un paramètre de décalage inconnu.

1. Calculer $\mathbb{E}_\theta[X]$ et $\text{Var}_\theta(X)$. Construire un estimateur T_n de θ à partir d'un échantillon, X_1, \dots, X_n , de n variables aléatoires indépendantes et de même loi que X .
2. Construire un intervalle de confiance de niveau asymptotique 95% de θ pour $n = 200$.

△

Exercice X.2 (Loi uniforme).

Soit X_1, \dots, X_n un échantillon de n variables aléatoires indépendantes de loi uniforme sur $[0, \theta]$ où $\theta > 0$ est un paramètre inconnu. Soit $\theta_0 > 0$. On veut tester l'hypothèse nulle $H_0 = \{\theta \leq \theta_0\}$ contre l'hypothèse alternative $H_1 = \{\theta > \theta_0\}$ au seuil α .

1. Donner le test pur de Neyman correspondant et la région critique W_n associée.
2. Calculer la puissance du test $\rho(\theta)$.
3. On choisit $\theta_0 = \frac{1}{2}$. 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 considère les hypothèses $H'_0 = \{\theta = \theta_0\}$ et $H'_1 = \{\theta > \theta_0\}$. Quel test proposer ?

6. Soit $\alpha' > 0$ donné. Calculer $k_n \geq 1$ tel que $\mathbb{P}_\theta(\theta \leq k_n \max_{1 \leq i \leq n} X_i)$ soit égal 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.3 (Intervalle de confiance d'une fréquence).

On considère des variables aléatoires $(X_n, n \in \mathbb{N}^*)$ indépendantes de loi de Bernoulli de même paramètre $p \in]0, 1[$ inconnu. Soit $\alpha \in]0, 1[$. On désire donner plusieurs intervalles de confiance pour p de niveau exact ou approché $1 - \alpha$, construits à partir de la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ de l'échantillon de taille n .

1. Vérifier que $\text{Var}_p(\bar{X}_n) \leq 1/4n$. Déduire de l'inégalité de Tchebychev, un intervalle de confiance (symétrique) de niveau par excès $1 - \alpha$ pour p .
2. Montrer que $\bar{X}_n(1 - \bar{X}_n)$ est un estimateur convergent de $\sigma^2 = \text{Var}_p(X_1)$. Déduire du théorème central limite un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour p .
3. On considère la fonction $g_n(x, p) = \frac{\sqrt{n}(x - p)}{\sqrt{p(1 - p)}}$. Montrer que la suite $(g_n(\bar{X}_n, p), n \in \mathbb{N}^*)$ converge en loi vers une loi indépendante du paramètre p (la suite de fonction $(g_n, n \in \mathbb{N}^*)$ est asymptotiquement pivotale). En considérant l'inverse de $p \mapsto g_n(x, p)$, donner un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour p .
4. En utilisant la majoration $x(1 - x) \leq 1/4$, déduire du théorème central limite un intervalle de confiance de niveau asymptotique par excès $1 - \alpha$ pour p .
5. Trouver une fonction g telle que $g'(\mathbb{E}_p[X_1])^2 \text{Var}_p(X_1)$ soit constant pour tout $p \in]0, 1[$. Montrer que la suite $(\sqrt{n}(g(\bar{X}_n) - g(p)), n \in \mathbb{N}^*)$ converge en loi et déterminer la limite. En déduire un intervalle de confiance de niveau asymptotique $1 - \alpha$ pour p (méthode de stabilisation de la variance).
6. Soit $1 > a > p$. Montrer, en utilisant l'inégalité de Markov, que $\mathbb{P}_p(\bar{X}_n > a) \leq \mathbb{E}_p \left[e^{\lambda \bar{X}_n - \lambda a} \right]$ pour tout $\lambda > 0$. En déduire que :

$$\mathbb{P}_p(\bar{X}_n > a) \leq \exp -n \left(a \log(a/p) + (1 - a) \log((1 - a)/(1 - p)) \right).$$

Déduire de cette majoration une majoration de $\mathbb{P}_p(\bar{X}_n < a)$ pour $0 < a < p$. Construire un intervalle de confiance de niveau par excès $1 - \alpha$ pour p .

7. Donner les intervalles de confiance lorsque l'on observe $\bar{X}_n = 0$ pour n grand. Conclure.

△

XI

Problèmes (probabilités)

XI.1 Loi uniforme sur la sphère

Exercice XI.1 (Loi gaussienne et loi uniforme sur la sphère).

Le but de l'exercice est de démontrer le résultat suivant, attribué (à tort¹) à H. Poincaré : en dimension n , la première coordonnée d'une variable aléatoire de loi uniforme sur la sphère de rayon \sqrt{n} , est approximativement de loi gaussienne $\mathcal{N}(0, 1)$.

Pour $n \geq 2$ et $y_n = (x_1, \dots, x_n) \in \mathbb{R}^n$, on note $|y_n| = \sqrt{\sum_{k=1}^n x_k^2}$ la norme de y_n , et $\theta_{n-1} = y_n/|y_n|$ la direction du vecteur y_n , avec θ_{n-1} un élément de $\mathcal{S}_{n-1} = \{z_n \in \mathbb{R}^n; |z_n| = 1\}$ la sphère de rayon 1 de \mathbb{R}^n .

Soit $(X_n, n \in \mathbb{N}^*)$ des variables aléatoires indépendantes de loi gaussienne centrée réduite $\mathcal{N}(0, 1)$. On considère $Y_n = (X_1, \dots, X_n)$ le vecteur aléatoire dans \mathbb{R}^n et on pose :

$$R_n = |Y_n| = \sqrt{\sum_{k=1}^n X_k^2} \quad \text{et} \quad \Theta_{n-1} = \frac{Y_n}{R_n}.$$

1. Après avoir étudié la convergence de $(\frac{1}{n} \sum_{k=1}^n X_k^2, n \geq 2)$, montrer que la suite $(R_n/\sqrt{n}, n \geq 2)$ converge p.s. vers 1. Autrement dit, en grande dimension, Y_n se concentre p.s. sur la sphère de rayon \sqrt{n} .

On note $\sigma_{n-1}(d\theta_{n-1})$, avec $\theta_{n-1} \in \mathcal{S}_{n-1}$, la mesure de Lebesgue sur \mathcal{S}_{n-1} . On dit qu'une variable aléatoire Z_n à valeurs dans \mathbb{R}^n est de loi uniforme sur \mathcal{S}_{n-1} si pour toute fonction bornée g définie sur \mathcal{S}_{n-1} , on a :

$$\mathbb{E}[g(Z_n)] = \frac{1}{c_{n-1}} \int_{\mathcal{S}_{n-1}} \sigma_{n-1}(d\theta_{n-1}) g(\theta_{n-1}),$$

1. P. Diaconis and D. Freedman. A dozen de Finetti-style results in search of a theory. *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 23(2) : pp. 397-423, (1987).

où $c_{n-1} = \sigma_{n-1}(\mathcal{S}_{n-1})$ est la surface de la sphère \mathcal{S}_{n-1} .

On rappelle la formule de changement de variable suivante, obtenue en posant $r_n = |y_n|$ et $\theta_{n-1} = y_n/r_n$, pour toute fonction h mesurable intégrable sur \mathbb{R}^n :

$$\int_{\mathbb{R}^n} h(y_n) dy_n = \int_{]0, +\infty[} r_n^{n-1} dr_n \int_{\mathcal{S}_{n-1}} \sigma_{n-1}(d\theta_{n-1}) h(r_n \theta_{n-1}).$$

2. Donner la densité de Y_n , puis montrer que $\Theta_{n-1} = Y_n/|Y_n|$ est de loi uniforme sur \mathcal{S}_{n-1} .
3. Pour $n \geq 2$, soit $U^{(n)} = (U_1^{(n)}, \dots, U_n^{(n)})$ une variable aléatoire de loi uniforme sur la sphère de \mathbb{R}^n de rayon \sqrt{n} , autrement dit $U^{(n)}/\sqrt{n}$ est de loi uniforme sur \mathcal{S}_{n-1} . Remarquer que $U^{(n)}$ a même loi que $\sqrt{n} Y_n/|Y_n|$. Puis, montrer que $(U_1^{(n)}, n \geq 2)$ converge en loi vers la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.

△

XI.2 Le collectionneur

Exercice XI.2 (Le collectionneur).

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 $n \geq 1$ images différentes et qu'elles sont équitablement réparties, à raison d'une par tablette. On note T_n le plus petit nombre de tablettes qu'il faut acheter pour obtenir une image en double.

1. Donner un modèle probabiliste élémentaire qui permette d'étudier T_n . Montrer que pour $k \in \{1, \dots, n\}$, on a :

$$\mathbb{P}(T_n > k) = \prod_{i=1}^k \left(1 - \frac{i-1}{n}\right).$$

2. Montrer en utilisant les fonctions de répartition que $(T_n/\sqrt{n}, n \geq 1)$ converge en loi vers une variable aléatoire X . (On rappelle que $\log(1+x) = x + O(x^2)$ au voisinage de 0.)
3. Montrer que la loi de X possède une densité et la calculer.
4. Déterminer et reconnaître la loi de X^2 .
5. On considère un groupe de k personnes. On suppose qu'il n'y a pas d'année bissextile. Donner une approximation de la probabilité, p_k , pour que deux personnes du groupe au moins aient la même date d'anniversaire. On traitera les valeurs suivantes de k : $k = 20$, $k = 40$ et $k = 366$.



Exercice XI.3 (Le collectionneur).

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 $X_i \in \{1, \dots, n\}$ 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 $k - 1$. Le but de cet exercice est l'étude du nombre de tablettes, N_n , à acheter pour avoir la collection complète. Ce problème est connu sous le nom du problème du collectionneur de coupons^{2,3} ou "coupon collector" en anglais.

1. Quelle loi proposez-vous pour la suite de variables aléatoires $(X_i, i \in \mathbb{N}^*)$?
2. Calculer $\mathbb{P}(T_2 = \ell)$ pour $\ell \in \mathbb{N}^*$. En déduire que T_2 suit une loi géométrique dont on précisera le paramètre.
3. Montrer que pour $\ell_2, \ell_3 \in \mathbb{N}^*$:

$$\begin{aligned} \mathbb{P}(T_2 = \ell_2, T_3 = \ell_3) &= \sum_{\substack{1, j_2, j_3 \text{ distincts,} \\ j_1, j_2, j_3 \in \{1, \dots, n\}}} \mathbb{P}(X_1 = j_1, \dots, X_{\ell_2} = j_1, X_{\ell_2+1} = j_2, \\ &\quad \dots, X_{\ell_2+\ell_3} \in \{j_1, j_2\}, X_{\ell_2+\ell_3+1} = j_3). \end{aligned}$$

4. En déduire que T_3 suit une loi géométrique dont on précisera le paramètre.
5. Vérifier que T_2 et T_3 sont indépendants.
6. Décrire T_k comme premier instant de succès et en déduire sa loi.

On admet dorénavant que les variables aléatoires T_1, T_2, \dots, T_n sont indépendantes.

7. Calculer $\mathbb{E}[N_n]$ et vérifier que $\mathbb{E}[N_n] = n(\log(n) + O(1))$ où $O(1)$ désigne une fonction g telle que $\sup_{n \geq 1} |g(n)| \leq M < \infty$.
8. Calculer $\text{Var}(N_n)$ et en donner un équivalent quand n tend vers l'infini.
9. Soit $\varepsilon > 0$. Majorer $\mathbb{P}\left(\left|\frac{N_n}{\mathbb{E}[N_n]} - 1\right| > \varepsilon\right)$.

2. P. Flajolet and R. Sedgewick. *Analytic combinatorics, Symbolic Combinatorics*. Cambridge University Press, <http://algo.inria.fr/flajolet/Publications/books.pdf> (2009).

3. D. Aldous. *Probability approximations via the Poisson clumping heuristic*. Springer (1989).

10. Montrer que la suite $\left(\frac{N_n}{n \log n}, n \in \mathbb{N}^*\right)$ converge en probabilité vers 1.

△

Exercice XI.4 (Le collectionneur).

Soit $(X_k, k \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes de loi exponentielle $\mathcal{E}(\lambda)$ 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. Soit $n \geq 2$. On note $X_{(1)} \leq \dots \leq X_{(n)}$ le réordonnement croissant de X_1, \dots, X_n , appelé aussi statistique d'ordre. 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 $k \in \{2, \dots, n\}$,

$$Y_k = X_{(k)} - X_{(k-1)}.$$

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 tend vers l'infini. Dans un deuxième temps nous étudierons la loi du vecteur (Y_1, \dots, Y_n) . Enfin nous donnerons une application de ces deux résultats au problème du collectionneur 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, \dots, 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 $X_{(1)} < \dots < X_{(n)}$. 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 :

$$\mathbb{E}[g_1(Y_1)g_2(Y_2)] = \int g_1(y_1)g_2(y_2) 2\lambda^2 e^{-2\lambda y_1 - \lambda y_2} \mathbf{1}_{\{y_1 > 0, y_2 > 0\}} dy_1 dy_2.$$

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 $\{X_{\sigma(1)} < \dots < X_{\sigma(n)}\}$, où σ parcourt l'ensemble \mathcal{S}_n des permutations de $\{1, \dots, n\}$, montrer que pour $n \geq 3$:

$$\mathbb{E}[g_1(Y_1) \cdots g_n(Y_n)] = n! \mathbb{E}[g_1(X_1)g_2(X_2 - X_1) \cdots g_n(X_n - X_{n-1}) \mathbf{1}_{\{X_1 < \dots < X_n\}}],$$

où les fonctions g_1, \dots, g_n sont mesurables bornées. On pourra utiliser, sans le justifier, le fait que les vecteurs $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$ et (X_1, \dots, X_n) ont même loi.

6. Calculer la loi de Y_i , $i \in \{1, \dots, n\}$. Montrer que les variables aléatoires Y_1, \dots, Y_n sont indépendantes. Donner la loi du vecteur (Y_1, \dots, Y_n) .
7. En déduire la fonction caractéristique de $X_{(n)}$.

III Application

Votre petit frère collectionne les images de Pokémon 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 $T_{1,n} + \dots + T_{n,n} = N_n$ plaquettes. On admet (voir l'exercice XI.3) que les variables aléatoires $T_{1,n}, \dots, T_{n,n}$ sont indépendantes et que la loi de $T_{k,n}$ est la loi géométrique de paramètre $1 - \frac{k-1}{n}$. On admet de plus que $\mathbb{E}[N_n] \sim n \log n$ et que $\frac{N_n}{n \log n}$ converge en probabilité vers 1 quand $n \rightarrow +\infty$. Le but de cette partie est de déterminer à quelle vitesse a lieu cette convergence⁴ et d'en déduire un intervalle aléatoire 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 $\frac{1}{n} T_{n-k+1,n}$, où $k \in \{1, \dots, n\}$ et ψ_k la fonction caractéristique de la loi exponentielle de paramètre k .

4. P. Erdős and A. Rényi. On a classical problem of probability theory. *Publ. Math. Inst. Hung. Acad. Sci.*, Ser. A, vol. 6, pp. 215–220 (1961).

1. Montrer que la suite $\left(\frac{1}{n} T_{n-k+1,n}, n \geq k\right)$ converge en loi vers la loi exponentielle de paramètre k .

On admet que pour tout $u \in \mathbb{R}$, il existe $n(u)$ et C , tels que pour tout $n \geq n(u)$ et $k \in \{1, \dots, n\}$, on a :

$$|\psi_{k,n}(u) - \psi_k(u)| \leq \frac{1}{kn} C. \quad (\text{XI.1})$$

On rappelle que si $a_k, b_k, k \in \{1, \dots, n\}$ sont des complexes de modules inférieurs à 1 ($|a_k| \leq 1$ et $|b_k| \leq 1$) alors on a :

$$\left| \prod_{k=1}^n a_k - \prod_{k=1}^n b_k \right| \leq \sum_{k=1}^n |a_k - b_k|. \quad (\text{XI.2})$$

2. 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 a pour n grand et des variables aléatoires indépendantes :

$$\frac{1}{n} \text{Géom}(1) + \frac{1}{n} \text{Géom}\left(\frac{n-1}{n}\right) + \dots + \frac{1}{n} \text{Géom}\left(\frac{1}{n}\right) \stackrel{\text{loi}}{\simeq} \mathcal{E}(n) + \mathcal{E}(n-1) + \dots + \mathcal{E}(1).$$

3. Montrer 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 aléatoire, I_n , de niveau asymptotique $\alpha = 95\%$ pour N_n : $\mathbb{P}(N_n \in I_n) \simeq 95\%$ pour n grand.
5. Quel est le nombre moyen de plaquettes de chocolat que votre petit frère doit acheter pour avoir la collection de $n = 151$ Pokémons complète. Donner un intervalle aléatoire qui contienne avec une probabilité de 95% le 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 de la deuxième édition.

△

XI.3 Le paradoxe du bus

Exercice XI.5 (Paradoxe du bus).

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 cet exercice est de déterminer à l'aide d'un modèle simple qui, de la société de bus ou du client, a raison. Le paradoxe du bus est aussi connu dans la littérature sous les noms de "inspection paradox", "Feller's paradox"⁵ ou "waiting time paradox"⁶.

On note T_1 le temps de passage du premier bus et T_{i+1} le temps entre le passage du i -ème et du $i + 1$ -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. On note $z_+ = \max(z, 0)$. Vérifier que, pour $n \geq 1$:

$$\mathbb{P}(R_t \leq r, S_t \leq s, N_t = n) = e^{-\lambda t} (1 - e^{-\lambda s}) \frac{\lambda^n}{n!} [t^n - (t - r)_+^n].$$

5. W. Feller. *An introduction to probability and its applications*, vol. 2. Wiley (1966).

6. L. Takács. *Introduction to the theory of queues*. Oxford University Press (1962).

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 variables R_t et S_t sont indépendantes, autrement dit que les évènements $\{R_t \leq r\}$ et $\{S_t \leq s\}$ sont indépendants pour tout $(r, s) \in \mathbb{R}^2$.
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 $U_t = R_t + S_t$.

△

Exercice XI.6 (Paradoxe du bus).

Les parties I et II sont indépendantes.

Soit T une variable aléatoire p.s. strictement positive, i.e. $\mathbb{P}(T > 0) = 1$.

I Le temps d'attente à l'arrêt de bus

Soit $(T_n, n \geq 1)$ une suite de variables aléatoires indépendantes et de même loi que T . Soit $n \geq 2$ fixé. À l'instant $t = 0$ un premier bus (bus 1) passe devant l'arrêt de bus. Le temps écoulé entre le passage des k -ième et $(k + 1)$ -ième bus est modélisé par la variable aléatoire T_k . L'expérience est terminée lors du passage du $(n + 1)$ -ième bus. On pose, pour $k \in \mathbb{N}^*$, $S_k = \sum_{i=1}^k T_i$ le temps de passage du $(k + 1)$ -ième bus, et $S_0 = 0$.

On suppose qu'un client arrive au hasard après le premier bus ($t = 0$) et avant le dernier bus ($t = S_n$). On note N_n^* le numéro du dernier bus passé avant son arrivée, et $T_n^* = T_{N_n^*}$ le temps écoulé entre le passage du bus juste avant son arrivée et le passage du bus juste après son arrivée. Le but de cette partie est de déterminer la loi de T_n^* , et de donner son comportement asymptotique quand n tend vers l'infini.

Soit U une variable aléatoire de loi uniforme sur $[0, 1]$ et indépendante de $(T_n, n \geq 1)$. On modélise le temps d'arrivée du client par US_n .

1. Expliquer brièvement pourquoi on modélise le temps d'arrivée du client par US_n .
2. Montrer que les variables aléatoires T_k/S_n ont même loi.
3. En déduire $\mathbb{E}[X_n] = 1$, où $X_n = \frac{nT_1}{S_n}$.
4. Soit Z une variable aléatoire à valeurs dans \mathbb{R}^n de densité h , indépendante de U . Montrer que pour toute fonction φ bornée mesurable, on a :

$$\mathbb{E}[\varphi(U, Z)] = \mathbb{E} \left[\int_0^1 du \varphi(u, Z) \right]. \quad (\text{XI.3})$$

On admettra que (XI.3) est vraie pour toute variable aléatoire Z vectorielle indépendante de U .

5. Remarquer que pour $1 \leq k \leq n$, on a $\{N_n^* = k\} = \{U \in]S_{k-1}/S_n, S_k/S_n]\}$. Soit g une fonction bornée mesurable. Montrer, en décomposant suivant les valeurs de N_n^* et en utilisant (XI.3), que :

$$\mathbb{E}[g(T_n^*)] = \mathbb{E} \left[\frac{nT_1}{S_n} g(T_1) \right]. \quad (\text{XI.4})$$

On suppose que T est intégrable et on pose $\mu = \mathbb{E}[T]$.

6. Montrer que la suite $(X_n, n \geq 1)$ converge p.s. vers une limite X . Calculer $\mathbb{E}[X]$.
7. On définit la fonction $\varphi_+(x) = \max(x, 0)$. Vérifier que la fonction φ_+ est continue et que $\varphi_+(x - y) \leq x$ pour tout $x \geq 0, y \geq 0$. Montrer que $\lim_{n \rightarrow \infty} \mathbb{E}[\varphi_+(X - X_n)] = 0$.
8. Vérifier que $|x| = -x + 2\varphi_+(x)$ pour tout $x \in \mathbb{R}$. En déduire alors que $\lim_{n \rightarrow \infty} \mathbb{E}[|X - X_n|] = 0$.
9. Soit g une fonction bornée mesurable. Déduire de la question précédente que la limite $\lim_{n \rightarrow \infty} \mathbb{E}[g(T_n^*)]$ existe et la déterminer.
10. En utilisant (XI.4) pour un choix judicieux de fonctions g , montrer que la suite $(T_n^*, n \geq 1)$ converge en loi vers une limite T^* quand n tend vers l'infini. La loi de T^* est appelée loi de T biaisée par la taille.

II Loi biaisée par la taille

On suppose que T est intégrable, et on pose $\mu = \mathbb{E}[T]$. Soit T^* une variable aléatoire suivant la loi de T biaisée par la taille : pour toute fonction g positive mesurable :

$$\mathbb{E}[g(T^*)] = \frac{1}{\mu} \mathbb{E}[Tg(T)].$$

1. On suppose que T est de carré intégrable. On note σ^2 sa variance. Calculer $\mathbb{E}[T^*]$ et comparer avec $\mathbb{E}[T]$.
2. On suppose que la loi de T possède une densité f . Montrer que la loi de T^* possède une densité, notée f^* , et l'identifier.
3. Si T est une variable aléatoire de loi exponentielle de paramètre λ , montrer que T^* a même loi que $T + T'$ où T' est une variable aléatoire indépendante de T et de même loi que T . (La loi exponentielle est la seule loi de variable aléatoire strictement positive à posséder cette propriété.) En déduire en particulier que $\mathbb{E}[T^*] = 2\mathbb{E}[T]$.
4. On pose $G(a) = \mathbb{E}[(T - \mu)\mathbf{1}_{\{T \leq a\}}]$ pour $a \geq 0$. Montrer que G est décroissante sur $[0, \mu]$ et croissante sur $[\mu, \infty[$. En déduire que pour tout $a \geq 0$, $G(a) \leq 0$ puis que $\mathbb{P}(T^* \leq a) \leq \mathbb{P}(T \leq a)$.
5. Soit Z une variable aléatoire réelle de fonction de répartition F_Z . On note F_Z^{-1} l'inverse généralisé de F_Z : pour $p \in]0, 1[$, $F_Z^{-1}(p) = \inf\{x \in \mathbb{R}, F(x) \geq p\}$, où par convention $\inf \emptyset = +\infty$. (Si F_Z est continue strictement croissante, alors F_Z^{-1} correspond à l'inverse de F_Z .) On admet que

$$F_Z(x) \geq p \iff x \geq F_Z^{-1}(p).$$

Soit U une variable aléatoire uniforme sur $[0, 1]$. Montrer en utilisant les fonctions de répartition que $F_Z^{-1}(U)$ a même loi que Z .

6. Déduire des deux questions précédentes, qu'il existe des variables aléatoires \tilde{T} , de même loi que T , et \tilde{T}^* , de même loi que T^* , telles que p.s. $\tilde{T}^* \geq \tilde{T}$. (On dit que T^* domine stochastiquement T .)

Conclusion

On déduit des deux parties que si les temps écoulés entre les passages de deux bus consécutifs peuvent être modélisés par des variables aléatoires indépendantes de même loi et de moyenne μ , alors quand le client arrive au hasard à l'arrêt de bus, le temps moyen écoulé entre le passage du bus précédent et le passage du prochain bus est plus grand que μ (et égal à 2μ dans le cas de la loi exponentielle).

△

XI.4 La statistique de Mann et Whitney

Exercice XI.7 (Statistique de Mann et Whitney).

L'objectif est d'étudier le comportement asymptotique des statistiques de Mann et Whitney^{7 8}.

I Calculs préliminaires

Soit X, Y deux variables aléatoires réelles indépendantes de densité respective f et g . On introduit les fonctions de répartition :

$$F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^x f(u) du \quad \text{et} \quad G(y) = \mathbb{P}(Y \leq y) = \int_{-\infty}^y g(u) du.$$

On suppose que $p = \mathbb{P}(Y \leq X) \in]0, 1[$.

1. Quelle est la loi de $\mathbf{1}_{\{Y \leq X\}}$? Donner $\text{Var}(\mathbf{1}_{\{Y \leq X\}})$ 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 $S = \mathbb{E}[\mathbf{1}_{\{Y \leq X\}}|X]$ et $T = \mathbb{E}[\mathbf{1}_{\{Y \leq X\}}|Y]$. Déterminer S et T . Donner $\mathbb{E}[S]$ et $\mathbb{E}[T]$.
4. On pose $\alpha = \text{Var}(S)$ et $\beta = \text{Var}(T)$. Calculer α (respectivement β) en fonction de p, G et f (respectivement p, F et g).
5. Montrer que, si X et Y ont même loi, alors $\alpha = \beta$. Donner alors leur valeur.
6. Calculer $\text{Cov}(S, \mathbf{1}_{\{Y \leq X\}})$ et $\text{Cov}(T, \mathbf{1}_{\{Y \leq X\}})$.

On admet que $p \in]0, 1[$ implique que $(\alpha, \beta) \neq (0, 0)$.

II Étude de la projection de Hájek 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 est la variable définie pour $m \geq 1, n \geq 1$, par :

$$U_{m,n} = \sum_{i=1}^m \sum_{j=1}^n \mathbf{1}_{\{Y_j \leq X_i\}}.$$

7. H. B. Mann and D.R. Whitney. On a Test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.*, vol. 18(1), pp. 50–60 (1947).

8. P. Capéraà et B. van Custem. *Méthodes et modèles en statistique non paramétrique*. Dunod (1988).

On pose $U_{m,n}^* = U_{m,n} - \mathbb{E}[U_{m,n}] = \sum_{i=1}^m \sum_{j=1}^n \left(\mathbf{1}_{\{Y_j \leq X_i\}} - p \right)$. La projection de Hájek⁹ 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 $\text{Var}(H_{m,n})$ en fonction de α et β .
3. Déterminer la limite en loi des suites :

$$\left(V_m = \frac{1}{\sqrt{m}} \sum_{i=1}^m (S_i - p), m \geq 1 \right) \quad \text{et} \quad \left(W_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n (T_j - p), n \geq 1 \right).$$

Il est facile de vérifier, à partir de la démonstration du théorème central limite, la convergence uniforme locale des fonctions caractéristiques. Soit $(Z_k, k \geq 1)$ une suite de variables aléatoires indépendantes, de même loi et de carré intégrable, telles que $\mathbb{E}[Z_k] = 0$ et $\text{Var}(Z_k) = \sigma^2$. Alors on a :

$$\psi_{\frac{1}{\sqrt{k}} \sum_{i=1}^k Z_i}(u) = e^{-\sigma^2 u^2 / 2} + R_k(u),$$

et pour tout $K \geq 0$, $\lim_{k \rightarrow \infty} \sup_{|u| \leq K} |R_k(u)| = 0$.

3. Écrire $H_{m,n} / \sqrt{\text{Var}(H_{m,n})}$ comme une combinaison linéaire de V_m et W_n . En utilisant la propriété ci-dessus, montrer que, quand $\min(m, n)$ tend vers l'infini, la suite $\left(H_{m,n} / \sqrt{\text{Var}(H_{m,n})}, m \geq 1, n \geq 1 \right)$ converge en loi vers la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.
4. On admet la formule suivante (voir la partie IV pour une démonstration) :

$$\text{Var}(U_{m,n}) = mn^2\alpha + m^2n\beta + mn(p - p^2 - \alpha - \beta). \quad (\text{XI.5})$$

Déterminer la limite en loi de la suite $\left(H_{m,n} / \sqrt{\text{Var}(U_{m,n})}, m \geq 1, n \geq 1 \right)$ quand $\min(m, n)$ tend vers l'infini.

9. J. Hájek. Asymptotic normality of simple linear rank statistics under alternative. *Ann. Math. Stat.*, vol. 39, pp. 325-346 (1968).

III Convergence de la statistique de Mann et Whitney

1. Montrer que $\text{Cov}(H_{m,n}, U_{m,n}^*) = mn^2 \text{Cov}(S, \mathbf{1}_{\{Y \leq X\}}) + nm^2 \text{Cov}(T, \mathbf{1}_{\{Y \leq X\}})$.
2. En déduire $\text{Var}(H_{m,n} - U_{m,n}^*)$.
3. Calculer la limite de $\text{Var}(H_{m,n} - U_{m,n}^*) / \text{Var}(U_{m,n})$ quand $\min(m, n)$ tend vers l'infini.
4. En déduire que la suite $\left(\frac{H_{m,n} - U_{m,n}^*}{\sqrt{\text{Var}(U_{m,n})}}, m \geq 1, n \geq 1 \right)$ converge en probabilité vers 0 quand $\min(m, n)$ tend vers l'infini.
5. Montrer que la suite :

$$\left(\frac{U_{m,n} - mnp}{\sqrt{\text{Var}(U_{m,n})}}, m \geq 1, n \geq 1 \right)$$

converge en loi quand $\min(m, n)$ tend vers l'infini. Déterminer la loi limite.

IV Calcul de la variance de la statistique de Mann et Whitney

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 $\text{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y' \leq X\}}) = \alpha$ et $\text{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y \leq X'\}}) = \beta$.
3. Calculer la variance de $U_{m,n}$ et vérifier ainsi la formule (XI.5).
4. Donner $\text{Var}(U_{m,n})$ dans le cas où les variables aléatoires X et Y ont même loi.

△

XI.5 Le processus de Galton Watson

Exercice XI.8 (Modèle de population).

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 détaillée¹⁰.

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 masculins 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 définit la probabilité d'extinction de la population :

$$\eta = \mathbb{P}(\text{il existe } n \geq 0 \text{ tel que } Z_n = 0).$$

Pour $k \in \mathbb{N}$, on note $p_k = \mathbb{P}(\xi = k)$, et l'on suppose que $\boxed{p_0 > 0}$ (si $p_0 = 0$, alors la probabilité d'extinction est nulle).

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.* p.s. la population s'éteint.

On note ϕ la fonction génératrice de ξ , $\phi_0(z) = z$ pour $z \in [0, 1]$ et, pour $n \in \mathbb{N}^*$, ϕ_n la fonction génératrice de Z_n .

10. D. Kendall. Branching processes since 1873. *J. London Math. Soc.*, vol. 41, pp. 385-406 (1966).

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 :

$$\phi(x) = x. \quad (\text{XI.6})$$

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.6) 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 10 en bas de page 82) sont telles que l'on peut modéliser la loi de ξ sous la forme :

$$p_0 = \alpha, \quad \text{et pour } k \geq 1, \quad p_k = (1 - \alpha)(1 - \beta)^{k-1}, \quad (\text{XI.7})$$

avec $0 < \alpha < \beta < 1$. On suppose dorénavant que la loi de ξ est donnée par (XI.7).

1. Calculer $m = \mathbb{E}[\xi]$, vérifier que $m > 1$ et calculer η , l'unique solution de (XI.6) dans $]0, 1[$, où ϕ est la fonction génératrice de ξ . Application numérique (cf la note 10 en bas de page 82) : $\alpha = 0.4813$ et $\beta = 0.5586$.
2. Vérifier que $\frac{\phi(z) - 1}{\phi(z) - \eta} = m \frac{z - 1}{z - \eta}$. En déduire $\phi_n(z)$.
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 Loi de Bose-Einstein

Exercice XI.9 (Loi de Bose-Einstein).

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.) ont un spin entier et obéissent à la statistique de Bose-Einstein : plusieurs particules peuvent occuper le même état, et les particules sont indiscernables.

Le but de cet exercice est, après avoir établi la loi de Bose-Einstein^{11 12}, 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. 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 $p(k) = \lim_{n \rightarrow \infty} p_n(k)$, 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

11. W. Feller. *An introduction to probability and its applications*, vol. 1. Wiley, 3rd ed. (1968).
 12. Y. Ijiri and H. A. Simon. Some distributions associated with Bose-Einstein statistics. *Proc. Nat. Acad. Sci.*, vol. 72(5), pp. 1654–1657 (1975).

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 $\frac{(n+r-1)!}{r!(n-1)!}$ 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 $X_{n,r} = (X_{n,r}^{(1)}, \dots, X_{n,r}^{(n)})$, 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 $r \rightarrow \infty$, $n \rightarrow \infty$ et $r/n \rightarrow \theta \in]0, \infty[$. 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 $i \in \{1, \dots, n\}$. Calculer $\sum_{i=1}^n X_{n,r}^{(i)}$. En déduire $\mathbb{E}[X_{n,r}^{(1)}]$.
5. Vérifier que si $r \rightarrow \infty$, $n \rightarrow \infty$ et $r/n \rightarrow \theta \in]0, \infty[$, alors $\mathbb{E}[X_{n,r}^{(1)}]$ converge vers $\mathbb{E}[X]$, où la variable X est définie à la question II.3.
6. On suppose $r \geq 1$. Donner une relation 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. En s'inspirant de la question précédente calculer également $\mathbb{E}[(X_{n,r}^{(1)})^2]$ pour $r \geq 2$. Vérifier que si $r \rightarrow \infty$, $n \rightarrow \infty$ et $r/n \rightarrow \theta \in]0, \infty[$, 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 système.

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 Sondages (II)

Exercice XI.10 (Sondages).

On considère un sondage^{13 14} pour le deuxième tour de l'élection présidentielle française effectué sur n personnes parmi la population des N électeurs, dont $N_A \geq 2$ (resp. $N_B \geq 2$) votent pour le candidat A (resp. B), avec $N = N_A + N_B$. Le but du sondage est d'estimer la proportion, $p = N_A/N$, de personnes qui votent pour A . On se propose de comparer¹⁵ les méthodes de sondage avec remise et de sondage sans remise.

I Sondage avec remise

Dans le sondage **avec remise**, on suppose que la k -ième personne interrogée est choisie au hasard parmi les N électeurs, indépendamment des personnes précédemment choisies. (En particulier, une personne peut être interrogée plusieurs fois.) La réponse de la k -ième personne interrogée est modélisée par une variable aléatoire Y_k : $Y_k = 1$ (resp. $Y_k = 0$) si la personne interrogée vote pour le candidat A (resp. B). On estime p à l'aide de la moyenne empirique, $\bar{Y}_n = \frac{1}{n} \sum_{k=1}^n Y_k$.

1. Donner la loi des variables aléatoires Y_1, \dots, Y_n . En déduire la loi de $n\bar{Y}_n$. Calculer $\mathbb{E}[\bar{Y}_n]$ et $\text{Var}(\bar{Y}_n)$. Montrer que $(\bar{Y}_n, n \geq 1)$ converge p.s. vers p . Montrer, à l'aide du théorème de Slutsky, que $(\sqrt{n}(\bar{Y}_n - p)/\sqrt{\bar{Y}_n(1 - \bar{Y}_n)}, n \geq 1)$ converge en loi vers une limite que l'on précisera.
2. Donner un intervalle de confiance sur p de niveau asymptotique $1 - \alpha$. Pour $n = 1000$ et $\alpha = 5\%$, quelle est la précision (i.e. la demi largeur maximale) de l'intervalle de confiance ?

II Sondage sans remise

Dans le sondage **sans remise**, on suppose que la k -ième personne interrogée est choisie au hasard parmi les $N - k + 1$ personnes qui n'ont pas encore été interrogées. La réponse de la k -ième personne interrogée est modélisée par une

13. Y. Tillé. *La théorie des sondages*. Dunod (2001).

14. W. Cochran. *Sampling techniques*. Wiley, thrid ed. (1977).

15. S.N. Lahiri, A. Chatterjee and T. Maiti. Normal approximation to the hypergeometric distribution in nonstandard cases and a sub-Gaussian Berry-Essèen theorem. *J. Stat. Planning and Inference*, vol. 137, pp. 3570–3590 (2007).

variable aléatoire $X_k : X_k = 1$ (resp. $X_k = 0$) si la personne interrogée vote pour le candidat A (resp. B). On suppose que $1 \leq n \leq N$, et on estime p , à l'aide de la moyenne empirique, $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$.

II.1 Loi faible des grands nombres

1. Montrer que :

$$\mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n)) = \frac{N_A! N_B! (N - n)!}{(N_A - s)! (N_B - (n - s))! N!},$$

où $s = \sum_{k=1}^n x_k$ et $(x_1, \dots, x_n) \in \{0, 1\}^n$. On précisera les valeurs possibles pour s .

- Vérifier que la loi de $(X_{\tau(1)}, \dots, X_{\tau(n)})$ ne dépend pas de la permutation τ de $\{1, \dots, n\}$.
- Déduire de la question II.1.2 que $\mathbb{E}[X_k] = p$ pour tout $1 \leq k \leq N$, puis $\mathbb{E}[\bar{X}_n]$.
- Calculer $\mathbb{E}[X_1^2]$ et $\mathbb{E}[X_1 X_2]$. Montrer, avec la question II.1.2, que $\text{Var}(\bar{X}_n) = \frac{1}{n} p(1-p) \left(1 - \frac{n-1}{N-1}\right)$. Que se passe-t-il pour $n = N$?
- Montrer, en utilisant l'inégalité de Tchebychev, que $\bar{X}_n - p$ converge en probabilité vers 0 quand N et n tendent vers l'infini.

II.2 Théorème central limite

Soit $(V_k, k \geq 1)$ des variables aléatoires indépendantes de loi de Bernoulli de paramètre θ où $\theta \in]0, 1[$ est fixé. On rappelle que $p = \frac{N_A}{N}$.

- Montrer que, quand N tend vers l'infini avec $\lim_{N \rightarrow \infty} \frac{N_A}{N} = \theta$ et n reste fixe, pour tout $x \in \{0, 1\}^n$, $\mathbb{P}((X_1, \dots, X_n) = x)$ converge vers $\mathbb{P}((V_1, \dots, V_n) = x)$ et en déduire que (X_1, \dots, X_n) converge en loi vers (V_1, \dots, V_n) .

On admet qu'il existe $C > 0$ tels que pour tous $1 \leq n^2 \leq \min(N_A, N_B)$ et $(x_1, \dots, x_n) \in \{0, 1\}^n$, on a, avec $s = \sum_{k=1}^n x_k$,

$$\left| \mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n)) - p^s (1-p)^{n-s} \right| \leq C \frac{n^2}{\min(N_A, N_B)} p^s (1-p)^{n-s}. \quad (\text{XI.8})$$

Soit $(g_n, n \geq 1)$ une suite de fonctions mesurables bornées par M , avec g_n définie sur \mathbb{R}^n .

2. Soit (Z_1, \dots, Z_n) des variables aléatoires indépendantes de loi de Bernoulli de paramètre p . Montrer que, pour tout $1 \leq n^2 \leq \min(N_A, N_B)$:

$$|\mathbb{E}[g_n(X_1, \dots, X_n)] - \mathbb{E}[g_n(Z_1, \dots, Z_n)]| \leq MC \frac{n^2}{\min(N_A, N_B)}.$$

3. En utilisant la majoration :

$$\begin{aligned} & |p^s(1-p)^{n-s} - \theta^s(1-\theta)^{n-s}| \\ & \leq \int_{[p, \theta]} sx^s(1-x)^{n-s} \frac{dx}{x} + \int_{[1-p, 1-\theta]} (n-s)x^{n-s}(1-x)^s \frac{dx}{x}, \end{aligned}$$

avec la convention $\int_{[a, b]} f(x) dx = \int_{\min(a, b)}^{\max(a, b)} f(x) dx$, montrer que :

$$|\mathbb{E}[g_n(Z_1, \dots, Z_n)] - \mathbb{E}[g_n(V_1, \dots, V_n)]| \leq 2Mn|p - \theta|.$$

4. En déduire que quand N et n tendent vers l'infini, n^2/N et $n(p - \theta)$ tendent vers 0, alors $|\mathbb{E}[g_n(X_1, \dots, X_n)] - \mathbb{E}[g_n(V_1, \dots, V_n)]|$ tend vers 0. (On rappelle que $0 < \theta < 1$.)
5. En déduire que quand N et n tendent vers l'infini, et n^2/N et $n(p - \theta)$ tendent vers 0, alors $\sqrt{n}(\bar{X}_n - \theta) / \sqrt{\bar{X}_n(1 - \bar{X}_n)}$ converge en loi vers une variable aléatoire gaussienne centrée. (On rappelle que $0 < \theta < 1$.)
6. Déduire de la question précédente un intervalle de confiance sur p de niveau asymptotique $1 - \alpha$. Quelle est la différence entre un sondage avec remise et un sondage sans remise quand N est grand et n^2/N petit ? Que pensez-vous de la précision d'un sondage sans remise pour le deuxième tour de l'élection présidentielle française en 2007 portant sur $n = 1\,000$ personnes (avec $N = 44\,472\,834$ inscrits).

△

XI.8 Loi de Yule (I)

Exercice XI.11 (Loi de Yule).

La loi de Yule¹⁶ permet de modéliser de nombreux phénomènes¹⁷ en économie, sociologie, en biologie ou encore en physique. Par exemple elle permet de représenter

16. U. Yule. A mathematical theory of evolution based on the conclusion of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society (B)*, vol. 213, pp. 21-87 (1924).

17. H. Simon. On a class of skew distribution functions. *Biometrika*, vol. 42, pp. 425-440 (1955).

la loi du nombre de consultations d'une page web, du nombre de personnes vivant dans une ville, du nombre de publications d'un scientifique, du nombre d'occurrences d'un mot dans un texte, du salaire (comprendre nombre d'euros gagnés) d'un individu, du nombre de disques vendus, ...

Pour illustrer la problématique, on considère le cas du nombre de disques vendus. On fait l'hypothèse suivante : le n -ème disque vendu est,

- avec probabilité $\alpha \in]0, 1[$, celui d'un nouvel album (pour lequel aucun disque n'avait encore été vendu).
- avec probabilité proportionnelle à k , celui d'un album pour lequel $k (\geq 1)$ disques ont déjà été vendus ;

Alors, le nombre de disques vendus pour un album, pris au hasard parmi les albums ayant au moins un disque vendu, suit asymptotiquement quand le nombre de disques vendus est grand la loi de Yule. La première partie de l'exercice permet de démontrer ce résultat. La deuxième partie de l'exercice étudie la loi de Yule. Les deux parties sont indépendantes.

I Étude asymptotique

On considère le modèle suivant. On dispose d'une infinité de boîtes vides. À l'instant $n = 1$, on met une boule dans une boîte. À l'instant $n > 1$, on met une nouvelle boule dans une boîte vide avec probabilité $\alpha \in]0, 1[$. Sinon, on met la nouvelle boule dans une boîte non vide avec une probabilité proportionnelle au nombre de boules dans cette boîte. À l'instant n on dispose de n boules réparties dans N_n boîtes non vides. On note $Z_n = 1$ si on met la boule dans une boîte vide à l'instant $n \geq 1$ et $Z_n = 0$ sinon. On a donc $N_n = \sum_{k=1}^n Z_k$. Remarquer que $Z_1 = 1$.

(Dans l'exemple précédent, une boîte correspond à un album et le nombre de boules dans la boîte au nombre de disques vendus de cet album. Le nombre total de disques vendus est n , et le nombre d'albums (différents) vendus est N_n , $Z_n = 1$ si la n -ième vente est la première vente d'un album.)

1. Donner la loi de $(Z_n, n \geq 2)$.
2. En déduire que la suite $(N_n/n, n \geq 1)$ converge en un sens à préciser vers une limite que l'on déterminera.

Pour $k \in \mathbb{N}^*$, on note $F_n^{(k)}$ le nombre de boîtes contenant exactement k boules à l'instant n .

3. Calculer $\sum_{k=1}^n k F_n^{(k)}$, $\sum_{k=1}^n F_n^{(k)}$ et $\sum_{k=1}^n \mathbb{E}[F_n^{(k)}]$.

Pour $n \in \mathbb{N}^*$, on note Y_n le nombre de boules à l'instant $n - 1$ de la boîte à laquelle on rajoute la n -ième boule. En particulier, on a $Z_n = 1$ si et seulement si $Y_n = 0$.

4. Pour $k \in \mathbb{N}^*$, exprimer $F_{n+1}^{(k)}$ à l'aide de $F_n^{(k)}$, $\mathbf{1}_{\{Y_{n+1}=k-1\}}$ et de $\mathbf{1}_{\{Y_{n+1}=k\}}$.

Pour $k \in \mathbb{N}^*$, on pose $p_n(k) = \frac{\mathbb{E}[F_n^{(k)}]}{\alpha n}$.

5. Montrer que, pour $k \geq 2$, $\mathbb{P}(Y_{n+1} = k | Z_{n+1} = 0) = \alpha k p_n(k)$.

6. En déduire que $\alpha(n+1)p_{n+1}(1) = \alpha n p_n(1) + \alpha - (1-\alpha)\alpha p_n(1)$. Et donner une relation entre $p_{n+1}(k)$, $p_n(k)$ et $p_n(k-1)$ pour $k \geq 2$.

On admet que pour tout $k \in \mathbb{N}^*$, il existe $p(k) \in [0, \infty[$ tel que $\lim_{n \rightarrow \infty} p_n(k) = p(k)$ où $p(k)$ est une solution stationnaire des équations de la question précédente. En fait, on peut facilement montrer que pour tout $k \geq 1$, $(F_n^{(k)})/\alpha n, n \geq 1$ converge en probabilité (et donc en moyenne car $F_n^{(k)}/n$ est borné) vers $p(k)$. On pose $\rho = 1/(1-\alpha)$.

7. Déduire de ce qui précède que :

$$p(1) = \rho/(1+\rho), \quad \text{et pour tout } k \geq 2, \quad p(k) = \frac{k-1}{k+\rho} p(k-1). \quad (\text{XI.9})$$

II Loi de Yule

On rappelle la définition et quelques propriétés de la fonction Γ : pour tout $a > 0, b > 0$:

$$\Gamma(a) = \int_{\mathbb{R}_+} x^{a-1} e^{-x} dx, \quad \Gamma(1) = 1, \quad \Gamma(a+1) = a\Gamma(a)$$

$$\lim_{a \rightarrow \infty} \frac{\Gamma(a)}{a^{a-\frac{1}{2}} e^{-a} \sqrt{2\pi}} = 1, \quad B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_{]0,1[} x^{a-1}(1-x)^{b-1} dx.$$

Soit $(p(k), k \in \mathbb{N}^*)$ défini par (XI.9), avec $\rho \in]0, \infty[$.

1. Vérifier que pour $k \in \mathbb{N}^*$, $p(k) = \rho B(k, \rho + 1)$. Montrer que $(p(k), k \in \mathbb{N}^*)$ définit la loi d'une variable aléatoire discrète à valeurs dans \mathbb{N}^* . Cette loi est appelée loi de Yule.
2. Montrer que $\sum_{i=k}^{\infty} p(i) = \rho B(k, \rho)$. Donner un équivalent de $p(k)$ et de $\sum_{i=k}^{\infty} p(i)$ quand k tend vers l'infini. On dit que la loi de Yule est une loi en puissance sur \mathbb{N}^* .
3. Calculer la moyenne et la variance de la loi de Yule. (On pourra utiliser le calcul de la moyenne et de la variance de la loi géométrique pour obtenir le résultat.)

△

XI.9 Mathématiques financières

Exercice XI.12 (Mathématiques pour la finance).

On présente une application du modèle de Cox-Ross-Rubinstein¹⁸ qui est une version discrète du modèle de Black-Sholes¹⁹, modèle fondateur en mathématiques pour la finance. On considère un marché avec un actif sans risque (par exemple un placement sur un livret) et un actif risqué (une action, une obligation, ...) à des instants discrets $n \in \mathbb{N}$. On note :

$$S_n^0 = (1 + r)^n S_0^0$$

le prix de l'actif sans risque à l'instant n , où r représente le taux d'intérêt fixe. On note S_n le prix de l'actif risqué à l'instant n , et on suppose que l'évolution de l'actif risqué est donné par :

$$S_{n+1} = X_{n+1} S_n,$$

où les variables aléatoires $(X_n, n \in \mathbb{N}^*)$ sont à valeurs dans $\{1 + d, 1 + m\}$ avec $-1 < d < r < m$.

Soit h une fonction positive. L'option h de maturité N promet le gain $h(S_N)$ à l'instant N . On peut distinguer deux notions de prix pour cette option :

- Le prix du gain moyen espéré : $\mathbb{E}[h(S_N)]$. Ce prix correspond à un point de vue spéculatif. Il dépend de la loi, dite historique, de (X_1, \dots, X_N) *a priori* non connue. Il ne sera pas considéré ici.
- Le prix de couverture qui correspond au coût d'une stratégie autofinancée de couverture, que l'on détaille dans ce qui suit.

Une stratégie est définie par $\phi = ((\phi_n^0, \phi_n), n \in \{0, \dots, N\})$, où ϕ_n^0 représente la quantité d'actif sans risque détenu à l'instant n et ϕ_n la quantité d'actif risqué détenu à l'instant n . La valeur de la stratégie ϕ à l'instant n est :

$$V_n = \phi_n^0 S_n^0 + \phi_n S_n. \quad (\text{XI.10})$$

La stratégie ϕ est une stratégie de couverture pour l'option h de maturité N si elle donne la même valeur que l'option à la maturité :

$$V_N = h(S_N). \quad (\text{XI.11})$$

La stratégie ϕ est dite autofinancée si les variations de la valeur de la stratégie sont dues aux variations des actifs : on change à l'instant n la répartition (sans coût de

18. J. C. Cox, S. A. Ross and M. Rubinstein. Option pricing : a simplified approach. *Journal of Financial Economics*, vol. 7, pp. 229-263 (1979).

19. F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, vol. 81(3), pp. 637-654 (1973).

transaction) entre actif sans risque et actif risqué en fonction des cours observés jusque là, mais on ne rajoute ni ne prélève d'actifs : pour tout $n \in \{0, \dots, N-1\}$:

$$V_{n+1} = \phi_n^0 S_{n+1}^0 + \phi_n S_{n+1}. \quad (\text{XI.12})$$

Le prix de couverture est V_0 : le coût à l'instant initial qui permet la mise en œuvre de la stratégie autofinancée de couverture. Quelle que soit l'évolution de l'actif risqué, la stratégie autofinancée de couverture donne $V_N = h(S_N)$ à la maturité N .

Le but de l'exercice est de déterminer la stratégie autofinancée de couverture et le prix de couverture d'une option h de maturité N .

On introduit la probabilité risque neutre \mathbb{P}^* (*a priori* différente de la probabilité réelle \mathbb{P} , dite probabilité historique), et \mathbb{E}^* l'espérance correspondante, sous laquelle les variables aléatoires $(X_n, n \in \mathbb{N}^*)$ sont indépendantes et de même loi définie par :

$$\mathbb{P}^*(X_1 = 1 + m) = (r - d)/(m - d) \quad \text{et} \quad \mathbb{P}^*(X_1 = 1 + d) = (m - r)/(m - d).$$

Soit une option h de maturité N . On pose $v(N, s) = h(s)$ et pour $n \in \{0, \dots, N-1\}$:

$$v(n, s) = (1 + r)^{-N+n} \mathbb{E}^* \left[h(s \prod_{k=n+1}^N X_k) \right]$$

et :

$$\varphi(n, s) = \frac{1}{s} \frac{v(n, (1 + m)s) - v(n, (1 + d)s)}{m - d}.$$

Soit ϕ une stratégie autofinancée de couverture pour l'option h de maturité N .

I Le cas à une période : de $N - 1$ à N

1. Dédurre de (XI.11) et (XI.12) les équations suivantes :

$$\begin{aligned} \phi_{N-1}^0 (1 + r) S_{N-1}^0 + \phi_{N-1} (1 + m) S_{N-1} &= h((1 + m) S_{N-1}), \\ \phi_{N-1}^0 (1 + r) S_{N-1}^0 + \phi_{N-1} (1 + d) S_{N-1} &= h((1 + d) S_{N-1}). \end{aligned}$$

2. Vérifier que $\phi_{N-1} = \varphi(N, S_{N-1})$.

3. Montrer que $V_{N-1} = v(N - 1, S_{N-1})$.

II Le cas général

1. Montrer que pour $n \in \{1, \dots, N\}$, on a : $v(n - 1, s) = (1 + r)^{-1} \mathbb{E}^* [v(n, sX_n)]$.

2. Montrer que, pour $n \in \{0, \dots, N - 1\}$, on a :

$$V_n = v(n, S_n) \quad \text{et} \quad \phi_n = \varphi(n + 1, S_n). \quad (\text{XI.13})$$

3. En déduire qu'il existe une unique stratégie autofinancée de couverture pour l'option h de maturité N . Montrer que le coût initial de cette stratégie, *i.e.* le prix de couverture de l'option, est :

$$V_0 = (1 + r)^{-N} \mathbb{E}^*[h(S_N)]. \quad (\text{XI.14})$$

4. Le prix de couverture de l'option donné par la stratégie autofinancée de couverture semble sans risque. Qu'en pensez-vous ? (Répondre en moins de 5 lignes.)

III Call et Put

1. Calculer $\mathbb{E}^*[\prod_{k=n+1}^N X_k]$ pour $0 \leq n \leq N - 1$.
2. Déduire de (XI.14) le prix de couverture de l'option qui promet S_N à la maturité N . Donner, à l'aide de (XI.13), la stratégie autofinancée de couverture correspondante.

On note $x_+ = \max(x, 0)$ la partie positive de x . Soit $K > 0$. On considère deux options très répandues sur les marchés financiers : le Call de strike K et de maturité N qui promet le gain $(S_N - K)_+$ à l'instant N , et le Put de strike K et de maturité N qui promet le gain $(K - S_N)_+$ à l'instant N . On note $C(s, K, N)$ le prix de couverture de ce Call et $P(s, K, N)$ le prix de couverture de ce Put quand la valeur initiale de l'actif risqué est $S_0 = s$.

3. Montrer, à l'aide de (XI.14), la relation de parité²⁰ Call-Put :

$$C(S_0, K, N) - P(S_0, K, N) = S_0 - (1 + r)^{-N} K.$$

4. Montrer la formule du prix du Call :

$$\begin{aligned} & C(s, K, N) \\ &= (1 + r)^{-N} \sum_{k=0}^N \binom{N}{k} \frac{(r - d)^k (m - r)^{N-k}}{(m - d)^N} \left((1 + m)^k (1 + d)^{N-k} s - K \right)_+. \end{aligned}$$

△

20. Pour des raisons d'absence d'arbitrage sur les marchés, cette relation est toujours vérifiée. Les modèles mathématiques sur l'évolution de l'actif risqué doivent donc permettre de la retrouver.

XI.10 Transmission de message

Exercice XI.13 (Transmission de message).

On souhaite envoyer un message formé de 0 ou de 1 en utilisant un canal bruité. Si on envoie un message $z = (z_1, \dots, z_n)$ de longueur n par le canal, on reçoit le message y noté $z+E$, où $E = (E_1, \dots, E_n)$ représente les erreurs et $y = (y_1, \dots, y_m)$ avec, pour $1 \leq i \leq n$:

$$y_i = z_i + E_i \pmod 2 = \begin{cases} z_i & \text{si } E_i = 0, \\ 1 - z_i & \text{si } E_i = 1. \end{cases} \quad (\text{XI.15})$$

Si on a $E_i = 0$ alors le signal z_i est correctement transmis.

On suppose que les erreurs dues au canal ($E_i, i \in \mathbb{N}^*$) sont des variables aléatoires indépendantes, indépendantes du message envoyé et de même loi de Bernoulli de paramètre $p \in]0, 1/2[$. Le paramètre p s'interprète comme la probabilité d'erreur du canal.

Afin d'améliorer la qualité de la transmission, on utilise un schéma de codage : on code le message initial de longueur m en un message de longueur $n \geq m$; ce message de longueur n est envoyé par le canal puis décodé en un message de longueur m .

Plus précisément un schéma de codage (ou code) est défini par une fonction de codage $f : \{0, 1\}^m \rightarrow \{0, 1\}^n$, et une fonction de décodage $g : \{0, 1\}^n \rightarrow \{0, 1\}^m$. Un message $x \in \{0, 1\}^m$ de longueur m est codé par $z = f(x)$, puis transmis par le canal. On reçoit le message bruité $f(x) + E = z + E$ (défini par (XI.15)), où $E = (E_1, \dots, E_n)$ représente les erreurs dues au canal. Le message reçu est décodé à l'aide de la fonction g : le message reçu est alors $x' = g(y(x, E))$. Le **taux de transmission** du schéma de codage est défini par m/n et sa **probabilité d'erreur** par :

$$p_{f,g} = \mathbb{P}\left(g(f(X) + E) \neq X\right),$$

où X est de la loi uniforme sur $\{0, 1\}^m$ et est indépendant de E . La probabilité d'erreur du schéma de codage représente la probabilité d'obtenir un message ne correspondant pas au message initial, pour un message initial pris au hasard.

L'objectif du problème est de montrer que l'on peut trouver asymptotiquement (pour m grand) des schémas de codage dont les probabilités d'erreur sont arbitrairement proches de 0 et dont le taux de transmission peut être arbitrairement proche d'une constante c_p définie par (XI.19) non nulle. De plus on peut montrer que cette constante est optimale. Ces résultats sont dus à Shannon²¹ ; il s'agit du théorème fondamental de la théorie de l'information.

21. C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, vol. 27, pp. 379-423 and 623-656, 1948.

I Code à répétition

On suppose $m = 1$. On choisit le code à répétition suivant : $n = 3$, $f(x) = (x, x, x)$ pour $x \in \{0, 1\}$ et la règle de la majorité pour le décodage :

$$g(y_1, y_2, y_3) = \begin{cases} 0 & \text{si } y_1 + y_2 + y_3 \leq 1, \\ 1 & \text{si } y_1 + y_2 + y_3 \geq 2. \end{cases}$$

1. Calculer la probabilité d'erreur du code à répétition.
2. Donner un code dont la probabilité d'erreur soit inférieure à $\varepsilon > 0$ fixé. Que devient le taux de transmission quand ε tend vers 0 ?

II Probabilités d'évènements rares

Soit $(V_n, n \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes et de même loi de Bernoulli de paramètre $\rho \in]0, 1[$. On considère la moyenne empirique $\bar{V}_n = \frac{1}{n} \sum_{i=1}^n V_i$. Soit $a \in]\rho, 1[$.

1. Expliquer pourquoi intuitivement l'évènement $\{\bar{V}_n > a\}$ est un évènement rare c'est-à-dire de faible probabilité.
2. Montrer que pour $\lambda > 0$, on a $\mathbb{P}(\bar{V}_n \geq a) \leq \mathbb{E} \left[e^{\lambda V_1} \right]^n e^{-an\lambda}$.

On considère la transformée de Legendre de la log-Laplace de la loi de Bernoulli :

$$A_\rho^*(v) = v \log \left(v/\rho \right) + (1 - v) \log \left((1 - v)/(1 - \rho) \right), \quad v \in]0, 1[. \quad (\text{XI.16})$$

3. Montrer que la fonction A_ρ^* est strictement convexe sur $]0, 1[$, nulle en ρ et atteint son minimum en ρ .
4. Montrer que :

$$\mathbb{P}(\bar{V}_n > a) \leq \mathbb{P}(\bar{V}_n \geq a) \leq e^{-nA_\rho^*(a)}. \quad (\text{XI.17})$$

5. Dédurre de la question précédente que pour $b \in]0, \rho[$:

$$\mathbb{P}(\bar{V}_n \leq b) \leq e^{-nA_\rho^*(b)}. \quad (\text{XI.18})$$

III Codes presque optimaux

Soit $m \leq n$. On définit la distance de Hamming δ sur l'ensemble $\{0, 1\}^n$ par :

$$\delta(y, y') = \sum_{i=1}^n |y_i - y'_i| \quad \text{pour } y = (y_1, \dots, y_n) \text{ et } y' = (y'_1, \dots, y'_n) \in \{0, 1\}^n.$$

Soit $(Z_x, x \in \{0, 1\}^m)$ une famille de variables aléatoires indépendantes de même loi uniforme sur $\{0, 1\}^n$. Soit $p < r < 1/2$.

1. Soit $x \in \{0, 1\}^m$ et $z \in \{0, 1\}^n$. Déterminer la loi de $\delta(Z_x, z)$.
2. Soit $x, x' \in \{0, 1\}^m$ distincts. Montrer que, pour $z \in \{0, 1\}^n$:

$$\mathbb{P}\left(\delta(Z_x, Z_{x'}) \leq nr \mid Z_x = z\right) \leq e^{-n\Lambda_{1/2}^*(r)}.$$

Pour $x \in \{0, 1\}^m$, on pose $h(x) = \mathbb{P}(\text{Il existe } x' \neq x \text{ tel que } \delta(Z_x, Z_{x'}) \leq nr)$.

3. Dédire de la question précédente que la fonction h est bornée par $2^m e^{-n\Lambda_{1/2}^*(r)}$.

On considère le codage (aléatoire) suivant. La fonction de codage f est définie par $f(x) = Z_x$ pour $x \in \{0, 1\}^m$ et la fonction de décodage, pour $y \in \{0, 1\}^n$, par :

$$g(y) = \begin{cases} x & \text{s'il existe un unique } x \in \{0, 1\}^m \text{ tel que } \delta(Z_x, y) \leq nr, \\ 0 & \text{sinon.} \end{cases}$$

Soit un message aléatoire de longueur m , X , de loi uniforme sur $\{0, 1\}^m$ et les erreurs $E = (E_1, \dots, E_n)$, où les variables aléatoires (E_1, \dots, E_n) sont indépendantes de loi Bernoulli de paramètre p , telles que X, E et $(Z_x, x \in \{0, 1\}^m)$ soient indépendantes.

4. Montrer que $\mathbb{P}(g(f(X) + E) \neq X) \leq \mathbb{E}[h(X)] + \mathbb{P}(\delta(0, E) > nr)$.
5. Dédire de (XI.17) que $\mathbb{P}(g(f(X) + E) \neq X) \leq 2^m e^{-n\Lambda_{1/2}^*(r)} + e^{-n\Lambda_p^*(r)}$.

On rappelle (XI.16) et on pose :

$$c_p = \frac{\Lambda_{1/2}^*(p)}{\log 2}. \tag{XI.19}$$

6. Montrer que pour tout $\varepsilon > 0$, il existe $m \leq n$ (grands) et un codage (déterministe) f, g tels que $\mathbb{P}(g(f(X) + E) \neq X) < \varepsilon$ et $c_p - \varepsilon \leq m/n < c_p$. (On choisira r proche de p .)
7. Conclure.

Il est toutefois difficile en pratique d'exhiber des codes optimaux simples à implémenter.

△

XI.11 Mariage d'un prince

Exercice XI.14 (Mariage d'un prince).

Le problème du choix de la princesse pour un prince, appelé problème du mariage

et dans un autre contexte problème du recrutement d'une secrétaire²², date des années 1960 et a donné lieu à de nombreuses analyses. Ce problème s'inscrit dans la théorie plus générale du contrôle optimal²³.

Le prince doit choisir une princesse pour son mariage. Il existe $n \geq 3$ candidates qui lui sont présentées pour la première fois. Les candidates sont présentées au hasard l'une après l'autre. À l'issue de l'entretien avec une princesse, le prince lui donne une réponse ferme. Soit il l'épouse et il arrête les entretiens ; soit il ne l'épouse pas (et il ne revient plus sur cette décision) et il rencontre la princesse suivante. On suppose que l'on peut classer les princesses de la meilleure à la pire sans avoir d'égalité. Et on souhaite trouver une stratégie optimale qui maximise la probabilité pour le prince d'épouser la meilleure princesse. (Si on modifie le critère à optimiser, alors la stratégie optimale est modifiée.)

On considère le modèle suivant. On note Σ_i le rang (de la meilleure à la pire) de la i -ème princesse rencontrée. Le prince souhaite recruter la princesse i telle que $\Sigma_i = 1$. Le vecteur des rangs $\Sigma = (\Sigma_1, \dots, \Sigma_n)$ est une variable aléatoire à valeurs dans l'ensemble des permutations de $\{1, \dots, n\}$, noté \mathcal{S}_n . Toutefois, lors de la rencontre avec la princesse k , son rang est inconnu (sauf si $k = n$). Mais le prince observe R_k son rang partiel parmi les k premières princesses rencontrées. Le vecteur des rangs partiels associé à Σ , noté $R = (R_1, \dots, R_n)$, peut être vu comme une fonction, h , de $\Sigma : R = h(\Sigma)$. On admet que h est une bijection de \mathcal{S}_n dans $E_n = \prod_{k=1}^n \{1, \dots, k\}$. On a en particulier $R_1 = 1$ et $R_n = \Sigma_n$. Par exemple, pour $n = 5$ et la permutation $\sigma = (2, 5, 3, 1, 4)$, les rangs partiels correspondants sont $h(\sigma) = (1, 2, 2, 1, 4)$.

Le but du problème est de montrer que la stratégie optimale (pour n grand) consiste à rencontrer 37% des princesses, puis à choisir la prochaine meilleure. Ceci assure d'épouser la meilleure princesse avec une probabilité d'environ 37%.

I Préliminaires

1. Justifier en quelques mots que la loi de Σ est la loi uniforme sur \mathcal{S}_n .
2. Déterminer la loi de R . Puis, en déduire que R_k est de loi uniforme sur $\{1, \dots, k\}$.
3. Vérifier que les variables aléatoires (R_1, \dots, R_n) sont indépendantes.
4. Vérifier que $\{\Sigma_n = 1\} = \{R_n = 1\}$ et que pour $k \in \{1, \dots, n-1\}$:

$$\{\Sigma_k = 1\} = \{R_k = 1, R_{k+1} > 1, \dots, R_n > 1\}.$$

22. T. Ferguson. Who solved the secretary problem. *Stat. Sciences*, vol. 4(3), pp. 282-296, (1989).

23. D.P. Bertsekas and S.E. Shreve. *Stochastic optimal control*. Academic Press Inc. (1978).

II Les stratégies de seuil

Soit $c \in \{0, \dots, n-1\}$. On considère la stratégie de seuil suivante : le prince rencontre les c premières princesses, puis à partir de la $c+1$ -ème rencontre, il choisit d'épouser la première princesse meilleure que les c premières et s'il n'y en a pas, il épouse la dernière princesse rencontrée. Il épouse donc la princesse :

$$\tau_c = \inf\{k \in \{c+1, \dots, n\}; R_k = 1\},$$

avec la convention $\tau_c = n$ si $\{k \in \{c+1, \dots, n\}; R_k = 1\} = \emptyset$. On cherche le seuil c qui maximise $\mathbb{P}(\Sigma_{\tau_c} = 1)$.

1. Montrer, en décomposant suivant les valeurs de τ_c , que pour $c \in \{1, \dots, n-1\}$:

$$\begin{aligned} \mathbb{P}(\Sigma_{\tau_c} = 1) &= \mathbb{P}(R_{c+1} = 1, \Sigma_{c+1} = 1) \\ &+ \sum_{k=c+2}^n \mathbb{P}(R_{c+1} > 1, \dots, R_{k-1} > 1, R_k = 1, \Sigma_k = 1). \end{aligned}$$

2. En utilisant la question I.4, montrer que $\mathbb{P}(\Sigma_{\tau_c} = 1) = g(c)$, où la fonction g est définie par $g(0) = 1/n$ et pour $c \in \{1, \dots, n-1\}$:

$$g(c) = \frac{c}{n} \sum_{k=c}^{n-1} \frac{1}{k}.$$

3. Étudier le signe de $f(c) = n(g(c) - g(c-1))$ pour $c \in \{1, \dots, n-1\}$ et en déduire que $g(c)$ est maximal pour $c = c_*$ défini par :

$$\sum_{k=c_*}^{n-1} \frac{1}{k} \geq 1 > \sum_{k=c_*+1}^{n-1} \frac{1}{k}.$$

4. En déduire que τ_{c_*} est la stratégie de seuil optimale.
5. On rappelle que $\sum_{k=c}^{n-1} 1/k = \log(n/c) + o(1/c)$. Montrer que :

$$\lim_{n \rightarrow +\infty} \frac{c_*}{n} = \frac{1}{e} \quad \text{et} \quad \lim_{n \rightarrow +\infty} \mathbb{P}(\Sigma_{\tau_{c_*}} = 1) = \frac{1}{e}.$$

III Les stratégies générales

Une stratégie τ correspond à une famille d'ensembles $A_k \subset E_k = \prod_{i=1}^k \{1, \dots, i\}$, pour $k \in \{1, \dots, n-1\}$ et le fait d'épouser la princesse :

$$\tau = \inf\{k \in \{1, \dots, n-1, \}; (R_1, \dots, R_k) \in A_k\},$$

avec la convention $\tau = n$ si $\{k \in \{1, \dots, n-1\}; (R_1, \dots, R_k) \in A_k\} = \emptyset$. (Les stratégies de seuil τ_c sont un cas particulier de stratégies générales avec $A_k = \emptyset$ pour $k \leq c$ et $A_k = E_{k-1} \times \{1\}$ pour $k > c$.) La variable aléatoire²⁴ τ est à valeurs dans $\{1, \dots, n\}$ et on peut facilement vérifier que :

$$\{\tau = k\} = \{(R_1, \dots, R_k) \in B_k\} \quad \text{et} \quad \{\tau > k\} = \{(R_1, \dots, R_k) \in C_k\}$$

pour des ensembles $B_k \subset E_k$ et $C_k \subset E_k$ que l'on ne cherchera pas à écrire.

On définit la probabilité de recruter la meilleure princesse à partir de l'instant k et conditionnellement à $\{R_1 = r_1, \dots, R_k = r_k\}$, pour $(r_1, \dots, r_k) \in E_1 \times \dots \times E_k$, par :

$$G_k(r_1, \dots, r_k) = \mathbb{P}(\Sigma_\tau = 1, \tau \geq k | R_1 = r_1, \dots, R_k = r_k).$$

Ainsi, la probabilité que le prince épouse la meilleure princesse avec la stratégie τ est $G_1(1) = \mathbb{P}(\Sigma_\tau = 1)$.

1. Calculer $G_n(r_1, \dots, r_n)$.
2. En utilisant la question I.4, montrer que :

$$G_k(r_1, \dots, r_k) = k! \sum_{i=k}^n \frac{i}{n} \mathbb{P}(R_i = 1, \tau = i, R_1 = r_1, \dots, R_k = r_k).$$

3. En déduire que pour $k \in \{1, \dots, n-1\}$:

$$G_k(r_1, \dots, r_k) = \frac{k}{n} \mathbf{1}_{\{r_k=1\}} \mathbf{1}_{\{(r_1, \dots, r_k) \in B_k\}} + \mathbb{E}[G_{k+1}(r_1, \dots, r_k, R_{k+1})] \mathbf{1}_{\{(r_1, \dots, r_k) \in C_k\}},$$

et donc :

$$G_k(r_1, \dots, r_k) \leq \max \left(\frac{k}{n} \mathbf{1}_{\{r_k=1\}}, \mathbb{E}[G_{k+1}(r_1, \dots, r_k, R_{k+1})] \right) \mathbf{1}_{\{(r_1, \dots, r_k) \in B_k \cup C_k\}}.$$

On considère la fonction g^* définie sur $\{1, \dots, n+1\}$ par la récursion descendante $g^*(n+1) = 0$ et pour $k \in \{1, \dots, n\}$:

$$g^*(k) = \mathbb{E} \left[\max \left(\frac{k}{n} \mathbf{1}_{\{R_k=1\}}, g^*(k+1) \right) \right].$$

On reprend les notations de la partie II.

4. Montrer que $\mathbb{E}[G_k(r_1, \dots, r_{k-1}, R_k)] \leq g^*(k)$ pour $k \in \{0, \dots, n-1\}$, et donc que $G_1(1) \leq g^*(1)$.

²⁴ On dit que τ est un temps d'arrêt associé aux variables aléatoires (R_1, \dots, R_n) .

5. Montrer que pour $k > c_*$ on a $g^*(k) = g(k - 1)$ et que pour $k \leq c_*$ on a $g^*(k) = g(c_*)$.
6. En déduire que la stratégie τ_{c_*} est la stratégie optimale.

△

XII

Problèmes (probabilités et statistique)

XII.1 Le modèle de Hardy-Weinberg

Exercice XII.1 (Modèle de Hardy-Weinberg pour la distribution d'un génotype). Le but de cet exercice est l'étude simplifiée du modèle de Hardy-Weinberg¹ pour la répartition d'un génotype dans la population humaine².

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 dès 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 $\mathbb{P}(E = aa | M = aa, P = aa) = 1$.

1. G. Hardy. Mendelian proportions in a mixed population. *Science*, vol. 28(706) pp. 49–50 (1908).

2. D. L. Hartl and E. W. Jones. *Genetics : analysis of genes and genomes*. Jones and Bartlett Publishers, 5th ed. (2001).

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 $\mathbb{P}(E = AA) = \left(u_3 + \frac{u_2}{2}\right)^2$.

On pose donc $\theta = u_1 + \frac{u_2}{2}$.

4. Montrer, sans calcul, que la répartition du génotype à la première génération est :

$$\begin{cases} \text{proportion de } aa : & q_1 = \theta^2, \\ \text{proportion de } aA : & q_2 = 2\theta(1 - \theta), \\ \text{proportion de } AA : & q_3 = (1 - \theta)^2. \end{cases} \quad (\text{XII.1})$$

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 :

$$N_j = N_j[n] = \sum_{i=1}^n \mathbf{1}_{\{X_i=j\}}, \quad (\text{XII.2})$$

le nombre de personnes de l'échantillon possédant le génotype j .

1. Donner la loi $\mathbf{1}_{\{X_i=j\}}$. En déduire la loi de N_j .
2. Donner $\mathbb{E}[N_j]$ et $\text{Var}(N_j)$.
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 :

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \frac{n!}{n_1!n_2!n_3!} q_1^{n_1} q_2^{n_2} q_3^{n_3}. \quad (\text{XII.3})$$

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 $\frac{n!}{n_1!n_2!n_3!}$.

6. Calculer la matrice de covariance du vecteur $(\mathbf{1}_{\{X_1=1\}}, \mathbf{1}_{\{X_1=2\}})$. En déduire $\text{Cov}(N_1, N_2)$.
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 tend vers l'infini.

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. L'estimateur de θ , $\hat{\theta}_n$, 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. Calculer $(\hat{q}_1, \hat{q}_2, \hat{q}_3)$ l'estimateur du maximum de vraisemblance de (q_1, q_2, q_3) . On considère la statistique de test :

$$\zeta_n = n \sum_{j=1}^3 \frac{(\hat{q}_j - q_j(\hat{\theta}_n))^2}{\hat{q}_j}.$$

2. Donner un test asymptotique à partir de la statistique de test ζ_n .
3. En déduire un test asymptotique convergent.
4. La mucoviscidose est une maladie génétique. 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 $n \simeq 3.88 \cdot 10^6$, nombre de nouveau-nés atteints de la maladie $n_1 = 1580$, 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 Estimation de la taille d'une population

Exercice XII.2 (Capture et recapture).

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, s_n 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 l'étude de l'estimateur de Bailey³ de N défini par $n_0(n+1)/(s_n+1)$. (On ne considèrera pas l'estimateur de Petersen n_0n/s_n , qui n'est pas défini si $s_n = 0$.)

On pose $p = n_0/N \in]0, 1[$. Comme n_0 est connu, on remarque que donner une estimation de $1/p$ permet de donner une estimation 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, \dots, X_n .

1. Au vu des hypothèses du modèle, quelle est la loi des variables aléatoires X_1, \dots, X_n ?

3. G.A.F. Seber. *The Estimation of Animal Abundance and Related Parameters*. Caldwell (1982).

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 $x_1 \in \{0, 1\}$. En déduire la densité de l'échantillon.
3. Montrer que $S_n = \sum_{i=1}^n X_i$ est une statistique exhaustive. Quelle est sa loi ?
4. Soit h une fonction définie sur $\{0, \dots, n\}$ telle que $h(S_n)$ soit intégrable. Vérifier que $\lim_{p \rightarrow 0} \mathbb{E}_p[h(S_n)] = h(0)$. En déduire qu'il n'existe pas d'estimateur intégrable de $1/p$, fonction de S_n , qui soit sans biais.
5. Montrer alors qu'il n'existe pas d'estimateur intégrable de $1/p$ sans biais.

II Estimation asymptotique

On souhaite étudier l'estimateur de Bailey de N défini par $n_0 \delta_n$, où :

$$\delta_n = \frac{n+1}{S_n+1}.$$

1. Montrer que $(\delta_n, n \in \mathbb{N}^*)$ est une suite d'estimateurs convergente de $1/p$.
2. Quel est le biais de l'estimateur δ_n ?
3. Montrer que la suite $\left(\sqrt{n} \left(\frac{S_n}{n} - p\right), n \in \mathbb{N}^*\right)$ converge. Déterminer le mode de convergence et la limite.
4. Vérifier que la suite $\left(\sqrt{n} \left(\frac{S_n+1}{n+1} - \frac{S_n}{n}\right), n \in \mathbb{N}^*\right)$ converge p.s. vers 0. En déduire que $\left(\frac{S_n+1}{n+1}, n \in \mathbb{N}^*\right)$ est une suite d'estimateurs convergente de p asymptotiquement normale.
5. En déduire que la suite d'estimateurs $(\delta_n, n \in \mathbb{N}^*)$ est asymptotiquement normale de variance asymptotique $\sigma_p^2 = (1-p)/p^3$.
6. Calculer le score et l'information de Fisher de l'échantillon de taille 1. La suite d'estimateurs $(\delta_n, n \in \mathbb{N}^*)$ est-elle asymptotiquement efficace ?

III Intervalle de confiance

1. Donner un estimateur convergent de σ_p^2 . En déduire un intervalle de confiance pour $1/p$ de niveau asymptotique $1 - \alpha$, où $\alpha \in]0, 1[$.
2. Les données numériques suivantes proviennent d'un site d'études au Danemark⁴ dans les années 1960 : $n_0 = 74$, $n = 462$ et $s_n = 340$. Donner une estimation de $1/p$ puis de N .
3. Donner un intervalle de confiance pour $1/p$ puis pour N de niveau asymptotique 95%.

4. G.A.F. Seber. *The Estimation of Animal Abundance and Related Parameters*. Caldwell (1982).

IV Tests

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 $N_1 > N_0$.

1. Calculer le rapport de vraisemblance $Z(x)$, où $x \in \{0, 1\}^n$ et déterminer la statistique de test correspondante.
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 de la question III.2. 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 du tableau XII.1 de la fonction de répartition de S_n , avec $n = 462$ et $p_0 = 74/96$:

| c | 338 | 339 | 340 | 341 | 342 |
|--------------------------------|-----------|-----------|-----------|-----------|-----------|
| $\mathbb{P}_{p_0}(S_n \leq c)$ | 0.0270918 | 0.0344991 | 0.0435125 | 0.0543594 | 0.0672678 |

Table XII.1. Fonction de répartition de S_n , avec $n = 462$ et $p_0 = 74/96$.

△

XII.3 Comparaison de traitements

Exercice XII.3 (Données censurées et comparaison de traitements).

On considère une population de souris en présence de produits toxiques. Le but de cet exercice est d'étudier l'effet d'un pré-traitement⁵ sur le temps d'apparition des effets dus aux produits toxiques, tout en tenant compte du fait que certaines observations sont censurées (i.e. certaines souris sont retirées de l'expérience avant l'apparition des effets).

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 de paramètres $w \geq 0$, $k > 0$ et $\theta > 0$. Plus précisément, en posant $\bar{F}_\theta(t) = \mathbb{P}_\theta(T > t)$, on a :

$$F_\theta(t) = \begin{cases} e^{-\theta(t-w)^k} & \text{pour } t \geq w, \\ 1 & \text{pour } t < w. \end{cases}$$

5. M. Pike. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, vol. 22, pp. 142-161 (1966).

Le paramètre w correspond à la période de latence des produits toxiques. Le paramètre k est lié à l'évolution du taux de mortalité des souris en fonction de leur âge. Enfin le paramètre θ correspond plus directement à la sensibilité des souris aux produits toxiques. En particulier, on souhaite vérifier si le paramètre θ peut être modifié par un pré-traitement. 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, \dots, T_n un échantillon de n variables aléatoires indépendantes de même loi que T . Donner la densité $p_n^*(t; \theta)$ de l'échantillon où $t = (t_1, \dots, t_n) \in]w, +\infty[^n$.
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 .

On rappelle que la fonction Γ est définie, pour $\alpha > 0$, par $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$. Et si α est entier, alors on a $\Gamma(\alpha) = (\alpha - 1)!$.

5. Calculer $\mathbb{E}_\theta[(T - w)^k]$ et $\mathbb{E}_\theta[(T - w)^{2k}]$.
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 $n = 17$ observations dont les valeurs sont présentées dans le tableau XII.2. Pour ces observations, on admet que $w = 100$ et $k = 3$. On donne $\sum_{i=1}^n (t_i - 100)^3 = 33\,175\,533$. Donner une estimation de θ et un intervalle de confiance de niveau asymptotique 95%.

| | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 143 | 164 | 188 | 188 | 190 | 192 | 206 | 209 | 213 |
| 216 | 220 | 227 | 230 | 234 | 246 | 265 | 304 | |

Table XII.2. Temps d'apparition des effets toxiques pour $n = 17$ souris.

II Données censurées

Durant l'expérience, certaines souris 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 donc pas le

temps d'apparition des premiers effets, T , mais $R = \min(T, S)$, où S est le temps aléatoire où la souris est retirée de la population étudiée. On observe également la variable $X = \mathbf{1}_{\{T \leq S\}}$, où $X = 0$ si la souris est retirée de l'expérience avant que les effets dus aux produits toxiques apparaissent et $X = 1$ si les effets dus aux produits toxiques apparaissent avant que la souris soit retirée de l'expérience. On suppose que les variables T et S sont indépendantes et que S est une variable aléatoire positive 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 .

On note $p = \mathbb{P}_\theta(X = 1)$.

1. Quelle est la loi de X ?

La densité des données censurées $p(r, x; \theta)$ est la densité de la loi de (R, X) . Elle est définie pour tous $x \in \{0, 1\}$, $r \in \mathbb{R}$ par :

$$\int_r^\infty p(u, x; \theta) du = \mathbb{P}_\theta(R > r, X = x).$$

2. Calculer $\mathbb{P}_\theta(R > r, X = x)$, où $x \in \{0, 1\}$ et $r \in \mathbb{R}$. Puis, vérifier que $p(r, x; \theta) = \theta^x e^{[-\theta(r-w)_+^k]} c(r, x)$, où $z_+ = \max(z, 0)$ désigne la partie positive de z et où la fonction c est indépendante de θ .

Soit $(R_1, X_1), \dots, (R_n, X_n)$ un échantillon de n variables aléatoires indépendantes de même loi que (R, X) . Cet échantillon modélise les données censurées. On note $N_n = \sum_{i=1}^n X_i$.

3. 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 l'information de Fisher, $I(\theta)$, de θ pour l'échantillon de taille 1.

On admet dorénavant que l'estimateur du maximum de vraisemblance est un estimateur convergent asymptotiquement efficace (voir la partie IV).

7. Dédire des questions précédentes un estimateur convergent de $I(\theta)$.

8. Dans la question I.8, on a omis les 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 $k = 3$. On donne la valeur de $\sum_{i=n+1}^{n+2} (s_i - 100)^3 = 4\,546\,880$.

Donner une estimation de θ et un intervalle de confiance de niveau asymptotique 95%. Comparer cet intervalle de confiance avec celui obtenu à la question I.8.

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 la valeur du paramètre θ 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 $(R_1^A, X_1^A), \dots, (R_{n^A}^A, X_{n^A}^A)$ les données censurées de la population de n^A souris ayant subi le pré-traitement A et $(R_1^B, X_1^B), \dots, (R_{n^B}^B, X_{n^B}^B)$ 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 $1 \leq i \leq n^j$ et $j \in \{A, B\}$, sont indépendantes. On note :

$$p^A = \mathbb{P}_{(\theta^A, \theta^B)}(X_1^A = 1) \quad \text{et} \quad p^B = \mathbb{P}_{(\theta^A, \theta^B)}(X_1^B = 1).$$

1. Donner la densité de l'échantillon de taille $n^A + n^B$:

$$((R_1^A, X_1^A), \dots, (R_{n^A}^A, X_{n^A}^A), (R_1^B, X_1^B), \dots, (R_{n^B}^B, X_{n^B}^B)).$$

2. Sous l'hypothèse $H_0 = \{\theta^A = \theta^B\}$ donner l'estimateur du maximum de vraisemblance $\hat{\theta}_n$ de $\theta = \theta^A = \theta^B$.
3. En utilisant la question II.7, donner les deux tests asymptotiques convergents de Hausman pour tester l'hypothèse nulle H_0 contre l'hypothèse alternative $H_1 = \{\theta^A \neq \theta^B\}$.

Les données de l'échantillon A correspondent à celles de la question II.8. 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_1^B, \dots, t_{19}^B donnés dans le tableau XII.3, $\sum_{i=1}^{19} (t_i^B - 100)^3 = 64\,024\,591$,
- 2 souris qui ont été retirées de l'expérience aux instants s_i^B : 204 et 344, avec $\sum_{i=1}^2 (s_i^B - 100)^3 = 15\,651\,648$.

| | | | | | | | | | |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 142 | 156 | 163 | 198 | 205 | 232 | 232 | 233 | 233 | 233 |
| 233 | 239 | 240 | 261 | 280 | 280 | 296 | 296 | 323 | |

Table XII.3. Temps d'apparition des effets pour $n = 19$ souris avec le pré-traitement B .

4. Calculer $\hat{\theta}_n^A$, $\hat{\theta}_n^B$, et $\hat{\theta}_n$. Les pré-traitements A et B sont-ils différents ?

5. Donner des intervalles de confiance $J_{n^A}^A$ pour θ^A et $J_{n^B}^B$ pour θ^B tels que la confiance asymptotique sur les deux intervalles soit d'au moins 95% (*i.e.* pour n^A et n^B asymptotiquement grand, $\mathbb{P}_{(\theta^A, \theta^B)}(\theta^A \in J_{n^A}^A, \theta^B \in J_{n^B}^B) \geq 95\%$). Retrouvez-vous la conclusion de la question précédente ?

IV Propriétés asymptotiques de l'estimateur $\hat{\theta}_n$

1. Exprimer p comme une intégrale fonction de f_θ et \bar{G} .
2. Calculer $\mathbb{P}_\theta(R > r)$ à l'aide de \bar{F}_θ 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 $\beta = \mathbb{E}_\theta[\mathbf{1}_{\{X=1\}}(R - w)_+^k]$. Montrer 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.

△

XII.4 Ensemencement des nuages

Exercice XII.4 (Ensemencement des nuages).

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'agrègent 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. On signale 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 en 1975⁶ si l'ensemencement par iodure d'argent est efficace ou non.

On dispose des données des volumes de pluie déversée en 10^7 m³, voir les tableaux XII.4 et XII.5, concernant 23 jours similaires dont $m = 11$ jours avec ensemencement correspondant aux réalisations des variables aléatoires X_1, \dots, X_m et $n = 12$ jours sans ensemencement correspondant aux réalisations des variables aléatoires Y_1, \dots, Y_n . On suppose que les variables aléatoires X_1, \dots, X_m ont même loi, que les variables aléatoires Y_1, \dots, Y_n ont même loi, et que les variables $X_1, \dots, X_m, Y_1, \dots, Y_n$ 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 |

Table XII.4. Volume de pluie en 10^7 m³ déversée avec ensemencement.

| | | | | | | | | | | | | |
|-------|-------|-------|------|------|------|------|------|------|------|------|------|------|
| 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 |

Table XII.5. Volume de pluie en 10^7 m³ déversée sans ensemencement.

On considérera divers modèles pour la quantité d'eau déversée par les nuages afin de construire des tests pour l'hypothèse nulle $H_0 = \{\text{l'ensemencement n'accroît pas de manière significative la quantité d'eau déversée}\}$ contre l'hypothèse alternative $H_1 = \{\text{l'ensemencement accroît de manière significative la quantité d'eau déversée}\}$.

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 ?

6. W. L. Woodley, J. Simpson, R. Biondini and J. Berkeley. Rainfall results, 1970-1975 : Florida Area Cumuls Experiment. *Science*, vol. 195, pp. 735-742 (1977).

2. Donner la loi de $\hat{\nu}_n$. En déduire un intervalle de confiance de ν de niveau exact $1 - \alpha$. Faire l'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, \dots, X_m et Y_1, \dots, 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

$$\zeta_{m,n}^{(1)} = \sqrt{\frac{mn}{m+n}} \frac{\hat{\mu}_m - \hat{\nu}_n}{\sigma_0}.$$

Donner la loi de $\zeta_{m,n}^{(1)}$. En particulier, quelle est la loi de $\zeta_{m,n}^{(1)}$ sous H_0 ?

8. Montrer que pour tout $\mu > \nu$, on a p.s. :

$$\lim_{\min(m,n) \rightarrow \infty} \zeta_{m,n}^{(1)} = +\infty.$$

9. Déduire des questions précédentes un test convergent de niveau exact α . Déterminer la région critique correspondante.

10. On choisit $\alpha = 5\%$. Rejetez-vous l'hypothèse nulle ? Calculer $p\text{-val}_1$ 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 la loi de X_i possède une densité notée f . On ne fait pas d'hypothèse sur la forme de la densité ; on parle alors de modèle non-paramétrique. 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 Whitney :

$$U_{m,n} = \sum_{i=1}^m \sum_{j=1}^n \mathbf{1}_{\{Y_j \leq X_i\}}.$$

On pose $p = \mathbb{E}[\mathbf{1}_{\{Y_j \leq X_i\}}]$, et il vient $\mathbb{E}[U_{m,n}] = mnp$. On rappelle que :

$$\text{Var}(U_{m,n}) = mn^2\alpha' + m^2n\beta' + mn(p - p^2 - \alpha' - \beta'),$$

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 :

$$Z_{m,n} = \frac{U_{m,n} - mnp}{\sqrt{\text{Var}(U_{m,n})}}.$$

On rappelle que si $\min(m, n)$ tend vers l'infini, alors la suite $(Z_{m,n}, m \geq 1, n \geq 1)$ converge en loi vers la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.

On rappelle que sous H_0 (i.e. X_i et Y_j ont même loi), on a :

$$p = 1/2 \quad \text{et} \quad \text{Var}(U_{m,n}) = \frac{mn(m+n+1)}{12}.$$

On admet que la loi de $U_{m,n}$ sous H_0 ne dépend pas de F . On admet que sous H_1 , on a $p > 1/2$. 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 :

$$\zeta_{m,n}^{(2)} = \frac{U_{m,n} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}.$$

1. Montrer que :

$$\{\zeta_{m,n}^{(2)} \geq a\} = \{Z_{m,n} \geq b_{m,n}\},$$

où l'on déterminera $b_{m,n}$. Vérifier que sous H_1 , on a $\lim_{\min(m,n) \rightarrow \infty} b_{m,n} = -\infty$.

2. En déduire, que sous H_1 , on a pour tout $a > 0$, $\lim_{\min(m,n) \rightarrow \infty} \mathbb{P}(\zeta_{m,n}^{(2)} \geq a) = 1$.

3. Déduire des questions précédentes un test asymptotique convergent de niveau α . Ce test est appelé test de Mann et Whitney.

4. On choisit $\alpha = 5\%$. Rejetez-vous l'hypothèse nulle ? Calculer $p\text{-val}_2$ la p -valeur pour ce 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 :

$$\zeta_{m,n}^{(3)} = \sqrt{\frac{mn}{m+n}} \sup_{z \in \mathbb{R}} \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{X_i \leq z\}} - \frac{1}{n} \sum_{j=1}^n \mathbf{1}_{\{Y_j \leq z\}} \right|.$$

1. Donner un test asymptotique convergent de niveau α construit à partir de $\zeta_{m,n}^{(3)}$.
2. On observe la réalisation $\zeta_{m,n}^{(3),\text{obs}} \simeq 0.5082$. Rejetez-vous l'hypothèse nulle au niveau asymptotique de 5%? Calculer $p\text{-val}_3$ la p -valeur de ce modèle en utilisant les quelques valeurs, voir le tableau XII.6 de la fonction de répartition K de la loi limite sous H_0 de $\zeta_{m,n}^{(3)}$ quand $\min(m, n)$ tend vers l'infini.

| | | | | | |
|--------|-------|-------|-------|-------|-------|
| x | 0.519 | 0.571 | 0.827 | 1.223 | 1.358 |
| $K(x)$ | 0.05 | 0.10 | 0.50 | 0.90 | 0.95 |

Table XII.6. Quelques valeurs de la fonction de répartition K .

IV Conclusion

L'expérience d'ensemencement pratiquée en Floride est-elle concluante?

△

XII.5 Chaleur latente de fusion

Exercice XII.5 (Chaleur latente de fusion).

On considère deux méthodes A et B de mesure de la chaleur latente de fusion (ou enthalpie de changement d'état) de la glace en eau. Le tableau XII.7 donne pour les deux méthodes de mesure, les valeurs⁷ de la chaleur latente de fusion de la glace à -0.72°C en eau à 0°C en calories par gramme. On dispose de $n = 13$ mesures pour la méthode A et de $m = 8$ pour la méthode B . On souhaite savoir si la méthode B n'a pas tendance à donner des valeurs inférieures à celle de la méthode A .

Méthode A

| | | | | |
|-------|-------|-------|-------|-------|
| 79.98 | 80.04 | 80.02 | 80.04 | 80.03 |
| 80.03 | 80.04 | 79.97 | 80.05 | 80.03 |
| 80.02 | 80.00 | 80.02 | | |

Méthode B

| | | | | |
|-------|-------|-------|-------|-------|
| 80.02 | 79.94 | 79.98 | 79.97 | 79.97 |
| 80.03 | 79.95 | 79.97 | | |

Table XII.7. Mesures de la chaleur latente de fusion de l'eau en calories par gramme par deux méthodes différentes.

⁷ J. Rice. *Mathematical statistics and data analysis*. International Thomson Publishing, 2nd ed. (1995).

I Le modèle

On modélise la i -ème mesure par la méthode A (resp. B) par la variable aléatoire X_i (resp. Y_i) et on note x_i (resp. y_i) l'observation correspondante. On suppose que : les variables aléatoires $(X_i, i \geq 1)$ sont indépendantes de loi gaussienne de moyenne μ_1 et de variance σ_1^2 , les variables aléatoires $(Y_j, j \geq 1)$ sont indépendantes de loi gaussienne de moyenne μ_2 et de variance σ_2^2 , et les variables aléatoires $(X_i, i \geq 1)$ et $(Y_j, j \geq 1)$ sont indépendantes. On note n le nombre de mesures faites avec la méthode A et m avec la méthode B . 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.

On pose :

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i, \quad V_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{n}{n-1} \left(\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}_n^2 \right),$$

$$\bar{Y}_m = \frac{1}{m} \sum_{j=1}^m Y_j, \quad W_m = \frac{1}{m-1} \sum_{j=1}^m (Y_j - \bar{Y}_m)^2 = \frac{m}{m-1} \left(\frac{1}{m} \sum_{j=1}^m Y_j^2 - \bar{Y}_m^2 \right).$$

On note $\bar{x}_n, v_n, \bar{y}_m, w_m$ les valeurs de $\bar{X}_n, V_n, \bar{Y}_m, W_m$ évaluées en les observations. On donne les valeurs numériques à 10^{-3} près correspondant aux observations :

$$n = 13, \quad \bar{x}_n \simeq 80.02, \quad v_n \simeq 5.7 \cdot 10^4, \quad (\text{XII.4})$$

$$m = 8, \quad \bar{y}_m \simeq 79.98, \quad w_m \simeq 9.8 \cdot 10^{-4}. \quad (\text{XII.5})$$

1. Décrire simplement, pour ce modèle, les hypothèses nulle et alternative correspondant à la question posée.
2. Donner la loi de $(X_1, \dots, X_n, Y_1, \dots, Y_m)$.
3. Rappeler la loi de $\left(\bar{X}_n, \frac{n-1}{\sigma_1^2} V_n \right)$.
4. Dédurre des deux questions précédentes la loi du couple $\left(\frac{n-1}{\sigma_1^2} V_n, \frac{m-1}{\sigma_2^2} W_m \right)$ et 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 ?

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 $\min(n, m)$ tend vers l'infini ?

Soit $\alpha_1, \alpha_2 \in]0, 1/2[$. On note a_{n,m,α_1} et b_{n,m,α_2} les quantiles d'ordre α_1 et $1 - \alpha_2$ de $F_{n,m}$ de loi de Fisher-Snedecor de paramètre (n, m) *i.e.* :

$$\mathbb{P}(F_{n,m} \leq a_{n,m,\alpha_1}) = \alpha_1 \quad \text{et} \quad \mathbb{P}(F_{n,m} \geq b_{n,m,\alpha_2}) = \alpha_2.$$

On admet les convergences suivantes :

$$\lim_{\min(n,m) \rightarrow \infty} a_{n,m,\alpha_1} = \lim_{\min(n,m) \rightarrow \infty} b_{n,m,\alpha_2} = 1. \quad (\text{XII.6})$$

3. Montrer que le test pur de région critique :

$$\tilde{W}_{n,m} = \{Z_{n,m} \notin]a_{n-1,m-1,\alpha_1}, b_{n-1,m-1,\alpha_2}[\}$$

est un test convergent, quand $\min(n, m)$ tend vers l'infini, pour tester H_0 contre H_1 . Déterminer son niveau, et vérifier qu'il ne dépend pas de (n, m) .

4. On choisit usuellement $\alpha_1 = \alpha_2$. Déterminer, en utilisant les tables XII.8 et XII.9, la région critique de niveau $\alpha = 5\%$. Calculer la p -valeur du test. Conclure.
5. Établir les convergences (XII.6).

| $(k, l) \backslash \alpha$ | 0.400 | 0.300 | 0.200 | 0.100 | 0.050 | 0.025 |
|----------------------------|-------|-------|-------|-------|-------|-------|
| (12, 7) | 1.244 | 1.510 | 1.906 | 2.667 | 3.574 | 4.666 |
| (12, 8) | 1.219 | 1.465 | 1.825 | 2.502 | 3.284 | 4.200 |
| (13, 7) | 1.247 | 1.510 | 1.901 | 2.655 | 3.550 | 4.628 |
| (13, 8) | 1.221 | 1.464 | 1.820 | 2.487 | 3.259 | 4.162 |
| (7, 12) | 1.142 | 1.373 | 1.700 | 2.283 | 2.913 | 3.607 |
| (7, 13) | 1.134 | 1.359 | 1.676 | 2.234 | 2.832 | 3.483 |
| (8, 12) | 1.148 | 1.371 | 1.686 | 2.245 | 2.848 | 3.512 |
| (8, 13) | 1.140 | 1.357 | 1.661 | 2.195 | 2.767 | 3.388 |

Table XII.8. 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 $\mathbb{P}(F_{12,7} \geq 1.244) \simeq 0.4$.

| $(k, l) \backslash \alpha$ | 0.025 | 0.050 | 0.100 | 0.200 | 0.300 | 0.400 |
|----------------------------|-------|-------|-------|-------|-------|-------|
| (12, 7) | 0.277 | 0.343 | 0.438 | 0.588 | 0.728 | 0.875 |
| (12, 8) | 0.285 | 0.351 | 0.446 | 0.593 | 0.729 | 0.871 |
| (13, 7) | 0.287 | 0.353 | 0.448 | 0.597 | 0.736 | 0.882 |
| (13, 8) | 0.295 | 0.361 | 0.456 | 0.602 | 0.737 | 0.877 |
| (7, 12) | 0.214 | 0.280 | 0.375 | 0.525 | 0.662 | 0.804 |
| (7, 13) | 0.216 | 0.282 | 0.377 | 0.526 | 0.662 | 0.802 |
| (8, 12) | 0.238 | 0.305 | 0.400 | 0.548 | 0.683 | 0.820 |
| (8, 13) | 0.240 | 0.307 | 0.402 | 0.550 | 0.683 | 0.819 |

Table XII.9. 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 $\mathbb{P}(F_{12,7} \leq 0.277) \simeq 0.025$.

III Comparaison de moyenne

On suppose dorénavant que $\sigma_1 = \sigma_2$, et on note σ la valeur commune (inconnue). On considère les hypothèses nulle $H_0 = \{\mu_1 = \mu_2, \sigma > 0\}$ et alternative $H_1 = \{\mu_1 > \mu_2, \sigma > 0\}$. On pose :

$$S_{n,m} = \frac{(n-1)V_n + (m-1)W_m}{n+m-2},$$

et on considère la statistique de test :

$$T_{n,m} = \sqrt{\frac{nm}{n+m}} \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{S_{n,m}}}.$$

- Déduire de la partie I la loi de $\frac{n+m-2}{\sigma^2} S_{n,m}$, et la limite de la suite $(S_{n,m}, n \geq 2, m \geq 2)$ quand $\min(n, m)$ tend vers l'infini.
- Déduire de la partie I la loi de $\frac{1}{\sigma} \sqrt{\frac{nm}{n+m}} (\bar{X}_n - \bar{Y}_m)$.
- Vérifier que sous H_0 , la loi de $T_{n,m}$ est une loi de Student dont on précisera les paramètres.
- Déterminer la limite sous H_1 de la suite $(\bar{X}_n - \bar{Y}_m, n \geq 1, m \geq 1)$ quand $\min(n, m)$ tend vers l'infini.
- En déduire sous H_1 la valeur de $\lim_{\min(n,m) \rightarrow \infty} T_{n,m}$.
- Construire un test pur convergent quand $\min(n, m)$ tend vers l'infini de niveau α pour tester H_0 contre H_1 .
- On choisit $\alpha = 5\%$. Déterminer la région critique du test asymptotique. Donner, en utilisant la table XII.10 une valeur approchée de la p -valeur. Conclure.

| $n \backslash \alpha$ | 0.05000 | 0.02500 | 0.01000 | 0.00500 | 0.00250 | 0.00100 | 0.00050 | 0.00025 | 0.00010 |
|-----------------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| 19 | 1.729 | 2.093 | 2.539 | 2.861 | 3.174 | 3.579 | 3.883 | 4.187 | 4.590 |
| 20 | 1.725 | 2.086 | 2.528 | 2.845 | 3.153 | 3.552 | 3.850 | 4.146 | 4.538 |
| 21 | 1.721 | 2.080 | 2.518 | 2.831 | 3.135 | 3.527 | 3.819 | 4.110 | 4.493 |

Table XII.10. 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 $\mathbb{P}(T_{19} \geq 1.729) \simeq 0.05$.

IV Variante sur les hypothèses du test

On reprend les notations de la partie III. Si on ne présuppose pas, $\mu_1 \geq \mu_2$, alors on considère l'hypothèse alternative $H'_1 = \{\mu_1 \neq \mu_2, \sigma > 0\}$ au lieu de H_1 .

- Quelles sont sous H'_1 les valeurs de $\lim_{\min(n,m) \rightarrow \infty} T_{n,m}$?
- 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.
- On considère l'hypothèse nulle $H'_0 = \{\mu_1 \leq \mu_2, \sigma > 0\}$ et l'hypothèse alternative H_1 . Vérifier que pour $c \in \mathbb{R}$:

$$\mathbb{P}_{(\mu_1, \sigma, \mu_2, \sigma)}(T_{n,m} \geq c) = \mathbb{P}_{(0, \sigma, 0, \sigma)}\left(T_{n,m} \geq c - \sqrt{\frac{nm}{n+m}} \frac{\mu_1 - \mu_2}{\sqrt{S_{n,m}}}\right)$$

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.

△

XII.6 Taille des grandes villes

Exercice XII.6 (Loi de Pareto et taille des villes).

Les lois en puissance semblent correspondre à de nombreux phénomènes⁸ : nombre d'habitants des villes, nombre de téléchargements des pages web, nombre d'occurrences des mots du langage... L'objectif de ce problème est d'étudier la famille des lois de Pareto, et de comparer à l'aide d'un tel modèle les nombres, convenablement renormalisés, d'habitants des plus grandes villes européennes et américaines.

On considère la loi de Pareto (réduite) de paramètre $\alpha > 0$, de densité :

$$f_\alpha(x) = \frac{\alpha}{x^{\alpha+1}} \mathbf{1}_{\{x>1\}}.$$

La partie I est consacrée à la recherche d'estimateurs du paramètre α . Dans la partie II, on construit un test pour comparer les paramètres α provenant de deux échantillons différents (villes européennes et villes américaines). La partie III concerne l'application numérique. La partie IV est indépendante des autres parties, et permet de comprendre que les données sur les plus grandes villes sont suffisantes et naturelles pour l'estimation du paramètre α .

I Estimations et intervalles de confiance du paramètre de la loi de Pareto

Soit $(X_k, k \in \mathbb{N}^*)$ une suite de variables aléatoires indépendantes de même loi de Pareto de paramètre $\alpha > 0$.

1. Calculer $\mathbb{E}_\alpha[X_1]$, puis $\mathbb{E}_\alpha[X_1^2]$.
2. Déduire du calcul de $\mathbb{E}_\alpha[X_1]$ un estimateur $\tilde{\alpha}_n$ de α , construit à partir de X_1, \dots, X_n . Vérifier que, pour $\alpha > 1$, l'estimateur $(\tilde{\alpha}_n, n \geq 1)$ est convergent.
3. Donner la vraisemblance du modèle et déterminer une statistique exhaustive. Que dire de l'estimateur $\tilde{\alpha}_n$?
4. Montrer que l'estimateur du maximum de vraisemblance de α , construit à partir de X_1, \dots, X_n , est $\hat{\alpha}_n = \frac{n}{\sum_{k=1}^n \log(X_k)}$.

8. M. E. J. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, vol. 46(5), pp. 323-351 (2005).

5. Montrer que la loi de $\log(X_1)$ est une loi exponentielle dont on précisera le paramètre. En déduire $\mathbb{E}_\alpha[\log(X_1)]$ et $\mathbb{E}_\alpha[\log(X_1)^2]$.
6. Vérifier directement, à l'aide de la question I.5, que la suite $(\hat{\alpha}_n, n \geq 1)$ est un estimateur convergent de α .
7. Montrer directement, à l'aide de la question I.5, que l'estimateur $(\hat{\alpha}_n, n \geq 1)$ est asymptotiquement normal. Montrer que sa variance asymptotique est α^2 .
8. Calculer l'information de Fisher. L'estimateur est-il asymptotiquement efficace ?
9. Construire un intervalle de confiance de niveau asymptotique $1 - \eta$ pour α .
10. Montrer, à l'aide des fonctions caractéristiques, que la loi de $\alpha\hat{\alpha}_n^{-1}$ est une loi gamma dont les paramètres ne dépendent pas de α . Construire alors un intervalle de confiance de niveau exact $1 - \eta$ pour α , utilisant les quantiles des lois gamma.

II Comparaison d'échantillons

Pour une région du globe r , on suppose que les nombres d'habitants des villes suivent approximativement une loi de Pareto générale qui sera introduite dans la partie IV. Ce modèle est raisonnable pour les grandes villes⁹. On note $\tilde{x}_{(1)}^r > \dots > \tilde{x}_{(n^r+1)}^r$ les nombres d'habitants des $n^r + 1$ plus grandes villes¹⁰. On vérifiera dans la partie IV, que les observations renormalisées par le minimum $\frac{\tilde{x}_{(1)}^r}{\tilde{x}_{(n^r+1)}^r} > \dots > \frac{\tilde{x}_{(n^r)}^r}{\tilde{x}_{(n^r+1)}^r}$ correspondent alors au réordonnement décroissant de réalisations de n variables aléatoires, (X_1^r, \dots, X_n^r) , indépendantes et de même loi de Pareto de paramètre α^r .

Le paramètre α^r peut s'interpréter comme le rapport du taux de naissance plus le taux d'apparition de nouvelles grandes villes sur le taux de naissance de la région r . Plus α^r est grand et plus la probabilité d'observer des (relativement) très grandes villes est faible.

On dispose des nombres d'habitants des 266 ($n^{\text{UE}} = 265$) plus grandes villes de l'Union Européenne (UE) et des 200 ($n^{\text{USA}} = 199$) plus grandes villes des États-Unis d'Amérique (USA). Les histogrammes des figures XII.1 (UE) et XII.2 (USA) concernent les données estimées en 2005. On a également porté la densité de la loi de Pareto avec le paramètre estimé, pour se convaincre que la modélisation est crédible.

9. Y. Ijiri and H. A. Simon. Some distributions associated with Bose-Einstein statistics. *Proc. Nat. Acad. Sci. U.S.A.*, vol. 72, pp. 1654-1657 (1975).

10. Données disponibles sur <http://www.citymayors.com/>

Pour simplifier l'écriture, on pose $n = n^{\text{UE}}$ et $m = n^{\text{USA}}$.

Le but de cette partie est de savoir s'il existe relativement moins de très grandes villes dans l'UE qu'aux USA. Pour cela on considère les hypothèses $H_0 = \{\alpha^{\text{UE}} \leq \alpha^{\text{USA}}\}$ et $H_1 = \{\alpha^{\text{UE}} > \alpha^{\text{USA}}\}$.

1. Existe-t-il un lien entre les variables $(X_1^{\text{UE}}, \dots, X_n^{\text{UE}})$ et $(X_1^{\text{USA}}, \dots, X_m^{\text{USA}})$?
2. Comme les nombres d'habitants des villes sont donnés par ordre décroissant, cela signifie que l'on observe seulement une réalisation du réordonnement décroissant. Vérifier que l'estimateur du maximum de vraisemblance de α^r , $\hat{\alpha}_{n^r}^r$, défini dans la partie I, peut s'écrire comme une fonction du réordonnement décroissant.

Pour $k \in \mathbb{N}^*$, on pose $Z_k^r = \sqrt{k}(\hat{\alpha}_k^r - \alpha^r)$ et on note ψ_k^r la fonction caractéristique de Z_k^r . La question I.7 assure la convergence simple de $(\psi_k^r, k \in \mathbb{N}^*)$ vers la fonction caractéristique de la loi gaussienne $\mathcal{N}(0, (\alpha^r)^2)$. On admet que la convergence est en fait uniforme sur les compacts, pour tout $u \in \mathbb{R}$:

$$\lim_{k \rightarrow \infty} \sup_{\varepsilon \in [0,1]} \left| \psi_k^r(\varepsilon u) - e^{-\varepsilon^2 u^2 (\alpha^r)^2 / 2} \right| = 0.$$

3. Montrer que si $\alpha^{\text{UE}} = \alpha^{\text{USA}}$, alors la suite :

$$\left(\sqrt{\frac{m}{n+m}} Z_n^{\text{UE}} - \sqrt{\frac{n}{n+m}} Z_m^{\text{USA}}, n \in \mathbb{N}^*, m \in \mathbb{N}^* \right)$$

converge en loi, quand $\min(n, m)$ tend vers l'infini, vers une loi gaussienne de moyenne nulle et de variance α^2 .

On considère :

$$\hat{\sigma}_{n,m}^2 = \frac{n(\hat{\alpha}_n^{\text{UE}})^2 + m(\hat{\alpha}_m^{\text{USA}})^2}{n+m},$$

et la statistique de test :

$$\zeta_{n,m} = \sqrt{\frac{nm}{n+m}} \frac{\hat{\alpha}_n^{\text{UE}} - \hat{\alpha}_m^{\text{USA}}}{\hat{\sigma}_{n,m}} = \sqrt{nm} \frac{\hat{\alpha}_n^{\text{UE}} - \hat{\alpha}_m^{\text{USA}}}{\sqrt{n(\hat{\alpha}_n^{\text{UE}})^2 + m(\hat{\alpha}_m^{\text{USA}})^2}}.$$

4. Dédurre de la question précédente que, si $\alpha^{\text{UE}} = \alpha^{\text{USA}}$, alors la suite $(\zeta_{n,m}, n \in \mathbb{N}^*, m \in \mathbb{N}^*)$ converge en loi, quand $\min(n, m)$ tend vers l'infini, vers la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.
5. Si $\alpha^{\text{UE}} < \alpha^{\text{USA}}$, donner le comportement asymptotique de la suite $(\zeta_{n,m}, n \in \mathbb{N}^*, m \in \mathbb{N}^*)$ quand $\min(n, m)$ tend vers l'infini.
6. Si $\alpha^{\text{UE}} > \alpha^{\text{USA}}$, donner le comportement asymptotique de la suite $(\zeta_{n,m}, n \in \mathbb{N}^*, m \in \mathbb{N}^*)$ quand $\min(n, m)$ tend vers l'infini.

7. En déduire la forme de la région critique pour tester asymptotiquement H_0 contre H_1 .
8. Montrer, en utilisant le fait que la loi de $\hat{\alpha}_{ni}^r/\alpha^r$ ne dépend pas de α^r (cf question I.10), que l'erreur de première espèce est maximale pour $\alpha^{\text{UE}} = \alpha^{\text{USA}}$.
9. En admettant le résultat de la question précédente, donner la région critique de niveau asymptotique η pour tester H_0 contre H_1 .
10. Ce test est-il convergent ?
11. Comment calculer la p -valeur asymptotique de ce test ?

III Application numérique

On donne $\eta = 5\%$ et dans le tableau XII.11 les statistiques pour les nombres d'habitants des plus grandes villes de l'Union Européenne et des États-Unis d'Amérique.

| | | |
|---|----------------|-----------------|
| Région : r | UE | USA |
| Nombre de données : n^r | 265 | 199 |
| Plus grande ville : $\tilde{x}_{(1)}$ | (Londres) 7.07 | (New York) 8.08 |
| Plus petite ville : $\tilde{x}_{(n^r+1)}$ | 0.15 | 0.12 |
| $\sum_{k=1}^{n^r} x_k$ | 676.8 | 620.8 |
| $\sum_{k=1}^{n^r} x_k^2$ | 5584.6 | 8704.6 |
| $\sum_{k=1}^{n^r} \log(x_k)$ | 166.6 | 147.6 |
| $\sum_{k=1}^{n^r} \log(x_k)^2$ | 210.2 | 207.6 |

Table XII.11. Données 2005 sur les plus grandes villes de l'UE et des USA. Les nombres d'habitants sont en millions.

1. Donner les estimations par maximum de vraisemblance de α^{UE} , α^{USA} et leurs intervalles de confiance asymptotiques de niveau $1 - \eta$ (cf. partie I).
2. Calculer la p -valeur du test présenté dans la partie II.
3. Conclusion ?

IV Réduction des lois de Pareto

Soit une suite $(\tilde{X}_n, n \in \mathbb{N}^*)$ de variables aléatoires indépendantes de loi de Pareto de paramètre $(\alpha, \beta) \in]0, \infty[^2$ de densité :

$$f_{\alpha, \beta}(x) = \alpha \frac{\beta^\alpha}{x^{\alpha+1}} \mathbf{1}_{\{x > \beta\}}.$$

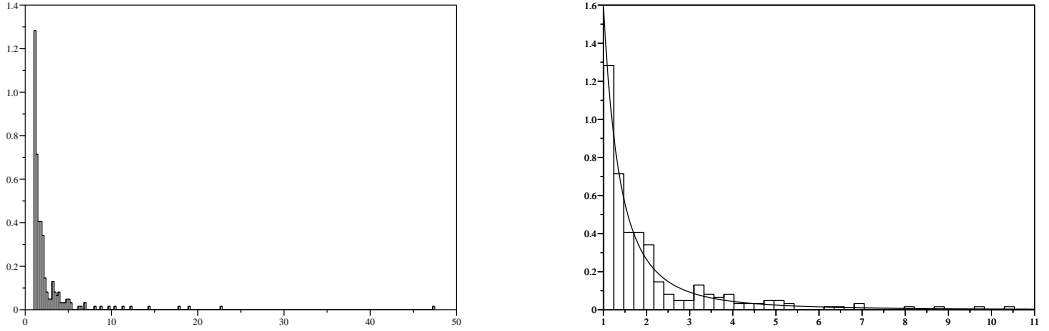


Figure XII.1. Histogramme des nombres d'habitants des $n = 265$ plus grandes villes de l'Union Européenne (divisées par la taille de la $(n + 1)$ -ième grande ville), et densité de la loi de Pareto de paramètre $\hat{\alpha}_n^{\text{UE}}$ (sur le graphique de droite seulement).

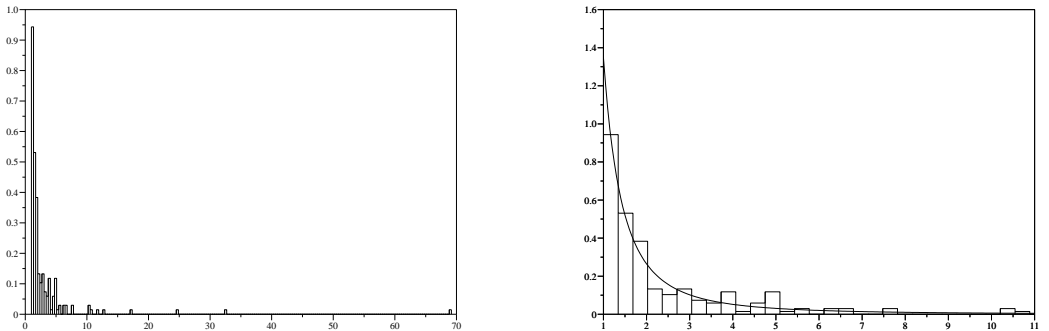


Figure XII.2. Histogramme des nombres d'habitants des $m = 199$ plus grandes villes des États Unis d'Amérique (divisées par la taille de la $(m + 1)$ -ième grande ville), et densité de la loi de Pareto de paramètre $\hat{\alpha}_m^{\text{USA}}$ (sur le graphique de droite seulement).

1. Calculer la fonction de répartition, $F_{\alpha, \beta}$, de la loi de Pareto de paramètre (α, β) .
2. Calculer $\mathbb{P}(\tilde{X}_1/y > x \mid \tilde{X}_1 > y)$ pour $y > \beta$ et $x > 1$. En déduire la fonction de répartition, puis la loi, de \tilde{X}_1/y sachant $\tilde{X}_1 > y$, où $y > \beta$.

Soit $n, k \in \mathbb{N}^*$. On considère le réordonnement décroissant, appelé aussi statistique d'ordre, de $(\tilde{X}_1, \dots, \tilde{X}_{n+k})$, que l'on note $(\tilde{X}_{(1)}, \dots, \tilde{X}_{(n+k)})$ et qui, on l'admet, est p.s. uniquement défini par :

$$\tilde{X}_{(1)} > \dots > \tilde{X}_{(n+k)} \quad \text{et} \quad \{\tilde{X}_1, \dots, \tilde{X}_{n+k}\} = \{\tilde{X}_{(1)}, \dots, \tilde{X}_{(n+k)}\}.$$

En particulier on a $\tilde{X}_{(1)} = \max_{1 \leq i \leq n+k} \tilde{X}_i$ et $\tilde{X}_{(n+k)} = \min_{1 \leq i \leq n+k} \tilde{X}_i$. On admet que le réordonnement décroissant est un vecteur de loi continue et de densité :

$$g_{n+k}(x_1, \dots, x_{n+k}) = (n+k)! \mathbf{1}_{\{x_1 > \dots > x_{n+k}\}} \prod_{i=1}^{n+k} f_{\alpha, \beta}(x_i).$$

Le but de ce qui suit est de déterminer la loi des n plus grandes valeurs divisées par la $(n+1)$ -ième, c'est-à-dire de $(Y_1, \dots, Y_n) = \left(\frac{\tilde{X}_{(1)}}{\tilde{X}_{(n+1)}}, \dots, \frac{\tilde{X}_{(n)}}{\tilde{X}_{(n+1)}} \right)$.

1. Montrer que la densité de la loi de $(\tilde{X}_{(1)}, \dots, \tilde{X}_{(n+1)})$ est :

$$\frac{(n+k)!}{(k-1)!} \mathbf{1}_{\{x_1 > \dots > x_{n+1}\}} F_{\alpha, \beta}(x_{n+1})^{k-1} f_{\alpha, \beta}(x_{n+1}) \prod_{i=1}^n f_{\alpha, \beta}(x_i).$$

2. Montrer que (Y_1, \dots, Y_n) a même loi que le réordonnement décroissant de (X_1, \dots, X_n) , où les variables X_1, \dots, X_n sont indépendantes de même loi de Pareto de paramètre $(\alpha, 1)$. Vérifier ainsi que la loi de (Y_1, \dots, Y_n) ne dépend ni de β ni de k .

Quitte à considérer les $n+1$ plus grandes valeurs de la suite $(\tilde{X}_i, 1 \leq i \leq n+k)$, on peut les remplacer par les n plus grandes divisées par la $(n+1)$ -ième, et supposer ainsi que l'on considère le réordonnement décroissant de n variables aléatoires indépendantes de loi de Pareto de paramètre $(\alpha, 1)$.

△

XII.7 Résistance d'une céramique

Exercice XII.7 (Résistance d'une céramique et loi de Weibull).

Les céramiques possèdent de nombreux défauts comme des pores ou des microfissures qui sont répartis aléatoirement. Leur résistance à la rupture est modélisée par une variable aléatoire de loi de Weibull¹¹.

La première partie de l'exercice permet de manipuler la loi de Weibull. La deuxième permet d'expliquer le choix de la loi de Weibull pour la loi de la résistance à la rupture. Les troisième et quatrième parties abordent l'estimation des paramètres de la loi de Weibull¹². La partie II d'une part, et les parties III et IV d'autre part, sont indépendantes.

11. Norme ISO 20501 : 2003, <http://www.iso.org/iso/fr/>

12. P. Murthy, M. Xie and R. Jiang. *Weibull models*. Wiley Series in Probability and Statistics (2004).

La densité de la loi de Weibull de paramètre $(\alpha, \beta) \in]0, \infty[^2$ est définie sur \mathbb{R} par :

$$f_{\alpha, \beta}(x) = \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^\alpha\right) \mathbf{1}_{\{x>0\}}$$

et la fonction de répartition par $F_{\alpha, \beta}(x) = 0$ si $x \leq 0$ et :

$$F_{\alpha, \beta}(x) = 1 - \exp\left(-\left(\frac{x}{\beta}\right)^\alpha\right) \quad \text{pour } x \geq 0.$$

Le paramètre de forme α est, en analyse des résistances des céramiques, appelé module de Weibull. Ce paramètre, qui caractérise l'homogénéité de la céramique, est adimensionnel. Il est d'autant plus élevé que la céramique est homogène. Les valeurs actuelles sont de l'ordre de 5 à 30. Le paramètre d'échelle β est parfois appelé contrainte caractéristique ($F_{\alpha, \beta}(\beta) = 1 - e^{-1} \simeq 63.21\%$). Il est exprimé dans les mêmes unités que les contraintes (*i.e.* en Pascal), il dépend de la qualité et de la taille de la céramique.

I Étude sommaire de la loi de Weibull

On rappelle la définition de la fonction Gamma, pour $r > 0$:

$$\Gamma(r) = \int \mathbf{1}_{\{t>0\}} t^{r-1} e^{-t} dt.$$

On a $\Gamma(r + 1) = r\Gamma(r)$ pour $r > 0$, et $\Gamma(r + 1) = r!$ pour $r \in \mathbb{N}^*$.

Soit X une variable aléatoire de loi de Weibull de paramètre $(\alpha, \beta) \in]0, \infty[^2$.

1. Soit $c > 0$. Déterminer, à l'aide des fonctions de répartition, la loi de cX .
2. Montrer que $\mathbb{E}_{(\alpha, \beta)}[X^\alpha] = \beta^\alpha$ et $\mathbb{E}_{(\alpha, \beta)}[X^{2\alpha}] = 2\beta^{2\alpha}$.
3. Calculer la loi de X^α et reconnaître une loi exponentielle dont on déterminera le paramètre. Retrouver les résultats de la question précédente.

II Pourquoi la loi de Weibull ?

On considère une barre de céramique de longueur L soumise à une force de traction. On modélise la résistance de la barre de céramique, *i.e.* la valeur de la force de traction qui fait se rompre la barre, par une variable aléatoire $X^{(L)}$. On décompose la barre de céramique en n tranches de longueur L/n , et on note $X_i^{(L/n)}$ la résistance de la tranche $i \in \{1, \dots, n\}$. La résistance de la barre de céramique est simplement la résistance de la tranche la plus faible :

$$X^{(L)} = \min_{1 \leq i \leq n} X_i^{(L/n)}. \tag{XII.7}$$

Ce modèle de tranche la plus faible intervient également en fiabilité quand on considère la durée de fonctionnement d'un appareil formé de composants en série, *i.e.* quand la durée de fonctionnement de l'appareil est la plus petite durée de fonctionnement de ses composants. On suppose que :

- (A) Les variables aléatoires $(X_i^{(L/n)}, i \in \{1, \dots, n\})$ sont indépendantes de même loi.
- (B) Pour tout $\ell > 0$, $X^{(\ell)}$ a même loi que $c_\ell X$, où $c_\ell > 0$ est une constante qui dépend de ℓ (et de la qualité de la céramique) et X est une variable aléatoire strictement positive.

Le but des questions qui suivent est d'identifier les lois possibles de X telles que les hypothèses (A) et (B) ainsi que l'équation (XII.7) soient satisfaites.

Soit $(X_k, k \geq 1)$ une suite de variables aléatoires indépendantes et de même loi.

1. On note F la fonction de répartition de X_1 et F_n celle de $\min_{1 \leq k \leq n} X_k$. Montrer que $1 - F_n(x) = (1 - F(x))^n$ pour tout $x \in \mathbb{R}$.
2. Dédire de la question précédente et de la question I.1 que si la loi de X_1 est la loi de Weibull de paramètres (α, β) , alors $\min_{1 \leq k \leq n} X_k$ a même loi que $n^{-1/\alpha} X_1$.
3. Montrer que si X suit la loi de Weibull de paramètres (α, β) , alors sous les hypothèses (A) et (B), l'égalité (XII.7) est satisfaite en loi dès que $c_\ell = c_1 \ell^{-1/\alpha}$ pour tout $\ell > 0$. Déterminer la loi de $X^{(L)}$.

On suppose que la fonction de répartition de X , H , possède l'équivalent suivant quand x tend vers 0 par valeurs supérieures :

$$H(x) = bx^a + o(x^a) \quad (x > 0),$$

où $a > 0$ et $b > 0$. On suppose également que la résistance est inversement proportionnelle à la longueur de la barre de céramique : $c_\ell = c_1 \ell^{-1/\alpha}$, avec $\alpha > 0$.

4. Donner la limite de la fonction de répartition de la variable aléatoire $(\min_{1 \leq i \leq n} X_i^{(L/n)}, n \geq 1)$. En déduire que l'égalité (XII.7) implique que $a = \alpha$ ainsi que $X^{(L)}$ et X suivent des lois de Weibull.

En fait la théorie des lois de valeurs extrêmes^{13 14} assure que si l'on suppose les hypothèses (A) et (B), l'égalité (XII.7) et le fait que la résistance $X^{(L)}$ n'est pas déterministe (et donc sans hypothèse sur la fonction de répartition de X ni sur c_ℓ) alors $X^{(L)}$ et X suivent des lois de Weibull. En particulier l'indépendance des résistances des tranches conduisent à modéliser la résistance d'une céramique par une variable aléatoire de Weibull.

13. J. Beirlant, Y. Goegebeur, J. Teugels and J. Segers. *Statistics of extremes*. Wiley Series in Probability and Statistics (2004).

14. P. Embrechts, C. Klueppelberg and T. Mikosch. *Modelling extremal events for insurance and finance*. Springer (1997).

III Estimation du paramètre d'échelle

Soit $(X_k, k \geq 1)$ une suite de variables aléatoires indépendantes de même loi de Weibull de paramètre $(\alpha_0, \beta) \in]0, \infty[^2$, où α_0 est supposé **connu**.

1. Montrer que la vraisemblance associée à un échantillon X_1, \dots, X_n de taille n s'écrit, avec $x = (x_k, k \in \{1, \dots, n\})$:

$$p_n(x; \beta) = \frac{\alpha_0^n}{\beta^{n\alpha_0}} e^{-\beta^{-\alpha_0} \sum_{i=1}^n x_i^{\alpha_0}} \prod_{j=1}^n x_j^{\alpha_0-1} \prod_{l=1}^n \mathbf{1}_{\{x_l > 0\}}.$$

2. En utilisant le théorème de factorisation, donner une statistique exhaustive pertinente.
3. Déterminer l'estimateur du maximum de vraisemblance, $\hat{\beta}_n$, de β .
4. En utilisant la question I.3, calculer le biais de $\hat{\beta}_n$. Vérifier qu'il est nul si $\alpha_0 = 1$.
5. À l'aide des résultats de la question I.2, montrer directement que $\hat{\beta}_n^{\alpha_0}$ est un estimateur convergent et asymptotiquement normal de β^{α_0} .
6. En déduire que $\hat{\beta}_n$ est un estimateur convergent et asymptotiquement normal de β . Donner la variance asymptotique.

IV Estimation du paramètre de forme

Soit $(X_k, k \geq 1)$ une suite de variables aléatoires indépendantes de même loi de Weibull de paramètre $(\alpha, \beta) \in]0, \infty[^2$. On suppose que le paramètre $(\alpha, \beta) \in]0, \infty[^2$ est inconnu.

1. Pour un échantillon de taille n , donner la vraisemblance et la log-vraisemblance, $L_n(x; (\alpha, \beta))$ où $x = (x_k, k \in \{1, \dots, n\})$.
2. Peut-on résumer les données à l'aide d'une statistique exhaustive ?
3. Donner les équations satisfaites par tout extremum, $(\tilde{\alpha}_n, \tilde{\beta}_n)$, de la log-vraisemblance.
4. Vérifier que $\tilde{\beta}_n = u_n(\tilde{\alpha}_n)$, pour une fonction u_n qui dépend de x . Que pensez-vous de l'estimation de β que α soit connu ou non ?

On définit la fonction h_n sur $]0, \infty[$ par :

$$h_n(a) = -\frac{1}{n} \sum_{i=1}^n \log(x_i) + \frac{\sum_{j=1}^n \log(x_j) x_j^a}{\sum_{l=1}^n x_l^a},$$

où $x_i > 0$ pour tout $i \in \{1, \dots, n\}$. On suppose $n \geq 2$ et qu'il existe $i, j \in \{1, \dots, n\}$ tels que $x_i \neq x_j$. On admet que la fonction h_n est strictement croissante.

5. Vérifier, en utilisant $\tilde{\beta}_n = u_n(\tilde{\alpha}_n)$, que $\tilde{\alpha}_n$ est l'unique solution de l'équation $h_n(a) = \frac{1}{a}$.
6. Montrer que la fonction $g : a \mapsto L_n(x; (a, u_n(a)))$ est strictement concave. En déduire que la log-vraisemblance atteint son unique maximum en $(\tilde{\alpha}_n, \tilde{\beta}_n)$.
7. Montrer que h_n , où l'on remplace x_i par X_i , converge p.s. vers une fonction h que l'on déterminera. Vérifier que $h(\alpha) = 1/\alpha$.

Ce dernier résultat permet de montrer la convergence p.s. de l'estimateur $(\tilde{\alpha}_n, \tilde{\beta}_n)$ vers (α, β) . Par ailleurs les estimateurs sont fortement biaisés.

△

XII.8 Sondages (II)

Exercice XII.8 (Sondage et stratification).

On considère un sondage pour le deuxième tour de l'élection présidentielle française effectué sur n personnes parmi la population des N électeurs, dont $N_A \geq 2$ (resp. $N_B \geq 2$) votent pour le candidat A (resp. B), avec $N = N_A + N_B$. Le but du sondage est d'estimer la proportion, $p = N_A/N$, de personnes qui votent pour A .

Comme N (environ 44 Millions d'inscrits pour l'élection présidentielle française de 2007) est grand devant n (de l'ordre de 1000), on considère uniquement des sondages avec remise (en particulier une personne peut être interrogée plusieurs fois). Le but de cet exercice est l'étude des méthodes de stratification.

I Modèle de référence

La réponse de la k -ième personne interrogée est modélisée par une variable aléatoire $Z_k : Z_k = 1$ (resp. $Z_k = 0$) si la personne interrogée vote pour le candidat A (resp. B). Les variables aléatoires $(Z_k, k \geq 1)$ sont indépendantes de loi de Bernoulli de paramètre p . On estime p à l'aide de la moyenne empirique, $\bar{Z}_n = \frac{1}{n} \sum_{k=1}^n Z_k$. Rappelons que \bar{Z}_n est un estimateur sans biais de p convergent et asymptotiquement normal.

1. Calculer le risque quadratique $R(\bar{Z}_n, p) = \mathbb{E}_p[(\bar{Z}_n - p)^2]$ pour l'estimation de p par \bar{Z}_n .
2. Vérifier que la vraisemblance du modèle est $p_n(z; p) = p^{n\bar{z}_n} (1-p)^{n-n\bar{z}_n}$, avec $z = (z_1, \dots, z_n) \in \{0, 1\}^n$ et $\bar{z}_n = \frac{1}{n} \sum_{k=1}^n z_k$. En déduire que \bar{Z}_n est l'estimateur du maximum de vraisemblance de p .
3. Calculer l'information de Fisher du modèle. Montrer que \bar{Z}_n est un estimateur efficace de p (dans la classe des estimateurs sans biais de p).

II Méthode de stratification

L'objectif des méthodes de stratification est d'améliorer la précision de l'estimation de p , tout en conservant le même nombre, n , de personnes interrogées. Pour cela on considère H strates (ou groupes) fixées. La strate h comporte $N_h \geq 1$ individus (N_h est connu), et on note p_h la proportion **inconnue** de personnes de cette strate qui votent pour A , n_h le nombre de personnes interrogées dans cette strate. Le nombre total de personnes interrogées est $n = \sum_{h=1}^H n_h$. On suppose que $p_h \in]0, 1[$ pour tout $1 \leq h \leq H$. On considère la proportion des personnes interrogées dans la strate h qui votent pour A :

$$Y_h = \frac{1}{n_h} \sum_{i=1}^{n_h} X_i^{(h)},$$

où la variable aléatoire $X_i^{(h)}$ modélise la réponse de la i -ème personne interrogée dans la strate h : $X_i^{(h)} = 1$ (resp. $X_i^{(h)} = 0$) si la personne interrogée vote pour le candidat A (resp. B). La variable aléatoire $X_i^{(h)}$ suit une loi de Bernoulli de paramètre p_h . On suppose que les variables aléatoires $(X_i^{(h)}, 1 \leq i, 1 \leq h \leq H)$ sont indépendantes.

L'estimateur de Horvitz-Thompson¹⁵ (initialement construit pour des sondages sans remise) de p est défini par :

$$Y = \sum_{h=1}^H \frac{N_h}{N} Y_h.$$

II.1 Étude à horizon fini

1. On choisit un individu au hasard. Quelle est la probabilité qu'il soit dans la strate h ? En déduire que $\sum_{h=1}^H \frac{N_h}{N} p_h = p$.
2. Vérifier que l'estimateur de Horvitz-Thompson est sans biais : $\mathbb{E}_p[Y] = p$.
3. Calculer le risque quadratique de Y : $R(Y, p) = \mathbb{E}_p[(Y - p)^2]$ en fonction des variances $\sigma_h^2 = p_h(1 - p_h)$.
4. Pourquoi dit-on que la stratification est mauvaise si $n \text{Var}_p(Y) > p(1 - p)$?
5. Donner un exemple de mauvaise stratification.

15. D. G. Horvitz and D.J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, vol. 47, pp.663-685 (1952).

On répondra aux trois questions suivantes en admettant que n_h peut prendre toutes les valeurs réelles positives et pas seulement des valeurs entières. On rappelle l'inégalité de Cauchy-Schwarz. Soit $a_i, b_i \in \mathbb{R}$ pour $i \in I$ et I fini, on a :

$$\left(\sum_{i \in I} a_i b_i \right)^2 \leq \left(\sum_{i \in I} a_i^2 \right) \left(\sum_{i \in I} b_i^2 \right),$$

avec égalité si et seulement si il existe $(\alpha, \beta) \in \mathbb{R}^2 \setminus \{(0, 0)\}$ tel que $\alpha a_i = \beta b_i$ pour tout $i \in I$.

6. Montrer que pour l'allocation proportionnelle, $n_h = n \frac{N_h}{N}$, on a $n \text{Var}_p(Y) \leq p(1-p)$. Donner une condition nécessaire et suffisante pour que $n \text{Var}_p(Y) < p(1-p)$.
7. Montrer que l'allocation optimale (*i.e.* celle pour laquelle $\text{Var}_p(Y)$ est minimal) correspond à l'allocation de Neyman où n_h est proportionnel à l'écart-type de la strate h : $N_h \sigma_h$. Donner une condition nécessaire et suffisante pour que l'allocation de Neyman soit strictement meilleure que l'allocation proportionnelle.
8. Dédire de ce qui précède que sous des conditions assez générales, on peut construire un estimateur de p sans biais qui est strictement préférable à l'estimateur efficace (dans la classe des estimateurs sans biais) \bar{Z}_n . En quoi cela n'est-il pas contradictoire ?

II.2 Étude asymptotique

1. Montrer que $(Y_h, n_h \geq 1)$ converge p.s. vers p_h et que $(\sqrt{n_h}(Y_h - p_h), n_h \geq 1)$ converge en loi vers une loi gaussienne dont on précisera les paramètres.
2. Montrer que p.s. Y converge vers p quand $\min_{1 \leq h \leq H} n_h$ tend vers l'infini.

On rappelle le résultat suivant sur la convergence uniforme locale des fonctions caractéristiques. Soit $(V_k, k \geq 1)$ une suite de variables aléatoires indépendantes, de même loi et de carré intégrable, avec $\mathbb{E}[V_k] = \mu$ et $\text{Var}(V_k) = \sigma^2$. Alors, on a :

$$\psi_{\sqrt{k}(\bar{V}_k - \mu)}(u) = e^{-\sigma^2 u^2 / 2} + R_k(u), \quad \text{où } \bar{V}_k = \frac{1}{k} \sum_{i=1}^k V_i,$$

et pour tout $K \geq 0$, $\lim_{k \rightarrow \infty} \sup_{|u| \leq K} |R_k(u)| = 0$.

3. Montrer que, si $\min_{1 \leq h \leq H} n_h$ tend vers l'infini, alors $(Y - p) / \sqrt{\text{Var}_p(Y)}$ converge en loi vers une loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.

4. Montrer que :

$$\frac{1}{\text{Var}_p(Y)} \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{Y_h(1 - Y_h)}{n_h}$$

converge p.s. vers 1, quand $\min_{1 \leq h \leq H} n_h$ tend vers l'infini. On pourra, après l'avoir justifié, utiliser que pour tout $\varepsilon > 0$ on a $|Y_h(1 - Y_h) - \sigma_h^2| \leq \varepsilon \sigma_h^2$ pour $\min_{1 \leq h \leq H} n_h$ suffisamment grand.

5. Dédurre de la question précédente que :

$$I = \left[Y \pm \phi_{1-\alpha/2} \sqrt{\sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{Y_h(1 - Y_h)}{n_h}} \right],$$

où ϕ_r est le quantile d'ordre r de la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$, est un intervalle de confiance sur p de niveau asymptotique $1 - \alpha$ ($\alpha \in]0, 1[$) quand $\min_{1 \leq h \leq H} n_h$ tend vers l'infini.

△

XII.9 Loi de Yule (II)

Exercice XII.9 (Disques d'or et loi de Yule).

On suppose que les ventes des disques d'un artiste dépendent de sa réputation seulement (*i.e.* du nombre de disques déjà vendus) et donc pas de son talent¹⁶. Dans ce cas, il est naturel de modéliser le nombre de ventes pour un artiste par la loi de Yule¹⁷, voir l'exercice (XI.11). On mesure la notoriété d'un artiste par le nombre de disques d'or qu'il a obtenus. Le tableau XII.12 donne le nombre d'artistes ayant eu un nombre de disques d'or fixé pendant la période de 1959 à 1989. Pour information, le tableau XII.13 indique les artistes ayant eu au moins 14 disques d'or. On compte $n = 1377$ artistes ayant eu un ou des disques d'or.

Si on estime que la probabilité d'acheter un disque d'un artiste inconnu est très faible, alors on peut modéliser le nombre de disques d'or d'un artiste par une variable aléatoire Y de loi de Yule de paramètre $\rho > 0$. On rappelle que si Y suit une loi de Yule de paramètre $\rho \in]0, \infty[$, alors on a :

$$\mathbb{P}_\rho(Y = k) = \rho B(k, \rho + 1), \quad k \in \mathbb{N}^*,$$

16. K. H. Chung and R. A. K. Cox. A Stochastic Model of Superstardom : an application of the Yule distribution. *The Review of Economics and Statistics*, vol. 76, pp. 771–775 (1994).

17. L. Spierdijk and M. Voorneveld. Superstars without talent? The Yule distribution controversy. *SSE/EFI Working Paper Series in Economics and Finance*. <http://swopec.hhs.se/hastef/abs/hastef0658.htm> (2007).

où pour $a, b \in]0, +\infty[$:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_{]0,1[} x^{a-1}(1-x)^{b-1} dx.$$

1. Déterminer la loi de Yule de paramètre $\rho = 1$. Comparer les données observées du tableau XII.12 avec les données théoriques attendues si le nombre de disques d'or pour un artiste pris au hasard suit la loi de Yule de paramètre $\rho = 1$.
2. Après avoir précisé le modèle, indiquer en détail une méthode pour tester si les données observées correspondent à la réalisation d'un échantillon de taille n de loi $\min(Y, m)$, où $m = 17$ et Y suit une loi de Yule de paramètre $\rho = 1$.
3. Justifier le fait que, pour répondre à la question 2, on ait regroupé dans le tableau XII.12 les artistes ayant eu plus de 17 disques d'or.

On suppose que les données observées correspondent à la réalisation d'un échantillon de taille n de loi de Yule de paramètre ρ inconnu. On note N_k le nombre d'artistes ayant eu au moins k disques d'or ; ainsi N_1 correspond au nombre d'artistes ayant eu au moins un disque d'or.

4. Exprimer la log-vraisemblance du modèle en fonction de $(N_k, k \in \mathbb{N}^*)$.
5. Trouver la plus grande constante $c > 0$ telle que $x^2 \geq (x+c)(x+c-1)$ pour tout $x > 1$. En déduire que :

$$\sum_{k=1}^{\infty} \frac{1}{(\rho+k)^2} \leq \frac{1}{\rho + \frac{1}{2}}.$$

Montrer que la log-vraisemblance de la question précédente possède un seul maximum local sur $]0, (1 + \sqrt{3})/2]$ et, si $N_2 \geq 1$, alors elle possède au moins un maximum sur $]0, \infty[$.

Pour les données du tableau XII.12, on obtient la valeur approchée suivante pour l'estimateur du maximum de vraisemblance de ρ : 1.137.

6. Après avoir précisé le modèle, indiquer en détail une méthode pour vérifier si les données observées correspondent à la réalisation d'un échantillon de taille n de loi $\min(Y, m)$, où $m = 17$ et Y suit une loi de Yule.
7. Les p -valeurs pour les questions 2 et 6 sont plus petites que 10^{-5} . Quelle est votre conclusion concernant le talent des artistes ?
8. Les auteurs de l'étude proposent de ne pas tenir compte des artistes ayant plus de 18 disques d'or et de considérer la loi de Yule conditionnée à prendre des valeurs inférieures à 18. Il obtiennent alors une p -valeur de l'ordre de 17%. Quelle est leur conclusion, concernant le talent des artistes ?

△

| Nombre de disques d'or | Nombre d'artistes | Nombre de disques d'or | Nombre d'artistes |
|------------------------|-------------------|------------------------|-------------------|
| 1 | 668 | 10 | 14 |
| 2 | 244 | 11 | 16 |
| 3 | 119 | 12 | 13 |
| 4 | 78 | 13 | 11 |
| 5 | 55 | 14 | 5 |
| 6 | 40 | 15 | 4 |
| 7 | 24 | 16 | 4 |
| 8 | 32 | 17 et + | 26 |
| 9 | 24 | | |

Table XII.12. Nombres d'artistes par nombre de disques d'or.

| Artiste | Nombre de disques d'or | Artiste | Nombre de disques d'or |
|-----------------------------|------------------------|--------------------|------------------------|
| Beatles | 46 | John Denver | 18 |
| Elvis Presley | 45 | Kiss | 18 |
| Elton John | 37 | Kool and the Gang | 18 |
| Rolling Stones | 36 | Rod Stewart | 18 |
| Barbra Streisand | 34 | Andy Williams | 17 |
| Neil Diamond | 29 | Frank Sinatra | 17 |
| Aretha Franklin | 24 | Billy Joel | 16 |
| Chicago | 23 | Hank Williams, Jr. | 16 |
| Donna Summer | 22 | Linda Ronstadt | 16 |
| Kenny Rogers | 22 | Willie Nelson | 16 |
| Olivia Newton-John | 22 | Doors | 15 |
| Beach Boys | 21 | Glen Campbell | 15 |
| Bob Dylan | 20 | Queen | 15 |
| Earth, Wind and Fire | 20 | REO Speedwagon | 15 |
| Hall and Oates | 20 | Anne Murray | 14 |
| Barry Manilow | 19 | Doobie Brothers | 14 |
| Three Dog Night | 19 | Jethro Tull | 14 |
| Bee Gees | 18 | Johnny Mathis | 14 |
| Carpenters | 18 | O'Jays | 14 |
| Credence Clearwater Revival | 18 | | |

Table XII.13. Artistes ayant eu au moins 14 disques d'or entre 1958 et 1989.

XII.10 Sexe des anges

Exercice XII.10 (Sexe des anges).

Soit p_0 la probabilité (inconnue) d'avoir un garçon à la naissance (p_0 est de l'ordre de 0.51 environ). On remarque, à l'aide d'un test du χ^2 sur les données¹⁸ du tableau XII.14 (p -valeurs asymptotiques de l'ordre de 5%), que le nombre de garçons

18. M.-P. Schützenberger. Résultats d'une enquête sur la distribution du sexe dans les familles nombreuses. *Semaine des Hôpitaux de Paris*, vol. 25(61), pp. 2579–2582 (1949).

d'une famille à $n \geq 2$ enfants ne suit pas une loi binomiale de paramètre (n, p_0) (voir l'exercice IX.18). Une des raisons possibles est que la probabilité d'avoir un garçon dépende de la famille considérée, et qu'elle soit donc aléatoire. Le modèle que nous étudions est un modèle classique de la statistique bayésienne¹⁹.

On note P la probabilité (aléatoire) d'avoir un garçon pour une famille choisie au hasard. On modélise P par une variable aléatoire de loi béta de paramètre $\theta = (a, b) \in]0, \infty[^2$, dont la densité est :

$$\frac{1}{B(a, b)} x^{a-1} (1-x)^{b-1} \mathbf{1}_{]0,1[}(x).$$

On note X_n le nombre de garçons parmi les n premiers enfants d'une famille choisie au hasard (parmi les familles ayant au moins n enfants).

I Préliminaires

1. Calculer $\mathbb{E}_\theta[P]$ et $\mathbb{E}_\theta[P^v(1-P)^w]$ pour $v, w \in [0, \infty[$.
2. Justifier l'égalité $p_0 = a/(a+b)$.
3. Quelle est la loi conditionnelle de X_n sachant P ? Déterminer la fonction ψ telle que $\mathbb{E}[g(X_n) | P] = \psi(P)$ en fonction de g .
4. Montrer que $\mathbb{E}_\theta[X_n] = n\mathbb{E}_\theta[P]$ et $\text{Var}_\theta(X_n) = n\mathbb{E}_\theta[P(1-P)] + n^2 \text{Var}_\theta(P)$.
5. En déduire que :

$$\mathbb{E}_\theta[X_n] = n \frac{a}{a+b} \quad \text{et} \quad \text{Var}_\theta(X_n) = \frac{1}{n} \mathbb{E}_\theta[X_n] (n - \mathbb{E}_\theta[X_n]) \left(1 + \frac{n-1}{a+b+1} \right). \tag{XII.8}$$

II Estimation des paramètres

On considère les familles à n enfants. On note $X_n^{(k)}$ le nombre de garçons de la k -ème famille à n enfants. On suppose que les variables aléatoires $(X_n^{(k)}, k \geq 1)$ sont indépendantes et de même loi que X_n .

1. Montrer que si $n = 1$, alors la loi de X_n ne dépend que de p_0 . En déduire que le modèle n'est pas identifiable.

On suppose $n \geq 2$.

2. Construire, en utilisant (XII.8), un estimateur $\hat{\theta}_K$ de θ à partir de :

$$M_K = \frac{1}{K} \sum_{k=1}^K X_n^{(k)} \quad \text{et} \quad V_K = \frac{1}{K} \sum_{k=1}^K \left(X_n^{(k)} \right)^2 - M_K^2.$$

¹⁹. C. Robert. *Le choix bayésien : Principes et pratique*. Springer (2006).

3. Montrer que $(\hat{\theta}_K, K \geq 1)$ est un estimateur convergent de θ .
4. Montrer, sans calculer explicitement la matrice de covariance asymptotique, que l'estimateur $(\hat{\theta}_K, K \geq 1)$ de θ est asymptotiquement normal.
5. On pose $N_r = \sum_{k=1}^K \mathbf{1}_{\{X_n^{(k)}=r\}}$. Vérifier que :

$$\sum_{k=1}^K X_n^{(k)} = \sum_{r=0}^n r N_r \quad \text{et} \quad \sum_{k=1}^K \left(X_n^{(k)}\right)^2 = \sum_{r=0}^n r^2 N_r.$$

Donner une valeur numérique pour l'estimation de θ à partir du tableau XII.14. Pour des raisons de temps de calcul, on ne donnera pas d'intervalle de confiance asymptotique pour θ .

| | | | | | | | |
|---------------------------------|-------|-------|-------|-------|-------|-------|-------|
| Nombres de garçons et de filles | (5,0) | (4,1) | (3,2) | (2,3) | (1,4) | (0,5) | Total |
| Nombre de familles | 204 | 841 | 1585 | 1544 | 810 | 144 | 5128 |

Table XII.14. Nombres de garçons et de filles dans 5128 familles de cinq enfants.

III Estimation de la probabilité d'avoir un garçon

On note $P_n = \mathbb{E}_\theta[P \mid X_n]$.

1. Soit h une fonction bornée mesurable et $k \in \{0, \dots, n\}$. Calculer $\mathbb{E}_\theta[h(P) \mid X_n = k]$ défini par :

$$\mathbb{E}_\theta[h(P) \mid X_n = k] = \frac{\mathbb{E}_\theta[h(P)\mathbf{1}_{\{X_n=k\}}]}{\mathbb{P}_\theta(X_n = k)}.$$

2. Calculer et reconnaître la loi conditionnelle de P sachant X_n . Vérifier que :

$$P_n = \frac{a + X_n}{a + b + n}.$$

3. Montrer que la suite $(X_n/n, n \in \mathbb{N})$ converge p.s. vers P .
4. En déduire que P_n est un estimateur convergent de P : p.s. $\lim_{n \rightarrow \infty} P_n = P$.
5. Montrer que P_n est le meilleur estimateur de P au sens quadratique : pour toute fonction h mesurable bornée, on a $\mathbb{E}_\theta[(P - h(X_n))^2] \geq \mathbb{E}_\theta[(P - P_n)^2]$. (On pourra faire intervenir la quantité $P_n - h(X_n)$ pour réécrire le membre de droite de l'inégalité ci-dessus.)
6. Déduire de la question III.2 un intervalle de confiance de niveau exact 90% de la forme $[c, 1]$ sur P en fonction de X_n .

On considère n^* la valeur minimale de n , telle que pour une famille avec n enfants qui sont tous des garçons, l'intervalle de confiance à 90% de P est $[c, 1]$ où $c > 1/2$. En particulier, on peut affirmer que dans cette famille le $n^* + 1$ -ième enfant (à venir) à plus d'une chance sur deux d'être un garçon (avec une confiance de 90%).

Pour les valeurs numériques de θ de la question II.5, on trouve $n^* = 12$ ($c \simeq 0.501$ et $P_n \simeq 0.56$). Ce résultat est sensible aux erreurs d'estimation de θ . Pour $\theta = (1, 0.97)$ et donc $p_0 \simeq 0.508$, on obtient $n^* = 3$ ($c \simeq 0.57$ et $P_n \simeq 0.80$).

△

XII.11 Comparaison d'échantillons appariés

Exercice XII.11 (Test des signes et test de Wilcoxon).

On considère le résultat d'une intervention chirurgicale sur des patients atteints de troubles de la rétine²⁰. Les données consistent en des mesures de la fonction rétinienne à l'aide d'un électro-rétinogramme 3 mois avant l'opération et 3 à 5 mois après l'opération pour $n = 10$ patients. Pour le patient k , on modélise par X_k le résultat de la mesure avant l'opération et par Y_k celui de la mesure après l'opération. Les deux variables aléatoires X_k et Y_k sont appariées et ne peuvent être traitées séparément. Le tableau XII.15 donne les mesures effectuées sur 10 patients (oeil gauche).

| X_k | Y_k | $V_k = X_k - Y_k$ | $S_k = \text{sgn}(V_k)$ | $R_n(k)$ |
|-------|-------|-------------------|-------------------------|----------|
| 8.12 | 6.50 | 1.62 | 1 | 10 |
| 2.45 | 2.10 | 0.35 | 1 | 8 |
| 1.05 | 0.84 | 0.21 | 1 | 6 |
| 6.86 | 5.32 | 1.54 | 1 | 9 |
| 0.14 | 0.11 | 0.03 | 1 | 3 |
| 0.11 | 0.17 | -0.06 | -1 | 4 |
| 0.19 | 0.16 | 0.03 | 1 | 2 |
| 0.23 | 0.16 | 0.07 | 1 | 5 |
| 1.89 | 1.54 | 0.35 | 1 | 7 |
| 0.07 | 0.05 | 0.02 | 1 | 1 |

Table XII.15. Mesures en micro Volt avant (X_k) et après (Y_k) l'opération de l'activité rétinienne de l'oeil gauche par électro-rétinogramme pour $n = 10$ patients. La variable S_k représente le signe de $V_k = X_k - Y_k$ et $R_n(k)$ le rang de $|V_k|$ (voir la définition dans la partie III).

On considère le modèle suivant. Soit X une variable aléatoire réelle de loi inconnue. On suppose que $((X_k, Y_k), k \in \mathbb{N}^*)$ est une suite de vecteurs aléatoires

²⁰ B. Rosner, R. J. Glynn and M.-L. T. Lee. The Wilcoxon signed rank test for paires comparisons of clustered data. *Biometrics*, vol. 62, pp. 185-192 (2006).

indépendants de même loi, que X_k et Y_k sont indépendantes, que X_k a même loi que X et que Y_k a même loi que $X - \mu$ où $\mu \in \mathbb{R}$. Le paramètre de décalage μ est a priori inconnu. On désire construire un test pour les hypothèses suivantes : $H_0 = \{\mu = 0\}$ (l'opération ne diminue pas la mesure de l'électro-rétinogramme) et $H_1 = \{\mu > 0\}$ (l'opération diminue la mesure de l'électro-rétinogramme). On peut construire un test à partir de la différence des moyennes empiriques avant l'opération et après l'opération, mais le peu de données et l'absence d'information sur la loi de X rendent cette approche peu pertinente. On a alors recours soit au test des signes soit au test de Wilcoxon²¹. Ces deux tests ne dépendent pas de la loi de X , on parle de tests non-paramétriques. Le test de Wilcoxon est en général préféré au test des signes, car plus puissant. Nous étudions dans les parties II et III les comportements asymptotiques de ces deux tests. En pratique, les tailles des échantillons étant faibles, on utilise la loi exacte des statistiques de test obtenue soit par calcul direct soit par simulation.

Les résultats de la partie préliminaire I, peuvent être directement utilisés dans les parties II et III. Les parties II et III sont indépendantes. Dans les parties II et III, **on suppose que la fonction de répartition de X est continue.**

Pour $v \in \mathbb{R}$, on note $\text{sgn}(v) = \mathbf{1}_{\{v>0\}} - \mathbf{1}_{\{v<0\}}$ le signe de v . On pose $V_k = X_k - Y_k$ et $S_k = \text{sgn}(V_k)$.

I Préliminaires

Soit X' une variable aléatoire indépendante de X et de même loi que X . On pose $V = X - X'$. On rappelle que la médiane de V est son quantile d'ordre $1/2$ défini par $\inf\{v \in \mathbb{R}; \mathbb{P}(V \leq v) \geq 1/2\}$.

1. Montrer que V et $-V$ ont même loi et que, pour tout $x \in \mathbb{R}$:

$$\mathbb{P}(V \leq x) = 1 - \mathbb{P}(V < -x). \tag{XII.9}$$

2. Dédire de (XII.9) que $2\mathbb{P}(V \leq 0) = 1 + \mathbb{P}(V = 0)$ et que, pour tout $\varepsilon > 0$, $2\mathbb{P}(V \leq -\varepsilon) = 1 - \mathbb{P}(V \in]-\varepsilon, \varepsilon])$.
3. Montrer que pour tout $\varepsilon > 0$, il existe $\delta > 0$ tel que pour tout $a \in \mathbb{R}$, $\mathbb{P}(V \in]-\varepsilon, \varepsilon]) \geq \mathbb{P}(X \in]a - \delta, a + \delta[, X' \in]a - \delta, a + \delta])$. En déduire que pour tout $\varepsilon > 0$:

$$\mathbb{P}(V \in]-\varepsilon, \varepsilon]) > 0. \tag{XII.10}$$

4. Dédire des questions précédentes que la médiane de V est nulle.

21. F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, vol. 1(6), pp. 80-83 (1945).

On admet que si ϕ est une fonction bornée mesurable définie sur \mathbb{R}^2 alors, si Z et Z' sont deux variables aléatoires réelles indépendantes, on a :

$$\mathbb{E}[\phi(Z, Z')] = \mathbb{E}[\varphi(Z)], \quad \text{avec, pour } z \in \mathbb{R}, \quad \varphi(z) = \mathbb{E}[\phi(z, Z')]. \quad (\text{XII.11})$$

On suppose que **la fonction de répartition de X est continue** : $\forall x \in \mathbb{R}, \mathbb{P}(X = x) = 0$.

5. Montrer en utilisant (XII.11) que la fonction de répartition de V est continue.
6. Montrer que $\text{sgn}(V)$ est de loi uniforme sur $\{-1, 1\}$.
7. En étudiant $\mathbb{E}[g(\text{sgn}(V))f(|V|)]$, pour des fonctions f et g mesurables bornées, montrer que les variables aléatoires $\text{sgn}(V)$ et $|V|$ sont indépendantes.

II Test des signes

On rappelle que $S_k = \text{sgn}(V_k)$. On considère la statistique de test sur l'échantillon de taille n définie par :

$$\zeta_n = \frac{1}{\sqrt{n}} \sum_{k=1}^n S_k.$$

1. Dédire de la question I.6 que, sous H_0 , la suite $(\zeta_n, n \in \mathbb{N}^*)$ converge en loi vers une variable aléatoire gaussienne de loi $\mathcal{N}(0, 1)$.
2. Dédire de (XII.10), que sous H_1 , $\mathbb{E}[S_k] > 0$. En déduire que, sous H_1 , la suite $(\zeta_n, n \in \mathbb{N}^*)$ converge p.s. vers $+\infty$.
3. Déterminer la région critique W_n du test pur, construit à partir de la statistique de test ζ_n , de niveau asymptotique $\alpha \in]0, 1[$.
4. Vérifier que le test pur de région critique W_n est convergent.
5. Calculer, à partir des observations, la p -valeur asymptotique du test pur de région critique W_{10} . Rejetez vous H_0 ?
6. Calculer, à partir des observations, la p -valeur exacte du test pur de région critique W_{10} . Commenter alors le résultat de la question II.5.

III Test de Wilcoxon

On note \mathcal{S}_n l'ensemble des permutations de $\{1, \dots, n\}$. On rappelle :

$$\sum_{k=1}^n k = \frac{n(n+1)}{2} \quad \text{et} \quad \sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6}.$$

1. Montrer en utilisant la question I.5 que pour $k \neq \ell$, $\mathbb{P}(|V_k| = |V_\ell|) = 0$. En déduire qu'il existe une unique permutation aléatoire $\tau_n \in \mathcal{S}_n$ telle que p.s. :

$$|V_{\tau_n(1)}| < \cdots < |V_{\tau_n(n)}|. \quad (\text{XII.12})$$

Pour les questions III.2 à III.7, on se place sous H_0 .

2. Vérifier en utilisant la question I.7 que τ_n et (S_1, \dots, S_n) sont indépendants.

3. Montrer que $(S_{\tau_n(1)}, \dots, S_{\tau_n(n)})$ a même loi que (S_1, \dots, S_n) .

On note R_n l'inverse de τ_n : $R_n(i) = j \Leftrightarrow \tau_n(j) = i$. On remarque que $R_n(k)$ est le rang de $|V_k|$ parmi $(|V_1|, \dots, |V_n|)$. En particulier $R_n(k) = 1$ si $|V_k|$ est le minimum et $R_n(k) = n$ si $|V_k|$ est le maximum. On définit :

$$T_n = \sum_{k=1}^n R_n(k) \mathbf{1}_{\{S_k > 0\}}$$

et la statistique de test de Wilcoxon :

$$\xi_n = \frac{T_n - \frac{n(n+1)}{4}}{\sigma_n}, \quad \text{avec } \sigma_n \geq 0 \text{ et } \sigma_n^2 = \frac{n(n+1)(2n+1)}{24}.$$

4. Montrer, à l'aide de la question III.3 et I.6 que T_n a même loi que :

$$T'_n = \sum_{k=1}^n k Z_k,$$

où les variables aléatoires $(Z_n, n \in \mathbb{N}^*)$ sont indépendantes de loi de Bernoulli de paramètre $1/2$.

5. Calculer $\mathbb{E}[T'_n]$ et $\text{Var}(T'_n)$.
6. Montrer que la fonction caractéristique ψ_n de ξ_n est :

$$\psi_n(u) = \prod_{k=1}^n \mathbb{E} \left[e^{iuk(Z_k - \frac{1}{2})/\sigma_n} \right] = \exp \left(\sum_{k=1}^n \log \left(\cos \left(\frac{uk}{2\sigma_n} \right) \right) \right).$$

7. Montrer que, sous H_0 , la suite $(\xi_n, n \in \mathbb{N}^*)$ converge en loi vers une variable aléatoire gaussienne de loi $\mathcal{N}(0, 1)$.

On admet que sous H_1 la suite $(\xi_n, n \in \mathbb{N}^*)$ converge en probabilité²² vers $+\infty$: pour tout $a \in \mathbb{R}$, $\lim_{n \rightarrow \infty} \mathbb{P}(\xi_n \geq a) = 1$.

8. Déterminer la région critique W'_n du test pur, construit à partir de la statistique de test ξ_n , de niveau asymptotique α . Vérifier que le test est convergent.
9. Calculer la p -valeur asymptotique du test pur de région critique W'_n . Rejetez-vous H_0 ? Comparer avec la question II.5.

△

²² P. Capéraà et B. Van Custen. *Méthodes et modèles en statistique non paramétrique*. Dunod (1988).

XII.12 Modèle auto-régressif pour la température

Exercice XII.12 (Modèle de température).

On désire étudier un modèle de série temporelle^{23 24} pour les températures journalières où la température du jour dépend de manière linéaire de la température de la veille. Plus précisément, on modélise par X_n la température du jour $n \in \mathbb{N}$ à la station de mesure et on suppose que pour $n \in \mathbb{N}$:

$$X_{n+1} = \alpha X_n + \beta + \varepsilon_{n+1}, \quad (\text{XII.13})$$

où $\alpha, \beta \in \mathbb{R}$, X_0 est une constante connue et $(\varepsilon_n, n \in \mathbb{N}^*)$ est une suite de variables aléatoires indépendantes de même loi gaussienne $\mathcal{N}(0, \sigma^2)$ avec $\sigma > 0$. On suppose les paramètres α, β et σ inconnus. Le modèle considéré est appelé modèle auto-régressif avec bruits gaussiens.

On note G une variable aléatoire gaussienne de loi $\mathcal{N}(0, 1)$.

I Préliminaires

1. Montrer que pour tout $v < 1$, on a :

$$\mathbb{E} \left[\exp \left(\frac{v}{2} G^2 \right) \right] = \frac{1}{\sqrt{1-v}}.$$

On pose $\varepsilon_0 = 0$ et pour $n \in \mathbb{N}^*$:

$$V_n = \sum_{k=1}^n \varepsilon_k^2 \quad \text{et} \quad T_n = \sum_{k=1}^n \varepsilon_{k-1} \varepsilon_k.$$

On souhaite étudier la convergence de la suite $(T_n, n \in \mathbb{N}^*)$ convenablement renormalisée.

Pour $x \geq 0$, on note $[x]$ l'entier m tel que $m \leq x < m+1$ et $\lceil x \rceil$ l'entier m' tel que $m' - 1 < x \leq m'$.

2. En écrivant $T_n = \left(\sum_{k=1}^{\lceil n/2 \rceil} \varepsilon_{2k-2} \varepsilon_{2k-1} \right) + \left(\sum_{k=1}^{\lfloor n/2 \rfloor} \varepsilon_{2k-1} \varepsilon_{2k} \right)$, montrer que la suite $(T_n/n, n \in \mathbb{N}^*)$ converge p.s. vers une limite que l'on précisera.

Soit $u \in \mathbb{R}$. On pose :

$$N_n(u) = \exp \left(\frac{u^2 \sigma^2 V_{n-1}}{2n} \right) \quad \text{et} \quad M_n(u) = N_n(u) \exp \left(iu \frac{T_n}{\sqrt{n}} \right).$$

On suppose que $u^2 < n/2\sigma^4$.

23. G. Box and G. Jenkins. *Time series analysis : Forecasting and control*. Holden-Day (1970).

24. P. Brockwell and R. Davis. *Time Series : Theory and Methods*. Springer-Verlag (1987).

3. Calculer $\mathbb{E}[N_n(u)]$ et $\mathbb{E}[N_n(u)^2]$, puis montrer que :

$$\lim_{n \rightarrow +\infty} \text{Var}(N_n(u)) = 0.$$

4. Pour $n \geq 2$, calculer $\mathbb{E}[M_n(u)|\varepsilon_1, \dots, \varepsilon_{n-1}]$.

5. Montrer, en utilisant la question précédente, que $\mathbb{E}[M_n(u)] = 1$.

On note ψ_n la fonction caractéristique de T_n/\sqrt{n} .

6. Montrer, en utilisant deux fois l'inégalité de Jensen, que :

$$\left| \psi_n(u)\mathbb{E}[N_n(u)] - \mathbb{E}[M_n(u)] \right| \leq \sqrt{\text{Var}(N_n(u))}.$$

7. Dédurre de ce qui précède que la suite $(T_n/\sqrt{n}, n \in \mathbb{N}^*)$ converge en loi vers $\sigma^2 G$.

II Estimation des paramètres

1. Justifier que la vraisemblance du modèle associé aux variables aléatoires $(\varepsilon_1, \dots, \varepsilon_n)$ s'écrit :

$$p_n(\varepsilon_1, \dots, \varepsilon_n; \alpha, \beta, \sigma) = (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{k=1}^n (x_k - \alpha x_{k-1} - \beta)^2 / 2\sigma^2\right),$$

où (x_0, \dots, x_n) correspond à une réalisation de (X_0, \dots, X_n) .

On pose $\delta_n = \frac{1}{n}(X_n - X_0)$ et :

$$\bar{\varepsilon}_n = \frac{1}{n} \sum_{k=1}^n \varepsilon_k, \quad \bar{X}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} X_k, \quad \overline{X^2}_{n-1} = \frac{1}{n} \sum_{k=0}^{n-1} X_k^2, \quad \text{et} \quad \Delta_n = \frac{1}{n} \sum_{k=1}^n X_{k-1} X_k.$$

2. Montrer que les estimateurs du maximum de vraisemblance de α et β sont :

$$\hat{\alpha}_n = \frac{\Delta_n - (\bar{X}_{n-1} + \delta_n)\bar{X}_{n-1}}{\overline{X^2}_{n-1} - \bar{X}_{n-1}^2} \quad \text{et} \quad \hat{\beta}_n = \bar{X}_{n-1} + \delta_n - \hat{\alpha}_n \bar{X}_{n-1}.$$

3. Montrer en utilisant (XII.13) que :

$$\bar{X}_{n-1}(1 - \alpha) = \beta + \bar{\varepsilon}_n - \delta_n.$$

On suppose que $\alpha \in]-1, 1[$ et on admet alors que p.s. $\lim_{n \rightarrow +\infty} X_n/\sqrt{n} = 0$ et que p.s. $\lim_{n \rightarrow +\infty} I_n = 0$, où :

$$I_n = \frac{1}{n} \sum_{k=1}^n X_{k-1} \varepsilon_k.$$

4. Montrer que $(\bar{X}_n, n \in \mathbb{N}^*)$ converge p.s. vers $a = \beta/(1 - \alpha)$.

5. On rappelle la notation : $V_n = \sum_{k=1}^n \varepsilon_k^2$. Montrer que :

$$\overline{X^2}_{n-1}(1 - \alpha^2) = \beta^2 + \frac{V_n}{n} + 2\beta\alpha\bar{X}_{n-1} + \left(\frac{X_0^2}{n} - \frac{X_n^2}{n} \right) + 2\beta\bar{\varepsilon}_n + 2\alpha I_n.$$

6. Montrer que $(\overline{X^2}_n, n \in \mathbb{N}^*)$ converge p.s. vers une limite b que l'on calculera.

7. En s'inspirant des questions précédentes, montrer que $(\Delta_n, n \in \mathbb{N}^*)$ converge p.s. vers $\alpha b + \beta a$.

8. Montrer que les estimateurs $\hat{\alpha}_n$ et $\hat{\beta}_n$ sont convergents.

9. Calculer $\hat{\sigma}_n^2$ l'estimateur du maximum de vraisemblance de σ^2 .

10. Montrer que $\hat{\sigma}_n^2$ est un estimateur convergent de σ^2 .

On peut également démontrer que l'estimateur $(\hat{\alpha}_n, \hat{\beta}_n, \hat{\sigma}_n^2)$ de $(\alpha, \beta, \sigma^2)$ est asymptotiquement normal.

III Test d'utilité du coefficient d'auto-régression

On utilise les notations des paragraphes précédents. On considère l'hypothèse nulle $H_0 = \{\alpha = 0\}$ et son alternative $H_1 = \{\alpha \in]-1, 0[\cup]0, 1[\}$. On introduit la statistique de test :

$$\zeta_n = \sqrt{n}\hat{\alpha}_n.$$

1. Vérifier que sous H_0 , on a :

$$\zeta_n = \frac{T_n/\sqrt{n} + R'_n}{V_{n-1}/n + R_n},$$

où les suites $(R'_n, n \in \mathbb{N}^*)$ et $(R_n, n \in \mathbb{N}^*)$ convergent en loi vers 0.

2. Montrer en utilisant la question I.7 que la suite $(\zeta_n, n \in \mathbb{N}^*)$ converge en loi sous H_0 vers G de loi gaussienne $\mathcal{N}(0, 1)$.

3. Donner, en utilisant la question II.8, le comportement asymptotique de ζ_n sous H_1 . Puis déterminer la région critique pour un test de niveau asymptotique $\eta \in]0, 1[$.

4. Donner la formule de la p -valeur asymptotique.

5. On observe les températures journalières moyennes de la station Météo France de Lille du 31/12/2001 au 31/12/2002 ($n = 365$), et on obtient :

$$\bar{X}_{n-1} \simeq 11.349521, \quad \delta_n = 0.0102397, \quad \overline{X^2}_{n-1} \simeq 158.79738, \quad \Delta_n \simeq 156.86773.$$

Faire l'application numérique (avec par exemple $\eta = 5\%$) et conclure.

Les figures XII.4 et XII.3 donnent respectivement pour la période du 2/01/1994 au 31/12/2002 ($n = 32\ 285$) et du 01/01/2002 au 31/12/2002 ($n = 365$) :

- À gauche : les valeurs des températures journalières moyennes ($X_k, k \in \{1, \dots, n\}$) relevées à la station de Météo France de Lille, ainsi que les valeurs des résidus correspondants ($\hat{\varepsilon}_k, k \in \{1, \dots, n\}$), où $\hat{\varepsilon}_k = X_k - \hat{\alpha}_n X_{k-1} - \hat{\beta}_n$.
- À droite : l’histogramme des résidus et la densité de la loi gaussienne $\mathcal{N}(0, \hat{\sigma}_n^2)$.

Un modèle plus réaliste pour l’évolution journalière des températures devrait tenir compte de la saisonnalité (*cf* la périodicité du signal dans la figure XII.4 à gauche).

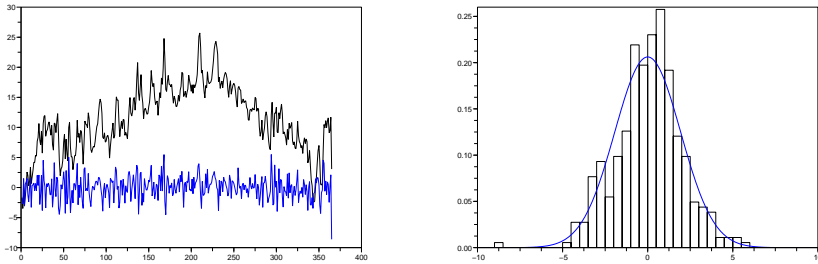


Figure XII.3. À gauche : températures journalières moyennes en 2002, et valeurs des résidus correspondants. À droite : histogramme des résidus et densité gaussienne associée.

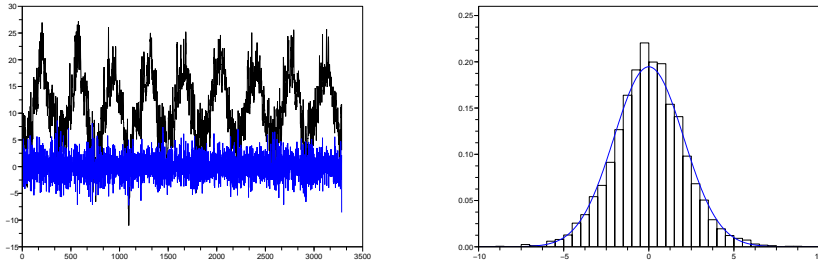


Figure XII.4. À gauche : températures journalières moyenne de 1994 à 2002, et valeurs des résidus correspondants. À droite : histogramme des résidus et densité gaussienne associée.

△

XII.13 Mutation de l'ADN mitochondrial

Exercice XII.13 (Mutation de l'ADN mitochondrial).

L'ADN mitochondrial (ADNmt) humain est une molécule circulaire double brin d'environ 16 500 paires de bases. L'ADNmt est dans la cellule mais hors du noyau contrairement à l'ADN chromosomique ; il est de plus transmis uniquement par la mère. Pour l'étude du taux de mutation de l'ADNmt, on s'intéresse donc uniquement à la population féminine. On suppose que la population féminine comporte à chaque génération N individus et que l'unité de temps est choisie égale à N générations. Si N est grand, alors l'arbre généalogique de l'ADNmt de n individus vivant à l'instant $t = 0$ peut être modélisé par l'approximation suivante due à Kingman²⁵ :

- L'arbre généalogique est binaire. Autrement dit, au plus 2 soeurs dans le passé ont des descendants aujourd'hui.
- La durée pendant laquelle les n individus considérés ont $k \in \{2, \dots, n\}$ ancêtres est modélisée par T_k , où **les variables aléatoires** $(T_k, k \geq 2)$ **sont indépendantes et T_k est de loi exponentielle de paramètre $k(k-1)/2$.**

La profondeur de l'arbre généalogique :

$$T_n^* = \sum_{k=2}^n T_k$$

représente, au signe près, la date de naissance du dernier ancêtre commun des n individus vivant à l'instant $t = 0$. La sous-population des n individus vivant à l'instant $t = 0$ possède, sur l'intervalle de temps $I_k =] - \sum_{r=k}^n T_r, - \sum_{r=k+1}^n T_r]$ de longueur T_k , exactement k ancêtres. (On utilise la convention $\sum_{r=n+1}^n T_r = 0$.) On définit également la longueur totale de l'arbre généalogique :

$$L_n = \sum_{k=2}^n k T_k.$$

On suppose que les mutations de l'ADN sont rares : la probabilité d'observer une mutation lors d'une reproduction est $\theta/2N$, avec $\theta > 0$. Pour une branche de l'arbre généalogique, on observe sur 1 unité de temps (soit N générations) en moyenne $\theta/2$ mutations. Le paramètre $\theta/2$ s'interprète comme un taux de mutation. On suppose de plus que les mutations sont indépendantes et affectent des nucléotides ou sites différents de l'ADN. Ceci revient à modéliser le nombre de mutations survenues durant l'intervalle de temps I_k et sur la branche $j \in \{1, \dots, k\}$ de l'arbre généalogique par une variable aléatoire $Y_{j,k}$ où **les variables aléatoires** $((T_k, Y_{1,k}, \dots, Y_{k,k}), k \geq 2)$ **sont indépendantes et, conditionnellement à T_k ,**

25. J. F. C. Kingman. The coalescent. *Stoch. Process. Appl.*, vol. 13(3), pp. 235–248 (1982).

les variables aléatoires $(Y_{1,k}, \dots, Y_{k,k})$ sont indépendantes de loi de Poisson de paramètre $\theta T_k/2$. Le nombre total de mutations observées est donc :

$$S_n = \sum_{k=2}^n Y_k, \quad \text{où} \quad Y_k = \sum_{j=1}^k Y_{j,k}$$

représente le nombre de mutations observées sur l'arbre généalogique durant l'intervalle de temps I_k . Une simulation de l'arbre généalogique et des mutations est représentée dans la figure XII.5.

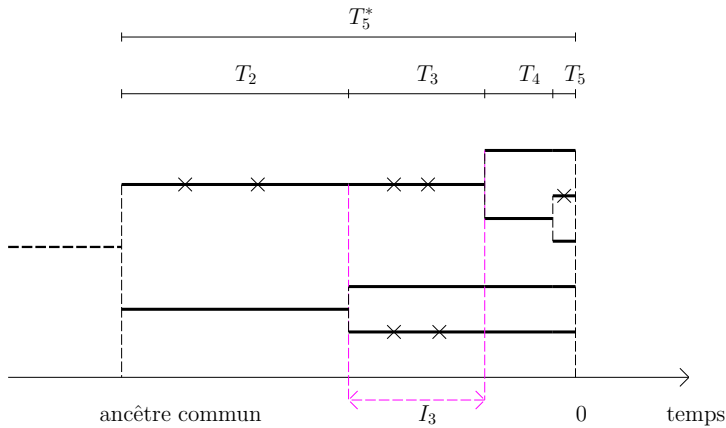


Figure XII.5. Simulation d'un arbre généalogique (trait plein) pour $n = 5$ et des mutations (représentées par des croix). On observe $Y_5 = 1$, $Y_4 = 0$, $Y_3 = 4$ (4 mutations dans l'intervalle de temps I_3), $Y_2 = 2$ et donc $S_n = 7$.

Le but du problème est d'étudier les propriétés de l'estimateur de Watterson²⁶ du paramètre θ :

$$\hat{\theta}_n = \frac{S_n}{h_{n-1}}, \quad \text{où} \quad h_{n-1} = \sum_{k=1}^{n-1} \frac{1}{k}.$$

On rappelle que $h_{n-1} = \log(n) + O(1)$.

²⁶ Y.-X. Fu and W.-H. Li. Maximum likelihood estimation of population parameters. *Genetics*, vol. 134(4), pp. 1261–1270 (1993).

I Préliminaires

1. Calculer $\lim_{n \rightarrow +\infty} \mathbb{E}[T_n^*]$. Pour la population humaine, on considère que $N \simeq 10\,000$ (et on néglige l'effet des millénaires les plus récents) et une génération est d'environ 20 ans. En déduire une estimation moyenne de la date de naissance du dernier ancêtre commun de toute la population humaine.
2. Calculer $\mathbb{E}[L_n]$ et $\text{Var}(L_n)$. En déduire que la suite $(L_n/h_{n-1}, n \geq 2)$ converge en probabilité vers 2.
3. Montrer que $\mathbb{E}_\theta[Y_k] = \theta/(k-1)$ pour $k \geq 2$.
4. Calculer la fonction caractéristique de $(Y_{j,k}, T_k)$.
5. Déduire de la question précédente la fonction caractéristique de (Y_k, T_k) :

$$\psi_{(Y_k, T_k)}(u, v) = \frac{(k-1)}{(k-1) + \theta(1 - e^{iu}) - 2i\frac{v}{k}}.$$

6. Montrer que $1 + Y_k$ est de loi géométrique et déterminer son paramètre.
7. Calculer $\mathbb{E}_\theta[S_n]$ et montrer que :

$$\text{Var}_\theta(S_n) = \theta h_{n-1} + \theta^2 \sum_{k=1}^{n-1} \frac{1}{k^2}.$$

II Propriétés de l'estimateur de Watterson

1. Déduire de la question I.7 que l'estimateur de Watterson est sans biais.
2. Déduire de la question I.7 que l'estimateur de Watterson converge en probabilité vers θ .
3. On rappelle que pour tout $M > 0$, il existe c tel que pour tout $z \in \mathbb{C}$ tel que $|z| \leq M$, on a $|e^z - 1 - z| \leq cz^2$ et $|e^z - 1 - z - z^2/2| \leq cz^3$. Montrer que :

$$\mathbb{E}_\theta \left[\exp \left(i \frac{u}{\sqrt{\theta h_{n-1}}} \left(Y_k - \frac{\theta}{k-1} \right) \right) \right] = \exp \left(-\frac{u^2}{2(k-1)h_{n-1}} \right) + r'_{k,n},$$

où $|r'_{k,n}| \leq c/[(k-1)h_{n-1}^{3/2} + (k-1)^2 h_{n-1}]$ et c est une constante qui dépend de u et θ mais pas de k ni de n .

On pose pour $n \geq 2$:

$$Z_n(\theta) = \frac{S_n - \theta h_{n-1}}{\sqrt{\theta h_{n-1}}}.$$

4. On rappelle que si $(a_k, k \in \mathbb{N}^*)$ et $(b_k, k \in \mathbb{N}^*)$ sont des suites de nombres complexes de modules inférieurs à 1 ($|a_k| \leq 1$ et $|b_k| \leq 1$ pour tout $k \in \mathbb{N}^*$), alors on a $|\prod_{k=1}^n a_k - \prod_{k=1}^n b_k| \leq \sum_{k=1}^n |a_k - b_k|$. Dédurre de la question II.3 que :

$$\mathbb{E}_\theta \left[e^{iuZ_n(\theta)} \right] = e^{-u^2/2} + O(\log(n)^{-1/2}).$$

5. Montrer, à l'aide de la question II.4, que la suite $((S_n - \theta h_{n-1})/\sqrt{S_n}, n \geq 2)$ converge en loi vers G , de loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.
6. Construire un intervalle de confiance pour θ de niveau asymptotique $1 - \alpha = 95\%$.
7. Donner une estimation et un intervalle de confiance pour le taux de mutation par nucléotide, θ/K , avec les données suivantes concernant le séquençage de $K = 360$ nucléotides de l'ADNmt²⁷ : $n = 63$, $S_n^{\text{obs}} = 26$. On donne également $h_{n-1} \simeq 4.7124$.

III Comparaison d'estimateurs

Soit $x = ((t_k, y_{1,k}, \dots, y_{k,k}), k \in \{2, \dots, n\})$, élément de $\prod_{k=2}^n (]0, +\infty[\times \mathbb{N}^k)$ une réalisation de l'arbre généalogique avec mutations $((T_k, Y_{1,k}, \dots, Y_{k,k}), k \in \{2, \dots, n\})$.

1. Justifier que la vraisemblance du modèle $p_n(x; \theta)$ est égale à :

$$\prod_{k=2}^n \frac{k(k-1)}{2} \exp \left(-\frac{k}{2}(k-1+\theta)t_k + \log(\theta t_k/2) \left(\sum_{i=1}^k y_{i,k} \right) - \sum_{i=1}^k \log(y_{i,k}!) \right).$$

2. Donner une statistique à valeurs dans \mathbb{R}^2 qui soit exhaustive.
3. Calculer $\tilde{\theta}_n$ l'estimateur du maximum de vraisemblance de θ en utilisant la log-vraisemblance $\mathcal{L}_n(x; \theta) = \log(p_n(x; \theta))$. À l'aide des questions I.2 et II.2, vérifier que $(\tilde{\theta}_n, n \geq 2)$ converge en probabilité vers θ .
4. Après avoir calculé l'information de Fisher : $I_n(\theta) = -\mathbb{E}_\theta[\partial_\theta^2 \mathcal{L}_n(X; \theta)]$, dire si l'estimateur de Watterson $\hat{\theta}_n$ est efficace. Peut-il être amélioré ?
5. Vérifier que l'estimateur de Watterson est asymptotiquement normal : la suite $(s_n(\hat{\theta}_n - \theta), n \geq 2)$ converge en loi vers une variable aléatoire gaussienne de loi $\mathcal{N}(0, \Sigma(\theta))$. Déterminer la vitesse de convergence s_n et la variance asymptotique $\Sigma(\theta)$. Vérifier que l'estimateur de Watterson est asymptotiquement efficace : $s_n^2/I_n(\theta) \sim \Sigma(\theta)$ quand n tend vers l'infini.

27. R. H. Ward, B. L. Frazier, K. Dew-Jager and S. Pääbo. Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl. Acad. Sci. USA*. vol. 88, pp. 8720–8724 (1991).

En pratique il est impossible d'observer l'arbre généalogique et donc d'observer la valeur de L_n . En revanche, comme on suppose que les mutations affectent des sites différents, on observe la valeur de S_n .

6. Justifier en peu de mots que, bien que l'on n'observe pas L_n et donc $\tilde{\theta}_n$, l'estimateur de Watterson donne une bonne estimation de θ .

IV Comparaison du taux de mutation

On souhaite vérifier que le taux de mutation θ de l'ADNmt est supérieur au taux de mutation de l'ADN chromosomique, θ_0 supposé connu. Pour cela on considère la statistique de test :

$$\zeta_n = Z_n(\theta_0)$$

et l'hypothèse nulle $H_0 = \{\theta = \theta_0\}$.

1. Donner l'hypothèse alternative H_1 et justifier pourquoi elle est unilatérale.
2. En utilisant les résultats de la partie II, donner le comportement asymptotique de la statistique de test sous H_0 et sous H_1 .
3. Donner la région critique de niveau asymptotique α .
4. Déterminer la p -valeur asymptotique. Faire l'application numérique avec les données de la question II.6 et un taux de mutation de l'ADN par nucléotide²⁸ de 10^{-4} soit $\theta_0 = 360 \times 10^{-4}$. Conclusion.

△

28. Le taux de mutation par nucléotide varie entre 10^{-4} et 10^{-8}

Deuxième partie

Corrections

XIII

Corrections

XIII.1 Espaces probabilisés

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) $\frac{2}{52} + 2 \frac{12}{52} \frac{3}{51} = \frac{29}{26 * 17}$. ▲

Exercice I.2. 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 $\mathbb{P}(A \text{ gagne})$, on compte le nombre de cas favorables divisé par le nombre de cas possibles ($\text{Card } \Omega = 6 \cdot 6 \cdot 12$). Le nombre de cas favorables est :

$$\begin{aligned} \text{Card } \{(i, j, k) \in \Omega; i + j > k\} &= \sum_{1 \leq i, j \leq 6} \sum_{k=1}^{12} \mathbf{1}_{\{i+j > k\}} = \sum_{1 \leq i, j \leq 6} (i + j - 1) \\ &= 2 * 6 \sum_{i=1}^6 i - 36 = 36 * 7 - 36 = 36 * 6. \end{aligned}$$

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

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

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

Exercice I.3. 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 :

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

On obtient les valeurs numériques suivantes :

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

(Pour les petites valeurs de n , on a l'approximation suivante :

$$\begin{aligned}
\frac{365!}{(365-n)!365^n} &= \prod_{k=1}^{n-1} \left(1 - \frac{k}{365}\right) = e^{\sum_{k=1}^{n-1} \log(1 - \frac{k}{365})} \\
&\simeq e^{-\sum_{k=1}^{n-1} \frac{k}{365}} = e^{-n(n-1)/730} \simeq e^{-n^2/730}.
\end{aligned}$$

On obtient $p_n \simeq 1/2$ pour $e^{-n^2/730} \simeq 1/2$ soit $n \simeq \sqrt{730 \log(2)}$. Comme $\log(2) \simeq 0.7$, il vient $n \simeq \sqrt{511}$ soit $n \in \{22, 23\}$.

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 :

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



Exercice I.4. 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 $(p_n, n \geq 1)$: $p_1 = p_2 = 0$ 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 $n \geq 3$, 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, il vient $p_4 = 3/8$ et $p_5 = 1/2$. Eugène s'est donc réjoui un peu vite...

On peut vérifier par récurrence que pour $n \geq 1$,

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

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

Exercice I.5. 1. On a $\mathbb{P}((p_r, p_b)) = \binom{p_r + p_b}{p_r} \left(\frac{r}{r+b} \right)^{p_r} \left(\frac{b}{r+b} \right)^{p_b}$.

2. On a

$$\begin{aligned} \mathbb{P}((p_r, p_b)) &= \frac{\binom{r}{p_r} \binom{b}{p_b}}{\binom{r+b}{p_r+p_b}} = \frac{\binom{p}{p_r} \binom{r+b-p}{r-p_r}}{\binom{r+b}{r}} \\ &= \binom{p_r + p_b}{p_r} \prod_{k=1}^{p_r} \frac{r-k+1}{r+b-k+1} \prod_{\ell=1}^{p_b} \frac{r-\ell+1}{r+b-\ell+1}. \end{aligned}$$

Pour la première égalité, on considère qu'il faut choisir p_r boules rouges parmi r et p_b boules bleues parmi b . Pour la deuxième égalité, on considère qu'il faut qu'il y ait p_r boules rouges parmi les p premières boules et $r - p_r$ dans les $r + b - p$ autres.

3. Dans les deux cas, on obtient :

$$\lim_{r, b \rightarrow +\infty; \frac{r}{r+b} \rightarrow \theta} \mathbb{P}((p_r, p_b)) = \binom{p_r + p_b}{p_r} \theta^{p_r} (1 - \theta)^{p_b}.$$

▲

Exercice I.6. 1. La correction est élémentaire.

2. On a vérifié (I.1) pour $n = 2$. On suppose (I.1) vraie pour n , et on la démontre pour $n + 1$. On a

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

où l'on a utilisé (I.1) avec $n = 2$ pour la première égalité et (I.1) avec n deux fois pour la deuxième égalité. L'égalité (I.1) est donc vraie au rang $n + 1$. Elle est donc vérifiée par récurrence.

3. On pose $I_{m,n} = \sum_{p=1}^m (-1)^{p+1} \sum_{1 \leq i_1 < \dots < i_p \leq n} \mathbb{P}(A_{i_1} \cap \dots \cap A_{i_p})$. On a pour $n = 2$:

$$I_{2,2} = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2) \leq \mathbb{P}(A_1 \cup A_2) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) = I_{1,2}.$$

Soit $n \geq 2$. On suppose la relation de récurrence vraie au rang n : pour tout $1 \leq m \leq n$, on a $I_{m,n} \leq \mathbb{P}(\bigcup_{i=1}^n A_i)$ si m est pair et $\mathbb{P}(\bigcup_{i=1}^n A_i) \leq I_{m,n}$ si m est impair. Soit $2 \leq m \leq n$ impair. On a

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

où l'on a utilisé (I.1) avec $n = 2$ pour la première égalité et l'hypothèse de récurrence sur n avec m et $m - 1$ pour l'inégalité. Un raisonnement similaire

assure que $I_{m,n} \leq \mathbb{P}(\bigcup_{i=1}^n A_i)$ pour m pair avec $2 \leq m \leq n$. L'inégalité pour $m = 1$ est immédiate. L'inégalité pour $m = n + 1$ est en fait une égalité d'après (I.1). La relation de récurrence est vraie au rang $n + 1$. Elle est donc vérifiée par récurrence. ▲

Exercice I.7. Le nombre de combinaisons de k exercices est $N = \binom{m}{k}$.

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

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

5. $N = 6$ et $p_1 \simeq 2.7 \cdot 10^{-16}$; $p_2 \simeq 1.6 \cdot 10^{-15}$; $p_3 \simeq 2.6\%$; $p_4 \simeq 15.2\%$.

Si les élèves choisissent au hasard, il est quasiment impossible qu'ils choisissent tous la même combinaison. S'ils ont tous choisi la même combinaison, alors le choix n'était pas fait au hasard (travail en groupe, exercices de difficulté ou d'intérêt différent, ...). ▲

Exercice I.8. 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 $\text{Card}(F) - \text{Card}(\bigcup_{i=1}^k A_i)$. On a $\text{Card}(F) = k^n$. D'après la formule du crible, on a

$$\begin{aligned} \text{Card}(\bigcup_{i=1}^k A_i) &= \sum_{j=1}^k (-1)^{j+1} \sum_{1 \leq i_1 \leq \dots \leq i_j} \text{Card}(A_{i_1} \cap \dots \cap A_{i_j}) \\ &= \sum_{j=1}^k (-1)^{j+1} \binom{k}{j} (k-j)^n. \end{aligned}$$

On obtient $N = \sum_{j=0}^k (-1)^j \binom{k}{j} (k-j)^n$.

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

$$\left\{ \begin{matrix} n \\ k \end{matrix} \right\} = \frac{N}{k!} = \frac{1}{k!} \sum_{j=0}^k (-1)^j \binom{k}{j} (k-j)^n.$$

3. Il est clair que $\left\{ \begin{matrix} n \\ 1 \end{matrix} \right\} = 1$ et $\left\{ \begin{matrix} n \\ k \end{matrix} \right\} = 0$ si $k > n$. Quand on dispose de $n > k > 1$ éléments à répartir en k sous-ensembles non vides, alors :

- Soit le dernier élément forme un sous-ensemble à lui tout seul et les $n - 1$ premiers éléments sont répartis en $k - 1$ sous-ensembles non vides. Ceci représente $\left\{ \begin{matrix} n - 1 \\ k - 1 \end{matrix} \right\}$ 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-ensembles. Ceci représente $k \left\{ \begin{matrix} n - 1 \\ k \end{matrix} \right\}$ cas possibles.

On en déduit donc la formule de récurrence. ▲

Exercice I.9. L'espace d'état est l'ensemble des permutations de $\{1, \dots, n\}$, la probabilité sur cet espace est la probabilité uniforme.

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

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

2. On note $\gamma(n)$ le nombre de permutations de $\{1, \dots, n\}$ sans point fixe. On a d'après la question précédente $1 - \frac{\gamma(n)}{n!} = \sum_{p=1}^n (-1)^{p+1} \frac{1}{p!}$ soit $\gamma(n) =$

$$n! \sum_{p=0}^n (-1)^p \frac{1}{p!}.$$

3. On remarque que

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

Il existe $\binom{n}{k}$ possibilités pour les k points fixes. On en déduit

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

4. On a $\lim_{n \rightarrow \infty} \pi_n(k) = \pi(k) = \frac{1}{k!} e^{-1}$. On retrouve la loi de Poisson de paramètre 1. En fait, on peut montrer le résultat¹ suivant sur la vitesse de convergence des probabilités π_n vers π : pour tout $B \subset \mathbb{N}$, on a (avec la convention $\pi_n(k) = 0$ si $k > n$)

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

▲

Exercice I.10. 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. $\mathbb{P}(\exists G) = \frac{\text{Card} \{(G, F), (F, G), (G, G)\}}{\text{Card} \{(F, F), (G, F), (F, G), (G, G)\}} = 3/4$.
2. $\mathbb{P}(\text{cadet} = G | \text{aînée} = F) = \frac{\mathbb{P}(\text{cadet} = G, \text{aînée} = F)}{\mathbb{P}(\text{aînée} = F)} = \frac{1/4}{1/2} = 1/2$.
3. $\mathbb{P}(\exists G | \exists F) = \frac{\mathbb{P}(\exists G, \exists F)}{\mathbb{P}(\exists F)} = \frac{1/2}{3/4} = 2/3$.
4. On a $\mathbb{P}(\exists G | \exists F, F \text{ décroche}) = \frac{\mathbb{P}(\exists G, \exists F, F \text{ décroche})}{\mathbb{P}(\exists F, F \text{ décroche})}$. On calcule d'abord le numérateur

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

On remarque ensuite que

$$\begin{aligned} \{\exists F, F \text{ décroche}\} &= \{(F, F), F \text{ décroche}\} \cup \{\exists G, \exists F, F \text{ décroche}\} \\ &= \{(F, F)\} \cup \{\exists G, \exists F, F \text{ décroche}\}, \end{aligned}$$

et les deux événements du dernier membre de droite sont disjoints. Ainsi on obtient :

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

1. A. D. Barbour, L. Holst and S. Janson. *Poisson approximation*, chap. 4. Oxford University Press (1992).

5. 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 (suivant que l'aîné(e) ou le cadet(te) ouvre la porte) :

$$\begin{aligned} \mathbb{P}(\exists G, \exists F, F \text{ ouvre la porte}) &= \mathbb{P}(\text{aînée ouvre la porte}, (F, G)) \\ &\quad + \mathbb{P}(\text{cadette ouvre la porte}, (G, F)). \end{aligned}$$

Par indépendance de $\{\text{aîné(e) ouvre la porte}\}$ et $\{(F, G)\}$, on a :

$$\begin{aligned} \mathbb{P}(\text{aînée ouvre la porte}, (F, G)) \\ = \mathbb{P}(\text{aîné(e) ouvre la porte})\mathbb{P}((F, G)) = p \frac{1}{4}. \end{aligned}$$

Comme

$$\begin{aligned} \mathbb{P}(\text{cadette ouvre la porte}, (G, F)) \\ = \mathbb{P}((G, F)) - \mathbb{P}(\text{aîné ouvre la porte}, (G, F)), \end{aligned}$$

on a par indépendance de $\{\text{aîné(e) ouvre la porte}\}$ et $\{(G, F)\}$ que

$$\begin{aligned} \mathbb{P}(\text{cadette ouvre la porte}, (G, F)) \\ = \frac{1}{4} - \mathbb{P}(\text{aîné(e) ouvre la porte})\mathbb{P}((G, F)) = (1 - p) \frac{1}{4}. \end{aligned}$$

On en déduit donc que

$$\mathbb{P}(\exists G, \exists F, F \text{ ouvre la porte}) = p \frac{1}{4} + (1 - p) \frac{1}{4} = \frac{1}{4}.$$

De manière similaire, on obtient

$$\begin{aligned} \mathbb{P}(\exists F, F \text{ ouvre la porte}) \\ = \mathbb{P}(\text{aîné(e) ouvre la porte})\mathbb{P}((F, G) \text{ ou } (F, F)) \\ \quad + \mathbb{P}(\text{cadet(te) ouvre la porte})\mathbb{P}((G, F) \text{ ou } (F, F)) \\ = p \frac{1}{2} + (1 - p) \frac{1}{2} = \frac{1}{2}. \end{aligned}$$

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



Exercice I.11. 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 $\mathbb{P}(M)$ que l'on note τ . On déduit de la formule de Bayes que :

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

On déduit également de la formule de Bayes que :

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

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.12. 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 . Le génotype des parents d'Adrien est mb .

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

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

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 $\mathbb{P}(2 = B | 1 = M) = \frac{\mathbb{P}(1 = M, 2 = B)}{\mathbb{P}(1 = M)} = \frac{1}{4}$.

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.13. 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

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

▲

Exercice I.14. On note r_A la probabilité que vous choisissiez la porte A . La probabilité, $p_{B|A;C}$, pour que le cadeau soit derrière la porte B ("cadeau en B ") sachant que vous avez choisi la porte A ("choisir A ") et que le présentateur ouvre la porte C ("ouvrir C ") est égale à

$$\frac{\mathbb{P}(\text{choisir } A, \text{ cadeau en } B, \text{ ouvrir } C)}{\mathbb{P}(\text{choisir } A, \text{ ouvrir } C)}.$$

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

$$\begin{aligned} & \mathbb{P}(\text{choisir } A, \text{ cadeau en } A, \text{ ouvrir } C) \\ &= \mathbb{P}(\text{ouvrir } C | \text{choisir } A, \text{ cadeau en } A) \\ & \qquad \qquad \qquad \mathbb{P}(\text{choisir } A, \text{ cadeau en } A). \end{aligned}$$

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

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

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 le cadeau est en B . On note $p_{C|A;B}$ la probabilité pour que le cadeau 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 le cadeau 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 .



XIII.2 Variables aléatoires discrètes

Exercice II.1.

| Loi | $\mathbb{E}[X]$ | $\text{Var}(X)$ | $\phi_X(z)$ |
|---|-----------------|---------------------|---------------------------|
| 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 | $e^{-\theta(1-z)}$ |



Exercice II.2. 1. Comme $X \geq 0$ p.s., on en déduit que p.s. $\left| \frac{1}{1 + X} \right| = \frac{1}{1 + X} \leq 1$.

Donc la v.a. $\frac{1}{1 + X}$ est intégrable. On a :

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

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

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

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

$$\mathbb{E}\left[\frac{1}{2+X}\right] = \mathbb{E}\left[\frac{1}{1+X}\right] - \mathbb{E}\left[\frac{1}{(1+X)(2+X)}\right] = \frac{\theta-1+e^{-\theta}}{\theta^2}.$$

▲

Exercice II.3. 1. La position de la bonne clef est au hasard sur les positions possibles. La loi de X est donc la loi uniforme sur $\{1, \dots, n\}$. On a

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=1}^n k \mathbb{P}(X=k) = \frac{1}{n} \sum_{k=1}^n k = \frac{n+1}{2}, \\ \mathbb{E}[X^2] &= \sum_{k=1}^n k^2 \mathbb{P}(X=k) = \frac{1}{n} \sum_{k=1}^n k^2 = \frac{(n+1)(2n+1)}{6}, \\ \text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 = \frac{n^2-1}{12}.\end{aligned}$$

2. Le gardien essaie les clefs éventuellement plusieurs fois chacune. (Il s'agit d'un tirage avec remise, alors que dans la question précédente il s'agissait d'un tirage

sans remise). Soit A_k l'événement : "le gardien choisit la bonne clef lors de la k -ième tentative". Les événements $(A_k, k \geq 1)$ sont indépendants et de probabilité $1/n$. La loi de X est donc la loi géométrique de paramètre $p = 1/n$. On a $X \in \mathbb{N}^*$ et pour $k \geq 1$,

$$\mathbb{P}(X = k) = \left(1 - \frac{1}{n}\right)^{k-1} \frac{1}{n}.$$

On sait que si une variable aléatoire X suit une loi géométrique de paramètre p alors $\mathbb{E}[X] = 1/p$ et $\text{Var}(X) = (1-p)/p^2$. En remplaçant p par $1/n$, on trouve $\mathbb{E}[X] = n$ et $\text{Var}(X) = n(n-1)$.

3. On note par I l'événement : "le gardien est ivre". Par la formule de Bayes, on obtient

$$\mathbb{P}(I|X = n) = \frac{\mathbb{P}(X = n|I)\mathbb{P}(I)}{\mathbb{P}(X = n|I)\mathbb{P}(I) + \mathbb{P}(X = n|I^c)\mathbb{P}(I^c)} = \frac{1}{1 + 2\left(\frac{n}{n-1}\right)^{n-1}}.$$

On a $\lim_{n \rightarrow \infty} \mathbb{P}(I|X = n) = \frac{1}{1 + 2e} \simeq 0.16$.

▲

Exercice II.4. 1. Soit X_i le nombre de jets nécessaires pour que le $i^{\text{ème}}$ dé amène un six pour la première fois. Les variables aléatoires $(X_i, 1 \leq i \leq 5)$ sont indépendantes et suivent une loi géométrique de paramètre $1/6$. Comme $X = \max_{1 \leq i \leq 5} X_i$, on obtient pour $k \geq 1$

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

Cette formule est valable pour $k = 0$: $\mathbb{P}(X \leq 0) = 0$.

2. On a

$$\sum_{k=1}^{\infty} \mathbf{1}_{\{Y \geq k\}} = \sum_{k=1}^{\infty} \sum_{j=k}^{\infty} \mathbf{1}_{\{Y=j\}} = \sum_{j=1}^{\infty} j \mathbf{1}_{\{Y=j\}}.$$

En prenant l'espérance dans les égalités ci-dessus, et en utilisant le théorème de convergence monotone pour permuter les sommations et l'espérance, on obtient

$$\sum_{k=1}^{\infty} \mathbb{P}(Y \geq k) = \sum_{j=1}^{\infty} j \mathbb{P}(Y = j) = \mathbb{E}[Y].$$

3. On a $\mathbb{E}[X] = \sum_{k=1}^{\infty} (1 - \mathbb{P}(X \leq k - 1)) = \sum_{k=0}^{\infty} \left[1 - \left(1 - \left(\frac{5}{6} \right)^k \right)^5 \right]$. Il vient $\mathbb{E}[X] \simeq 13.02$.

▲

Exercice II.5. 1. On a $\phi_Y(z) = \frac{az}{1 - (1-a)z}$ et $\phi_Z(z) = e^{-\theta(1-z)}$.

2. On a $U = Y\mathbf{1}_{\{X=1\}}$, $z^U = \mathbf{1}_{\{X=0\}} + z^Y\mathbf{1}_{\{X=1\}}$. Il vient

$$\phi_U(z) = \mathbb{E}[z^U] = \mathbb{E}[\mathbf{1}_{\{X=0\}} + z^Y\mathbf{1}_{\{X=1\}}] = 1 - p + p\phi_Y(z).$$

On a

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

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

$$\phi_V(z) = \mathbb{E}[z^V] = \mathbb{E}[\mathbf{1}_{\{X=0\}}z^Y + \mathbf{1}_{\{X=1\}}z^Z] = (1-p)\phi_Y(z) + p\phi_Z(z).$$

On a

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

▲

Exercice II.6. 1. On note X_i le résultat du candidat à la question i . Les v.a.d. $(X_i, 1 \leq i \leq 20)$ 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. $(Y_i, 1 \leq i \leq 20)$ 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 :

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

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. Le nombre total de points 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.7. 1. La probabilité pour que lors du i -ème tirage, les deux boules n'aient pas la même couleur est $p = \frac{2RB}{(R+B)^2}$. On définit les variables aléatoires X_i par $X_i = 1$ si lors du i -ème tirage, les deux boules n'ont pas la même couleur et $X_i = 0$ sinon. Les variables aléatoires $(X_i, 1 \leq i \leq R+B)$ suivent la loi de Bernoulli de paramètre p . (Elles ne sont pas indépendantes.) On a $X = \sum_{i=1}^{R+B} X_i$. On en déduit par linéarité que

$$\mathbb{E}[X] = \sum_{i=1}^{R+B} \mathbb{E}[X_i] = (R+B)p = \frac{2RB}{R+B}.$$

2. La variable aléatoire X prend des valeurs paires. Les valeurs possibles de X sont $\{2d; 0 \leq d \leq \min(R, B)\}$. Pour obtenir $2d$ tirages de couleurs différentes, il faut choisir d boules parmi les rouges et d boules parmi les bleues. Enfin, une fois les boules de couleurs différentes choisies, l'ordre du tirage de la première urne (il y a $\frac{(R+B)!}{R!B!}$ possibilités) n'a pas d'importance. On a donc $\frac{(R+B)!}{R!B!} \frac{R!B!}{(R-d)!d!(B-d)!d!}$ cas favorables parmi $\left(\frac{(R+B)!}{R!B!}\right)^2$ cas possibles. Donc, pour $0 \leq d \leq \min(R, B)$, on a

$$\mathbb{P}(X = 2d) = \frac{R!^2 B!^2}{(R+B)!(R-d)!(B-d)!d!^2}.$$
▲

Exercice II.8. On note par X_1, X_2, \dots , et X_n les résultats des n boules. Ces variables aléatoires sont indépendantes de loi uniforme sur $\{1, \dots, N\}$.

1. Soit $x \in \{1, \dots, N\}$. On a $\{X \geq x\} = \cap_{k=1}^n \{X_k \geq x\}$. Comme les variables aléatoires X_1, \dots, X_n sont indépendantes et identiquement distribuées, il vient

$$\mathbb{P}(X \geq x) = \mathbb{P}(X_1 \geq x, \dots, X_n \geq x) = \mathbb{P}(X_1 \geq x)^n = \frac{(N-x+1)^n}{N^n}.$$

On en déduit que pour $x \in \{1, \dots, N-1\}$,

$$\mathbb{P}(X = x) = \mathbb{P}(X \geq x) - \mathbb{P}(X \geq x+1) = \frac{(N-x+1)^n - (N-x)^n}{N^n}.$$

Cette formule est encore valable pour $x = N$, car $\mathbb{P}(X = N) = \frac{1}{N^n}$.

2. Par symétrie, Y a même loi que $N - X + 1$. Pour $y \in \{1, \dots, N\}$, on a $\mathbb{P}(Y = y) = \frac{y^n - (y-1)^n}{N^n}$.
3. Soit $(x, y) \in \{0, \dots, N\}^2$. Si $x \geq y$, on a

$$\mathbb{P}(X > x, Y \leq y) = \mathbb{P}(x < X \leq Y \leq y) = 0.$$

Si $x < y$, on a

$$\begin{aligned} \mathbb{P}(X > x, Y \leq y) &= \mathbb{P}(x < X \leq Y \leq y) \\ &= \mathbb{P}(x < X_i \leq y, \forall i \in \{1, \dots, n\}) \\ &= \mathbb{P}(x < X_1 \leq y)^n = \frac{(y-x)^n}{N^n}. \end{aligned}$$

Cette formule est encore valable pour $x = y$. Pour la loi du couple (X, Y) , on a pour $(x, y) \in \{1, \dots, N\}^2$,

- si $x > y$, $\mathbb{P}(X = x, Y = y) = 0$ car on a toujours $X \leq Y$.
- si $x = y$, $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X_i = x, \forall i \in \{1, \dots, n\}) = \frac{1}{N^n}$.
- 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.9. 1. La fonction génératrice associée au résultat d'un lancer du dé à onze faces est :

$$\phi(z) = \frac{1}{11} \sum_{k=2}^{12} z^k = \frac{1}{11} z^2 \frac{1-z^{11}}{1-z}.$$

Toutes les racines du polynôme $1 - z^{11}$ sont complexes sauf la racine 1 qui est réelle. Le polynôme ϕ admet donc deux racines réelles : 0 (de multiplicité 2) et 1.

2. La fonction génératrice associée à la somme d'un lancer de deux dés à six faces est, par indépendance, de la forme

$$\phi_2(z) = \left(\sum_{k=1}^6 p_k z^k\right) \left(\sum_{k=1}^6 q_k z^k\right) = z^2 \left(\sum_{k=1}^6 p_k z^{k-1}\right) \left(\sum_{k=1}^6 q_k z^{k-1}\right),$$

où p_k (resp. q_k) est la probabilité d'obtenir k avec le premier (resp. deuxième) dé. Si $\phi = \phi_2$, alors le polynôme $\sum_{k=1}^6 p_k z^{k-1}$ est de degré 5. Comme il est à coefficients réels, il admet au moins une racine réelle. Il en est de même pour le polynôme $\sum_{k=1}^6 q_k z^{k-1}$. Le polynôme ϕ_2 admet donc au moins quatre racines réelles. Il ne peut donc être égal à ϕ qui n'admet que deux racines réelles dont une de multiplicité 2.

On ne peut donc pas reproduire le résultat d'un lancer d'un dé équilibré à onze faces, numérotées de 2 à 12, comme la somme d'un lancer de deux dés à six faces, numérotées de 1 à 6, éventuellement différemment biaisés. ▲

Exercice II.10. Soit X et Y les variables aléatoires discrètes qui représentent le nombre de piles obtenus par chacun des joueurs au cours des n lancers. Les variables X et Y sont indépendantes et suivent des lois binomiales $\mathcal{B}(n, 1/2)$. On a, en utilisant la formule de décomposition puis l'indépendance de X et Y ,

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

Pour démontrer l'égalité $\sum_{k=0}^n \binom{n}{k}^2 = \binom{2n}{n}$, 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.11. 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 boîte, il reste k allumettes dans l'autre boîte", 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

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

2. La probabilité cherchée est $p_0 = \binom{2N}{N} p^N(1-p)^N$. Si on suppose que $p = 1/2$, alors $p_0 = \binom{2N}{N}/2^{2N}$. 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, en prenant $p = 1/2$, la relation suivante non

$$\text{triviale sur les coefficients binomiaux : } \sum_{k=0}^N 2^k \binom{2N-k}{N} = 2^{2N}.$$

▲

Exercice II.12. 1. Soit $m, n \geq 1$. L’évènement $\{T_1 = m, T_2 - T_1 = n\}$ est égal à $\{X_1 = 0, \dots, X_{m-1} = 0, X_m = 1, X_{m+1} = 0, \dots, X_{m+n-1} = 0, X_{m+n} = 1\}$. Par indépendance des variables aléatoires X_i , on a donc

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

En utilisant la formule des lois marginales, il vient

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

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 variables aléatoires 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 $n_1, n_2, \dots, n_{k+1} \geq 1$, notons $I = \{n_1, n_1 + n_2, \dots, n_1 + \dots + n_{k+1}\}$ et $J = \{1, \dots, n_1 + \dots + n_{k+1}\} \setminus I$. L'évènement $\{T_1 - T_0 = n_1, T_2 - T_1 = n_2, \dots, T_{k+1} - T_k = n_{k+1}\}$ est alors égal à

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

Par indépendance des variables aléatoires X_i , on a donc

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

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

$$\mathbb{P}(T_{j+1} - T_j = n_{j+1}) = p(1-p)^{n_{j+1}-1}. \quad (\text{XIII.1})$$

Par conséquent, il vient

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

ce qui prouve que les variables aléatoires $T_1 - T_0, T_2 - T_1, \dots, T_{k+1} - T_k$ sont indépendantes. De plus, d'après (XIII.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, on a

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

Par indépendance des variables aléatoires $T_{j+1} - T_j$, il vient

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

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 variables aléatoires X_i , on a donc

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

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

Par indépendance des variables aléatoires $T_{j+1} - T_j$, la fonction génératrice ϕ_{T_k} de T_k est le produit des fonctions génératrices des variables aléatoires $T_{j+1} - T_j$ ($j \in \{0, \dots, k-1\}$). 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 $\{\tau = n\}$ et on utilise le fait que τ est indépendant de $(T_n, n \geq 1)$:

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

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

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

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. 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 $p\rho$. On retrouve que T_τ suit la loi géométrique de paramètre $p\rho$.

▲

Exercice II.13. 1. On a

$$\begin{aligned} \mathbb{P}(T = +\infty) &= \mathbb{P}\left(\bigcup_{n \geq 0} \{X_k = 0, 1 \leq k \leq n\} \cap \{X_k = 1, k > n\}\right) \\ &= \sum_{n=0}^{\infty} \mathbb{P}(\{X_k = 0, 1 \leq k \leq n\} \cap \{X_k = 1, k > n\}). \end{aligned}$$

Par ailleurs, on a

$$\begin{aligned} &\mathbb{P}(\{X_k = 0, 1 \leq k \leq n\} \cap \{X_k = 1, k > n\}) \\ &= \lim_{N \rightarrow \infty} \mathbb{P}(\{X_k = 0, 1 \leq k \leq n\} \cap \{X_k = 1, N \geq k > n\}) \\ &= \lim_{N \rightarrow \infty} (1-p)^n p^{N-n} = 0. \end{aligned}$$

On en déduit donc que $\mathbb{P}(T = +\infty) = 0$.

On propose une deuxième méthode de portée plus générale : on cherche à majorer T par une variable aléatoire géométrique. 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 fini 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 :

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

Donc T a même loi que $U+V$. (En fait U est le premier temps d'occurrence de 1, et à partir de cet instant on attend la première occurrence de 0, qui correspond à V . Le temps total d'attente de l'occurrence 10 est donc $T = U + V$.)

3. On a $\phi_U(z) = \frac{zp}{1 - (1-p)z}$, $\phi_V(z) = \frac{z(1-p)}{1-pz}$. On en déduit, par indépendance de U et V , que $\phi_T(z) = \phi_{U+V}(z) = \phi_U(z)\phi_V(z)$.
4. On a $\mathbb{E}[T] = \mathbb{E}[U] + \mathbb{E}[V] = \frac{1}{p(1-p)}$. On a également par indépendance $\text{Var}(T) = \text{Var}(U) + \text{Var}(V) = \frac{1-3p(1-p)}{p^2(1-p)^2}$.

▲

Exercice II.14. 1. On calcule la fonction génératrice du couple $(S, N - S)$. Pour $v, z \in [-1, 1]$, on a par indépendance

$$\begin{aligned} \phi_{(S, N-S)}(v, z) &= \mathbb{E}[v^S z^{N-S}] = \mathbb{P}(N = 0) + \sum_{n=1}^{\infty} \mathbb{P}(N = n) z^n \mathbb{E}[(v/z)^{\sum_{k=1}^n X_k}] \\ &= \mathbb{P}(N = 0) + \sum_{n=1}^{\infty} \mathbb{P}(N = n) (pv + (1-p)z)^n \\ &= \phi(pv + (1-p)z) = e^{-\theta p(1-v)} e^{-\theta(1-p)(1-z)}. \end{aligned}$$

On remarque que $\phi_{(S, N-S)}(v, z) = \phi_V(v)\phi_Z(z)$, où V et Z suivent des lois de Poisson de paramètre respectif θp et $\theta(1-p)$. En particulier, cela implique que S et $N - S$ sont indépendants.

2. Soit $z \in [-1, 1]$. On a, par indépendance entre N et $N - S$, $\phi(z) = \mathbb{E}[z^S z^{N-S}] = \mathbb{E}[z^S] \mathbb{E}[z^{N-S}]$. Par indépendance entre N et $(X_n, n \geq 1)$, on a

$$\begin{aligned} \mathbb{E}[z^S] &= \mathbb{P}(N = 0) + \sum_{n=1}^{\infty} \mathbb{P}(N = n) \mathbb{E}[z^{\sum_{k=1}^n X_k}] \\ &= \mathbb{P}(N = 0) + \sum_{n=1}^{\infty} \mathbb{P}(N = n) ((1-p) + pz)^n = \phi((1-p) + pz), \end{aligned}$$

et par symétrie $\mathbb{E}[z^{N-S}] = \phi(p + (1-p)z)$. On en déduit donc que $\phi(z) = \phi((1-p) + pz)\phi(p + (1-p)z)$. On remarque que h est bien définie sur $]0, 1[$, car ϕ est strictement positif pour $z > 0$ et ϕ' est définie a priori sur $] - 1, 1[$. La relation $h(z) = ph((1-p) + pz) + (1-p)h(p + (1-p)z)$ est immédiate.

3. Rappelons que ϕ est continue en 1, non nulle en 1 (en fait égale à 1 en 1), et que ϕ' est croissante et admet une limite à gauche en 1, éventuellement égale à l'infini. En particulier, h admet une limite à gauche en 1, éventuellement égale à l'infini.

Pour $a \in [0, 1[$, on pose $f_a(z) = 1 - a + az$ pour $z \in \mathbb{R}$. L'application f_a est contractante (avec constante de Lipschitz a) et admet 1 comme seul point fixe.

La suite $(f_a^n(z), n \geq 1)$ converge en croissant vers 1 pour $z \in [0, 1[$, où f_a^n désigne l'itéré n -ième de f_a .

Comme $h = h \circ f_{1/2}$, on en déduit que $h = h \circ f_{1/2}^n$ et par continuité $h(z) = \lim_{r \rightarrow 1^-} h(r)$ pour tout $z \in [0, 1[$. Ceci implique que h est constante sur $[0, 1[$.

Comme h est positive et finie sur $[0, 1[$, on en déduit qu'il existe $c \geq 0$ tel que $\phi' = c\phi$ sur $[0, 1[$ et donc sur $[0, 1]$ par continuité. Si $c = 0$, on a ϕ constant et donc, comme $\phi(1) = 1$, on a $\phi = 1$ c'est-à-dire p.s. $N = 0$. Si $c > 0$, la solution de l'équation différentielle ordinaire $\phi' = c\phi$ sur $[0, 1]$ avec condition $\phi(1) = 1$ donne $\phi(z) = e^{-c(1-z)}$. Donc, si $c > 0$, alors N suit la loi de Poisson de paramètre c .

4. Comme $p < 1/2$, on a $p + (1-p)z \leq 1 - p + pz$ pour $z \in [0, 1[$. On déduit donc de $h(z) \geq \min(h((1-p) + pz), (1-p)h(p + (1-p)z))$ que $h(z) \geq \inf\{h(u); u \geq p + (1-p)z\}$. Comme $h(z) \geq \inf\{h(u); u \geq f_{1-p}(z)\}$, on en déduit que $h(z) \geq \inf\{h(u); u \geq f_{1-p}^n(z)\}$, et donc $h(z) \geq \liminf_{r \rightarrow 1^-} h(r)$ pour tout $z \in [0, 1[$.

Un raisonnement similaire à ce qui précède assure que $h(z) \leq \max(h((1-p) + pz), (1-p)h(p + (1-p)z))$, puis $h(z) \leq \sup\{h(u); u \geq f_p(z)\}$ et $h(z) \leq \sup\{h(u); u \geq f_p^n(z)\}$, et donc $h(z) \leq \limsup_{r \rightarrow 1^-} h(r)$ pour tout $z \in [0, 1[$.

Comme h admet une limite à gauche en 1, éventuellement égale à l'infini (cf le début de la démonstration de la question précédente), on en déduit que $h(z) = \lim_{r \rightarrow 1^-} h(r)$, et donc h est constant sur $[0, 1[$. La fin de la démonstration de la question précédente permet de conclure.

▲

Exercice II.15. 1. On a $\mathbb{P}(X_1 = k_1, \dots, X_n = k_n \mid S_n = k) = 0$ si $\sum_{i=1}^n k_i \neq k$ et sinon

$$\mathbb{P}(X_1 = k_1, \dots, X_n = k_n \mid S_n = k) = \frac{\prod_{i;k_i=1} p \prod_{j;k_j=0} (1-p)}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{1}{\binom{n}{k}}.$$

On en déduit que la loi conditionnelle de (X_1, \dots, X_n) sachant S_n est la loi uniforme sur $\{(k_1, \dots, k_n) \in \{0, 1\}^n; \sum_{i=1}^n k_i = k\}$.

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 $\mathbb{P}(X_i = 1 \mid S_n = k) = \binom{n-1}{k-1} / \binom{n}{k} = k/n$ pour $k \geq 1$, et $\mathbb{P}(X_i = 1 \mid S_n = 0) = 0$, on en déduit que $\tilde{p} = S_n/n$.
3. On a $\mathbb{P}(X_1 = 1, X_2 = 1 \mid S_n) = S_n(S_n - 1) / (n(n - 1))$. En particulier $\mathbb{P}(X_1 = 1, X_2 = 1 \mid S_n) \neq \mathbb{P}(X_1 = 1 \mid S_n)\mathbb{P}(X_2 = 1 \mid S_n)$. Conditionnellement à S_n , les variables X_1 et X_2 ne sont pas indépendantes.

▲

Exercice II.16. 1. On utilise les fonctions génératrices. Par indépendance, on a pour $z \in [-1, 1]$:

$$\phi_{X_1+X_2}(z) = \phi_{X_1}(z)\phi_{X_2}(z) = e^{-\theta_1(1-z)-\theta_2(1-z)} = e^{-(\theta_1+\theta_2)(1-z)}.$$

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. En utilisant l'indépendance, on calcule, pour $0 \leq k \leq n$:

$$\begin{aligned} \mathbb{P}(X_1 = k | X_1 + X_2 = n) &= \frac{\mathbb{P}(X_1 = k, X_2 = n - k)}{\mathbb{P}(X_1 + X_2 = n)} \\ &= \frac{\mathbb{P}(X_1 = k)\mathbb{P}(X_2 = n - k)}{\mathbb{P}(X_1 + X_2 = n)} \\ &= e^{-\theta_1} \frac{\theta_1^k}{k!} e^{-\theta_2} \frac{\theta_2^{n-k}}{(n-k)!} e^{(\theta_1+\theta_2)} \frac{n!}{(\theta_1 + \theta_2)^n} \\ &= C_n^k \left(\frac{\theta_1}{\theta_1 + \theta_2} \right)^k \left(\frac{\theta_2}{\theta_1 + \theta_2} \right)^{n-k}. \end{aligned}$$

La loi de X_1 sachant $\{X_1 + X_2 = n\}$ est une loi binomiale $\mathcal{B}(n, p)$, 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 $\mathbb{E}[Z] = np$. On déduit de la question précédente que :

$$\mathbb{E}[X_1 | X_1 + X_2] = (X_1 + X_2) \frac{\theta_1}{\theta_1 + \theta_2}.$$

▲

Exercice II.17. 1. On note ϕ_X, ϕ_Y et ϕ_S les fonctions génératrices de X, Y et S . On a $\phi_X(z) = \phi_Y(z) = \frac{pz}{1 - (1-p)z}$. Par indépendance, on a $\phi_S(z) =$

$\phi_X(z)\phi_Y(z) = \frac{p^2 z^2}{(1 - (1-p)z)^2}$. 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 $n \geq 2$, $\mathbb{P}(S = n) = p^2 (n-1)(1-p)^{n-2}$.

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

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

Ainsi, conditionnellement à S , X suit la loi uniforme sur $\{1, \dots, S-1\}$. 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 $\mathbb{E}[\mathbb{E}[X|S]] = \frac{\mathbb{E}[S]}{2} = \frac{\mathbb{E}[X+Y]}{2} = \mathbb{E}[X]$. ▲

Exercice II.18. Le coût de la première stratégie est $c_1 = cN$. On note C_n le coût du test sur un groupe de n personnes en mélangeant les prélèvements. On a $\mathbb{P}(C_n = c) = (1-p)^n$ et $\mathbb{P}(C_n = c + nc) = 1 - (1-p)^n$. On suppose que N/n est entier et on regroupe les N personnes en N/n groupes de n personnes. Le coût moyen total du test, en utilisant la deuxième stratégie, est $c_2 = \frac{N}{n}\mathbb{E}[C_n]$ soit :

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

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 stratégie. On peut vérifier que pour $p \ll 1$, en prenant $n = \lceil 1/\sqrt{p} \rceil$, 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 :

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

Exercice II.19. 1. Soit X une variable aléatoire géométrique de paramètre $p \in]0, 1[$.

On a pour $n \in \mathbb{N}$,

$$\mathbb{P}(X > n) = \sum_{k \geq n+1} \mathbb{P}(X = k) = \sum_{k \geq n+1} p(1-p)^{k-1} = (1-p)^n.$$

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. On pose $q = \mathbb{P}(X > 1) \in [0, 1]$. Comme $\mathbb{P}(X > 1 + n | X > n)$ est indépendant de n , on a

$$\mathbb{P}(X > 1 + n | X > n) = \mathbb{P}(X > 1 | X > 0) = \mathbb{P}(X > 1) = q,$$

car p.s. $X \in \mathbb{N}^*$. Si $q = 0$, alors $\mathbb{P}(X > n) = 0$, et les probabilités conditionnelles ne sont pas définies. On a donc $q > 0$. Comme $\mathbb{P}(X > 1 + n | X > n) = \frac{\mathbb{P}(X > 1 + n)}{\mathbb{P}(X > n)}$, il vient $\mathbb{P}(X > n + 1) = q\mathbb{P}(X > n)$ i.e. $\mathbb{P}(X > n + 1) = q^{n+1}\mathbb{P}(X > 0)$, et donc $\mathbb{P}(X > n) = q^n$ car $\mathbb{P}(X > 0) = 1$. 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 $\alpha = \mathbb{P}(X > 0)$. La question précédente traite 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 pose $q = \mathbb{P}(X > 1 | X > 0) \in [0, 1]$. L'absence de mémoire implique que $\mathbb{P}(X > 1 + n | X > n) = q$ soit $\mathbb{P}(X > 1 + n) = q\mathbb{P}(X > n)$ et donc en itérant $\mathbb{P}(X > 1 + n) = q^{n+1}\mathbb{P}(X > 0) = q^{n+1}\alpha$. 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 variable aléatoire de Bernoulli de paramètre α , et Z est une variable aléatoire 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.20. Si les sportifs de haut niveau sont uniformément répartis dans la population, les médailles le sont aussi. Chaque individu a donc, environ, une probabilité $p_m \simeq 928/6 \cdot 10^9$ d'avoir une médaille et $p_o \simeq 301/6 \cdot 10^9$ d'avoir une médaille d'or. Le nombre de médailles françaises suit donc une loi binomiale de paramètre (n, p_m) , avec $n \simeq 60 \cdot 10^6$. 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 médailles 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 n'est donc pas réaliste.

▲

Exercice II.21. 1. On considère des variables aléatoires $(X_n, n \in \mathbb{N}^*)$ indépendantes de loi de Bernoulli de paramètre p . On pose $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$. On a :

$$\mathbb{P}(C_n^c) = \mathbb{P}(|\bar{X}_n - p| > \delta_n) \leq \frac{\mathbb{E}[(\bar{X}_n - p)^2]}{\delta_n^2} = \frac{\text{Var}(\bar{X}_n)}{\delta_n^2} = \frac{p(1-p)}{n\delta_n^2},$$

où l'on a utilisé l'inégalité de Tchebychev pour l'inégalité et l'indépendance pour la dernière égalité. On en déduit le résultat.

2. Un calcul élémentaire assure que l'entropie H_p est maximale pour $p = 1/2$.
3. On a $\mathbb{P}(\{\omega\}) = e^{n(\frac{1}{n} \sum_{i=1}^n \omega_i \log(p) + (1 - \frac{1}{n} \sum_{i=1}^n \omega_i) \log(1-p))}$. Pour $\omega \in C_n$, il vient :

$$\begin{aligned} n(p \log(p) + (1-p) \log(1-p) + \delta_n \log(p(1-p))) \\ \leq \log(\mathbb{P}(\{\omega\})) \leq n(p \log(p) + (1-p) \log(1-p) - \delta_n \log(p(1-p))). \end{aligned}$$

On a donc :

$$e^{-n(H_p + \beta_n)} \leq \mathbb{P}(\{\omega\}) \leq e^{-n(H_p - \beta_n/2)},$$

où $0 \leq \beta_n = -2\delta_n \log(p(1-p))$.

4. En sommant l'inégalité précédente sur $\omega \in C_n$, on obtient :

$$\text{Card}(C_n) e^{-n(H_p + \beta_n)} \leq \mathbb{P}(C_n) \leq 1 \quad \text{et} \quad \mathbb{P}(C_n) \leq e^{-n(H_p - \beta_n/2)} \text{Card}(C_n).$$

Comme $\mathbb{P}(C_n) \geq 1 - \alpha$ pour n grand et que $\lim_{n \rightarrow +\infty} n\beta_n = +\infty$, on en déduit que pour n suffisamment grand on a $\mathbb{P}(C_n) \geq e^{-n\beta_n/2}$. Pour n suffisamment grand, il vient :

$$e^{n(H_p - \beta_n)} \leq \text{Card}(C_n) \leq e^{n(H_p + \beta_n)}.$$

5. Pour $p = 1/2$, l'entropie est maximale et vaut $H_p = \log(2)$. On en déduit que le cardinal de l'ensemble des suites typiques est de l'ordre de $2^n = \text{Card}(\Omega)$. Pour $p \simeq 1$, on obtient que le cardinal de l'ensemble des suites typiques est de l'ordre de $e^{n(H_p \pm \beta_n)}$ qui est beaucoup plus petit que le cardinal de Ω . (Le résultat est similaire pour $p \simeq 0$).

▲

Exercice II.22. 1. Comme $0 \leq u_n, v_n \leq 1$, on en déduit que les séries U et V sont absolument convergentes sur $] -1, 1[$.

2. On note $H_n = \{\text{le phénomène se produit à l'instant } n\}$. On a

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

La dernière 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

$$\begin{aligned} \mathbb{P}(\sum_{i=2}^k X_i = n - \ell, \text{ pour un } k \geq 2) \\ = \mathbb{P}(\sum_{i=2}^k X_i = n - \ell, \text{ pour un } k \in \{2, \dots, n - \ell + 1\}) = u_{n-\ell}. \end{aligned}$$

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 $V = BU$.

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 $U(s) - 1 = F(s)U(s)$ et donc $V = B/(1 - F)$.

4. On a $V(s) = p/(1 - s)$. En évaluant (II.1) en $s = 0$, on obtient $b_0 = p$. On a $F(s) = 1 - \frac{1}{p}(1 - s)B(s)$. En dérivant $n \geq 1$ fois, il vient $F^{(n)}(s) = \frac{1}{p}(nB^{(n-1)}(s) - (1 - s)B^{(n)}(s))$. En évaluant en $s = 0$, on obtient $f_n = (b_{n-1} - b_n)/p$ pour tout $n \geq 1$.

5. Comme $\mathbb{P}(X_1 = n | X_1 > 0) = \mathbb{P}(X_2 = n) = f_n$ pour $n \geq 1$, il vient pour $n \geq 1$,

$$b_n = \mathbb{P}(X_1 = n) = \mathbb{P}(X_1 = n | X_1 > 0) \mathbb{P}(X_1 > 0) = f_n(1 - p).$$

On en déduit donc que $b_n = (1 - p)b_{n-1}$, puis $b_n = (1 - p)^n b_0 = (1 - p)^n p$ pour $n \geq 0$ et $f_n = (1 - p)^{n-1} p$ pour $n \geq 1$. La loi de X_2 est la loi géométrique de paramètre p , et X_1 a même loi que $X_2 - 1$.



XIII.3 Variables aléatoires continues

Exercice III.1. 1. La fonction f définie sur \mathbb{R} est positive, mesurable (car continue) et

$$\int_{-\infty}^{+\infty} f(x) dx = \int_0^{+\infty} x e^{-x^2/2} dx = 1.$$

Donc f est une densité de probabilité.

2. Soit g une fonction mesurable bornée. On a

$$\mathbb{E}[g(Y)] = \mathbb{E}[g(X^2)] = \int_0^{+\infty} g(x^2) x e^{-x^2/2} dx = \int_0^{+\infty} g(y) \frac{1}{2} e^{-y/2} dy,$$

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 on a $\mathbb{E}[Y] = 2$ et $\text{Var}(Y) = 4$. ▲

Exercice III.2. La probabilité, p , de toucher la fève correspond à la probabilité que le centre de la fève, qui est uniformément réparti sur le disque de rayon $R-r$ (et de surface $\pi(R-r)^2$) soit à une distance inférieure à r d'un rayon donné. On trouve

$$p = \frac{1}{\pi(R-r)^2} \left[\frac{\pi r^2}{2} + r\sqrt{(R-r)^2 - r^2} + (R-r)^2 \arcsin\left(\frac{r}{R-r}\right) \right],$$

et pour r petit, on obtient $p \sim \frac{2r}{\pi R}$. ▲

Exercice III.3. 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 $2\pi - |\Theta_1 - \Theta_2|$.

1. 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 :

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

La probabilité pour que la part contenant la cerise soit la plus petite est $1/4$.

2. La longueur moyenne de la part contenant la cerise est égale à $2\pi - \mathbb{E}[|\Theta_1 - \Theta_2|]$.

On calcule :

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

La longueur moyenne de la part contenant la cerise est donc $4\pi/3$.

La part qui contient la cerise est plus grande en moyenne et elle est également plus grande dans 75% des cas. Pour voir que ces résultats ne contredisent 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.4. 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 et seulement si aucune des longueurs des morceaux n'est plus grande que la somme des deux autres, ou ce qui revient au même, si et seulement si la longueur de chaque morceau est plus petite que $1/2$. Cela est équivalent aux trois conditions suivantes :

$$1 - \max(X, Y) \leq 1/2, \quad \min(X, Y) \leq 1/2 \quad \text{et} \quad \max(X, Y) - \min(X, Y) \leq 1/2.$$

On obtient :

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

▲

Exercice III.5. 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

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

D'autre part, par indépendance, on a

$$\mathbb{E} \left[e^{-\sum_{n \geq 1} X_n} \right] = \prod_{n \geq 1} \mathbb{E} \left[e^{-X_n} \right] = \prod_{n \geq 1} \frac{\lambda_n}{1 + \lambda_n} = e^{-\sum_{n \geq 1} \log(1 + \lambda_n^{-1})}.$$

On en déduit que $\sum_{n \geq 1} \log(1 + \lambda_n^{-1}) < \infty$. Cela implique $\lim_{n \rightarrow \infty} \lambda_n = \infty$ ainsi que $\sum_{n \geq 1} \lambda_n^{-1} < \infty$. On a donc montré que $3 \Rightarrow 1$. ▲

Exercice III.6. Soit g une fonction mesurable bornée. En utilisant $\varepsilon = \mathbf{1}_{\{\varepsilon=1\}} - \mathbf{1}_{\{\varepsilon=-1\}}$ p.s., puis l'indépendance, on a

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

Donc Z est une variable aléatoire continue dont la loi a pour densité $f(z) = \frac{1}{2} \lambda e^{-\lambda|z|}$, $z \in \mathbb{R}$. ▲

Exercice III.7. 1. Soit $S = X + Y$ et $T = \frac{X}{X + Y}$. Comme $X + Y > 0$ p.s., T est une variable aléatoire réelle bien définie. 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 $\Delta = \{(s, t) \in \mathbb{R}^2; s > 0, 0 < t < 1\}$. De plus la fonction φ est de classe C^1 ainsi que son inverse :

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

La matrice jacobienne de ce changement de variable est :

$$\begin{pmatrix} 1 & 1 \\ \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \end{pmatrix}.$$

La valeur absolue du déterminant de la matrice jacobienne est $|\text{Jac}[\varphi](x, y)| = \frac{1}{x+y}$. On en déduit que $dt ds = |\text{Jac}[\varphi](x, y)| dx dy$ i.e. $s ds dt = dx dy$. On obtient

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

Donc, la densité du couple (S, T) , f , est définie par

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

2. La fonction f est le produit d'une fonction de s et d'une fonction de t . Les variables aléatoires S et T sont donc indépendantes. On obtient la densité de T et de S par la formule des lois marginales. La densité de T est la fonction f_T définie par

$$f_T(t) = \int f(s, t) ds = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} t^{a-1} (1-t)^{b-1} \mathbf{1}_{]0, 1[}(t).$$

On reconnaît la densité de la loi bêta de paramètres a et b . La densité de S est la fonction f_S définie par

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

On reconnaît la densité de la loi gamma de paramètres λ et $a+b$. ▲

Exercice III.8. 1. On remarque que p.s. $X \neq 0$. La variable aléatoire $1/X$ est donc bien définie. Soit g une fonction mesurable bornée. On a :

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

où l'on a fait le changement de variable $y = 1/x$ (C^1 difféomorphisme de \mathbb{R}^* dans lui-même). On obtient donc que $1/X$ est de loi de Cauchy.

2. Comme $Z \neq 0$ p.s., la variable aléatoire Y/Z est bien définie. Soit g une fonction mesurable bornée. On a :

$$\begin{aligned} \mathbb{E}[g(Y/Z)] &= \int_{-\infty}^{+\infty} dz \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \int_{-\infty}^{+\infty} dy \frac{1}{\sqrt{2\pi}} e^{-y^2/2} g(y/z) \\ &= 2 \int_0^{+\infty} dz \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \int_{-\infty}^{+\infty} dy \frac{1}{\sqrt{2\pi}} e^{-y^2/2} g(y/z) \\ &= \frac{1}{\pi} \int_0^{+\infty} dz e^{-z^2/2} \int_{-\infty}^{+\infty} du z e^{-u^2 z^2/2} g(u) \\ &= \frac{1}{\pi} \int_0^{+\infty} dz z e^{-(1+u^2)z^2/2} \int_{-\infty}^{+\infty} du g(u) \\ &= \int_{-\infty}^{+\infty} \frac{1}{\pi} \frac{1}{1+u^2} g(u) du, \end{aligned}$$

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

3. Si X est de loi de Cauchy, alors X a même loi que Y/Z . Donc $1/X$ a même loi que Z/Y . Comme (Z, Y) a même loi que (Y, Z) on en déduit que Y/Z a même loi que Z/Y . On retrouve bien que $1/X$ a même loi que X . ▲

Exercice III.9. Soit g une fonction mesurable bornée.

1. On a :

$$\begin{aligned} \mathbb{E}[g(X_1^2)] &= \int_{-\infty}^{+\infty} g(x^2) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 2 \int_0^{+\infty} g(x^2) \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &= \int_0^{+\infty} g(y) \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2} dy, \end{aligned}$$

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

2. On a :

$$\begin{aligned} \mathbb{E}[g(X_1^2 + X_2^2)] &= \int_{\mathbb{R}^2} dx_1 dx_2 \frac{1}{2\pi} e^{-(x_1^2+x_2^2)/2} g(x_1^2 + x_2^2) \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\theta \int_0^{+\infty} r dr g(r^2) e^{-r^2/2} \\ &= \int_0^{+\infty} g(y) \frac{1}{2} e^{-y/2} dy, \end{aligned}$$

où l'on a utilisé l'indépendance de X_1 et X_2 pour écrire la densité du couple comme produit des densités dans la première égalité, le changement de variable en coordonnées polaires dans la deuxième puis le changement de variable $y = r^2$ sur $]0, \infty[$ dans la dernière. On en déduit que $X_1^2 + X_2^2$ suit la loi $\chi^2(2)$. ▲

Exercice III.10. 1. Soit $D = \{(x_1, \dots, x_n) \in]0, 1[^n; x_i \neq x_j \text{ pour tout } i \neq j\}$. Les ensembles $\Delta_\sigma = \{(x_1, \dots, x_n) \in]0, 1[^n; x_{\sigma(1)} < \dots < x_{\sigma(n)}\}$, 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 :

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

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

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

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 béta $\beta(1, n)$. 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. La variable aléatoire Y étant intégrable, l'espérance conditionnelle $\mathbb{E}[Y|Z]$ a un sens. On a, pour $z \in]0, 1[$:

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

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

4. Soit g une fonction mesurable bornée. 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 mesurable 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 :

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

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

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

▲

Exercice III.11. 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 :

$$f_{X,Y}(x, y) = f_Y(y) f_{X|Y}(x|y) = \frac{1}{\pi} e^{-y(1+x^2)} \mathbf{1}_{\{y > 0\}}.$$

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 X est de loi de Cauchy. Par définition, on a :

$$f_{Y|X}(y|x) = f_{X,Y}(x,y)/f_X(x) = (1+x^2)e^{-y(1+x^2)} \mathbf{1}_{\{y>0\}}.$$

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 III.12. 1. La densité de probabilité f 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 :

$$\mathbb{P}(T \geq 3 | T \geq 1) = \frac{\mathbb{P}(T \geq 3, T \geq 1)}{\mathbb{P}(T \geq 1)} = \frac{\mathbb{P}(T \geq 3)}{\mathbb{P}(T \geq 1)} = \frac{e^{-\frac{9}{2}}}{e^{-\frac{1}{2}}} = e^{-4}.$$

Comme $\mathbb{P}(T \geq 3 | T \geq 1) \neq e^{-2} = \mathbb{P}(T \geq 2)$ la loi n'est pas sans mémoire.

3. Les variables aléatoires X_1, \dots, X_{10} sont indépendantes et suivent une loi de Bernoulli de paramètre :

$$\mathbb{P}(T \leq 1) = F(1) = 1 - e^{-\frac{1}{2}}.$$

On en déduit que la loi de N est la loi binomiale de paramètre $(10, 1 - e^{-\frac{1}{2}})$.

4. La probabilité que l'équipement en série soit défaillant avant 1 an vaut :

$$\mathbb{P}(N \geq 1) = 1 - \mathbb{P}(N = 0) = 1 - e^{-\frac{10}{2}} \simeq 9.9 \cdot 10^{-1} = 99\%.$$

5. La probabilité que l'équipement en parallèle soit défaillant avant 1 an vaut :

$$\mathbb{P}(N = 10) = \left(1 - e^{-\frac{1}{2}}\right)^{10} \simeq 8.9 \cdot 10^{-5}.$$

▲

Exercice III.13. 1. On note Z_k la variable aléatoire réelle donnant la hauteur de la k -ième perforation. Les variables aléatoires $(Z_k, k \geq 1)$ sont indépendantes de même loi uniforme sur $[0, h]$, 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 $n > 0$:

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

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 pourcentage τ_N de liquide qu'on peut espérer conserver sachant le nombre de perforations N , correspond à l'espérance de $\mathbb{E}[Z|N]$ rapportée à la hauteur h du fût. On a $\tau_N = \mathbb{E}[Z|N]/h$. Il vient pour $n \geq 1$:

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

On constate que cette formule couvre aussi le cas où $n = 0$. En effet, dans ce cas on conserve la totalité du liquide (le fût n'est pas perforé). On a donc $\tau_N = 1/(N + 1)$.

3. Soit $Z^{(1)}, \dots, Z^{(n)}$ les hauteurs des perforations les plus basses des fûts de 1 à n . Le pourcentage de la cargaison que l'on sauve est $\frac{1}{n} \sum_{i=1}^n Z^{(i)}$. Si on suppose

les variables aléatoires $Z^{(1)}, \dots, Z^{(n)}$ indépendantes et de même loi que Z , ce pourcentage est, pour n grand, à peu près égal à $\mathbb{E}[Z]/h$ d'après la loi faible (ou forte) des grands nombres. Le pourcentage moyen τ que l'on peut espérer conserver est donc

$$\tau = \frac{\mathbb{E}[Z]}{h} = \frac{1}{h} \mathbb{E}[\mathbb{E}[Z|N]] = \mathbb{E} \left[\frac{1}{N+1} \right] = \sum_{k=0}^{+\infty} \frac{1}{1+k} e^{-\theta} \frac{\theta^k}{k!} = \frac{1 - e^{-\theta}}{\theta}.$$

Pour $\theta = 5$, on obtient $\tau \simeq 19.8\%$. ▲

Exercice III.14. Soit f une fonction mesurable bornée.

1. On note H la distance de J à O . Par une application du théorème de Pythagore, on trouve $L = 2\sqrt{r^2 - H^2}$. Pour calculer sa loi, on utilise la méthode de la fonction muette. Comme H est uniforme sur $[0, r]$, on a :

$$\mathbb{E}[f(L)] = \frac{1}{r} \int_0^r f\left(2\sqrt{r^2 - h^2}\right) dh = \int_0^{2r} f(u) \frac{udu}{4r\sqrt{r^2 - \frac{u^2}{4}}},$$

où l'on a effectué le changement de variable $u = 2\sqrt{r^2 - h^2}$. 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 :

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

où l'on a effectué le changement de variable $u = 2r \sin(t)$. En utilisant le même changement de variable, on calcule :

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

La variance est donc $\text{Var}(L) = \frac{8r^2}{3} - \frac{r^2\pi^2}{4}$.

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 $[0, r]$, il vient $p = \mathbb{P}(H \geq r/2) = 1/2$.

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 abscisses sont uniformes sur $] -\pi, \pi[$. On note $\alpha = \widehat{AOB}$. On a $\alpha = (\theta_A - \theta_B)$ modulo 2π . La longueur de la corde est alors $L = 2r |\sin(\alpha/2)|$. On utilise la méthode de la fonction muette pour calculer la loi de α . On a :

$$\begin{aligned} \mathbb{E}[f(\alpha)] &= \mathbb{E}[f(\theta_A - \theta_B) \mathbf{1}_{]0,2\pi[}(\theta_A - \theta_B) + f(2\pi + \theta_A - \theta_B) \mathbf{1}_{]-2\pi,0[}(\theta_A - \theta_B)] \\ &= \frac{1}{4\pi^2} \int_{]0,2\pi[^2} dq_A dq_B (f(q_A - q_B) \mathbf{1}_{]0,2\pi[}(q_A - q_B) \\ &\quad + f(2\pi + q_A - q_B) \mathbf{1}_{]-2\pi,0[}(q_A - q_B)) \\ &= \frac{1}{4\pi^2} \int_{]0,2\pi[^2} dq_A dq_B (f(q_A - q_B) + f(2\pi - q_A + q_B)) \mathbf{1}_{]0,2\pi[}(q_A - q_B) \\ &= \frac{1}{4\pi^2} \int_{]0,2\pi[} dq_B \int_0^{2\pi - q_B} dt (f(t) + f(2\pi - t)) \\ &= \frac{1}{4\pi^2} \int_{]0,2\pi[} dt (2\pi - t)(f(t) + f(2\pi - t)) = \frac{1}{2\pi} \int_{]0,2\pi[} dt f(t), \end{aligned}$$

où l'on a effectué le changement de variable $t = q_A - q_B$ à q_B fixé. La loi de α est donc la loi uniforme sur $]0, 2\pi[$.

On utilise, encore la méthode de la fonction muette pour déterminer la loi de L :

$$\begin{aligned} \mathbb{E}[f(L)] &= \frac{1}{2\pi} \int_0^{2\pi} dt f(2r |\sin(t/2)|) = \frac{2}{\pi} \int_0^{\pi/2} f(2r \sin(t)) dt \\ &= \int_0^{2r} f(u) \frac{2du}{\pi\sqrt{4r^2 - u^2}}, \end{aligned} \quad (\text{XIII.2})$$

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

Pour calculer l'espérance et la variance de L , on utilise (XIII.2) :

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

et pour la variance :

$$\text{Var}(L) = \mathbb{E}[L^2] - (\mathbb{E}[L])^2 = \frac{8r^2}{\pi} \int_0^{\pi/2} \sin^2(t) dt - \frac{16r^2}{\pi^2} = 2r^2 - \frac{16r^2}{\pi^2}.$$

On a $p = \mathbb{P}(\alpha \geq 2\pi/3) = 1/3$.

3. On calcule la longueur de la corde comme à 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 :

$$\begin{aligned} \mathbb{E}[f(L)] &= \frac{1}{\pi r^2} \int_{-r}^r dx \int_{-r}^r dy f\left(2\sqrt{r^2 - x^2 - y^2}\right) \mathbf{1}_{\{x^2 + y^2 \leq r^2\}} \\ &= \frac{1}{\pi r^2} \int_0^r \rho d\rho \int_{-\pi}^{\pi} d\theta f\left(2\sqrt{r^2 - \rho^2}\right) \\ &= \frac{1}{2r^2} \int_0^{2r} f(u) u du. \end{aligned}$$

où on a effectué un changement de coordonnées polaires dans \mathbb{R}^2 , puis le changement de variable $u = 2\sqrt{r^2 - \rho^2}$. La variable L est continue de densité $\frac{u}{2r^2} \mathbf{1}_{]0, 2r[}(u)$.

La moyenne et la variance de cette loi se calculent immédiatement :

$$\mathbb{E}[L] = \frac{1}{2r^2} \int_0^{2r} u^2 du = \frac{4r}{3},$$

$$\text{Var}(L) = \frac{1}{2r^2} \int_0^{2r} u^3 du - \frac{16r^2}{9} = 2r^2 - \frac{16r^2}{9} = \frac{2r^2}{9}.$$

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

$$p = \frac{1}{2r^2} \int_{\sqrt{3}r}^{2r} u du = \frac{1}{2r^2} \frac{4r^2 - 3r^2}{2} = \frac{1}{4}.$$

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 | 0.20 | 1/2 |
| 2 | 1.27 | 0.38 | 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) a pour densité :

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

2. On considère le changement de variables en coordonnées polaires. Vérifions qu'il s'agit d'un C^1 -difféomorphisme. On considère la fonction φ définie sur $\Delta = \{(r, \theta) \in \mathbb{R}^2; r > 0, -\pi < \theta < \pi\}$:

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

Il est immédiat de vérifier que φ est une bijection de Δ dans $\mathcal{D} = \mathbb{R}^2 \setminus (\mathbb{R}^- \times \{0\})$, dont l'inverse a pour expression :

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

avec la convention que $\text{sgn}(0) = 1$. La matrice jacobienne de φ existe pour tout $(r, \theta) \in \Delta$ et ses éléments sont des fonctions C^1 de (r, θ) :

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

On a également que φ^{-1} est C^1 sur $\mathcal{D} \setminus (]0, \infty[\times \{0\})$ de matrice jacobienne

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

La vérification du caractère C^1 de φ^{-1} sur $]0, \infty[\times \{0\}$ se fait par une étude directe. Pour $x > 0$, on a $\varphi^{-1}(x, 0) = (x, \operatorname{sgn}(0) \arccos(1)) = (x, 0)$ et :

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

Ceci implique à la fois la continuité, la différentiabilité et la continuité des dérivées partielles de φ^{-1} en $(x, 0)$. On a donc vérifié que le changement de variables en coordonnées polaires est un C^1 -difféomorphisme de \mathcal{D} dans Δ .

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

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

Comme $dx dy = |\operatorname{Jac}[\varphi](r, \theta)| dr d\theta = r dr d\theta$, il vient :

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

On en déduit que (R, Θ) est une v.a.c. de densité :

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

3. 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 indépendants. La loi de Θ est la loi uniforme sur $[-\pi, \pi]$ et la loi de R a pour densité f_R .

4. L'espérance de R vaut :

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

On a $\mathbb{E}[R^2] = \mathbb{E}[X^2 + Y^2] = 2\sigma^2$. La variance de R vaut donc $\text{Var}(R) = (2 - \pi/2)\sigma^2$. ▲

Exercice III.16. 1. La direction θ est à valeurs dans $[-\pi/2, \pi/2[$. 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 $|X| > \frac{d}{2} - \frac{\ell}{2} \cos(\theta)$.

2. Au vu des hypothèses, il est naturel de supposer 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 :

$$\begin{aligned}\mathbb{P}(|X| > \frac{d}{2} \cos(\theta) - \frac{\ell}{2}) &= \int \mathbf{1}_{\{|x| > \frac{d}{2} - \frac{\ell}{2} \cos(\theta)\}} \frac{1}{\pi d} \mathbf{1}_{[-d/2, d/2] \times [-\pi/2, \pi/2]}(x, \theta) dx d\theta \\ &= 2 \int_{-\pi/2}^{\pi/2} \left(\int_{\frac{d}{2} - \frac{\ell}{2} \cos(\theta)}^{\frac{d}{2}} \frac{1}{\pi d} dx \right) d\theta \\ &= \frac{2}{\pi d} \int_{-\pi/2}^{\pi/2} \frac{\ell}{2} \cos \theta d\theta \\ &= \frac{2\ell}{\pi d}.\end{aligned}$$

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 $p = 2\ell/\pi d$. D'après la loi faible des grands nombres, la moyenne empirique $\frac{N_n}{n}$ converge en probabilité vers l'espérance de Y_1 , c'est-à-dire vers $2\ell/\pi d$. En fait la loi forte des grands nombres assure que cette convergence est presque sûre. On a ainsi un moyen expérimental de calculer une approximation de $1/\pi$ et donc de π .

4. On veut trouver n tel que $\mathbb{P}(|\frac{N_n}{n} - \frac{1}{\pi}| > 10^{-2}) \leq 5\%$. On déduit de l'inégalité de Tchebychev que

$$\mathbb{P}\left(\left|\frac{N_n}{n} - \frac{1}{\pi}\right| > a\right) \leq \frac{\mathbb{E}\left[\left(\frac{N_n}{n} - \frac{1}{\pi}\right)^2\right]}{a^2} \leq \frac{\frac{1}{\pi}\left(1 - \frac{1}{\pi}\right)}{na^2}.$$

On trouve $n = 43\,398$. 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 lancers 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 ($113/355$ 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 suggère que le choix de 355 lancers et peut-être même du nombre "aléatoire" de 113 lancers avec intersection est intentionnel.



Exercice III.17. 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 : $X = 0$ s'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 indépendants. 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 aléatoires S et T indépendantes et de même loi de fonction de répartition F . 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 $G = \{Z \geq \min(S, T), X = 0\} \cup \{Z \leq \max(S, T), X = 1\}$. En utilisant l'indépendance de S, T, Z et X , il vient

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

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 en utilisant l'indépendance entre S et T :

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

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. (Ce résultat est effectivement vrai, mais nous ne le démontrons pas en toute généralité.) 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 p.s. $Z = x_{1/2}$.

Vérifions que $Z = x_{1/2}$ est optimal. Un calcul direct donne, si $Z = x_{1/2}$ p.s.,

$$\mathbb{P}(G) = \frac{1}{2} + \frac{1}{2} \mathbb{P}(\min(S, T) \leq x_{1/2} \leq \max(S, T)) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

Il reste 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 qu'il existe $\varepsilon > 0$ et $\eta > 0$, tels que $\mathbb{P}(|Z - x_{1/2}| > \eta) > \varepsilon$. On décompose, pour n fixé, suivant $Z \in [\frac{k}{n}, \frac{k+1}{n}[$, pour $k \in \mathbb{N}$. On calcule :

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

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 :

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

En laissant n tendre vers l'infini, et en utilisant le fait que F est continue, on obtient :

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

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 $\mathbb{P}(G)$ est maximale si et seulement si p.s. $Z = x_{1/2}$. Et dans ce cas on a $\mathbb{P}(G) = 3/4$.

▲

Exercice III.18. 1. La densité du triplet (X, Y, Z) étant invariante par rotation, elle est de la forme $f_{(X, Y, Z)}(x, y, z) = \phi(x^2 + y^2 + z^2)$ avec ϕ une fonction positive définie sur \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

$$f(x)f(y)f(z) = \phi(x^2 + y^2 + z^2), \quad \text{pour tout } (x, y, z) \in \mathbb{R}^3$$

implique $f(x)f'(y)f(z) = 2y\phi'(x^2 + y^2 + z^2)$ et $f'(x)f(y)f(z) = 2x\phi'(x^2 + y^2 + z^2)$. On en déduit que $xf(x)f'(y)f(z) = f'(x)yf(y)f(z)$ pour tout $(x, y, z) \in \mathbb{R}^3$. Comme f est une densité sur \mathbb{R} , f est non constante. En particulier il existe $y_0 \neq 0, z_0 \in \mathbb{R}$ tels que $f(y_0)f(z_0) > 0$. On en déduit donc qu'il existe une constante c telle que pour tout $x \in \mathbb{R}$, $f'(x) = 2cx f(x)$. Ceci implique que $f(x) = a e^{cx^2}$ pour une certaine constante a . Comme f est une densité de probabilité, on a $f \geq 0$ et $\int_{\mathbb{R}} f(x) dx = 1$. On en déduit que

$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$ avec σ une constante strictement positive. Finalement, les variables aléatoires X, Y et Z suivent une loi gaussienne centrée.

2. L'énergie cinétique moyenne est définie par $E_c = \frac{1}{2} m \mathbb{E}[|V|^2] = \frac{1}{2} m \mathbb{E}[V_1^2 + V_2^2 + V_3^2]$. Comme $\mathbb{E}[V_1^2] = \sigma^2$, on en déduit que $E_c = \frac{3}{2} m \sigma^2$. On obtient $\sigma^2 = \frac{kT}{m}$.
3. La loi de $X + Y$ est la loi $\Gamma(\lambda, a + b)$, cf. l'exercice (III.7).
4. À l'aide de la méthode de la fonction muette, on montre que V_i^2 suit 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 $\frac{1}{\sqrt{2\pi}\sigma^3} \sqrt{z} e^{-z/2\sigma^2} \mathbf{1}_{\{z>0\}}$. À l'aide de la méthode de la fonction muette, on montre que $|V|$ est une variable continue de densité

$$\frac{\sqrt{2}}{\sqrt{\pi}} \left(\frac{m}{kT}\right)^{3/2} v^2 e^{-mv^2/2kT} \mathbf{1}_{\{v>0\}}.$$

Il s'agit de la densité de la loi de Maxwell.

▲

Exercice III.19. 1. On utilise la méthode de la fonction muette. Soit g une fonction mesurable bornée. On a avec $U = \sqrt{XY}$ et $V = \sqrt{Y}$:

$$\begin{aligned} \mathbb{E}[g(U, V)] &= \mathbb{E}[g(\sqrt{XY}, \sqrt{Y})] \\ &= \int g(\sqrt{xy}, \sqrt{y}) f_{X,Y}(x, y) \, dx dy \\ &= \int g(\sqrt{xy}, \sqrt{y}) f_X(x) f_Y(y) \, dx dy \\ &= \frac{\lambda^{2\alpha+1/2}}{\Gamma(\alpha)\Gamma(\alpha+1/2)} \int_{]0, \infty[^2} g(\sqrt{xy}, \sqrt{y}) (xy)^{\alpha-1/2} e^{-\lambda(x+y)} \frac{dx dy}{\sqrt{x}}, \end{aligned}$$

où, pour la deuxième égalité, on a utilisé que, par indépendance, la densité de la loi de (X, Y) est le produit des densités des lois de X et de Y .

On considère le changement de variable $u = \sqrt{xy}$, $v = \sqrt{y}$ qui est un C^1 difféomorphisme de $]0, \infty[^2$ dans lui-même. Le calcul du déterminant du jacobien de la transformation donne :

$$dudv = \frac{dx dy}{4\sqrt{x}}.$$

On en déduit donc que :

$$\mathbb{E}[g(\sqrt{XY}, \sqrt{Y})] = \frac{4\lambda^{2\alpha+1/2}}{\Gamma(\alpha)\Gamma(\alpha+1/2)} \int_{]0, \infty[^2} g(u, v) u^{2\alpha-1} e^{-\lambda(v^2 + \frac{u^2}{v^2})} dudv.$$

Ainsi la densité de la loi de (U, V) est :

$$f_{(U, V)}(u, v) = \frac{4\lambda^{2\alpha+1/2}}{\Gamma(\alpha)\Gamma(\alpha+1/2)} u^{2\alpha-1} e^{-\lambda(v^2 + \frac{u^2}{v^2})} \mathbf{1}_{\{u>0, v>0\}}.$$

2. Par la formule des lois marginales, la densité, f_U , de la loi de U est :

$$\begin{aligned} f_U(u) &= \int f_{(U, V)}(u, v) dv = \frac{4\lambda^{2\alpha+1/2}}{\Gamma(\alpha)\Gamma(\alpha+1/2)} \frac{\sqrt{\pi}}{2\sqrt{\lambda}} u^{2\alpha-1} e^{-2\lambda u} \mathbf{1}_{\{u>0\}} \\ &= \frac{2\sqrt{\pi}}{\Gamma(\alpha)\Gamma(\alpha+1/2)} \lambda^{2\alpha} u^{2\alpha-1} e^{-2\lambda u} \mathbf{1}_{\{u>0\}}. \end{aligned}$$

3. La densité est égale à $\lambda^{2\alpha} u^{2\alpha-1} e^{-2\lambda u} \mathbf{1}_{\{u>0\}}$, à une constante multiplicative près. Il s'agit donc de la densité de la loi $\Gamma(2\lambda, 2\alpha)$. En particulier, cette constante multiplicative est égale à $2^{2\alpha}/\Gamma(2\alpha)$. On en déduit alors la formule de duplication de la fonction Γ .

Il est facile de démontrer que $h(u) = \int_0^\infty e^{-\lambda(v^2 + \frac{u^2}{v^2})} dv$ est en fait égal à $\frac{\sqrt{\pi}}{2\sqrt{\lambda}} e^{-2\lambda u}$. En effet, on a :

$$h'(u) = -2\lambda u \int_0^\infty e^{-\lambda(v^2 + \frac{u^2}{v^2})} \frac{dv}{v^2} = -2\lambda \int_0^\infty e^{-\lambda(\frac{u^2}{w^2} + w^2)} dw = -2\lambda h(u),$$

où l'on a fait le changement de variable $w = u/v$. Ainsi $h(u) = h(0) e^{-2\lambda u}$. Il reste à calculer $h(0)$. On a :

$$h(0) = \frac{1}{2} \int_{-\infty}^\infty e^{-\lambda v^2} dv = \frac{\sqrt{\pi}}{2\sqrt{\lambda}} \int_{-\infty}^\infty e^{-v^2/2(1/2\lambda)} \frac{dv}{\sqrt{2\pi(1/2\lambda)}} = \frac{\sqrt{\pi}}{2\sqrt{\lambda}},$$

où l'on a utilisé la parité de la fonction intégrée pour la première égalité et le fait que l'intégrale de la densité de la loi $\mathcal{N}(0, 1/2\lambda)$ vaut 1 pour la dernière égalité. ▲

XIII.4 Fonctions caractéristiques

Exercice IV.1. 1. 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 $\Gamma(\lambda, \alpha_1 + \alpha_2)$.

2. Par récurrence, on déduit de la question précédente que la loi de $\sum_{i=1}^n X_i$ est la loi $\Gamma(\lambda, n)$. Si Z est de loi $\Gamma(a, b)$ alors pour $c > 0$, on a :

$$\psi_{cZ}(u) = \left(\frac{a}{a - icu} \right)^b = \left(\frac{a/c}{a/c - iu} \right)^b.$$

Ainsi cZ est de loi $\Gamma(a/c, b)$. On en déduit que \bar{X}_n est de loi $\Gamma(n\lambda, n)$. ▲

Exercice IV.2.

1. La densité de la loi de Z , f_Z , a été calculée dans l'exercice III.6 : $f_Z(z) = \frac{\lambda}{2} e^{-\lambda|z|}$. On utilise la formule de décomposition et l'indépendance entre Y et ε pour obtenir

$$\begin{aligned} \psi_Z(u) &= \mathbb{E} [e^{iuY} \mathbf{1}_{\{\varepsilon=1\}}] + \mathbb{E} [e^{-iuY} \mathbf{1}_{\{\varepsilon=-1\}}] \\ &= \frac{1}{2} \left[\frac{\lambda}{\lambda - iu} + \overline{\left(\frac{\lambda}{\lambda - iu} \right)} \right] \\ &= \frac{\lambda^2}{\lambda^2 + u^2}. \end{aligned}$$

2. On remarque que $\frac{1}{\lambda\pi} \psi_Z$ 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 p.p. $f_Z(z) = \int_{\mathbb{R}} e^{-iuz} \psi_Z(u) \frac{du}{2\pi}$. Comme les membres de droite et de gauche sont des fonctions continues, on a l'égalité pour tout $z \in \mathbb{R}$. On a donc $\frac{\lambda}{2} e^{-\lambda|z|} = \int_{\mathbb{R}} e^{-iuz} \frac{\lambda^2}{\lambda^2 + u^2} \frac{du}{2\pi}$. On en déduit ainsi la fonction caractéristique, ψ , de la loi de Cauchy de paramètre λ : pour tout $z \in \mathbb{R}$,

$$\psi(z) = \int_{\mathbb{R}} e^{iuz} \frac{1}{\pi} \frac{\lambda}{\lambda^2 + u^2} du = e^{-\lambda|z|}. \quad \blacktriangle$$

Exercice IV.3. 1. Par définition on a $\mathbb{E}[\varphi(V, W) | W] = h(W)$, 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 variables aléatoires 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 déduit de la question précédente que :

$$\mathbb{E} [e^{iuY}] = \mathbb{E} [e^{iuX_1X_4 - iuX_2X_3} \mid X_1, X_2] = h(X_1, X_2),$$

où $h(x_1, x_2) = \mathbb{E}[e^{iux_1X_4 - iux_2X_3}]$. En utilisant l'indépendance puis la fonction caractéristique de la loi $\mathcal{N}(0, 1)$, il vient :

$$h(x_1, x_2) = \mathbb{E}[e^{iux_1X_4}]\mathbb{E}[e^{-iux_2X_3}] = e^{-\frac{u^2x_1^2}{2} - \frac{u^2x_2^2}{2}}.$$

On a , en utilisant l'indépendance :

$$\begin{aligned} \mathbb{E} [e^{iuY}] &= \mathbb{E} [e^{iuX_1X_4 - iuX_2X_3}] = \mathbb{E} [\mathbb{E} [e^{iuX_1X_4 - iuX_2X_3} \mid X_1, X_2]] \\ &= \mathbb{E} [h(X_1, X_2)] \\ &= \mathbb{E} \left[e^{-\frac{u^2X_1^2}{2} - \frac{u^2X_2^2}{2}} \right] \\ &= \mathbb{E} \left[e^{-\frac{u^2X_1^2}{2}} \right]^2 \\ &= \left(\int_{\mathbb{R}} e^{-\frac{u^2x^2}{2} - \frac{x^2}{2}} \frac{dx}{\sqrt{2\pi}} \right)^2 \\ &= \left[\frac{1}{\sqrt{u^2 + 1}} \right]^2 \\ &= \frac{1}{u^2 + 1}. \end{aligned}$$

On reconnaît la fonction caractéristique de la loi exponentielle symétrique (cf exercice IV.2). ▲

Exercice IV.4. 1. En utilisant l'espérance conditionnelle, on a pour $n \in \mathbb{N}$:

$$\mathbb{E}[S_N \mid N = n] = \mathbb{E}[S_n \mid N = n] = \mathbb{E}[S_n] = n\mathbb{E}[X_1].$$

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

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

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

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

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

2. De la même manière, on a :

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

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

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

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

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

On retrouve ainsi les résultats de la question précédente. ▲

Exercice IV.5. On note $\Im(z)$ la partie imaginaire de $z \in \mathbb{C}$.

1. Si X est symétrique alors $\Im(\psi_X(u)) = \frac{1}{2}(\psi_X(u) - \bar{\psi}_X(u)) = \frac{1}{2}(\mathbb{E}[e^{iuX}] - \mathbb{E}[e^{-iuX}]) = 0$. Si $\Im(\psi_X(u)) = 0$ alors le calcul précédent montre que les fonctions caractéristiques de X et $-X$ coïncident, donc X et $-X$ sont égales en loi.
2. Si Y est indépendant de X et de même loi que X , alors la fonction caractéristique de $X - Y$ est $\psi_X(u)\bar{\psi}_X(u) = |\psi_X(u)|^2$.
3. La loi de 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.6. 1. Soit $u \in \mathbb{R}^d$. On note $\langle \cdot, \cdot \rangle$ le produit scalaire de \mathbb{R}^d . Comme la loi du vecteur est invariante par rotation, on en déduit que $\langle u, X \rangle / \|u\|$ a même loi que X_1 . On a par indépendance et en utilisant que X_1, \dots, X_d ont même loi (grâce à l'invariance de la loi de X par rotation)

$$\psi_X(u) = \mathbb{E}[e^{i\langle u, X \rangle}] = \prod_{k=1}^d \psi_{X_1}(u_k).$$

On a également $\psi_X(u) = \psi_{\langle u, X \rangle / \|u\|}(\|u\|) = \psi_{X_1}(\|u\|)$. Comme X_1 est $-X_1$ ont même loi (car la loi de X est invariante par rotation), on a $\psi_{X_1}(u) = \psi_{-X_1}(u) = \psi_{X_1}(-u) = \bar{\psi}_{X_1}(u)$. En particulier la fonction ψ_{X_1} est réelle symétrique. On pose $g(x) = \psi_{X_1}(\sqrt{x})$ pour $x \geq 0$. On en déduit que la fonction réelle g vérifie (IV.1).

2. Rappelons que les fonctions caractéristiques sont continues, prennent la valeur 1 en 0, sont en valeurs absolue inférieures à 1 et caractérisent la loi. En particulier g est continue, $g(0) = 1$ et $|g(v_1)| \leq 1$. En intégrant (IV.1) en v_2 , on en déduit que $g(v_1)$ est dérivable, puis que g'/g est constant et donc $g(v_1) = \alpha e^{\beta v_1}$. Les conditions $g(0) = 1$ et $|g(v_1)| \leq 1$ impliquent $\alpha = 0$ et $\beta = -\sigma^2/2$ où $\sigma \geq 0$. On en déduit que $\psi_{X_1}(u_1) = e^{-\sigma^2 u_1^2/2}$. Ceci implique que :
- Si $\sigma = 0$ alors p.s. $X_1 = 0$ et donc p.s. $X_k = 0$ pour tout $k \in \{1, \dots, d\}$.
 - Si $\sigma > 0$ alors X_1 , et donc X_2, \dots, X_n , sont de loi gaussienne centrée de variance σ^2 .

▲

Exercice IV.7. 1. En utilisant les fonctions caractéristiques, il vient par indépendance, pour $u \in \mathbb{R}$:

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

On en déduit que la loi de $X - Y$ est la loi gaussienne de moyenne $\mu = \mu_X - \mu_Y$ et de variance $\sigma^2 = \sigma_X^2 + \sigma_Y^2$.

2. On utilise la méthode de la fonction muette. Soit g une fonction mesurable bornée, on a, en notant

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

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

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

On en déduit que Z est une variable aléatoire continue et la densité de sa loi est donnée par $[f(v) + f(-v)]\mathbf{1}_{\{v>0\}}$.

3. On remarque que Z est égal en loi à $|\sigma G + \mu|$, où G est une variable aléatoire de loi gaussienne $\mathcal{N}(0, 1)$. En particulier, on a :

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

où Φ est la fonction de répartition de la loi $\mathcal{N}(0, 1)$, et

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

Enfin, la variance de Z est égale à $\mathbb{E}[Z^2] - \mathbb{E}[Z]^2$.



XIII.5 Théorèmes limites

Exercice V.1. Soit g continue bornée.

1. On a $\mathbb{E}[g(X_n)] = \int_0^\infty \lambda_n e^{-\lambda_n x} g(x) dx$. 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 $|\lambda_n e^{-\lambda_n x} g(x)| \leq \|g\|_\infty \lambda_+ e^{-\lambda_- x} = h(x)$. La fonction h est intégrable sur $[0, \infty[$. On a $\lim_{n \rightarrow \infty} \lambda_n e^{-\lambda_n x} g(x) = \lambda e^{-\lambda x} g(x)$. On déduit du théorème de convergence dominée que :

$$\mathbb{E}[g(X_n)] \xrightarrow{n \rightarrow \infty} \int_0^\infty \lambda e^{-\lambda x} g(x) dx.$$

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

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

$$\mathbb{E}[g(X_n)] \xrightarrow{n \rightarrow \infty} g(0) = \mathbb{E}[g(X)],$$

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

3. Si la suite $(X_n, n \in \mathbb{N}^*)$ converge en loi vers une variable aléatoire X , alors les fonctions caractéristiques $\psi_{X_n}(u)$ convergent vers $\psi_X(u)$ pour tout $u \in \mathbb{R}$. On a :

$$\lim_{n \rightarrow \infty} \psi_{X_n}(u) = \lim_{n \rightarrow \infty} \frac{\lambda_n}{\lambda_n - iu} = \mathbf{1}_{\{u=0\}}.$$

La fonction $u \mapsto \mathbf{1}_{\{u=0\}}$ n'est pas continue en 0. Or les fonctions caractéristiques sont continues. Par contraposée, ce n'est donc pas la fonction caractéristique d'une variable aléatoire et la suite $(X_n, n \in \mathbb{N}^*)$ ne converge pas en loi. ▲

Exercice V.2. On a :

$$\psi_{\frac{T_n}{n}}(u) = \psi_{T_n}\left(\frac{u}{n}\right) = \frac{\theta e^{iu/n}}{n - (n - \theta)e^{iu/n}}.$$

Rappelons que $e^{iu/n} = 1 + \frac{iu}{n} + o\left(\frac{1}{n}\right)$. On en déduit que :

$$\psi_{\frac{T_n}{n}}(u) = \frac{\theta + o(1)}{\theta - iu + o(1)} \xrightarrow{n \rightarrow \infty} \frac{\theta}{\theta - iu}.$$

On reconnaît dans le membre 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 E de loi exponentielle de paramètre $\theta > 0$. ▲

Exercice V.3. 1. La suite $(M_n, n \geq 1)$ est croissante et bornée p.s. par θ . Elle converge donc p.s. vers une limite M . Soit $\varepsilon \in]0, \theta]$, on a :

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

Donc on a :

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

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

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

$$\mathbb{P}(n(\theta - M_n) \leq a) = \mathbb{P}(M_n > \theta - \frac{a}{n}) = 1 - \left(1 - \frac{a}{\theta n}\right)^n \xrightarrow{n \rightarrow \infty} 1 - e^{-\frac{a}{\theta}}.$$

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

Exercice V.4. On rappelle que $\psi_{X_n}(u) = e^{-a|u|}$.

1. On a :

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

où l'on a utilisé l'indépendance des variables aléatoires pour la deuxième égalité. Donc on en déduit que :

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

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 :

$$\psi_{S_n/n^2}(u) = \psi_{S_n}(u/n^2) = (e^{-a|u|/n^2})^n = e^{-a|u|/n}.$$

Donc on en déduit que :

$$\psi_{S_n/n^2}(u) \xrightarrow[n \rightarrow \infty]{} 1.$$

La suite converge en loi vers la variable aléatoire constante égale à 0.

Un résultat général assure que la convergence en loi vers une constante implique la convergence en probabilité vers cette constante. On en donne une preuve élémentaire. Soit $(Z_n, n \geq 1)$ une suite de variables aléatoires qui converge en loi vers une constante c . Quitte à considérer $Z_n - c$, on peut supposer que $c = 0$. Soit $\varepsilon > 0$. On a $\mathbb{P}(|Z_n| \geq \varepsilon) \leq \mathbb{E}[g(Z_n)]$, où $g(z) = \min(|z|/\varepsilon, 1)$. La fonction g est continue bornée. On déduit de la convergence en loi que $\lim_{n \rightarrow \infty} \mathbb{E}[g(Z_n)] = \mathbb{E}[g(0)] = 0$. Ceci implique que $\lim_{n \rightarrow \infty} \mathbb{P}(|Z_n| \geq \varepsilon) = 0$. (Plus directement, on peut dire que 0 est un point de continuité de la fonction $\mathbf{1}_{\{|z| \geq \varepsilon\}}$, et donc $\lim_{n \rightarrow \infty} \mathbb{P}(|Z_n| \geq \varepsilon) = \lim_{n \rightarrow \infty} \mathbb{E}[\mathbf{1}_{\{|Z_n| \geq \varepsilon\}}] = \mathbb{E}[\mathbf{1}_{\{0 \geq \varepsilon\}}] = 0$.) Ainsi la suite $(Z_n, n \geq 1)$ converge en probabilité vers 0.

La suite $(S_n/n^2, n \geq 1)$ converge donc en probabilité vers 0.

3. On a :

$$\psi_{S_n/n}(u) = \psi_{S_n}(u/n) = (e^{-a|u|/n})^n = e^{-a|u|}.$$

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 :

$$\frac{S_{2n}}{2n} - \frac{S_n}{n} = \frac{X_{n+1} + \dots + X_{2n}}{2n} - \frac{X_1 + \dots + X_n}{2n}.$$

On a donc par indépendance, puis en utilisant le fait que $\frac{X_{n+1} + \dots + X_{2n}}{2n}$ et $\frac{X_1 + \dots + X_n}{2n}$ ont même loi que $S_n/2n$, que :

$$\psi_{\left(\frac{S_{2n}-S_n}{2n}\right)}(u) = \psi_{\frac{X_{n+1}+\dots+X_{2n}}{2n}}(u)\psi_{\frac{X_1+\dots+X_n}{2n}}(u) = \psi_{S_n/2n}(u)^2 = e^{-a|u|}.$$

Donc pour tout n , $\frac{S_{2n}}{2n} - \frac{S_n}{n}$ est une variable aléatoire de Cauchy de paramètre a .

Raisonnons par l'absurde pour la convergence en probabilité. Si $\left(\frac{S_n}{n}, n \geq 1\right)$ convergerait 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 :

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

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

Exercice V.5. 1. On déduit du TCL que la suite $(Z_n, n \in \mathbb{N}^*)$ converge en loi vers σG , où G est de loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.

2. On a :

$$Z_{2n} - Z_n = \left(\frac{1}{\sqrt{2}} - 1\right) Z_n + \frac{1}{\sqrt{2}} Z'_n,$$

où $Z'_n = \frac{1}{\sqrt{n}} \sum_{k=n+1}^{2n} (X_k - \mu)$ est indépendant de Z_n et a même loi que Z_n . En

utilisant l'indépendance, il vient :

$$\begin{aligned} \psi_{Z_{2n}-Z_n}(u) &= \psi_{\left(\frac{1}{\sqrt{2}}-1\right)Z_n} \psi_{\frac{1}{\sqrt{2}}Z'_n}(u) \\ &= \psi_{Z_n} \left(\left(\frac{1}{\sqrt{2}} - 1 \right) u \right) \psi_{Z_n} \left(\frac{1}{\sqrt{2}} u \right) \\ &\xrightarrow{n \rightarrow \infty} \psi_{\sigma G} \left(\left(\frac{1}{\sqrt{2}} - 1 \right) u \right) \psi_{\sigma G} \left(\frac{1}{\sqrt{2}} u \right) \\ &= \exp \left(-\frac{u^2}{2} \sigma^2 \left(\left(\frac{1}{\sqrt{2}} - 1 \right)^2 + \frac{1}{2} \right) \right). \end{aligned}$$

On en déduit que la suite $(Z_{2n} - Z_n, n \in \mathbb{N}^*)$ converge en loi vers $\sigma\sqrt{2}\sqrt{1 - \frac{1}{\sqrt{2}}}G$.

3. Si la suite $(Z_n, n \in \mathbb{N}^*)$ converge en probabilité, alors la suite $(Z_{2n} - Z_n, n \in \mathbb{N}^*)$ converge en probabilité vers 0. On déduit alors de la question précédente que $\sigma = 0$ (et donc $X_k = 0$ p.s et $Z_n = 0$ p.s.). Par contraposé, si $\sigma > 0$, alors la convergence du TCL n'est pas une convergence en probabilité. ▲

Exercice V.6. 1. 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 Bernoulli de paramètre $1/2$. En particulier $\mathbb{E}[X_i] = 1/2$ et $\text{Var}(X_i) = 1/4$. La loi de S_m est une loi binomiale $\mathcal{B}(m, 1/2)$. On note :

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

On remarque que si m est grand, alors, d'après le théorème central limite, la loi de $\frac{S_m - \frac{1}{2}m}{\frac{1}{2}\sqrt{m}}$ est asymptotiquement la loi gaussienne centrée. On a alors :

$$p_m \simeq \mathbb{P}(|G| > \sqrt{m}0.1),$$

où G est de loi $\mathcal{N}(0, 1)$ gaussienne centrée.

Les fréquences empiriques de pile lors de n séries de m tirages sont égales en loi à S_m/m . Notons que les variables aléatoires F_1, \dots, F_n sont indépendantes et de même loi. On a que, pour tout i , $\mathbb{P}(F_i \notin]0.45, 0.55[) = p_m$. En particulier, pour $m = 400$ tirages, on a à l'aide des tables de la fonction de répartition de la loi normale $\mathcal{N}(0, 1)$:

$$\mathbb{P}(F_i \notin]0.45, 0.55[) = p_{400} \simeq \mathbb{P}(|G| \geq 2) \simeq 0.05.$$

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

$$\xi_i = \begin{cases} 1 & \text{si } F_i \notin]0.45, 0.55[, \\ 0 & \text{sinon.} \end{cases}$$

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_m) .

2. Comme p_m est petit, la loi binomiale de paramètre (n, p_m) peut être approchée par la loi de Poisson de paramètre $np_m = 20p_{400} \simeq 1$ (on a la convergence en loi de la loi binomiale de paramètre (n, p) vers la loi de Poisson de paramètre θ quand n tend vers l'infini et np vers $\theta > 0$). En particulier, on a :

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

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

Exercice V.7. 1. On a $\mathbb{P}(\exists k \in \{1, \dots, n\}; U_k \leq 0) \leq \sum_{k=1}^n \mathbb{P}(U_k \leq 0) = 0$. La variable aléatoire X_n est donc p.s. strictement positive. On peut donc considérer $V_n = \log(X_n) = \frac{1}{n} \sum_{k=1}^n \alpha \log(U_k)$. Les variables aléatoires $(\alpha \log(U_k), k \in \mathbb{N}^*)$ sont indépendantes et de même loi. Elles sont également intégrables :

$$\mathbb{E}[|\log(U_1)|] = \int |\log(u)| \mathbf{1}_{]0,1[}(u) du = - \int_{]0,1[} \log(u) du = 1.$$

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 $e^{-\alpha}$.

2. On a :

$$\log(Y_n) = \sqrt{n}(\log(X_n) + \alpha) = \sqrt{n} \left(\frac{1}{n} \sum_{k=1}^n \alpha \log(U_k) - \mathbb{E}[\alpha \log(U_k)] \right).$$

Vérifions que les variables aléatoires $\alpha \log(U_k)$ sont de carré intégrable. On a :

$$\mathbb{E}[\log(U_1)^2] = \int \log(u)^2 \mathbf{1}_{]0,1[}(u) du = [u \log(u)^2 - 2u \log(u) + 2u]_0^1 = 2.$$

On déduit du théorème central limite que la suite $(\log(Y_n), n \in \mathbb{N}^*)$ converge en loi vers Z de loi gaussienne $\mathcal{N}(0, \sigma^2)$, où $\sigma^2 = \text{Var}(\alpha \log(U_1)) = \alpha^2(2 - 1) = \alpha^2$. 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 $Y = e^Z$. On utilise la méthode de la fonction muette. Soit g une fonction mesurable bornée. On a :

$$\begin{aligned} \mathbb{E}[g(e^Z)] &= \int g(e^z) f_Z(z) dz = \int g(e^z) e^{-z^2/2\alpha^2} \frac{dz}{\sqrt{2\pi\alpha^2}} \\ &= \int g(y) \frac{1}{y\sqrt{2\pi\alpha^2}} e^{-\log(y)^2/2\alpha^2} \mathbf{1}_{]0,\infty[}(y) dy, \end{aligned}$$

où l'on a effectué le changement de variable $y = e^z$ de \mathbb{R} dans $]0, \infty[$. On en déduit donc que $Y = e^Z$ est une variable aléatoire continue de densité $\frac{1}{y\sqrt{2\pi\alpha^2}} e^{-\log(y)^2/2\alpha^2} \mathbf{1}_{]0, \infty[}(y)$. La loi de Y est appelée loi log-normale. ▲

Exercice V.8. 1. On considère pour Ω l'ensemble des résultats des tirages de n boules. Les tirages sont équiprobables. On a $\text{Card}(\Omega) = \binom{N}{n}$. Parmi ces tirages, on considère ceux qui donnent k boules blanches. Il faut donc choisir k boules blanches parmi les m boules blanches et $n - k$ boules noires parmi les $N - m$ boules noires. Il existe donc $\binom{m}{k} \binom{N-m}{n-k}$ configurations possibles. On en déduit que $\mathbb{P}(X_N = k) = \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}}$.

On peut également considérer pour Ω l'ensemble des résultats des tirages exhaustifs de N boules, et regarder les positions des m boules blanches. Les tirages sont équiprobables. On a $\text{Card}(\Omega) = \binom{N}{m}$. Parmi ces tirages, on considère ceux qui donnent k boules blanches dans les n premières boules. Il faut donc choisir k positions pour les boules blanches parmi les n premières boules, et $m - k$ positions pour les boules blanches parmi les $N - n$ dernières. Il existe donc $\binom{n}{k} \binom{N-n}{m-k}$ configurations possibles. On en déduit que $\mathbb{P}(X_N = k) = \frac{\binom{n}{k} \binom{N-n}{m-k}}{\binom{N}{m}}$.

Il est facile de vérifier que :

$$\frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} = \frac{\binom{n}{k} \binom{N-n}{m-k}}{\binom{N}{m}}.$$

2. Pour N et n suffisamment grands, on a $n - N + m \leq 0$ et les conditions $n - N + m \leq k \leq m$ et $n \geq k \geq 0$ deviennent $0 \leq k \leq m$. On note $p_N(k) = \mathbb{P}(X_N = k)$. On remarque que :

$$\begin{aligned} p_N(k) &= \frac{\binom{m}{k} \binom{N-m}{n-k}}{\binom{N}{n}} \\ &= \binom{m}{k} \left(\prod_{i=0}^{k-1} \frac{n-i}{N-i} \right) \left(\prod_{j=0}^{m-k-1} \frac{N-n-j}{N-k-j} \right) \\ &= \binom{m}{k} \left(\prod_{i=0}^{k-1} (p + o(1)) \right) \left(\prod_{j=0}^{m-k-1} (1 - p + o(1)) \right) \\ &= \binom{m}{k} p^k (1 - p)^{m-k} + o(1). \end{aligned}$$

On en déduit que pour tout $k \in \{0, \dots, m\}$, on a $\lim_{N \rightarrow \infty} p_N(k) = \binom{m}{k} p^k (1-p)^{m-k}$. Soit g une fonction continue bornée. Comme la somme porte sur un nombre fini de termes, on a :

$$\mathbb{E}[g(X_N)] = \sum_{k=0}^m p_N(k) g(k) \xrightarrow{N \rightarrow \infty} \sum_{k=0}^m \binom{m}{k} p^k (1-p)^{m-k} g(k) = \mathbb{E}[g(S_m)],$$

où $\lim_{N \rightarrow \infty} n/N = p$ et S_m suit la loi binomiale de paramètre (m, p) . On en déduit que la suite $(X_N, N \in \mathbb{N}^*)$ converge en loi, quand $\lim_{N \rightarrow \infty} n/N = p$, vers la loi binomiale de paramètre (m, p) .

3. On a $p_N(k) = \frac{\binom{n}{k} \binom{N-n}{m-k}}{\binom{N}{m}}$ pour $n - N + m \leq k \leq m$ et $n \geq k \geq 0$. Cette expression correspond à celle de la question précédente, où l'on a échangé m et n . Les calculs de la question précédente assurent donc que la suite $(X_N, N \in \mathbb{N}^*)$ converge en loi vers la loi binomiale de paramètre (n, θ) . ▲

Exercice V.9. 1. La fonction $\theta \mapsto e^{\theta X}$ est de classe C^∞ et sur tout intervalle borné de \mathbb{R} cette fonction et ses dérivées sont p.s. bornées (car X est bornée). Ceci assure que Φ est de classe C^∞ et que $\Phi^{(n)}(\theta) = \mathbb{E}[X^n e^{\theta X}]$.

2. L'inégalité de Cauchy-Schwarz $\mathbb{E}[YZ]^2 \leq \mathbb{E}[Y^2]\mathbb{E}[Z^2]$ avec $Y = X e^{\theta X/2}$ et $Z = e^{\theta X/2}$ implique que $(\Phi')^2 \leq \Phi \Phi''$. Comme $\lambda'' = \frac{\Phi \Phi'' - (\Phi')^2}{\Phi^2}$, on en déduit que $\lambda'' \geq 0$ et donc λ est convexe. On a $\lambda(0) = \log(1) = 0$.

3. Comme $I(x) \geq -\lambda(0)$ et que $\lambda(0) = 0$, on en déduit que I est positive. Comme λ est convexe et de classe C^1 , on en déduit que $\theta\mu - \lambda(\theta)$ est maximal en θ tel que $\mu = \lambda'(\theta)$ soit encore $\Phi'(\theta) = \mu\Phi(\theta)$. On remarque que $\Phi(0) = 1$ et $\Phi'(0) = \mathbb{E}[X] = \mu$. Ceci assure que $I(\mu) = 0$.

Soit $\alpha \in]0, 1[$, $x, y \in \mathbb{R}$. On a :

$$\begin{aligned} I(\alpha x + (1-\alpha)y) &= \sup_{\theta \in \mathbb{R}} \{(\alpha x + (1-\alpha)y)\theta - \lambda(\theta)\} \\ &= \sup_{\theta \in \mathbb{R}} \{\alpha(\theta x - \lambda(\theta)) + (1-\alpha)(\theta y - \lambda(\theta))\} \\ &\leq \alpha \sup_{\theta \in \mathbb{R}} \{\theta x - \lambda(\theta)\} + (1-\alpha) \sup_{\theta \in \mathbb{R}} \{\theta y - \lambda(\theta)\} \\ &= \alpha I(x) + (1-\alpha)I(y). \end{aligned}$$

4. On a, pour $\theta \geq 0$:

$$\mathbb{P}(\bar{X}_n \geq \mu + \varepsilon) = \mathbb{P}(e^{\theta \bar{X}_n - \theta(\mu + \varepsilon)} \geq 1) \leq \mathbb{E}[e^{\theta \bar{X}_n - \theta(\mu + \varepsilon)}] = \mathbb{E}[e^{\theta X/n}]^n e^{-\theta(\mu + \varepsilon)},$$

où l'on a utilisé l'inégalité de Markov puis l'indépendance pour la seconde égalité.

On peut optimiser en $\theta \geq 0$ et obtenir :

$$\mathbb{P}(\bar{X}_n \geq \mu + \varepsilon) \leq e^{-\sup_{\theta \geq 0} \{\theta(\mu + \varepsilon) - n\lambda(\theta/n)\}}.$$

On considère la fonction g définie par $g(\theta) = \theta(\mu + \varepsilon) - \lambda(\theta)$. Cette fonction est concave (car λ est convexe) et :

$$g'(0) = \mu + \varepsilon - \frac{\Phi'(0)}{\Phi(0)} = \mu + \varepsilon - \mu = \varepsilon \geq 0.$$

En particulier le supremum de g est obtenu sur $[0, \infty[$ et donc

$$\sup_{\theta \geq 0} \{\theta(\mu + \varepsilon) - n\lambda(\theta/n)\} = \sup_{\theta \geq 0} ng(\theta/n) = n \sup_{\theta' \in \mathbb{R}} g(\theta') = nI(\mu + \varepsilon),$$

avec le changement de variable $\theta' = \theta/n$. Comme la fonction I est convexe et atteint son minimum en μ , on en déduit que $I(\mu + \varepsilon) = \inf_{x \geq \mu + \varepsilon} I(x)$. Il vient donc :

$$\mathbb{P}(\bar{X}_n \geq \mu + \varepsilon) \leq e^{-n \inf\{I(x); x \geq \mu + \varepsilon\}}.$$

5. En considérant $-X$ au lieu de X , on déduit de l'inégalité précédente que :

$$\mathbb{P}(\bar{X}_n \leq \mu - \varepsilon) \leq e^{-n \inf\{I(x); x \leq \mu - \varepsilon\}}.$$

Ceci implique donc $\mathbb{P}(|\bar{X}_n - \mu| \geq \varepsilon) \leq 2e^{-n \inf\{I(x); |x - \mu| \geq \varepsilon\}}$.

6. Pour la loi de Bernoulli de paramètre $p \in]0, 1[$, on a $\mu = p$, $\Phi(\theta) = 1 - p + pe^\theta$, $\lambda(\theta) = \log(1 - p + pe^\theta)$ et $I(x) = x \log(x/p) + (1 - x) \log((1 - x)/(1 - p))$ si $x \in [0, 1]$ et $I(x) = +\infty$ sinon. Pour $\varepsilon \in]0, \min(p, 1 - p)[$, on a :

$$\mathbb{P}(\bar{X}_n \leq p - \varepsilon) \leq e^{-n \inf\{I(x); x \leq p - \varepsilon\}},$$

et $\inf\{I(x); x \leq p - \varepsilon\} = \frac{\varepsilon^2}{2p(1 - p)} + O(\varepsilon^3)$. En particulier, on retrouve la loi faible des grands nombres avec en plus une majoration exponentielle de la vitesse de convergence. ▲

Exercice V.10. 1. Il suffit d'appliquer l'espérance à l'égalité $X = \sum_{k=1}^{+\infty} \mathbf{1}_{\{X \geq k\}}$. On peut ensuite intervertir l'espérance et la somme en utilisant le théorème de convergence dominée car tous les termes de la somme sont positifs.

2. En appliquant le résultat de la question précédente à la variable aléatoire mX (qui est intégrable et à valeurs dans \mathbb{N}), il vient :

$$\mathbb{E}[mX] = \sum_{n=1}^{\infty} \mathbb{P}(mX_n \geq n) = \sum_{n=1}^{\infty} \mathbb{P}\left(\frac{X_n}{n} \geq \frac{1}{m}\right) = \mathbb{E}[Y_m].$$

L'interversion de la somme et de l'espérance dans la deuxième égalité est justifiée car tous les termes sont positifs. La variable aléatoire Y_m est d'espérance finie, elle est donc p.s. finie.

3. L'évènement $A_m = \{Y_m < +\infty\}$ est de probabilité 1. Par conséquent, l'évènement $A = \bigcap_{m \geq 1} A_m$ est lui aussi de probabilité 1. Pour $\omega \in A$, on a que pour tout m , $Y_m(\omega)$ est fini et donc $\frac{X_n(\omega)}{n} \leq \frac{1}{m}$ pour n assez grand. On en déduit donc que $\lim_{n \rightarrow \infty} \frac{X_n(\omega)}{n} = 0$ pour tout $\omega \in A$. Comme $\mathbb{P}(A) = 1$, on en déduit que $\frac{X_n}{n}$ converge p.s. vers 0 quand n tend vers l'infini.
4. Soit $(a_n, n \geq 1)$ une suite de termes positifs telle que $\lim_{n \rightarrow \infty} a_n/n = 0$. En particulier la suite est bornée par une constante $M \geq 0$. Pour tout $\varepsilon > 0$, il existe n_0 tel que $a_n/n \leq \varepsilon$ pour tout $n \geq n_0$. On a pour tout $n \geq \max(n_0, M/\varepsilon) = n_1$:

$$\frac{1}{n} \max_{1 \leq k \leq n} a_k \leq \frac{1}{n} \max_{1 \leq k \leq n_1} a_k + \max_{n_1 < k \leq n} \frac{a_k}{k} \leq \frac{M}{n} + \varepsilon \leq 2\varepsilon.$$

On en déduit donc que $\lim_{n \rightarrow \infty} \frac{1}{n} \max_{1 \leq k \leq n} a_k = 0$. Comme p.s. $\lim_{n \rightarrow \infty} X_n/n = 0$, on déduit de ce résultat déterministe que p.s. $\lim_{n \rightarrow \infty} \frac{1}{n} \max_{1 \leq k \leq n} X_k = 0$.

5. On note $[x]$ l'entier relatif n tel que $n - 1 < x \leq n$. La variable aléatoire $[|X|]$ est à valeurs dans \mathbb{N} . Elle est intégrable car $[|X|] \leq |X| + 1$. On déduit de la question précédente que p.s. $\lim_{n \rightarrow \infty} \frac{1}{n} \max_{1 \leq k \leq n} [|X_k|] = 0$. Comme $[|X_k|] \geq |X_k|$, on a donc p.s. $\lim_{n \rightarrow \infty} \frac{1}{n} \max_{1 \leq k \leq n} |X_k| = 0$ et donc p.s. $\lim_{n \rightarrow \infty} \frac{1}{n} \max_{1 \leq k \leq n} X_k = 0$. ▲

Exercice V.11. 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 :

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

D'autre part, on a :

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

On en déduit donc :

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

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

2. On vérifie facilement à l'aide des fonctions caractéristiques que $\sum_{k=1}^n Y_k$ est une variable aléatoire de loi de Poisson de paramètre $n\alpha$. En effet, en utilisant l'indépendance des variables aléatoires Y_1, \dots, Y_n , on a :

$$\psi_{Z_n}(u) = \prod_{k=1}^n \psi_{Y_k}(u) = e^{-n\alpha(1-e^{iu})}.$$

On reconnaît dans le membre de droite la fonction caractéristique de la loi de Poisson de paramètre $n\alpha$.

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 :

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

On en déduit que :

$$\lim_{n \rightarrow \infty} \sum_{k \geq 0} e^{-\alpha n} \frac{(\alpha n)^k}{k!} f\left(\frac{k}{n}\right) = f(\alpha).$$

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

▲

Exercice V.12. 1. Les variables aléatoires X_n sont indépendantes, de même loi et de carré intégrable. On a $\mathbb{E}[X_n] = 1$, $\text{Var}(X_n) = 1$. Le théorème central limite implique que $(\frac{S_n - n}{\sqrt{n}}, n \geq 1)$ converge en loi vers la loi gaussienne $\mathcal{N}(0, 1)$.

2. On vérifie facilement à l'aide des fonctions caractéristiques que S_n est une variable aléatoire de loi de Poisson de paramètre n .

Soit $F_n(x)$ la fonction de répartition de $\frac{S_n - n}{\sqrt{n}}$. La question précédente assure que la suite $(F_n(x), n \geq 1)$ converge vers la fonction de répartition de la loi gaussienne $\mathcal{N}(0, 1)$, $F(x)$, pour tout x point de continuité de F . Comme F est une fonction continue, on obtient pour $x = 0$:

$$\mathbb{P}\left(\frac{S_n - n}{\sqrt{n}} \leq 0\right) \xrightarrow{n \rightarrow \infty} F(0) = \frac{1}{2}.$$

Comme S_n est de loi de Poisson de paramètre n , il vient :

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

On en déduit le résultat. ▲

Exercice V.13. 1. Les variables aléatoires $(X_n, n \geq 1)$ sont indépendantes et de même loi avec :

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

On observe que $\mathbb{E}[X_1] = \infty$. Les variables aléatoires étant positives indépendantes et de même loi, on a par la loi forte des grands nombres que p.s. $\lim_{n \rightarrow \infty} \frac{S_n}{n} = \mathbb{E}[X_1] = \infty$. Le prix équitable du jeu correspond pour le casino à la moyenne des pertes : $\lim_{n \rightarrow \infty} \frac{S_n}{n}$, c'est-à-dire l'infini.

2. La majoration (V.2) est évidente. On considère le deuxième terme du membre de droite de (V.2). On note $[x]$ la partie entière de x . On a :

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

et, pour n suffisamment grand :

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

Ceci implique que $\lim_{n \rightarrow \infty} \frac{\mathbb{E}[S'_n]}{n \log_2 n} = 1$. Donc pour n assez grand (dépendant de ε), on a $\left| \frac{\mathbb{E}[S'_n]}{n \log_2 n} - 1 \right| \leq \varepsilon/2$. Pour n assez grand, le deuxième terme du membre de droite de (V.2) est donc nul.

On considère le premier terme du membre de droite de (V.2). L'inégalité de Tchebychev implique :

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

En utilisant l'indépendance des variables $(X_k, 1 \leq k \leq n)$, il vient :

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

De plus, on a :

$$\begin{aligned} \mathbb{E}[X_1^2 \mathbf{1}_{\{X_1 \leq n \log_2 n\}}] &= \sum_{k \geq 1; 2^k \leq n \log_2 n} 2^k \leq \sum_{1 \leq k \leq \lfloor \log_2(n \log_2 n) \rfloor} 2^k \\ &\leq 2^{\lfloor \log_2(n \log_2 n) \rfloor + 1} \\ &\leq 2n \log_2 n. \end{aligned}$$

On en conclut que $\frac{4 \text{Var}(S'_n)}{(\varepsilon n \log_2 n)^2} \leq \frac{8}{\varepsilon^2 \log_2 n}$. On en déduit donc que pour tout $\varepsilon > 0$, $\lim_{n \rightarrow \infty} \mathbb{P}\left(\left|\frac{S'_n}{n \log_2 n} - 1\right| > \varepsilon\right) = 0$. Donc la suite $\left(\frac{S'_n}{n \log_2 n}, n \geq 1\right)$ converge en probabilité vers 1.

3. On observe que :

$$\mathbb{P}(X_1 > n \log_2 n) = \sum_{k \geq 1; 2^k > n \log_2 n} 2^{-k} \leq \sum_{k \geq \lfloor \log_2(n \log_2 n) \rfloor} 2^{-k} = 2^{-\lfloor \log_2(n \log_2 n) \rfloor + 1}.$$

On remarque que :

$$\lfloor \log_2(n \log_2 n) \rfloor = \lfloor \log_2 n + \log_2(\log_2 n) \rfloor \geq \log_2 n + \log_2(\log_2 n) - 1,$$

et donc :

$$\begin{aligned} \mathbb{P}(S_n \neq S'_n) &= \mathbb{P}(\cup_{k=1}^n \{X_k \neq X_k^n\}) \leq \sum_{k=1}^n \mathbb{P}(X_k \neq X_k^n) \\ &= n \mathbb{P}(X_1 > n \log_2 n) \leq \frac{4}{\log_2 n} \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

4. Soit $\varepsilon > 0$. On a :

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

Il vient donc :

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

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

On peut en fait montrer que la suite $(2^{-n}S_{2^n} - n, n \geq 1)$ converge en loi². ▲

Exercice V.14. 1. On note X_i la réponse de la i -ème personne interrogée ($X_i = 1$ si elle vote pour le candidat A et $X_i = 0$ si elle vote pour le candidat B). Les variables aléatoires X_1, \dots, 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 le nombre de personnes interrogées, n , alors les résultats de convergence et en particulier l'intervalle de confiance sont les mêmes que pour le tirage avec remise). On estime p avec 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%, on a :

$$p \in \left[\bar{X}_n \pm 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \subset \left[\bar{X}_n \pm \frac{1.96}{2\sqrt{n}}\right] \simeq \left[\bar{X}_n \pm \frac{1}{\sqrt{n}}\right].$$

En particulier pour $n = 1\,000$, on obtient une précision asymptotique (par excès) d'environ $\pm 1/\sqrt{n}$ c'est-à-dire ± 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 $p = 0.5 + 3.3 \cdot 10^{-4}$. 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 $p - \frac{2}{\sqrt{n}} \geq \frac{1}{2}$ avec une certitude d'au moins 95%. On en déduit $\frac{2}{\sqrt{n}} = 3.3 \cdot 10^{-4}$ soit $n = 36\,000\,000$. Comme n est plus grand que la taille de la population, on ne peut donc pas déterminer le vainqueur en faisant un sondage. ▲

Exercice V.15. 1. On considère les fonctions caractéristiques :

2. A. Martin-Löf. A limit theorem which clarifies the "Petersburg paradox". *J. Appl. Prob.*, vol. 22, pp. 634-643 (1985).

$$\psi_{Y_m}(u) = (1 - p_m + p_m e^{iu})^m = \left(1 - \frac{mp_m(1 - e^{iu})}{m}\right)^m.$$

Rappelons que $(1 - \frac{x_n}{n})^n$ converge vers e^{-x} si x_n converge dans \mathbb{C} vers x , quand n tend vers l'infini. On en déduit que $\lim_{n \rightarrow \infty} \psi_{Y_m}(u) = e^{-\theta(1 - e^{iu})}$. On reconnaît la fonction caractéristique 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 $\mathcal{B}(537, 1/576)$ soit approximativement une loi de Poisson de paramètre $\theta \simeq 0.9323$. 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 . On compare N_k et $N p_k$ dans la table (XIII.1).

| | | | | | | |
|--------|-------|-------|------|------|-----|-----|
| k | 0 | 1 | 2 | 3 | 4 | 5+ |
| N_k | 229 | 211 | 93 | 35 | 7 | 1 |
| Np_k | 226.7 | 211.4 | 98.5 | 30.6 | 7.1 | 1.6 |

Table XIII.1. Nombres N_k et $N p_k$ en fonction de k .

Les valeurs sont très proches. En fait, en faisant un test statistique du χ^2 , on peut vérifier qu'il est raisonnable d'accepter le modèle. On peut donc dire que les bombes qui sont tombées dans le sud de Londres sont réparties au hasard. ▲

Exercice V.16. L'ensemble des fractions rationnelles $F = \{a/b^n; a \in \mathbb{N}, n \in \mathbb{N}^*\} \cap [0, 1]$ 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 $x_1, \dots, x_n \in \{0, \dots, b - 1\}$. On a $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \mathbb{P}(X - \sum_{k=1}^n \frac{x_k}{b^k} \in [0, b^{-n}[)$. Comme X est de loi uniforme, on a $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = b^{-n}$. La loi de (X_1, \dots, X_n) est donc la loi uniforme sur $\{0, \dots, b - 1\}^n$.
2. Par la formule des lois marginales, on obtient $\mathbb{P}(X_n = x_n) = 1/b$, et donc X_n est de loi uniforme sur $\{0, \dots, b - 1\}$. On a $\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{k=1}^n \mathbb{P}(X_k = x_k)$ pour tout $x_1, \dots, x_n \in \{0, \dots, b - 1\}$. Cela implique que les variables aléatoires X_1, \dots, X_n sont indépendantes. Ceci étant vrai pour tout $n \geq 2$, cela signifie que les variables aléatoires $(X_n, n \geq 1)$ sont indépendantes.
3. Soit $a \in \{0, \dots, b - 1\}$. On pose $Y_k = \mathbf{1}_{\{X_k=a\}}$. La quantité $\sum_{k=1}^n Y_k$ compte le nombre d'apparitions 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 :

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n Y_k = \mathbb{E}[Y_1] = \frac{1}{b} \quad \text{p.s..}$$

Ceci est vrai pour tout $a \in \{0, \dots, b-1\}$. On en déduit que p.s. X est simplement normal en base b .

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 :

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

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.17. 1. On déduit de l'inégalité de Tchebychev que :

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

où l'on utilise que $x \in [0, 1]$ pour la dernière inégalité.

2. La fonction h est uniformément continue sur $[0, 1]$ car elle est continue sur le compact $[0, 1]$ (théorème de Heine). Soit $\varepsilon > 0$, il existe donc $\delta > 0$, tel que pour tous $x, y \in [0, 1]$, si $|x - y| \leq \delta$, alors on a $|h(x) - h(y)| \leq \varepsilon$. D'autre part, h étant continue sur $[0, 1]$, elle est bornée par une constante $M > 0$. En utilisant l'inégalité de Jensen, il vient :

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

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

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 $|h(x) - \mathbb{E}[h(\bar{X}_n)]| \leq 2\varepsilon$ pour tout $x \in [0, 1]$.
En conclusion, il vient :

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

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

4. On a :

$$\mathbb{E}[h(\bar{X}_n)] = \mathbb{E}[h(n\bar{X}_n/n)] = \sum_{k=1}^n h(k/n) \binom{n}{k} x^k (1-x)^{n-k}.$$

On déduit de la question 5, que la suite de polynômes $(\sum_{k=1}^n h(k/n) \binom{n}{k} x^k (1-x)^{n-k}, n \geq 1)$, 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 x , et on utilise le fait que $\sum_{k=1}^n X_k$ suit une loi de Poisson de paramètre nx . On obtient :

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

La convergence n'est pas uniforme. En effet, pour l'exemple donné, il vient :

$$\lim_{n \rightarrow \infty} \left| f(x_n) - \sum_{k=0}^{\infty} e^{-nx_n} \frac{(nx_n)^k}{k!} f(k/n) \right| = 1 - e^{-\pi/2}.$$

▲

Exercice V.18. 1. En utilisant l'indépendance, pour tout $y \in \mathbb{R}$, on a :

$$\mathbb{P}(Z_n \leq y) = \mathbb{P}(X_1 \leq n^{1/\lambda} y)^n = (1 - \varepsilon_n)^n,$$

où $\varepsilon_n = \mathbb{P}(X_1 > n^{1/\lambda} y)$. Pour $y \leq 0$, on a $\varepsilon_n > \eta > 0$, et donc $\lim_{n \rightarrow +\infty} \mathbb{P}(Z_n \leq y) = 0$. Pour $y > 0$, on a $\varepsilon_n \sim \frac{\alpha y^{-\lambda}}{n}$ quand n tend vers l'infini. Par conséquent, $\lim_{n \rightarrow +\infty} \mathbb{P}(Z_n \leq y) = \exp(-\alpha y^{-\lambda})$. Cela démontre que la fonction de répartition de Z_n converge vers $\exp(-\alpha y^{-\lambda}) \mathbf{1}_{\{y>0\}}$ et donc la suite $(Z_n, n \geq 1)$ converge en loi vers Y de loi de Fréchet de paramètre (α, λ) .

2. Soit n individus. On note X_k le niveau annuel d'exposition au mercure de l'individu k . On suppose les variables aléatoires X_k indépendantes et de même loi que X . Pour un individu de 70 kg, la dose annuelle admissible fixée par l'OMS est $s = 18.14 \cdot 10^{-3}$ g. La probabilité qu'un individu au moins ait un niveau de mercure trop élevé est

$$\begin{aligned} p_n &= \mathbb{P}(\max(X_1, \dots, X_n) > s) = 1 - \mathbb{P}(\max(X_1, \dots, X_n) \leq s) \\ &= 1 - e^{n \ln(1 - \alpha s^{-\lambda})} \\ &\simeq 1 - \exp(-\alpha s^{-\lambda} n). \end{aligned}$$

Si $n_1 = 225\ 362$, $n_2 = 100$, on a $p_{n_1} \simeq 1$ et $p_{n_2} \simeq 0.01$. Enfin, $1 - \exp(-\alpha s^{-\lambda} n) \geq 0.95$ si et seulement si $n \geq \frac{\log(20)}{\alpha s^{-\lambda}}$, i.e. $n \geq 27\ 214$. ▲

XIII.6 Vecteurs gaussiens

Exercice VI.1. 1. On voit que quel que soit $i = 1, 2, 4$, on a $\text{Cov}(X_3, X_i) = 0$, 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 :

$$\text{Cov}(X_1, X_2) = \mathbb{E}[X_1 X_2] = a\mathbb{E}[X_2^2] + \mathbb{E}[X_2]\mathbb{E}[W] = a,$$

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 $\mathbb{E}[(X - \mathbb{E}[X|Y])Y] = 0$, 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. Le vecteur $(X_1 - X_2, X_4 - X_2)$ est gaussien et on a $\mathbb{E}[(X_1 - X_2)(X_4 - X_2)] = \text{Cov}(X_1, X_4) - \text{Cov}(X_2, X_4) - \text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_2) = 0$. Donc ses marginales variables sont indépendantes. On pose $Y_1 = (X_1 - X_2, 0, 0, 0)$, $Y_2 = (X_2, X_2, 0, X_2)$, $Y_3 = (0, 0, X_3, 0)$ et $Y_4 = (0, 0, 0, X_4 - X_2)$. Les vecteurs gaussiens Y_1, Y_2, Y_3 et Y_4 sont indépendants et leur somme vaut X . ▲

Exercice VI.2. 1. Soit g une fonction mesurable bornée. On a par indépendance :

$$\begin{aligned}\mathbb{E}[g(Y)] &= \mathbb{E}[g(ZX)] \\ &= \mathbb{E}[g(X)\mathbf{1}_{\{Z=1\}} + g(-X)\mathbf{1}_{\{Z=-1\}}] \\ &= \mathbb{E}[g(X)]\mathbb{P}(Z=1) + \mathbb{E}[g(-X)]\mathbb{P}(Z=-1) \\ &= \frac{1}{2}(\mathbb{E}[g(X)] + \mathbb{E}[g(-X)]) \\ &= \mathbb{E}[g(X)],\end{aligned}$$

car la loi de X est symétrique. Donc X et Y sont de même loi $\mathcal{N}(0, 1)$.

2. On a :

$$\text{Cov}(X, Y) = \mathbb{E}[XY] = \mathbb{E}[X^2]\mathbb{P}(Z=1) + \mathbb{E}[-X^2]\mathbb{P}(Z=-1) = 0.$$

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 on a dans cet exercice deux variables aléatoires gaussiennes de covariance nulle qui ne sont pas indépendantes.

▲

Exercice VI.3. 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 :

$$\mathbb{E}[(X+Y)(X-Y)] - \mathbb{E}[X+Y]\mathbb{E}[X-Y] = \text{Var}(X) - \text{Var}(Y).$$

On conclut donc que $X+Y$ et $X-Y$ sont indépendantes si et seulement si $\text{Var}(X) = \text{Var}(Y)$.

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 on a $\text{Var}\left(\frac{X-Y}{\sqrt{2}}\right) = \text{Var}\left(\frac{X+Y}{\sqrt{2}}\right) = 1$. Ceci assure que $\frac{X-Y}{\sqrt{2}}$ et $\frac{X+Y}{\sqrt{2}}$ sont de loi gaussienne centrée réduite.

▲

- Exercice VI.4.* 1. En utilisant la matrice de changement de base, on a $Y = UX$ où $X = (X_1, \dots, X_n)$. 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 $m_i = \sum_{k=1}^{i-1} n_k$ 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 :

$$\|X_H\|^2 = \|X - X_\Delta\|^2 = \sum_{i=1}^n |X_i - \bar{X}_n|^2 = T_n.$$

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 VI.5.* 1. La variable aléatoire $X = (X_1, \dots, X_n)$ 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 $\text{Var}(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{\sigma^2}{n}$. 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 $\text{Cov}(\bar{X}_n, X_i - \bar{X}_n) = 0$, ce qui est le cas car :

$$\text{Cov}(\bar{X}_n, X_i) = \text{Cov}\left(\frac{1}{n} \sum_{j=1}^n X_j, X_i\right) = \frac{1}{n} \text{Cov}(X_i, X_i) = \frac{1}{n} \sigma^2 = \text{Var}(\bar{X}_n).$$

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 :

$$\sum_{i=1}^n X_i^2 = (n-1)V_n + (\sqrt{n}\bar{X}_n)^2.$$

La loi de $\sum_{i=1}^n X_i^2/\sigma^2$ est la loi $\chi^2(n)$. Comme $\sqrt{n}\bar{X}_n/\sigma$ est une variable aléatoire gaussienne centrée réduite, la loi de $(\sqrt{n}\bar{X}_n/\sigma)^2$ est donc la loi $\chi^2(1)$. Par indépendance, on obtient pour les fonctions caractéristiques :

$$\begin{aligned} \left(\frac{1}{1-2iu}\right)^{n/2} &= \psi_{\frac{\sum_{i=1}^n X_i^2}{\sigma^2}}(u) = \psi_{\frac{(n-1)V_n}{\sigma^2}}(u) \psi_{\frac{\sqrt{n}\bar{X}_n}{\sigma}}(u) \\ &= \psi_{\frac{(n-1)V_n}{\sigma^2}}(u) \left(\frac{1}{1-2iu}\right)^{1/2}. \end{aligned}$$

On en déduit donc que $\psi_{(n-1)V_n/\sigma^2}(u) = \left(\frac{1}{1-2iu}\right)^{(n-1)/2}$. Ainsi $(n-1)V_n/\sigma^2$ est de loi $\chi^2(n-1)$. 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.6. 1. On a :

$$\mathbb{E}[(n-1)V_n] = \mathbb{E}\left[\sum_{j=1}^n (X_j^2 - 2\bar{X}_n X_j + \bar{X}_n^2)\right] = n\sigma^2 - 2n\frac{\sigma^2}{n} + n^2\frac{\sigma^2}{n^2} = (n-1)\sigma^2.$$

Comme V_n et \bar{X}_n sont indépendants, on a :

$$\mathbb{E}[(n-1)V_n e^{itn\bar{X}_n}] = \mathbb{E}[(n-1)V_n] \mathbb{E}[e^{itn\bar{X}_n}] = (n-1)\sigma^2 \psi(t)^n.$$

2. En développant, il vient :

$$\begin{aligned}
 \mathbb{E}[(n-1)V_n e^{itn\bar{X}_n}] &= \mathbb{E}\left[\left(1 - \frac{1}{n}\right) \sum_{j=1}^n X_j^2 - \frac{2}{n} \sum_{1 \leq j < k \leq n} X_j X_k\right] e^{it \sum_{j=1}^n X_j} \\
 &= \left(1 - \frac{1}{n}\right) \sum_{j=1}^n \mathbb{E}[X_j^2 e^{itX_j}] \mathbb{E}[e^{itX_1}]^{n-1} \\
 &\quad - \frac{2}{n} \sum_{1 \leq j < k \leq n} \mathbb{E}[X_j e^{itX_j}] \mathbb{E}[X_k e^{itX_k}] \mathbb{E}[e^{itX_1}]^{n-2} \\
 &= -\left(1 - \frac{1}{n}\right) n \psi''(t) \psi(t)^{n-1} - \frac{2}{n} \frac{n(n-1)}{2} (-i\psi'(t))^2 \psi(t)^{n-2} \\
 &= -(n-1) \psi''(t) \psi(t)^{n-1} + (n-1) \psi'(t)^2 \psi(t)^{n-2}.
 \end{aligned}$$

3. Soit l'ouvert $\{t; \psi(t) \neq 0\}$ et O la composante connexe de cet ouvert contenant 0. On déduit de la question précédente que ψ est solution de l'équation différentielle :

$$\begin{cases} \frac{\psi''}{\psi} - \left(\frac{\psi'}{\psi}\right)^2 = -\sigma^2 & \text{sur } O, \\ \psi(0) = 1, \psi'(0) = 0. \end{cases}$$

4. On pose $\varphi = \psi'/\psi$ sur O . 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 $\psi(t) = e^{-t^2 \sigma^2 / 2}$ sur O . On cherche alors les solutions sous la forme $\psi(t) = e^{-t^2 \sigma^2 / 2} h(t)$ (méthode de la variation de la constante). On en déduit que $h' = 0$. La condition $\psi(0) = 1$ implique que $e^{-t^2 \sigma^2 / 2}$ est la seule solution de l'équation différentielle considérée et $O = \mathbb{R}$. 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.7. 1. On sait que la variable aléatoire \bar{X}_n est de loi gaussienne avec $\mathbb{E}[\bar{X}_n] = \theta$ et $\text{Var}(\bar{X}_n) = \frac{\theta}{n}$. 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. On sait que $(n-1)V_n/\theta$ suit la loi $\chi^2(n-1)$, et on a $\mathbb{E}[V_n] = \theta$ et $\text{Var}(V_n) = \frac{2\theta^2}{(n-1)}$. On remarque 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 $\text{Var}(X_1) = \theta$.

3. Les variables aléatoires \bar{X}_n et V_n sont indépendantes. 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 :

$$\begin{aligned}\mathbb{E}[T_n^\lambda] &= \lambda \mathbb{E}[\bar{X}_n] + (1 - \lambda) \mathbb{E}[V_n] = \theta, \\ \text{Var}(T_n^\lambda) &= \lambda^2 \text{Var}(\bar{X}_n) + (1 - \lambda)^2 \text{Var}(V_n) + 2 \text{Cov}(\bar{X}_n, V_n) \\ &= \lambda^2 \frac{\theta}{n} + (1 - \lambda)^2 \frac{2\theta^2}{(n - 1)},\end{aligned}$$

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 $\lambda \mathbb{E}[X_1] + (1 - \lambda) \text{Var}(X_1) = \theta$.

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 $\text{Var}(X_k) = \theta$. On déduit du théorème central limite, que la suite $(\sqrt{n}(\bar{X}_n - \theta), n \geq 1)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \theta)$.
6. La variable aléatoire $\frac{n-1}{\theta} V_n$ est distribuée 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 :

$$\begin{aligned}\sqrt{n}(V_n - \theta) &= \sqrt{\frac{n}{n-1}} \theta \left(\sqrt{n-1} \left(\frac{1}{\theta} V_n - 1 \right) \right) \\ &\stackrel{\text{loi}}{=} \sqrt{\frac{n}{n-1}} \theta \left(\sqrt{n-1} \left(\frac{1}{n-1} \sum_{k=1}^{n-1} G_k^2 - 1 \right) \right).\end{aligned}$$

Comme $\mathbb{E}[G_k^2] = 1$ et $\text{Var}(G_k^2) = 2$, le théorème central limite implique 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

$\mathcal{N}(0, 2)$. On remarque 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 :

$$\lim_{n \rightarrow \infty} \psi_{(Y_n, W_n)}(v, w) = \lim_{n \rightarrow \infty} \psi_{Y_n}(v) \psi_{W_n}(w) = e^{-\theta v^2/2} e^{-2\theta^2 w^2/2}$$

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 $(a, b) \rightarrow \lambda a + (1 - \lambda)b$, 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 $\frac{\sqrt{n}T_n^\lambda - \theta}{\sigma}$ converge en loi vers une loi gaussienne $\mathcal{N}(0, 1)$, et que la loi gaussienne est une loi à densité, on a :

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\theta \in [T_n^\lambda - a \frac{\sigma}{\sqrt{n}}, T_n^\lambda + a \frac{\sigma}{\sqrt{n}}] \right) = \frac{1}{\sqrt{2\pi}} \int_{-a}^a e^{-\frac{x^2}{2}} dx.$$

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 $\frac{1}{\sqrt{2\pi}} \int_{-a}^a e^{-\frac{x^2}{2}} dx$. Pour le niveau de 95%, on trouve $a \simeq 1.96$, 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 obtient que $(\frac{\sqrt{n}T_n^\lambda - \theta}{\sigma_n}, n \geq 2)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, 1)$. On a :

$$\lim_{n \rightarrow \infty} \mathbb{P} \left(\theta \in [T_n^\lambda - a \frac{\sigma_n}{\sqrt{n}}, T_n^\lambda + a \frac{\sigma_n}{\sqrt{n}}] \right) = \frac{1}{\sqrt{2\pi}} \int_{-a}^a e^{-\frac{x^2}{2}} dx.$$

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 $\frac{1}{\sqrt{2\pi}} \int_{-a}^a e^{-\frac{x^2}{2}} dx$. Pour le niveau de 95%, on trouve $a \simeq 1.96$, 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 $\tilde{I}_n \simeq [3.42, 4.60]$.

11. En minimisant la fonction $\lambda \rightarrow \lambda^2\theta + 2(1 - \lambda)^2\theta^2$, on trouve $\lambda^* = \frac{2\theta}{1 + 2\theta}$. Par la convergence presque sûre de la suite V_n vers θ et la continuité de la

fonction $x \rightarrow \frac{2x}{1+2x}$ 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 $(\lambda', a, b) \rightarrow \lambda'a + (1-\lambda')b$ 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 $\mathcal{N}(0, \frac{2\theta^2}{1+2\theta})$. En remarquant que $T_n^{\lambda_n^*}$ converge presque sûrement vers θ , il résulte du théorème de Slutsky que $(\sqrt{n(1+2T_n^{\lambda_n^*})} \frac{T_n^{\lambda_n^*} - \theta}{\sqrt{2T_n^{\lambda_n^*}}}, n \geq 2)$ 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 asymptotique 95% :

$$\tilde{I}_n^* = [T_n^{\lambda_n^*} - 1.96 \frac{\sqrt{2T_n^{\lambda_n^*}}}{\sqrt{n(1+2T_n^{\lambda_n^*})}}, T_n^{\lambda_n^*} + 1.96 \frac{\sqrt{2T_n^{\lambda_n^*}}}{\sqrt{n(1+2T_n^{\lambda_n^*})}}].$$

On a $\lambda_n^* \simeq 0.885$, $T_n^{\lambda_n^*} \simeq 4.015$, $1.96 \frac{\sqrt{2T_n^{\lambda_n^*}}}{\sqrt{n(1+2T_n^{\lambda_n^*})}} \simeq 0.370$. On obtient l'intervalle $\tilde{I}_n^* \simeq [3.64, 4.39]$.

▲

XIII.7 Simulation

Exercice VII.1. 1. On calcule pour $x \geq 0$:

$$\mathbb{P}(T > x) = \mathbb{P}(U < e^{-\lambda x}) = e^{-\lambda x}.$$

On en déduit que T est de loi exponentielle de paramètre λ .

2. On calcule pour $n \in \mathbb{N}^*$:

$$\begin{aligned} \mathbb{P}(X = n) &= \mathbb{P}(n \leq T + 1 < n + 1) = \mathbb{P}(T \leq n - 1) - \mathbb{P}(T \leq n) \\ &= (e^{-\lambda})^{n-1} (1 - e^{-\lambda}). \end{aligned}$$

On en déduit que X est de loi géométrique de paramètre $p = 1 - e^{-\lambda}$.

▲

Exercice VII.2. 1. On calcule la fonction de répartition de $X_k = -\log(U_k)/\theta$, qui est une variable aléatoire positive : pour $x > 0$,

$$\mathbb{P}(-\log(U_k)/\theta \leq x) = \mathbb{P}(U_k \geq e^{-\theta x}) = 1 - e^{-\theta x}.$$

On en déduit que la loi de X_k est la loi exponentielle de paramètre θ .

2. En utilisant les fonctions caractéristiques, on en déduit que la loi de $\sum_{k=1}^n X_k$ est la loi $\Gamma(\theta, n)$.
3. Pour $n = 0$, on a $\mathbb{P}(N = 0) = \mathbb{P}(U_1 \leq e^{-\theta}) = e^{-\theta}$, et pour $n \geq 1$,

$$\begin{aligned} \mathbb{P}(N = n) &= \mathbb{P}\left(\prod_{k=1}^n U_k \geq e^{-\theta} > \prod_{k=1}^{n+1} U_k\right) \\ &= \mathbb{P}\left(\sum_{k=1}^n X_k \leq 1 < \sum_{k=1}^{n+1} X_k\right) \\ &= \mathbb{P}(1 < \sum_{k=1}^{n+1} X_k) - \mathbb{P}(1 < \sum_{k=1}^n X_k) \\ &= \int_1^\infty \frac{n!}{\theta^{n+1}} x^n e^{-\theta x} dx - \int_1^\infty \frac{(n-1)!}{\theta^n} x^{n-1} e^{-\theta x} dx \\ &= \left[-\frac{n!}{\theta^n} x^n e^{-\theta x} \right]_1^\infty \\ &= \frac{n!}{\theta^n} e^{-\theta}. \end{aligned}$$

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, \dots , d'une suite de variables aléatoires indépendantes de loi uniforme sur $[0, 1]$. On déduit de ce qui précède que $\inf \left\{ n \in \mathbb{N}; \prod_{k=1}^{n+1} x_k < e^{-\theta} \right\}$ est la réalisation d'une variable aléatoire de loi de Poisson de paramètre θ . ▲

Exercice VII.3. 1. On utilise la méthode de la fonction muette. Soit g une fonction mesurable bornée définie sur \mathbb{R}^2 . Soit φ le C^1 -difféomorphisme défini sur $]0, +\infty[\times]0, 2\pi[$ à valeurs dans $\mathbb{R}^2 \setminus \{(x, y); y = 0 \text{ et } x \geq 0\}$ par :

$$\begin{pmatrix} x \\ y \end{pmatrix} = \varphi(r, \theta) = \begin{pmatrix} r \cos(\theta) \\ r \sin(\theta) \end{pmatrix}.$$

Le jacobien de φ est $\text{Jac}[\varphi](r, \theta) = r$. On en déduit donc que :

$$\mathbb{E}[g(R, \Theta)] = \int_{]0, +\infty[\times]0, 2\pi[} r dr d\theta \frac{1}{2\pi} e^{-r^2/2} g(r, \theta).$$

On en déduit que (R, Θ) a pour densité $\frac{1}{2\pi} r e^{-r^2/2} \mathbf{1}_{]0, +\infty[\times]0, 2\pi[}(r, \theta)$. Comme la densité est sous forme du produit d'une fonction de r par une fonction de θ , on en déduit que R et Θ sont indépendants, R a pour densité $r e^{-r^2/2} \mathbf{1}_{r>0}$, et Θ est de loi uniforme sur $[0, 2\pi]$.

2. On utilise la méthode de la fonction muette. Soit g une fonction mesurable bornée définie sur \mathbb{R} . En utilisant le changement de variable $z = e^{-r^2/2}$, il vient :

$$\mathbb{E}[g(e^{-R^2/2})] = \int_{]0, +\infty[} g(e^{-r^2/2}) r e^{-r^2} dr = \int_{]0, 1[} g(z) dz.$$

On en déduit que $e^{-R^2/2}$ est de loi uniforme sur $[0, 1]$.

3. On déduit des questions précédentes que (X', Y') a même loi que (X, Y) . ▲

Exercice VII.4. 1. Soit φ une fonction mesurable bornée et $n \in \mathbb{N}^*$. On a, en utilisant l'indépendance des variables aléatoires :

$$\begin{aligned} \mathbb{E}[\varphi(Y)\mathbf{1}_{\{T=n\}}] &= \mathbb{E}[\varphi(X_k)\mathbf{1}_{\{X_1 \notin A, \dots, X_{n-1} \notin A, X_n \in A\}}] \\ &= \mathbb{P}(X \notin A)^{n-1} \mathbb{E}[\varphi(X)\mathbf{1}_{\{X \in A\}}] \end{aligned} \tag{XIII.3}$$

Cette espérance est sous forme produit. On en déduit que Y et T sont indépendants.

2. En prenant $\varphi = 1$ dans (XIII.3), on obtient :

$$\mathbb{P}(T = n) = \mathbb{P}(X \in A) \left(1 - \mathbb{P}(X \in A)\right)^{n-1}.$$

On en déduit que T est de loi géométrique de paramètre $\mathbb{P}(X \in A)$.

3. En sommant (XIII.3) sur $n \in \mathbb{N}^*$, il vient :

$$\mathbb{E}[\varphi(Y)] = \mathbb{E}[\varphi(X)\mathbf{1}_{\{X \in A\}}] / \mathbb{P}(X \in A) = \mathbb{E}[\varphi(X) | X \in A].$$

La loi de Y est donc la loi de X conditionnellement à $\{X \in A\}$.

4. On a :

$$\begin{aligned} \mathbb{P}((Z_1, U_1) \in A') &= \mathbb{P}(U_1 \leq ch(Z_1)/g(Z_1)) \\ &= \int g(z)\mathbf{1}_{[0,1]}(u) dz du \mathbf{1}_{\{u \leq ch(z)/g(z)\}} \\ &= c \int h(z) dz = c. \end{aligned}$$

5. Soit φ une fonction mesurable bornée. On déduit de ce qui précède que :

$$\begin{aligned} \mathbb{E}[\varphi(Z_{T'})] &= \mathbb{E}[\varphi(Z_1) | (U_1, Z_1) \in A'] \\ &= \frac{1}{c} \mathbb{E}[\varphi(Z_1) \mathbf{1}_{\{U_1 \leq ch(Z_1)/g(Z_1)\}}] \\ &= \frac{1}{c} \int g(z) \mathbf{1}_{[0,1]}(u) dz du \varphi(z) \mathbf{1}_{\{u \leq ch(z)/g(z)\}} \\ &= \int h(z) dz \varphi(z). \end{aligned}$$

On en déduit que $Z_{T'}$ est une variable aléatoire de densité h . ▲

XIII.8 Estimateurs

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 $\sum_{i=1}^n a_i = 1$. Comme les variables aléatoires sont indépendantes, on a $\text{Var}(\hat{\theta}_n) = \sum_{i=1}^n a_i^2 \sigma^2$, avec $\sigma^2 = \text{Var}(X_1)$. D'après l'inégalité de Cauchy-Schwarz on a $(\sum_{i=1}^n a_i^2)(\sum_{i=1}^n \frac{1}{n^2}) \geq (\sum_{i=1}^n \frac{a_i}{n})^2$. Puisque $\sum_{i=1}^n a_i = 1$, on déduit que $\sum_{i=1}^n a_i^2 \geq \frac{1}{n}$ avec égalité si et seulement si $a_i = \frac{1}{n}$ pour tout $i \in \{1, \dots, n\}$. L'estimateur de variance minimale dans la classe des estimateurs linéaires et sans biais est donc la moyenne empirique $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. ▲

Exercice VIII.2. 1. La loi forte des grands nombres implique que $(\bar{X}_n, n \geq 1)$, où $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, converge p.s. vers $\mathbb{E}_\lambda[X_1] = 1/\lambda$. Par continuité de la fonction $x \mapsto 1/x$ sur $]0, \infty[$, on en déduit que $\hat{\lambda}_n = \frac{1}{\bar{X}_n}$ est un estimateur convergent de λ .

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 :

$$\frac{\partial}{\partial \lambda} L_n(x; \lambda) = \frac{n}{\lambda} - \sum_{i=1}^n x_i = 0 \iff \lambda = \frac{n}{\sum_{i=1}^n x_i}.$$

Comme $\frac{\partial^2}{\partial \lambda^2} L_n(\lambda, x) > 0$, la log-vraisemblance est strictement convexe. On en déduit qu'elle 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}_λ . 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 $\Gamma(n-1, \lambda)$ dans l'intégrale pour la dernière égalité. 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 :

$$\hat{\lambda}_n^* = \frac{n-1}{\sum_{i=1}^n X_i}.$$

On calcule le deuxième moment de la même manière que le premier, pour $n \geq 3$:

$$\mathbb{E}_\lambda \left[\frac{1}{(\sum_{i=1}^n X_i)^2} \right] = \frac{\lambda^2}{(n-1)(n-2)} \int_0^\infty \frac{\lambda^{n-2}}{(n-3)!} s^{n-3} e^{-\lambda s} ds = \frac{\lambda^2}{(n-1)(n-2)}.$$

Donc pour tout $c > 0$, on a :

$$\begin{aligned} R(\hat{\lambda}_n^{(c)}, \lambda) &= \mathbb{E}_\lambda \left[\left(\frac{c}{\sum_{i=0}^n X_i} - \lambda \right)^2 \right] \\ &= \frac{\lambda^2}{(n-1)(n-2)} (c^2 - 2(n-2)c + (n-1)(n-2)). \end{aligned}$$

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 les formules d'interversion entre l'intégrale en x et la différentiation 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, \dots, 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 $C(\lambda) = \lambda^n$, $h(x) = \prod_{i=1}^n \mathbf{1}_{\{x_i > 0\}}$, $Q(\lambda) = -\lambda$ 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 $n \geq 3$.

- L'estimateur $\hat{\lambda}_n^*$ a été choisi sans biais.
- L'estimateur $\hat{\lambda}_n^*$ est optimal car il est fonction de la statistique exhaustive et totale $S(X)$ (Théorème de Lehman-Sheffé).
- L'estimateur sans biais $\hat{\lambda}_n^*$ a comme risque quadratique

$$\begin{aligned} \text{Var}_\lambda(\hat{\lambda}_n^*) &= R(\hat{\lambda}_n^*, \lambda) = \frac{\lambda^2}{(n-1)(n-2)} \left((n-1)^2 - (n-1)(n-2) \right) \\ &= \frac{\lambda^2}{n-2}. \end{aligned}$$

Il n'atteint donc pas la borne FDCR. Il n'est donc pas efficace. Il n'existe pas d'estimateur efficace, parce que tout estimateur efficace serait optimal et alors (p.s.) égal à $\hat{\lambda}^*$ par l'unicité (p.s.) de l'estimateur optimal.

- L'estimateur $\hat{\lambda}_n^*$ est préférable à $\hat{\lambda}_n$ car $R(\hat{\lambda}_n^*, \lambda) < R(\hat{\lambda}_n, \lambda)$ pour tout $\lambda > 0$.

- e) L'estimateur $\hat{\lambda}_n^*$ est inadmissible car $R(\hat{\lambda}_n^\circ, \lambda) < R(\hat{\lambda}_n^*, \lambda)$ pour tout $\lambda > 0$.
- f) L'estimateur $\hat{\lambda}_n^*$ 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 différentiation 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. $\sqrt{n}(\hat{\lambda}_n - \lambda)$ 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 efficace.

▲

Exercice VIII.3. 1. On a :

$$\mathbb{P}_\theta(X_1 = k_1, \dots, X_n = k_n) = e^{-n\theta} \left(\prod_{i=1}^n \frac{1}{k_i!} \right) \exp \left(\log(\theta) \sum_{i=1}^n k_i \right)$$

et on identifie la statistique canonique $S_n = \sum_{i=1}^n X_i$. La statistique canonique d'un modèle exponentiel est toujours exhaustive et totale. On vérifie à l'aide des fonctions caractéristiques que la variable aléatoire S_n suit une loi de Poisson de paramètre $n\theta$ sous \mathbb{P}_θ .

- 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 tous $k, s \in \mathbb{N}$, les probabilités conditionnelles :

$$\begin{aligned} \mathbb{P}_\theta(X_1 = k | S_n = s) &= \frac{\mathbb{P}_\theta(X_1 = k, S_n - X_1 = s - k)}{\mathbb{P}_\theta(S_n = s)} \\ &= \frac{e^{-\theta} \theta^k / k! e^{-(n-1)\theta} ((n-1)\theta)^{s-k} / (s-k)!}{e^{-n\theta} (n\theta)^s / s!} \\ &= \binom{n}{k} \left(\frac{n-1}{n} \right)^{s-k} \left(\frac{1}{n} \right)^k \end{aligned}$$

et on identifie les probabilités binomiales de paramètres $(s, \frac{1}{n})$. La loi de X_1 conditionnellement à S_n est donc la loi binomiale de paramètre $(S_n, 1/n)$.

- 4. On applique le Théorème de Lehman-Sheffé pour $\delta = \mathbf{1}_{\{X_1=0\}}$ et on calcule l'estimateur optimal δ_{S_n} par la méthode de Rao-Blackwell :

$$\delta_{S_n} = \mathbb{E}[\delta | S_n] = \mathbb{E}[\mathbf{1}_{\{X_1=0\}} | S_n] = \mathbb{P}(X_1 = 0 | S_n) = \left(1 - \frac{1}{n} \right)^{S_n}.$$

Par la loi forte des grands nombres on a p.s. $\lim_{n \rightarrow \infty} S_n/n = \theta$, en particulier p.s.

$\lim_{n \rightarrow \infty} S_n = \infty$, et, comme $\delta_{S_n} = \left(1 - \frac{1}{n} \right)^{S_n} = \left(1 - \frac{S_n/n}{S_n} \right)^{S_n}$, on déduit que p.s. $\lim_{n \rightarrow \infty} \delta_{S_n} = e^{-\theta}$. Ainsi, δ_{S_n} est un estimateur convergent de $e^{-\theta}$.

5. On vérifie la régularité du modèle (support, dérivabilité, interchangeabilité et existence de $I(\theta)$), et on calcule :

$$V_\theta(k) = \frac{\partial}{\partial \theta} \log \left(e^{-\theta} \theta^k / k! \right) = \frac{\partial}{\partial \theta} (k \log(\theta)) - \theta = \frac{k}{\theta} - 1,$$

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} V_\theta(X_1) \right] = \frac{\mathbb{E}_\theta[X_1]}{\theta^2} = \frac{1}{\theta}.$$

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_n} on utilise la fonction génératrice $\mathbb{E}[z^Y] = e^{-\lambda(1-z)}$ pour une variable Poisson Y de paramètre λ . Il vient :

$$\begin{aligned} \text{Var}_\theta(\delta_{S_n}) &= \mathbb{E}_\theta \left[\left(\left(1 - \frac{1}{n} \right)^2 \right)^{S_n} \right] - e^{-2\theta} \\ &= \exp \left(-n\theta \left(1 - \left(1 - \frac{1}{n} \right)^2 \right) \right) - e^{-2\theta} = e^{-2\theta} \left(e^{\theta/n} - 1 \right). \end{aligned}$$

La borne FDCR n'est donc pas atteinte. L'estimateur δ_{S_n} est donc optimal mais pas efficace. ▲

Exercice VIII.4. 1. La densité de la loi $\beta(1, 1/\theta)$ peut s'écrire :

$$p(x; \theta) = \frac{1}{\theta} (1-x)^{-1} \exp\left(\frac{1}{\theta} \log(1-x)\right) \mathbf{1}_{]0,1[}(x), \quad \theta \in \mathbb{R}_*^+.$$

Le modèle $\mathcal{P} = \{\beta(1, 1/\theta), \theta > 0\}$ 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 :

$$L_n(x_1, \dots, x_n; \theta) = \left(\frac{1}{\theta} - 1\right) \sum_{i=1}^n \log(1-x_i) - n \log(\theta) + \prod_{i=1}^n \log(\mathbf{1}_{]0,1[}(x_i)).$$

La log-vraisemblance vaut $-\infty$ sur la frontière de $]0, +\infty[$. Son maximum est donc atteint au point θ qui annule sa dérivée. On obtient l'estimateur du maximum 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 :

$$\mathbb{E}_\theta[h(-\log(1 - X_i))] = \int_0^1 h(-\log(1 - x)) \frac{1}{\theta} (1 - x)^{\frac{1}{\theta} - 1} dx = \int_0^\infty h(y) \frac{1}{\theta} e^{-y/\theta} dy.$$

On en déduit donc que $-\log(1 - X_i)$ suit une loi exponentielle de paramètre $1/\theta$.

4. a) L'estimateur T_n est sans biais car $\mathbb{E}_\theta[T_n] = \mathbb{E}_\theta[-\log(1 - X_1)] = \theta$.

b) Le risque quadratique de T_n est :

$$R(T_n, \theta) = \mathbb{E}_\theta[T_n - \theta]^2 = \text{Var}_\theta(T_n) = \frac{\theta^2}{n}.$$

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-Sheffé, on en déduit que T_n est un estimateur optimal de θ .

e) On calcule l'information de Fisher :

$$I_1(\theta) = -\mathbb{E} \left[\frac{\partial^2}{\partial \theta^2} L_1(X_1, \theta) \right] = -\frac{2}{\theta^3} \mathbb{E}[\log(1 - X_1)] - \frac{1}{\theta^2} = \frac{1}{\theta^2}.$$

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. Les variables aléatoires $(-\log(1 - X_i), i \geq 1)$ sont également de carré intégrable. Par le théorème central limite, la suite $(\sqrt{n}(T_n - \theta), n \geq 1)$ converge en loi vers la loi gaussienne centrée de variance θ^2 . L'estimateur T_n est donc asymptotiquement normal 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 efficace, *i.e.* asymptotiquement normal de variance l'inverse de l'information de Fisher : $1/I_1(\theta) = \theta^2$.

▲

Exercice VIII.5. 1. La densité du couple (Z_1, Y_1) est :

$$p_1(z_1, y_1; \lambda, \mu) = \lambda \mu e^{(-\lambda z_1 - \mu y_1)} \mathbf{1}_{\{z_1 > 0, y_1 > 0\}}.$$

Il s'agit d'un modèle exponentiel. La vraisemblance de l'échantillon de taille n vaut pour $z = (z_1, \dots, z_n)$ et $y = (y_1, \dots, y_n)$:

$$p_n(z, y; \lambda\mu) = \lambda^n \mu^n e^{\left(-\lambda \sum_{i=1}^n z_i - \mu \sum_{i=1}^n y_i\right)} \left(\prod_{i=1}^n \mathbf{1}_{\{z_i > 0, y_i > 0\}} \right).$$

Une statistique exhaustive (et totale) est $T = \left(\sum_{i=1}^n Z_i, \sum_{i=1}^n Y_i \right)$.

2. 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 la dérive 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 $\partial_\lambda^2 L_n(z, y; \lambda, \mu) = -n / \lambda^2$, on en déduit que la fonction $\lambda \mapsto L_n(z, y; \lambda, \mu)$ 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 $\hat{\mu}_n = n / \sum_{i=1}^n Y_i$.

3. 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 :

$$\sqrt{n} \begin{pmatrix} \hat{\lambda}_n - \lambda \\ \hat{\mu}_n - \mu \end{pmatrix} \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda^2 & 0 \\ 0 & \mu^2 \end{pmatrix} \right).$$

4. Par indépendance, on a :

$$\mathbb{P}(X_i > t) = \mathbb{P}(Z_i > t, Y_i > t) = \mathbb{P}(Z_i > t) \mathbb{P}(Y_i > t) = \exp(-(\lambda + \mu)t),$$

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 $\gamma = \lambda + \mu$.

5. Le modèle statistique est $\mathcal{P} = \{\mathcal{E}(\lambda + \mu), \lambda > 0, \mu > 0\}$, où $\mathcal{E}(\gamma)$ désigne la loi exponentielle de paramètre γ . 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 :

$$\prod_{i=1}^n p(x_i; \gamma) = \prod_{i=1}^n \gamma \exp(-\gamma x_i) \mathbf{1}_{\{x_i > 0\}}.$$

6. Cas a : par analogie avec la question 2, on trouve $\hat{\gamma}_n = n / \sum_{i=1}^n X_i$.

Cas b : 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.

7. Comme pour la question 3, on a :

$$\sqrt{n}(\hat{\gamma}_n - \lambda - \mu) \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}(0, (\lambda + \mu)^2).$$

On déduit de la question 3 que :

$$\sqrt{n}(\hat{\lambda}_n + \hat{\mu}_n - \lambda - \mu) \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}(0, \lambda^2 + \mu^2).$$

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 asymptotiquement préférable à $\hat{\gamma}_n$. (En fait on peut facilement vérifier que pour $n \geq 3$ les estimateurs sont de carré intégrable et que l'estimateur $\hat{\lambda}_n + \hat{\mu}_n$ est 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 $i = 1, \dots, n$. Les X_i sont alors indépendantes et identiquement distribuées selon une loi binomiale de même paramètre (N, θ) , N connu, $\theta \in (0, 1)$ inconnu. On peut écrire :

$$\mathbb{P}(X_i = k) = \binom{N}{k} \theta^k (1 - \theta)^{N-k} = (1 - \theta)^N \binom{N}{k} \exp\left(\log\left(\frac{\theta}{1 - \theta}\right) k\right)$$

et conclure qu'il s'agit d'un modèle exponentiel.

2. La statistique canonique du modèle est $S(X) = \sum_{i=1}^n X_i$. Elle est exhaustive et totale. Elle est de loi binomiale de paramètres (Nn, θ) .
3. L'estimateur $\delta = \mathbf{1}_{\{X_1 \leq k\}}$ est un estimateur sans biais de $\mathbb{P}_\theta(X_i \leq k)$.
4. On utilise l'amélioration de Rao-Blackwell. On calcule d'abord pour $k \in \{\max(0, N + S - nN), \dots, \min(N, s)\}$:

$$\begin{aligned} \mathbb{P}_\theta(X_1 = k | S = s) &= \frac{\mathbb{P}_\theta(X_1 = k, S - X_1 = s - k)}{\mathbb{P}_\theta(S = s)} \\ &= \frac{\binom{N}{k} \theta^k (1 - \theta)^{N-k} \binom{N(n-1)}{s-k} \theta^{s-k} (1 - \theta)^{N(n-1)-(s-k)}}{\binom{Nn}{s} \theta^s (1 - \theta)^{Nn-s}} \\ &= \frac{\binom{N}{k} \binom{Nn-N}{s-k}}{\binom{Nn}{s}}. \end{aligned}$$

La loi de X_1 sachant $\{S = s\}$ est la loi hypergéométrique de paramètre (Nn, s, N) . Donc, on obtient :

$$\delta_S = \mathbb{E}[\delta | S] = \sum_{j=0}^k \mathbb{E}[\mathbf{1}_{\{X_1=j\}} | S] = \sum_{j=\max(0, N+S-nN)}^k \frac{\binom{N}{j} \binom{Nn-N}{S-j}}{\binom{Nn}{S}}.$$

D'après le théorème de Lehman-Sheffé, δ_S est optimal.

Lorsque k varie, δ_S est la valeur en k de la fonction de répartition de la loi hypergéométrique de paramètres (Nn, S, N) . ▲

Exercice VIII.7. 1. La vraisemblance est $p(x; \theta) = (1 - \theta)^{x-1} \theta$. Il s'agit d'un modèle exponentiel avec $S(X) = X$ comme statistique canonique. La statistique canonique S est exhaustive et totale.

2. On a $\frac{\partial}{\partial \theta} \log p(x; \theta) = \frac{1}{\theta} - \frac{x-1}{1-\theta}$ et $\frac{\partial^2}{\partial^2 \theta} \log p(x; \theta) = -\frac{1}{\theta^2} - \frac{(x-1)}{(1-\theta)^2}$. On en déduit :

$$I(\theta) = \frac{1}{\theta^2} + \frac{1}{(1-\theta)^2} (\mathbb{E}_\theta[X] - 1) = \frac{1}{\theta^2} + \frac{1}{(1-\theta)\theta} = \frac{1}{\theta^2(1-\theta)}.$$

3. La vraisemblance du n échantillon est :

$$p_n(x_1, \dots, x_n; \theta) = \theta^n (1 - \theta)^{\sum_{i=1}^n x_i - n}.$$

On regarde les zéros de la dérivée en θ de la log-vraisemblance $L_n = \log p_n$, et on vérifie que $\hat{\theta}_n = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}_n}$ est l'estimateur du maximum de vraisemblance de θ .

4. On peut appliquer le théorème central limite. Comme $\text{Var}_\theta(X) = (1 - \theta)/\theta^2$, il vient :

$$\sqrt{n} \left(\bar{X}_n - \frac{1}{\theta} \right) \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N} \left(0, \frac{1 - \theta}{\theta^2} \right).$$

Comme la fonction $h(u) = 1/u$ est de classe \mathcal{C}^1 pour $u > 0$, on a la convergence en loi suivante :

$$\sqrt{n}\left(\frac{1}{\bar{X}_n} - \theta\right) \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}\left(0, \frac{1-\theta}{\theta^2}\theta^4\right) = \mathcal{N}(0, \theta^2(1-\theta)).$$

(Le modèle étant régulier, on pouvait également en déduire que l'EMV est asymptotiquement efficace.)

5. La suite $(\hat{\theta}_n \sqrt{1 - \hat{\theta}_n}, n \geq 1)$ converge presque sûrement vers $\theta \sqrt{1 - \theta}$ d'une part, et la suite $\left(\frac{\sqrt{n}}{\theta \sqrt{1 - \theta}}(\hat{\theta}_n - \theta), n \geq 1\right)$ converge en loi vers une gaussienne centrée réduite $\mathcal{N}(0, 1)$ d'autre part. On déduit du théorème de Slutsky que :

$$\frac{\sqrt{n}}{\hat{\theta}_n \sqrt{1 - \hat{\theta}_n}}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}(0, 1).$$

On note ϕ_α le quantile d'ordre α de la loi gaussienne centrée $\mathcal{N}(0, 1)$. Un intervalle de confiance pour θ de niveau asymptotique $1 - \alpha$ est :

$$\left[\hat{\theta}_n \pm \phi_{1-\alpha/2} \frac{\hat{\theta}_n \sqrt{1 - \hat{\theta}_n}}{\sqrt{n}} \right].$$

6. Les valeurs numériques sont $n = 40$ et $\sum_{i=1}^n x_i = 1779$. On en déduit l'estimation $\hat{\theta}_n \simeq 0.0225$ et l'intervalle de confiance de niveau asymptotique 95% (avec $\alpha = 5\%$ et $\phi_{1-\alpha/2} \simeq 1.96$) : $[0.016; 0.028]$. Le nombre de fraudeurs estimé est $n_f \simeq n_0 \hat{\theta}_n \in [320; 560]$.

▲

Exercice VIII.8. 1. Notons x_1, \dots, x_n les $n = 8$ observations. La vraisemblance s'écrit :

$$p_n(x_1, \dots, x_n; a) = \frac{1}{a^n} \left(\prod_{i=1}^n x_i \right) \exp\left(-\frac{1}{2a} \sum_{i=1}^n x_i^2\right).$$

Son maximum est atteint pour :

$$\hat{a}_n = \frac{1}{2n} \sum_{i=1}^n x_i^2.$$

L'application numérique donne $\hat{a}_8 \simeq 2.42$.

2. On vérifie que l'estimateur \hat{a}_n est :

a) Sans biais.

b) Optimal. En effet, le modèle est exponentiel et la statistique exhaustive et totale est $S_n = \frac{1}{2} \sum_{i=1}^n X_i^2$. L'estimateur sans biais est fonction de la statistique exhaustive et totale. Il est donc optimal.

- c) Efficace. Le modèle exponentiel sous sa forme naturelle donne $\lambda = -1/a$ et $\varphi(\lambda) = \log(-1/\lambda)$. Dans un modèle exponentiel, sous sa forme naturelle, S_n est un estimateur efficace de $\varphi'(\lambda) = -1/\lambda = a$.
- d) Asymptotiquement normal. Une application du théorème central limite assure que $(\sqrt{n}(\hat{a}_n - a), n \geq 1)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, a^2)$.
3. La probabilité qu'une catastrophe se produise durant une année est donnée par :

$$p(a) = \int_6^{+\infty} \frac{x}{a} \exp\left(-\frac{x^2}{2a}\right) dx = e^{-18/a}.$$

Elle est estimée par $\hat{p}_n = p(\hat{a}_n)$, soit $\hat{p}_8 \simeq 5.8 \cdot 10^{-4}$. Mais il faut en donner un intervalle de confiance. Comme p est une fonction croissante de a et que l'on souhaite majorer p , on recherche un intervalle de confiance de a de la forme $[0, c]$. On déduit de la normalité asymptotique de \hat{a}_n qu'un intervalle de confiance de niveau asymptotique $1 - \alpha$ est donné par $I_n = [0, \hat{a}_n(1 + \phi_{1-\alpha}/\sqrt{n})]$, où ϕ_α est le quantile d'ordre α de la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$. Pour $\alpha = 5\%$, on a $u_\alpha \simeq 1.65$ et $I_8 \simeq [0; 3.83]$. Et donc un intervalle de confiance de niveau asymptotique 95% de p est $[0; 9.1 \cdot 10^{-3}]$.

Par indépendance, la probabilité d'avoir strictement plus d'une catastrophe en mille ans vaut :

$$q(a) = 1 - ((1 - p(a))^{1000} + 1000p(a)(1 - p(a))^{999}).$$

On obtient $q(\hat{a}_8) \simeq 11\%$ et l'intervalle de confiance à 95% de $q(a)$ est $[0; 0.999]$. L'incertitude due au petit nombre d'observations est donc très importante. ▲

Exercice VIII.9. 1. La loi de X_i est la loi $\mathcal{N}(m_i + \alpha, \sigma^2)$. La log-vraisemblance s'écrit :

$$L_n(x_1, \dots, x_n; \alpha) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - m_i - \alpha)^2.$$

On obtient :

$$\frac{\partial}{\partial \alpha} L_n(x_1, \dots, x_n; \alpha) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - m_i - \alpha),$$

et donc :

$$\frac{\partial}{\partial \alpha} L_n(x_1, \dots, x_n; \alpha) = 0 \iff \alpha = \frac{1}{n} \sum_{i=1}^n (x_i - m_i).$$

L'étude du signe de la dérivée de la log-vraisemblance, montre qu'elle atteint son maximum en $\frac{1}{n} \sum_{i=1}^n (x_i - m_i)$. 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 qu'il est efficace. On a également :

$$\frac{\partial^2}{\partial^2 \alpha} L_n(x_1, \dots, x_n; \alpha) = -\frac{n}{\sigma^2},$$

et, on en déduit que l'information de Fisher est :

$$I_n = \mathbb{E}_\alpha \left[-\frac{\partial^2}{\partial^2 \alpha} L_n(X_1, \dots, X_n; \alpha) \right] = \frac{n}{\sigma^2}.$$

Comme $\text{Var}_\alpha(\hat{\alpha}_n) = \frac{\sigma^2}{n} = \frac{1}{I_n}$, 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 nombres, 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 $\mathcal{N}(\beta, \frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{m_i^2})$. Ainsi $\tilde{\beta}_n$ est un estimateur sans biais de β . On vérifie qu'il est efficace. La log-vraisemblance s'écrit :

$$L_n(x_1, \dots, x_n; \beta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \beta m_i)^2.$$

On a :

$$\frac{\partial^2}{\partial^2 \beta} L_n(x_1, \dots, x_n; \beta) = -\frac{1}{\sigma^2} \sum_{i=1}^n m_i^2,$$

et donc :

$$I_n = \mathbb{E}_\beta \left[-\frac{\partial^2}{\partial^2 \beta} L_n(X_1, \dots, X_n; \beta) \right] = \frac{1}{\sigma^2} \sum_{i=1}^n m_i^2.$$

D'autre part on a :

$$\text{Var}_\beta(\tilde{\beta}_n) = \frac{\sigma^2}{n^2} \sum_{i=1}^n \frac{1}{m_i^2}.$$

Par Cauchy-Schwarz, on a $\frac{1}{n^2} \sum_{i=1}^n \frac{1}{m_i^2} \geq \frac{1}{\sum_{i=1}^n m_i^2}$, et l'inégalité est stricte dès qu'il existe $m_i \neq m_j$. En particulier $\text{Var}_\beta(\tilde{\beta}_n) > \frac{1}{I_n}$, s'il existe $m_i \neq m_j$, et l'estimateur n'est alors pas efficace.

3. On a :

$$\frac{\partial}{\partial \beta} L_n(x_1, \dots, x_n; \beta) = \frac{1}{\sigma^2} \sum_{i=1}^n m_i (x_i - \beta m_i).$$

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 $\mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n m_i^2}\right)$. En particulier, cet estimateur est sans biais et il est efficace. Il est préférable à $\tilde{\beta}_n$.

4. On obtient les estimations avec les intervalles de confiance de niveau exact 95% : $\hat{\alpha}_n \simeq 88.6 \pm 6.1$, $\tilde{\beta}_n \simeq 1.088 \pm 0.006$ et $\hat{\beta}_n \simeq 1.087 \pm 0.006$. (La théorie des tests permet de déterminer lequel des deux effets (additif ou multiplicatif) modélise au mieux les données observées.)

▲

Exercice VIII.10. 1. On a $\pi = \mathbb{P}(X_k = 1) = \mathbb{P}(Z_k = 0) = e^{-\lambda}$ et $\mathbb{P}(X_k = 0) = 1 - e^{-\lambda}$. La loi de X_k est la loi de Bernoulli de paramètre π . La loi de Y , comme somme de variables aléatoires indépendantes de même loi de Bernoulli de paramètre π , suit une loi binomiale $\mathcal{B}(n, \pi)$ de paramètre (n, π) .

2. La vraisemblance s'écrit : $p(y; \pi) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$, avec $y = \sum_{i=1}^n x_i$. La log-vraisemblance est donnée par :

$$L(y; \pi) = \log(p(y; \pi)) = \log \binom{n}{y} + y \log(\pi) + (n - y) \log(1 - \pi).$$

On a $\frac{\partial}{\partial \pi} L(y; \pi) = \frac{y}{\pi} - \frac{n - y}{1 - \pi}$. On vérifie que l'estimateur du maximum de vraisemblance de π est $\hat{\pi} = Y/n$. On en déduit l'estimateur du maximum de vraisemblance de λ :

$$\hat{\lambda} = -\log(\hat{\pi}) = -\log(Y/n).$$

3. On remarque que $\hat{\pi}$ est un estimateur convergent de π (quand n tend vers l'infini). Le théorème central limite et le théorème de Slutsky assurent que l'intervalle de confiance pour π :

$$IC_{1-\alpha}(\pi) = \left[\hat{\pi} \pm \phi_{1-\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right],$$

où ϕ_r désigne le quantile d'ordre r de la loi normale centrée réduite $\mathcal{N}(0, 1)$, est de niveau asymptotique $1 - \alpha$. On en déduit que l'intervalle de confiance pour λ de niveau asymptotique $1 - \alpha$ est :

$$IC_{1-\alpha}(\lambda) = \left[-\log \left(\hat{\pi} \mp \phi_{1-\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \right) \right].$$

4. On trouve : $\hat{\pi} = 0.6$, $\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \simeq 0.155$, $\hat{\lambda} \simeq 0.51$, $IC_{95\%}(\pi) = [0.3; 0.9]$ et $IC_{95\%}(\lambda) = [0.11; 1.20]$.
5. Si la densité λ est très forte (resp. faible) alors π est très proche de 0 (resp. 1), ce qui implique que la probabilité d'avoir des tubes négatifs est très proche de 0 (resp. 1). L'estimateur de π , et donc de λ , est très mauvais car l'approximation du théorème de la limite centrale n'est pas valide. (En effet la borne dans le théorème de Berry-Esséen indique qu'il faut $np(1-p)$ grand pour une bonne approximation de la loi binomiale $\mathcal{B}(n, p)$ par la loi gaussienne $\mathcal{N}(np, np(1-p))$.)
6. La loi de Y_i est la loi binomiale $\mathcal{B}(n_i, \pi_i)$ où $\pi_i = e^{-\lambda d_i}$.
7. La vraisemblance s'écrit :

$$\begin{aligned} p(y_1, \dots, y_N; \lambda) &= \prod_{i=1}^N \binom{n_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i} \\ &= \prod_{i=1}^N \binom{n_i}{y_i} e^{-\lambda d_i y_i} (1 - e^{-\lambda d_i})^{n_i - y_i}, \end{aligned}$$

et la log-vraisemblance :

$$\begin{aligned} L(y_1, \dots, y_N; \lambda) &= \log p(y_1, \dots, y_N; \lambda) \\ &= \sum_{i=1}^N \log \left(\binom{n_i}{y_i} \right) + \sum_{i=1}^N y_i \log(\pi_i) + \sum_{i=1}^N (n_i - y_i) \log(1 - \pi_i) \\ &= \sum_{i=1}^N \log \left(\binom{n_i}{y_i} \right) - \lambda \sum_{i=1}^N y_i d_i + \sum_{i=1}^N (n_i - y_i) \log(1 - e^{-\lambda d_i}). \end{aligned}$$

La dérivée de la log-vraisemblance s'écrit donc :

$$\frac{\partial L(y_1, \dots, y_N; \lambda)}{\partial \lambda} = - \sum_{i=1}^N y_i d_i + \sum_{i=1}^N (n_i - y_i) d_i \frac{e^{-\lambda d_i}}{(1 - e^{-\lambda d_i})}.$$

Elle est décroissante, elle tend vers $+\infty$ en 0 et elle est négative en $+\infty$. Elle s'annule en un seul point λ , unique solution de :

$$\sum_{i=1}^N y_i d_i = \sum_{i=1}^N (n_i - y_i) d_i \left(\frac{e^{-\lambda d_i}}{1 - e^{-\lambda d_i}} \right).$$

La résolution de cette équation est en général numérique. L'estimateur obtenu est appelé MPN pour "most probable number".

8. L'information de Fisher du modèle est :

$$I_N(\lambda) = -\mathbb{E} \left[\frac{\partial^2}{\partial \lambda^2} L(Y_1, \dots, Y_N; \lambda) \right] = \sum_{i=1}^N n_i d_i^2 \frac{e^{-\lambda d_i}}{1 - e^{-\lambda d_i}}.$$

En s'inspirant des démonstrations pour les modèles réguliers, on peut montrer que la variance de l'estimateur du maximum de vraisemblance est asymptotiquement équivalente (quand $\sum_{i=1}^N n_i$ tend vers l'infini) à :

$$V(\lambda) = \frac{1}{I_N(\lambda)} = \left(\sum_{i=1}^N n_i d_i^2 \frac{e^{-\lambda d_i}}{1 - e^{-\lambda d_i}} \right)^{-1}.$$

9. Avec $n = n_1 = n_2$, on a :

$$\frac{\partial}{\partial \lambda} L(y_1, y_2; \lambda) = -y_1 + (n - y_1) \frac{e^{-\lambda}}{1 - e^{-\lambda}} - \frac{y_2}{2} + \frac{n - y_2}{2} \frac{e^{-\frac{\lambda}{2}}}{1 - e^{-\frac{\lambda}{2}}}.$$

Après avoir résolu une équation du second degré, on obtient :

$$\hat{\lambda} = 2 \log(6n) - 2 \log \left(- (n - Y_2) + \sqrt{(n - Y_2)^2 + 12n(2Y_1 + Y_2)} \right).$$

On peut montrer que dans ce modèle l'intervalle de confiance pour λ :

$$IC'_{1-\alpha}(\lambda) = \left[\hat{\lambda} \pm \phi_{1-\alpha/2} \sqrt{V(\hat{\lambda})} \right]$$

est de niveau asymptotique $1 - \alpha$. Numériquement, on obtient : $\hat{\pi} \simeq 0.569$, $\hat{\lambda} \simeq 1.23$ et $IC'_{95\%}(\lambda) \simeq [0.44; 1.81]$. ▲

XIII.9 Tests

Exercice IX.1. 1. Le test de Neyman de niveau α pour tester l'hypothèse simple $H_0 = \{\text{type 2}\}$ contre l'hypothèse alternative simple $H_1 = \{\text{type 1}\}$ est UPP. Il est défini par la région critique :

$$W_1 = \{(x, y); p_1(x, y)/p_2(x, y) \geq k_\alpha\} = \{(x, y) \in [-2; 2]^2; x^2 + y^2 \geq z_{1-\alpha}\},$$

où $z_{1-\alpha}$ (et donc k_α) est défini par $\mathbb{P}_{H_0}(X^2 + Y^2 \leq z_{1-\alpha}) = \alpha$. Sous l'hypothèse H_0 , $X^2 + Y^2$ suit une loi du $\chi^2(2)$; donc $z_{1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi du $\chi^2(2)$. Pour $\alpha = 5\%$, on obtient $z_{1-\alpha} \simeq 5.99$ et la région critique est l'intersection de l'extérieur du disque de rayon $\sqrt{5.99}$ avec le carré $[-2; 2]^2$.

2. Pour $\alpha = 1\%$, on obtient $z_{1-\alpha} \simeq 9.21$ et la région critique est vide. On accepte toujours H_0 . Dans ce cas une seule observation ne permet pas de rejeter H_0 avec une erreur de première espèce inférieure à 1% .

▲

Exercice IX.2. 1. La vraisemblance de l'échantillon de taille 1 est :

$$p(x_1; \theta) = \frac{1}{\theta} e^{-x_1/\theta} \mathbf{1}_{\{x_1 > 0\}}.$$

Le modèle est exponentiel de la forme :

$$p(x_1; \theta) = C(\theta)h(x_1)e^{Q(\theta)S(x_1)},$$

avec $S(x_1) = x_1$ et $Q(\theta) = 1/\theta$. La statistique de test canonique est $S_n(x) = \sum_{i=1}^n x_i$, où $x = (x_1, \dots, x_n)$. On remarque que la loi de $S_n(X)$ est la loi gamma de paramètre $(1/\theta, n)$. En particulier elle est à densité. On en déduit que le test UPP de niveau α est donné par la région critique :

$$W_n = \{x; S_n(x) \geq \gamma_{1-\alpha}\},$$

où $\gamma_{1-\alpha}$ est tel que $\mathbb{P}_{\theta_0}(W_n) = \mathbb{P}_{\theta_0}(S_n(X) \geq \gamma_{1-\alpha}) = \alpha$. On obtient que $\gamma_{1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi $\Gamma(1/\theta_0, n)$. On remarque que sous H_0 , $Z_n = 2S_n(X)/\theta_0$ est de loi $\Gamma(1/2, n)$ c'est à dire de loi $\chi^2(2n)$. Autrement dit $\gamma_{1-\alpha} = \theta_0 z_{1-\alpha}/2$, où $z_{1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi χ^2 à $2n$ degrés de liberté.

2. Le test bilatère UPPS de niveau α est donné par la région critique :

$$W_n = \{x; S_n(x) \notin]c_1, c_2[\},$$

avec $\mathbb{P}_{\theta_0}(W_n) = \alpha$. Cette seule équation ne permet pas de déterminer les constantes c_1 et c_2 . La puissance du test est :

$$\rho(\theta) = \mathbb{P}_\theta(S_n(X) \notin]c_1, c_2[) = \mathbb{P}(Z_n \notin]2c_1/\theta, 2c_2/\theta[) = 1 - \int_{2c_1/\theta}^{2c_2/\theta} f_{2n}(x) dx,$$

où Z_n est de loi $\chi^2(2n)$ de densité f_{2n} . Par définition, la puissance est minimale pour $\theta = \theta_0$ et de plus $\rho(\theta_0) = \alpha$. Les constantes c_1 et c_2 sont donc déterminées par $\rho(\theta_0) = \alpha$ et $\rho'(\theta_0) = 0$. La condition $\rho'(\theta_0) = 0$ implique :

$$c_1^n e^{-c_1/\theta_0} = c_2^n e^{-c_2/\theta_0}. \quad (\text{XIII.4})$$

On peut alors résoudre numériquement les deux équations et obtenir c_1 et c_2 . On peut cependant donner une approximation de c_1 et c_2 . Le théorème central limite implique que $\sqrt{n}(S_n(X)/n - \theta)$ converge en loi vers θG , où G est de loi gaussienne $\mathcal{N}(0, 1)$. On en déduit que :

$$\mathbb{P}_\theta(S_n(X) \in]c_1, c_2]) \simeq \mathbb{P}(G \in]\sqrt{n}(c_1/\theta n - 1), \sqrt{n}(c_2/\theta n - 1)[). \quad (\text{XIII.5})$$

Une solution approchée (pour n grand) consiste alors à choisir $c'_1 = n\theta_0(1 - \phi_{1-\alpha/2}/\sqrt{n})$ comme approximation de c_1 et $c'_2 = n\theta_0(1 + \phi_{1-\alpha/2}/\sqrt{n})$ comme approximation de c_2 , où ϕ_r est le quantile d'ordre r de la loi gaussienne $\mathcal{N}(0, 1)$. On remarque que c'_1 et c'_2 satisfont l'égalité (XIII.4) à $O(1/\sqrt{n})$ près et que (XIII.5) implique que $\rho(\theta_0) = \alpha$ à $o(1)$ près (en fait à $O(1/\sqrt{n})$ près en utilisant le développement d'Edgeworth). ▲

Exercice IX.3. 1. Le vecteur des paramètres est $\theta = (\mu_1, \mu_2, \sigma_1, \sigma_2)^t$. On utilise 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 est $\hat{\theta}_n = (\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2)^t$ avec :

$$\hat{\mu}_1 = \frac{1}{n} \sum_{i=1}^n X_i, \quad \hat{\mu}_2 = \frac{1}{n} \sum_{i=1}^n Y_i, \quad \hat{\sigma}_1^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu}_1)^2, \quad \hat{\sigma}_2^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\mu}_2)^2.$$

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 en déduit la matrice d'information de Fisher :

$$I(\theta) = \begin{pmatrix} 1/\sigma_1^2 & 0 & 0 & 0 \\ 0 & 1/\sigma_2^2 & 0 & 0 \\ 0 & 0 & 2/\sigma_1^2 & 0 \\ 0 & 0 & 0 & 2/\sigma_2^2 \end{pmatrix}.$$

La variance asymptotique de l'estimateur $g(\hat{\theta}_n)$ de $\mu_1 - \mu_2$ est :

$$\Sigma(\theta) = \frac{\partial g}{\partial \theta}(\theta) I^{-1}(\theta) \left(\frac{\partial g}{\partial \theta}(\theta) \right)^t = \sigma_1^2 + \sigma_2^2.$$

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. Sous H_1 , ζ_n converge p.s. vers $+\infty$. On en déduit qu'une région critique de niveau asymptotique α est donnée par $\{\zeta_n > z_{1-\alpha}\}$, où z_r est le quantile d'ordre r de la loi $\chi^2(1)$. Le test est convergent. La p -valeur asymptotique (non-uniforme) du test est donnée par $\mathbb{P}(V \geq \zeta_n^{\text{obs}})$, où V suit la loi $\chi^2(1)$ et ζ_n^{obs} est la valeur de la statistique de test évaluée sur les données.

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 . On remarque 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 $Z_{2n-2} = n(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/\sigma^2$ 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 $Z'_1 = \frac{n(\hat{\mu}_1 - \hat{\mu}_2)^2}{2\sigma^2}$ a pour loi $\chi^2(1)$. On remarque que $\zeta_n = 2n \frac{Z'_1}{Z_{2n-2}}$.

On déduit de l'indépendance de $\hat{\mu}_1$ avec $\hat{\sigma}_1^2$ et de $\hat{\mu}_2$ avec $\hat{\sigma}_2^2$, que Z'_1 et Z_{2n-2} sont indépendants. Ainsi la loi de $\frac{n-1}{n}\zeta_n = \frac{Z'_1}{Z_{2n-2}/(2n-2)}$ est la loi de Fisher-Snedecor de paramètre $(1, 2n-2)$. La région critique de niveau exact α est donnée par $\{\zeta_n > \frac{n}{n-1}z'_{1-\alpha}\}$, où z'_r est le quantile d'ordre r de la loi de Fisher-Snedecor de paramètre $(1, 2n-2)$.

Pour $n = 15$ et $\alpha = 5\%$, on a $z_{1-\alpha} \simeq 3.84$ et $z'_{1-\alpha} \simeq 4.20$ soit $\frac{n}{n-1}z'_{1-\alpha} \simeq 4.50$.

On remarque par ailleurs 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.4. 1. La vraisemblance du modèle est, pour $x = (x_1, \dots, x_n) \in \mathbb{R}^n$:

$$p_n(x; \theta) = \prod_{i=1}^n \frac{e^{-(x_i - \theta)^2/2\theta}}{\sqrt{2\pi\theta}},$$

et la log-vraisemblance est donnée par :

$$L_n(x; \theta) = \log(p_n(x; \theta)) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\theta) - \sum_{i=1}^n \frac{(x_i - \theta)^2}{2\theta}.$$

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 $\lim_{\theta \rightarrow 0} L_n(x; \theta) = \lim_{\theta \rightarrow \infty} L_n(x; \theta) = -\infty$, 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'estimateurs $(\hat{\theta}_n, n \geq 1)$ converge p.s. vers $f(\mathbb{E}_\theta[X_1^2]) = f(\theta + \theta^2) = \theta$. La suite d'estimateurs est donc convergente. Les variables $(X_n^2, n \geq 1)$ sont indépendantes et de carré intégrable avec :

$$\text{Var}_\theta(X_k^2) = \mathbb{E}[X_k^4] - \mathbb{E}[X_k^2]^2 = \theta^4 + 6\theta^3 + 3\theta^2 - (\theta + \theta^2)^2 = 4\theta^3 + 2\theta^2. \quad (\text{XIII.6})$$

De plus, la fonction f est de classe C^1 sur \mathbb{R}^+ . On en déduit donc du théorème central limite, que la suite d'estimateurs $(\hat{\theta}_n, n \geq 1)$ est asymptotiquement normale de variance asymptotique :

$$\sigma^2 = f'(\mathbb{E}_\theta[X_1^2])^2 \text{Var}_\theta(X_1^2) = \frac{2\theta^2}{1 + 2\theta}.$$

De plus, on a :

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} L_1(X_1; \theta) \right] = -\frac{1}{2\theta^2} + \frac{1}{\theta^3} \mathbb{E}_\theta[X_1^2] = \frac{1 + 2\theta}{2\theta^2}.$$

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 $(\zeta'_n, n \geq 1)$ converge en loi vers la loi 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 données pour $n \geq 1$ par :

$$W'_n = \{\zeta'_n \geq c\}$$

définissent un test asymptotique convergent de niveau asymptotique $\alpha = \mathbb{P}(G > c)$, où G est de loi gaussienne $\mathcal{N}(0, 1)$.

3. Par la loi forte des grands nombres, la suite $(\bar{X}_n, n \geq 1)$ converge p.s. 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 p.s. vers $\mathbb{E}_\theta[X_1^2]$. On en déduit que la suite $(V_n, n \geq 2)$ converge p.s. vers $\mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2 = \text{Var}_\theta(X_1) = \theta$. On a :

$$\begin{aligned} \mathbb{E}_\theta[T_n^\lambda] &= \lambda \mathbb{E}_\theta[\bar{X}_n] + (1 - \lambda) \mathbb{E}_\theta[V_n] = \theta, \\ \text{Var}_\theta(T_n^\lambda) &= \lambda^2 \text{Var}_\theta(\bar{X}_n) + (1 - \lambda)^2 \text{Var}_\theta(V_n) + 2 \text{Cov}_\theta(\bar{X}_n, V_n) \\ &= \lambda^2 \frac{\theta}{n} + (1 - \lambda)^2 \frac{2\theta^2}{(n - 1)}, \end{aligned}$$

car \bar{X}_n et V_n sont indépendantes et $\frac{n-1}{\theta} V_n$ suit la loi $\chi^2(n-1)$ de moyenne $n-1$ et de variance $2(n-1)$. Les estimateurs T_n^λ sont sans biais car $\mathbb{E}_\theta[T_n^\lambda] = \theta$. 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 p.s. vers $\lambda \mathbb{E}_\theta[X_1] + (1 - \lambda) \text{Var}_\theta(X_1) = \theta$. La suite d'estimateurs $(T_n^\lambda, n \geq 2)$ 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 $\mathbb{E}_\theta[X_k^2] = \text{Var}_\theta(X_k) + \mathbb{E}_\theta[X_k]^2 = \theta + \theta^2$. 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, en utilisant

$$\begin{aligned} \text{Var}_\theta(X_k) &= \theta, \\ \text{Cov}_\theta(X_k, X_k^2) &= \mathbb{E}[X_k^3] - \mathbb{E}_\theta[X_k] \mathbb{E}_\theta[X_k^2] \\ &= \theta^3 + 3\theta^2 - \theta(\theta + \theta^2) = 2\theta^2, \end{aligned}$$

et (XIII.6) que :

$$\Sigma = \begin{pmatrix} \theta & 2\theta^2 \\ 2\theta^2 & 4\theta^3 + 2\theta^2 \end{pmatrix}.$$

5. La suite $(\sqrt{n}(\bar{X}_n - \theta), n \geq 1)$ converge en loi vers la loi gaussienne $\mathcal{N}(0, \theta)$. (En fait la suite est constante en loi, égale en loi à $\mathcal{N}(0, \theta)$.)

On remarque que $\frac{n-1}{n} V_n$ est l'image de $(\bar{X}_n, \frac{1}{n} \sum_{k=1}^n X_k^2)$ par la fonction h , de classe C^1 définie par $h(y) = y_2 - y_1^2$, avec $y = (y_1, y_2)$. En particulier, la suite $(\sqrt{n}(\frac{n-1}{n} V_n - \theta), n \geq 2)$ converge en loi vers la loi gaussienne centrée de variance :

$$\frac{\partial h}{\partial y}(\theta, \theta + \theta^2) \Sigma \left(\frac{\partial h}{\partial y}(\theta, \theta + \theta^2) \right)^t = 2\theta^2.$$

Enfin, on remarque que $\sqrt{n}V_n - \sqrt{n}\frac{n-1}{n}V_n = \frac{V_n}{\sqrt{n}}$. En particulier, la suite $\left(\sqrt{n}V_n - \sqrt{n}\frac{n-1}{n}V_n, n \geq 2\right)$ 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 $U_n = \sqrt{n}(\bar{X}_n - \theta)$ et $W_n = \sqrt{n}(V_n - \theta)$. En utilisant l'indépendance entre U_n et W_n , on obtient, pour tout $u, w \in \mathbb{R}$:

$$\lim_{n \rightarrow \infty} \psi_{(U_n, W_n)}(u, w) = \lim_{n \rightarrow \infty} \psi_{U_n}(u)\psi_{W_n}(w) = e^{-\theta u^2/2} e^{-2\theta^2 w^2/2}.$$

On reconnaît la fonction caractéristique du couple (U, W) , où U 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 (U, W) . Par continuité de l'application $(a, b) \rightarrow \lambda a + (1 - \lambda)b$, 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 $\lambda U + (1 - \lambda)W$. Autrement dit, la suite $(\sqrt{n}(T_n^\lambda - \theta), n \geq 2)$ converge en loi vers la loi gaussienne $\mathcal{N}(0, \lambda^2\theta + 2(1 - \lambda)^2\theta^2)$. Ainsi la suite d'estimateurs $(\sqrt{n}(T_n^\lambda - \theta), n \geq 2)$ 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 $(\zeta_n^\lambda, n \geq 2)$ converge en loi vers la loi 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 :

$$W_n = \{\zeta_n^\lambda \geq c\}$$

définissent un test asymptotique convergent. Ces tests sont de niveau asymptotique $\alpha = \mathbb{P}(G \geq c)$, où G est de loi gaussienne $\mathcal{N}(0, 1)$.

7. On trouve $\lambda^* = 2\theta_0/(1 + 2\theta_0)$, $\sigma_{\lambda^*}^2 = \frac{2\theta_0^2}{1 + 2\theta_0}$ et $T_n^{\lambda^*} = \frac{1}{1 + 2\theta_0}(2\theta_0\bar{X}_n + V_n)$. Les suites d'estimateurs $(\hat{\theta}_n, n \geq 1)$ et $(T_n^{\lambda^*}, n \geq 2)$ ont même variance asymptotique. On a :

$$\hat{\theta}_n = -\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{n-1}{n}V_n + \bar{X}_n^2} = -\frac{1}{2} + \sqrt{\left(\theta + \frac{1}{2}\right)^2 + Q_n},$$

où

$$Q_n = \frac{n-1}{n}V_n - \theta + (\bar{X}_n^2 - \theta^2).$$

On déduit de ce qui précède que la suite $(Q_n, n \geq 2)$ converge p.s. vers 0. En utilisant $1 + x/2 - x^2/8 \leq \sqrt{1+x} \leq 1 + x/2$ pour $x \geq 0$, on obtient :

$$|\sqrt{1+x} - (1+x/2)| \leq x^2/8.$$

On en déduit que :

$$\hat{\theta}_n = \theta + \frac{Q_n}{1+2\theta} + g(Q_n),$$

où $|g(q)| \leq q^2/(1+2\theta)^3$. Il vient que sous H_0 :

$$\hat{\theta}_n - T_n^{\lambda^*} = \frac{1}{1+2\theta_0} \frac{1}{n-1} V_n + (\bar{X}_n - \theta_0)^2 + g(Q_n).$$

En particulier, pour tout $\varepsilon \in]0, 1/2[$, on a $\lim_{n \rightarrow \infty} n^{1-\varepsilon}(\hat{\theta}_n - T_n^{\lambda^*}) = 0$ en probabilité et donc $\lim_{n \rightarrow \infty} n^{1/2-\varepsilon}(\zeta'_n - \zeta_n^{\lambda^*}) = 0$ en probabilité. Ainsi les deux statistiques de test ζ'_n et $\zeta_n^{\lambda^*}$ définissent asymptotiquement les mêmes régions critiques.

8. 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 asymptotique de 2.3%. On obtient également $\zeta_n^{\lambda^*} = 1.86$ et une p -valeur asymptotique de 3.2%. Dans les deux cas, on rejette H_0 au niveau de 5%.

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 statistique 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 $\{\sum_{k=1}^n X_k^2 \geq c\}$, 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 est en fait UPP à horizon fini.

▲

Exercice IX.5. 1. La vraisemblance du modèle est :

$$p_n(x, \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^n (x_i - \theta)^2 / (2\sigma^2)}.$$

On pose $\mu_n = \sqrt{n}(\theta_1 - \theta_0)/\sigma$. Le rapport de vraisemblance est :

$$\frac{p_n(x, \theta_1)}{p_n(x, \theta_0)} = \exp\left(-\frac{1}{2}\mu_n^2 + \mu_n \zeta_n(x)\right),$$

avec la statistique de test :

$$\zeta_n(x) = \sqrt{n} \frac{\frac{1}{n} \sum_{i=1}^n x_i - \theta_0}{\sigma}.$$

Le test pur de Neyman est UPP. Comme $\theta_1 > \theta_0$ et donc $\mu_n > 0$, il est défini par sa région critique :

$$W_n = \{x; p_n(x, \theta_1)/p_n(x, \theta_0) \geq c\} = \{x; \zeta_n(x) \geq z\}.$$

On pose $\zeta_n = \zeta_n(X)$. Sous H_0 , la statistique de test ζ_n a même loi que G de loi gaussienne centrée réduite $\mathcal{N}(0, 1)$. Le niveau du test de Neyman est :

$$\mathbb{P}_{\theta_0}(W_n) = \mathbb{P}_{\theta_0}(\zeta_n \geq z) = \mathbb{P}(G \geq z).$$

Pour avoir un test UPP de niveau α , il faut choisir $z = \phi_{1-\alpha}$, où ϕ_r est le quantile d'ordre r de la loi $\mathcal{N}(0, 1)$.

2. L'erreur de deuxième espèce est :

$$1 - \mathbb{P}_{\theta_1}(W_n) = \mathbb{P}_{\theta_1}(\zeta_n < \phi_{1-\alpha}) = \mathbb{P}(G < \phi_{1-\alpha} - \mu_n).$$

On souhaite donc trouver n tel que $1 - \mathbb{P}_{\theta_1}(W_n) \leq \beta$ soit $\phi_{1-\alpha} - \mu_n \leq -\phi_{1-\beta}$, autrement dit :

$$n \geq \sigma^2 \left(\frac{\phi_{1-\alpha} + \phi_{1-\beta}}{\theta_1 - \theta_0} \right)^2.$$

3. Si ce test existe, il est en particulier plus puissant que le test de H_0 contre H_1 pour tout $\theta_1 \neq \theta_0$. Or pour $\theta_1 > \theta_0$ et $\theta_1 < \theta_0$, les régions critiques définies par le test UPP de Neyman sont distinctes. Il n'existe donc pas de test UPP de H_0 contre H'_1 .

4. L'ensemble W'_n peut aussi s'écrire :

$$W'_n = \left\{ x; e^{-\mu_n^2/2 + \mu_n \zeta_n(x)} \geq c(\theta_1) + c^*(\theta_1) \frac{\sqrt{n}}{\sigma} \zeta_n(x) \right\}.$$

En choisissant $c^*(\theta_1)$ strictement positif (resp. négatif) si $\theta_1 > \theta_0$ (resp. $\theta_1 < \theta_0$), on obtient que la fonction :

$$h(r) = e^{-\mu_n^2/2 + \mu_n r} - c(\theta_1) - c^*(\theta_1) \frac{\sqrt{n}}{\sigma} r$$

est strictement convexe et $\lim_{r \rightarrow \pm\infty} h(r) = +\infty$. On peut donc trouver deux constantes $c(\theta_1)$ et $c^*(\theta_1)$ tels que :

$$\{r; h(r) \geq 0\} =] -\phi_{1-\alpha/2}, \phi_{1-\alpha/2}[^c,$$

et cet ensemble est indépendant de θ_1 . On a alors :

$$\mathbb{P}_{\theta_0}(W'_n) = \mathbb{P}_{\theta_0}(|\zeta_n| \geq \phi_{1-\alpha/2}) = \alpha.$$

On a également :

$$\begin{aligned}
 \frac{\partial}{\partial \theta} \mathbb{P}_\theta(W'_n)_{|\theta=\theta_0} &= \frac{\partial}{\partial \theta} \mathbb{P}_\theta(|\zeta_n| \geq \phi_{1-\alpha/2})_{|\theta=\theta_0} \\
 &= \frac{\partial}{\partial \theta} \mathbb{P}_\theta(|\zeta_n| < \phi_{1-\alpha/2})_{|\theta=\theta_0} \\
 &= \frac{\partial}{\partial \theta} \mathbb{P} \left(\left| G - \frac{\theta_0 - \theta}{\sigma} \right| < \phi_{1-\alpha/2} \right)_{|\theta=\theta_0} \\
 &= \frac{\partial}{\partial \theta} \mathbb{P} \left(-\phi_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma} < G < \phi_{1-\alpha/2} + \frac{\theta_0 - \theta}{\sigma} \right)_{|\theta=\theta_0} \\
 &= f(\phi_{1-\alpha/2}) - f(-\phi_{1-\alpha/2}) = 0,
 \end{aligned}$$

où f est la densité de la loi de G .

5. Soit W''_n une région critique définissant un test pur de niveau α sans biais. On déduit de la régularité des densités gaussiennes que :

$$\frac{\partial}{\partial \theta} \mathbb{P}_\theta(W''_n)_{|\theta=\theta_0} = 0.$$

On en déduit $\mathbb{P}_{\theta_0}(W'_n \cap (W''_n)^c) = \mathbb{P}_{\theta_0}(W''_n \cap (W'_n)^c)$ et :

$$\frac{\partial}{\partial \theta} \mathbb{P}_\theta(W'_n \cap (W''_n)^c)_{|\theta=\theta_0} = \frac{\partial}{\partial \theta} \mathbb{P}_\theta(W''_n \cap (W'_n)^c)_{|\theta=\theta_0}.$$

Pour $x \in W'_n$, on a :

$$p_n(x, \theta) \geq c(\theta)p_n(x, \theta_0) + c^*(\theta) \frac{\partial p_n}{\partial \theta}(x; \theta_0).$$

En intégrant sur $W'_n \cap (W''_n)^c$, et en permutant l'intégrale en x et la dérivation en θ , il vient :

$$\begin{aligned}
 \mathbb{P}_\theta(W'_n \cap (W''_n)^c) &\geq c(\theta) \mathbb{P}_{\theta_0}(W'_n \cap (W''_n)^c) + c^*(\theta) \frac{\partial}{\partial \theta} \mathbb{P}_\theta(W'_n \cap (W''_n)^c)_{|\theta=\theta_0} \\
 &= c(\theta) \mathbb{P}_{\theta_0}(W''_n \cap (W'_n)^c) + c^*(\theta) \frac{\partial}{\partial \theta} \mathbb{P}_\theta(W''_n \cap (W'_n)^c)_{|\theta=\theta_0}.
 \end{aligned}$$

Pour $x \notin W'_n$, on a :

$$c(\theta)p_n(x, \theta_0) + c^*(\theta) \frac{\partial p_n}{\partial \theta}(x; \theta_0) > p_n(x, \theta).$$

En intégrant sur $W''_n \cap (W'_n)^c$, et en permutant l'intégrale en x et la dérivation en θ , il vient :

$$c(\theta) \mathbb{P}_{\theta_0}(W''_n \cap (W'_n)^c) + c^*(\theta) \frac{\partial}{\partial \theta} \mathbb{P}_\theta(W''_n \cap (W'_n)^c)_{|\theta=\theta_0} \geq \mathbb{P}_\theta(W''_n \cap (W'_n)^c).$$

On en déduit donc que :

$$\mathbb{P}_\theta(W'_n \cap (W''_n)^c) \geq \mathbb{P}_\theta(W''_n \cap (W'_n)^c),$$

et donc en ajoutant $\mathbb{P}_\theta(W'_n \cap W''_n)$:

$$\mathbb{P}_\theta(W'_n) \geq \mathbb{P}_\theta(W''_n).$$

Le test pur de région critique W'_n est donc plus puissant que n'importe quel autre test pur sans biais. ▲

Exercice IX.6. 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_n la variable aléatoire donnant le nombre de femmes des $n = 700$ jurés. Sous H_0 , X_n suit une loi binomiale de paramètres $\mathcal{B}(n, p_0)$. On considère la statistique de test :

$$\zeta_n = \frac{X_n - np_0}{\sqrt{np_0(1-p_0)}}.$$

Sous H_0 , la statistique ζ_n converge en loi vers G de la loi gaussienne $\mathcal{N}(0, 1)$. Sous H_1 , on a que p.s. $\lim_{n \rightarrow \infty} |\zeta_n| = +\infty$. On considère donc les régions critiques de la forme $\{|\zeta_n| \geq a\}$. La région critique $W = \{|\zeta_n| \geq \phi_{1-\alpha/2}\}$, où $\phi_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi normale centrée réduite, définit un test convergent de niveau asymptotique α . La p -valeur asymptotique associée est $p\text{-val} = \mathbb{P}(|G| \geq |\zeta_n^{\text{obs}}|)$.

Application numérique. On a $\zeta_n \simeq -8.16$ et $p\text{-val} \simeq 3 \cdot 10^{-16}$. On rejette donc l'hypothèse d'impartialité du juge.

2. Pour $n = 40$, on a $\zeta_n \simeq -1.95$ et $p\text{-val} \simeq 5\%$. Comme $\phi_{1-\alpha/2} \simeq 1.96$ pour $\alpha = 5\%$, on ne peut donc pas rejeter l'hypothèse d'impartialité du juge au niveau 5% . ▲

Exercice IX.7.

- D'après les données de l'énoncé, $p_0 = 1/310\,000 \simeq 3.23 \cdot 10^{-6}$.
- En supposant les X_i indépendants, la variable aléatoire N suit une loi binomiale de paramètres (n, p) avec $n = 300\,533$ et p de l'ordre de $p_0 = 3.23 \cdot 10^{-6}$. On peut donc approcher cette loi par une loi de Poisson de paramètre $\theta = np$. Dans le cas des études antérieures, on a $\theta_0 = np_0 \simeq 0.969$.
- On construit un test de la forme $\varphi(N) = \mathbf{1}_{\{N \geq N_{\text{obs}}\}}$. Sous l'hypothèse H_0 , N suit une loi de Poisson de paramètre θ_0 . La p -valeur est donnée par :

$$p\text{-val} = \mathbb{E}_{\theta_0}[\varphi(N)] = \mathbb{P}_{\theta_0}(N \geq N^{\text{obs}}) = 1 - \sum_{k < N^{\text{obs}}} \frac{\theta_0^k}{k!} e^{-\theta_0}.$$

On obtient $p\text{-val} \simeq 7.47\%$ pour $N^{\text{obs}} = 3$ et $p\text{-val} \simeq 1.71\%$ pour $N^{\text{obs}} = 4$. Comme on a observé 4 cas de dommages, la p -valeur de ce test est environ 1.71%. En particulier, on rejette l'hypothèse H_0 au seuil de 5%. ▲

Exercice IX.8. 1. On utilise l'indépendance des variables X_i :

$$\text{Var}_p(\hat{p}_n) = \text{Var}_p\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}_p(X_i) = \frac{p(1-p)}{n}.$$

Comme $p \in [0, 1]$, on a $p(1-p) \leq 1/4$. On en déduit que $\text{Var}_p(\hat{p}_n) \leq 1/(4n)$.

2. On modélise le vote d'un électeur par une loi de Bernoulli : la variable X_i désigne le vote de l'individu i ; elle vaut 1 s'il vote pour l'ancien maire et 0 sinon. Le théorème central limite assure la convergence en loi suivante :

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}(0, p(1-p)).$$

Soit $\alpha \in]0, 1[$. Un intervalle de confiance pour p de niveau asymptotique $1 - \alpha$ est donné par :

$$\left[\hat{p}_n \pm \phi_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} \right],$$

où ϕ_r est le quantile d'ordre r de la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$. Comme $p(1-p) \leq 1/4$, on en déduit un intervalle de confiance pour p de niveau asymptotique par excès $1 - \alpha$:

$$\left[\hat{p}_n \pm \frac{\phi_{1-\alpha/2}}{2\sqrt{n}} \right].$$

On souhaite $\phi_{1-\alpha/2}/(2\sqrt{n}) \leq 0.03$ soit $n \geq (\phi_{1-\alpha/2}/0.06)^2$. Pour la valeur usuelle $\alpha = 5\%$, on obtient $n \geq 1068$.

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 .

Comme les variables \hat{p}_n et \hat{q}_n sont indépendantes, on déduit du théorème central limite que :

$$\sqrt{n}(\hat{p}_n - p, \hat{q}_n - q) \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}(0, \Sigma),$$

où la matrice de covariance est $\Sigma = \begin{pmatrix} p(1-p) & 0 \\ 0 & q(1-q) \end{pmatrix}$. Sous H_0 , on en déduit donc que :

$$\sqrt{n}(\hat{p}_n - \hat{q}_n) \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}(0, p(1-p) + q(1-q)).$$

Un estimateur convergent de la variance $p(1-p) + q(1-q)$ est $\hat{p}_n(1-\hat{p}_n) + \hat{q}_n(1-\hat{q}_n)$. Le théorème de Slutsky assure que :

$$\frac{\sqrt{n}(\hat{p}_n - \hat{q}_n)}{\sqrt{\hat{p}_n(1-\hat{p}_n) + \hat{q}_n(1-\hat{q}_n)}} \xrightarrow[n \rightarrow \infty]{\text{en loi}} \mathcal{N}(0, 1).$$

À la vue de ce résultat, il est naturel de considérer la statistique de test :

$$T_n = \frac{\sqrt{n}(\hat{p}_n - \hat{q}_n)}{\sqrt{\hat{p}_n(1-\hat{p}_n) + \hat{q}_n(1-\hat{q}_n)}}.$$

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, sous H_0 :

$$\mathbb{P}_{p,p}(W_n) = \mathbb{P}_{p,p}(|T_n| > c) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(|G| > c),$$

où G est une variable aléatoire de loi $\mathcal{N}(0, 1)$. Pour obtenir un test de niveau asymptotique α , on choisit donc $c = \phi_{1-\alpha/2}$.

On rejette donc H_0 avec un risque asymptotique (non uniforme) de première espèce α si $|T_n| > \phi_{1-\alpha/2}$.

Application numérique. Avec $\alpha = 0.05$: $\phi_{1-\alpha/2} \simeq 1.96$ et $t_n \simeq -0.93$. On ne peut pas rejeter H_0 au niveau 5%, infirmant ainsi la position du journaliste. La p -valeur asymptotique (non-uniforme) de ce test est $\mathbb{P}(|G| > t_n) \simeq 0.18$.

▲

Exercice IX.9. 1. Les variables aléatoires sont indépendantes et de même loi $\mathcal{N}(\mu, \sigma^2)$, avec μ inconnu.

2. L'estimateur de la moyenne empirique $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i$ est un estimateur sans biais de μ . Dans ce modèle, c'est également l'estimateur du maximum de vraisemblance. La loi de $\hat{\mu}_n$ est la loi gaussienne $\mathcal{N}(\mu, \sigma^2/n)$.
3. On considère l'hypothèse nulle $H_0 = \{\mu \leq \mu_0\}$ (le nouveau modèle n'est pas plus performant que l'ancien) et l'hypothèse alternative $H_1 = \{\mu > \mu_0\}$. L'erreur de première espèce est d'adopter le nouveau modèle alors qu'il n'est pas plus performant que l'ancien. Le modèle est exponentiel et la statistique de test $\hat{\mu}_n$ est continue. Le test unilatéral UPP est donc un test pur défini par la région critique :

$$W_n = \{\hat{\mu}_n \geq z\}.$$

L'erreur de première espèce étant α , la valeur de z est déterminée par :

$$\alpha = \sup_{\mu \in H_0} \mathbb{P}_\mu(W_n) = \sup_{\mu \leq \mu_0} \mathbb{P}_\mu(W_n).$$

Comme $\hat{\mu}_n$ est de loi gaussienne $\mathcal{N}(\mu, \sigma^2/n)$, il vient :

$$\mathbb{P}_\mu(W_n) = \mathbb{P}_\mu(\hat{\mu}_n \geq z) = \mathbb{P}\left(G \geq \sqrt{n} \frac{z - \mu}{\sigma}\right),$$

où G est de loi gaussienne $\mathcal{N}(0, 1)$. On remarque que $\mu \mapsto \mathbb{P}_\mu(W_n)$ est une fonction croissante de μ et donc $\sup_{\mu \leq \mu_0} \mathbb{P}_\mu(W) = \mathbb{P}_{\mu_0}(W)$. Pour un seuil α , z est donc déterminé par l'équation :

$$\alpha = \mathbb{P}\left(G \geq \sqrt{n} \frac{z - \mu_0}{\sigma}\right),$$

soit $z = \mu_0 + \sigma \phi_{1-\alpha}/\sqrt{n}$, où ϕ_r est le quantile d'ordre r de la loi gaussienne $\mathcal{N}(0, 1)$.

Pour l'application numérique, on a $\alpha = 5\%$, $\phi_{1-\alpha} \simeq 1.64$ et $z \simeq 123.9 > \hat{\mu}_n = 123.5$. On accepte donc l'hypothèse H_0 : le nouveau modèle n'est pas plus performant que l'ancien. La p -valeur de ce test est $p\text{-val} \simeq 0.07$. Ceci confirme que l'on accepte H_0 au niveau de 5%, mais la p -valeur indique également que le choix de 5% est presque critique.

4. On évalue l'erreur de deuxième espèce, β , pour $\mu_1 = 1.05 \mu_0$, c'est-à-dire la probabilité de rejeter le nouveau modèle sachant que l'annonce du représentant est exacte et donc que le nouveau modèle est plus performant de 5%. Elle est définie par :

$$\beta = 1 - \mathbb{P}_{\mu_1}(W_n) = \mathbb{P}_{\mu_1}(\hat{\mu}_n < z) = \mathbb{P}(G < \phi_{1-\alpha} + \sqrt{n}(\mu_0 - \mu_1)/\sigma).$$

L'application numérique donne $\beta \simeq 19.5\%$. 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 nulle est alors $H'_0 = \{\mu = \mu_1\}$, et l'hypothèse alternative $H_1 = \{\mu \leq \mu_0\}$. Dans ce cas la région critique est $W'_n = \{\hat{\mu}_n \leq z'\}$ et l'erreur de première espèce est :

$$\mathbb{P}_{\mu_1}(W'_n) = \mathbb{P}(G \leq \sqrt{n}(z' - \mu_1)/\sigma).$$

Pour un niveau α , on obtient :

$$z' = \mu_1 - \frac{1}{\sqrt{n}} \sigma \phi_{1-\alpha} = 1.05\mu_0 - \frac{1}{\sqrt{n}} \sigma \phi_{1-\alpha}.$$

Pour $\alpha = 5\%$, on obtient $z' \simeq 122.0 < \hat{\mu}_n = 123.5$. On accepte donc H'_0 : le nouveau modèle est donc plus performant que l'ancien de 5% en moyenne. La p -valeur de ce test est $p\text{-val} = 0.19$; elle confirme qu'il n'est pas raisonnable de rejeter H_0 .

Le risque de deuxième espèce maximal est alors ;

$$\beta' = \sup_{\mu \leq \mu_0} (1 - \mathbb{P}_{\mu}(W'_n)) = \mathbb{P}_{\mu_0}(\hat{\mu}_n > z') = \mathbb{P}(G > -\phi_{1-\alpha} - \sqrt{n}(\mu_0 - \mu_1)/\sigma) = \beta.$$

L'acheteur a donc environ 20% de chance d'accepter l'annonce du représentant alors que celle-ci est fautive.

6. Dans la question 3, le risque de l'acheteur est $\mathbb{P}_{\mu_0}(\hat{\mu}_n \geq z'')$ et dans la question 4, celui du vendeur est $\mathbb{P}_{\mu_1}(\hat{\mu}_n \leq z'')$. On peut donc chercher z'' tel que ces deux risques soient égaux ; l'acheteur et le vendeur ayant alors la même région critique. On a :

$$\mathbb{P}_{\mu_0}(\hat{\mu}_n \geq z'') = \mathbb{P}(G \geq \sqrt{n}(z'' - \mu_0)/\sigma),$$

et

$$\mathbb{P}_{\mu_1}(\hat{\mu}_n \leq z'') = \mathbb{P}(G \leq \sqrt{n}(z'' - \mu_1)/\sigma) = \mathbb{P}(G \geq \sqrt{n}(\mu_1 - z'')/\sigma).$$

Ceci implique $z'' - \mu_0 = \mu_1 - z''$ soit $z'' = (\mu_0 + \mu_1)/2$. L'application numérique donne $z'' = 123$ et le risque de l'acheteur et du vendeur est $\mathbb{P}(G \geq \sqrt{n}(\mu_1 - \mu_0)/(2\sigma)) \simeq 11\%$.

▲

Exercice IX.10. 1. On note Y_1, \dots, Y_m la durée de vie observée des pièces produites par la machine de l'autre société. Les variables aléatoires Y_1, \dots, Y_m sont indépendantes de loi gaussienne $\mathcal{N}(\nu, \sigma^2)$ et indépendantes de X_1, \dots, X_n . Un calcul élémentaire assure que l'estimateur du maximum de vraisemblance de (μ, ν) est $(\hat{\mu}_n, \hat{\nu}_m)$, où :

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{et} \quad \hat{\nu}_m = \frac{1}{m} \sum_{i=1}^m Y_i.$$

2. La variable aléatoire $\hat{\mu}_n - \hat{\nu}_m$ est de loi gaussienne $\mathcal{N}(\mu_\nu, \sigma^2(1/n + 1/m))$. On considère l'hypothèse nulle $H_0 = \{\mu = \nu\}$, l'hypothèse alternative $H_1 = \{\mu \neq \nu\}$ et la statistique de test :

$$\zeta_{n,m} = \sqrt{\frac{n+m}{nm}} \frac{\hat{\mu}_n - \hat{\nu}_m}{\sigma}.$$

Sous H_0 , la statistique de test $\zeta_{n,m}$ est de loi gaussienne $\mathcal{N}(0, 1)$. Sous H_1 , la statistique de test $\zeta_{n,m}$ converge p.s. vers $+\infty$ ou $-\infty$ quand $\min(n, m)$ tend vers l'infini. On en déduit que le test pur de région critique :

$$W_{n,m} = \{|\zeta_{n,m}| \geq z\},$$

est convergent. Ce test est de niveau uniforme :

$$\sup_{H_0} \mathbb{P}_{(\mu, \nu)}(W_{m,n}) = \mathbb{P}(|G| \geq z),$$

où G est de loi gaussienne $\mathcal{N}(0, 1)$. On en déduit que pour $z = \phi_{1-\alpha/2}$, le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, le test est de niveau uniforme α . La p -valeur (uniforme) de ce test est définie par :

$$p\text{-val} = \sup_{H_0} \mathbb{P}_{(\mu, \nu)}(|\zeta_{n,m}| \geq |\zeta_{n,m}^{\text{obs}}|) = \mathbb{P}(|G| \geq |\zeta_{n,m}^{\text{obs}}|),$$

où $\zeta_{n,m}^{\text{obs}}$ est la valeur de la statistique de test évaluée en les données observées. L'application numérique donne $\zeta_{n,m} = 0.36$ et $p\text{-val} \simeq 0.72$. On accepte donc H_0 : les deux machines sont équivalentes. ▲

Exercice IX.11. 1. La vraisemblance est :

$$p_1(x_1; \ell) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x_1 - \ell)^2 / (2\sigma_0^2)} = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x_1^2 + \ell^2) / (2\sigma_0^2)} e^{-x_1 \ell / \sigma_0^2}.$$

Il s'agit donc bien d'un modèle exponentiel de statistique canonique $S_1(x_1) = x_1$. La statistique canonique de l'échantillon de taille n est donc $S_n = \sum_{i=1}^n L_i$. L'estimateur du maximum de vraisemblance de ℓ est $\bar{L}_n = S_n/n$.

2. La statistique canonique est continue. Le test UPPS au seuil α pour tester H_0 contre H_1 est donc un test pur de région critique :

$$W_n = \{x = (x_1, \dots, x_n); \sum_{i=1}^n x_i/n \notin]c_1, c_2[\},$$

où les constantes $c_1 < c_2$ sont déterminées par les relations $\mathbb{P}_{\ell_0 \pm \delta \ell_0}(W_n) = \alpha$. On remarque que sous \mathbb{P}_ℓ , la variable \bar{L}_n suit une loi gaussienne $\mathcal{N}(\ell, \sigma_0^2/n)$.

On pose $\varepsilon_j = \sqrt{n}(\ell_0 - c_j)/\sigma_0$ pour $j \in \{1, 2\}$ et $\mu_0 = \sqrt{n} \delta\ell_0/\sigma_0$. On a $\varepsilon_1 < \varepsilon_2$. On remarque que la condition $\mathbb{P}_{\ell_0 + \delta\ell_0}(W_n) = \alpha$ est équivalente à $\mathbb{P}_{\ell_0 + \delta\ell_0}(\bar{L}_n \in]c_1, c_2]) = 1 - \alpha$ soit :

$$\int_{\mu_0 + \varepsilon_1}^{\mu_0 + \varepsilon_2} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy = 1 - \alpha,$$

et la condition $\mathbb{P}_{\ell_0 - \delta\ell_0}(W_n) = \alpha$ est équivalente à :

$$\int_{-\mu_0 + \varepsilon_1}^{-\mu_0 + \varepsilon_2} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy = 1 - \alpha.$$

Comme $\delta\ell_0 \neq 0$, on déduit du changement de variable $u = -y$ dans la première intégrale que $\varepsilon_2 = -\varepsilon_1$. On note $\varepsilon = \varepsilon_2 > 0$ et ε est caractérisé par la condition :

$$\int_{-\mu_0 - \varepsilon}^{-\mu_0 + \varepsilon} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy = 1 - \alpha.$$

3. L'application numérique donne $\mu_0 = 10.61$ et donc :

$$1 - \alpha = \int_{-\mu_0 - \varepsilon}^{-\mu_0 + \varepsilon} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy \simeq \int_{-\infty}^{\mu_0 + \varepsilon} \frac{e^{-y^2/2}}{\sqrt{2\pi}} dy,$$

soit $\varepsilon \simeq \phi_{1-\alpha} + \mu_0$, où ϕ_r est le quantile d'ordre r de la loi gaussienne $\mathcal{N}(0, 1)$. Pour $\alpha = 5\%$, on obtient $\phi_{1-\alpha} \simeq 1.64$ et $\varepsilon \simeq 12.25$, puis $c_1 \simeq 783.3$ et $c_2 \simeq 786.7$. Comme $\bar{L}_n^{\text{obs}} = 788.3 \notin [c_1, c_2]$, on rejette l'hypothèse H_0 au niveau de 5%. La machine nécessite donc d'être révisée. La p -valeur associée à ce test est de l'ordre de 10^{-37} . Ceci indique qu'il est justifié de rejeter H_0 .

4. Comme la variable \bar{L}_n suit une loi gaussienne $\mathcal{N}(\ell, \sigma_0^2/n)$, on en déduit que l'intervalle de confiance de ℓ , $I_n = [\bar{L}_n \pm \phi_{1-\alpha/2}\sigma_0/\sqrt{n}]$, est de niveau exact $1 - \alpha$. Pour $\alpha = 5\%$, on a $\phi_{1-\alpha/2} \simeq 1.96$ et donc $I_n \simeq [788.0, 788.6]$. Comme $[\ell_0 \pm \delta\ell_0] \cap I_n = \emptyset$, on retrouve le fait que la machine doit être révisée. ▲

Exercice IX.12. On suppose que le générateur donne des réalisations de variables aléatoires indépendantes et de même loi. On considère l'hypothèse nulle, H_0 : le générateur donne des réalisations de variables aléatoires de loi uniforme sur $\{0, \dots, 9\}$, et l'hypothèse alternative, H_1 : le générateur donne des réalisations de variables aléatoires de loi non uniforme sur $\{0, \dots, 9\}$. Les statistiques de test :

$$\zeta_n^{(1)} = n \sum_{j=0}^{j=9} \frac{(\hat{p}_j - p_j)^2}{\hat{p}_j} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{j=0}^{j=9} \frac{(\hat{p}_j - p_j)^2}{p_j},$$

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 :

$$W_i = \{\zeta_n^{(i)} \geq z_{1-\alpha}\}, \quad i \in \{1, 2\},$$

où z_r est le quantile d'ordre r de la loi $\chi^2(9)$, sont convergents de niveau asymptotique α . L'application numérique donne $\zeta_n^{(1),\text{obs}} \simeq 11.1$ et $\zeta_n^{(2),\text{obs}} \simeq 10.4$. Au seuil $\alpha = 5\%$, on lit dans la table des quantiles de la loi du χ^2 : $z_{1-\alpha} \simeq 6.92$. Comme $z_{1-\alpha} > \zeta_n^{(i),\text{obs}}$ pour $i \in \{1, 2\}$, on accepte l'hypothèse nulle H_0 . Les p -valeurs asymptotiques (non-uniformes) sont :

$$p\text{-val}^{(i)} = \mathbb{P}(Z \geq \zeta_n^{(i),\text{obs}}),$$

où Z est de loi $\chi^2(9)$. On obtient $p\text{-val}^{(1)} \simeq 0.27$ et $p\text{-val}^{(2)} \simeq 0.32$. Ces valeurs confirment qu'il n'est pas raisonnable de rejeter H_0 . ▲

Exercice IX.13. 1. L'estimateur du maximum de vraisemblance, \hat{p} , de p est le vecteur des fréquences empiriques. On a donc $\hat{p} \simeq (0.606; 0.186; 0.173; 0.035)$.

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 $\hat{p}_J \simeq 0.779$ et $\hat{p}_N \simeq 0.792$. La statistique du χ^2 est :

$$\begin{aligned} \zeta_n^{(1)} = n & \frac{(\hat{p}_{J,N} - \hat{p}_J \hat{p}_N)^2}{\hat{p}_{J,N}} + n \frac{(\hat{p}_{W,N} - (1 - \hat{p}_J) \hat{p}_N)^2}{\hat{p}_{W,N}} \\ & + n \frac{(\hat{p}_{J,C} - \hat{p}_J (1 - \hat{p}_N))^2}{\hat{p}_{J,C}} + n \frac{(\hat{p}_{W,C} - (1 - \hat{p}_J) (1 - \hat{p}_N))^2}{\hat{p}_{W,C}}. \end{aligned}$$

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 asymptotique de 0.1% . (On aurait également pu utiliser la statistique $\zeta_n^{(2)}$, avec dans notre cas particulier $\zeta_n^{(2)} \simeq 15\,594$.) La p -valeur asymptotique est nulle. On rejette donc H_0 .

3. On teste $H_0 = \{p = p^0\}$, avec $p^0 = (0.605; 0.189; 0.17; 0.036)$ contre son alternative $H_1 = \{p \neq p^0\}$. Le nombre de degrés de liberté du test du χ^2 est donc $q = 4 - 1 = 3$. La statistique du χ^2 est :

$$\zeta_n^{(1)} = n \frac{(\hat{p}_{J,N} - p_{J,N}^0)^2}{\hat{p}_{J,N}} + n \frac{(\hat{p}_{W,N} - p_{W,N}^0)^2}{\hat{p}_{W,N}} + n \frac{(\hat{p}_{J,C} - p_{J,C}^0)^2}{\hat{p}_{J,C}} + n \frac{(\hat{p}_{W,C} - p_{W,C}^0)^2}{\hat{p}_{W,C}}.$$

On obtient $\zeta_n^{(1)} \simeq 502$. 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 0.1%. 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 $\zeta_n^{(2)} \simeq 500$.) La p -valeur asymptotique est de l'ordre de 10^{-100} .

▲

Exercice IX.14. 1. Comme les variables aléatoires $(w_k, 1 \leq k \leq 2K + 1)$ sont indépendantes, on a que les variables aléatoires $((w_{2k-1}, w_{2k}), 1 \leq k \leq K)$ sont indépendantes. En revanche, les variables ne sont pas indépendantes $((w_k, w_{k+1}), 1 \leq k \leq 2K)$ ne sont pas indépendantes, et on ne peut pas appliquer le test d'adéquation de loi du χ^2 .

Il s'agit d'effectuer un test d'indépendance du χ^2 à $(2 - 1)(2 - 1) = 1$ degré de liberté. Les statistiques de test du χ^2 sont :

$$\zeta_n^{(1)} = n \sum_{i,j \in \{C,L\}} \frac{(\hat{p}_{i,j} - \hat{p}_{i\cdot} \hat{p}_{\cdot j})^2}{\hat{p}_{i,j}} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i,j \in \{C,L\}} \frac{(\hat{p}_{i,j} - \hat{p}_{i\cdot} \hat{p}_{\cdot j})^2}{\hat{p}_{i\cdot} \hat{p}_{\cdot j}},$$

où les fréquences empiriques $\hat{p}_{i,j} = N_{i,j}/n$, $\hat{p}_{i\cdot} = (N_{i,C} + N_{i,L})/n$ et $\hat{p}_{\cdot j} = (N_{C,j} + N_{L,j})/n$ sont les estimateurs du maximum de vraisemblance des fréquences $\mathbb{P}(X_k = i, Y_k = j)$, $\mathbb{P}(X_k = i)$ et $\mathbb{P}(Y_k = j)$. On obtient $\zeta_n^{(1)} \simeq 130$ et $\zeta_n^{(2)} \simeq 68$. On lit dans les tables que $\mathbb{P}(Z > 11) \leq 0.1\%$, où Z suit la loi $\chi^2(1)$. On rejette donc l'hypothèse d'indépendance au niveau de 99.9%. (Les p -valeurs sont inférieures à 10^{-15} ; ceci confirme que l'on rejette l'hypothèse d'indépendance.)

2. Sous l'hypothèse nulle, les variables aléatoires X_k et Y_k sont indépendantes et de même loi. La loi de (X_k, Y_k) ne dépend que du paramètre $\rho = \mathbb{P}(X_k = C)$. La vraisemblance du modèle sous l'hypothèse nulle s'écrit $\rho^N (1 - \rho)^{2K - N}$ avec $N = 2N_{C,C} + N_{C,L} + N_{L,C}$. L'estimateur du maximum de vraisemblance de ρ est $\hat{\rho} = N/2K$.

Le nombre de valeurs possibles pour la variable aléatoire (X_k, Y_k) est $m = 4$, et on estime un paramètre de dimension 1. On effectue un test d'adéquation de loi à une famille à un paramètre. Il s'agit d'un test du χ^2 à $m - 1 - 1 = 2$ degrés de liberté. Les statistiques de test du χ^2 sont :

$$\zeta_n^{(1)} = n \sum_{i,j \in \{C,L\}} \frac{(\hat{p}_{i,j} - p_{i,j}(\hat{\rho}))^2}{\hat{p}_{i,j}} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i,j \in \{C,L\}} \frac{(\hat{p}_{i,j} - p_{i,j}(\hat{\rho}))^2}{p_{i,j}(\hat{\rho})},$$

où $p_{i,j}(r) = (r\mathbf{1}_{\{i=C\}} + (1-r)\mathbf{1}_{\{i=L\}})(r\mathbf{1}_{\{j=C\}} + (1-r)\mathbf{1}_{\{j=L\}})$. On obtient $\zeta_n^{(1)} \simeq 135$ et $\zeta_n^{(2)} \simeq 72$. On lit dans les tables que $\mathbb{P}(Z > 14) \leq 0.1\%$, où Z suit la loi $\chi^2(2)$. On rejette donc l'hypothèse nulle au niveau de 99.9%. (Les p -valeurs sont inférieures à 10^{-15} ; ceci confirme que l'on rejette l'hypothèse d'indépendance.).

L'hypothèse nulle est plus contrainte que dans la question précédente. Il n'est pas étonnant que la statistique prenne une valeur plus élevée. ▲

Exercice IX.15. 1. Sous H_0 , la loi des réponses des deux juges est la même. Notons β la probabilité de répondre par l'affirmative : $\beta = p_+^{(1)} = p_+^{(2)}$. On a alors $p_-^{(1)} = p_-^{(2)} = 1 - \beta$. On note $p_{\varepsilon_1\varepsilon_2}$, avec $\varepsilon_i \in \{+, -\}$ la probabilité que le juge 1 réponde ε_1 et le juge 2 réponde ε_2 . On pose $\alpha = p_{+-}$. On remarque que $\beta = p_+^{(1)} = p_{++} + p_{+-}$, soit $p_{++} = \beta - \alpha$, et de même $p_{--} = 1 - \beta - \alpha$, puis finalement $p_{-+} = \beta - p_{++} = \alpha$.

2. On considère l'échantillon $x = (x_1, \dots, x_n)$, où $x_i \in \{-, +\}^2$. On définit $N_{\varepsilon_1\varepsilon_2} = \sum_{i=1}^n \mathbf{1}_{\{x_i=(\varepsilon_1,\varepsilon_2)\}}$, pour $\varepsilon_i \in \{+, -\}$, le nombre d'occurrences de $(\varepsilon_1, \varepsilon_2)$. La vraisemblance de l'échantillon est donnée par :

$$p_n(x; \alpha, \beta) = (\beta - \alpha)^{N_{++}} \alpha^{N_{-++} + N_{+-}} (1 - \beta - \alpha)^{N_{--}} \\ = (\beta - \alpha)^{N_{++}} \alpha^{n - N_{++} + N_{--}} (1 - \beta - \alpha)^{N_{--}},$$

car $N_{++} + N_{--} + N_{+-} + N_{-+} = n$. Le couple (N_{++}, N_{--}) est donc la statistique canonique du modèle. La log-vraisemblance du modèle est donc :

$$L_n(x, \alpha, \beta) = N_{++} \log(\beta - \alpha) + (n - N_{--} - N_{++}) \log \alpha + N_{++} \log(1 - \beta - \alpha).$$

L'annulation des deux dérivées partielles donne :

$$\frac{-N_{++}}{\beta - \alpha} + \frac{n - N_{--} - N_{++}}{\alpha} - \frac{N_{--}}{1 - \beta - \alpha} = 0 \quad \text{et} \quad \frac{N_{++}}{\beta - \alpha} - \frac{N_{--}}{1 - \beta - \alpha} = 0.$$

Ce système se résout alors facilement et on trouve :

$$\alpha = \frac{n - N_{--} - N_{++}}{2n} \quad \text{et} \quad \beta = \frac{n + N_{++} - N_{--}}{2n}.$$

De plus, la log-vraisemblance tend vers $-\infty$ sur la frontière du domaine $\{(\alpha, \beta); 0 < \alpha < \beta < 1 - \alpha\}$. On en déduit donc que l'estimateur du maximum de vraisemblance de (α, β) est :

$$(\hat{\alpha}_n, \hat{\beta}_n) = \frac{1}{2n}(n - N_{--} - N_{++}, n + N_{++} - N_{--}).$$

3. L'hypothèse H_0 est équivalente à $\{p_{-+} = p_{+-}\}$. L'estimateur du maximum de vraisemblance des fréquences $p_{\varepsilon_1\varepsilon_2}$ sont les fréquences empiriques $\hat{p}_{\varepsilon_1\varepsilon_2} = N_{\varepsilon_1\varepsilon_2}/n$. Sous H_0 , l'estimateur du maximum de vraisemblance de $p_{\varepsilon_1\varepsilon_2}$ est $p_{\varepsilon_1\varepsilon_2}(\hat{\alpha}_n, \hat{\beta}_n)$ soit :

$$\begin{aligned} p_{++}(\hat{\alpha}_n, \hat{\beta}_n) &= \hat{\beta} - \hat{\alpha} = \frac{N_{++}}{n} = \hat{p}_{++}, \\ p_{--}(\hat{\alpha}_n, \hat{\beta}_n) &= 1 - \hat{\beta} - \hat{\alpha} = \frac{N_{--}}{n} = \hat{p}_{--}, \\ p_{-+}(\hat{\alpha}_n, \hat{\beta}_n) &= \hat{\alpha}, \\ p_{+-}(\hat{\alpha}_n, \hat{\beta}_n) &= \hat{\alpha}. \end{aligned}$$

On peut réécrire $\hat{\alpha}_n$ de la manière suivante :

$$\hat{\alpha}_n = \frac{N_{+-} + N_{-+}}{2n}.$$

Une statistique du test du χ^2 est donc :

$$\begin{aligned} \zeta_n &= n \sum_{\varepsilon_1, \varepsilon_2 \in \{+, -\}} \frac{\left(\hat{p}_{\varepsilon_1\varepsilon_2} - p_{\varepsilon_1\varepsilon_2}(\hat{\alpha}_n, \hat{\beta}_n)\right)^2}{p_{\varepsilon_1\varepsilon_2}(\hat{\alpha}_n, \hat{\beta}_n)} \\ &= n \left(\frac{\left(\frac{N_{+-}}{n} - \frac{N_{+-} + N_{-+}}{2n}\right)^2}{\frac{N_{+-} + N_{-+}}{2n}} + \frac{\left(\frac{N_{-+}}{n} - \frac{N_{+-} + N_{-+}}{2n}\right)^2}{\frac{N_{-+} + N_{+-}}{2n}} \right) \\ &= \frac{(N_{+-} - N_{-+})^2}{N_{-+} - N_{+-}}. \end{aligned}$$

Sous H_0 , la statistique de test $(\zeta_n, n \geq 1)$ tend en loi vers un χ^2 à 4-1-2=1 degré de liberté (on retire 2 degrés de liberté pour l'estimation des paramètres α et β). Le test défini par la région critique :

$$W = \{\zeta_n \geq z_{1-\alpha}\},$$

où z_r est le quantile d'ordre r de la loi $\chi^2(1)$, est convergent de niveau asymptotique α .

4. L'application numérique donne $\zeta_n^{\text{obs}} = 5$. Au seuil $\alpha = 5\%$, on lit dans la table des quantiles de la loi du χ^2 : $z_{1-\alpha} \simeq 3.84$. Comme $z_{1-\alpha} \leq \zeta_n^{\text{obs}}$, on rejette l'hypothèse nulle H_0 . La p -valeur asymptotique (non-uniforme) est :

$$p\text{-val} = \mathbb{P}(Z \geq \zeta_n^{\text{obs}}),$$

où Z est de loi $\chi^2(1)$. On obtient $p\text{-val} \simeq 0.025$. Cette valeur confirme qu'il est raisonnable de rejeter H_0 .

▲

Exercice IX.16. Le nombre d'observations d'abcès étant insuffisant (≤ 5), on les regroupe avec les observations d'ulcération. On considère donc le tableau XIII.2.

| Vaccin | Réaction légère | Réaction moyenne | Ulcération ou Abcès | Total |
|--------|-----------------|------------------|---------------------|-------|
| A | 12 | 156 | 9 | 177 |
| B | 29 | 135 | 7 | 171 |
| Total | 41 | 291 | 16 | 348 |

Table XIII.2. Réactions à deux vaccins de B.C.G..

On désire savoir si la réaction due au vaccin dépend du type de vaccin (hypothèse alternative H_1) ou non (hypothèse nulle H_0). Il s'agit d'un test du χ^2 d'indépendance : $H_0 = \{p_{i,j} = p_{i \cdot} p_{\cdot j}; i \in \{l, m, u\}, j \in \{A, B\}\}$, où $p_{i \cdot} = \sum_{j \in \{A, B\}} p_{i,j}$ et $p_{\cdot j} = \sum_{i \in \{l, m, u\}} p_{i,j}$, avec l'hypothèse alternative $H_1 = \{\text{Il existe } i, j \text{ tel que } p_{i,j} \neq p_{i \cdot} p_{\cdot j}\}$. Les statistiques de tests :

$$\zeta_n^{(1)} = n \sum_{\substack{i \in \{l, m, u\}, \\ j \in \{A, B\}}} \frac{(\hat{p}_{i,j} - \hat{p}_{i \cdot} \hat{p}_{\cdot j})^2}{\hat{p}_{i,j}} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{\substack{i \in \{l, m, u\}, \\ j \in \{A, B\}}} \frac{(\hat{p}_{i,j} - \hat{p}_{i \cdot} \hat{p}_{\cdot j})^2}{\hat{p}_{i \cdot} \hat{p}_{\cdot j}}$$

convergent (quand $n \rightarrow \infty$) sous H_0 vers un χ^2 à est $d = (3-1)(2-1) = 2$ degrés de liberté, et divergent sous H_1 . Le test de région critique $W^{(k)} = \{\zeta_n^{(k)} > z_{1-\alpha}(d)\}$ est un test convergent de niveau asymptotique α , où $z_{1-\alpha}(d)$ est le quantile d'ordre $1 - \alpha$ de la loi $\chi^2(d)$. Les p -valeurs asymptotiques (non-uniformes) sont $p^{(k)} = \mathbb{P}(Z \geq \zeta_n^{(k), \text{obs}})$, où Z est de loi $\chi^2(d)$.

L'application numérique donne : $n = 348$, $\zeta_n^{(1), \text{obs}} \simeq 10.3$ et $\zeta_n^{(2), \text{obs}} \simeq 8.7$, $z_{1-\alpha}(d) \simeq 5.99$ pour le seuil usuel $\alpha = 5\%$. Les p -valeurs asymptotiques de ces tests sont $p\text{-val}^{(1)} \simeq 0.006$ et $p^{(2), \text{obs}} \simeq 0.013$. On rejette donc H_0 . ▲

Exercice IX.17. L'estimateur du maximum de vraisemblance du paramètre $\theta > 0$ de la loi de Poisson est $\hat{\theta} = \sum_{i=1}^n X_i/n$. On considère les statistiques de tests :

$$\zeta_n^{(1)} = n \sum_{j=0}^4 \frac{(\hat{p}_j - p_j(\hat{\theta}_n))^2}{\hat{p}_j} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{j=0}^4 \frac{(\hat{p}_j - p_j(\hat{\theta}_n))^2}{p_j(\hat{\theta}_n)}$$

où (\hat{p}_j) sont les fréquences empiriques et $(p_j(\hat{\theta}_n) = e^{-\hat{\theta}_n} \hat{\theta}_n^j / j!)$ sont les fréquences théoriques de la loi de Poisson de paramètre $\hat{\theta}_n$ (on prendra $p_4(\hat{\theta}_n) = 1 - \sum_{j=0}^3 p_j(\hat{\theta}_n)$). Étant donné qu'une case ne contient que 2 observations, il faudrait la regrouper avec une autre case ; mais cela ne change pas ici significativement le résultat. Les statistiques ci-dessus convergent en loi vers Z de loi du χ^2 à

$5 - 1 - 1 = 3$ degrés de liberté (on a retranché un degré de liberté supplémentaire du fait de l'estimation de θ). Les tests définis par les régions critiques :

$$W^{(i)} = \{\zeta_n^{(i)} \geq z_{1-\alpha}(3)\},$$

pour $i \in \{1, 2\}$ et $z_{1-\alpha}(r)$ le quantile d'ordre $1 - \alpha$ de la loi $\chi^2(r)$, sont convergents de niveau asymptotique α . Les p -valeurs asymptotiques correspondantes sont $p\text{-val}^{(i)} = \mathbb{P}(Z \geq \zeta_n^{(i),\text{obs}})$.

L'application numérique donne $\hat{\theta}_n = 0.7$, $\zeta_n^{(1)} \simeq 1.67$, $\zeta_n^{(2)} \simeq 1.98$ et les p -valeurs $p\text{-val}^{(1)} \simeq 0.64$, $p\text{-val}^{(2)} \simeq 0.58$. Au seuil $\alpha = 5\%$, on a $z_{1-\alpha}(3) \simeq 7.8$. Donc, on accepte l'hypothèse nulle : les données suivent une loi de Poisson. ▲

Exercice IX.18. 1. Soit n le nombre total d'enfants et n_g le nombre total de garçons. La proportion de garçons est $\hat{r} = n_g/n$. Les statistiques du χ^2 empirique sont :

$$\zeta_n^{(1)} = n \sum_{i=1}^2 \frac{(\hat{p}_i - p_i)^2}{\hat{p}_i} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i=1}^2 \frac{(\hat{p}_i - p_i)^2}{p_i},$$

où $\hat{p}_1 = \hat{r}$ est la fréquence empirique des garçons, $\hat{p}_2 = 1 - \hat{r}$, la fréquence empirique des filles, et $p_1 = p_2 = 1/2$. Les statistiques $(\zeta_n^{(i)}, n \geq 1)$ convergent en loi vers un χ^2 avec $2 - 1 = 1$ degré de liberté sous $H_0 = \{r = 1/2\}$, et divergent sous $H_1 = \{r \neq 1/2\}$. Donc, les régions critiques $W_n^{(i)} = \{\zeta_n^{(i)} \geq z_{1-\alpha}(1)\}$, où $z_\beta(k)$ est le quantile d'ordre β de la loi $\chi^2(k)$, définissent des tests convergents de niveau asymptotique (non-uniforme) α . Les p -valeurs asymptotiques (non uniformes) sont $p\text{-val}^{(i)} = \mathbb{P}(Z \geq \zeta_n^{(i),\text{obs}})$, où Z est de loi $\chi^2(1)$.

L'application numérique donne $n = 5 \times 5128 = 25640$, $n_g = 13037$, $\hat{r} \simeq 0.5085$, $\hat{p}_1 = \hat{r} \simeq 0.5085$, $\hat{p}_2 = 1 - \hat{r} \simeq 0.4915$, $p_1 = p_2 = 1/2$, $\zeta_n^{(1),\text{obs}} \simeq 7.35$ et $\zeta_n^{(2),\text{obs}} \simeq 7.35$. Pour $\alpha = 0.01$, on a $z_{1-\alpha}(1) \simeq 6.63$. Les p -valeurs asymptotiques de ces tests sont $p\text{-val}^{(1)} \simeq p\text{-val}^{(2)} \simeq 0.007$. Donc, on rejette l'hypothèse $r = 1/2$ au niveau asymptotique 1% : il n'y a pas autant de garçons que de filles à la naissance.

2. 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 la loi gaussienne $\mathcal{N}(0, 1)$. Un intervalle de confiance pour r de niveau asymptotique α est :

$$IC_\alpha = \left[\hat{r} \pm \phi_{1-\alpha/2} \sqrt{\frac{\hat{r}(1 - \hat{r})}{n}} \right],$$

où ϕ_β est le quantile d'ordre β de la loi gaussienne $\mathcal{N}(0, 1)$.

L'application numérique donne, pour $\alpha = 0.01$, $\phi_{1-\alpha/2} \simeq 2.58$ et :

$$IC_{0,01} = [0.5004; 0.5165].$$

En particulier, on a $1/2 \notin IC_{0,01}$. On vérifie qu'il y a équivalence entre $\{1/2 \notin IC_\alpha\}$ et $W_n^{(1)}$. On a :

$$\begin{aligned} \zeta_n^{(1)} &= n \sum_{i=1}^2 \frac{(\hat{p}_i - p_i)^2}{\hat{p}_i} \\ &= n \left(\frac{(\hat{r} - 1/2)^2}{\hat{r}} + \frac{(\hat{r} - 1/2)^2}{1 - \hat{r}} \right) \\ &= \left(\sqrt{n} \frac{\hat{r} - 1/2}{\sqrt{\hat{r}(1 - \hat{r})}} \right)^2. \end{aligned}$$

Comme la loi du carré d'une gaussienne $\mathcal{N}(0, 1)$ est la loi $\chi(1)^2$, on en déduit que $\phi_{1-\alpha/2}^2 = z_{1-\alpha}(1)$. On en déduit donc que $\{1/2 \notin IC_\alpha\}$ et $W_n^{(1)}$ coïncident (pour l'égalité, il faut prendre l'intervalle ouvert au lieu de l'intervalle fermé IC_α).

3. On calcule d'abord les probabilités théoriques, voir le tableau XIII.3.

| Répartition (G,F) | (5,0) | (4,1) | (3,2) | (2,3) | (1,4) | (0,5) | Total |
|-------------------------------|--------|--------|--------|--------|--------|--------|-------|
| Nombre de familles | 204 | 841 | 1585 | 1544 | 810 | 144 | 51280 |
| Prop. observées : \hat{p}_i | 0.0398 | 0.1640 | 0.3091 | 0.3011 | 0.1580 | 0.0281 | 1 |
| Prop. théoriques : p_i | 1/32 | 5/32 | 10/32 | 10/32 | 5/32 | 1/32 | 1 |

Table XIII.3. Proportions observées et théoriques

On teste si les données suivent la loi binomiale $p^0 = \mathcal{B}(5, 1/2) : H_0 = \{p = p^0\}$ contre l'hypothèse alternative $H_1 = \{p \neq p^0\}$. Les statistiques du χ^2 empirique sont :

$$\zeta_n^{(1)} = n \sum_{i=0}^5 \frac{(\hat{p}_i - p_i^0)^2}{\hat{p}_i} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i=0}^5 \frac{(\hat{p}_i - p_i^0)^2}{p_i}.$$

Les statistiques $(\zeta_n^{(i)}, n \geq 1)$ convergent en loi sous H_0 vers un χ^2 avec $6 - 1 = 5$ degrés de liberté et divergent sous H_1 . Donc, les régions critiques $W_n^{(i)} = \{\zeta_n^{(i)} \geq z_{1-\alpha}(5)\}$ définissent des tests convergents de niveau asymptotique α . Les p -valeurs asymptotiques sont $p\text{-val}^{(i)} = \mathbb{P}(Z \geq \zeta_n^{(i),\text{obs}})$, où Z est de loi $\chi^2(5)$.

L'application numérique donne $\zeta_n^{(1),\text{obs}} \simeq 15.6$ et $\zeta_n^{(2),\text{obs}} \simeq 18.0$. Pour $\alpha = 0.01$, $z_{1-\alpha}(5) \simeq 15.1$. Les p -valeurs asymptotiques de ces tests sont $p\text{-val}^{(1)} \simeq 0.008$ et $p\text{-val}^{(2)} \simeq 0.003$. Donc, on rejette l'hypothèse $r = 1/2$ au niveau $\alpha = 1\%$.

4. On teste si les données suivent la loi binomiale $p(r) = \mathcal{B}(5, r)$ et r inconnu : $H_0 = \{\text{il existe } r \in]0, 1[\text{ tel que } p = p(r)\}$ contre l'hypothèse alternative $H_1 = \{p \neq p(r), r \in]0, 1[\}$. Les probabilités théoriques sont alors $p_0(r) = r^5$, $p_1(r) = \binom{5}{1}(1-r)r^4, \dots, p_5(r) = (1-r)^5$. L'estimateur du maximum de vraisemblance de la fréquence r est la fréquence empirique \hat{r} . Les statistiques du χ^2 sont :

$$\zeta_n^{(1)} = n \sum_{i=0}^5 \frac{(\hat{p}_i - p_i(\hat{r}))^2}{\hat{p}_i} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i=0}^5 \frac{(\hat{p}_i - p_i(\hat{r}))^2}{p_i(\hat{r})}.$$

Elles convergent en loi sous H_0 vers la loi du χ^2 à $6 - 1 - 1 = 4$ degrés de liberté et elles divergent sous H_1 . (On a retiré un degré de liberté pour l'estimation du paramètre r de dimension 1.) Donc, les régions critiques $W_n^{(i)} = \{\zeta_n^{(i)} \geq z_{1-\alpha}(4)\}$ définissent des tests convergents de niveau asymptotique α . Les p -valeurs asymptotiques (non uniformes) sont $p\text{-val}^{(i)} = \mathbb{P}(Z \geq \zeta_n^{(i),\text{obs}})$, où Z est de loi $\chi^2(4)$.

L'application numérique donne $\zeta_n^{(1),\text{obs}} \simeq 9.2$ et $\zeta_n^{(2),\text{obs}} \simeq 10.1$ et pour $\alpha = 0.01$, $z_{1-\alpha}(4) \simeq 13.3$. Les p -valeurs asymptotiques de ces tests sont $p\text{-val}^{(1)} \simeq 0.057$ et $p\text{-val}^{(2)} \simeq 0.040$. Donc, on accepte l'hypothèse selon laquelle les naissances des garçons suivent une loi binomiale. ▲

Exercice IX.19. 1. On teste si les données suivent la loi binomiale $p^0 = \mathcal{B}(12, 1/2)$: $H_0 = \{p = p^0\}$ contre l'hypothèse alternative $H_1 = \{p \neq p^0\}$. Comme les nombres d'observations dans les cases 11 et 12 sont inférieures à 5, on regroupe les cases 10, 11 et 12, et on remplace $p^0(10)$ par $p_{10}^0 + p_{11}^0 + p_{12}^0$. Les statistiques du χ^2 empirique sont :

$$\zeta_n^{(1)} = n \sum_{i=0}^{10} \frac{(\hat{p}_i - p_i^0)^2}{\hat{p}_i} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i=0}^{10} \frac{(\hat{p}_i - p_i^0)^2}{p_i^0}.$$

Les statistiques $(\zeta_n^{(i)}, n \geq 1)$ convergent en loi sous H_0 vers un χ^2 avec $11 - 1 = 10$ degrés de liberté et divergent sous H_1 . Donc, les régions critiques $W_n^{(i)} = \{\zeta_n^{(i)} \geq z_{1-\alpha}(10)\}$, où $z_{1-\alpha}(k)$ est le quantile d'ordre $1 - \alpha$ de la loi du χ^2 à k degrés de liberté, définissent des tests convergents de niveau asymptotique α . Les p -valeurs asymptotiques sont $p\text{-val}^{(i)} = \mathbb{P}(Z \geq \zeta_n^{(i),\text{obs}})$, où Z est de loi $\chi^2(10)$.

L'application numérique donne $\zeta_n^{(1),\text{obs}} \simeq 34$ et $\zeta_n^{(2),\text{obs}} \simeq 35$. Pour $\alpha = 0.05$, $z_{1-\alpha}(10) \simeq 18$. Les p -valeurs asymptotiques de ces tests sont $p\text{-val}^{(1)} \simeq 0.0006$ et $p\text{-val}^{(2)} \simeq 0.0004$. Donc, on rejette l'hypothèse nulle.

2. On teste si les données suivent la loi binomiale $p(r) = \mathcal{B}(12, r)$ et r inconnu : $H_0 = \{\text{il existe } r \in]0, 1[\text{ tel que } p = p(r)\}$ contre l'hypothèse alternative $H_1 = \{p \neq p(r), r \in]0, 1[\}$. Les probabilités théoriques sont alors $p_0(r) = (1-r)^{12}$, $p_1(r) = \binom{12}{1}(1-r)^{11}r, \dots, p_{10}(r) = \binom{12}{10}(1-r)^2r^{10} + \binom{12}{11}(1-r)r^{11} + r^{12}$. L'estimateur du maximum de vraisemblance de la fréquence r , la probabilité d'obtenir un 5 ou un 6 lors d'un lancer de 12 dés, est la fréquence empirique :

$$\hat{r} = \frac{\sum_{i=0}^{12} iN_i}{12 \sum_{i=0}^{12} N_i}.$$

Les statistiques du χ^2 sont :

$$\zeta_n^{(1)} = n \sum_{i=0}^{10} \frac{(\hat{p}_i - p_i(\hat{r}))^2}{\hat{p}_i} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{i=0}^{10} \frac{(\hat{p}_i - p_i(\hat{r}))^2}{p_i(\hat{r})}.$$

Elles convergent en loi sous H_0 vers la loi du χ^2 à $11 - 1 - 1 = 9$ degrés de liberté et elles divergent sous H_1 . (On a retiré un degré de liberté pour l'estimation du paramètre r de dimension 1.) Donc, les régions critiques $W_n^{(i)} = \{\zeta_n^{(i)} \geq z_{1-\alpha}(9)\}$ définissent des tests convergents de niveau asymptotique α . Les p -valeurs asymptotiques (non uniformes) sont $p\text{-val}^{(i)} = \mathbb{P}(Z \geq \zeta_n^{(i), \text{obs}})$, où Z est de loi $\chi^2(9)$.

L'application numérique donne $\hat{r} \simeq 0.338$ (avec l'intervalle de confiance de niveau asymptotique à 95% $[0.332, 0.343]$), $\zeta_n^{(1), \text{obs}} \simeq 8.2$ et $\zeta_n^{(2), \text{obs}} \simeq 8.4$ et pour $\alpha = 0.01$, $z_{1-\alpha}(9) \simeq 16.9$. Les p -valeurs asymptotiques de ces tests sont $p\text{-val}^{(1)} \simeq 0.51$ et $p\text{-val}^{(2)} \simeq 0.50$. Donc, on accepte l'hypothèse selon laquelle les dés sont également biaisés. ▲

XIII.10 Intervalles et régions de confiance

Exercice X.1. 1. On a :

$$\mathbb{E}_\theta[X] = \frac{1}{2} \int_{+\infty}^{-\infty} x e^{-|x-\theta|} dx = \frac{1}{2} \int_{+\infty}^{-\infty} (u + \theta) e^{-|u|} dx = \theta.$$

On obtient également $\mathbb{E}_\theta[X^2] = 2 + \theta^2$ et donc $\text{Var}_\theta(X) = 2$. D'après la méthode des moments, on en déduit que la moyenne empirique $T_n = \bar{X}_n$ est un estimateur de θ . De plus, c'est un estimateur convergent par la loi forte des grands nombres.

2. Le théorème central limite assure que l'intervalle de confiance :

$$I_n = \left[\bar{X}_n \pm \phi_{1-\alpha/2} \sqrt{2/n} \right],$$

où $\phi_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$, est de niveau asymptotique $1 - \alpha$. Pour un niveau asymptotique de confiance de 95%, on obtient $\phi_{1-\alpha/2} \simeq 1.96$ et, avec $n = 200$, il vient :

$$I_n \simeq [\bar{X}_n \pm 0.2].$$

▲

Exercice X.2. 1. Soit $\theta' > \theta$. Le rapport de vraisemblance est :

$$\frac{p_n(x_1, \dots, x_n; \theta')}{p_n(x_1, \dots, x_n; \theta)} = \left(\frac{\theta}{\theta'} \right)^n \frac{\mathbf{1}_{\{\max_{1 \leq i \leq n} x_i \leq \theta', \inf_{1 \leq i \leq n} x_i \geq 0\}}}{\mathbf{1}_{\{\max_{1 \leq i \leq n} x_i \leq \theta, \inf_{1 \leq i \leq n} x_i \geq 0\}}}$$

Si $0 \leq \max_{1 \leq i \leq n} x_i \leq \theta$, 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 $\max_{1 \leq i \leq n} x_i \geq \theta$, 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 de test $\zeta_n = \max_{1 \leq i \leq n} X_i$. Comme la statistique de test possède une densité, il existe un test U.P.P donné par la région critique $W_n = \{\zeta_n \geq z_\alpha\}$, où z_α est déterminé par $\alpha = \sup_{\theta \leq \theta_0} \mathbb{P}_\theta(W_n)$. Il reste encore à déterminer z_α . Pour cela, on calcule la fonction de répartition de ζ_n :

$$\mathbb{P}_\theta(\zeta_n < x) = \mathbb{P}_\theta(X_1 < x)^n = \begin{cases} 0 & \text{si } x < 0, \\ (x/\theta)^n & \text{si } 0 \leq x \leq \theta, \\ 1 & \text{si } x > \theta. \end{cases}$$

Il vient donc $\alpha = \sup_{\theta \leq \theta_0} (1 - (z_\alpha/\theta)^n)$. Comme le maximum est atteint pour $\theta = \theta_0$, il vient :

$$z_\alpha = \theta_0 (1 - \alpha)^{1/n}.$$

En conclusion, on obtient la région critique suivante :

$$W_n = \{\zeta_n \geq \theta_0(1 - \alpha)^{1/n}\}.$$

2. Comme $0 < z_\alpha \leq \theta$, on a :

$$\rho(\theta) = \mathbb{P}_\theta(W_n) = 1 - \mathbb{P}_\theta(\zeta_n < z_\alpha) = 1 - (z_\alpha/\theta)^n = 1 - (1 - \alpha)(\theta_0/\theta)^n.$$

3. Pour $\theta_0 = \frac{1}{2}$ et $\alpha = 5\%$, on obtient $z_\alpha = \frac{1}{2}(0.95)^{1/n}$.

4. Il faut choisir n tel que :

$$1 - 0.95 \left(\frac{1/2}{3/4} \right)^n \geq 0.98,$$

soit $n = 10$. Si $n = 2$, on obtient $\rho(3/4) \simeq 0.578$. Plus n est grand, plus $\rho(\theta)$ est proche de 1. L'erreur de deuxième espèce est proche de 0 quand n est grand. (En effet le test associé à la région critique W_n est convergent.)

5. C'est un cas particulier du test précédent. On conserve le même test.

6. D'après ce qui précède, on a :

$$1 - \alpha' = \mathbb{P}_\theta \left(\zeta_n \geq \frac{\theta}{k_n} \right) = 1 - \mathbb{P}_\theta \left(\zeta_n < \frac{\theta}{k_n} \right).$$

Comme $k_n \geq 1$, 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 $\theta \geq \zeta_n$, on obtient l'intervalle de confiance de niveau (exact) $1 - \alpha'$ pour θ :

$$I_n = \left[\zeta_n, \left(\frac{1}{\alpha'} \right)^{1/n} \zeta_n \right].$$

7. La fonction de répartition F_n de $n(\theta - \zeta_n)$ est, pour $x \geq 0$:

$$F_n(x) = 1 - \mathbb{P}_\theta(n(\theta - \zeta_n) > x) = 1 - \left(1 - \frac{x}{n\theta} \right)^n.$$

En particulier, la suite de fonctions de répartition $(F_n, n \in \mathbb{N}^*)$ converge vers la fonction F définie par :

$$F(x) = 1 - e^{-x/\theta} \quad \text{si } x \geq 0, \quad \text{et} \quad F(x) = 0 \quad \text{si } x \leq 0.$$

Ainsi la suite $(n(\theta - \zeta_n), n \in \mathbb{N}^*)$ converge en loi vers une loi exponentielle de paramètre $1/\theta$. En particulier la suite $(n(1 - \frac{1}{\theta} \zeta_n), n \in \mathbb{N}^*)$ converge en loi vers une variable aléatoire Y de loi exponentielle de paramètre 1. Comme les variables aléatoires exponentielles sont continues, on a :

$$\mathbb{P}_\theta \left(0 \leq n \left(1 - \frac{1}{\theta} \zeta_n \right) \leq a \right) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(0 \leq Y \leq a) = 1 - e^{-a}.$$

Pour $a = -\log(\alpha')$, on a $1 - e^{-a} = 1 - \alpha'$. On en déduit donc que l'intervalle de confiance :

$$J_n = \begin{cases} [\zeta_n; +\infty] & \text{si } n \leq -\log(\alpha'), \\ [\zeta_n; \zeta_n/(1 + \log(\alpha')/n)] & \text{si } n > -\log(\alpha'). \end{cases}$$

est de niveau asymptotique $1 - \alpha'$. Comme $(1/\alpha')^{1/n} \leq 1/(1 + \log(\alpha')/n)$ pour tout α', n tels que $n > -\log(\alpha')$, on en déduit que $I_n \subset J_n$ et donc l'intervalle de confiance J_n est en fait de niveau $1 - \alpha'$ par excès. ▲

Exercice X.3. 1. On a par indépendance :

$$\text{Var}_p(\bar{X}_n) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}_p(X_i) = \frac{p(1-p)}{n} \leq \frac{1}{4n}.$$

On pose :

$$I_n^{(1)} = \left[\bar{X}_n \pm \frac{1}{2\sqrt{\alpha n}} \right].$$

Comme $\mathbb{E}_p[\bar{X}_n] = p$, il vient :

$$\mathbb{P}_p(p \notin I_n^{(1)}) = \mathbb{P}_p \left(|\bar{X}_n - p| > \frac{1}{2\sqrt{\alpha n}} \right) \leq 4\alpha n \text{Var}_p(\bar{X}_n) = \alpha.$$

On en déduit que $I_n^{(1)}$ est un intervalle de confiance de p de niveau par excès $1 - \alpha$.

2. On déduit de la loi forte des grands nombres que $\bar{X}_n(1 - \bar{X}_n)$ est un estimateur convergent de $\sigma^2 = p(1-p)$. On déduit du théorème central limite et du théorème de Slutsky que :

$$\sqrt{n} \frac{\bar{X}_n - p}{\sqrt{\bar{X}_n(1 - \bar{X}_n)}} \xrightarrow[n \rightarrow \infty]{\text{en loi}} G,$$

où G est de loi gaussienne centrée $\mathcal{N}(0, 1)$. On en déduit donc que :

$$I_n^{(2)} = \left[\bar{X}_n \pm \frac{\phi_{1-\alpha/2}}{\sqrt{n}} \sqrt{\bar{X}_n(1 - \bar{X}_n)} \right],$$

où $\phi_{1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, est un intervalle de confiance de p de niveau asymptotique $1 - \alpha$.

3. On déduit du théorème central limite que $(g_n(\bar{X}_n, p), n \in \mathbb{N}^*)$ converge en loi vers G . On considère :

$$I_n^{(3)} = \left[\frac{\bar{X}_n + \frac{\phi_{1-\alpha/2}^2}{2n}}{1 + \frac{\phi_{1-\alpha/2}^2}{n}} \pm \frac{\phi_{1-\alpha/2}}{\sqrt{n}} \frac{\sqrt{\bar{X}_n(1 - \bar{X}_n) + \frac{\phi_{1-\alpha/2}^2}{4n}}}{1 + \frac{\phi_{1-\alpha/2}^2}{n}} \right].$$

On a $p \in I_n^{(3)}$ si et seulement si $g_n(\bar{X}_n, p) \in [\pm\phi_{1-\alpha/2}]$. On en déduit que $I_n^{(3)}$ est un intervalle de confiance de p de niveau asymptotique $1 - \alpha$.

4. On pose :

$$I_n^{(4)} = \left[\bar{X}_n \pm \frac{\phi_{1-\alpha/2}}{2\sqrt{n}} \right].$$

Comme $\bar{X}_n \in [0, 1]$, on en déduit que $I_n^{(3)} \subset I_n^{(4)}$ et donc $I_n^{(4)}$ est un intervalle de confiance de p de niveau asymptotique par excès $1 - \alpha$.

5. Soit $g(x) = \arcsin(2x - 1)$. On a $g'(x)\sqrt{x(1-x)} = 1$ pour $x \in]0, 1[$. On déduit du théorème central limite que $(\sqrt{n}(g(\bar{X}_n) - g(p)), n \in \mathbb{N}^*)$ converge en loi vers $\sigma_p G$, où $\sigma_p^2 = g'(p)^2 p(1-p) = 1$. On en déduit donc, en inversant la fonction g que, pour $\left[\arcsin(2\bar{X}_n - 1) \pm \frac{\phi_{1-\alpha/2}}{\sqrt{n}} \right] \subset [-\pi/2, \pi/2]$:

$$I_n^{(5)} = \left[\frac{1}{2} + \frac{1}{2} \sin \left(\arcsin(2\bar{X}_n - 1) \pm \frac{\phi_{1-\alpha/2}}{\sqrt{n}} \right) \right]$$

est un intervalle de confiance de p de niveau asymptotique $1 - \alpha$. Si la condition $\left[\arcsin(2\bar{X}_n - 1) \pm \frac{\phi_{1-\alpha/2}}{\sqrt{n}} \right] \subset [-\pi/2, \pi/2]$ n'est pas satisfaite, il faut être plus précis lors de l'inversion de la fonction $p \mapsto \arcsin(2p - 1)$.

6. On déduit de l'inégalité de Markov que, pour $\lambda > 0$:

$$\mathbb{P}_p(\bar{X}_n > a) = \mathbb{P}_p \left(e^{\lambda\bar{X}_n - \lambda a} > 1 \right) \leq \mathbb{E}_p \left[e^{\lambda\bar{X}_n - \lambda a} \right] = e^{-\lambda a} \left(1 - p + p e^{\lambda/n} \right)^n.$$

Le terme de droite est minimal pour $\lambda = n \log(a/p) - n \log((1-a)/(1-p))$, et on obtient :

$$\mathbb{P}_p(\bar{X}_n > a) \leq e^{-nH_p(a)}, \quad \text{où } H_p(a) = a \log(a/p) + (1-a) \log((1-a)/(1-p)).$$

Par symétrie, on obtient pour $p > a > 0$:

$$\mathbb{P}_p(\bar{X}_n < a) = \mathbb{P}_{1-p}(\bar{X}_n > 1-a) \leq e^{-nH_{1-p}(1-a)} = e^{-nH_p(a)}.$$

La fonction $a \mapsto H_p(a)$ est convexe ($\partial_a^2 H_p(a) > 0$), atteint son minimum en p ($\partial_a H_p(p) = 0$) qui est nul, et diverge en 0 et 1. Il existe donc a_n^+ et a_n^- tels que $0 < a_n^- < p < a_n^+ < 1$ et $H_p(a^\pm) = \log(2/\alpha)/n$. On en déduit que :

$$I_n^{(6)} = \left[\bar{X}_n - (a_n^+ - p), \bar{X}_n - (a_n^- - p) \right]$$

est un intervalle de confiance pour p de niveau par excès $1 - \alpha$.

7. Comme $p \in [0, 1]$, on considère la restriction des intervalles de confiance sont sur $[0, 1]$:

$$\begin{aligned}
 I_n^{(1)} &= \left[0, \frac{1}{2\sqrt{\alpha n}} \right], \\
 I_n^{(2)} &= \{0\}, \\
 I_n^{(3)} &= \left[0, \frac{\phi_{1-\alpha/2}^2}{n} \frac{1}{1 + \frac{\phi_{1-\alpha/2}^2}{n}} \right] \simeq \left[0, \frac{\phi_{1-\alpha/2}^2}{n} \right], \\
 I_n^{(4)} &= \left[0, \frac{\phi_{1-\alpha/2}}{2\sqrt{n}} \right], \\
 I_n^{(5)} &= \left[0, \frac{1}{2} - \frac{1}{2} \cos \left(\frac{\phi_{1-\alpha/2}}{\sqrt{n}} \right) \right] \simeq \left[0, \frac{\phi_{1-\alpha/2}^2}{4n} \right], \\
 I_n^{(6)} &= [0, p - a_n^-] \simeq \left[0, \frac{\sqrt{2p(1-p)} \log(2/\alpha)}{\sqrt{n}} \right].
 \end{aligned}$$

Par ailleurs, on a $\mathbb{P}_p(\bar{X}_n = 0) = (1 - p)^n$. Les valeurs de p , pour lesquelles il est raisonnable d'observer $\bar{X}_n = 0$ (avec au moins une probabilité $1 - \alpha$) sont telles que $p \leq 1 - (1 - \alpha)^{1/n}$ soit p dans l'intervalle :

$$I_n^{(7)} = [0, 1 - (1 - \alpha)^{1/n}] \simeq [0, -\log(1 - \alpha)/n].$$

Ceci signifie que pour $\bar{X}_n = 0$, les intervalles de confiance $I_n^{(3)}$ et $I_n^{(5)}$ donnent le bon ordre de grandeur. ▲

XIII.11 Problèmes (probabilités)

Exercice XI.1. 1. La variable aléatoire X_1 est de carré intégrable. La loi forte des grands nombres assure que $(R_n^2/n, n \geq 1)$ converge p.s. vers $\mathbb{E}[X_1^2] = 1$. Comme la fonction $a \mapsto \sqrt{a}$ est continue sur $]0, +\infty[$, on en déduit que $(R_n/\sqrt{n}, n \geq 1)$ converge p.s. vers 1.

2. Par indépendance, la variable aléatoire Y_n a pour densité :

$$f(y_n) = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{k=1}^n x_k^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-|y_n|^2/2}.$$

On a :

$$\begin{aligned}
 \mathbb{E}[g(\Theta_{n-1})] &= \mathbb{E}[g(Y_n/|Y_n|)] \\
 &= \int_{\mathbb{R}^n} g(y_n/|y_n|) f(y_n) dy_n \\
 &= \int_{\mathbb{R}^n} g(y_n/|y_n|) \frac{1}{(2\pi)^{n/2}} e^{-|y_n|^2/2} dy_n \\
 &= \int_{]0,+\infty[} \frac{r^{n-1}}{(2\pi)^{n/2}} e^{-r^2/2} dr \int_{\mathcal{S}_{n-1}} \sigma_{n-1}(d\theta_{n-1}) g(\theta_{n-1}) \\
 &= \frac{1}{\alpha_{n-1}} \int_{\mathcal{S}_{n-1}} \sigma_{n-1}(d\theta_{n-1}) g(\theta_{n-1}),
 \end{aligned}$$

où $1/\alpha_{n-1} = \int_{]0,+\infty[} \frac{r^{n-1}}{(2\pi)^{n/2}} e^{-r^2/2} dr$. On vérifie que $\alpha_{n-1} = c_{n-1}$ en prenant $g = 1$. On remarque que :

$$\frac{1}{c_{n-1}} = \frac{1}{\alpha_{n-1}} = \int_{]0,+\infty[} \frac{r^{n-1}}{(2\pi)^{n/2}} e^{-r^2/2} dr = \int_{]0,+\infty[} \frac{v^{n/2-1}}{(2\pi)^{n/2}} e^{-v} dv = \frac{\Gamma(n/2)}{(2\pi)^{n/2}},$$

soit $c_{n-1} = (2\pi)^{n/2}/\Gamma(n/2)$.

3. La question précédente assure que $U^{(n)}$ a même loi que $\sqrt{n}\Theta_{n-1} = \sqrt{n}Y_n/|Y_n|$. Ainsi, $U_1^{(n)}$ a même loi que $\sqrt{n}X_1/R_n$. La question 1 assure donc que p.s. $\lim_{n \rightarrow +\infty} \sqrt{n}X_1/R_n = X_1$. Par convergence dominée, on a pour $u \in \mathbb{R}$:

$$\psi_{U_1^{(n)}}(u) = \psi_{\sqrt{n}X_1/R_n}(u) = \mathbb{E} \left[e^{iu\sqrt{n}X_1/R_n} \right] \xrightarrow[n \rightarrow \infty]{} \mathbb{E} [e^{iuX_1}].$$

Ceci assure que $(U_1^{(n)}, n \geq 2)$ converge en loi vers X_1 . (On aurait pu aussi démontrer le résultat en utilisant le théorème de Slutsky.)

▲

Exercice XI.2. 1. On note X_k le numéro de l'image dans la tablette k . Les variables aléatoires $(X_k, k \in \mathbb{N}^*)$ sont indépendantes et de loi uniforme sur $\{1, \dots, n\}$. On a $T_n = \inf\{k \geq 2; X_k \in \{X_1, \dots, X_{k-1}\}\}$. Soit $k \in \{2, \dots, n\}$. On a $\{T_n > k\} = \{T_n > k-1, X_k \notin \{X_1, \dots, X_{k-1}\}\}$. Par indépendance, on a $\mathbb{P}(X_k \notin \{X_1, \dots, X_{k-1}\} | T_n > k-1) = 1 - \frac{k-1}{N}$. Il vient donc $\mathbb{P}(T_n > k) = \mathbb{P}(T_n > k-1) (1 - \frac{k-1}{N})$. Comme $\mathbb{P}(T_n > 1) = 1$, on en déduit donc que $\mathbb{P}(T_n > k) = \prod_{i=1}^k (1 - \frac{i-1}{n})$. Cette formule est valable pour $k \in \mathbb{N}^*$.

2. On a, pour $x \leq 0$, $\mathbb{P}(T_n/\sqrt{n} \leq x) = 0$. Soit $x > 0$. On a :

$$\mathbb{P}(T_n/\sqrt{n} \leq x) = 1 - \mathbb{P}(T_n > [\sqrt{nx}]) = 1 - e^{\sum_{i=1}^{[\sqrt{nx}]} \log(1 - \frac{i-1}{n})}.$$

Pour $i \in \{1, \dots, [\sqrt{nx}]\}$, on a $\log(1 - \frac{i-1}{n}) = -\frac{i-1}{n} + \frac{(i-1)^2}{n^2}g((i-1)/n)$, où g est une fonction bornée. On en déduit que :

$$\sum_{i=1}^{[\sqrt{nx}]} \log\left(1 - \frac{i-1}{n}\right) = -\frac{1}{2n}[\sqrt{nx}]([\sqrt{nx}] - 1) + O(n^{-1/2}).$$

On en déduit donc que pour tout $x \in \mathbb{R}$, on a $\lim_{n \rightarrow \infty} \mathbb{P}(T_n/\sqrt{n} \leq x) = F(x)$ avec $F(x) = (1 - e^{-x^2/2})\mathbf{1}_{x>0}$. La fonction F est croissante nulle en $-\infty$ et vaut 1 en $+\infty$. Il s'agit de la fonction de répartition d'une variable aléatoire réelle X .

3. Comme la fonction F est continue et de classe C^1 sauf en 0, on en déduit que la loi de X possède la densité $F'(x) = 2x e^{-x^2} \mathbf{1}_{\{x>0\}}$.
4. Comme $\mathbb{P}(X^2 \leq x) = \mathbb{P}(X \leq \sqrt{x})\mathbf{1}_{\{x>0\}} = (1 - e^{-x/2})\mathbf{1}_{\{x>0\}}$, on en déduit que X^2 est de loi exponentielle de paramètre 1/2.
5. On a $p_k = \mathbb{P}(T_n > k)$ avec $n = 365$. On obtient les approximations suivantes :

$$p_k = \mathbb{P}(T_n > k) = \mathbb{P}(T_n/\sqrt{n} > k/\sqrt{n}) \simeq e^{-k^2/2n}.$$

En utilisant cette approximation, on obtient pour $n = 365$:

$$p_{23} \simeq 0.484, \quad p_{40} \simeq 0.112 \quad \text{et} \quad p_{366} \simeq 2 \cdot 10^{-80}.$$

Les valeurs exactes (à 10^{-3} près) sont :

$$p_{23} \simeq 0.493, \quad p_{40} \simeq 0.109 \quad \text{et} \quad p_{366} = 0.$$

▲

Exercice XI.3. 1. Les variables aléatoires sont indépendantes et de même loi uniforme sur $\{1, \dots, n\}$.

2. On décompose suivant la valeur de la première image, et en utilisant l'indépendance, il vient :

$$\begin{aligned} \mathbb{P}(T_2 = \ell) &= \sum_{1 \leq j \leq n} \mathbb{P}(X_1 = j; \dots, X_\ell = j; X_{\ell+1} \neq j) \\ &= \sum_{1 \leq j \leq n} \mathbb{P}(X_1 = j) \cdots \mathbb{P}(X_{\ell+1} \neq j) \\ &= \sum_{1 \leq j \leq n} \left(\frac{1}{n}\right)^\ell \left(1 - \frac{1}{n}\right) \\ &= \left(1 - \frac{1}{n}\right) \left(\frac{1}{n}\right)^{\ell-1}. \end{aligned}$$

Donc la loi de T_2 est la loi géométrique de paramètre $(1 - \frac{1}{n})$.

3. On décompose suivant les cas possibles pour les images.
4. On déduit de la formule précédente en utilisant l'indépendance que :

$$\begin{aligned} \mathbb{P}(T_2 = \ell_2, T_3 = \ell_3) &= n(n-1)(n-2) \left(\frac{1}{n}\right)^{\ell_2} \frac{1}{n} \left(\frac{2}{n}\right)^{\ell_3-1} \frac{1}{n} \\ &= \left(1 - \frac{1}{n}\right) \left(\frac{1}{n}\right)^{\ell_2-1} \left(1 - \frac{2}{n}\right) \left(\frac{2}{n}\right)^{\ell_3-1}. \end{aligned}$$

En utilisant la formule des lois marginales, on obtient :

$$\begin{aligned} \mathbb{P}(T_3 = \ell_3) &= \sum_{\ell_2 \geq 1} \mathbb{P}(T_2 = \ell_2, T_3 = \ell_3) \\ &= \left(1 - \frac{2}{n}\right) \left(\frac{2}{n}\right)^{\ell_3-1}. \end{aligned}$$

Donc la loi de T_3 est la loi géométrique de paramètre $(1 - \frac{2}{n})$.

5. On a pour tous $\ell_2 \in \mathbb{N}^*$ et $\ell_3 \in \mathbb{N}^*$ que $\mathbb{P}(T_2 = \ell_2, T_3 = \ell_3) = \mathbb{P}(T_2 = \ell_2)\mathbb{P}(T_3 = \ell_3)$. Donc T_2 et T_3 sont indépendants.
6. La variable aléatoire T_k représente le premier instant où l'on obtient une nouvelle image alors que l'on en possède déjà $k-1$. Elle est donc de loi géométrique de paramètre p , où p est la probabilité de succès : obtenir une nouvelle carte alors que l'on en possède déjà $k-1$. On a ainsi $p = 1 - \frac{k-1}{n}$.
7. On a par linéarité de l'espérance :

$$\mathbb{E}[N_n] = \mathbb{E}\left[\sum_{k=1}^n T_k\right] = \sum_{k=1}^n \mathbb{E}[T_k] = \sum_{k=1}^n \frac{n}{n-k+1} = n \sum_{j=1}^n \frac{1}{j}.$$

On en déduit que $\mathbb{E}[N_n] = n(\log(n) + O(1))$.

8. On a par indépendance :

$$\begin{aligned} \text{Var}(N_n) &= \text{Var}\left(\sum_{k=1}^n T_k\right) = \sum_{k=1}^n \text{Var}(T_k) = \sum_{k=1}^n \frac{n(k-1)}{(n-k+1)^2} \\ &= \sum_{k=1}^n \frac{n^2}{(n-k+1)^2} - \frac{n}{(n-k+1)} \\ &= n^2 \sum_{j=1}^n \frac{1}{j^2} - n \sum_{j=1}^n \frac{1}{j} \\ &= n^2 \frac{\pi^2}{6} + o(n^2). \end{aligned}$$

9. 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 :

$$\mathbb{P} \left(\left| \frac{N_n}{\mathbb{E}[N_n]} - 1 \right| > \varepsilon \right) = \mathbb{E} \left[\mathbf{1}_{\left\{ \left| \frac{N_n}{\mathbb{E}[N_n]} - 1 \right|^2 > \varepsilon^2 \right\}} \right] \leq \varepsilon^{-2} \mathbb{E} \left[\left(\frac{N_n}{\mathbb{E}[N_n]} - 1 \right)^2 \right].$$

10. On a :

$$\mathbb{E} \left[\left(\frac{N_n}{\mathbb{E}[N_n]} - 1 \right)^2 \right] = \frac{1}{\mathbb{E}[N_n]^2} \text{Var}(N_n) = O \left(\frac{n^2}{n^2 (\log n)^2} \right) = O((\log n)^{-2}).$$

On en déduit que pour tout $\varepsilon > 0$, $\lim_{n \rightarrow \infty} \mathbb{P} \left(\left| \frac{N_n}{\mathbb{E}[N_n]} - 1 \right| > \varepsilon \right) = 0$. Donc la suite

$\left(\frac{N_n}{\mathbb{E}[N_n]}, n \in \mathbb{N}^* \right)$ converge en probabilité vers 1. Comme $\lim_{n \rightarrow \infty} \frac{\mathbb{E}[N_n]}{n \log n} = 1$, 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 $f(x) = e^{-x} e^{-e^{-x}}$, $x \in \mathbb{R}$ appelée loi de Gumbel (voir l'exercice XI.4). ▲

Exercice XI.4. I Comportement asymptotique de $X_{(n)} = \sum_{i=1}^n Y_i$.

1. En utilisant l'indépendance, on a pour $x \geq 0$:

$$\mathbb{P}(X_{(n)} \leq x) = \mathbb{P}(\max_{1 \leq i \leq n} X_i \leq x) = \mathbb{P}(X_1 \leq x) \cdots \mathbb{P}(X_n \leq x) = (1 - e^{-\lambda x})^n.$$

2. On détermine la fonction de répartition. On a pour $x + \lambda^{-1} \log n \geq 0$:

$$\mathbb{P}(X_{(n)} - \lambda^{-1} \log n \leq x) = \mathbb{P}(X_{(n)} \leq x + \lambda^{-1} \log n) = \left(1 - \frac{e^{-\lambda x}}{n} \right)^n.$$

Cette quantité converge vers $F(x) = e^{-e^{-\lambda x}}$ pour tout $x \in \mathbb{R}$. Comme $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow +\infty} F(x) = 1$, 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 . La loi de Z porte le nom de loi de Gumbel.

3. On a $f(x) = e^{-x} e^{-e^{-x}}$ pour $x \in \mathbb{R}$; la fonction f est représentée dans la figure XIII.1. On obtient $e^{-e^{-a}} = 0.025$ soit $a \simeq -1.3$ et $e^{-e^{-b}} = 1 - 0.025$ soit $b \simeq 3.7$.

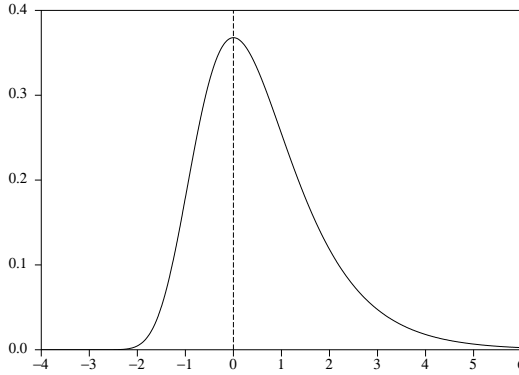


Figure XIII.1. Densité de la loi de Gumbel.

II Loi du vecteur (Y_1, \dots, Y_n) .

1. Comme les lois possèdent des densités et que les variables aléatoires sont indépendantes, $\mathbb{P}(X_i = X_j) = 0$ si $i \neq j$. En effet, on a :

$$\mathbb{P}(X_i = X_j) = \mathbb{E}[\mathbf{1}_{\{X_i=X_j\}}] = \int_{\mathbb{R}_+^2} \mathbf{1}_{\{x=y\}} \lambda^2 e^{-\lambda x - \lambda y} dx dy = 0.$$

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 :

$$\begin{aligned} \mathbb{E}[g_1(Y_1)g_2(Y_2)] &= \mathbb{E}[g_1(X_1)g_2(X_2 - X_1)\mathbf{1}_{\{X_1 < X_2\}}] + \mathbb{E}[g_1(X_2)g_2(X_1 - X_2)\mathbf{1}_{\{X_2 < X_1\}}] \\ &= 2 \int g_1(x_1)g_2(x_2 - x_1)\mathbf{1}_{\{0 < x_1 < x_2\}} \lambda^2 e^{-\lambda x_1 - \lambda x_2} dx_1 dx_2 \\ &= 2 \int g_1(x_1)g_2(y_2)\mathbf{1}_{\{0 < x_1\}}\mathbf{1}_{\{0 < y_2\}} \lambda^2 e^{-\lambda x_1 - \lambda(y_2 + x_1)} dx_1 dy_2, \end{aligned}$$

où on a posé $y_2 = x_2 - x_1$, à x_1 fixé. Il vient :

$$\mathbb{E}[g_1(Y_1)g_2(Y_2)] = \int g_1(y_1)g_2(y_2) 2\lambda^2 e^{-2\lambda y_1 - \lambda y_2} \mathbf{1}_{\{y_1 > 0, y_2 > 0\}} dy_1 dy_2,$$

où on a posé $y_1 = x_1$. En faisant $g_2 = \mathbf{1}$, on obtient :

$$\begin{aligned}\mathbb{E}[g_1(Y_1)] &= \int g_1(y_1) 2\lambda^2 e^{-2\lambda y_1 - \lambda y_2} \mathbf{1}_{\{y_1 > 0, y_2 > 0\}} dy_1 dy_2 \\ &= \int g_1(y_1) 2\lambda e^{-2\lambda y_1} \mathbf{1}_{\{y_1 > 0\}} dy_1.\end{aligned}$$

On en déduit que la loi de Y_1 a pour densité $f_{Y_1}(y_1) = 2\lambda e^{-2\lambda y_1} \mathbf{1}_{\{y_1 > 0\}}$. 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 : $f_{Y_2}(y_2) = \lambda e^{-\lambda y_2} \mathbf{1}_{\{y_2 > 0\}}$. 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 variables aléatoires 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 événements $\{X_{\sigma(1)} < \dots < X_{\sigma(n)}\}$, et en utilisant le fait que $(X_{\sigma(1)}, \dots, X_{\sigma(n)})$ et (X_1, \dots, X_n) ont même loi, on a :

$$\begin{aligned}\mathbb{E}[g_1(Y_1) \cdots g_n(Y_n)] &= \sum_{\sigma \in \mathcal{S}_n} \mathbb{E}[g_1(Y_1) \cdots g_n(Y_n) \mathbf{1}_{\{X_{\sigma(1)} < \dots < X_{\sigma(n)}\}}] \\ &= \sum_{\sigma \in \mathcal{S}_n} \mathbb{E}[g_1(X_{\sigma(1)}) \cdots g_n(X_{\sigma(n)} - X_{\sigma(n-1)}) \mathbf{1}_{\{X_{\sigma(1)} < \dots < X_{\sigma(n)}\}}] \\ &= \sum_{\sigma \in \mathcal{S}_n} \mathbb{E}[g_1(X_1) \cdots g_n(X_n - X_{n-1}) \mathbf{1}_{\{X_1 < \dots < X_n\}}] \\ &= n! \mathbb{E}[g_1(X_1) \cdots g_n(X_n - X_{n-1}) \mathbf{1}_{\{X_1 < \dots < X_n\}}] \\ &= n! \int g_1(x_1) \cdots g_n(x_n - x_{n-1}) \\ &\quad \mathbf{1}_{\{0 < x_1 < \dots < x_n\}} \lambda^n e^{-\lambda \sum_{i=1}^n x_i} dx_1 \cdots dx_n.\end{aligned}$$

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, en utilisant la notation $u_{\sigma^{-1}} = (u_{\sigma^{-1}(1)}, \dots, u_{\sigma^{-1}(n)})$, l'indépendance et enfin le fait que les variables aléatoires X_1, \dots, X_n ont même loi :

$$\begin{aligned}
 \psi_{(X_{\sigma(1)}, \dots, X_{\sigma(n)})}(u) &= \mathbb{E} \left[e^{i \sum_{k=1}^n X_{\sigma(k)} u_k} \right] = \mathbb{E} \left[e^{i \sum_{j=1}^n X_j u_{\sigma^{-1}(j)}} \right] \\
 &= \prod_{j=1}^n \mathbb{E} \left[e^{i X_j u_{\sigma^{-1}(j)}} \right] \\
 &= \prod_{k=1}^n \mathbb{E} \left[e^{i X_1 u_k} \right] \\
 &= \psi_{(X_1, \dots, X_n)}(u).
 \end{aligned}$$

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 $\{(x_1, \dots, x_n); 0 < x_1 < \dots < x_n\}$ dans l'ouvert $\{(y_1, \dots, y_n) \in]0, +\infty[^n\}$. Le Jacobien de l'application φ est :

$$\text{Jac}[\varphi](x) = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix}.$$

Son déterminant est $\det(\text{Jac}[\varphi](x)) = 1$. On remarque que $x_i = \sum_{j=1}^i y_j$, et donc il vient :

$$\sum_{i=1}^n x_i = \sum_{i=1}^n \sum_{j=1}^i y_j = \sum_{j=1}^n y_j \sum_{i=j}^n 1 = \sum_{j=1}^n (n - j + 1) y_j.$$

On en déduit en faisant le changement de variables $(y_1, \dots, y_n) = \varphi(x_1, \dots, x_n)$, que :

$$\begin{aligned}
 &\mathbb{E}[g_1(Y_1) \cdots g_n(Y_n)] \\
 &= n! \int g_1(x_1) \cdots g_n(x_n - x_{n-1}) \mathbf{1}_{\{x_1 < \dots < x_n\}} \lambda^n e^{-\lambda \sum_{i=1}^n x_i} dx_1 \cdots dx_n \\
 &= n! \int g_1(y_1) \cdots g_n(y_n) \mathbf{1}_{\{0 < y_1, \dots, 0 < y_n\}} \lambda^n e^{-\lambda \sum_{j=1}^n (n-j+1)y_j} dy_1 \cdots dy_n.
 \end{aligned}$$

En posant $g_j = \mathbf{1}$ si $j \neq i$, on obtient :

$$\begin{aligned}
 \mathbb{E}[g_i(Y_i)] &= n! \int g_i(y_i) \mathbf{1}_{\{0 < y_1, \dots, 0 < y_n\}} \lambda^n e^{-\lambda \sum_{j=1}^n (n-j+1)y_j} dy_1 \cdots dy_n \\
 &= \int g_i(y_i) \mathbf{1}_{\{0 < y_i\}} (n - i + 1) \lambda e^{-\lambda(n-i+1)y_i} dy_i.
 \end{aligned}$$

La loi de Y_i est donc une loi exponentielle de paramètre $n - i + 1$. On remarque 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 $\frac{1}{n} T_{n-k+1, n}$ est pour $u \in \mathbb{R}$:

$$\psi_{k, n}(u) = \frac{k}{n} \frac{e^{iu/n}}{1 - (1 - \frac{k}{n}) e^{iu/n}} = \frac{k}{k - iu} (1 + o(1)).$$

On pose $\psi_k(u) = \frac{k}{k - iu}$. La suite $(\psi_{k, n}, n > k)$ converge vers ψ_k , fonction caractéristique de la loi exponentielle de paramètre k . On en déduit que la suite $\left(\frac{1}{n} T_{n-k+1, n}, n \geq k\right)$ converge en loi vers la loi exponentielle de paramètre k .

On regarde plus en détail $\psi_{k, n} - \psi_k$. On a :

$$\begin{aligned} \psi_{k, n}(u) - \psi_k(u) &= \frac{k}{n} \frac{e^{iu/n}}{1 - (1 - \frac{k}{n}) e^{iu/n}} - \frac{k}{k - iu} \\ &= \frac{k}{k - n(1 - e^{-iu/n})} - \frac{k}{k - iu} \\ &= k \frac{-iu + n(1 - e^{-iu/n})}{(k - iu)(k - n(1 - e^{-iu/n}))}. \end{aligned}$$

On a $\left| \frac{k}{k - iu} \right| \leq 1$ et $\left| -iu + n(1 - e^{-iu/n}) \right| \leq C \frac{u^2}{n}$, où C est une constante indépendante de n et u . Enfin, on a $\left| k - n(1 - e^{-iu/n}) \right| \geq |k - n(1 - \cos(u/n))|$. 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

$\left|k - n(1 - e^{-iu/n})\right| \geq k - 1/2 \geq k/2$. 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 $2Cu^2$, on obtient bien la majoration (XI.1).

2. Comme les variables aléatoires $T_{1,n}, \dots, T_{n,n}$ sont indépendantes, on a donc

$\psi_{N_n/n}(u) = \prod_{k=1}^n \psi_{k,n}(u)$. 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 obtient donc en utilisant (XI.2) :

$$\left| \psi_{N_n/n}(u) - \psi_{X_{(n)}}(u) \right| \leq \sum_{k=1}^n |\psi_{k,n}(u) - \psi_k(u)| \leq C \frac{\log n}{n}.$$

3. On utilise les fonctions caractéristiques. Il vient :

$$\begin{aligned} & \left| \psi_{(N_n - n \log n)/n}(u) - \psi_Z(u) \right| \\ & \leq \left| \psi_{(N_n - n \log n)/n}(u) - \psi_{(X_{(n)} - n \log n)}(u) \right| + \left| \psi_{(X_{(n)} - n \log n)}(u) - \psi_Z(u) \right|. \end{aligned}$$

Après avoir remarqué que :

$$\psi_{(N_n - n \log n)/n}(u) - \psi_{(X_{(n)} - n \log n)}(u) = e^{-iu n \log n} \left(\psi_{N_n/n}(u) - \psi_{X_{(n)}}(u) \right),$$

on en déduit que :

$$\begin{aligned} & \left| \psi_{(N_n - n \log n)/n}(u) - \psi_Z(u) \right| \\ & \leq \left| \psi_{N_n/n}(u) - \psi_{X_{(n)}}(u) \right| + \left| \psi_{(X_{(n)} - n \log n)}(u) - \psi_Z(u) \right|. \end{aligned}$$

Soit $u \in \mathbb{R}$ fixé. Le premier terme du membre de droite converge vers 0 quand n tend vers l'infini. 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 tend vers l'infini. Par conséquent, on a :

$$\lim_{n \rightarrow \infty} \psi_{(N_n - n \log n)/n}(u) = \psi_Z(u).$$

La suite $(n^{-1}(N_n - n \log n), n \in \mathbb{N}^*)$ converge en loi vers 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 :

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}((N_n - n \log n)/n \in [a, b]) &= \mathbb{P}(Z \in [a, b]) \\ &= 1 - \int_{-\infty}^a f(x) dx - \int_b^{+\infty} f(x) dx. \end{aligned}$$

Comme $\mathbb{P}((N_n - n \log n)/n \in [a, b]) = \mathbb{P}(N_n \in [an + n \log n, bn + n \log n])$, on en déduit que $I_n = [an + n \log n, bn + n \log n]$ est un intervalle aléatoire contenant N_n avec une probabilité asymptotique $\int_a^b f(x) dx$. 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 aléatoire contenant avec une probabilité asymptotique de 95% le nombre, N_n , de plaquettes que l'on doit acheter pour avoir une collection complète. Les valeurs de r_n et I_n en fonction de n sont données dans le tableau XIII.4.

| n | r_n | I_n |
|-----|-------|--------------|
| 151 | 758 | [560, 1313] |
| 250 | 1380 | [1054, 2299] |

Table XIII.4. Valeurs de r_n et I_n en fonction de n .



Exercice XI.5. I Préliminaires

1. Le temps moyen est $\mathbb{E}[T_1] = 1/\lambda$.
2. On considère les fonctions caractéristiques. Par indépendance, on a pour $n \geq 1$ et $u \in \mathbb{R}$:

$$\psi_{V_n}(u) = \psi_{\sum_{i=1}^n T_i}(u) = \prod \psi_{T_i}(u) \quad \text{p} = \left(\frac{\lambda}{\lambda - iu} \right)^n.$$

On reconnaît la fonction caractéristique de la loi gamma de paramètre (λ, n) .

3. On a $\mathbb{P}(N_t = 0) = \mathbb{P}(T_1 > t) = e^{-\lambda t}$.
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 :

$$\begin{aligned}
 \mathbb{P}(N_t = n) &= \mathbb{P}(V_n \leq t < V_n + T_{n+1}) \\
 &= \int \mathbf{1}_{\{v \leq t < v+w\}} \frac{1}{(n-1)!} \lambda^n v^{n-1} e^{-\lambda v} \mathbf{1}_{\{v>0\}} \lambda e^{-\lambda w} \mathbf{1}_{\{w>0\}} dv dw \\
 &= \int \frac{1}{(n-1)!} \lambda^n v^{n-1} e^{-\lambda v} \lambda e^{-\lambda w} \mathbf{1}_{\{0 < v \leq t < v+w\}} dv dw \\
 &= \int \frac{1}{(n-1)!} \lambda^n v^{n-1} e^{-\lambda v} e^{-\lambda(t-v)} \mathbf{1}_{\{0 < v \leq t\}} dv \\
 &= \frac{1}{(n-1)!} \lambda^n e^{-\lambda t} \int v^{n-1} \mathbf{1}_{\{0 < v \leq t\}} dv \\
 &= e^{-\lambda t} \frac{(\lambda t)^n}{n!}.
 \end{aligned}$$

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 p.s. $R_t \leq t$. 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 on a :

$$\mathbb{P}(R_t \leq r, S_t \leq s, N_t = 0) = \mathbb{P}(t < T_1 \leq t + s) = e^{-\lambda t} (1 - e^{-\lambda s}).$$

2. Soit $n \geq 1$. On pose $I = \mathbb{P}(R_t \leq r, S_t \leq s, N_t = n)$. On a :

$$\begin{aligned}
 I &= \mathbb{P}(t - V_n \leq r, V_n + T_{n+1} \leq t + s, V_n \leq t < V_n + T_{n+1}) \\
 &= \mathbb{P}((t - r) \leq V_n \leq t < V_n + T_{n+1} \leq t + s) \\
 &= \int \mathbf{1}_{\{t-r \leq v \leq t < v+w \leq t+s\}} \frac{1}{(n-1)!} \lambda^n v^{n-1} e^{-\lambda v} \mathbf{1}_{\{0 < v\}} \lambda e^{-\lambda w} \mathbf{1}_{\{0 < w\}} dv dw,
 \end{aligned}$$

car les variables V_n et T_{n+1} sont indépendantes. Il vient :

$$\begin{aligned}
 I &= \int \mathbf{1}_{\{(t-r)_+ \leq v \leq t\}} \frac{1}{(n-1)!} \lambda^n v^{n-1} e^{-\lambda v} \mathbf{1}_{\{t-v < w \leq t+s-v\}} \lambda e^{-\lambda w} dv dw \\
 &= \int \mathbf{1}_{\{(t-r)_+ \leq v \leq t\}} \frac{1}{(n-1)!} \lambda^n v^{n-1} e^{-\lambda v} e^{-\lambda(t-v)} (1 - e^{-\lambda s}) dv \\
 &= \frac{1}{(n-1)!} \lambda^n e^{-\lambda t} (1 - e^{-\lambda s}) \int \mathbf{1}_{\{(t-r)_+ \leq v \leq t\}} v^{n-1} dv \\
 &= e^{-\lambda t} (1 - e^{-\lambda s}) \frac{\lambda^n}{n!} [t^n - (t-r)_+^n].
 \end{aligned}$$

3. En décomposant suivant les valeurs possibles de N_t , on obtient :

$$\begin{aligned}
 \mathbb{P}(R_t \leq r, S_t \leq s) &= \sum_{n=0}^{\infty} \mathbb{P}(R_t \leq r, S_t \leq s, N_t = n) \\
 &= e^{-\lambda t} (1 - e^{-\lambda s}) \left\{ \mathbf{1}_{\{r \geq t\}} + \sum_{n=1}^{\infty} \frac{\lambda^n}{n!} [t^n - (t-r)_+^n] \right\} \\
 &= e^{-\lambda t} (1 - e^{-\lambda s}) \left\{ e^{\lambda t} - e^{\lambda(t-r)} \mathbf{1}_{\{r < t\}} \right\} \\
 &= (1 - e^{-\lambda s}) (1 - e^{-\lambda r} \mathbf{1}_{\{r < t\}}).
 \end{aligned}$$

4. Comme $R_t \leq t$ p.s., on a $F(s) = \mathbb{P}(S_t \leq s) = \mathbb{P}(R_t \leq t, S_t \leq s) = (1 - e^{-\lambda s})$. Comme S_t est positif p.s., on en déduit que sa fonction de répartition est $F(s) = (1 - e^{-\lambda s}) \mathbf{1}_{\{s > 0\}}$. On reconnaît la fonction de la loi exponentielle de paramètre λ . Donc la loi de S_t est la loi exponentielle de paramètre λ .
5. La variable aléatoire S_t est p.s. finie. On en déduit que :

$$\mathbb{P}(R_t \leq r) = \lim_{s \rightarrow \infty} \mathbb{P}(R_t \leq r, S_t \leq s) = (1 - e^{-\lambda r} \mathbf{1}_{\{r < t\}}).$$

On calcule la fonction de répartition de $\min(T_1, t)$. On a :

$$\mathbb{P}(\min(T_1, t) \leq r) = \mathbf{1}_{\{r \geq t\}} + \mathbf{1}_{\{r < t\}} \mathbb{P}(T_1 \leq r) = (1 - e^{-\lambda r} \mathbf{1}_{\{r < t\}}).$$

Comme les fonctions de répartition 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 :

$$\begin{aligned}
 \mathbb{E}[R_t] &= \mathbb{E}[\min(T_1, t)] = \int \min(w, t) \lambda e^{-\lambda w} \mathbf{1}_{\{w > 0\}} dw \\
 &= \int_0^t w \lambda e^{-\lambda w} dw + \int_t^{\infty} t \lambda e^{-\lambda w} dw \\
 &= \frac{1}{\lambda} (1 - e^{-\lambda t}).
 \end{aligned}$$

On en déduit donc que $\mathbb{E}[S_t + R_t] = \frac{2}{\lambda} - \frac{1}{\lambda} e^{-\lambda t}$. 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 $\mathbb{P}(R_t \leq r, S_t \leq s) = \mathbb{P}(R_t \leq r)\mathbb{P}(S_t \leq s)$ pour tout $r, s \in \mathbb{R}$. On en déduit que les variables R_t et S_t sont indépendantes.
2. On remarque que U_t a même loi que $W_t = \min(T_1, t) + T_2$. Comme T_1 est une variable aléatoire finie p.s., on en déduit que $\lim_{t \rightarrow \infty} W_t = T_1 + T_2$ 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 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 :

$$\begin{aligned}
 \mathbb{E}[g(U_t)] &= \mathbb{E}[g(R_t + S_t)] = \mathbb{E}[g(\min(T_1, t) + T_2)] \\
 &= \int g(\min(t_1, t) + t_2) \lambda^2 e^{-\lambda(t_1+t_2)} \mathbf{1}_{\{0 < t_1, 0 < t_2\}} dt_1 dt_2 \\
 &= e^{-\lambda t} \int g(t + t_2) \lambda e^{-\lambda t_2} \mathbf{1}_{\{0 < t_2\}} dt_2 \\
 &\quad + \int g(t_1 + t_2) \lambda^2 e^{-\lambda(t_1+t_2)} \mathbf{1}_{\{0 < t_1 < t, 0 < t_2\}} dt_1 dt_2 \\
 &= \int g(u) \lambda e^{-\lambda u} \mathbf{1}_{\{t < u\}} du \\
 &\quad + \int g(u) \lambda^2 e^{-\lambda u} \mathbf{1}_{\{t_2 < u < t+t_2, 0 < t_2\}} du dt_2 \\
 &= \int g(u) \lambda e^{-\lambda u} (\mathbf{1}_{\{t < u\}} + \lambda(u - (u - t)_+) \mathbf{1}_{\{0 < u\}}) du \\
 &= \int g(u) \lambda e^{-\lambda u} (\lambda u \mathbf{1}_{\{0 < u \leq t\}} + (\lambda t + 1) \mathbf{1}_{\{t < u\}}) du.
 \end{aligned}$$

Donc la loi de la variable aléatoire U_t est de densité $\lambda e^{-\lambda u} [\lambda u \mathbf{1}_{\{0 < u \leq t\}} + (\lambda t + 1) \mathbf{1}_{\{t < u\}}]$.



Exercice XI.6. I Le temps d'attente à l'arrêt de bus

1. La variable aléatoire aU est de loi uniforme sur $[0, a]$. Il est donc naturel de choisir US_n , avec U indépendante de S_n , pour modéliser une variable aléatoire uniforme sur $[0, S_n]$.
2. Pour toute permutation σ de $\{1, \dots, n\}$, la variable aléatoire $(T_{\sigma(i)}, 1 \leq i \leq n)$ a même loi que $(T_i, 1 \leq i \leq n)$. Comme $S_n = \sum_{i=1}^n T_{\sigma(i)}$, on en déduit que $T_{\sigma(1)}/S_n$ a même loi que X_1/S_n .
3. On en déduit que :

$$\mathbb{E} \left[\frac{nT_1}{S_n} \right] = \sum_{k=1}^n \mathbb{E} \left[\frac{T_1}{S_n} \right] = \sum_{k=1}^n \mathbb{E} \left[\frac{T_k}{S_n} \right] = 1.$$

4. Par indépendance (U, Z) est une variable continue de densité $\mathbf{1}_{[0,1]}(u)f_Z(z)$, où f_Z est la densité de la loi de Z . Par Fubini, on a :

$$\begin{aligned} \mathbb{E}[\varphi(U, Z)] &= \int \varphi(u, z) \mathbf{1}_{[0,1]}(u) f_Z(z) \, dudz = \int \left(\int_0^1 \varphi(u, z) \, du \right) f_Z(z) dz \\ &= \mathbb{E} \left[\int_0^1 du \varphi(u, Z) \right]. \end{aligned}$$

Pour $1 \leq k \leq n$, on a $\{N_n^* = k\} = \{U \in]S_{k-1}/S_n, S_k/S_n]\}$. On en déduit d'après (XI.3) que :

$$\mathbb{P}(N_n^* = k) = \mathbb{E} [\mathbf{1}_{\{U \in]S_{k-1}/S_n, S_k/S_n]\}}] = \mathbb{E} \left[\frac{S_k}{S_n} - \frac{S_{k-1}}{S_n} \right] = \mathbb{E} \left[\frac{T_k}{S_n} \right] = \frac{1}{n}.$$

On observe que la loi de N_n^* est la loi uniforme sur $\{1, \dots, n\}$.

5. On a :

$$\begin{aligned} \mathbb{E}[g(T_n^*)] &= \sum_{k=1}^n \mathbb{E} [\mathbf{1}_{\{N_n^*=k\}} g(T_k)] = \sum_{k=1}^n \mathbb{E} [\mathbf{1}_{\{U \in]S_k/S_n, S_{k+1}/S_n]\}} g(T_k)] \\ &= \sum_{k=1}^n \mathbb{E} \left[\frac{T_k}{S_n} g(T_k) \right] \\ &= \mathbb{E} \left[\frac{nT_1}{S_n} g(T_1) \right], \end{aligned}$$

où l'on a utilisé (XI.3) pour la troisième égalité et le fait que $(T_k, T_k/S_n)$ a même loi que $(T_1, T_1/S_n)$ pour la dernière égalité.

6. Par la loi forte des grands nombres, on a p.s. $\lim_{n \rightarrow \infty} S_n/n = \mu$. Comme $T > 0$ p.s., on en déduit $\mu > 0$. La fonction $x \mapsto T_1/x$ est continue en μ . On en déduit que p.s. $\lim_{n \rightarrow \infty} nT_1/S_n = X$ avec $X = T_1/\mu$. On a $\mathbb{E}[X] = 1$.
7. On a $|\varphi_+(x) - \varphi_+(y)| \leq |x - y|$ pour tout $x, y \in \mathbb{R}$. La fonction φ_+ est donc continue. Si $x \geq y \geq 0$, on a $\varphi_+(x - y) = x - y \leq x$. Si $y \geq x \geq 0$, on a $\varphi_+(x - y) = 0 \leq x$. Donc, pour tout $x \geq 0, y \geq 0$, on a $\varphi_+(x - y) \leq x$. On considère $Y_n = \varphi_+(X - X_n)$. D'après la continuité de φ_+ , on a p.s. $\lim_{n \rightarrow \infty} Y_n = 0$. De plus $|Y_n| \leq X$, et X est intégrable. On déduit du théorème de convergence dominée que $\lim_{n \rightarrow \infty} \mathbb{E}[Y_n] = \mathbb{E}[\lim_{n \rightarrow \infty} Y_n] = 0$.
8. L'égalité $|x| = -x + 2\varphi_+(x)$ est évidente. On a :

$$\mathbb{E}[|X - X_n|] = \mathbb{E}[-X + X_n + 2\varphi_+(X - X_n)] = 2\mathbb{E}[\varphi_+(X - X_n)].$$

On en déduit donc que $\lim_{n \rightarrow \infty} \mathbb{E}[|X - X_n|] = 0$.

9. On pose $c = \sup\{|g(x)|, x \in \mathbb{R}\}$. On a $\mathbb{E}[g(T_n^*)] = \mathbb{E}[Xg(T_1)] + \mathbb{E}[(X_n - X)g(T_1)]$. On remarque que :

$$|\mathbb{E}[(X_n - X)g(T_1)]| \leq \mathbb{E}[|X_n - X| |g(T_1)|] \leq c\mathbb{E}[|X_n - X|].$$

Ceci assure que $\lim_{n \rightarrow \infty} \mathbb{E}[g(T_n^*)] = \mathbb{E}[Xg(T_1)] = \frac{1}{\mu} \mathbb{E}[Tg(T)]$.

10. On choisit $g(x) = e^{iux}$. D'après la question précédente, on a $\lim_{n \rightarrow \infty} \psi_{T_n^*}(u) = \frac{1}{\mu} \mathbb{E}[T e^{iuT}]$. Le théorème de convergence dominée assure que l'application $u \mapsto \frac{1}{\mu} \mathbb{E}[T e^{iuT}]$ est continue en 0. Cela implique que la suite $(T_n^*, n \geq 1)$ converge en loi vers une limite, T^* , de fonction caractéristique $\psi_{T^*}(u) = \frac{1}{\mu} \mathbb{E}[T e^{iuT}]$.

II Loi biaisée par la taille

1. On a $\mathbb{E}[T^*] = \frac{1}{\mu} \mathbb{E}[T^2] = \frac{\text{Var}(T) + \mathbb{E}[T]^2}{\mu} = \mu + \frac{\sigma^2}{\mu} \geq \mu$. L'égalité a lieu si et seulement si $\sigma^2 = 0$, c'est-à-dire que p.s. T est égal à une constante.
2. On a $\mathbb{E}[g(T^*)] = \frac{1}{\mu} \mathbb{E}[Tg(T)] = \frac{1}{\mu} \int g(t) t f(t) dt$ pour toute fonction g bornée mesurable. Ceci implique que T^* est une variable aléatoire continue de densité $f^*(t) = \frac{t}{\mu} f(t)$.
3. On a $\mu = 1/\lambda$, et T^* a pour densité $\lambda^2 t e^{-\lambda t} \mathbf{1}_{\{t > 0\}}$. On reconnaît la densité de la loi $\Gamma(\lambda, 2)$. En utilisant les fonctions caractéristiques, on a que la loi de $T + T'$ est la loi $\Gamma(\lambda, 2)$. Donc T^* a même loi que $T + T'$. Il vient $\mathbb{E}[T^*] = \mathbb{E}[T + T'] = 2\mathbb{E}[T]$.

4. Soit $a > a'$ dans $[0, \mu]$. Pour $x \in]a', a]$, on a $(x - \mu) \leq 0$. Il vient :

$$G(a) - G(a') = \mathbb{E}[(T - \mu)\mathbf{1}_{]a', a]}(T)] \leq 0.$$

Ceci assure que G est décroissante sur $[0, \mu]$. Un raisonnement similaire assure que G est croissante sur $[\mu, \infty[$. On en déduit que G est majorée par $\max(G(0), \lim_{a \rightarrow \infty} G(a))$. Comme cette dernière quantité est nulle, on en déduit que $G(a) \leq 0$. Comme :

$$G(a) = \mathbb{E}[T\mathbf{1}_{\{T \leq a\}}] - \mu\mathbb{E}[\mathbf{1}_{\{T \leq a\}}] = \mu(\mathbb{P}(T^* \leq a) - \mathbb{P}(T \leq a)),$$

on en déduit que $\mathbb{P}(T^* \leq a) \leq \mathbb{P}(T \leq a)$ pour tout $a \in \mathbb{R}$.

5. Soit $z \in \mathbb{R}$, on a :

$$\mathbb{P}(F_Z^{-1}(U) \leq z) = \mathbb{P}(U \leq F_Z(z)) = F_Z(z),$$

car $F_Z(z) \in [0, 1]$. On en déduit que $F_Z^{-1}(U)$ et Z ont même fonction de répartition et donc même loi.

6. Soit F (resp. F_*) la fonction de répartition de T (resp. T^*). Soit U une variable aléatoire uniforme sur $[0, 1]$. La variable $\tilde{T} = F^{-1}(U)$ (resp. $\tilde{T}^* = F_*^{-1}(U)$) a même loi que T (resp. T^*). On a vérifié que $F(x) \geq F_*(x)$ pour tout $x \in \mathbb{R}$. Ceci assure que pour tout $u \in]0, 1[$, on a $F^{-1}(u) \leq F_*^{-1}(u)$, et donc p.s. $\tilde{T} \leq \tilde{T}^*$. ▲

Exercice XI.7. 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 $p = \mathbb{P}(\mathbf{1}_{\{Y \leq X\}} = 1) = \mathbb{E}[\mathbf{1}_{\{Y \leq X\}}] = \mathbb{P}(Y \leq X)$. Sa variance est $\text{Var}(\mathbf{1}_{\{Y \leq X\}}) = p(1 - p)$.
2. On a :

$$\begin{aligned} p = \mathbb{E}[\mathbf{1}_{\{Y \leq X\}}] &= \int \mathbf{1}_{\{y \leq x\}} f(x)g(y) \, dx dy = \int \left(\int_{-\infty}^x g(y) \, dy \right) f(x) \, dx \\ &= \int G(x)f(x) \, dx. \end{aligned}$$

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 :

$$p = \int F(x)f(x) \, dx = [F(x)^2/2]_{-\infty}^{+\infty} = 1/2.$$

3. Comme X et Y sont indépendants, la densité conditionnelle de Y sachant X est la densité de Y . On a donc :

$$\mathbb{E} [\mathbf{1}_{\{Y \leq X\}} | X = x] = \int \mathbf{1}_{\{y \leq x\}} g(y) dy = [G(y)]_{-\infty}^x = G(x).$$

Cela implique $S = \mathbb{E} [\mathbf{1}_{\{Y \leq X\}} | X] = G(X)$. On a :

$$\mathbb{E}[S] = \mathbb{E} [\mathbb{E} [\mathbf{1}_{\{Y \leq X\}} | X]] = \mathbb{E} [\mathbf{1}_{\{Y \leq X\}}] = p.$$

Des calculs similaires donnent $T = 1 - F(Y)$ et $\mathbb{E}[T] = p$.

4. On a :

$$\alpha = \text{Var}(S) = \mathbb{E}[G(X)^2] - p^2 = \int G(x)^2 f(x) dx - p^2.$$

De façon similaire, on obtient $\beta = \text{Var}(T) = \int (1 - F(y))^2 g(y) dy - p^2$.

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 ont même loi. En particulier, on a $\beta = \alpha = 1/12$.

6. On a :

$$\begin{aligned} \mathbb{E} [S \mathbf{1}_{\{Y \leq X\}}] &= \mathbb{E} [G(X) \mathbf{1}_{\{Y \leq X\}}] = \int G(x) \mathbf{1}_{\{y \leq x\}} g(y) f(x) dx dy \\ &= \int G(x)^2 f(x) dx. \end{aligned}$$

On en déduit donc que :

$$\text{Cov}(S, \mathbf{1}_{\{Y \leq X\}}) = \int G(x)^2 f(x) dx - \mathbb{E}[S] \mathbb{E} [\mathbf{1}_{\{Y \leq X\}}] = \alpha.$$

Des calculs similaires donnent $\text{Cov}(T, \mathbf{1}_{\{Y \leq X\}}) = \beta$.

Vérifions par l'absurde que $(\alpha, \beta) \neq (0, 0)$ si $p \in]0, 1[$. Si $\alpha = 0$, alors $\text{Var}(S) = 0$ 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 p.s. $Y \notin [a, b]$. 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 on a p.s. $F(Y) = 1$ ou p.s. $F(Y) = 0$. 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 $p \in]0, 1[$.

II Étude de la projection de Hájek de la statistique de Mann et Withney

1. Par indépendance, pour $\ell \neq i$, on a :

$$\mathbb{E} \left[\mathbf{1}_{\{Y_j \leq X_\ell\}} | X_i \right] = \mathbb{E} \left[\mathbf{1}_{\{Y_j \leq X_\ell\}} \right] = p.$$

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 :

$$\begin{aligned} \text{Var}(H_{m,n}) &= \text{Var} \left(\sum_{i=1}^m n(S_i - p) + \sum_{j=1}^n m(T_j - p) \right) \\ &= \sum_{i=1}^m \text{Var}(n(S_i - p)) + \sum_{j=1}^n \text{Var}(m(T_j - p)) \\ &= \sum_{i=1}^m n^2 \text{Var}(S_i) + \sum_{j=1}^n m^2 \text{Var}(T_j) \\ &= mn^2\alpha + nm^2\beta. \end{aligned}$$

3. Les variables aléatoires $((S_i - p), i \geq 1)$ sont indépendantes, de même loi, et de carré intégrable. On remarque que $\mathbb{E}[S_i - p] = 0$ et $\text{Var}(S_i - p) = \alpha$. On déduit du théorème central limite, que la suite $(V_m, m \geq 1)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \alpha)$. Un raisonnement similaire assure que la suite $(W_n, n \geq 1)$ converge en loi vers la loi gaussienne $\mathcal{N}(0, \beta)$.

4. On a $\psi_{V_m}(u) = e^{-\alpha u^2/2} + R_m^{(1)}(u)$ et $\psi_{W_n}(u) = e^{-\beta u^2/2} + R_n^{(2)}(u)$, avec :

$$\lim_{m \rightarrow \infty} \sup_{|u| \leq K} |R_m^{(1)}(u)| = \lim_{n \rightarrow \infty} \sup_{|u| \leq K} |R_n^{(2)}(u)| = 0. \quad (\text{XIII.7})$$

On pose $Z_{m,n} = \frac{H_{m,n}}{\sqrt{\text{Var}(H_{m,n})}}$. On remarque que $Z_{m,n} = a_{m,n}V_m + b_{m,n}W_n$ avec :

$$a_{m,n} = \frac{n\sqrt{m}}{\sqrt{mn^2\alpha + mn^2\beta}} \quad \text{et} \quad b_{m,n} = \frac{m\sqrt{n}}{\sqrt{mn^2\alpha + mn^2\beta}}.$$

En utilisant l'indépendance entre V_m et W_n pour la première égalité, il vient

$$\begin{aligned} \psi_{Z_{m,n}}(u) &= \psi_{V_m}(a_{m,n}u)\psi_{W_n}(b_{m,n}u) = e^{-\frac{(\alpha a_{m,n}^2 + \beta b_{m,n}^2)u^2}{2}} + R_{m,n}^{(3)}(u) \\ &= e^{-u^2/2} + R_{m,n}^{(3)}(u), \end{aligned}$$

où $R_{m,n}^{(3)}(u)$ est égal à :

$$e^{-\alpha a_{m,n}^2 \frac{u^2}{2}} R_n^{(2)}(b_{m,n}u) + e^{-\beta b_{m,n}^2 \frac{u^2}{2}} R_m^{(1)}(a_{m,n}u) + R_m^{(1)}(a_{m,n}u)R_n^{(2)}(b_{m,n}u).$$

On a $a_{m,n} \leq 1/\alpha$ et $b_{m,n} \leq 1/\beta$. On déduit de (XIII.7) que :

$$\lim_{\min(m,n) \rightarrow \infty} \sup_{|u| \leq K} |R_{m,n}^{(3)}(u)| = 0.$$

Ceci implique que $\lim_{\min(m,n) \rightarrow \infty} \psi_{Z_{m,n}}(u) = e^{-u^2/2}$. Quand $\min(m, n)$ tend vers l'infini, la suite $(Z_{m,n}, m \geq 1, n \geq 1)$ converge en loi vers Z de loi gaussienne centrée réduite $\mathcal{N}(0, 1)$.

5. On pose $c_{m,n} = \sqrt{\text{Var}(H_{m,n})/\text{Var}(U_{m,n})}$. Comme $\lim_{\min(m,n) \rightarrow \infty} c_{m,n} = 1$, on déduit du théorème de Slutsky que la suite $((c_{m,n}, Z_{m,n}), m \geq 1, n \geq 1)$ 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 $(c_{m,n}Z_{m,n}, m \geq 1, n \geq 1)$ vers Z quand $\min(m, n)$ tend vers l'infini. Donc la suite $\left(H_{m,n}/\sqrt{\text{Var}(U_{m,n})}, m \geq 1, n \geq 1\right)$ converge en loi vers la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$ quand $\min(m, n)$ tend vers l'infini.

III Convergence de la statistique de Mann et Whitney

1. On a :

$$\begin{aligned}
 \text{Cov}(H_{m,n}, U_{m,n}^*) &= \text{Cov}(H_{m,n}, U_{m,n}) \\
 &= \text{Cov}\left(n \sum_{i=1}^m S_i, U_{m,n}\right) + \text{Cov}\left(m \sum_{j=1}^n T_j, U_{m,n}\right) \\
 &= n \sum_{i=1}^m \sum_{j=1}^n \text{Cov}(S_i, \mathbf{1}_{\{Y_j \leq X_i\}}) + m \sum_{i=1}^m \sum_{j=1}^n \text{Cov}(T_j, \mathbf{1}_{\{Y_j \leq X_i\}}) \\
 &= mn^2 \text{Cov}(S, \mathbf{1}_{\{Y \leq X\}}) + nm^2 \text{Cov}(T, \mathbf{1}_{\{Y \leq X\}}).
 \end{aligned}$$

2. On déduit de ce qui précède que :

$$\begin{aligned}
 \text{Var}(H_{m,n} - U_{m,n}^*) &= \text{Var}(H_{m,n}) + \text{Var}(U_{m,n}^*) - 2 \text{Cov}(H_{m,n}, U_{m,n}^*) \\
 &= mn^2\alpha + nm^2\beta + mn^2\alpha \\
 &\quad + nm^2\beta + mn(p - p^2 - \alpha - \beta) - 2mn^2\alpha - 2nm^2\beta \\
 &= mn(p - p^2 - \alpha - \beta).
 \end{aligned}$$

3. En particulier, on a :

$$\lim_{\min(m,n) \rightarrow \infty} \frac{\text{Var}(H_{m,n} - U_{m,n}^*)}{\text{Var}(U_{m,n})} = 0.$$

4. On pose $Q_{m,n} = \frac{H_{m,n} - U_{m,n}^*}{\sqrt{\text{Var}(U_{m,n})}}$. En utilisant l'inégalité de Tchebychev, on a pour tout $\varepsilon > 0$:

$$\mathbb{P}(|Q_{m,n}| > \varepsilon) \leq \mathbb{E}[Q_{m,n}^2]/\varepsilon^2.$$

Comme $\lim_{\min(m,n) \rightarrow \infty} \mathbb{E}[Q_{m,n}^2] = 0$, on déduit que la suite $(Q_{m,n}, m \geq 1, n \geq 1)$ converge en probabilité vers 0 quand $\min(m, n)$ tend vers l'infini.

5. On a :

$$\frac{U_{m,n} - mnp}{\sqrt{\text{Var}(U_{m,n})}} = c_{m,n}Z_{m,n} - Q_{m,n}.$$

On déduit du théorème de Slutsky que la suite $((c_{m,n}Z_{m,n}, -Q_{m,n}), m \geq 1, n \geq 1)$ converge en loi vers $(Z, 0)$ quand $\min(m, n)$ tend vers l'infini. Par continuité de l'addition, on en déduit que la suite $(c_{m,n}Z_{m,n} - Q_{m,n}, m \geq 1, n \geq 1)$ converge en loi vers Z quand $\min(m, n)$ tend vers l'infini. On a donc montré que, quand $\min(m, n)$ tend vers l'infini, la suite :

$$\left(\frac{U_{m,n} - mnp}{\sqrt{\text{Var}(U_{m,n})}}, m \geq 1, n \geq 1 \right)$$

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 $\text{Card}(\Delta_1) = mn$, $\text{Card}(\Delta_2) = m(m-1)n$, $\text{Card}(\Delta_3) = n(n-1)m$ et $\text{Card}(\Delta_4) = m(m-1)n(n-1)$.

2. On a :

$$\begin{aligned} \text{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y' \leq X\}}) &= \mathbb{E}[\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y' \leq X\}}] - \mathbb{E}[\mathbf{1}_{\{Y \leq X\}}] \mathbb{E}[\mathbf{1}_{\{Y' \leq X\}}] \\ &= \int \mathbf{1}_{\{y \leq x\}}, \mathbf{1}_{\{y' \leq x\}} g(y)g(y')f(x) dydy'dx - p^2 \\ &= \int G(x)^2 f(x) dx - p^2 = \alpha. \end{aligned}$$

Un calcul similaire donne $\text{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y \leq X'\}}) = \beta$.

3. On pose $A_{i,i',j,j'} = \mathbf{1}_{\{Y_j \leq X_i\}} - p)(\mathbf{1}_{\{Y_{j'} \leq X_{i'}\}} - p$. On a :

$$\text{Var}(U_{m,n}) = \mathbb{E}[(U_{m,n}^*)^2] = \sum_{(i,i',j,j') \in \Delta} \mathbb{E}[A_{i,i',j,j'}].$$

On décompose la somme sur les ensembles d'indices $\Delta_1, \Delta_2, \Delta_3$ et Δ_4 . Il vient :

$$\begin{aligned} \sum_{(i,i',j,j') \in \Delta_1} \mathbb{E}[A_{i,i',j,j'}] &= \text{Card}(\Delta_1) \mathbb{E}[(\mathbf{1}_{\{Y \leq X\}} - p)^2] \\ &= \text{Card}(\Delta_1) \text{Var}(\mathbf{1}_{\{Y \leq X\}}) \\ &= mnp(1-p), \\ \sum_{(i,i',j,j') \in \Delta_2} \mathbb{E}[A_{i,i',j,j'}] &= \text{Card}(\Delta_2) \mathbb{E}[(\mathbf{1}_{\{Y \leq X\}} - p)(\mathbf{1}_{\{Y \leq X'\}} - p)] \\ &= \text{Card}(\Delta_2) \text{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y \leq X'\}}) \\ &= m(m-1)n\beta, \\ \sum_{(i,i',j,j') \in \Delta_3} \mathbb{E}[A_{i,i',j,j'}] &= \text{Card}(\Delta_3) \mathbb{E}[(\mathbf{1}_{\{Y \leq X\}} - p)(\mathbf{1}_{\{Y' \leq X\}} - p)] \\ &= \text{Card}(\Delta_3) \text{Cov}(\mathbf{1}_{\{Y \leq X\}}, \mathbf{1}_{\{Y' \leq X\}}) \\ &= n(n-1)m\alpha, \\ \sum_{(i,i',j,j') \in \Delta_4} \mathbb{E}[A_{i,i',j,j'}] &= \text{Card}(\Delta_4) \mathbb{E}[(\mathbf{1}_{\{Y \leq X\}} - p)(\mathbf{1}_{\{Y' \leq X'\}} - p)] = 0. \end{aligned}$$

La dernière égalité s'obtient en utilisant l'indépendance entre (X, Y) et (X', Y') .

On obtient donc :

$$\text{Var}(U_{m,n}) = mn^2\alpha + m^2n\beta + mn(p - p^2 - \alpha - \beta).$$

4. Dans le cas où les variables aléatoires $(X_i, i \geq 1)$ et $(Y_j, j \geq 1)$ ont même loi, on a :

$$\text{Var}(U_{m,n}) = \frac{mn(m+n+1)}{12}.$$

▲

Exercice XI.8. I Calcul de la probabilité d'extinction

1. On a $\{Z_n = 0\} \subset \{Z_{n+1} = 0\}$ pour tout $n \geq 0$. Ceci implique que $\{Z_n = 0\} = \cup_{k=1}^n \{Z_k = 0\}$. On en déduit que $\{\text{il existe } n \geq 0 \text{ tel que } Z_n = 0\} = \cup_{k \geq 1} \{Z_k = 0\}$ est la limite croissante de la suite d'évènements $(\{Z_n = 0\}, n \geq 0)$. On déduit donc du théorème de convergence monotone que η est la limite croissante de la suite $(\mathbb{P}(Z_n = 0), n \geq 0)$.
2. On a pour $k \geq 1$:

$$\begin{aligned} \mathbb{E}[Z_{n+1}|Z_n = k] &= \mathbb{E}\left[\sum_{i=1}^{Z_n} \xi_{i,n} | Z_n = k\right] = \mathbb{E}\left[\sum_{i=1}^k \xi_{i,n} | Z_n = k\right] = \mathbb{E}\left[\sum_{i=1}^k \xi_{i,n}\right] \\ &= km, \end{aligned}$$

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 $\mathbb{E}[Z_{n+1}|Z_n] = mZ_n$, et que $\mathbb{E}[Z_{n+1}] = m\mathbb{E}[Z_n]$. En itérant, on obtient que $\mathbb{E}[Z_n] = m^n$.

3. On remarque que $\mathbb{P}(Z_n > 0) \leq \mathbb{E}[Z_n]$. Si $m < 1$, alors on a $\lim_{n \rightarrow \infty} \mathbb{P}(Z_n > 0) = 0$, et donc $\eta = 1 - \lim_{n \rightarrow \infty} \mathbb{P}(Z_n > 0) = 1$.
4. On a pour $k \geq 1$:

$$\begin{aligned} \mathbb{E}[z^{Z_{n+1}}|Z_n = k] &= \mathbb{E}\left[z^{\sum_{i=1}^{Z_n} \xi_{i,n}} | Z_n = k\right] = \mathbb{E}\left[z^{\sum_{i=1}^k \xi_{i,n}} | Z_n = k\right] \\ &= \mathbb{E}\left[z^{\sum_{i=1}^k \xi_{i,n}}\right] \\ &= \phi(z)^k, \end{aligned}$$

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 tend vers l'infini, que η est solution de l'équation (XI.6).

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 $\phi''(z) > 0$. 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 (XI.6) 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. On remarque que $\mathbb{P}(Z_0 = 0) = 0 \leq x_0$. On suppose 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 $\eta = \lim_{n \rightarrow \infty} \mathbb{P}(Z_n = 0) \leq x_0$. Comme $\eta \in \{x_0, 1\}$, on en déduit que $\eta = x_0$.

II Comportement asymptotique sur un exemple

1. On obtient :

$$\phi(z) = \alpha + \frac{(1 - \alpha)(1 - \beta)z}{1 - \beta z}. \tag{XIII.8}$$

On a $m = \phi'(1) = \frac{1 - \alpha}{1 - \beta} > 1$. 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 :

$$\psi_{XY}(u) = \mathbb{E}[e^{iuXY}] = \mathbb{E}[\mathbf{1}_{\{X=0\}}] + \mathbb{E}[\mathbf{1}_{\{X=1\}} e^{iuY}] = (1-p) + p \frac{\lambda}{\lambda - iu}.$$

4. La fonction caractéristique, ψ , de ξ se déduit de (XIII.8), en remarquant que $\psi(u) = \phi(e^{iu})$ pour $u \in \mathbb{R}$. En reproduisant les calculs qui permettent de déterminer ϕ_n , on obtient $\psi_{Z_n}(u) = \phi_n(e^{iu})$. Il vient :

$$\begin{aligned} \psi_{m^{-n}Z_n}(u) &= \psi_{Z_n}(m^{-n}u) \\ &= \phi_n(e^{im^{-n}u}) \\ &= \frac{1 - \eta - m^n \eta(1 + im^{-n}u) + m^n \eta + o(1)}{1 - \eta - m^n(1 + im^{-n}u) + m^n + o(1)} \\ &= \frac{1 - \eta - iu\eta + o(1)}{1 - \eta - iu + o(1)}. \end{aligned}$$

On en déduit donc que :

$$\lim_{n \rightarrow \infty} \psi_{m^{-n}Z_n}(u) = \frac{1 - \eta - iu\eta}{1 - \eta - iu} = \eta + (1 - \eta) \frac{1 - \eta}{1 - \eta - iu}.$$

On en déduit 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.9. 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 $\lim_{n \rightarrow \infty} p_n(k) = p(k)$ 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 $\lim_{n \rightarrow \infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$. 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 :

$$\lim_{n \rightarrow \infty} \mathbb{E}[G(X_n)] = \mathbb{E}[G(X)] = \sum_{k=0}^{\infty} \mathbb{E}[g_k(X)] = \sum_{k=0}^{\infty} p(k),$$

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 $\binom{n+r-1}{n-1} = \frac{(n+r-1)!}{r!(n-1)!}$ tirages possibles. Les configurations étant équiprobables, on en déduit que la probabilité d’obtenir une configuration donnée est $\frac{r!(n-1)!}{(n+r-1)!}$.
2. Si $X_{n,r}^{(1)} = k$, alors il faut répartir $r - k$ particules indiscernables dans $n - 1$ boîtes. Il existe donc $\binom{n+r-2-k}{n-2} = \frac{(n+r-2-k)!}{(r-k)!(n-2)!}$ configurations possibles. On en déduit donc que pour $k \in \{0, \dots, r\}$:

$$\begin{aligned} \mathbb{P}(X_{n,r}^{(1)} = k) &= (n-1) \frac{r!(n+r-2-k)!}{(r-k)!(n+r-1)!} \\ &= \frac{(n-1)}{n+r-1} \frac{r!}{(r-k)!} \frac{(n+r-2-k)!}{(n+r-2)!}. \end{aligned}$$

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 :

$$p(0) = \lim_{n \rightarrow \infty} p_{n,r}(0) = \lim_{n \rightarrow \infty} \frac{1 - \frac{1}{n}}{1 + \frac{r}{n} - \frac{1}{n}} = \frac{1}{1 + \theta},$$

et pour $k \geq 1$,

$$\begin{aligned} p(k) &= \lim_{n \rightarrow \infty} p_{n,r}(k) = \lim_{n \rightarrow \infty} \frac{1 - \frac{1}{n}}{1 + \frac{r}{n} - \frac{1}{n}} \frac{\frac{r}{n}}{1 + \frac{r}{n} - \frac{2}{n}} \cdots \frac{\frac{r}{n} - \frac{k-1}{n}}{1 + \frac{r}{n} - \frac{2}{n} - \frac{k-1}{n}} \\ &= \frac{\theta^k}{(1 + \theta)^{k+1}}. \end{aligned}$$

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 $k \in \mathbb{N}$, $\mathbb{P}(X = k) = \frac{\theta^k}{(1 + \theta)^{k+1}}$. 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 en déduit :

$$r = \mathbb{E}\left[\sum_{i=1}^n X_{n,r}^{(i)}\right] = \sum_{i=1}^n \mathbb{E}[X_{n,r}^{(i)}] = n\mathbb{E}[X_{n,r}^{(1)}].$$

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 $\mathbb{E}[X] = \mathbb{E}[Y - 1] = \frac{1}{\rho} - 1 = \theta$.
6. On a pour $k \in \{0, \dots, r - 1\}$:

$$\mathbb{P}(X_{n+1,r-1}^{(1)} = k) = n \frac{(r-1)!(n+r-2-k)!}{(r-1-k)!(n+r-1)!} = \frac{n}{n-1} \frac{r-k}{r} \mathbb{P}(X_{n,r}^{(1)} = k),$$

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 $\mathbb{E}[X_{n,r}^{(1)}] = r - r \frac{n-1}{n} = \frac{r}{n}$.

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 $\mathbb{E}[X^2] = \mathbb{E}[Y^2] - 2\mathbb{E}[Y] + 1 = \text{Var}(Y) + \mathbb{E}[Y]^2 - 2\mathbb{E}[Y] + 1$.
 Comme $\mathbb{E}[Y] = \frac{1}{\rho} = 1 + \theta$ et $\text{Var}(Y) = \frac{1-\rho}{\rho^2} = \theta(1 + \theta)$, 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 $a = (a_1, \dots, a_n) \in \mathbb{N}^n$, $b = (b_1, \dots, b_n) \in \mathbb{N}^n$ 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 $j \in \{1, \dots, n\}$ 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 :

$$\begin{aligned} \mathbb{P}(X_{n,r+1} = b) &= \sum_{\substack{a=(a_1, \dots, a_n) \in \mathbb{N}^n, \\ \sum_{i=1}^n a_i = r}} \mathbb{P}(X_{n,r+1} = b | X_{n,r} = a) \mathbb{P}(X_{n,r} = a) \\ &= \frac{r!(n-1)!}{(n+r-1)!} \sum_{\substack{a=(a_1, \dots, a_n) \in \mathbb{N}^n, \\ \sum_{i=1}^n a_i = r}} \mathbb{P}(X_{n,r+1} = b | X_{n,r} = a) \\ &= \frac{r!(n-1)!}{(n+r-1)!} \sum_{j=1}^n \frac{b_j}{n+r} \\ &= \frac{r!(n-1)!}{(n+r-1)!} \frac{r+1}{n+r} \\ &= \frac{(r+1)!(n-1)!}{(n+r)!}. \end{aligned}$$

La loi de $X_{n,r+1}$ est la loi de Bose-Einstein pour $r + 1$ particules et n boîtes.

▲

Exercice XI.10. I Sondage avec remise

1. La loi de $n\bar{Y}_n$ est la loi binomiale de paramètre (n, p) . On a $\mathbb{E}[\bar{Y}_n] = p$ et $\text{Var}(Y_n) = p(1-p)/n$. Les variables aléatoires $(Y_n, n \geq 1)$ sont indépendantes de même loi et bornées (donc en particulier intégrables et de carré intégrable). La loi forte des grands nombres assure que $(\bar{Y}_n, n \geq 1)$ converge p.s. vers p . On a $\text{Var}(Y_1) = p(1-p)$. Le théorème central limite assure que $(\sqrt{n}(\bar{Y}_n - p), n \geq 1)$ converge en loi vers $\sqrt{p(1-p)}G$, où G est de loi $\mathcal{N}(0, 1)$. Le théorème de Slutsky implique que $((\sqrt{n}(\bar{Y}_n - p), \bar{Y}_n), n \geq 1)$ converge en loi vers $(\sqrt{p(1-p)}G, p)$. La continuité de la fonction $(x, y) \mapsto x/\sqrt{y(1-y)}$ sur $\mathbb{R} \times]0, 1[$ implique que $(\sqrt{n}(\bar{Y}_n - p)/\sqrt{\bar{Y}_n(1-\bar{Y}_n)}, n \geq 1)$ converge en loi vers G (on a utilisé le fait que $0 < p < 1$).
2. Soit $I_n = [\bar{Y}_n \pm a_\alpha \sqrt{\bar{Y}_n(1-\bar{Y}_n)}/\sqrt{n}]$, où a_α est le quantile de la loi $\mathcal{N}(0, 1)$ d'ordre $1 - \alpha/2$. Comme $(\sqrt{n}(\bar{Y}_n - p)/\sqrt{\bar{Y}_n(1-\bar{Y}_n)}, n \geq 1)$ converge en loi vers G et que G est une variable continue, on en déduit que :

$$\begin{aligned} \lim_{n \rightarrow \infty} \mathbb{P}(p \in I_n) &= \lim_{n \rightarrow \infty} \mathbb{P} \left(\sqrt{n} \frac{\bar{Y}_n - p}{\sqrt{\bar{Y}_n(1-\bar{Y}_n)}} \in [-a_\alpha, a_\alpha] \right) \\ &= \mathbb{P}(G \in [-a_\alpha, a_\alpha]) = 1 - \alpha. \end{aligned}$$

Ainsi I_n est un intervalle de confiance sur p de niveau asymptotique $1 - \alpha$. Comme $\bar{Y}_n(1-\bar{Y}_n) \leq 1/4$ car $\bar{Y}_n \in [0, 1]$, on en déduit que la demi largeur de I_n est plus petite que $a_\alpha/2\sqrt{n}$. Pour $\alpha = 5\%$ et $n = 1000$, on trouve $a_\alpha \simeq 1.96$, et $\frac{a_\alpha}{2\sqrt{n}} \simeq 0.031$. La précision du sondage est d'environ 3 points.

II Sondage sans remise**II.1 Loi faible des grands nombres**

1. On note $x^+ = \max(x, 0)$. La formule est évidente pour $n = 1$. Pour $n \geq 2$, on a :

$$\begin{aligned} \mathbb{P}(X_n = 1 | X_1 = x_1, \dots, X_{n-1} = x_{n-1}) &= \frac{(N_A - \sum_{i=1}^{n-1} x_i)^+}{N - n + 1}, \\ \mathbb{P}(X_n = 0 | X_1 = x_1, \dots, X_{n-1} = x_{n-1}) &= \frac{(N_B - (n-1 - \sum_{i=1}^{n-1} x_i))^+}{N - n + 1}. \end{aligned}$$

La formule s'obtient alors par récurrence. Elle est valable pour $(n - N_B)^+ \leq s \leq \min(n, N_A)$.

2. On remarque que $\mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n))$ ne dépend que de $\sum_{i=1}^n x_i$. En particulier la loi de (X_1, \dots, X_n) est inchangée par permutation des variables aléatoires.
3. La variable aléatoire X_1 est de loi de Bernoulli de paramètre p : $\mathbb{E}[X_1] = p$. D'après la question II.1.2, X_i a même loi que X_1 . Par linéarité, on a donc $\mathbb{E}[\bar{X}_n] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = p$.
4. On a $X_1 = X_1^2$ et donc $\mathbb{E}[X_1^2] = p$. On a :

$$\mathbb{E}[X_1 X_2] = \mathbb{P}(X_1 = 1, X_2 = 1) = \frac{N_A(N_A - 1)}{N(N - 1)} = \frac{N}{N - 1} p \left(p - \frac{1}{N} \right).$$

D'après la question II.1.2, pour tout $i \neq j$ (X_i, X_j) a même loi que (X_1, X_2) et donc $\mathbb{E}[X_1 X_2] = \mathbb{E}[X_i X_j]$. Il vient :

$$\begin{aligned} \mathbb{E}[\bar{X}_n^2] &= \frac{1}{n^2} \left(\sum_{i=1}^n \mathbb{E}[X_i^2] + 2 \sum_{1 \leq i < j \leq n} \mathbb{E}[X_i X_j] \right) \\ &= \frac{1}{n^2} \left(np + n(n - 1) \frac{N}{N - 1} p \left(p - \frac{1}{N} \right) \right). \end{aligned}$$

On obtient donc :

$$\text{Var}(\bar{X}_n) = \mathbb{E}[\bar{X}_n^2] - \mathbb{E}[\bar{X}_n]^2 = \left(1 - \frac{n - 1}{N - 1} \right) \frac{1}{n} p(1 - p).$$

Pour $n = N$, la variance est nulle : l'estimation est exacte car on a interrogé tous les électeurs une et une seule fois.

5. Soit $\varepsilon > 0$. L'inégalité de Tchebychev donne :

$$\mathbb{P}(|\bar{X}_n - p| \geq \varepsilon) \leq \frac{\mathbb{E}[(\bar{X}_n - p)^2]}{\varepsilon^2} = \left(1 - \frac{n - 1}{N - 1} \right) \frac{p(1 - p)}{n\varepsilon^2}.$$

Le membre de droite converge vers 0 quand N et n tendent vers l'infini. Ceci assure la convergence en probabilité de $\bar{X}_n - p$ vers 0.

II.2 Théorème central limite

1. On a avec $s = \sum_{i=1}^n x_i$:

$$\begin{aligned} \mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n)) &= \prod_{i=1}^s \frac{N_A - i + 1}{N - i + 1} \prod_{j=1}^{n-s} \frac{N_B - j + 1}{N - s - j + 1} \\ &= \prod_{i=1}^s \frac{N_A - i + 1}{N - i + 1} \prod_{j=1}^{n-s} \left(1 - \frac{N_A - s}{N - s - j + 1} \right) \\ &\xrightarrow{N \rightarrow \infty} \theta^s (1 - \theta)^{n-s}. \end{aligned}$$

On en déduit que pour toute fonction continue bornée g , on a :

$$\begin{aligned} \mathbb{E}[g(X_1, \dots, X_n)] &= \sum_{x_1, \dots, x_n \in \{0,1\}} g(x_1, \dots, x_n) \mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n)) \\ &\xrightarrow{N \rightarrow \infty} \sum_{x_1, \dots, x_n \in \{0,1\}} g(x_1, \dots, x_n) \theta^s (1 - \theta)^{n-s} \\ &= \mathbb{E}[g(V_1, \dots, V_n)]. \end{aligned}$$

Ceci implique donc (X_1, \dots, X_n) converge en loi vers (V_1, \dots, V_n) .

On montre (XI.8). On a :

$$\begin{aligned} \mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n)) &= \prod_{i=1}^s \frac{p - (i-1)/N}{1 - (i-1)/N} \prod_{j=1}^{n-s} \frac{(1-p) - (j-1)/N}{1 - (s+j-1)/N} \\ &= p^s (1-p)^{n-s} \prod_{k=1}^{n-1} \frac{1}{1 - \frac{k}{N}} \prod_{i=1}^s \left(1 - \frac{i-1}{N_A}\right) \prod_{j=1}^{n-s} \left(1 - \frac{j-1}{N_B}\right). \end{aligned}$$

Il existe une constante c_1 telle que pour $n^2/N \leq 1$, on a $1 \geq \prod_{k=1}^{n-1} (1 - \frac{k}{N}) \geq (1 - \frac{n}{N})^n \geq 1 - c_1 \frac{n^2}{N}$. Par ailleurs si $(a_k, k \in \mathbb{N}^*)$ et $(b_k, k \in \mathbb{N}^*)$ sont des suites de nombres complexes de modules inférieurs à 1 ($|a_k| \leq 1$ et $|b_k| \leq 1$ pour tout $k \in \mathbb{N}^*$), il est facile de vérifier que $|\prod_{k=1}^n a_k - \prod_{k=1}^n b_k| \leq \sum_{k=1}^n |a_k - b_k|$. En particulier, on a pour $n \leq \min(N_A, N_B)$,

$$\left| \prod_{i=1}^s \left(1 - \frac{i-1}{N_A}\right) \prod_{j=1}^{n-s} \left(1 - \frac{j-1}{N_B}\right) - 1 \right| \leq \sum_{i=1}^s \frac{i-1}{N_A} + \sum_{j=1}^{n-s} \frac{j-1}{N_B} \leq \frac{2n^2}{\min(N_A, N_B)}.$$

On en déduit donc (XI.8).

2. On a :

$$\begin{aligned} &|\mathbb{E}[g_n(X_1, \dots, X_n)] - \mathbb{E}[g_n(Z_1, \dots, Z_n)]| \\ &\leq \sum_{x_1, \dots, x_n \in \{0,1\}} |g_n(x_1, \dots, x_n)| \left| \mathbb{P}((X_1, \dots, X_n) = (x_1, \dots, x_n)) - p^s (1-p)^{n-s} \right| \\ &\leq MC \frac{n^2}{\min(N_A, N_B)} \sum_{x_1, \dots, x_n \in \{0,1\}} p^s (1-p)^{n-s} \\ &= MC \frac{n^2}{\min(N_A, N_B)}. \end{aligned}$$

3. On a :

$$\begin{aligned}
 & |\mathbb{E}[g_n(Z_1, \dots, Z_n)] - \mathbb{E}[g_n(V_1, \dots, V_n)]| \\
 & \leq \sum_{x_1, \dots, x_n \in \{0,1\}} |g_n(x_1, \dots, x_n)| |p^s(1-p)^{n-s} - \theta^s(1-\theta)^{n-s}| \\
 & \leq M \int_{[p,\theta]} s \binom{n}{s} x^s (1-x)^{n-s} \frac{dx}{x} + \int_{[1-p,1-\theta]} (n-s) \binom{n}{n-s} x^{n-s} (1-x)^s \frac{dx}{x} \\
 & = M \int_{[p,\theta]} \mathbb{E}[S_x] \frac{dx}{x} + \int_{[1-p,1-\theta]} \mathbb{E}[S_x] \frac{dx}{x} \\
 & = 2Mn|p-\theta|,
 \end{aligned}$$

où S_x est une variable aléatoire de loi binomiale (n, x) .

4. Les conditions impliquent que $n^2 / \min(N_A, N_B)$ tend vers 0. On déduit des questions précédentes que $|\mathbb{E}[g_n(X_1, \dots, X_n)] - \mathbb{E}[g_n(V_1, \dots, V_n)]|$ tend vers 0 pour toute fonction continue bornée.
5. Soit G une variable aléatoire de loi gaussienne centrée réduite. On a :

$$\begin{aligned}
 |\psi_{\sqrt{n}(\bar{X}_n - \theta)}(u) - \psi_{\sqrt{\theta(1-\theta)}G}(u)| & \leq |\psi_{\sqrt{n}(\bar{X}_n - \theta)}(u) - \psi_{\sqrt{n}(\bar{V}_n - \theta)}(u)| \\
 & \quad + |\psi_{\sqrt{n}(\bar{V}_n - \theta)}(u) - \psi_{\sqrt{\theta(1-\theta)}G}(u)|,
 \end{aligned}$$

où $\bar{V}_n = \frac{1}{n} \sum_{k=1}^n V_k$. Comme dans la première partie, $(\sqrt{n}(\bar{V}_n - \theta), n \geq 1)$ converge en loi vers $\sqrt{\theta(1-\theta)}G$. On déduit donc des questions précédentes, et de la continuité de l'exponentielle complexe, que $|\psi_{\sqrt{n}(\bar{X}_n - \theta)}(u) - \psi_{\sqrt{\theta(1-\theta)}G}(u)|$ tend vers 0. Ceci assure la convergence en loi de $\sqrt{n}(\bar{X}_n - \theta)$ vers $\sqrt{\theta(1-\theta)}G$. Par ailleurs $|p - \theta|$ tend vers 0 et $\bar{X}_n - p$ tend en probabilité vers 0, donc \bar{X}_n converge en probabilité vers θ . On déduit du théorème de Slutsky que $(\sqrt{n}(\bar{X}_n - \theta), \bar{X}_n)$ converge en loi vers $(\sqrt{\theta(1-\theta)}G, \theta)$. Comme dans la première partie, on en déduit donc que $\sqrt{n}(\bar{X}_n - \theta) / \sqrt{\bar{X}_n(1 - \bar{X}_n)}$ converge en loi vers G .

6. On en déduit que $J_n = [\bar{X}_n \pm a_\alpha \sqrt{\bar{X}_n(1 - \bar{X}_n)} / \sqrt{n}]$, où a_α est le quantile de la loi $\mathcal{N}(0, 1)$ d'ordre $1 - \alpha/2$ est un intervalle de confiance sur θ de niveau asymptotique $1 - \alpha$. Comme $p = \theta + o(1/n)$, on en déduit que J_n est également un intervalle de confiance sur p de niveau asymptotique $1 - \alpha$. En conclusion si N est grand devant n , il n'y a pas de différence sur la précision d'un sondage sans remise et d'un sondage avec remise.

Si N et n tendent vers l'infini et $\sigma^2 = \text{Var}(\sum_{k=1}^n X_k) = np(1-p)(1 - \frac{n-1}{N-1})$ tend également vers l'infini, alors on peut montrer que $\sqrt{n}(\bar{X}_n - p) / \sigma$ converge en loi vers la loi gaussienne centrée réduite³. Ainsi, si n est grand mais négligeable devant

3. J. Hájek. Limiting distributions in simple random sampling from finite population. *Pub. Math. Inst. Hungarian Acad. Sci.*, vol. 5, pp. 361-374 (1960).

N , le comportement asymptotique de $\sqrt{n}(\bar{X}_n - p)$ est le même pour un sondage avec remise et sans remise, et donc on obtient le même intervalle de confiance pour un niveau asymptotique donné. ▲

Exercice XI.11. I Étude asymptotique

1. Il s'agit de variables aléatoires indépendantes de loi de Bernoulli de paramètre α .
2. On a $\frac{N_n}{n} = \frac{1}{n} + \frac{n-1}{n} \frac{1}{n-1} \sum_{k=2}^n Z_k$. On déduit de la loi forte des grands nombres que la suite $(N_n/n, n \geq 1)$ converge p.s. vers $\mathbb{E}[Z_2] = \alpha$.
3. On a $\sum_{k=1}^n kF_n^{(k)} = n$, $\sum_{k=1}^n F_n^{(k)} = N_n$ et $\sum_{k=1}^n \mathbb{E}[F_n^{(k)}] = 1 + \alpha(n-1)$.
4. Le nombre de boîtes contenant une boule augmente de 1 si on rajoute une boule à une boîte ayant $k-1$ boules et diminue de 1 si on rajoute une boule à une boîte contenant déjà k boules. On a donc :

$$F_{n+1}^{(k)} = F_n^{(k)} + \mathbf{1}_{\{Y_{n+1}=k-1\}} - \mathbf{1}_{\{Y_{n+1}=k\}}. \tag{XIII.9}$$

5. Par construction, conditionnellement à $\{Z_{n+1} = 0\}$ et à $(F_n^{(k)}, k \in \mathbb{N}^*)$, la probabilité de choisir une boîte est proportionnelle à son nombre de boules. En particulier la probabilité de choisir une boîte contenant k boules est proportionnelle à $kF_n^{(k)}$, car il y a $F_n^{(k)}$ boîtes contenant k boules. Comme $\sum_{k=1}^n kF_n^{(k)} = n$, on en déduit que la probabilité de choisir une boîte contenant k boules est $kF_n^{(k)}/n$. On en déduit donc que conditionnellement à $\{Z_{n+1} = 0\}$ la probabilité de choisir une boîte contenant k boules est $\mathbb{E}[kF_n^{(k)}/n] = \alpha k p_n(k)$.
6. En prenant l'espérance dans (XIII.9), on en déduit :

$$\alpha(n+1)p_{n+1}(1) = \alpha n p_n(1) + \alpha - (1-\alpha)\alpha p_n(1), \tag{XIII.10}$$

et pour $k \geq 2$:

$$\alpha(n+1)p_{n+1}(k) = \alpha n p_n(k) + (1-\alpha) (\alpha(k-1)p_n(k-1) - \alpha k p_n(k)). \tag{XIII.11}$$

7. Les équations stationnaires donnent pour (XIII.10) : $\alpha p(1) = \alpha - (1-\alpha)p(1)$ soit :

$$p(1) = \frac{1}{2-\alpha} = \frac{\rho}{1+\rho}$$

et pour (XIII.11) : $(1 + (1-\alpha)k)p(k) = (1-\alpha)(k-1)p(k-1)$ soit $p(k) = \frac{k-1}{k+\rho} p(k-1)$.

Pour compléter le résultat, on montre la convergence en probabilité de la suite $(F_n^{(k)}/\alpha n, n \geq 1)$ vers $p(k)$. Pour $k \geq 2$, on pose $f_n(k) = \alpha n p(k)$, où $p(k)$ est défini par (XI.9) et $\Delta_n(k) = F_n^{(k)} - f_n(k)$. On déduit de (XIII.9) que :

$$\Delta_{n+1}(k) = \Delta_n(k) + \mathbf{1}_{\{Y_{n+1}=k-1\}} - \mathbf{1}_{\{Y_{n+1}=k\}} - \alpha p(k).$$

et donc :

$$\begin{aligned} \Delta_{n+1}(k)^2 &= \Delta_n(k)^2 + \mathbf{1}_{\{Y_{n+1}=k-1\}} + \mathbf{1}_{\{Y_{n+1}=k\}} + \alpha^2 p(k)^2 \\ &\quad + 2\Delta_n(k) (\mathbf{1}_{\{Y_{n+1}=k-1\}} - \mathbf{1}_{\{Y_{n+1}=k\}}) \\ &\quad - 2\Delta_n(k)\alpha p(k) - 2\alpha (\mathbf{1}_{\{Y_{n+1}=k-1\}} - \mathbf{1}_{\{Y_{n+1}=k\}}) p(k). \end{aligned}$$

On pose :

$$h_n = \sum_{k \in \mathbb{N}^*} \mathbb{E}[\Delta_n(k)^2]$$

et par convention $\Delta_n(k-1) = p(k-1) = 0$ pour $k=1$. Il vient :

$$\begin{aligned} h_{n+1} &= \sum_{k \in \mathbb{N}^*} \mathbb{E}[\Delta_{n+1}(k)^2] \\ &= h_n + \sum_{k \in \mathbb{N}^*} (\mathbb{P}(Y_{n+1} = k-1) + \mathbb{P}(Y_{n+1} = k) + \alpha^2 p(k)^2) \\ &\quad + 2\frac{1-\alpha}{n} \sum_{k \in \mathbb{N}^*} \mathbb{E}[\Delta_n(k) ((k-1)\Delta_n(k-1) - k\Delta_n(k))] \\ &\quad + 2(1-\alpha)\alpha \sum_{k \in \mathbb{N}^*} \mathbb{E}[\Delta_n(k) ((k-1)p(k-1) - kp(k))] + 2\alpha \mathbb{E}[\Delta_n(1)] \\ &\quad - 2\alpha \sum_{k \in \mathbb{N}^*} p(k)\mathbb{E}[\Delta_n(k)] - 2\alpha \sum_{k \in \mathbb{N}^*} p(k) (\mathbb{P}(Y_{n+1} = k-1) - \mathbb{P}(Y_{n+1} = k)) \\ &= h_n \left(1 - \frac{1-\alpha}{n}\right) + 2 - \alpha + \alpha^2 \sum_{k \in \mathbb{N}^*} p(k)^2 \\ &\quad - 2\alpha \sum_{k \in \mathbb{N}^*} p(k) (\mathbb{P}(Y_{n+1} = k-1) - \mathbb{P}(Y_{n+1} = k)) \\ &\quad - \frac{1-\alpha}{n} \sum_{k \in \mathbb{N}^*} (k-1)\mathbb{E}[(\Delta_n(k) - \Delta_n(k-1))^2] \\ &\leq h_n \left(1 - \frac{1-\alpha}{n}\right) + 4. \end{aligned}$$

On remarque que $h_1 = \mathbb{E}[(1 - \alpha p(1))^2] + \alpha^2 \sum_{k \geq 2} p(k)^2 \leq 2$. Il est alors immédiat de montrer par récurrence que $h_n \geq 4n/(2 - \alpha)$. On en déduit donc que

$(\frac{1}{n^2} \sum_{k \in \mathbb{N}^*} \Delta_n(k)^2, n \geq 1)$ converge dans L^1 et en probabilité vers 0. En particulier, la suite $(\sup_{k \in \mathbb{N}^*} \left| \frac{F_n^{(k)}}{\alpha n} - p(k) \right|, n \geq 1)$ converge dans L^1 et en probabilité vers 0. Ce résultat est plus fort que celui utilisé dans l'énoncé de l'exercice.

II Loi de Yule

1. On a :

$$p(k) = \frac{(k-1) \cdots 1}{(k+\rho) \cdots (2+\rho)} p(1) = \rho \frac{\Gamma(k) \Gamma(\rho+1)}{\Gamma(k+\rho+1)} = \rho B(k, \rho+1).$$

On a également :

$$\sum_{r=k}^{\infty} p(r) = \rho \int_{]0,1[} \sum_{r=k}^{\infty} x^{r-1} (1-x)^\rho dx = \rho \int_{]0,1[} x^{k-1} (1-x)^{\rho-1} dx = \rho B(k, \rho).$$

En particulier $\sum_{r=1}^{\infty} p(r) = 1$. Comme de plus $p(k) \geq 0$ pour $k \in \mathbb{N}^*$, on en déduit que $(p(k), k \in \mathbb{N}^*)$ définit une probabilité sur \mathbb{N}^* .

2. Par la formule de Stirling, on a $p(k) = \rho B(k, \rho+1) = \rho \frac{\Gamma(\rho+1)}{k^{\rho+1}} (1 + o(1))$ et

$$\sum_{r=k}^{\infty} p(r) = \rho B(k, \rho) = \frac{\Gamma(\rho+1)}{k^\rho} (1 + o(1)).$$

3. Soit Y de loi de Yule de paramètre $\rho > 0$. Soit V_x une variable aléatoire de loi géométrique de paramètre $1-x \in]0,1[$. On rappelle que $\mathbb{E}[V_x] = 1/(1-x)$ et $\mathbb{E}[V_x^2] = \text{Var}(V_x) + \mathbb{E}[V_x]^2 = (1+x)/(1-x)^2$. On a, en utilisant Fubini pour permuter l'intégration (en x) et la sommation (en k) :

$$\begin{aligned} \mathbb{E}[Y] &= \rho \int_{]0,1[} \sum_{k \in \mathbb{N}^*} k x^{k-1} (1-x)^\rho dx = \rho \int_{]0,1[} \mathbb{E}[V_x] (1-x)^{\rho-1} dx \\ &= \rho \int_{]0,1[} (1-x)^{\rho-2} dx. \end{aligned}$$

On obtient donc $\mathbb{E}[Y] = +\infty$ si $\rho \in]0,1[$ et $\mathbb{E}[Y] = \frac{\rho}{\rho-1}$ si $\rho \in]1, \infty[$. On a également :

$$\mathbb{E}[Y^2] = \rho \int_{]0,1[} \sum_{k \in \mathbb{N}^*} k^2 x^{k-1} (1-x)^\rho dx = \rho \int_{]0,1[} \mathbb{E}[V_x^2] (1-x)^{\rho-1} dx$$

et donc $\mathbb{E}[Y^2] = \rho \int_{]0,1[} (1+x)(1-x)^{\rho-3} dx$. Il vient $\mathbb{E}[Y^2] = +\infty$ pour $\rho \in]0,2[$ et pour $\rho \in]2, \infty[$:

$$\mathbb{E}[Y^2] = 2\rho \int_{]0,1[} (1-x)^{\rho-3} dx - \rho \int_{]0,1[} (1-x)^{\rho-2} dx = \frac{\rho^2}{(\rho-1)(\rho-2)}$$

Pour $\rho > 2$, on obtient $\text{Var}(Y) = \frac{\rho^2}{(\rho-1)^2(\rho-2)}$.

▲

Exercice XI.12. I Le cas à une période : de $N-1$ à N

1. On déduit de (XI.12) en $n = N-1$ et de $V_N = h(S_N)$ que $\phi_{N-1}^0(1+r)S_{N-1}^0 + \phi_{N-1}X_N S_{N-1} = h(X_N S_{N-1})$. On obtient les deux équations en considérant les événements $\{X_N = 1+m\}$ et $\{X_N = 1+d\}$.
2. La résolution du système linéaire élémentaire donne :

$$\phi_{N-1} = \frac{1}{S_{N-1}} \frac{h((1+m)S_{N-1}) - h((1+d)S_{N-1})}{m-d} = \varphi(N, S_{N-1}).$$

3. On a :

$$\begin{aligned} V_{N-1} &= \phi_{N-1}^0 S_{N-1}^0 + \phi_{N-1} S_{N-1} \\ &= \frac{1}{1+r} h((1+m)S_{N-1}) - \frac{1}{1+r} \phi_{N-1}(1+m)S_{N-1} + \phi_{N-1} S_{N-1} \\ &= (1+r)^{-1} \left(\frac{r-d}{m-d} h((1+m)S_{N-1}) + \frac{m-r}{m-d} h((1+d)S_{N-1}) \right) \\ &= v(N-1, S_{N-1}). \end{aligned}$$

II Le cas général

1. On a, en utilisant l'indépendance des variables X_n et (X_{n+1}, \dots, X_N) sous \mathbb{P}^* :

$$\begin{aligned}
 & v(n-1, s) \\
 &= (1+r)^{-N+n-1} \mathbb{E}^* \left[h\left(s \prod_{k=n}^N X_k\right) \right] \\
 &= (1+r)^{-N+n-1} \sum_{x_n, \dots, x_N \in \{1+d, 1+m\}} h\left(s \prod_{k=n}^N x_k\right) \mathbb{P}^*(X_n = x_n, \dots, X_N = x_N) \\
 &= (1+r)^{-1} \sum_{x_n \in \{1+d, 1+m\}} \mathbb{P}^*(X_n = x_n) (1+r)^{-N+n} \\
 &\quad \sum_{x_{n+1}, \dots, x_N \in \{1+d, 1+m\}} h\left(s x_n \prod_{k=n+1}^N x_k\right) \mathbb{P}^*(X_{n+1} = x_{n+1}, \dots, X_N = x_N) \\
 &= (1+r)^{-1} \sum_{x_n \in \{1+d, 1+m\}} \mathbb{P}^*(X_n = x_n) v(n, s x_n) \\
 &= (1+r)^{-1} \mathbb{E}^* [v(n, s X_n)].
 \end{aligned}$$

2. On démontre le résultat par récurrence descendante. Le résultat est vrai au rang $N-1$ d'après la partie I. On suppose qu'il est vrai au rang $N' \in \{1, \dots, N-1\}$. Les calculs de la partie avec N remplacé par N' et h par $v(N', \cdot)$ assurent que $\phi_{N'-1} = \varphi(N', S_{N'-1})$ et $V_{N'-1} = w(S_{N'-1})$ où $w(s) = (1+r)^{-1} \mathbb{E} [v(N', s X_{N'})] = v(N'-1, s)$ d'après la question précédente. L'hypothèse de récurrence est satisfaite au rang $N'-1$. Le résultat est donc vrai.
3. La construction par récurrence descendante assure que la stratégie autofinancée de couverture pour l'option h existe et est unique. Son prix à l'instant initial est :

$$V_0 = v(0, S_0) = (1+r)^{-N} \mathbb{E}^* \left[h\left(S_0 \prod_{k=1}^N X_k\right) \right] = (1+r)^{-N} \mathbb{E}^* [h(S_N)].$$

4. Si le modèle est juste, il n'y a pas de risque. Mais le modèle est-il juste ? Les taux d'intérêt ne sont pas fixes, les incréments relatifs d'un actif risqué ne prennent pas que les deux valeurs $1+m$ et $1+d, \dots$. Il existe donc un risque de modèle. Ce risque est important en pratique, c'est une des causes de l'ampleur de la crise financière actuelle.

III Call et Put

1. On a par indépendance :

$$\mathbb{E}^* \left[\prod_{k=n+1}^N X_k \right] = \prod_{k=n+1}^N \mathbb{E}^*[X_k] = \mathbb{E}^*[X_1]^{N-n} = (1+r)^{N-n}.$$

2. Avec pour h la fonction identité, il vient $v(n, s) = s$ et $\varphi(n, s) = 1$. On en déduit que la valeur de la stratégie à l'instant n est S_n et qu'il faut détenir à cet instant un actif risqué. La stratégie autofinancée de couverture pour l'option qui promet un actif risqué à l'instant N consiste à acheter un actif risqué à l'instant 0 et de le garder jusqu'à l'instant N . Le prix initial de cette stratégie est S_0 .
3. Après avoir remarqué que $(x - K)_+ - (K - x)_+ = (x - K)$, on déduit de (XI.14) et de la question III.1 avec $n = 0$ que :

$$\begin{aligned} C(S_0, K, N) - P(S_0, K, N) &= (1 + r)^{-N} \mathbb{E}^* [(S_N - K)_+ - (K - S_N)_+] \\ &= (1 + r)^{-N} \mathbb{E}^* [S_N - K] \\ &= (1 + r)^{-N} \mathbb{E}^* \left[\prod_{k=1}^N X_k \right] S_0 - (1 + r)^{-N} K \\ &= S_0 - (1 + r)^{-N} K. \end{aligned}$$

4. Soit $(Y_n, n \in \mathbb{N}^*)$ des variables aléatoires indépendantes de loi de Bernoulli de paramètre $(r - d)/(m - d)$. On pose $\tilde{X}_n = (1 + m)^{Y_k} (1 + d)^{1 - Y_k}$ et on remarque que les variables aléatoires $(\tilde{X}_n, n \in \mathbb{N}^*)$ ont même loi que $(X_n, n \in \mathbb{N}^*)$. On en déduit que :

$$\begin{aligned} \mathbb{E}^* \left[\left(s \prod_{k=1}^N X_k - K \right)_+ \right] &= \mathbb{E} \left[\left(s \prod_{k=1}^N \tilde{X}_k - K \right)_+ \right] \\ &= \mathbb{E} \left[\left(s(1 + m)^{Z_N} (1 + d)^{N - Z_N} - K \right)_+ \right], \end{aligned}$$

où $Z_N = \sum_{k=1}^N Y_k$ est de loi binomiale de paramètre $(N, (r - d)/(m - d))$. La formule en découle.

▲

Exercice XI.13. I Code à répétition

1. Par construction, on a :

$$p_{f,g} = \mathbb{P}(E_1 + E_2 + E_3 \geq 2) = p^3 + 3p^2(1 - p) = p^2(3 - 2p).$$

Comme $p \in]0, 1/2[$, on peut vérifier que $p_{f,g} < p$.

2. Si on considère un code à répétition de longueur impaire $n = 2n' + 1$ avec règle de la majorité pour le décodage, on obtient une probabilité d'erreur pour un message de longueur $m = 1$:

$$p_{f,g} = \mathbb{P}\left(\sum_{k=1}^n E_k \geq n' + 1\right) \leq \mathbb{P}(\bar{E}_n \geq 1/2),$$

où $\bar{E}_n = \frac{1}{n} \sum_{k=1}^n E_k$. La loi forte des grands nombres implique que p.s. $\lim_{n \rightarrow \infty} \bar{E}_n = p < 1/2$. Le théorème de convergence dominée implique que $\lim_{n \rightarrow \infty} \mathbb{P}(\bar{E}_n \geq 1/2) = 0$. Pour $\varepsilon > 0$ fixé, on peut trouver, en choisissant n suffisamment grand un code de répétition dont la probabilité d'erreur soit inférieure à ε . Le taux de transmission $1/n$ est alors proche de 0 car n est grand.

II Probabilités d'évènements rares

1. La loi forte des grands nombres assure (cf. la réponse à la question I.2) que la probabilité de l'évènement $\{\bar{V}_n \geq a\}$ converge vers 0 quand n tend vers l'infini. On remarque que le théorème de la limite centrale assure que les évènements $\{\bar{V}_n \geq \rho + c/\sqrt{n}\}$ ont une probabilité qui converge vers une limite non nulle.
2. On remarque que $\mathbf{1}_{\{\bar{V}_n \geq a\}} \leq e^{\lambda \sum_{i=1}^n V_i - \lambda na}$. En prenant l'espérance puis en utilisant l'indépendance, il vient :

$$\mathbb{P}(\bar{V}_n \geq a) \leq \mathbb{E} \left[e^{\lambda \sum_{i=1}^n V_i - \lambda na} \right] = \mathbb{E} \left[e^{\lambda V_1} \right]^n e^{-an\lambda}.$$

3. On a :

$$\frac{\partial_2 A_\rho^*(v)}{\partial^2 v} = \frac{1}{v} + \frac{1}{1-v} > 0.$$

La fonction A_ρ^* est donc strictement convexe sur $]0, 1[$. Il est immédiat de vérifier qu'elle s'annule en ρ ainsi que sa dérivée. Elle atteint donc son minimum en 0.

4. On considère la fonction h définie pour $\lambda > 0$ par :

$$h(\lambda) = \mathbb{E} \left[e^{\lambda V_1} \right] e^{-a\lambda} = p e^{\lambda(1-a)} + (1-p) e^{-a\lambda}.$$

La fonction h est strictement convexe (car $h'' > 0$). On a $h'(\lambda) = 0$ pour $\lambda = \lambda_0$ défini par $e^{\lambda_0} = a(1-p)/(p(1-a))$. On remarque que $h(\lambda_0) = e^{-A_\rho^*(a)}$. On déduit de la question précédente que :

$$\mathbb{P}(\bar{V}_n > a) \leq \mathbb{P}(\bar{V}_n \geq a) \leq h(\lambda_0)^n = e^{-nA_\rho^*(a)}.$$

5. On pose $V'_i = 1 - V_i$. Les variables aléatoires $(V'_n, n \in \mathbb{N}^*)$ sont indépendantes de loi de Bernoulli de paramètre $1 - \rho$. En utilisant (XI.17) avec V' au lieu de V et $a = 1 - b$, on obtient :

$$\mathbb{P}(\bar{V}_n \leq b) = \mathbb{P}(\bar{V}'_n \geq 1 - b) \leq e^{-nA_{1-\rho}^*(1-b)} = e^{-nA_\rho^*(b)}.$$

III Codes presque optimaux

1. Soit (V_1, \dots, V_n) des variables aléatoires indépendantes de loi uniforme sur $\{0, 1\}$. On remarque que $|Z_{x_i} - z_i|$ a même loi que V_i . On en déduit que $\delta(Z_x, z)$ a même loi que $\sum_{i=1}^n V_i$ et donc suit une loi binomiale de paramètre $(n, 1/2)$.
2. En utilisant (XI.18), il vient :

$$\mathbb{P}\left(\delta(Z_x, Z_{x'}) \leq nr \mid Z_x = z\right) = \mathbb{P}(\delta(Z_x, z) \leq nr) = \mathbb{P}(\bar{V}_n \leq r) \leq e^{-n\Lambda_{1/2}^*(r)}.$$

3. On a :

$$\begin{aligned} \mathbb{P}\left(\delta(Z_x, Z_{x'}) \leq nr\right) &= \sum_{z \in \{0,1\}^n} \mathbb{P}\left(\delta(Z_x, Z_{x'}) \leq nr \mid Z_x = z\right) \mathbb{P}(Z_x = z) \\ &\leq e^{-n\Lambda_{1/2}^*(r)}. \end{aligned}$$

On en déduit que :

$$h(x) \leq \sum_{x' \in \{0,1\}^n, x' \neq x} \mathbb{P}\left(\delta(Z_x, Z_{x'}) \leq nr\right) \leq 2^m e^{-n\Lambda_{1/2}^*(r)}.$$

4. Soit $x \in \{0, 1\}^m$. On remarque que si pour tout $x' \neq x$, on a $\delta(Z_x, Z_{x'}) > nr$ et $\delta(0, E) \leq nr$, alors on a $g(f(x) + E) = x$. En particulier $\{g(f(x) + E) \neq x\}$ est inclus dans la réunion de $\{\text{Il existe } x' \neq x \text{ tel que } \delta(Z_x, Z_{x'}) \leq nr\}$ et de $\{\delta(Z_x, Z_x + E) \leq nr\} = \{\delta(E, 0) \leq nr\}$. On en déduit donc que :

$$\mathbb{P}\left(g(f(x) + E) \neq x\right) \leq h(x) + \mathbb{P}(\delta(0, E) > nr).$$

Comme X est indépendant de E et $(Z_x, x \in \{0, 1\}^m)$, on en déduit que :

$$\begin{aligned} \mathbb{P}\left(g(f(X) + E) \neq X\right) &= \sum_{x \in \{0,1\}^m} \mathbb{P}\left(g(f(X) + E) \neq X \mid X = x\right) \mathbb{P}(X = x) \\ &= \sum_{x \in \{0,1\}^m} \mathbb{P}\left(g(f(x) + E) \neq x\right) \mathbb{P}(X = x) \\ &\leq \sum_{x \in \{0,1\}^m} (h(x) + \mathbb{P}(\delta(0, E) > nr)) \mathbb{P}(X = x) \\ &= \mathbb{E}[h(X)] + \mathbb{P}(\delta(0, E) > nr). \end{aligned}$$

5. On a $\mathbb{E}[h(X)] \leq 2^m e^{-n\Lambda_{1/2}^*(r)}$ et $\mathbb{P}(\delta(0, E) > nr) = \mathbb{P}(\bar{E}_n > r) \leq e^{-n\Lambda_p^*(r)}$ d'après (XI.17).

6. On a :

$$2^m e^{-n\Lambda_{1/2}^*(r)} = e^{n\left(\frac{m}{n} - \Lambda_{1/2}^*(r)\right)}.$$

Cette quantité converge vers 0 quand n tend vers l'infini dès que $\frac{m}{n} < \Lambda_{1/2}^*(r) - \varepsilon_0$ pour $\varepsilon_0 > 0$ fixé. Par ailleurs, la question II.3 assure que $\Lambda_{1/2}^*(r) - \Lambda_{1/2}^*(p)$ est positif, mais peut être choisi aussi petit que l'on veut (avec r proche de p). Donc, pour $\varepsilon > 0$ fixé, on peut trouver r tel que $0 < \Lambda_{1/2}^*(r) - \Lambda_{1/2}^*(p) < \varepsilon/3$. Pour m et n qui tendent vers l'infini et tels que $c_p - \varepsilon \leq m/n < c_p - 2\varepsilon/3$, on a que $2^m e^{-n\Lambda_{1/2}^*(r)}$ converge vers 0. La question II.3 assure également que $\Lambda_p^*(r)$ est strictement positif et donc $e^{-n\Lambda_p^*(r)}$ converge vers 0 quand n tend vers l'infini.

On en déduit que $\mathbb{P}\left(g(f(X) + E) \neq X\right) < \varepsilon$ pour m et n qui suffisamment grands et tels que $c_p - \varepsilon \leq m/n < c_p - 2\varepsilon/3$. Dans la probabilité d'erreur précédente, les fonctions de codage et de décodage sont aléatoires et indépendantes du message et des erreurs. On en déduit donc qu'il en existe au moins une (déterministe) telle que sa probabilité d'erreur soit très faible pour un taux de transmission proche de c_p .

▲

Exercice XI.14. Pour deux permutations différentes correspondent des rangs partiels différents. La fonction h est donc une injection de \mathcal{S}_n dans E_n . Comme les ensembles \mathcal{S}_n et E_n ont même cardinal, la fonction h est donc une bijection.

I Préliminaires

1. Les candidates étant présentées au hasard, on en déduit que toutes les configurations sont équiprobables. On modélise donc la loi de Σ par la loi uniforme sur \mathcal{S}_n .
2. Comme h est une bijection, on en déduit que R est de loi uniforme sur E_n . En particulier, pour $(r_1, \dots, r_n) \in E_n$, on a $\mathbb{P}(R_1 = r_1, \dots, R_n = r_n) = 1/n!$. La formule des lois marginales implique alors que $\mathbb{P}(R_k = r_k) = 1/k$, et donc R_k est de loi uniforme sur $\{1, \dots, k\}$.
3. On en déduit que :

$$\mathbb{P}(R_1 = r_1, \dots, R_n = r_n) = \prod_{k=1}^n \mathbb{P}(R_k = r_k).$$

Ceci assure que les variables aléatoires (R_1, \dots, R_n) sont indépendantes.

4. Si la princesse k est la meilleure alors son rang partiel vaut 1 et les princesses suivantes étant moins bien, leur rang partiel est donc strictement supérieur à 1. Comme h est une bijection et qu'il existe toujours un rang partiel égal à 1 (car $R_1 = 1$), on en déduit l'égalité demandée.

II Les stratégies de seuil

1. On décompose suivant les valeurs de τ_c :

$$\begin{aligned}
 \mathbb{P}(\Sigma_{\tau_c} = 1) &= \sum_{k=1}^n \mathbb{P}(\Sigma_k = 1, \tau_c = k) \\
 &= \sum_{k=c+2}^{n-1} \mathbb{P}(R_{c+1} > 1, \dots, R_{k-1} > 1, R_k = 1, \Sigma_k = 1) \\
 &\quad + \mathbb{P}(R_{c+1} = 1, \Sigma_{c+1} = 1) + \mathbb{P}(R_{c+1} > 1, \dots, R_{n-1} > 1, \Sigma_n = n) \\
 &= \mathbb{P}(R_{c+1} = 1, \Sigma_{c+1} = 1) \\
 &\quad + \sum_{k=c+2}^n \mathbb{P}(R_{c+1} > 1, \dots, R_{k-1} > 1, R_k = 1, \Sigma_k = 1).
 \end{aligned}$$

2. En utilisant l'indépendance des variables (R_1, \dots, R_n) et le fait que R_i est de loi uniforme sur $\{1, \dots, i\}$, on a pour $c \geq 1$:

$$\begin{aligned}
 \mathbb{P}(R_{c+1} = 1, \Sigma_{c+1} = 1) &= \mathbb{P}(R_{c+1} = 1, R_{c+2} > 1, \dots, R_n > 1) \\
 &= \frac{1}{c+1} \left(\prod_{j=c+2}^n \frac{j-1}{j} \right) = \frac{1}{n},
 \end{aligned}$$

et pour $k \geq c+2$:

$$\begin{aligned}
 &\mathbb{P}(R_{c+1} > 1, \dots, R_{k-1} > 1, R_k = 1, \Sigma_k = 1) \\
 &= \mathbb{P}(R_{c+1} > 1, \dots, R_{k-1} > 1, R_k = 1, R_{k+1} > 1, \dots, R_n > 1) \\
 &= \left(\prod_{i=c+1}^{k-1} \frac{i-1}{i} \right) \frac{1}{k} \left(\prod_{j=k+1}^n \frac{j-1}{j} \right) \\
 &= \frac{c}{n} \frac{1}{k-1}.
 \end{aligned}$$

On obtient donc :

$$\mathbb{P}(\Sigma_{\tau_c} = 1) = \sum_{k=c+1}^n \frac{c}{n} \frac{1}{k-1} = \frac{c}{n} \sum_{k=c}^{n-1} \frac{1}{k}.$$

Pour $c = 0$, on a $\tau_c = 1$ et donc :

$$\mathbb{P}(\Sigma_{\tau_c} = 1) = \mathbb{P}(\Sigma_1 = 1) = \frac{1}{n}.$$

3. Pour $c \in \{0, \dots, n-1\}$, on a :

$$f(c) = n(g(c) - g(c-1)) = \sum_{k=c}^{n-1} \frac{1}{k} - 1.$$

Comme $n \geq 3$, on a $f(1) = \sum_{k=2}^{n-1} \frac{1}{k} > 0$. Pour $c > c_*$ on a $f(c) < 0$ et donc $g(c) < g(c_*)$. Pour $c \leq c_*$ on a $f(c) \geq 0$ et donc $g(c) \leq g(c_*)$. On en déduit que $g(c)$ est maximal pour $c = c_*$.

4. Par construction, τ_{c_*} est la stratégie de seuil optimale. Elle consiste à choisir la première princesse meilleure que les c_* premières princesses rencontrées.
5. Comme la somme $\sum_{k=c}^{n-1} 1/k$ diverge quand n tend vers l'infini, on en déduit que c_* tend vers l'infini quand n tend vers l'infini. On déduit de la définition de c_* que :

$$\log(n/c_*) + o(1/c_*) \geq 1 > \log(n/c_*) + o(1/c_*) - \frac{1}{c_*} = \log(n/c_*) + o(1/c_*).$$

Ceci implique que $\log(n/c_*) + o(1/c_*) = 1$ et donc $n/c_* = e + o(1/c_*)$ soit $c_*/n = 1/e + o(1/n)$. On a également $g(c_*) = (1 + o(1/c_*))c_*/n = 1/e + o(1/n)$.

III Les stratégies générales

On remarque que $\{\tau > k\} = \bigcap_{j=1}^k \{(R_1, \dots, R_j) \notin A_j\} = \{(R_1, \dots, R_k) \in C_k\}$ avec :

$$C_k = \bigcap_{j=1}^k \left(A_j^c \times \prod_{i=j+1}^k E_i \right),$$

ainsi que $\{\tau = k\} = \{\tau \geq k\} \cap \{\tau > k\}^c = \{(R_1, \dots, R_k) \in B_k\}$ avec :

$$B_k = (C_{k-1} \times E_k) \cap C_k^c.$$

1. On a $G_n(r_1, \dots, r_n) = \mathbf{1}_{\{r_n=1\}} \mathbf{1}_{\{(r_1, \dots, r_n) \in B_n\}}$.
2. En utilisant la question I.4, il vient par indépendance :

$$\begin{aligned}
 G_k(r_1, \dots, r_k) &= k! \sum_{i=k}^n \mathbb{P}(\Sigma_i = 1, \tau = i, R_1 = r_1, \dots, R_k = r_k) \\
 &= k! \sum_{i=k}^n \mathbb{P}(R_n > 1, \dots, R_{i+1} > 1, R_i = 1, \tau = i, R_1 = r_1, \dots, R_k = r_k) \\
 &= k! \sum_{i=k}^n \mathbb{P}(R_n > 1, \dots, R_{i+1} > 1) \mathbb{P}(R_i = 1, \tau = i, R_1 = r_1, \dots, R_k = r_k) \\
 &= k! \sum_{i=k}^n \frac{i}{n} \mathbb{P}(R_i = 1, \tau = i, R_1 = r_1, \dots, R_k = r_k).
 \end{aligned}$$

3. On a, en utilisant la question précédente pour l'avant-dernière égalité :

$$\begin{aligned}
 G_k(r_1, \dots, r_k) &= k! \sum_{i=k}^n \frac{i}{n} \mathbb{P}(R_i = 1, \tau = i, R_1 = r_1, \dots, R_k = r_k) \\
 &= \frac{k}{n} \mathbf{1}_{\{r_k=1\}} \mathbf{1}_{\{(r_1, \dots, r_k) \in B_k\}} k! \mathbb{P}(R_1 = r_1, \dots, R_k = r_k) \\
 &\quad + k! \sum_{i=k+1}^n \frac{i}{n} \mathbb{P}(R_i = 1, \tau = i, R_1 = r_1, \dots, R_k = r_k) \\
 &= \frac{k}{n} \mathbf{1}_{\{r_k=1\}} \mathbf{1}_{\{(r_1, \dots, r_k) \in B_k\}} \\
 &\quad + \frac{1}{k+1} \sum_{r_{k+1}=1}^{k+1} (k+1)! \sum_{i=k+1}^n \frac{i}{n} \mathbb{P}(R_i = 1, \tau = i, R_1 = r_1, \dots, R_{k+1} = r_{k+1}) \\
 &= \frac{k}{n} \mathbf{1}_{\{r_k=1\}} \mathbf{1}_{\{(r_1, \dots, r_k) \in B_k\}} + \frac{1}{k+1} \sum_{r_{k+1}=1}^{k+1} G_{k+1}(r_1, \dots, r_k, r_{k+1}) \mathbf{1}_{\{(r_1, \dots, r_k) \in C_k\}} \\
 &= \frac{k}{n} \mathbf{1}_{\{r_k=1\}} \mathbf{1}_{\{(r_1, \dots, r_k) \in B_k\}} + \mathbb{E}[G_{k+1}(r_1, \dots, r_k, R_{k+1})] \mathbf{1}_{\{(r_1, \dots, r_k) \in C_k\}}.
 \end{aligned}$$

L'inégalité est alors évidente.

4. On remarque que $g^*(n) = 1/n$ et $\mathbb{E}[G_n(r_1, \dots, r_{n-1}, R_n)] \leq \mathbb{P}(R_n = 1) = 1/n$. Les inégalités se démontrent alors par récurrence descendante en utilisant l'inégalité de la question précédente. Pour conclure, il suffit de remarquer que $G_1(1) = \mathbb{E}[G_1(R_1)]$.
5. Par définition de g^* , on a $g^*(n) = 1/n = g(n-1)$ et pour $k \in \{1, \dots, n-1\}$:

$$g^*(k) = \frac{k-1}{k}g^*(k+1) + \max\left(\frac{1}{n}, \frac{g^*(k+1)}{k}\right). \quad (\text{XIII.12})$$

Pour $k > c_*$, on a :

$$\begin{aligned} \frac{k-1}{k}g(k) + \max\left(\frac{1}{n}, \frac{g(k)}{k}\right) &= \frac{k-1}{n} \sum_{j=k}^{n-1} \frac{1}{j} + \frac{1}{n} \max\left(1, \sum_{j=k}^{n-1} \frac{1}{j}\right) \\ &= \frac{k-1}{n} \sum_{j=k}^{n-1} \frac{1}{j} + 1 \\ &= \frac{k-1}{n} \sum_{j=k-1}^{n-1} \frac{1}{j} = g(k-1). \end{aligned}$$

Pour $k \leq c_*$, on a $\sum_{j=c_*}^{n-1} 1/j \geq 1 \geq k/c_*$ et donc $g(c_*)/k \geq 1/n$. Il vient pour $k \leq c_*$:

$$\frac{k-1}{k}g(c_*) + \max\left(\frac{1}{n}, \frac{g(c_*)}{k}\right) = g(c_*).$$

On en déduit que $g^*(k)$ et $g(\max(k-1, c_*))$ vérifient l'équation (XIII.12) avec la même condition initiale en n . Ces deux fonctions sont donc égales.

6. La question précédente assure que $g^*(1) = g(c_*)$ et donc $\mathbb{P}(\Sigma_\tau = 1) \leq g(c_*) = \mathbb{P}(\Sigma_{\tau_{c_*}} = 1)$. La stratégie τ_{c_*} est la stratégie qui maximise la probabilité d'épouser la meilleure princesse. C'est donc la stratégie optimale.

▲

XIII.12 Problèmes (probabilités et statistique)

I Distribution du génotype : le modèle de Hardy-Weinberg

Exercice XII.1. 1. On note (E_1, E_2) les gènes de l'enfant, où 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. On remarque 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 :

$$\begin{aligned} \mathbb{P}(E = aa|M = aA, P = aA) \\ = \frac{\text{Card} \{(E_1, E_2) = (a, a); E_1 \in \{A, a\} \text{ et } E_2 \in \{a, A\}\}}{\text{Card} \{(E_1, E_2); E_1 \in \{A, a\} \text{ et } E_2 \in \{a, A\}\}} = \frac{1}{4}, \end{aligned}$$

$$\begin{aligned} \mathbb{P}(E = aa|M = aa, P = aA) \\ = \frac{\text{Card} \{(E_1, E_2) = (a, a); E_1 \in \{a\} \text{ et } E_2 \in \{a, A\}\}}{\text{Card} \{(E_1, E_2); E_1 \in \{a\} \text{ et } E_2 \in \{a, A\}\}} = \frac{1}{2}, \end{aligned}$$

et par symétrie $\mathbb{P}(E = aa|M = aA, P = aa) = \frac{1}{2}$ et $\mathbb{P}(E = aa|M = aa, P = aa) = 1$.

- 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 :

$$\begin{aligned} \mathbb{P}(E = aa) &= \mathbb{P}(E = aa|M = aA, P = aA)\mathbb{P}(M = aA, P = aA) \\ &\quad + \mathbb{P}(E = aa|M = aa, P = aA)\mathbb{P}(M = aa, P = aA) \\ &\quad + \mathbb{P}(E = aa|M = aA, P = aa)\mathbb{P}(M = aA, P = aa) \\ &\quad + \mathbb{P}(E = aa|M = aa, P = aa)\mathbb{P}(M = aa, P = aa). \end{aligned}$$

On suppose les mariages aléatoires. En supposant le génotype de la mère et du père indépendant et de même répartition, on a $\mathbb{P}(M = i, P = j) = \mathbb{P}(M = i)\mathbb{P}(P = j) = u_i u_j$. On obtient $\mathbb{P}(E = aa) = (u_1 + \frac{u_2}{2})^2$.

- Par symétrie on a $\mathbb{P}(E = AA) = (u_3 + \frac{u_2}{2})^2$.
- En passant au complémentaire on a $\mathbb{P}(E = aA) = 1 - \mathbb{P}(E \neq aA) = 1 - \mathbb{P}(E = aa) - \mathbb{P}(E = AA)$ car les événements $\{E = aa\}$ et $\{E = AA\}$ sont disjoints. Donc on a $\mathbb{P}(E = aA) = 1 - \theta^2 - (1 - \theta)^2 = 2\theta(1 - \theta)$.
- 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

- Les variables aléatoires $(\mathbf{1}_{\{X_i=j\}}, i \in \{1, \dots, n\})$ sont indépendantes et de même loi : la loi de Bernoulli de paramètre q_j . La loi de N_j est donc une loi binomiale de paramètre (n, q_j) .
- On a $\mathbb{E}[N_j] = nq_j$ et $\text{Var}(N_j) = nq_j(1 - q_j)$.
- L'estimateur N_j/n de q_j est sans biais. 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 $\text{Var}(\mathbf{1}_{\{X_i=j\}}) = q_j(1 - q_j)$, on déduit du théorème central limite que $\sqrt{n} \left(\frac{N_j}{n} - q_j \right)$ converge en loi vers $\mathcal{N}(0, q_j(1 - q_j))$. L'estimateur N_j/n est donc un estimateur asymptotiquement normal de variance asymptotique $q_j(1 - q_j)$.

5. On a :

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3) = \sum_{\substack{(x_1, \dots, x_n) \in \{1, 2, 3\}^n \\ \text{Card } \{i; x_i=1\} = n_1 \\ \text{Card } \{i; x_i=2\} = n_2}} \mathbb{P}(X_1 = x_1, \dots, X_n = x_n)$$

par indépendance, et comme les X_i ont même loi :

$$\begin{aligned} \mathbb{P}(N_1 = n_1, N_2 = n_2, N_3 = n_3) &= \sum_{\substack{(x_1, \dots, x_n) \in \{1, 2, 3\}^n \\ \text{Card } \{i; x_i=1\} = n_1 \\ \text{Card } \{i; x_i=2\} = n_2}} q_1^{n_1} q_2^{n_2} q_3^{n - n_1 - n_2} \\ &= \frac{n!}{n_1! n_2! n_3!} q_1^{n_1} q_2^{n_2} q_3^{n_3}. \end{aligned}$$

6. La matrice de covariance de $Z_1 = (\mathbf{1}_{\{X_i=1\}}, \mathbf{1}_{\{X_i=2\}})$ est :

$$\begin{aligned} \Sigma &= \begin{pmatrix} \text{Var } \mathbf{1}_{\{X_1=1\}} & \text{Cov}(\mathbf{1}_{\{X_1=1\}}, \mathbf{1}_{\{X_1=2\}}) \\ \text{Cov}(\mathbf{1}_{\{X_1=1\}}, \mathbf{1}_{\{X_1=2\}}) & \text{Var } \mathbf{1}_{\{X_1=2\}} \end{pmatrix} \\ &= \begin{pmatrix} q_1(1 - q_1) & -q_1 q_2 \\ -q_1 q_2 & q_2(1 - q_2) \end{pmatrix}. \end{aligned}$$

Par indépendance, on a $\text{Cov}(N_1, N_2) = n \text{Cov}(\mathbf{1}_{\{X_1=1\}}, \mathbf{1}_{\{X_1=2\}}) = -n q_1 q_2$. 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 central limite assure que :

$$\sqrt{n} \left(\sum_{i=1}^n Z_i - n \mathbb{E}[Z_i] \right) = \left(\frac{N_1[n] - n q_1}{\sqrt{n}}, \frac{N_2[n] - n q_2}{\sqrt{n}} \right)$$

converge en loi, quand n tend vers l'infini, vers la loi gaussienne $\mathcal{N}(0, \Sigma)$.

III Estimation de θ à l'aide du génotype

1. On a $c = \log \frac{n!}{n_1! n_2! n_3!} + n_2 \log 2$.

2. On a $V_n = \frac{\partial L_n(N_1, N_2, N_3; \theta)}{\partial \theta} = \frac{2N_1}{\theta} + \frac{N_2}{\theta} - \frac{N_2}{1 - \theta} - \frac{2N_3}{1 - \theta}$.

3. On a :

$$\begin{aligned} I_n(\theta) &= -\mathbb{E}_\theta \left[\frac{\partial^2 L_n(N_1, N_2, N_3; \theta)}{\partial \theta^2} \right] \\ &= \mathbb{E}_\theta \left[\frac{2N_1}{\theta^2} + \frac{N_2}{\theta^2} + \frac{N_2}{(1-\theta)^2} + \frac{2N_3}{(1-\theta)^2} \right] = \frac{2n}{\theta(1-\theta)}. \end{aligned}$$

4. Si $(n_1, n_2) \neq (0, 0)$ et $(n_2, n_3) \neq (0, 0)$, alors on a :

$$\lim_{\theta \downarrow 0} L_n(n_1, n_2, n_3; \theta) = \lim_{\theta \uparrow 1} L_n(n_1, n_2, n_3; \theta) = -\infty.$$

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 : $\theta = \frac{n_1}{n} + \frac{n_2}{2n}$. Si $(n_1, n_2) = (0, 0)$, alors on a $L_n(n_1, n_2, n_3; \theta) = c + n \log \theta$. Le maximum est atteint pour $\theta = 0 = \frac{n_1}{n} + \frac{n_2}{2n}$. 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 $\theta = 1 = \frac{n_1}{n} + \frac{n_2}{2n}$. Dans tous les cas le maximum de la vraisemblance est atteint en un point unique $\theta = \frac{n_1}{n} + \frac{n_2}{2n}$. Donc $\hat{\theta}_n = \frac{N_1}{n} + \frac{N_2}{2n}$ est l'estimateur du maximum de vraisemblance de θ .

5. L'estimateur est sans biais car on déduit de II.2. que $\mathbb{E}_\theta [\hat{\theta}_n] = \theta^2 + \theta(1-\theta) = \theta$.

6. On a, grâce aux questions II.2 et II.6 :

$$\begin{aligned} \text{Var}(\hat{\theta}_n) &= \frac{1}{4n^2} \text{Var}(2N_1 + N_2) = \frac{1}{4n^2} (4 \text{Var}(N_1) + \text{Var}(N_2) + 4 \text{Cov}(N_1, N_2)) \\ &= \frac{1}{4n} (4q_1(1-q_1) + q_2(1-q_2) - 4q_1q_2) = \frac{\theta(1-\theta)}{2n}. \end{aligned}$$

La borne FCDD est $I_n(\theta)^{-1} = \frac{\theta(1-\theta)}{2n}$. L'estimateur est donc efficace. On peut remarquer sur l'équation (XII.3) 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 $\hat{\theta}_n = \sum_{i=1}^n (1, \frac{1}{2}) \cdot Z_i$, on déduit du II.7 que $\hat{\theta}_n$ est asymptotiquement normal de variance asymptotique $\lim_{n \rightarrow \infty} n \text{Var}(\hat{\theta}_n) = \frac{\theta(1-\theta)}{2}$.

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. On a ζ_n égal à la statistique $\zeta_n^{(1)}$ du test du χ^2 . Le test du χ^2 assure que le test de région critique $W_n^{(1)} = \{\zeta_n \geq z\}$ est un test convergent de niveau asymptotique $\alpha = \mathbb{P}(Z \geq z)$, où Z a pour loi la loi d'un χ^2 à 3-1-1=1 degré de liberté. On a enlevé un degré de liberté pour l'estimation de $\theta \in]0, 1[\subset \mathbb{R}$.
3. On a $n \simeq 3.88 \cdot 10^6$, $\hat{\theta}_n \simeq 1.9853 \cdot 10^{-2}$ et les valeurs suivantes :

$$\begin{aligned} n_1 &= 1580, & \hat{q}_1 &\simeq 4.0722 \cdot 10^{-4}, & q_1(\hat{\theta}_n) &\simeq 3.9415 \cdot 10^{-4}, \\ n_2 &= 150900, & \hat{q}_2 &\simeq 3.8892 \cdot 10^{-2}, & q_2(\hat{\theta}_n) &\simeq 3.8918 \cdot 10^{-2}, \\ n_3 &= 1580, & \hat{q}_3 &\simeq 9.6070 \cdot 10^{-1}, & q_3(\hat{\theta}_n) &\simeq 9.6068 \cdot 10^{-1}. \end{aligned}$$

On obtient $\zeta_n^{(1)} \simeq 1.69$. Pour $\alpha = 5\%$, on a $z \simeq 3.84$. Comme $\zeta_n^{(1)} \leq 3.84$, on accepte donc H_0 au seuil de 5%. La p -valeur asymptotique de ce test est $p\text{-val} = 0.19$. Ce qui confirme que l'on ne peut pas rejeter H_0 .

4. On obtient $\zeta_n^{(1)} \simeq 1163$. Comme $\zeta_n^{(1)} \geq 3.84$, on rejette donc H_0 au seuil de 5% (la p -valeur asymptotique est de l'ordre de 10^{-255}). 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.2. I Modélisation

1. Les variables aléatoires X_1, \dots, X_n sont indépendantes de loi de Bernoulli de paramètre p .
2. La variable aléatoire est discrète. Sa densité est donnée par $p(x_1; p) = \mathbb{P}_p(X_1 = x_1)$, avec $x_1 \in \{0, 1\}$. La densité de l'échantillon est :

$$p_n(x; p) = \left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i} (1-p)^n, \quad \text{où } x = (x_1, \dots, x_n) \in \{0, 1\}^n.$$

3. On remarque 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.

4. Pour que $h(S_n)$ soit intégrable, il faut et il suffit que h , définie sur $\{0, \dots, n\}$, 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 quand p tend vers 0, il vient $\lim_{p \rightarrow 0} \mathbb{E}_p[h(S_n)] = h(0)$. Soit $\delta = h(S_n)$ un estimateur intégrable sans biais de $1/p$. On a donc $\mathbb{E}_p[h(S_n)] = 1/p$. Quand p tend vers 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 .

On peut également démontrer directement que la statistique S_n est totale. En effet $p \mapsto \mathbb{E}_p[h(S_n)]$ est un polynôme en p de degré au plus n . S'il s'annule pour tout $p \in]0, 1[$, alors ses coefficients sont tous nuls. Ceci implique que $h(k) = 0$ pour tout $k \in \{0, \dots, n\}$ et donc \mathbb{P}_p p.s. on a $h(S_n) = 0$.

5. 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 $\lim_{n \rightarrow \infty} n/(n+1) = 1$ et $\lim_{n \rightarrow \infty} 1/(n+1) = 0$, on en déduit que \mathbb{P}_p -p.s. $\lim_{n \rightarrow \infty} (S_n + 1)/(n + 1) = p$. Comme la fonction $g(x) = 1/x$ est continue en p , on en déduit que \mathbb{P}_p -p.s. $\lim_{n \rightarrow \infty} \delta_n = \lim_{n \rightarrow \infty} (n + 1)/(S_n + 1) = p$. La suite d'estimateurs est donc convergente.
2. On a :

$$\begin{aligned} \mathbb{E}_p \left[\frac{n+1}{S_n+1} \right] &= \sum_{k=0}^n \frac{n!}{k!(n-k)!} \frac{n+1}{k+1} p^k (1-p)^{n-k} \\ &= \frac{1}{p} \sum_{k=0}^n \frac{(n+1)!}{(k+1)!(n-k)!} p^{k+1} (1-p)^{n-k} \\ &= \frac{1}{p} \sum_{j=0}^{n+1} \frac{(n+1)!}{j!(n+1-j)!} p^j (1-p)^{n+1-j} - \frac{1}{p} (1-p)^{n+1} \\ &= \frac{1}{p} (1 - (1-p)^{n+1}). \end{aligned}$$

Le biais est $b_n = -(1-p)^{n+1}/p$. On remarque que $-pb_n = (1 - \frac{n_0}{N})^{n+1} \simeq e^{-nn_0/N}$ si $n_0 \ll N$.

3. Les variables aléatoires $(X_i, i \in \mathbb{N}^*)$ 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 $s_p G_0$, où G_0 est de loi gaussienne $\mathcal{N}(0, 1)$ et $s_p^2 = \text{Var}_p(X_1) = p(1-p)$.

4. On a :

$$0 \leq \sqrt{n} \left(\frac{S_n + 1}{n + 1} - \frac{S_n}{n}\right) = \frac{n - S_n}{\sqrt{n}(n + 1)} \leq \frac{\sqrt{n}}{n + 1} \leq \frac{1}{\sqrt{n}}.$$

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 $(s_p G_0, 0)$. Par continuité, on en déduit que :

$$\sqrt{n} \left(\frac{S_n + 1}{n + 1} - p\right) = \sqrt{n} \left(\frac{S_n}{n} - p\right) + \sqrt{n} \left(\frac{S_n + 1}{n + 1} - \frac{S_n}{n}\right) \xrightarrow[n \rightarrow \infty]{\text{en loi}} s_p G_0.$$

La suite $(1/\delta_n, n \in \mathbb{N}^*)$ est donc une suite d'estimateurs de p convergente et asymptotiquement normale de variance asymptotique s_p^2 .

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 $\left(\sqrt{n} \left(\delta_n - \frac{1}{p}\right), n \in \mathbb{N}^*\right)$ converge en loi $\sigma_p G_0$ où $\sigma_p^2 = s_p^2 (g'(p))^2 = (1-p)/p^3$. La suite $(\delta_n, n \in \mathbb{N}^*)$ est une suite d'estimateurs de $1/p$ convergente et asymptotiquement normale de variance asymptotique σ_p^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 de l'échantillon de taille 1 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 :

$$I(p) = \text{Var}_p(V_1) = \text{Var}_p(X_1/p(1-p)) = \frac{1}{p^2(1-p)^2} p(1-p) = \frac{1}{p(1-p)}.$$

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 $(g'(p))^2/I(p) = (1-p)/p^3$. La variance asymptotique de l'estimateur δ_n 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 $(\delta_n, n \in \mathbb{N}^*)$ converge \mathbb{P}_p -p.s. vers $1/p$, on en déduit que :

$$\hat{\sigma}_n = \delta_n \sqrt{\delta_n - 1} \xrightarrow[n \rightarrow \infty]{\mathbb{P}_p\text{-p.s.}} \sigma_p.$$

On déduit du théorème de Slutsky que la suite $((\hat{\sigma}_n, \sqrt{n}(\delta_n - 1/p)), n \in \mathbb{N}^*)$ converge en loi vers $(\sigma_p, \sigma_p G_0)$, où G_0 est de loi gaussienne $\mathcal{N}(0, 1)$. Par continuité, on en déduit que :

$$\frac{\sqrt{n}}{\hat{\sigma}_n} \left(\delta_n - \frac{1}{p} \right) \xrightarrow[n \rightarrow \infty]{\text{en loi}} G_0,$$

où G_0 est de loi gaussienne $\mathcal{N}(0, 1)$. Soit $z > 0$. Comme G_0 est une variable aléatoire continue, on a :

$$\mathbb{P}_p \left(\frac{1}{p} \in \left[\delta_n \pm z \frac{\hat{\sigma}_n}{\sqrt{n}} \right] \right) = \mathbb{P} \left(\frac{\sqrt{n}}{\hat{\sigma}_n} \left(\delta_n - \frac{1}{p} \right) \in [-z, z] \right) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(G_0 \in [-z, z]).$$

Pour $z = \phi_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, on en déduit que $I_n = \left[\delta_n \pm \phi_{1-\alpha/2} \frac{\hat{\sigma}_n}{\sqrt{n}} \right]$ est un intervalle de confiance de niveau asymptotique $1 - \alpha$.

2. L'estimateur de $1/p$ est $\delta_n = (n + 1)/(S_n + 1) = 463/341 \simeq 1.36$. L'estimateur de N est $n_0 \delta_n \simeq 100$.
3. Pour $\alpha = 5\%$, on a $\phi_{1-\alpha/2} \simeq 1.96$. On a $\hat{\sigma}_n \simeq 0.81$ et $I_n \simeq [1.28; 1.43]$. 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 :

$$\begin{aligned} \varphi(x) &= 1 & \text{si } S_n(x) < c, \\ \varphi(x) &= \gamma & \text{si } S_n(x) = c, \\ \varphi(x) &= 0 & \text{si } S_n(x) > c, \end{aligned}$$

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 XII.1, on en déduit que $c = 341$. Et on obtient $\gamma \simeq 0.60$. Comme $s_n = 340 < 341$, on rejette donc l'hypothèse H_0 au niveau 5%.
Le même test permet en fait de rejeter l'hypothèse $N \leq 96$, car 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 (car on considère un test unilatéral mais un intervalle de confiance bilatéral). Il est également plus simple à calculer, dans la mesure où il ne fait pas intervenir la variance asymptotique σ_p^2 , et donc ne nécessite pas son estimation. Enfin, le test est de niveau exact alors que l'intervalle de confiance est de niveau asymptotique.

▲

Exercice XII.3. I L'estimateur du maximum de vraisemblance de θ

1. On a $f_\theta(t) = -\frac{\partial \bar{F}_\theta(t)}{\partial t} = \theta k(t-w)^{k-1} e^{-\theta(t-w)^k} \mathbf{1}_{\{t>w\}}$.
2. Comme les variables sont indépendantes, la densité de l'échantillon est :

$$p_n^*(t_1, \dots, t_n; \theta) = \theta^n k^n \left(\prod_{i=1}^n (t_i - w)^{k-1} \mathbf{1}_{\{t_i > w\}} \right) e^{-\theta \sum_{i=1}^n (t_i - w)^k}.$$

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 $\frac{\partial^2 L_n^*(t; \theta)}{\partial \theta^2} = -n/\theta^2 < 0$, on en déduit que la log-vraisemblance est concave ; elle est donc 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 $\alpha > 0$:

$$\mathbb{E}_\theta[(T - w)^{\alpha k}] = \int (t - w)^{\alpha k} f_\theta(t) dt = \int_w^\infty \theta k (t - w)^{\alpha k + k - 1} e^{-\theta(t-w)^k} dt.$$

En posant $y = \theta(t - w)^k$, il vient :

$$\mathbb{E}_\theta[(T - w)^{\alpha k}] = \theta^{-\alpha} \int_0^\infty y^\alpha e^{-y} dy = \theta^{-\alpha} \Gamma(\alpha + 1).$$

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 p.s. vers $\mathbb{E}_\theta[(T - w)^k] = \theta^{-1}$. On considère la fonction $h : x \mapsto 1/x$. On a $\hat{\theta}_n^* = h(Z_n)$. Comme la fonction h est continue en $\theta^{-1} > 0$, on en déduit que la suite $(\hat{\theta}_n^*, n \geq 1)$ converge p.s. vers $h(\theta^{-1}) = \theta$. L'estimateur du maximum de vraisemblance est donc convergent. Comme les variables aléatoires $(T_1 - w)^k, \dots, (T_n - w)^k$ sont de carré intégrable, on déduit du théorème central limite 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 $\sigma^2 = \text{Var}_\theta((T - w)^k) = \theta^{-2}$. Comme la fonction h est de classe C^1 en $\theta^{-1} > 0$, on en déduit que la suite $(\sqrt{n}(\hat{\theta}_n^* - \theta), n \geq 1)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \Sigma_\theta^2)$, où $\Sigma_\theta^2 = \sigma^2 h'(\theta^{-1})^2 = \theta^2$. L'estimateur du maximum de vraisemblance est donc asymptotiquement normal de variance asymptotique Σ_θ .

7. L'estimateur du maximum de vraisemblance est asymptotiquement efficace, car $\Sigma_\theta^2 = I(\theta)^{-1}$.

8. On remarque que $(\hat{\theta}_n^*)^2$ est un estimateur convergent de Σ_θ^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 G_0 de loi gaussienne $\mathcal{N}(0, 1)$. Soit $z > 0$. Comme G_0 est une variable aléatoire continue, on a :

$$\mathbb{P}_\theta \left(\theta \in \left[\hat{\theta}_n^* \pm z \frac{\hat{\theta}_n^*}{\sqrt{n}} \right] \right) = \mathbb{P} \left(\sqrt{n}(\hat{\theta}_n^* - \theta)/\hat{\theta}_n^* \in [-z, z] \right) \xrightarrow{n \rightarrow \infty} \mathbb{P}(G_0 \in [-z, z]).$$

Pour $z = \phi_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$, on en déduit que :

$$J_n^* = \left[\hat{\theta}_n^* \pm \phi_{1-\alpha/2} \frac{\hat{\theta}_n^*}{\sqrt{n}} \right]$$

est un intervalle de confiance de niveau asymptotique $1 - \alpha$. Pour $\alpha = 5\%$, on a $\phi_{1-\alpha/2} \simeq 1.96$. Il vient $\hat{\theta}_n^* = 17/33\ 175\ 533 \simeq 5.124 \cdot 10^{-7}$ et

$$J_n^* \simeq [5.124 \cdot 10^{-7} \pm 2.46 \cdot 10^{-7}] = [2.7 \cdot 10^{-7}, 7.6 \cdot 10^{-7}].$$

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 $x = 1$:

$$\begin{aligned} \mathbb{P}_\theta(R > r, X = 1) &= \mathbb{P}_\theta(S \geq T > r) = \int \mathbf{1}_{\{s \geq t > r\}} f_\theta(t) g(s) dt ds \\ &= \int_r^\infty f_\theta(t) \bar{G}(t) dt, \end{aligned}$$

et pour $x = 0$:

$$\begin{aligned} \mathbb{P}_\theta(R > r, X = 0) &= \mathbb{P}_\theta(T > S > r) = \int \mathbf{1}_{\{t > s > r\}} f_\theta(t) g(s) dt ds \\ &= \int_r^\infty g(t) \bar{F}_\theta(t) dt. \end{aligned}$$

On en déduit donc par dérivation que :

$$p(r, 1; \theta) = f_\theta(r) \bar{G}(r) = \theta e^{-\theta(r-w)_+^k} c(r, 1) \quad \text{avec} \quad c(r, 1) = k(w - r)_+^{k-1} \bar{G}(r),$$

et

$$p(r, 0; \theta) = g(r) \bar{F}_\theta(r) = e^{-\theta(r-w)_+^k} c(r, 0) \quad \text{avec} \quad c(r, 0) = g(r).$$

La fonction c est bien indépendante de θ .

3. La variable aléatoire 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 $r = (r_1, \dots, r_n) \in \mathbb{R}^n$, $x = (x_1, \dots, x_n) \in \{0, 1\}^n$,

$$p_n(r, x; \theta) = \prod_{i=1}^n p(r_i, x_i; \theta) = \theta^{\sum_{i=1}^n x_i} e^{-\theta \sum_{i=1}^n (r_i - w)_+^k} c_n(r, x),$$

où $c_n(r, x) = \prod_{i=1}^n c(r_i, x_i)$ ne dépend pas 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 question I.3, 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 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. On remarque également que la loi de S n'intervient pas dans $\hat{\theta}_n$. En particulier, il n'est pas nécessaire de connaître explicitement la fonction g ou \bar{G} .

5. Les variables aléatoires X_1, \dots, X_n sont indépendantes intégrables et de même loi de Bernoulli de paramètre p . On déduit de la loi forte des grands nombres 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 p.s. vers $h(p, \theta) = p/\theta^2 = I(\theta)$. L'estimateur $\frac{N_n}{n} \frac{1}{\hat{\theta}_n^2}$ 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 $\mathcal{N}(0, 1/I(\theta))$. 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 raisonnant comme dans la question I.8, on déduit que :

$$J_n = \left[\hat{\theta}_n \pm \phi_{1-\alpha/2} \frac{\hat{\theta}_n}{\sqrt{N_n}} \right]$$

est un intervalle de confiance de niveau asymptotique $1 - \alpha$. Pour $\alpha = 5\%$, on a $\phi_{1-\alpha/2} \simeq 1.96$. Il vient $N_n = 17$, $\hat{\theta}_n = 17/(33\ 175\ 533 + 4\ 546\ 880) \simeq 4.507\ 10^{-7}$ et :

$$J_n \simeq [4.507\ 10^{-7} \pm 2.142\ 10^{-7}] = [2.4\ 10^{-7}, 6.6\ 10^{-7}].$$

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.

III Comparaison de traitements

1. Par indépendance, la densité de l'échantillon p_{n^A, n^B} est le produit des intensités. Z_1 . Pour $r^A = (r_1^A, \dots, r_{n^A}^A) \in \mathbb{R}^{n^A}$, $r^B = (r_1^B, \dots, r_{n^B}^B) \in \mathbb{R}^{n^B}$, et $x^A = (x_1^A, \dots, x_{n^A}^A) \in \{0, 1\}^{n^A}$ et $x^B = (x_1^B, \dots, x_{n^B}^B) \in \{0, 1\}^{n^B}$, on a :

$$\begin{aligned} p_{n^A, n^B}(r^A, x^A, r^B, x^B; \theta^A, \theta^B) \\ = (\theta^A)^{\sum_{i=1}^{n^A} x_i^A} (\theta^B)^{\sum_{i=1}^{n^B} x_i^B} e^{-\theta^A \sum_{i=1}^{n^A} (r_i^A - w)_+^k - \theta^B \sum_{i=1}^{n^B} (r_i^B - w)_+^k} \\ c_{n^A}(r^A, x^A) c_{n^B}(r^B, x^B). \end{aligned}$$

2. On cherche θ qui maximise $\theta \mapsto p_{n^A, n^B}(r^A, x^A, r^B, x^B; \theta, \theta)$. Des calculs similaires à ceux de la question II.4 assurent que l'estimateur du maximum de vraisemblance de θ est :

$$\hat{\theta}_{n^A, n^B} = \frac{N_{n^A}^A + N_{n^B}^B}{\sum_{i=1}^{n^A} (R_i^A - w)_+^k + \sum_{i=1}^{n^B} (R_i^B - w)_+^k}.$$

3. On déduit de la question II.7 que, pour $j \in \{A, B\}$, l'estimateur $N_{n_j}^j / (n_j \hat{\theta}_{n_j}^j)$ est un estimateur convergent de $I(\theta^j)$. Les deux statistiques de test du test de Hausman sont :

$$\begin{aligned} \zeta_{n^A, n^B}^{(1)} &= n^A (\hat{\theta}_{n^A}^A - \hat{\theta}_{n^A, n^B})^2 \frac{N_{n^A}^A}{n^A (\hat{\theta}_{n^A}^A)^2} + n^B (\hat{\theta}_{n^B}^B - \hat{\theta}_{n^A, n^B})^2 \frac{N_{n^B}^B}{n^B (\hat{\theta}_{n^B}^B)^2} \\ &= N_{n^A}^A \frac{(\hat{\theta}_{n^A}^A - \hat{\theta}_{n^A, n^B})^2}{(\hat{\theta}_{n^A}^A)^2} + N_{n^B}^B \frac{(\hat{\theta}_{n^B}^B - \hat{\theta}_{n^A, n^B})^2}{(\hat{\theta}_{n^B}^B)^2}, \\ \zeta_{n^A, n^B}^{(2)} &= n^A (\hat{\theta}_{n^A}^A - \hat{\theta}_{n^A, n^B})^2 \frac{N_{n^A}^A}{n^A (\hat{\theta}_{n^A, n^B})^2} + n^B (\hat{\theta}_{n^B}^B - \hat{\theta}_{n^A, n^B})^2 \frac{N_{n^B}^B}{n^B (\hat{\theta}_{n^A, n^B})^2} \\ &= N_{n^A}^A \frac{(\hat{\theta}_{n^A}^A - \hat{\theta}_{n^A, n^B})^2}{(\hat{\theta}_{n^A, n^B})^2} + N_{n^B}^B \frac{(\hat{\theta}_{n^B}^B - \hat{\theta}_{n^A, n^B})^2}{(\hat{\theta}_{n^A, n^B})^2}. \end{aligned}$$

Les régions critiques associées au test de niveau asymptotique α sont $W_{n^A, n^B}^{(j)} = \{\zeta_{n^A, n^B}^{(j)} \geq z_{1-\alpha}\}$, pour $j \in \{1, 2\}$, où $z_{1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi du χ^2 à 1 degré de liberté. Les tests sont convergents (quand $\min(n^A, n^B)$ tend vers l'infini) et les p -valeurs asymptotiques sont, $j \in \{1, 2\}$:

$$p^{(j)\text{-val}} = \mathbb{P}(Z \geq \zeta_{n^A, n^B}^{(j), \text{obs}}),$$

où Z suit une loi du χ^2 à 1 degré de liberté.

4. On a déjà calculé $\hat{\theta}_{n^A}^A \simeq 4.507 \cdot 10^{-7}$. Il vient :

$$\hat{\theta}_{n^B}^B = \frac{19}{64\,024\,591 + 15\,651\,648} \simeq 2.385 \cdot 10^{-7} \quad \text{et} \quad \hat{\theta}_{n^A, n^B} \simeq 3.066 \cdot 10^{-7}.$$

Le quantile à 95% du χ^2 à 1 degré de liberté est $z_{1-\alpha} = 3.84$. On obtient $\zeta_n^{(1)} \simeq 3.3$ et $\zeta_n^{(2)} \simeq 4.7$ ainsi que $p^{(1)\text{-val}} \simeq 7\%$ et $p^{(2)\text{-val}} \simeq 3\%$. (On remarque que $\zeta_n^{(2)}$ est dans la région critique mais pas $\zeta_n^{(1)}$.) On ne peut donc pas rejeter H_0 au niveau de confiance de 95%. Les pré-traitements A et B ne sont pas significativement différents.

5. Pour $j \in \{A, B\}$, on déduit de la partie précédente que l'intervalle de confiance de niveau asymptotique $1 - \alpha^j$ pour θ^j est $J_{n^j}^j = \left[\hat{\theta}_{n^j}^j \pm \phi_{1-\alpha^j/2} \frac{\hat{\theta}_{n^j}^j}{\sqrt{N_{n^j}^j}} \right]$. Les

variables aléatoires $(\hat{\theta}_{n^A}^A, N_{n^A})$ et $(\hat{\theta}_{n^B}^B, N_{n^B})$ sont indépendantes. On en déduit donc que :

$$\mathbb{P}_{(\theta^A, \theta^B)}(\theta^A \in J_{n^A}^A, \theta^B \in J_{n^B}^B) = \mathbb{P}_{\theta^A}(\theta^A \in J_{n^A}^A) \mathbb{P}_{\theta^B}(\theta^B \in J_{n^B}^B) \xrightarrow{\min(n^A, n^B) \rightarrow \infty} (1 - \alpha^A)(1 - \alpha^B).$$

Pour tout $(\alpha^A, \alpha^B) \in]0, 1]^2$ tel que $(1 - \alpha^A)(1 - \alpha^B) = 1 - \alpha$, on obtient des intervalles de confiance $J_{n^A}^A$ pour θ^A et $J_{n^B}^B$ pour θ^B tels que le niveau asymptotique sur les deux intervalles soit de $1 - \alpha$. On peut choisir $\alpha' = \alpha^A = \alpha^B = 1 - \sqrt{1 - \alpha}$. Pour $\alpha = 5\%$, il vient $\alpha' \simeq 2.5\%$, $\phi_{1-\alpha'/2} \simeq 2.24$ et

$$J_{n^A}^A \simeq [2 \cdot 10^{-7}, 7 \cdot 10^{-7}] \quad \text{et} \quad J_{n^B}^B \simeq [1.2 \cdot 10^{-7}, 3.6 \cdot 10^{-7}].$$

Les deux intervalles $J_{n^A}^A$ et $J_{n^B}^B$ 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 $p = \mathbb{P}_\theta(T \leq S) = \int \mathbf{1}_{\{t \leq s\}} f_\theta(t) g(s) dt ds = \int f_\theta(t) \bar{G}(t) dt$.
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 :

$$\begin{aligned} \int (r - w)_+^k g(r) \bar{F}_\theta(r) dr &= \left[-\bar{G}(r) (r - w)_+^k F_\theta(r) \right]_{-\infty}^{+\infty} \\ &\quad + \int k (r - w)_+^{k-1} \bar{F}_\theta(r) \bar{G}(r) dr \\ &\quad - \int (r - w)_+^k f_\theta(r) \bar{G}(r) dr \\ &= \theta^{-1} \int f_\theta(r) \bar{G}(r) dr - \int (r - w)_+^k f_\theta(r) \bar{G}(r) dr, \end{aligned}$$

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 $(R_1 - w)_+^k, \dots, (R_n - w)_+^k$ 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 p.s. vers $\mathbb{E}_\theta[(R - w)_+^k] = p/\theta$. De plus

la suite $(N_n/n, n \geq 1)$ converge p.s. 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 p.s. 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 :

$$\begin{aligned} \int (r-w)_+^{2k} g(r) \bar{F}_\theta(r) dr &= \left[-\bar{G}(r)(r-w)_+^{2k} F_\theta(r) \right]_{-\infty}^{+\infty} \\ &\quad + \int 2k(r-w)_+^{2k-1} \bar{F}_\theta(r) \bar{G}(r) dr \\ &\quad - \int (r-w)_+^{2k} f_\theta(r) \bar{G}(r) dr \\ &= 2\theta^{-1} \int (r-w)_+^k f_\theta(r) \bar{G}(r) dr \\ &\quad - \int (r-w)_+^{2k} f_\theta(r) \bar{G}(r) dr. \end{aligned}$$

Comme

$$\begin{aligned} \beta = \mathbb{E}_\theta[\mathbf{1}_{\{X=1\}}(R-w)_+^k] &= \int \mathbf{1}_{\{t \leq s\}} (t-w)_+^k f_\theta(t) g(s) ds dt \\ &= \int (t-w)_+^k f_\theta(t) \bar{G}(t) dt, \end{aligned}$$

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 $((R_1-w)_+^k, X_1), \dots, ((R_n-w)_+^k, X_n)$ sont indépendantes de même loi et de carré intégrable. On déduit du théorème central limite 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 $\mathcal{N}(0, \sigma_\theta^2)$, où :

$$\begin{aligned}
\sigma_\theta^2 &= \left(\frac{\partial h}{\partial x}(p, p/\theta), \frac{\partial h}{\partial z}(p, p/\theta) \right) \Delta \left(\frac{\partial h}{\partial x}(p, p/\theta), \frac{\partial h}{\partial z}(p, p/\theta) \right)^t \\
&= (\theta/p, -\theta^2/p) \begin{pmatrix} p(1-p) & \beta - \frac{p^2}{\theta} \\ \beta - \frac{p^2}{\theta} & \frac{2\beta}{\theta} - \frac{p^2}{\theta^2} \end{pmatrix} (\theta/p, -\theta^2/p)^t \\
&= \theta^2/p.
\end{aligned}$$

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 $1/I(\theta) = \sigma_\theta^2$, on en déduit que l'estimateur $\hat{\theta}_n$ est asymptotiquement efficace. ▲

Exercice XII.4. 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, \dots, Y_n) est un vecteur gaussien car les variables aléatoires Y_1, \dots, Y_n sont gaussiennes et indépendantes. Comme $\hat{\nu}_n$ est une transformation linéaire du vecteur (Y_1, \dots, Y_n) , sa loi est une loi gaussienne. On a $\mathbb{E}_\nu[\hat{\nu}_n] = \nu$ et $\text{Var}_\nu(\hat{\nu}_n) = \frac{\sigma_0^2}{n}$. La loi de $\hat{\nu}_n$ est donc la loi $\mathcal{N}(\nu, \sigma_0^2/n)$. En particulier $\sqrt{n} \frac{\hat{\nu}_n - \nu}{\sigma_0}$ a même loi que G_0 de loi gaussienne centrée réduite $\mathcal{N}(0, 1)$. On en déduit que :

$$\mathbb{P} \left(\sqrt{n} \frac{\hat{\nu}_n - \nu}{\sigma_0} \in [-\phi_{1-\alpha/2}, \phi_{1-\alpha/2}] \right) = 1 - \alpha,$$

où ϕ_r est le quantile d'ordre r de la loi gaussienne centrée réduite $\mathcal{N}(0, 1)$. Donc l'intervalle :

$$I_n = \left[\hat{\nu}_n \pm \phi_{1-\alpha/2} \frac{\sigma_0}{\sqrt{n}} \right]$$

est un intervalle de confiance exact de ν de niveau $1 - \alpha$.

Pour $\alpha = 5\%$, on a $\phi_{1-\alpha/2} \simeq 1.96$. Il vient $\hat{\nu}_n \simeq 6.42$ et $I_n \simeq [4.38, 8.46]$.

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, pour $x = (x_1, \dots, x_m) \in \mathbb{R}^m$ et $y = (y_1, \dots, y_n) \in \mathbb{R}^n$:

$$p(x, y; \mu, \nu) = \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(x_i - \mu)^2 / 2\sigma_0^2} \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(y_j - \nu)^2 / 2\sigma_0^2}$$

$$L(x, y; \mu, \nu) = -\frac{m+n}{2} \log(2\pi\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{1}{2\sigma_0^2} \sum_{j=1}^n (y_j - \nu)^2.$$

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 $\hat{\mu}_m(x) = \frac{1}{m} \sum_{i=1}^m x_i$

et $\hat{\nu}_n(y) = \frac{1}{n} \sum_{j=1}^n y_j$. On en déduit que les estimateurs du maximum de vrai-

semblance de (μ, ν) sont $\hat{\mu}_m = \frac{1}{m} \sum_{i=1}^m X_i$ et $\hat{\nu}_n = \frac{1}{n} \sum_{j=1}^n Y_j$.

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 I.2, assurent que la loi de $\hat{\mu}_m$ est la loi $\mathcal{N}(\mu, \sigma_0^2/m)$. 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 $\zeta_{m,n}^{(1)}$ est une transformation linéaire du vecteur gaussien $(\hat{\mu}_m, \hat{\nu}_n)$. Il s'agit donc d'une variable aléatoire gaussienne. On obtient :

$$\mathbb{E}_{\mu, \nu}[\zeta_{m,n}^{(1)}] = \sqrt{\frac{mn}{m+n}} \frac{\mu - \nu}{\sigma_0},$$

et en utilisant l'indépendance de $\hat{\mu}_m$ et $\hat{\nu}_n$:

$$\text{Var}_{\mu, \nu}(\zeta_{m,n}^{(1)}) = \frac{mn}{m+n} \frac{\text{Var}_{\mu, \nu}(\hat{\mu}_m) + \text{Var}_{\mu, \nu}(\hat{\nu}_n)}{\sigma_0^2} = 1.$$

En particulier, sous H_0 , la statistique de test $\zeta_{m,n}^{(1)}$ a même loi que G_0 .

8. De même qu'à la question I.3, l'estimateur $\hat{\mu}_m$ est un estimateur convergent de μ . On remarque 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, on a :

$$\lim_{\min(m, n) \rightarrow \infty} \frac{mn}{m+n} = +\infty.$$

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 p.s. vers $(\mu, \nu, +\infty)$ quand $\min(m, n)$ tend vers l'infini. Si $\mu > \nu$, alors p.s. on a :

$$\lim_{\min(m, n) \rightarrow \infty} \zeta_{m, n}^{(1)} = +\infty.$$

9. Avec les notations de la question I.5, on définit la fonction sur \mathbb{R}^{m+n} :

$$\zeta_{m, n}^{(1)}(xy) = \sqrt{\frac{mn}{m+n}} \frac{\hat{\mu}_m(x) - \hat{\nu}_n(y)}{\sigma_0}.$$

En particulier, on a $\zeta_{m, n}^{(1)}(X_1, \dots, X_m, Y_1, \dots, Y_n) = \zeta_{m, n}^{(1)}$. On considère le test pur de région critique :

$$W_{m, n} = \{(x, y) \in \mathbb{R}^m \times \mathbb{R}^n; \zeta_{m, n}^{(1)}(x, y) \geq z\}.$$

Comme sous H_0 , $\zeta_{m, n}^{(1)}$ a même loi que G_0 , l'erreur de première espèce est donc :

$$\mathbb{P}_{\mu, \nu}(W_{m, n}) = \mathbb{P}_{\mu, \nu}(\zeta_{m, n}^{(1)} > z) = \mathbb{P}(G_0 \geq z).$$

Le test est de niveau exact α avec $z = \phi_{1-\alpha}$, le quantile d'ordre $1 - \alpha$ de la loi $\mathcal{N}(0, 1)$. On vérifie que le test est convergent. D'après la question précédente, par convergence dominée on a sous H_1 :

$$\lim_{\min(m, n) \rightarrow \infty} \mathbb{P}_{\mu, \nu}(W_{m, n}) = \lim_{\min(m, n) \rightarrow \infty} \mathbb{E}_{\mu, \nu}[\mathbf{1}_{\{\zeta_{m, n}^{(1)} \geq \phi_{1-\alpha}\}}] = 1.$$

Le test est donc convergent. La p -valeur du test est $p\text{-val}_1 = \mathbb{P}(G_0 \geq \zeta_{m, n}^{(1), \text{obs}})$, où $\zeta_{m, n}^{(1), \text{obs}} = \zeta_{m, n}^{(1)}(x, y)$ est la statistique de test évaluée en les observations.

10. Pour $\alpha = 5\%$, on a $\phi_{1-\alpha} \simeq 1.64$. Il vient $\hat{\mu}_m \simeq 6.71$, $\hat{\nu}_n \simeq 6.42$, $\zeta_{m, n}^{(1), \text{obs}} \simeq 0.187$ et $p\text{-val}_1 \simeq 0.426$. En particulier, on accepte H_0 au niveau de 5%.

II Modèle non paramétrique de décalage

On fait quelques remarques préliminaires sur les résultats admis dans l'énoncé. Sous H_0 , on a $p = \int \mathbf{1}_{\{y \leq x\}} f(x)f(y) dx dy = 1/2$. On vérifie que sous H_0 la loi de $U_{m,n}$ ne dépend pas de F . Soit a_r le quantile d'ordre r de F . Comme F est continue, on a $F(a_r) = r$. Soit X de fonction de répartition F . Comme a_r est un point de croissance à gauche de F , on a $\{X < a_r\} = \{F(X) < r\}$. On en déduit que pour $r \in]0, 1[$:

$$r = \mathbb{P}(X \leq a_r) = \mathbb{P}(X < a_r) = \mathbb{P}(F(X) < F(a_r)) = \mathbb{P}(F(X) < r),$$

où l'on a utilisé la définition du quantile pour la première égalité et la continuité de F pour la deuxième. On en déduit que $F(X)$ suit la loi uniforme sur $[0, 1]$. Soit Y de fonction de répartition F et indépendant de X . On a par croissance de F que $\{Y \leq X\} \subset \{F(Y) \leq F(X)\}$ et $\{F(Y) < F(X)\} \subset \{Y < X\} \subset \{Y \leq X\}$. D'autre part comme $(F(Y), F(X))$ a même loi qu'un couple de variables aléatoires de loi uniforme sur $[0, 1]$ et indépendantes, on en déduit que p.s. $\mathbf{1}_{\{F(Y)=F(X)\}} = 0$. Ainsi p.s. on a $\mathbf{1}_{\{Y \leq X\}} = \mathbf{1}_{\{F(Y) \leq F(X)\}}$. On en déduit donc que $U_{m,n}$ a même loi que $\sum_{i=1}^m \sum_{j=1}^n \mathbf{1}_{\{V_j \leq W_i\}}$ où $(V_j, W_i, i \geq 1, j \geq 1)$ sont des variables aléatoires indépendantes de loi uniforme sur $[0, 1]$. En particulier, sous H_0 , la loi de $U_{m,n}$ ne dépend pas de F .

On vérifie maintenant 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 $x \in [x_0 - \varepsilon, x_0]$:

$$F(x + \rho) = F(x) + \int_{[x, x+\rho]} f(x') dx' \geq F(x) + \int_{[x_0, x_0+\varepsilon]} f(x') dx' > F(x).$$

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

$$\begin{aligned} p &= \int \mathbf{1}_{\{y \leq x\}} f(x)g(y) dx dy = \int \mathbf{1}_{\{y \leq x\}} f(x)f(y + \rho) dx dy \\ &= \int F(x + \rho)f(x) dx \\ &> \int F(x)f(x) dx = \frac{1}{2}. \end{aligned}$$

On en déduit donc que les valeurs possibles de p sous H_1 sont $] \frac{1}{2}, 1]$.

Enfin, on renvoie à la correction de l'exercice XI.7 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 :

$$\begin{aligned} \{\zeta_{m,n}^{(2)} > a\} &= \left\{ \frac{U_{m,n} - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} > a \right\} \\ &= \left\{ \frac{U_{m,n} - mnp}{\sqrt{\text{Var}(U_{m,n})}} > \sqrt{\frac{mn(m+n+1)}{12 \text{Var}(U_{m,n})}} a - \frac{mn(p - \frac{1}{2})}{\sqrt{\text{Var}(U_{m,n})}} \right\}. \end{aligned}$$

On pose :

$$b_{m,n} = \sqrt{\frac{mn(m+n+1)}{12 \text{Var}(U_{m,n})}} a - \frac{mn(p - \frac{1}{2})}{\sqrt{\text{Var}(U_{m,n})}}.$$

Comme $p > 1/2$, on en déduit que $\lim_{\min(m,n) \rightarrow \infty} \frac{mn(p - \frac{1}{2})}{\sqrt{\text{Var}(U_{m,n})}} = +\infty$. De plus comme au moins l'un des deux termes α' ou β' est strictement positif, cela implique la convergence de la suite $\left(\frac{mn(m+n+1)}{12 \text{Var}(U_{m,n})}, m \geq 1, n \geq 1 \right)$ vers une limite finie, quand $\min(m, n)$ tend vers l'infini. On a donc :

$$\lim_{\min(m,n) \rightarrow \infty} b_{m,n} = -\infty.$$

2. De la question précédente, on en déduit que pour tout réel $M > 0$, il existe $k_0 \geq 1$ tel que pour tout $m \geq k_0, n \geq k_0, b_{m,n} < -M$. Cela implique que pour tout $m \geq k_0, n \geq k_0$:

$$\mathbb{P}(\zeta_{m,n}^{(2)} > a) \geq \mathbb{P}\left(\frac{U_{m,n} - mnp}{\sqrt{\text{Var}(U_{m,n})}} > -M \right).$$

Grâce à la convergence en loi de $\left(\frac{U_{m,n} - mnp}{\sqrt{\text{Var}(U_{m,n})}}, m \geq 1, n \geq 1 \right)$ vers G_0 de la loi continue, on en déduit que :

$$\lim_{\min(m,n) \rightarrow \infty} \mathbb{P}\left(\frac{U_{m,n} - mnp}{\sqrt{\text{Var}(U_{m,n})}} > -M \right) = \mathbb{P}(G_0 > -M).$$

On a pour tout $M, \liminf_{\min(m,n) \rightarrow \infty} \mathbb{P}(\zeta_{m,n}^{(2)} > a) \geq \mathbb{P}(G_0 > -M)$. Le choix de M étant arbitraire, cela implique que $\lim_{\min(m,n) \rightarrow \infty} \mathbb{P}(\zeta_{m,n}^{(2)} > a) = 1$.

3. On définit les fonctions :

$$u_{m,n}(x, y) = \sum_{i=1}^m \sum_{j=1}^n \mathbf{1}_{\{y_j \leq x_i\}},$$

$$\zeta_{m,n}^{(2)}(x, y) = \frac{u_{m,n}(x_1, \dots, x_m, y_1, \dots, y_n) - \frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}.$$

On a $\zeta_{m,n}^{(2)}(X_1, \dots, X_m, Y_1, \dots, Y_n) = \zeta_{m,n}^{(2)}$. On considère le test pur de région critique :

$$W_{m,n} = \{(x, y) \in \mathbb{R}^m \times \mathbb{R}^n; \zeta_{m,n}^{(2)}(x, y) \leq z\}.$$

L'erreur de première espèce est $\mathbb{P}(W_{m,n}) = \mathbb{P}(\zeta_{m,n}^{(2)} \geq z)$. Comme la loi de $\zeta_{m,n}^{(2)}$ sous H_0 est indépendante de F , on en déduit que $\mathbb{P}(\zeta_{m,n}^{(2)} \geq z)$ est constant sous H_0 . Comme, sous H_0 , $\zeta_{m,n}^{(2)}$ converge en loi vers G_0 (quand $\min(m, n)$ tend vers l'infini), l'erreur de première espèce converge vers $\mathbb{P}(G_0 \geq z)$. Pour que le test soit de niveau asymptotique α , on choisit $z = \phi_{1-\alpha}$. Le test est convergent d'après la question précédente. La p -valeur asymptotique du test est $p\text{-val}_2 = \mathbb{P}(G_0 \geq \zeta_{m,n}^{(2),\text{obs}})$, où $\zeta_{m,n}^{(2),\text{obs}} = \zeta_{m,n}^{(2)}(x, y)$ est la statistique de test évaluée en les observations.

4. Pour $\alpha = 5\%$, on a $\phi_{1-\alpha} \simeq 1.64$. Il vient $u_{m,n} = 70$, $\zeta_{m,n}^{(2),\text{obs}} \simeq 0.25$ et $p\text{-val}_2 \simeq 0.403$. En particulier, on accepte H_0 au niveau 5%.

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 :

$$W_{m,n} = \{(x, y) \in \mathbb{R}^m \times \mathbb{R}^n; \zeta_{m,n}^{(3)}(x, y) > z'\},$$

avec

$$\zeta_{m,n}^{(3)}(x, y) = \sqrt{\frac{mn}{m+n}} \sup_{z \in \mathbb{R}} \left| \frac{1}{m} \sum_{i=1}^m \mathbf{1}_{\{x_i \leq z\}} - \frac{1}{n} \sum_{j=1}^n \mathbf{1}_{\{y_j \leq z\}} \right|.$$

2. Comme le test est de niveau asymptotique α , on en déduit que $z' = z_{1-\alpha}$ le quantile d'ordre $1 - \alpha$ de la loi de fonction de répartition K , i.e. $K(z_\alpha) = 1 - \alpha$, où la fonction K est définie par

$$K(z) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 z^2}.$$

Pour $\alpha = 5\%$, on a $z_\alpha \simeq 1.358$. Il vient $\zeta_{m,n}^{(3),\text{obs}} \simeq 0.508$ et $p\text{-val}_3 = 1 - K(\zeta_{m,n}^{(3),\text{obs}}) \simeq 0.958$. On accepte sans aucun doute H_0 .

Conclusion

On a $p\text{-val}_1 \simeq p\text{-val}_2 < p\text{-val}_3$. On remarque que pour le modèle le plus général (modèle 3) la p -valeur est très élevée. En revanche, plus on fait d'hypothèses sur le modèle 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 reproduites partiellement ici ne permettent pas de conclure à l'efficacité de l'ensemencement des nuages par iodure d'argent. ▲

Exercice XII.5. I Le modèle

1. L'hypothèse nulle est $H_0 = \{\text{Les deux méthodes de mesure sont similaires}\}$ et l'hypothèse alternative $H_1 = \{\text{La méthode B donne des valeurs significativement inférieures à celle de la méthode A}\}$. On a donc :

$$H_0 = \{\mu_1 = \mu_2, \sigma_1 > 0, \sigma_2 > 0\} \quad \text{et} \quad H_1 = \{\mu_1 > \mu_2, \sigma_1 > 0, \sigma_2 > 0\}.$$

On espère rejeter 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 $\frac{n-1}{\sigma_1^2}V_n$ est la loi $\chi^2(n-1)$, enfin ces deux variables sont indépendantes. Cela détermine complètement la loi du couple $\left(\bar{X}_n, \frac{n-1}{\sigma_1^2}V_n\right)$.
4. Comme les variables X_1, \dots, X_n et Y_1, \dots, Y_m sont indépendantes, on en déduit que les variables $\frac{n-1}{\sigma_1^2}V_n$ et $\frac{m-1}{\sigma_2^2}W_m$ 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 :

$$\psi_{\frac{n-1}{\sigma_1^2}V_n + \frac{m-1}{\sigma_2^2}W_m}(u) = \psi_{\frac{n-1}{\sigma_1^2}V_n}(u) \psi_{\frac{m-1}{\sigma_2^2}W_m}(u) = \frac{1}{(1-2iu)^{(n+m-2)/2}}.$$

On en déduit donc que la loi de $\frac{n-1}{\sigma_1^2}V_n + \frac{m-1}{\sigma_2^2}W_m$ 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 $(\bar{X}_n, n \geq 1)$ convergent \mathbb{P}_θ -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}_θ -p.s. vers $\mathbb{E}_\theta[X_1^2] - \mathbb{E}_\theta[X_1]^2 = \text{Var}_\theta(X_1) = \sigma_1^2$. 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}_θ -p.s. vers σ_1^2/σ_2^2 quand $\min(n, m)$ tend vers l'infini. Sous H_0 cette limite vaut 1. Sous H_1 cette limite est différente de 1.
3. Sous H_1 , on a \mathbb{P}_θ -p.s. $\lim_{\min(n,m) \rightarrow \infty} Z_{n,m} \neq 1$. En particulier, on en déduit que sous H_1 , \mathbb{P}_θ -p.s. :

$$\lim_{\min(n,m) \rightarrow \infty} \mathbf{1}_{\{\tilde{W}_{n,m}\}} = 1.$$

Par convergence dominée, on en déduit que sous H_1 , $\lim_{\min(n,m) \rightarrow \infty} \mathbb{P}_\theta(\tilde{W}_{n,m}) = 1$. 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 $n = 13$ et $m = 8$, que $a_{n-1,m-1,\alpha_1} \simeq 0.277$ et $b_{n-1,m-1,\alpha_2} \simeq 4.666$. La région critique est donc

$$\tilde{W}_{n,m} \simeq [0, 0.277] \cup [4.666, +\infty[.$$

La valeur observée de la statistique de test est $Z_{n,m}^{\text{obs}} \simeq 0.584$ (à 10^{-3} près). D'après la table XII.9, la p -valeur du test :

$$p\text{-val} = 2 \min \left(\mathbb{P}(Z_{n,m} \leq Z_{n,m}^{\text{obs}}), \mathbb{P}(Z_{n,m} \geq Z_{n,m}^{\text{obs}}) \right),$$

est d'environ 0.4 (car pour $\alpha_1 = 0.2$, on a $a_{n-1,m-1,\alpha_1} \simeq 0.588$ soit environ $Z_{n,m}^{\text{obs}}$). Il est tout à fait raisonnable d'accepter H_0 ($Z_{n,m}^{\text{obs}}$ n'est pas dans la région critique et la p -valeur est très supérieure aux valeurs critiques classiques (telles que 0.1, 0.05 ou 0.01)).

5. 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 , $\lim_{\min(n,m) \rightarrow \infty} Z_{n,m} = 1$. Par convergence dominée, on en déduit que :

$$\lim_{\min(n,m) \rightarrow \infty} \mathbb{P}(Z_{n,m} \leq 1 - \varepsilon) = 0.$$

On a donc $\lim_{\min(n,m) \rightarrow \infty} \mathbb{P}(F_{n,m} \leq 1 - \varepsilon) = 0$. 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 $\lim_{\min(n,m) \rightarrow \infty} \mathbb{P}(F_{n,m} \geq 1 + \varepsilon) = 0$, 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 :

$$1 - \varepsilon < \liminf_{\min(n,m) \rightarrow \infty} a_{n,m,\alpha_1} \leq \limsup_{\min(n,m) \rightarrow \infty} b_{n,m,\alpha_2} < 1 + \varepsilon.$$

Comme $\varepsilon \in]0, 1[$ est arbitraire, on en déduit que :

$$\lim_{\min(n,m) \rightarrow \infty} a_{n,m,\alpha_1} = \lim_{\min(n,m) \rightarrow \infty} b_{n,m,\alpha_2} = 1.$$

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 $n+m-2$ degrés de liberté. On a \mathbb{P}_θ -p.s. $\lim_{\min(n,m) \rightarrow \infty} V_n = \lim_{\min(n,m) \rightarrow \infty} W_m = \sigma^2$. On en déduit donc que \mathbb{P}_θ -p.s. $\lim_{\min(n,m) \rightarrow \infty} S_{n,m} = \sigma^2$.

2. On déduit de I.2 et de I.3 que les variables aléatoires \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 :

$$\text{Var}_\theta \left(\frac{1}{\sigma} \sqrt{\frac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m) \right) = \frac{1}{\sigma^2} \frac{nm}{n+m} \left(\frac{\sigma^2}{n} + \frac{\sigma^2}{m} \right) = 1.$$

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 $n+m-2$ degrés de liberté. De plus, sous H_0 , $\frac{1}{\sigma} \sqrt{\frac{nm}{n+m}}(\bar{X}_n - \bar{Y}_m)$ suit une loi 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 $\min(n, m)$ tend vers l'infini est $\mu_1 - \mu_2$ \mathbb{P}_θ -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 \mathbb{P}_θ -p.s., quand $\min(n, m)$ tend vers l'infini, vers $(\mu_1 - \mu_2)/\sigma$. De plus, on a :

$$\lim_{\min(n,m) \rightarrow \infty} \sqrt{\frac{nm}{n+m}} = \lim_{\min(n,m) \rightarrow \infty} \frac{1}{\sqrt{\frac{1}{n} + \frac{1}{m}}} = +\infty.$$

On en déduit que sous H_1 , \mathbb{P}_θ -p.s., la suite $(T_{n,m}, n \geq 2, m \geq 2)$ converge, quand $\min(n, m)$ tend vers l'infini, vers $+\infty$ car $\mu_1 > \mu_2$.

6. On considère la région critique :

$$W_{n,m} = \{T_{n,m} \geq t_{n+m-2, 1-\alpha}\},$$

où $t_{k,r}$ est le quantile d'ordre r 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 $\min(n, m)$ tend vers

l'infini, 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_0 de loi gaussienne centrée réduite $\mathcal{N}(0, 1)$ quand $\min(n, m)$ tend vers l'infini. Comme G_0 est une variable continue, on en déduit que pour tout $c \in \mathbb{R}$:

$$\lim_{\min(n,m) \rightarrow \infty} \mathbb{P}(T_{n,m} \geq c) = \mathbb{P}(G_0 \geq c).$$

En particulier, cela implique que la suite $(t_{n+m-2, 1-\alpha}, n \geq 2, m \geq 2)$ converge vers $\phi_{1-\alpha}$ le quantile d'ordre $1 - \alpha$ de G_0 . Cette suite est donc majorée par une constante, disons c . Par convergence dominée, on déduit de la réponse à la question précédente que sous H_1 :

$$\liminf_{\min(n,m) \rightarrow \infty} \mathbb{P}_\theta(T_{n,m} \geq t_{n+m-2, 1-\alpha}) \geq \liminf_{\min(n,m) \rightarrow \infty} \mathbb{P}_\theta(T_{n,m} \geq c) = 1.$$

Donc on a sous H_1 :

$$\lim_{\min(n,m) \rightarrow \infty} \mathbb{P}_\theta(W_{n,m}) = 1,$$

et le test est convergent. De plus l'erreur de première espèce associée est, pour $\theta \in H_0$:

$$\mathbb{P}_\theta(W_{n,m}) = \mathbb{P}_\theta(T_{n,m} \geq t_{n+m-2, 1-\alpha}) = \alpha.$$

Le niveau de ce test est α .

7. On obtient pour $\alpha = 5\%$, $t_{n+m-2, 1-\alpha} \simeq 1.729$. La région critique est donc $[1.729, \infty[$. La valeur de la statistique en les observations est $T_{n,m}^{\text{obs}} \simeq 3.648$ (à 10^{-3} près). La p -valeur est $p\text{-val} = \mathbb{P}(T \geq T_{n,m}^{\text{obs}})$, où T est de loi de Student de paramètre $n + m - 2$, est comprise entre $1/1\ 000$ et $2.5/1\ 000$ (en fait on a $p\text{-val} \simeq 0.0013$).

La valeur observée de la statistique de test appartient à la région critique. De plus, la p -valeur est très inférieure aux valeurs critiques classiques (telles que $0.1, 0.05$ ou 0.01). On rejette donc H_0 .

IV Variante sur les hypothèses du test

1. On déduit de III.5 que sous H'_1 , \mathbb{P}_θ -p.s., la suite $(T_{n,m}, n \geq 2, m \geq 2)$ converge, quand $\min(n, m)$ tend vers l'infini, vers $+\infty$ si $\mu_1 > \mu_2$, et vers $-\infty$ si $\mu_1 < \mu_2$.
2. On considère la région critique :

$$W'_{n,m} = \{|T_{n,m}| \geq t_{n+m-2, 1-\alpha/2}\}.$$

Par convergence dominée, on déduit de la réponse à la question précédente que sous H'_1 :

$$\lim_{\min(n,m) \rightarrow \infty} \mathbb{P}_\theta(W'_{n,m}) = 1.$$

Le test est donc convergent. De plus l'erreur de première espèce associée est pour $\theta \in H_0$:

$$\mathbb{P}_\theta(W'_{n,m}) = \mathbb{P}_\theta(|T_{n,m}| \geq t_{n+m-2,1-\alpha/2}) = \alpha.$$

Le niveau de ce test est donc α . Pour $\alpha = 5\%$, on obtient $t_{n+m-2,1-\alpha/2} \simeq 2.093$. La p -valeur du test est $p\text{-val} = \mathbb{P}(|T| \geq |T_{n,m}^{\text{obs}}|)$; elle est donc comprise entre $2/1\ 000$ et $5/10\ 000$ (en fait $p\text{-val} \simeq 0.0026$).

La valeur observée de la statistique de test appartient à la région critique. De plus, la p -valeur est très inférieure aux valeurs critiques classiques (telles que 0.1, 0.05 ou 0.01). On rejette donc H_0 .

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 :

$$\begin{aligned} \mathbb{P}_{(\mu_1, \sigma, \mu_2, \sigma)}(T_{n,m} \leq c) &= \mathbb{P}_{(0, \sigma, 0, \sigma)}\left(T_{n,m} + \sqrt{\frac{nm}{n+m}} \frac{\mu_1 - \mu_2}{\sqrt{S_{n,m}}} \geq c\right) \\ &= \mathbb{P}_{(0, \sigma, 0, \sigma)}\left(T_{n,m} \geq c - \sqrt{\frac{nm}{n+m}} \frac{\mu_1 - \mu_2}{\sqrt{S_{n,m}}}\right) \\ &\leq \mathbb{P}_{(0, \sigma, 0, \sigma)}(T_{n,m} \geq c) \end{aligned}$$

car $\mu_1 \leq \mu_2$ pour la dernière égalité. En particulier, pour $c = t_{n+m-2,1-\alpha}$, l'erreur de première espèce du test pur de région critique $W_{n,m} = \{T_{n,m} \geq c\}$ est donc majorée par $\mathbb{P}_{(0, \sigma, 0, \sigma)}(T_{n,m} \geq c_{n+m-2,1-\alpha})$. 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 $13/10\ 000$.



Exercice XII.6. I Estimations et intervalles de confiance du paramètre de la loi de Pareto

1. Si $\alpha \leq 1$ alors X_1 n'est pas intégrable et $\mathbb{E}[X_1] = \infty$, si $\alpha > 1$ on a $\mathbb{E}_\alpha[X_1] = \alpha/(\alpha - 1)$. Si $\alpha \leq 2$ alors X_1 n'est pas de carré intégrable, si $\alpha > 2$ on a $\mathbb{E}_\alpha[X_1^2] = \alpha/(\alpha - 2)$.
2. La méthode des moments suggère de regarder l'estimateur $\tilde{\alpha}_n = g^{-1}(\bar{X}_n)$, où $g(\alpha) = \mathbb{E}_\alpha[X_1] = \alpha/(\alpha - 1)$ et $\bar{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$. Comme $g^{-1} = g$, on en déduit que $\tilde{\alpha}_n = \bar{X}_n/(\bar{X}_n - 1)$. Si $\alpha > 1$, alors $(\bar{X}_n, n \geq 1)$ converge p.s. vers $\mathbb{E}_\alpha[X_1] = g(\alpha) > 1$. Comme g^{-1} est continue sur $]1, +\infty[$, on en déduit que $(\hat{\alpha}_n, n \geq 1)$ est un estimateur convergent de α .

On remarque que si $\alpha \leq 1$, alors $(\bar{X}_n, n \geq 1)$ converge p.s. vers $\mathbb{E}_\alpha[X_1] = +\infty$, car les variables aléatoires X_i sont positives. On en déduit que la suite $(\hat{\alpha}_n, n \geq 1)$ converge p.s. vers 1. Donc l'estimateur n'est pas convergent si $\alpha < 1$, et il est convergent si $\alpha \geq 1$.

3. La vraisemblance du modèle est, par indépendance des variables aléatoires :

$$p_n(x_1, \dots, x_n; \alpha) = \alpha^n e^{-(\alpha+1) \sum_{i=1}^n \log(x_i)} \prod_{i=1}^n \mathbf{1}_{\{x_i > 1\}}.$$

Par le théorème de factorisation, on reconnaît que $S_n = \sum_{i=1}^n \log(X_i)$ est une statistique exhaustive. On peut également remarquer qu'il s'agit d'un modèle exponentiel. L'estimateur $\tilde{\alpha}_n$ n'est pas une fonction de la statistique exhaustive. Il peut donc être amélioré.

4. On calcule la log-vraisemblance :

$$L_n(x_1, \dots, x_n; \alpha) = n \log(\alpha) - (\alpha + 1) \sum_{i=1}^n \log(x_i) + \log\left(\prod_{i=1}^n \mathbf{1}_{\{x_i > 1\}}\right).$$

La dérivée de la log-vraisemblance est :

$$\frac{\partial L_n}{\partial \alpha}(x_1, \dots, x_n; \alpha) = \frac{n}{\alpha} - \sum_{i=1}^n \log(x_i).$$

On en déduit que $\frac{\partial L_n}{\partial \alpha} = 0$ pour $\alpha = \frac{n}{\sum_{i=1}^n \log(x_i)}$. De plus la dérivée est négative (resp. positive) strictement pour les valeurs plus petites (resp. grandes) du paramètre. On en déduit donc que la log-vraisemblance est maximale en $\alpha = \frac{n}{\sum_{i=1}^n \log(x_i)}$. L'estimateur du maximum de vraisemblance de α est donc :

$$\hat{\alpha}_n = \frac{n}{\sum_{i=1}^n \log(X_i)}.$$

5. Soit g une fonction mesurable bornée. Avec le changement de variable $y = \log(x)$ de $]1, +\infty[$ dans $]0, \infty[$, on a :

$$\mathbb{E}[g(\log(X_1))] = \int g(\log(x)) f_\alpha(x) dx = \int g(y) \alpha e^{-\alpha y} \mathbf{1}_{\{y > 0\}} dy.$$

On en déduit que la loi de $\log(X_1)$ est la loi exponentielle de paramètre α . En particulier, on a $\mathbb{E}_\alpha[\log(X_1)] = 1/\alpha$ et $\mathbb{E}_\alpha[\log(X_1)^2] = 2/\alpha^2$.

6. Comme $\log(X_1)$ est intégrable, on déduit de la loi forte des grands nombres que p.s. $\left(\frac{1}{n} \sum_{i=1}^n \log(X_i), n \geq 1\right)$ converge vers $\mathbb{E}_\alpha[\log(X_1)] = \alpha^{-1}$. Comme la fonction $g(x) = 1/x$ est continue en $1/\alpha \in]0, \infty[$, et que $\hat{\alpha}_n = g\left(\frac{1}{n} \sum_{i=1}^n \log(X_i)\right)$, on en déduit que la suite $(\hat{\alpha}_n, n \geq 1)$ converge p.s. vers α . L'estimateur du maximum de vraisemblance est donc convergent.
7. Comme $\log(X_1)$ est de carré intégrable, on déduit du théorème central limite que $\left(\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n \log(X_i) - \alpha^{-1}\right), n \geq 1\right)$ converge en loi vers la loi gaussienne $\mathcal{N}(0, \sigma^2)$, où $\sigma^2 = \text{Var}_\alpha(\log(X_1)) = \alpha^{-2}$. Comme la fonction $g(x) = 1/x$ est de classe C^1 en $1/\alpha \in]0, \infty[$, on en déduit que la suite $(\sqrt{n}(\hat{\alpha}_n - \alpha), n \geq 1)$ converge en loi vers la loi gaussienne centrée de variance $\sigma^2 g'(\alpha^{-1})^2 = \alpha^{-2}(\alpha^2)^2 = \alpha^2$. L'estimateur du maximum de vraisemblance est asymptotiquement normal de variance asymptotique α^2 .
8. On calcule la dérivée seconde de la log-vraisemblance :

$$\frac{\partial^2 L_n}{\partial \alpha^2}(x_1, \dots, x_n; \alpha) = -\frac{n}{\alpha^2}.$$

On en déduit donc que :

$$I_n = -\mathbb{E}_\alpha\left[\frac{\partial^2 L_n}{\partial \alpha^2}(X_1, \dots, X_n; \alpha)\right] = \frac{n}{\alpha^2}.$$

Donc on a $I_1 = \alpha^{-2}$. La variance asymptotique de $\hat{\alpha}_n$ est égale à $1/I$. L'estimateur du maximum de vraisemblance est donc asymptotiquement efficace.

9. On déduit du théorème de Slutsky que la suite $((\sqrt{n}(\hat{\alpha}_n - \alpha), \hat{\alpha}_n), n \geq 1)$ converge en loi vers la loi de $(\alpha G_0, \alpha)$, où G_0 est une variable gaussienne de loi $\mathcal{N}(0, 1)$. Comme la fonction $f(x, y) = x/y$ est continue sur $\mathbb{R} \times \mathbb{R}^*$, on en déduit que la suite $(\sqrt{n}(\hat{\alpha}_n - \alpha)\hat{\alpha}_n^{-1}, n \geq 1)$ converge en loi vers G_0 . Donc, si $\phi_{1-\eta/2}$ est le quantile d'ordre $1 - \eta/2$ de la loi $\mathcal{N}(0, 1)$, on en déduit que :

$$\lim_{n \rightarrow \infty} \mathbb{P}_\alpha \left(\sqrt{n}(\hat{\alpha}_n - \alpha)\hat{\alpha}_n^{-1} \in [\pm \phi_{1-\eta/2}] \right) = \mathbb{P}(G \in [\pm \phi_{1-\eta/2}]) = 1 - \eta.$$

De l'équivalence :

$$\sqrt{n}(\hat{\alpha}_n - \alpha)\hat{\alpha}_n^{-1} \in [\pm \phi_{1-\eta/2}] \iff \alpha \in \left[\hat{\alpha}_n \pm \phi_{1-\eta/2} \frac{\hat{\alpha}_n}{\sqrt{n}} \right],$$

on déduit que $I = \left[\hat{\alpha}_n \pm \phi_{1-\eta/2} \frac{\hat{\alpha}_n}{\sqrt{n}} \right]$ est un intervalle de confiance de α de niveau asymptotique $1 - \eta$.

10. On a par indépendance, et en utilisant le fait que la loi de $\log(X_i)$ est la loi exponentielle de paramètre α , que :

$$\psi_{\alpha \hat{\alpha}_n^{-1}}(u) = \prod_{i=1}^n \psi_{n^{-1}\alpha \log(X_i)}(u) = \left(\frac{\alpha}{\alpha - i\alpha u/n} \right)^n = \left(\frac{n}{n - iu} \right)^n.$$

On en déduit que $\alpha \hat{\alpha}_n^{-1}$ suit la loi gamma de paramètre (n, n) . Si z_- et z_+ sont les quantiles d'ordre $\eta/2$ et $1 - \eta/2$ de la loi gamma (n, n) , on en déduit que :

$$\mathbb{P}_\alpha(\alpha \in [\hat{\alpha}_n z_-, \hat{\alpha}_n z_+]) = \mathbb{P}_\alpha(\alpha \hat{\alpha}_n^{-1} \in [z_-, z_+]) = 1 - \eta.$$

Ainsi $I_* = [\hat{\alpha}_n z_-, \hat{\alpha}_n z_+]$ est un intervalle de confiance pour α de niveau exact $1 - \eta$.

II Comparaison d'échantillons

1. Les suites sont indépendantes *a priori*.
2. L'estimateur du maximum de vraisemblance s'écrit comme une fonction de la somme des logarithmes des observations. La somme reste inchangée si on classe par ordre décroissant les termes. En particulier, on peut donc calculer l'estimateur à partir des données classées.

Soit K un compact de \mathbb{R} . Si on reprend la démonstration du théorème central limite à l'aide des fonctions caractéristiques, on remarque que pour $u \in K$, on peut majorer uniformément les restes qui interviennent dans le développement à l'ordre de 2 des fonctions caractéristiques. Cela permet de vérifier que la convergence des fonctions caractéristiques est uniforme sur les compacts.

3. On note α la valeur commune de α^{UE} et α^{USA} . On utilise les fonctions caractéristiques. Par indépendance, on a :

$$\begin{aligned} \psi_{\sqrt{\frac{m}{n+m}} Z_n^{\text{UE}} - \sqrt{\frac{n}{n+m}} Z_m^{\text{USA}}}(u) &= \psi_{\sqrt{\frac{m}{n+m}} Z_n^{\text{UE}}}(u) \psi_{-\sqrt{\frac{n}{n+m}} Z_m^{\text{USA}}}(u) \\ &= \psi_n^{\text{UE}} \left(\left(\sqrt{\frac{m}{n+m}} u \right) \right) \psi_m^{\text{USA}} \left(-\sqrt{\frac{n}{n+m}} u \right) \\ &= \left(e^{-\frac{m}{n+m} u^2 \alpha^2 / 2} + g_{n,m}(u) \right) \left(e^{-\frac{n}{n+m} u^2 \alpha^2 / 2} + h_{n,m}(u) \right) \\ &= e^{-u^2 \alpha^2 / 2} + \varepsilon_{n,m}(u), \end{aligned}$$

où les fonctions $g_{n,m}$, $h_{n,m}$ et $\varepsilon_{n,m}$ sont telles que les égalités soient vraies. Comme $\frac{m}{n+m}, \frac{n}{n+m} \in [0, 1]$, on déduit de la convergence uniforme de ψ_n^{EU} et ψ_m^{USA} vers la fonction caractéristique de la loi $\mathcal{N}(0, \alpha^2)$, que :

$$\lim_{\min(n,m) \rightarrow \infty} g_{n,m}(u) = \lim_{\min(n,m) \rightarrow \infty} h_{n,m}(u) = \lim_{\min(n,m) \rightarrow \infty} \varepsilon_{n,m}(u) = 0.$$

On en déduit donc que :

$$\lim_{\min(n,m) \rightarrow \infty} \psi_{\sqrt{\frac{m}{n+m}}Z_n^{\text{UE}} - \sqrt{\frac{n}{n+m}}Z_m^{\text{USA}}}(u) = e^{-u^2\alpha^2/2}.$$

Donc la suite $\left(\sqrt{\frac{m}{n+m}}Z_n^{\text{UE}} - \sqrt{\frac{n}{n+m}}Z_m^{\text{USA}}, n \in \mathbb{N}^*, m \in \mathbb{N}^*\right)$ converge en loi, quand $\min(n, m)$ tend vers l'infini, vers une loi gaussienne $\mathcal{N}(0, \alpha^2)$.

4. Si $\alpha^{\text{UE}} = \alpha^{\text{USA}}$, on a :

$$\zeta_{n,m} = \frac{\sqrt{\frac{m}{n+m}}Z_n^{\text{UE}} - \sqrt{\frac{n}{n+m}}Z_m^{\text{USA}}}{\sigma_{n,m}}.$$

De la convergence des estimateurs, on déduit que $(\hat{\sigma}_{n,m}^2, n \in \mathbb{N}^*, m \in \mathbb{N}^*)$ est, quand $\min(n, m)$ tend vers l'infini, un estimateur convergent de α^2 . On déduit de la question précédente et du théorème de Slutsky que si $\alpha^{\text{UE}} = \alpha^{\text{USA}}$, la suite $(\zeta_{n,m}, n \in \mathbb{N}^*, m \in \mathbb{N}^*)$ converge en loi, quand $\min(n, m)$ tend vers l'infini, vers la loi $\mathcal{N}(0, 1)$.

5. Si $\alpha^{\text{UE}} < \alpha^{\text{USA}}$, on a p.s. :

$$\lim_{\min(n,m) \rightarrow \infty} \hat{\alpha}_n^{\text{UE}} - \hat{\alpha}_m^{\text{USA}} < 0, \quad \lim_{\min(n,m) \rightarrow \infty} \sqrt{\frac{mn}{m+n}} = +\infty,$$

et par la loi forte des grands nombres p.s. :

$$\min(\alpha^{\text{UE}}, \alpha^{\text{USA}}) \leq \liminf_{\min(n,m) \rightarrow \infty} \hat{\sigma}_{n,m} \leq \limsup_{\min(n,m) \rightarrow \infty} \hat{\sigma}_{n,m} \leq \max(\alpha^{\text{UE}}, \alpha^{\text{USA}}).$$

De sorte que p.s. on a $\lim_{\min(n,m) \rightarrow \infty} \zeta_{n,m} = -\infty$.

6. Par symétrie, p.s. on a $\lim_{\min(n,m) \rightarrow \infty} \zeta_{n,m} = +\infty$, si $\alpha^{\text{UE}} > \alpha^{\text{USA}}$.

7. Vu les comportements sous H_0 et sous H_1 de la statistique de test, il est naturel de considérer le test de région critique $W = \{\zeta_{n,m} \geq c\}$.

8. On note $T_r = \hat{\alpha}_{n^r}^r / \alpha^r$. Sa loi ne dépend que de n^r . On a $\zeta_{n,m} = h_{\alpha^{\text{UE}}}(\alpha^{\text{USA}})$, où :

$$h_a(x) = \sqrt{nm} \frac{aT_{\text{UE}} - xT_{\text{USA}}}{\sqrt{na^2T_{\text{UE}}^2 + mx^2T_{\text{USA}}^2}}.$$

En calculant la dérivée de h_a et en utilisant que p.s. $T_{\text{EU}} > 0$ et $T_{\text{USA}} > 0$, on vérifie que la fonction h_a est décroissante. On a pour $\alpha^{\text{UE}} \leq \alpha^{\text{USA}}$:

$$\begin{aligned} \mathbb{P}_{\alpha^{\text{UE}}, \alpha^{\text{USA}}}(\zeta_{n,m} \geq c) &= \mathbb{P}(h_{\alpha^{\text{UE}}}(\alpha^{\text{USA}}) \geq c) \\ &\leq \mathbb{P}(h_{\alpha^{\text{UE}}}(\alpha^{\text{UE}}) \geq c) \\ &= \mathbb{P}(h_1(1) \geq c), \end{aligned}$$

où pour la première égalité, on a utilisé que la loi de $h_a(x)$ ne dépend pas de α^{UE} ni de α^{USA} , la décroissance de h_a pour l'inégalité, et le fait que $h_a(a)$ est indépendant de a pour la dernière égalité. On en déduit donc que le niveau du test est donné par :

$$\sup_{0 < \alpha^{\text{UE}} \leq \alpha^{\text{USA}}} \mathbb{P}_{\alpha^{\text{UE}}, \alpha^{\text{USA}}}(\zeta_{n,m} \geq c) = \mathbb{P}_{1,1}(\zeta_{n,m} \geq c).$$

9. On déduit de la question précédente et de la question II.4, que :

$$\begin{aligned} \lim_{\min(n,m) \rightarrow \infty} \sup_{0 < \alpha^{\text{UE}} \leq \alpha^{\text{USA}}} \mathbb{P}_{\alpha^{\text{UE}}, \alpha^{\text{USA}}}(\zeta_{n,m} \geq c) \\ = \lim_{\min(n,m) \rightarrow \infty} \mathbb{P}_{1,1}(\zeta_{n,m} \geq c) = \mathbb{P}(G_0 \geq c), \end{aligned}$$

où G_0 est de loi $\mathcal{N}(0, 1)$. Le test est de niveau asymptotique η pour c égal à $\phi_{1-\eta}$, le quantile d'ordre $1 - \eta$ de la loi $\mathcal{N}(0, 1)$.

10. Par convergence dominée, on a pour $\alpha^{\text{UE}} > \alpha^{\text{USA}}$:

$$\lim_{\min(n,m) \rightarrow \infty} \mathbb{E}_{\alpha^{\text{UE}}, \alpha^{\text{USA}}}[\mathbf{1}_{\{\zeta_{n,m} \geq c\}}] = 1.$$

Le test est donc convergent.

11. La p -valeur asymptotique du test est $\mathbb{P}(G_0 \geq \zeta_{n,m}^{\text{obs}})$, où G_0 est de loi $\mathcal{N}(0, 1)$ et $\zeta_{n,m}^{\text{obs}}$ est la statistique de test évaluée en les observations. D'après la question II.8, la p -valeur asymptotique est uniforme.

III Application numérique

1. On obtient $\hat{\alpha}_n^{\text{UE}} \simeq 1.590$, $I^{\text{UE}} \simeq [1.399, 1.782]$, et $\hat{\alpha}_n^{\text{USA}} \simeq 1.348$, $I^{\text{USA}} \simeq [1.161, 1.536]$.

Les intervalles de confiance de niveau asymptotique 95% sont très proches des intervalles exacts (cf. la dernière question de la partie I) qui sont : $I_*^{\text{UE}} \simeq [1.404, 1.787]$ et $I_*^{\text{USA}} \simeq [1.168, 1.542]$.

2. On obtient $\zeta_{n,m} \simeq 1.728$ et une p -valeur asymptotique de 4.2% environ.

3. On rejette donc au niveau de 5% le fait qu'il existe relativement plus de très grandes villes dans l'UE qu'aux USA.

IV Réduction des lois de Pareto

1. La fonction de répartition de la loi de Pareto est nulle pour $x \leq \beta$ et pour $x > \beta$ on a :

$$F_{\alpha,\beta}(x) = \mathbb{P}(\tilde{X}_1 \leq x) = 1 - \frac{\beta^\alpha}{x^\alpha}.$$

2. Il vient pour $y > \beta$ et $x > 1$:

$$\mathbb{P}\left(\frac{\tilde{X}_1}{y} > x \mid \tilde{X}_1 > y\right) = \frac{1 - F_{\alpha,\beta}(yx)}{1 - F_{\alpha,\beta}(y)} = \frac{1}{x^\alpha} = 1 - F_{\alpha,1}(x).$$

La fonction de répartition de \tilde{X}_1/y sachant $\tilde{X}_1 > y$ est $1 - \mathbb{P}(\tilde{X}_1/y > x \mid \tilde{X}_1 > y) = F_{\alpha,1}(x)$ pour $x > 1$. Elle est nulle si $x \leq 1$. On en déduit donc que la loi de \tilde{X}_1/y sachant $\tilde{X}_1 > y$ est la loi de Pareto de paramètre $(\alpha, 1)$.

Nous démontrons les résultats admis de l'énoncé. On pose $m = n + k$. Le vecteur $(\tilde{X}_1, \dots, \tilde{X}_m)$ possède une densité. En particulier p.s. $\tilde{X}_i \neq \tilde{X}_j$ pour $1 \leq i < j \leq m$. On en déduit donc que p.s. le réordonnement décroissant est bien défini et qu'il est unique. De plus les événements $\{\tilde{X}_{\sigma(1)} > \dots > \tilde{X}_{\sigma(m)}\}$ où σ parcourt l'ensemble \mathcal{S}_m des permutations de $\{1, \dots, m\}$ sont disjoints deux à deux, et d'après ce qui précède leur union est de probabilité 1. Si h est une fonction mesurable bornée, on a donc par la formule de décomposition

$$\begin{aligned} & \mathbb{E}[h(\tilde{X}_{(1)}, \dots, \tilde{X}_{(m)})] \\ &= \sum_{\sigma \in \mathcal{S}_m} \mathbb{E}[h(\tilde{X}_{(1)}, \dots, \tilde{X}_{(m)}) \mathbf{1}_{\{\tilde{X}_{\sigma(1)} > \dots > \tilde{X}_{\sigma(m)}\}}] \\ &= \sum_{\sigma \in \mathcal{S}_m} \mathbb{E}[h(\tilde{X}_{\sigma(1)}, \dots, \tilde{X}_{\sigma(m)}) \mathbf{1}_{\{\tilde{X}_{\sigma(1)} > \dots > \tilde{X}_{\sigma(m)}\}}] \\ &= \sum_{\sigma \in \mathcal{S}_m} \int h(x_{\sigma(1)}, \dots, x_{\sigma(m)}) \mathbf{1}_{\{x_{\sigma(1)} > \dots > x_{\sigma(m)}\}} \prod_{i=1}^m f_{\alpha,\beta}(x_i) dx_1 \dots dx_m \\ &= m! \int h(x_1, \dots, x_m) \mathbf{1}_{\{x_1 > \dots > x_m\}} \prod_{i=1}^m f_{\alpha,\beta}(x_i) dx_1 \dots dx_m, \end{aligned}$$

où l'on a utilisé pour la troisième égalité le fait que les variables $\tilde{X}_1, \dots, \tilde{X}_m$ sont indépendantes et de densité $f_{\alpha,\beta}$ et la symétrie pour la quatrième égalité. On en déduit donc que le réordonnement décroissant est un vecteur de loi continue et de densité :

$$g_m(x_1, \dots, x_m) = m! \mathbf{1}_{\{x_1 > \dots > x_m\}} \prod_{i=1}^m f_{\alpha,\beta}(x_i).$$

3. Par la formule des lois marginales, la densité de $(\tilde{X}_{(1)}, \dots, \tilde{X}_{(n+1)})$ est :

$$\begin{aligned} f(x_1, \dots, x_{n+1}) &= \int g_m(x_1, \dots, x_m) dx_{n+2} \dots dx_m \\ &= \mathbf{1}_{\{x_1 > \dots > x_{n+1}\}} \prod_{i=1}^{n+1} f_{\alpha, \beta}(x_i) h(x_{n+1}), \end{aligned}$$

où $h(x_{n+1}) = \int \mathbf{1}_{\{x_{n+1} > x_{n+2} > \dots > x_m\}} \prod_{j=n+2}^m f_{\alpha, \beta}(x_j) dx_{n+2} \dots dx_m$. En intégrant d'abord sur x_m , il vient que $h(x_{n+1})$ est égal à :

$$\int \mathbf{1}_{\{x_{n+1} > x_{n+2} > \dots > x_{m-1}\}} F_{\alpha, \beta}(x_{m-1}) f_{\alpha, \beta}(x_{m-1}) \prod_{j=n+2}^{m-2} f_{\alpha, \beta}(x_j) dx_{n+2} \dots dx_{m-1}.$$

Une primitive de $F_{\alpha, \beta}(x) f_{\alpha, \beta}(x)$ est $F_{\alpha, \beta}(x)^2/2$. En intégrant sur x_{m-1} , il vient que $h(x_{n+1})$ est égal à :

$$\frac{1}{2} \int \mathbf{1}_{\{x_{n+1} > x_{n+2} > \dots > x_{m-2}\}} F_{\alpha, \beta}(x_{m-2})^2 f_{\alpha, \beta}(x_{m-2}) \prod_{j=n+2}^{m-3} f_{\alpha, \beta}(x_j) dx_{n+2} \dots dx_{m-2}.$$

Par une récurrence immédiate, on obtient :

$$h(x_{n+1}) = \frac{1}{(k-1)!} F_{\alpha, \beta}(x_{n+1})^{k-1} f_{\alpha, \beta}(x_{n+1}).$$

La densité de la loi de $(\tilde{X}_{(1)}, \dots, \tilde{X}_{(n+1)})$ est donc :

$$\frac{m!}{(k-1)!} \mathbf{1}_{\{x_1 > \dots > x_{n+1}\}} F_{\alpha, \beta}(x_{n+1})^{k-1} f_{\alpha, \beta}(x_{n+1}) \prod_{i=1}^n f_{\alpha, \beta}(x_i).$$

4. Soit h une fonction bornée mesurable. On a :

$$\begin{aligned}
 & \mathbb{E}[h(Y_1, \dots, Y_n)] \\
 &= \mathbb{E} \left[h \left(\frac{\tilde{X}_{(1)}}{\tilde{X}_{(n+1)}}, \dots, \frac{\tilde{X}_{(n)}}{\tilde{X}_{(n+1)}} \right) \right] \\
 &= \int h \left(\frac{x_1}{x_{n+1}}, \dots, \frac{x_n}{x_{n+1}} \right) \frac{m!}{(k-1)!} \mathbf{1}_{\{x_1 > \dots > x_{n+1}\}} F_{\alpha, \beta}(x_{n+1})^{k-1} \\
 & \quad f_{\alpha, \beta}(x_{n+1}) \prod_{i=1}^n f_{\alpha, \beta}(x_i) dx_1 \dots dx_{n+1} \\
 &= \int h \left(\frac{x_1}{x_{n+1}}, \dots, \frac{x_n}{x_{n+1}} \right) \frac{m!}{(k-1)!} \mathbf{1}_{\{x_1 > \dots > x_{n+1}\}} F_{\alpha, \beta}(x_{n+1})^{k-1} \\
 & \quad f_{\alpha, \beta}(x_{n+1}) \prod_{i=1}^n \frac{\alpha \beta^{\alpha+1}}{x_i^{\alpha+1}} dx_1 \dots dx_{n+1} \\
 &= \int h(y_1, \dots, y_n) \frac{m!}{(k-1)!} \mathbf{1}_{\{y_1 > \dots > y_n > 1\}} F_{\alpha, \beta}(x_{n+1})^{k-1} \\
 & \quad f_{\alpha, \beta}(x_{n+1}) \prod_{i=1}^n \frac{\alpha \beta^{\alpha}}{y_i^{\alpha+1} x_{n+1}^{\alpha+1}} x_{n+1}^n dy_1 \dots dy_n dx_{n+1} \\
 &= c \int h(y_1, \dots, y_n) \mathbf{1}_{\{y_1 > \dots > y_n > 1\}} \prod_{i=1}^n \frac{1}{y_i^{\alpha+1}} dy_1 \dots dy_n,
 \end{aligned}$$

où l'on a utilisé pour la troisième égalité que $x_i > x_{n+1} > \beta$ impliquait $x_i > \beta$ pour supprimer les indicatrices $\mathbf{1}_{\{x_i > \beta\}}$, le changement de variable $y_i = x_i/x_{n+1}$ dans la quatrième égalité pour $1 \leq i \leq n$, et où c ne dépend pas de y_1, \dots, y_n . La densité du vecteur (Y_1, \dots, Y_n) est donc :

$$f(y_1, \dots, y_n) = c \mathbf{1}_{\{y_1 > \dots > y_n > 1\}} \prod_{i=1}^n \frac{1}{y_i^{\alpha+1}}.$$

On reconnaît que f est proportionnel, et donc égal, à la densité du réordonnement décroissant de n variables aléatoires indépendantes de loi de Pareto de paramètre $(\alpha, 1)$. En particulier, on en déduit que $c = n!$ et que (Y_1, \dots, Y_n) a même loi que le réordonnement décroissant de n variables aléatoires indépendantes de loi de Pareto de paramètre $(\alpha, 1)$. La loi de (Y_1, \dots, Y_n) ne dépend donc ni de β ni de k . ▲

Exercice XII.7. I Étude sommaire de la loi de Weibull

1. Comme $c > 0$, on a pour tout $x \in \mathbb{R}$, $\mathbb{P}_{(\alpha, \beta)}(cX \leq x) = \mathbb{P}_{(\alpha, \beta)}(X \leq x/c) = F_{(\alpha, \beta)}(x/c) = F_{(\alpha, c\beta)}(x)$. La loi de cX est donc la loi de Weibull de paramètre $(\alpha, c\beta)$.

2. Par changement de variable $y = (x/\beta)^\alpha$, on obtient :

$$\begin{aligned}\mathbb{E}_{(\alpha,\beta)}[X^\alpha] &= \int x^\alpha f_{\beta,\alpha}(x) dx = \Gamma(2)\beta^\alpha = \beta^\alpha, \\ \mathbb{E}_{(\alpha,\beta)}[X^{2\alpha}] &= \int x^{2\alpha} f_{\beta,\alpha}(x) dx = \Gamma(3)\beta^{2\alpha} = 2\beta^{2\alpha}.\end{aligned}$$

3. On utilise la méthode de la fonction muette. Soit g mesurable bornée. On a :

$$\begin{aligned}\mathbb{E}_{(\alpha,\beta)}[g(X^\alpha)] &= \int_{\mathbb{R}} g(x^\alpha) f_{\alpha,\beta}(x) dx \\ &= \int_{]0,\infty[} g(x^\alpha) \frac{\alpha}{\beta} \left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left(-\left(\frac{x}{\beta}\right)^\alpha\right) dx \\ &= \int g(y) \beta^{-\alpha} e^{-\beta^{-\alpha}y} \mathbf{1}_{]0,\infty[}(y) dy,\end{aligned}$$

où l'on a effectué le changement de variable $y = x^\alpha$ sur $]0, \infty[$. On en déduit que $Y = X^\alpha$ est une variable aléatoire de densité $h(y) = \beta^{-\alpha} e^{-\beta^{-\alpha}y} \mathbf{1}_{]0,\infty[}(y)$. On reconnaît la loi exponentielle de paramètre $\beta^{-\alpha}$. Comme le moment d'ordre 1 (resp. 2) de la loi exponentielle de paramètre λ est $1/\lambda$ (resp. $2/\lambda^2$), on retrouve bien les résultats de la question précédente.

II Pourquoi la loi de Weibull ?

1. On a $1 - F_n(x) = \mathbb{P}(\min_{1 \leq i \leq n} X_i > x)$ et :

$$\mathbb{P}\left(\min_{1 \leq i \leq n} X_i > x\right) = \mathbb{P}\left(\bigcap_{i=1}^n \{X_i > x\}\right) = \prod_{i=1}^n \mathbb{P}(X_i > x) = (1 - F(x))^n,$$

où l'on a utilisé l'indépendance des variables $(X_i, i \geq 1)$ pour la deuxième égalité et le fait qu'elles ont même loi pour la dernière égalité.

2. La réponse à la question précédente assure que $\min_{1 \leq k \leq n} X_k$ suit la loi de Weibull de paramètre $(\alpha, n^{-1/\alpha}\beta)$. On déduit de la réponse à la question I.1, que $n^{-1/\alpha}X_1$ suit également la loi de Weibull de paramètre $(\alpha, n^{-1/\alpha}\beta)$.

3. D'après la question précédente, le membre de droite de (XII.7) suit la loi de Weibull de paramètre $(\alpha, c_L/n n^{-1/\alpha}\beta)$. Le membre de gauche suit la loi de Weibull de paramètre $(\alpha, c_L\beta)$. On en déduit que l'égalité (XII.7) est satisfaite en loi pour tout L dès que $c_\ell = c_1 \ell^{-1/\alpha}$ pour tout $\ell > 0$. La loi de $X^{(L)}$ est alors la loi de Weibull de paramètre $(\alpha, c_1 L^{-1/\alpha}\beta)$.

4. Si on note G_n la fonction de répartition de $X_1^{(L/n)}$, on a :

$$G_n(x) = P(X_1^{(L/n)} \leq x) = \mathbb{P}((L/n)^{-1/\alpha} X \leq x) = H((L/n)^{1/\alpha} x).$$

En particulier, il vient : pour $x < 0$, $\lim_{n \rightarrow \infty} (1 - G_n(x))^n = 0$; et, pour $x > 0$,

$$\lim_{n \rightarrow \infty} (1 - G_n(x))^n = \lim_{n \rightarrow \infty} \left(1 - b(L/n)^{a/\alpha} x^a + o(n^{-a/\alpha}) \right)^n = \begin{cases} 1 & \text{si } a > \alpha, \\ e^{-bLx^a} & \text{si } a = \alpha, \\ 0 & \text{si } a < \alpha. \end{cases}$$

Soit G la fonction de répartition de $X^{(L)}$. L'égalité (XII.7) implique que $G(x) = 1 - (1 - G_n(x))^n = 1 - \lim_{n \rightarrow \infty} (1 - G_n(x))^n$.

- Si $a > \alpha$, on obtient $G(x) = 0$ pour tout $x \in \mathbb{R}$. La fonction G n'est pas la fonction de répartition d'une variable aléatoire réelle.
- Si $a < \alpha$, on obtient $G(x) = \mathbf{1}_{]0, \infty[}(x)$. La fonction G n'est donc pas la fonction de répartition d'une variable aléatoire strictement positive.
- Pour $a = \alpha$, on a $G = F_{(\alpha, (bL)^{-1/\alpha})}$.

La fonction G correspond à une fonction de répartition si et seulement si $a = \alpha$. Il s'agit de la loi de Weibull de paramètre $(\alpha, (bL)^{-1/\alpha})$. La loi de X est alors la loi de Weibull de paramètre $(\alpha, b^{-1/\alpha})$.

III Estimation du paramètre d'échelle

1. Par indépendance, la densité de l'échantillon est le produit des densités. Il vient :

$$p_n(x; \beta) = \prod_{i=1}^n f_{(\alpha_0, \beta)}(x_i) = \frac{\alpha_0^n}{\beta^{n\alpha_0}} e^{-\beta^{-\alpha_0} \sum_{i=1}^n x_i^{\alpha_0}} \prod_{j=1}^n x_j^{\alpha_0 - 1} \prod_{l=1}^n \mathbf{1}_{\{x_l > 0\}}.$$

2. Le théorème de factorisation assure que $S = \sum_{i=1}^n X_i^{\alpha_0}$ est une statistique exhaustive.
3. On considère la log-vraisemblance $L_n(x; \beta) = \log(p_n(x; \beta))$. On a :

$$L_n(x; \beta) = n \log(\alpha_0) - n\alpha_0 \log(\beta) - \beta^{-\alpha_0} \sum_{i=1}^n x_i^{\alpha_0} + (\alpha_0 - 1) \sum_{j=1}^n \log(x_j \mathbf{1}_{\{x_j > 0\}}).$$

On a :

$$\frac{\partial L_n(x; \beta)}{\partial \beta} = -\frac{n\alpha_0}{\beta} + \alpha_0 \beta^{-\alpha_0 - 1} \sum_{i=1}^n x_i^{\alpha_0}.$$

Ainsi la dérivée de la log-vraisemblance est nulle en $\hat{\beta}_n = \left(\frac{1}{n} \sum_{i=1}^n x_i^{\alpha_0} \right)^{1/\alpha_0}$.

De plus, elle est positive pour $\beta < \hat{\beta}_n$ et négative pour $\beta > \hat{\beta}_n$. Ceci assure

que la log-vraisemblance atteint son unique maximum en $\hat{\beta}_n$. L'estimateur du maximum de vraisemblance est donc :

$$\hat{\beta}_n = \left(\frac{1}{n} \sum_{i=1}^n X_i^{\alpha_0} \right)^{1/\alpha_0}.$$

4. Les variables aléatoires $X_i^{\alpha_0}$ sont indépendantes de loi exponentielle de paramètres $\beta^{-\alpha_0}$. En utilisant les fonctions caractéristiques, on en déduit que la loi de $\frac{1}{n} \sum_{i=1}^n X_i^{\alpha_0}$ est la loi Γ de paramètre $(n\beta^{-\alpha_0}, n)$.

Si Z suit la loi $\Gamma(\lambda, a)$, on a :

$$\begin{aligned} \mathbb{E}[Z^b] &= \int_{]0, \infty[} x^b \frac{1}{\Gamma(a)} \lambda^a x^{a-1} e^{-\lambda x} dx \\ &= \frac{\Gamma(a+b)}{\lambda^b \Gamma(a)} \int_{]0, \infty[} \frac{1}{\Gamma(a+b)} \lambda^{a+b} x^{a+b-1} e^{-\lambda x} dx = \frac{\Gamma(a+b)}{\lambda^b \Gamma(a)}, \end{aligned}$$

où l'on a utilisé pour la dernière égalité le fait que l'intégrale de la densité de la loi $\Gamma(\lambda, a+b)$ vaut 1.

On en déduit donc, avec $Z = \frac{1}{n} \sum_{i=1}^n X_i^{\alpha_0}$, $\lambda = n\beta^{-\alpha_0}$, $a = n$, et $b = 1/\alpha_0$, que :

$$\mathbb{E}_{(\alpha_0, \beta)}[\hat{\beta}_n] = \frac{\Gamma(n + \frac{1}{\alpha_0})}{n^{1/\alpha_0} \Gamma(n)} \beta.$$

Le biais est donc égal à $\mathbb{E}[\hat{\beta}_n] - \beta = \left(\frac{\Gamma(n + \frac{1}{\alpha_0})}{n^{1/\alpha_0} \Gamma(n)} - 1 \right) \beta$. L'estimateur est sans biais si $\alpha_0 = 1$. On peut obtenir, à l'aide de la formule de Stirling à l'ordre 1 :

$$\Gamma(z) = e^{-z} z^{z-\frac{1}{2}} (2\pi)^{1/2} \left[1 + \frac{1}{12z} + O\left(\frac{1}{z^2}\right) \right],$$

que le biais, quand n tend vers l'infini, est asymptotiquement équivalent à $(1 - \alpha_0)/(2n\alpha_0^2)$.

5. On déduit de la loi forte des grands nombres que $(\hat{\beta}_n^{\alpha_0}, n \geq 1)$ converge p.s. vers $\mathbb{E}_{(\alpha_0, \beta)}[X^{\alpha_0}] = \beta^{\alpha_0}$. L'estimateur $\hat{\beta}_n^{\alpha_0}$ de β^{α_0} est donc convergent. On a $\text{Var}_{(\alpha_0, \beta)}(X^{\alpha_0}) = \beta^{2\alpha_0}$. Le théorème central limite assure que la suite $(\sqrt{n}(\hat{\beta}_n^{\alpha_0} - \beta^{\alpha_0}), n \geq 1)$ converge en loi vers la loi gaussienne centrée de variance $\beta^{2\alpha_0}$. L'estimateur $\hat{\beta}_n^{\alpha_0}$ de β^{α_0} est donc asymptotiquement normal de variance asymptotique $\beta^{2\alpha_0}$.

6. La fonction $g : x \mapsto x^{1/\alpha_0}$ est continue et de classe \mathcal{C}^1 sur $]0, \infty[$. On déduit de la question précédente que l'estimateur $\hat{\beta}_n = g(\hat{\beta}_n^{\alpha_0})$ de $\beta = g(\beta^{\alpha_0})$ est convergent et qu'il est asymptotiquement normal de variance asymptotique :

$$g'(\beta^{\alpha_0})^2 \beta^{2\alpha_0} = \left(\frac{\beta}{\alpha_0} \right)^2.$$

IV Estimation du paramètre de forme

1. La vraisemblance est :

$$p_n(x; (\alpha, \beta)) = \frac{\alpha^n}{\beta^{n\alpha}} e^{-\beta^{-\alpha} \sum_{i=1}^n x_i^\alpha + (\alpha-1) \sum_{j=1}^n \log(x_j)} \prod_{l=1}^n \mathbf{1}_{\{x_l > 0\}}.$$

La log-vraisemblance est :

$$\begin{aligned} L_n(x; (\alpha, \beta)) &= n \log(\alpha) - n\alpha \log(\beta) - \beta^{-\alpha} \sum_{i=1}^n x_i^\alpha \\ &\quad + (\alpha - 1) \sum_{j=1}^n \log(x_j) + \sum_{l=1}^n \log(\mathbf{1}_{\{x_l > 0\}}). \end{aligned}$$

2. Le théorème de factorisation ne permet pas de trouver d'autre statistique exhaustive que l'échantillon entier. Dans ce modèle, on ne peut pas résumer les données !
3. On a :

$$\begin{aligned} \frac{\partial L_n(x; (\alpha, \beta))}{\partial \alpha} &= \frac{n}{\alpha} - n \log(\beta) + \log(\beta) \beta^{-\alpha} \sum_{i=1}^n x_i^\alpha \\ &\quad - \beta^{-\alpha} \sum_{j=1}^n \log(x_j) x_j^\alpha + \sum_{l=1}^n \log(x_l), \end{aligned}$$

$$\frac{\partial L_n(x; (\alpha, \beta))}{\partial \beta} = -n \frac{\alpha}{\beta} + \alpha \beta^{-\alpha-1} \sum_{i=1}^n x_i^\alpha.$$

Pour tout extremum local, le gradient de la log-vraisemblance s'annule. On recherche donc $(\tilde{\alpha}_n, \tilde{\beta}_n)$ tel que $\frac{\partial L_n(x; (\tilde{\alpha}_n, \tilde{\beta}_n))}{\partial \alpha} = \frac{\partial L_n(x; (\tilde{\alpha}_n, \tilde{\beta}_n))}{\partial \beta} = 0$.

4. On obtient que $\tilde{\beta}_n = u_n(\tilde{\alpha}_n)$, avec u_n définie par :

$$u_n(a) = \left(\frac{1}{n} \sum_{i=1}^n x_i^a \right)^{1/a}.$$

D'après la partie III, à α fixé, $u_n(\alpha)$ maximise la log-vraisemblance. On en déduit que l'estimation de β est calculée de manière similaire que α soit connu ou non : soit par $u_n(\alpha_0)$ si α est connu égal à α_0 , soit, si α n'est pas connu, par $u(\hat{\alpha}_n)$, avec $\hat{\alpha}_n$ l'estimateur du maximum de vraisemblance de α (quand il existe).

Vérifions que la fonction h_n est strictement croissante. On a :

$$h'_n(a) = \frac{n}{\left(\sum_{i=1}^n x_i^a\right)^2} \left[\sum_{k=1}^n x_k^a \sum_{j=1}^n x_j^a \log(x_j)^2 - \left(\sum_{l=1}^n x_l^a \log(x_l) \right)^2 \right].$$

En utilisant Cauchy-Schwarz, il vient $\sum_{k=1}^n x_k^a \sum_{j=1}^n x_j^a \log(x_j)^2 \geq \left(\sum_{l=1}^n x_l^a \log(x_l) \right)^2$.

En particulier h'_n est strictement positive pour $a > 0$ car $n \geq 2$ et il existe $i, j \in \{1, \dots, n\}$ tels que $x_i \neq x_j$. Comme $h_n(0) = 0$, ceci assure que h_n est strictement positive sur $]0, +\infty[$.

5. Comme $\tilde{\beta}_n = u_n(\tilde{\alpha}_n)$, on obtient :

$$\frac{\partial L_n(x; (\tilde{\alpha}_n, \tilde{\beta}_n))}{\partial \alpha} = \frac{n}{\tilde{\alpha}_n} - n h_n(\tilde{\alpha}_n).$$

Comme $\frac{\partial L_n(x; (\tilde{\alpha}_n, \tilde{\beta}_n))}{\partial \alpha} = 0$, on en déduit que $\tilde{\alpha}_n$ est solution de $h_n(a) = 1/a$.

La fonction $a \mapsto h_n(a) - \frac{1}{a}$ est croissante, et on a $\lim_{a \rightarrow 0^+} h_n(a) - \frac{1}{a} = -\infty$ ainsi que $\lim_{a \rightarrow \infty} h_n(a) - \frac{1}{a} = \lim_{a \rightarrow \infty} h_n(a) > 0$, car h_n est strictement croissante et $h_n(0) = 0$. Donc il existe une unique valeur $\tilde{\alpha}_n > 0$ solution de $h_n(a) = \frac{1}{a}$.

6. On a :

$$g'(a) = \frac{dL_n(x; (a, u_n(a)))}{da} = \frac{\partial L_n(x; (a, u_n(a)))}{\partial \alpha} + \frac{\partial L_n(a, u_n(a))}{\partial \beta} u'_n(a).$$

Par définition de $u_n(a)$, on a $\frac{\partial L_n(a, u_n(a))}{\partial \beta} = 0$. Par définition de h_n , il

vient $g'(a) = n \left(\frac{1}{a} - h_n(a) \right)$, qui est une fonction strictement décroissante. La fonction g est donc strictement concave. Elle est donc maximale en $\tilde{\alpha}_n$ unique solution de $h_n(a) = 1/a$.

D'après la question III.3, on sait que la fonction $\beta \mapsto L_n(x; (\alpha, \beta))$ est maximale pour $\beta = u_n(\alpha)$. Ce qui précède assure que la fonction $\alpha \mapsto L_n(x; (\alpha, u_n(\alpha)))$ atteint son maximum en $\tilde{\alpha}_n$. On en déduit donc que la log-vraisemblance atteint son unique maximum en $(\tilde{\alpha}_n, \tilde{\beta}_n = u_n(\tilde{\alpha}_n))$.

7. Les variables $\log(X_1)$, $X_1^a \log(X_1)$ et X_1^a sont intégrables pour tout $a > 0$. On déduit de la loi forte des grands nombres que $h_n(a)$ converge vers :

$$h(a) = -\mathbb{E}_{(\alpha, \beta)}[\log(X_1)] + \frac{\mathbb{E}_{(\alpha, \beta)}[\log(X_1)X_1^a]}{\mathbb{E}[X_1^a]}.$$

Un intégration par partie assure que :

$$\begin{aligned} \mathbb{E}_{(\alpha, \beta)}[\log(X_1)X_1^\alpha] &= \int_0^\infty \log(x)x^\alpha f_{\alpha, \beta}(x)dx \\ &= [\log(x)x^\alpha (F_{\alpha, \beta}(x) - 1)]_0^\infty \\ &\quad - \int_0^\infty (1 + \alpha \log(x))x^{\alpha-1} (F_{\alpha, \beta}(x) - 1) dx \\ &= \int_0^\infty (1 + \alpha \log(x))x^{\alpha-1} \frac{\beta^\alpha}{\alpha x^{\alpha-1}} f_{\alpha, \beta}(x)dx \\ &= \beta^\alpha \left(\frac{1}{\alpha} + \mathbb{E}[\log(X_1)] \right). \end{aligned}$$

Comme $\mathbb{E}_{(\alpha, \beta)}[X_1^\alpha] = \beta^\alpha$, on en déduit que :

$$h(\alpha) = \frac{\mathbb{E}[\log(X_1)X_1^\alpha]}{\mathbb{E}[X_1^\alpha]} - \mathbb{E}[\log(X_1)] = \frac{1}{\alpha}.$$

On montre la convergence p.s. de l'estimateur $(\tilde{\alpha}_n, \tilde{\beta}_n)$.

Par convergence dominée, on obtient que h est continue et dérivable avec :

$$h'(a) = \frac{1}{\mathbb{E}_{(\alpha, \beta)}[X_1^a]^2} (\mathbb{E}_{(\alpha, \beta)}[X_1^a] \mathbb{E}_{(\alpha, \beta)}[X_1^a \log(X_1)^2] - \mathbb{E}_{(\alpha, \beta)}[X_1^a \log(X_1)]^2).$$

En utilisant Cauchy-Schwarz, il vient :

$$\mathbb{E}_{(\alpha, \beta)}[X_1^a] \mathbb{E}_{(\alpha, \beta)}[X_1^a \log(X_1)^2] > \mathbb{E}_{(\alpha, \beta)}[X_1^a \log(X_1)]^2$$

(avec inégalité stricte car pour tout $\lambda \in \mathbb{R}$, $\mathbb{P}_{(\alpha, \beta)}(\log(X_1)X_1^{a/2} = \lambda X_1^{a/2}) < 1$). En particulier h' est strictement positive pour $a > 0$. Comme $h(0) = 0$, ceci assure que h est strictement positive sur $]0, +\infty[$. Donc α est l'unique racine de $h(a) = 1/a$. Le théorème de Dini assure que p.s. h_n converge uniformément vers h sur tout compact de $[0, \infty[$. Donc, l'unique racine de $h_n(a) = 1/a$, $\tilde{\alpha}_n$, converge p.s. vers α .

La convergence p.s. de l'estimateur $\tilde{\beta}_n$ vers β se traite comme dans la partie III.

▲

Exercice XII.8. I Modèle de référence

1. Comme $\mathbb{E}_p[\bar{Z}_n] = p$, on a en utilisant l'indépendance des variables aléatoires :

$$R(\bar{Z}_n, p) = \mathbb{E}_p[(\bar{Z}_n - p)^2] = \text{Var}_p(\bar{Z}_n) = \frac{1}{n} \text{Var}_p(Z_1) = \frac{p(1-p)}{n}.$$

2. En utilisant l'indépendance des variables aléatoires (Z_1, \dots, Z_n) , il vient :

$$p_n(z; p) = \mathbb{P}_p(Z_1 = z_1, \dots, Z_n = z_n) = \prod_{i=1}^n \mathbb{P}_p(Z_i = z_i) = p^{n\bar{z}_n} (1-p)^{n-n\bar{z}_n}.$$

La dérivée de la log-vraisemblance est :

$$\frac{\partial \log p_n(z; p)}{\partial p} = \frac{n\bar{z}_n}{p} - \frac{n - n\bar{z}_n}{1-p} = n \frac{\bar{z}_n - p}{p(1-p)}.$$

Elle s'annule en $p = \bar{z}_n$, est strictement positive (resp. négative) pour $p < \bar{z}_n$ (resp. $p > \bar{z}_n$). On en déduit donc que la log-vraisemblance atteint son unique maximum pour $p = \bar{z}_n$. L'estimateur du maximum de vraisemblance est donc \bar{Z}_n .

3. L'information de Fisher du modèle est :

$$I_n(p) = \mathbb{E}_p \left[\left(\frac{\partial \log p_n(z; p)}{\partial p} \right)^2 \right] = \frac{n^2}{p^2(1-p)^2} \text{Var}_p(\bar{Z}_n) = \frac{n}{p(1-p)}.$$

Comme \bar{Z}_n est un estimateur sans biais de p et que $R(\bar{Z}_n, p) = \frac{1}{I_n(p)}$, on en déduit que \bar{Z}_n est un estimateur de p efficace (dans la classe des estimateurs sans biais).

II Méthode de stratification**II.1 Étude à horizon fini**

1. Soit X_1 la réponse d'une personne interrogée au hasard. On a $\mathbb{E}_p[X_1] = p$. Cette personne appartient à la strate h avec probabilité N_h/N . En décomposant suivant la strate, S_1 , à laquelle elle appartient, on a :

$$\mathbb{E}_p[X_1] = \mathbb{P}_p(X_1 = 1) = \sum_{h=1}^H \mathbb{P}_p(X_1 = 1 | S_1 = h) \mathbb{P}_p(S_1 = h) = \sum_{h=1}^H p_h \frac{N_h}{N}.$$

On en déduit que $p = \sum_{h=1}^H p_h N_h / N$.

2. On a $\mathbb{E}_p[Y_h] = p_h$ et par linéarité, on a $\mathbb{E}_p[Y] = \sum_{h=1}^H \frac{N_h}{N} p_h = p$.
3. Comme les variables aléatoires $(Y_h, 1 \leq h \leq H)$ sont indépendantes de variance σ_h^2/n_h , on a :

$$R(Y, p) = \text{Var}_p(Y) = \sum_{h=1}^H \text{Var} \left(\frac{N_h}{N} Y_h \right) = \sum_{h=1}^H \frac{N_h^2}{N^2} \frac{1}{n_h} \sigma_h^2.$$

On peut décomposer la variance $p(1-p)$ en la somme de deux termes : la variance inter-strate $\sum_{h=1}^H (N_h/N)(p_h - p)^2$ et la variance intra-strate $\sum_{h=1}^H (N_h/N)\sigma_h^2$.

4. La méthode de stratification est bonne si son risque quadratique, $\text{Var}_p(Y)$, est inférieur au risque quadratique de \bar{Z}_n , à savoir $p(1-p)/n$. Cela revient à dire que Y est préférable à \bar{Z}_n . (Remarquer que le nombre de personnes interrogées est le même.)
5. On choisit $p = 1/2$, $H = 2$, $p_1 = 1/4$, $p_2 = 3/4$, $N_1 = N_2 = N/2$, $n_2 = n_1^2$, $n = n_1 + n_2$. Pour $n_1 \geq 5$, on a $\text{Var}_p(Y) = \frac{1}{4} \frac{3}{16} \left(\frac{1}{n_1} + \frac{1}{n_1^2} \right) \geq \frac{1}{4} \frac{1}{n_1 + n_1^2} = \text{Var}_p(\bar{Z}_n)$. L'estimateur \bar{Z}_n est donc strictement préférable à Y (dans la méthode de stratification, l'estimation de p_1 n'est pas assez précise).

6. On a $n \text{Var}_p(Y) = n \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{N}{nN_h} \sigma_h^2 = \sum_{h=1}^H \frac{N_h}{N} p_h (1 - p_h)$. L'inégalité de Cauchy-Schwarz avec $a_h = \sqrt{N_h/N}$ et $b_h = p_h \sqrt{N_h/N}$ assure que :

$$\sum_{h=1}^H \frac{N_h}{N} p_h (1 - p_h) = p - \sum_{h=1}^H \frac{N_h}{N} p_h^2 \leq p - \left(\sum_{h=1}^H \frac{N_h}{N} \right)^{-1} \left(\sum_{h=1}^H \frac{N_h}{N} p_h \right)^2 = p(1-p).$$

L'égalité a lieu si et seulement si $p_h = p$ pour tout $1 \leq h \leq H$.

7. On minimise $\sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{n_h}$ en (n_1, \dots, n_H) , sous la contrainte $\sum_{h=1}^H n_h = n$. Cela revient à minimiser :

$$\varphi(n_1, \dots, n_{H-1}) = \sum_{h=1}^{H-1} \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{n_h} + \left(\frac{N_H}{N} \right)^2 \frac{\sigma_H^2}{n - \sum_{h=1}^{H-1} n_h}$$

sur le compact $\Delta = \{(n_1, \dots, n_{H-1}) \in \mathbb{R}^{H-1}, n_h \geq 0 \text{ pour tout } 1 \leq h \leq H-1 \text{ et } \sum_{h=1}^{H-1} n_h \leq n\}$. Comme $\sigma_h^2 > 0$, on en déduit que φ prend la valeur $+\infty$ sur la frontière de Δ . La fonction φ atteint donc son minimum dans l'intérieur de Δ . Comme la fonction φ est de classe C^1 dans l'intérieur de Δ , on recherche les

zéros de son gradient. Comme $\sigma_h^2 > 0$ pour tout $1 \leq h \leq H$, on vérifie que la nullité du gradient de φ implique que les termes $N_h \sigma_h / (N n_h)$ pour $1 \leq h \leq H$ sont tous égaux, c'est-à-dire n_h proportionnel à $N_h \sigma_h$. Comme $\sum_{h=1}^H n_h = n$, on en déduit qu'il existe un seul zéro du gradient de φ et que φ atteint son unique minimum en ce point. On obtient donc que $\text{Var}_p(Y)$ est minimal pour $n_h = N_h \sigma_h / (\sum_{j=1}^H N_j \sigma_j)$ pour tout $1 \leq h \leq H$. Ce minimum vaut :

$$A = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{N_h \sigma_h / (\sum_{j=1}^H N_j \sigma_j)} = \left(\sum_{h=1}^H \frac{N_h}{N} \sigma_h \right)^2.$$

Par ailleurs pour l'allocation proportionnelle $\text{Var}_p(Y)$ est donné par $B = \sum_{h=1}^H \sigma_h^2 N_h / N$. L'inégalité de Cauchy-Schwarz avec $a_h = \sqrt{N_h / N}$ et $b_h = \sigma_h \sqrt{N_h / N}$ assure que $A \leq B$ avec égalité si et seulement si σ_h est constant c'est-à-dire si et seulement si p_h est égal à p ou $1 - p$ pour tout $1 \leq h \leq H$.

8. Le modèle utilisé pour construire l'estimateur Y est plus riche que celui utilisé pour construire \bar{Z}_n . Il n'est donc pas contradictoire de trouver dans ce modèle plus riche des estimateurs sans biais strictement préférables à \bar{Z}_n .

II.2 Étude asymptotique

1. Il découle de la loi forte des grands nombres que $(Y_h, n_h \geq 1)$ converge p.s. vers p_h , et du théorème central limite que $(\sqrt{n_h}(Y_h - p_h), n_h \geq 1)$ converge en loi vers une loi gaussienne $\mathcal{N}(0, \sigma_h^2)$.
2. Comme il y a un nombre fini (fixe) de strates, le résultat découle de la question précédente.
3. En utilisant l'indépendance des variables Y_h , il vient :

$$\psi_{\frac{Y-p}{\sqrt{\text{Var}_p(Y)}}}(u) = \prod_{h=1}^H \psi_{\frac{N_h}{N} \frac{Y_h - p_h}{\sqrt{\text{Var}_p(Y)}}}(u) = \prod_{h=1}^H \psi_{\sqrt{n_h}(Y_h - p_h)}(c_h u),$$

où :

$$c_h = \frac{N_h}{N} \frac{1}{\sqrt{n_h \text{Var}_p(Y)}} = \left(\frac{n_h}{N_h^2} \sum_{j=1}^H \frac{N_j^2}{n_j} \sigma_j^2 \right)^{-1/2}.$$

On remarque que $0 \leq c_h \leq 1/\sigma_h$. En utilisant la convergence uniforme locale des fonctions caractéristiques, il vient pour tout $|u| \leq K \min_{1 \leq h \leq H} \sigma_h$:

$$\begin{aligned} \psi_{\frac{Y-p}{\sqrt{\text{Var}_p(Y)}}}(u) &= \prod_{h=1}^H \left(e^{-\sigma_h^2 c_h^2 u^2 / 2} + R_h(c_h u) \right) = \prod_{h=1}^H e^{-\sigma_h^2 c_h^2 u^2 / 2} + R(u) \\ &= e^{-u^2 / 2} + R(u), \end{aligned}$$

où la fonction R est définie par la deuxième égalité. On déduit de la convergence $\lim_{n_h \rightarrow \infty} \sup_{|u| \leq K} |R_h(u)| = 0$, que $\lim_{\min_{1 \leq h \leq H} n_h \rightarrow \infty} \sup_{|u| \leq K} |R(u)| = 0$. Ceci assure que $(Y - p)/\sqrt{\text{Var}_p(Y)}$ converge en loi vers une loi gaussienne centrée réduite $\mathcal{N}(0, 1)$ quand $\min_{1 \leq h \leq H} n_h$ tend vers l'infini.

4. Pour tout $1 \leq h \leq H$, on a que p.s. $(N_h/N)^2 Y_h(1 - Y_h)$ converge vers $(N_h/N)^2 \sigma_h^2$ quand $\min_{1 \leq h \leq H} n_h$ tend vers l'infini. En particulier, pour $\varepsilon > 0$ et pour $\min_{1 \leq h \leq H} n_h$ suffisamment grand, on a $|Y_h(1 - Y_h) - \sigma_h^2| \leq \varepsilon \sigma_h^2$ et donc :

$$\left| \frac{\sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \frac{Y_h(1-Y_h)}{n_h}}{\text{Var}_p(Y)} - 1 \right| \leq \frac{\sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \frac{|Y_h(1-Y_h) - \sigma_h^2|}{n_h}}{\text{Var}_p(Y)} \leq \varepsilon.$$

Ceci assure que $\frac{1}{\text{Var}_p(Y)} \sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \frac{Y_h(1 - Y_h)}{n_h}$ converge p.s. vers 1, quand $\min_{1 \leq h \leq H} n_h$ tend vers l'infini.

5. On pose :

$$S^2 = \sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \frac{Y_h(1 - Y_h)}{n_h}.$$

On déduit du théorème de Slutsky et de la question précédente que quand $\min_{1 \leq h \leq H} n_h$ tend vers l'infini, alors $(Y - p)/S$ converge en loi vers une loi gaussienne G centrée réduite $\mathcal{N}(0, 1)$. Comme la loi gaussienne est continue, on en déduit que quand $\min_{1 \leq h \leq H} n_h$ tend vers l'infini, alors $\mathbb{P}_p(p \in I) = \mathbb{P}_p((Y - p)/S \in [\pm\phi_{1-\alpha/2}])$ converge vers $\mathbb{P}(G \in [\pm\phi_{1-\alpha/2}]) = 1 - \alpha$. Ceci assure que I est un intervalle de confiance sur p de niveau asymptotique $1 - \alpha$.



Exercice XII.9. 1. On a $\mathbb{P}_1(Y = k) = 1/k(k + 1)$ pour $k \in \mathbb{N}^*$. Le nombre théorique attendu d'artistes ayant eu k disques d'or est $n/k(k + 1)$. Les données théoriques sont reportées dans le tableau XIII.5 .

2. On utilise le test du χ^2 d'adéquation à une loi. Soit Y de loi de Yule de paramètre $\rho = 1$.

a) Modèle : $(X_i, i \in \mathbb{N}^*)$ variables aléatoires indépendantes de même loi à valeurs dans $\{1, \dots, m\}$.

b) Hypothèses : $H_0 = \left\{ X_1 \stackrel{\text{en loi}}{=} \min(Y, m) \right\}$ et $H_1 = \left\{ X_1 \stackrel{\text{en loi}}{\neq} \min(Y, m) \right\}$.

c) Statistiques de test :

$$\zeta_n^{(1)} = n \sum_{k=1}^m \frac{(p_{\text{obs}}(k) - p_{\text{théo}}(k))^2}{p_{\text{théo}}(k)} \quad \text{et} \quad \zeta_n^{(2)} = n \sum_{k=1}^m \frac{(p_{\text{obs}}(k) - p_{\text{théo}}(k))^2}{p_{\text{obs}}(k)},$$

| Nombre de disques d'or | Nombre d'artistes | Nombre de disques d'or | Nombre d'artistes |
|------------------------|-------------------|------------------------|-------------------|
| 1 | 668 (688.5) | 10 | 14 (12.5) |
| 2 | 244 (229.5) | 11 | 16 (10.4) |
| 3 | 119 (114.7) | 12 | 13 (8.8) |
| 4 | 78 (68.8) | 13 | 11 (7.6) |
| 5 | 55 (45.9) | 14 | 5 (6.6) |
| 6 | 40 (32.8) | 15 | 4 (5.7) |
| 7 | 24 (24.6) | 16 | 4 (5.1) |
| 8 | 32 (19.1) | 17 et + | 26 (81.0) |
| 9 | 24 (15.3) | | |

Table XIII.5. Nombres d'artistes observés (et théoriques entre parenthèses, si le nombre de disques d'or suit une loi de Yule de paramètre 1) par nombre de disques d'or.

avec $p_{\text{obs}}(k) = n_k/n$ où n_k est le nombre d'artistes ayant eu k disques d'or si $1 \leq k \leq m-1$ et n_m est le nombre d'artistes ayant eu au moins m disques d'or, et $p_{\text{théo}}(k) = \mathbb{P}_1(\min(Y, m) = k)$.

- d) Les régions critiques de niveau asymptotique α sont $W_n^{(i)} = \{\zeta_n^{(i)} \geq z_{1-\alpha}\}$ où $z_{1-\alpha}$ est le quantile d'ordre $1 - \alpha$ de la loi du χ^2 à $m - 1$ degrés de liberté. Les tests sont convergents.
- e) Les p -valeurs asymptotiques sont $\mathbb{P}(Z \geq \zeta_n^{(i), \text{obs}})$ où Z suit la loi du χ^2 à $m - 1$ degrés de liberté et $\zeta_n^{(i), \text{obs}}$ est la valeur de la statistique de test $\zeta_n^{(i)}$ évaluée en les données.

3. On regroupe les cases k pour lesquelles on $np_k(1 - p_k) \leq 5$. Comme $p_k \simeq k^{-2}$ pour k grand, on regroupe les cases telles que $k \geq \sqrt{n/5}$ soit $k \geq 17$ environ.

4. On remarque que $\mathbb{P}(Y = \ell) = \Gamma(\ell)\rho \prod_{k=1}^{\ell} \frac{1}{\rho + k}$ et donc :

$$\log(\mathbb{P}(Y = \ell)) = \log(\Gamma(\ell)) + \log(\rho) - \sum_{k=1}^{\ell} \log(\rho + k).$$

Soit $y = (y_1, \dots, y_n) \in (\mathbb{N}^*)^n$. On en déduit que la log-vraisemblance est :

$$\begin{aligned} L_n(y; \rho) &= \sum_{i=1}^n \log(\Gamma(y_i)) + \log(\rho) - \sum_{k=1}^{y_i} \log(\rho + k) \\ &= N_1 \log(\rho) - \sum_{k=1}^{\infty} N_k \log(\rho + k) + \sum_{i=1}^n \log(\Gamma(y_i)). \end{aligned}$$

5. La condition $x^2 \geq (x+c)(x+c-1)$ est équivalente à $c-c^2 \geq (2c-1)x$. Cette condition a lieu pour tout $x > 1$ si et seulement si $2c-1 \leq 0$ et $c-c^2 \geq 0$ soit $c \leq 1/2$. On a :

$$\sum_{k=1}^{\infty} \frac{1}{(\rho+k)^2} \leq \sum_{k=1}^{\infty} \frac{1}{(\rho+k+\frac{1}{2})(\rho+k-\frac{1}{2})} = \frac{1}{\rho+\frac{1}{2}}.$$

On a :

$$\begin{aligned} \frac{\partial^2 L_n(y; \rho)}{\partial^2 \rho} &= \frac{-N_1}{\rho^2} + \sum_{k=1}^{\infty} \frac{N_k}{(\rho+k)^2} \\ &\leq N_1 \left(-\frac{1}{\rho^2} + \frac{1}{\rho+\frac{1}{2}} \right) \\ &= \frac{N_1}{\rho^2(2\rho+1)} (2\rho^2 - 2\rho - 1). \end{aligned}$$

La quantité $2\rho^2 - 2\rho - 1$ est négative sur $]0, (1 + \sqrt{3})/2]$, on en déduit que la fonction $L_n(y; \cdot)$ est convexe sur cette ensemble. Comme $\lim_{\rho \rightarrow 0} L_n(y; \rho) = -\infty$, on obtient que la log-vraisemblance possède un seul maximum local sur $]0, (1 + \sqrt{3})/2]$.

On remarque que :

$$L_n(y; \rho) \leq N_1 \log(1 + 1/\rho) - N_2 \log(\rho + 2) + \sum_{i=1}^n \log(\Gamma(y_i)).$$

En particulier, si $N_2 \geq 1$, on en déduit que $\lim_{\rho \rightarrow \infty} L_n(y; \rho) = -\infty$. Donc la log-vraisemblance possède au moins un maximum sur $]0, \infty[$.

6. On utilise le test du χ^2 d'adéquation à une famille de loi.
- Modèle : $(X_i, i \in \mathbb{N}^*)$ variables aléatoires indépendantes de même loi à valeurs dans $\{1, \dots, m\}$.
 - Hypothèses :

$$\begin{aligned} H_0 &= \left\{ X_1 \stackrel{\text{en loi}}{=} \min(Y, m) \text{ où } Y \text{ est de loi de Yule de paramètre } \rho > 0 \right\}, \\ H_1 &= \left\{ X_1 \stackrel{\text{en loi}}{\neq} \min(Y, m) \text{ pour tout } Y \text{ de loi de Yule de paramètre } \rho > 0 \right\}. \end{aligned}$$

- c) Statistiques de test :

$$\zeta_n^{(1)} = n \sum_{k=1}^m \frac{(p_{\text{obs}}(k) - \hat{p}_{\text{théo}}(k))^2}{\hat{p}_{\text{théo}}(k)} \quad \text{ou} \quad \zeta_n^{(2)} = n \sum_{k=1}^m \frac{(p_{\text{obs}}(k) - \hat{p}_{\text{théo}}(k))^2}{p_{\text{obs}}(k)},$$

avec $p_{\text{obs}}(k) = n_k/n$ où n_k est le nombre d'artistes ayant eu k disques d'or si $1 \leq k \leq m-1$ et n_m est le nombre d'artistes ayant eu au moins m disques d'or, et $\hat{p}_{\text{théo}}(k) = \mathbb{P}_1(\min(Y, m) = k)$, où Y est de loi de Yule de paramètre $\hat{\rho}$, l'estimateur du maximum de vraisemblance de ρ .

- d) Les régions critiques de niveau asymptotique α sont $W_n^{(i)} = \{\zeta_n^{(i)} \geq z_{1-\alpha}\}$ où $z_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ de la loi du χ^2 à $m-2$ degrés de liberté. Les tests sont convergents.
- e) Les p -valeurs asymptotiques sont $\mathbb{P}(Z \geq \zeta_n^{(i), \text{obs}})$ où Z suit la loi du χ^2 à $m-2$ degrés de liberté et $\zeta_n^{(i), \text{obs}}$ est la valeur de la statistique de test $\zeta_n^{(i)}$ évaluée en les données.
7. On rejette H_0 (au niveau 10^{-5}). La notoriété (que l'on a modélisée par la loi de Yule) ne suffit pas à expliquer le nombre d'or. Autrement dit, le talent existe.
8. La distribution du nombre d'artistes ayant moins de 18 disques d'or peut être modélisée par la loi de Yule (conditionnée à prendre des valeurs inférieures à 18), car la p -valeur est élevée. Une conclusion est que le talent existe pour les artistes ayant au moins 18 disques d'or.

▲

Exercice XII.10. I Préliminaires

1. On a $\mathbb{E}_\theta[P^v(1-P)^w] = \frac{1}{B(a,b)} \int_{]0,1[} x^{a+v-1}(1-x)^{b+w-1} = \frac{B(a+v, b+w)}{B(a, v)}$ et en particulier $\mathbb{E}_\theta[P] = a/(a+b)$.
2. Un enfant pris au hasard dans une famille donnée a une probabilité P d'être un garçon ; la famille étant choisie au hasard, l'enfant a une probabilité $\mathbb{E}_\theta[P]$ d'être un garçon.
3. La loi conditionnelle de X_n sachant P est la loi binomiale de paramètre (n, P) . Autrement dit $\mathbb{E}[g(X_n) | P] = \psi(P)$ où $\psi(p) = \mathbb{E}[g(Z_p)]$ où Z_p est une variable aléatoire de loi binomiale de paramètre (n, p) .
4. On a :

$$\begin{aligned} \mathbb{E}_\theta[X_n] &= \mathbb{E}_\theta[\mathbb{E}[X_n | P]] = \mathbb{E}_\theta[nP] = n\mathbb{E}_\theta[P], \\ \text{Var}_\theta(X_n) &= \mathbb{E}_\theta[X_n^2] - \mathbb{E}_\theta[X_n]^2 \\ &= \mathbb{E}_\theta[\mathbb{E}[X_n^2 | P] - \mathbb{E}[X_n | P]^2] + \mathbb{E}_\theta[nP^2] - \mathbb{E}_\theta[nP]^2 \\ &= \mathbb{E}_\theta[nP(1-P)] + n^2 \text{Var}_\theta(P), \end{aligned}$$

où l'on a utilisé pour la dernière égalité que la variance d'une loi binomiale de paramètre (n, p) est $np(1-p)$.

5. Cela découle des résultats suivants, voir la question I.1, $\mathbb{E}_\theta[P] = a/(a+b)$ et :

$$\mathbb{E}_\theta[P^2] = \frac{a(a+1)}{(a+b)(a+b+1)} \quad \text{et} \quad \mathbb{E}_\theta[P(1-P)] = \frac{ab}{(a+b)(a+b+1)}.$$

II Estimation des paramètres

1. Si $n = 1$, la loi de X_n est la loi de Bernoulli de paramètre $p_0 = a/(a+b)$. Le paramètre p_0 est identifiable, mais pas θ (il existe plusieurs valeurs possibles de θ pour une même valeur de p_0).
2. On utilise la méthode des moments : on résout les équations :

$$m = \mathbb{E}_\theta[P] = n \frac{a}{a+b} \quad \text{et} \quad v = \text{Var}_\theta(P) = \frac{m(n-m)}{n} \left(1 + \frac{n-1}{a+b+1} \right)$$

pour obtenir :

$$a = m \frac{m(n-m) - v}{nv - m(n-m)} \quad \text{et} \quad b = (n-m) \frac{m(n-m) - v}{nv - m(n-m)}.$$

L'estimateur des moments de $\theta = (a, b)$ est donc $\hat{\theta}_K = (\hat{a}_K, \hat{b}_K)$ avec :

$$\hat{a}_K = M_K \frac{M_K(n - M_K) - V_k}{nV_k - M_k(n - M_K)} \quad \text{et} \quad \hat{b}_K = (n - M_K) \frac{M_K(n - M_K) - V_k}{nV_k - M_k(n - M_K)}.$$

3. La convergence p.s. de l'estimateur des moments, $\hat{\theta}_K$, vers θ se déduit des convergences p.s. de la moyenne empirique et de la variance empirique, et de la continuité de la fonction $(m, v) \mapsto (a, b)$ en $(\mathbb{E}_\theta[P], \text{Var}_\theta(P))$.
4. La normalité asymptotique de l'estimateur des moments, $\hat{\theta}_K$, de θ se déduit de la normalité asymptotique du couple formé de la moyenne empirique et de la variance empirique ainsi que de la propriété \mathcal{C}^1 de la fonction $(m, v) \mapsto (a, b)$ en $(\mathbb{E}_\theta[P], \text{Var}_\theta(P))$.
5. Les vérifications sont évidentes. Il vient $KM_K = 13\,037$ et $V_K \simeq 1.23$, soit $a \simeq 50.7$, $b \simeq 48.9$ et $p_0 \simeq 0.508$.

III Estimation de la probabilité d'avoir un garçon

1. On a :

$$\mathbb{E}_\theta[h(P) \mid X_n = k] = \frac{\int_{]0,1[} h(p) p^{a+k-1} (1-p)^{b+n-k-1} dp}{\int_{]0,1[} p^{a+k-1} (1-p)^{b+n-k-1} dp} = \mathbb{E}_{(a+k, b+n-k)}[h(P)].$$

2. On déduit de la question précédente que la loi conditionnelle de P sachant X_n est la loi β de paramètre $(a + X_n, b + n - X_n)$. Comme l'espérance d'une variable aléatoire de loi β de paramètre (a', b') est $a'/(a' + b')$, on en déduit que $P_n = \mathbb{E}[P|X_n] = (a + X_n)/(a + b + n)$.
3. On note $Z_i = 1$ si le i -ème enfant de la famille est un garçon et $Z_i = 0$ sinon. On a $X_n = n\bar{Z}_n$ où $n\bar{Z}_n = \sum_{i=1}^n Z_i$. Conditionnellement à P , les variables aléatoires $(Z_i, i \in \mathbb{N}^*)$ sont indépendantes et de même loi de Bernoulli de paramètre P . La loi forte des grands nombres assure alors que, conditionnellement à P , la suite $(\bar{Z}_n, n \in \mathbb{N}^*)$ converge p.s. vers $P : \mathbb{E}[\mathbf{1}_{\{\lim_{n \rightarrow \infty} \bar{Z}_n = P\}} | P] = 1$. En prenant l'espérance dans cette dernière égalité, on obtient $\mathbb{P}_\theta(\lim_{n \rightarrow \infty} \bar{Z}_n = P) = 1$. Ceci assure que la suite $(X_n/n, n \in \mathbb{N})$ converge p.s. vers P .
4. Comme $P_n = \frac{\frac{X_n}{n} + \frac{a}{n}}{1 + \frac{a+b}{n}}$, on en déduit que p.s. $\lim_{n \rightarrow \infty} P_n = P$.
5. On a :

$$\begin{aligned} \mathbb{E}_\theta[(P - h(X_n))^2] &= \mathbb{E}_\theta[((P - P_n) + (P_n - h(X_n)))^2] \\ &= \mathbb{E}_\theta[(P - P_n)^2] + \mathbb{E}_\theta[(P_n - h(X_n))^2] \\ &\quad + 2\mathbb{E}_\theta[(P - P_n)(P_n - h(X_n))] \\ &= \mathbb{E}_\theta[(P - P_n)^2] + \mathbb{E}_\theta[(P_n - h(X_n))^2] \\ &\geq \mathbb{E}_\theta[(P - P_n)^2], \end{aligned}$$

où l'on a utilisé pour la troisième égalité, grâce aux propriétés de l'espérance conditionnelle, que :

$$\begin{aligned} \mathbb{E}_\theta[(P - P_n)(P_n - h(X_n))] &= \mathbb{E}_\theta \left[\mathbb{E}_\theta[(P - P_n)(P_n - h(X_n)) | X_n] \right] \\ &= \mathbb{E}_\theta \left[\mathbb{E}_\theta[(P - P_n) | X_n] (P_n - h(X_n)) \right] = 0. \end{aligned}$$

6. Comme la loi conditionnelle de P sachant X_n est la loi β de paramètre $(a + X_n, b + n - X_n)$, il faut choisir pour c le quantile d'ordre 10% de la loi β de paramètre $(a + X_n, b + n - X_n)$.

▲

Exercice XII.11. On note F la fonction de répartition de X .

I Préliminaires

1. Comme (X, X') a même loi que (X', X) , il vient :

$$\mathbb{P}(-V \leq x) = \mathbb{P}(X' - X \leq x) = \mathbb{P}(X - X' \leq x) = \mathbb{P}(V \leq x).$$

Les fonctions de répartition de V et $-V$ sont identiques, donc V et $-V$ ont même loi. On a $1 - \mathbb{P}(V < -x) = \mathbb{P}(V \geq -x) = \mathbb{P}(-V \leq -x) = \mathbb{P}(V \leq x)$.

2. On a :

$$2\mathbb{P}(V \leq 0) = 1 + \mathbb{P}(V \leq 0) - \mathbb{P}(V < 0) = 1 + \mathbb{P}(V = 0)$$

et :

$$2\mathbb{P}(V \leq -\varepsilon) = 1 + \mathbb{P}(V \leq -\varepsilon) - \mathbb{P}(V < \varepsilon) = 1 - \mathbb{P}(V \in]-\varepsilon, \varepsilon]).$$

3. La première inégalité découle de $\{X \in]a \pm \varepsilon/2[\} \cap \{X' \in]a \pm \varepsilon/2[\} \subset \{X - X' \in]-\varepsilon, \varepsilon[\}$. Comme X et X' sont indépendants et de même loi, on en déduit que pour tout $a \in \mathbb{R}$, $\varepsilon > 0$:

$$\mathbb{P}(V \in]-\varepsilon, \varepsilon]) \geq \mathbb{P}(X \in]a \pm \varepsilon/2])^2 \geq (F(a + \varepsilon/4) - F(a - \varepsilon/4))^2.$$

Comme $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow \infty} F(x) = 1$, on en déduit qu'il existe $a \in \mathbb{R}$ tel que $F(a + \varepsilon/4) > F(a - \varepsilon/4)$. Ceci assure la seconde inégalité.

4. On déduit de la question I.2 que $\mathbb{P}(V \leq 0) \geq 1/2$, ce qui assure que la médiane est inférieure ou égale à 0. On déduit des questions I.2 et I.3 que pour tout $\varepsilon > 0$, $\mathbb{P}(V \leq -\varepsilon) < 1/2$, ce qui assure que la médiane est supérieure à $-\varepsilon$. En conclusion, la médiane est nulle.

5. Soit $v \in \mathbb{R}$. On a :

$$\mathbb{P}(V = v) = \mathbb{E}[\mathbf{1}_{\{X - X' = v\}}] = \mathbb{E}[\varphi(X)],$$

où $\varphi(x) = \mathbb{E}[\mathbf{1}_{\{x - X' = v\}}] = \mathbb{P}(X' = x - v) = 0$. On en déduit que $\mathbb{P}(V = v) = 0$.

6. Comme $\mathbb{P}(V = 0) = 0$, il vient :

$$\mathbb{P}(V > 0) = \mathbb{P}(-V \leq 0) = \mathbb{P}(V \leq 0) = \mathbb{P}(V < 0) = \mathbb{P}(V \leq 0) = 1/2,$$

où l'on a utilisé que V et $-V$ ont même loi pour la deuxième égalité et (XII.9) pour la dernière. Comme $\{\text{sgn}(V) = 1\} = \{V > 0\}$ et $\{\text{sgn}(V) = -1\} = \{V < 0\}$, on en déduit que $\text{sgn}(V)$ est uniforme sur $\{-1, 1\}$.

7. On a, en utilisant que V et $-V$ ont même loi :

$$\begin{aligned} \mathbb{E}[g(\text{sgn}(V))f(|V|)] &= g(1)\mathbb{E}[f(V)\mathbf{1}_{\{V > 0\}}] + g(-1)\mathbb{E}[f(-V)\mathbf{1}_{\{V < 0\}}] \\ &= (g(1) + g(-1))\mathbb{E}[f(V)\mathbf{1}_{\{V > 0\}}] \\ &= \mathbb{E}[g(\text{sgn}(V))]2\mathbb{E}[f(V)\mathbf{1}_{\{V > 0\}}]. \end{aligned}$$

Avec $g = 1$ dans l'égalité précédente, il vient $\mathbb{E}[f(|V|)] = 2\mathbb{E}[f(V)\mathbf{1}_{\{V > 0\}}]$ et donc :

$$\mathbb{E}[g(\text{sgn}(V))f(|V|)] = \mathbb{E}[g(\text{sgn}(V))]\mathbb{E}[f(|V|)].$$

Ceci assure que $\text{sgn}(V)$ et $|V|$ sont indépendants.

II Test des signes

1. Sous H_0 , les variables aléatoires $(S_n, n \in \mathbb{N}^*)$ sont indépendantes de loi uniforme sur $\{-1, 1\}$. En particulier, elles sont de carré intégrable et on a $\mathbb{E}[S_k] = 0$ et $\text{Var}(S_k) = 1$. On déduit du théorème central limite que sous H_0 , la suite $(\zeta_n, n \in \mathbb{N}^*)$ converge en loi vers une variable aléatoire gaussienne de loi $\mathcal{N}(0, 1)$.
2. Avec les notations de la partie I, il vient :

$$\begin{aligned} \mathbb{E}[S_k] &= \mathbb{P}(X_k > Y_k) - \mathbb{P}(X_k < Y_k) = \mathbb{P}(X > X' - \mu) - \mathbb{P}(X < X' - \mu) \\ &= \mathbb{P}(-V > -\mu) - \mathbb{P}(V < -\mu) \\ &= \mathbb{P}(V < \mu) - \mathbb{P}(V \geq -\mu) \\ &= \mathbb{P}(V \in]-\mu, \mu[). \end{aligned}$$

L'inégalité (XII.10) assure alors que, sous H_1 , $\mathbb{E}[S_k] > 0$. La loi forte des grands nombres assure que p.s. $\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} \zeta_n = \mathbb{E}[S_1] > 0$. On en déduit donc que la suite $(\zeta_n, n \in \mathbb{N}^*)$ converge p.s. vers $+\infty$.

3. La région critique est $W_n = \{\zeta_n \geq a\}$. Pour $a = \phi_{1-\alpha}$, le quantile d'ordre $1 - \alpha$ de la loi gaussienne $\mathcal{N}(0, 1)$, la question II.1 assure que le test de région critique W_n est de niveau asymptotique α .
4. La question II.2 assure que sous H_1 , p.s. $\lim_{n \rightarrow \infty} \mathbf{1}_{\{\zeta_n \geq a\}} = 1$. On déduit du théorème de convergence dominée que, sous H_1 :

$$\lim_{n \rightarrow \infty} \mathbb{P}(\zeta_n \geq a) = 1.$$

Le test est donc convergent.

5. La p -valeur asymptotique est donnée par $p\text{-val} = \mathbb{P}(G \geq \zeta_n^{\text{obs}})$ où G est de loi gaussienne $\mathcal{N}(0, 1)$ et ζ_n^{obs} est la statistique de test évaluée sur les données. On a $\zeta_n^{\text{obs}} \simeq 2.53$, la p -valeur asymptotique correspondante est environ 5.7%. Il est raisonnable d'accepter H_0 car la p -valeur n'est pas très petite (elle est supérieure à 5%).
6. Sous H_0 , les variables aléatoires $((S_n + 1)/2, n \in \mathbb{N}^*)$ sont indépendantes de loi de Bernoulli. On en déduit que sous H_0 , $(n/2) + \sum_{k=1}^n (S_k/2)$ est de loi binomiale de paramètre $(n, 1/2)$. Soit Z de loi binomiale de paramètre $(n, 1/2)$. Après avoir remarqué que $n - Z$ a même loi que Z , on a (avec $n = 10$) :

$$\mathbb{P}(\zeta_n \geq \zeta_n^{\text{obs}}) = \mathbb{P}(Z \geq 9) = \mathbb{P}(Z \leq 1) = 2^{-10} + \binom{10}{1} 2^{-10} \simeq 0.1\%.$$

Comme $p\text{-val} \simeq 0.1\%$, on rejette donc H_0 . On voit que l'approche asymptotique est de mauvaise qualité pour des petits échantillons.

III Test de Wilcoxon

1. On déduit de la question I.5 avec $X = |V_k|$ et $X' = |V_\ell|$ et $v = 0$ que $\mathbb{P}(|V_k| = |V_\ell|) = 0$. Il vient :

$$\mathbb{P}(\exists k \neq \ell |V_k| = |V_\ell|) \leq \sum_{k \neq \ell} \mathbb{P}(|V_k| = |V_\ell|) = 0.$$

Ceci assure l'unicité de la permutation τ_n .

2. Les variables aléatoires $(|V_1|, \dots, |V_n|)$ et (S_1, \dots, S_n) sont indépendantes. Comme τ_n est une fonction de $(|V_1|, \dots, |V_n|)$, on en déduit que τ_n est indépendant de (S_1, \dots, S_n) .
3. Soit $s_1, \dots, s_n \in \{-1, 1\}$. Comme τ_n et (S_1, \dots, S_n) sont indépendants, on a :

$$\begin{aligned} \mathbb{P}(S_{\tau_n(1)} = s_1, \dots, S_{\tau_n(n)} = s_n) &= \sum_{\sigma \in \mathcal{S}_n} \mathbb{P}(\tau_n = \sigma, S_{\sigma(1)} = s_1, \dots, S_{\sigma(n)} = s_n) \\ &= \sum_{\sigma \in \mathcal{S}_n} \mathbb{P}(\tau_n = \sigma) \mathbb{P}(S_{\sigma(1)} = s_1, \dots, S_{\sigma(n)} = s_n) \\ &= \sum_{\sigma \in \mathcal{S}_n} \mathbb{P}(\tau_n = \sigma) 2^{-n} \\ &= 2^{-n} = \mathbb{P}(S_1 = s_1, \dots, S_n = s_n). \end{aligned}$$

On en déduit que $(S_{\tau_n(1)}, \dots, S_{\tau_n(n)})$ a même loi que (S_1, \dots, S_n) .

4. On a :

$$T_n = \sum_{k=1}^n R_n(k) \mathbf{1}_{\{S_k > 0\}} = \sum_{k=1}^n k \mathbf{1}_{\{S_{\tau_n(k)} > 0\}}.$$

On remarque que si U est uniforme sur $\{-1, 1\}$, alors $\mathbf{1}_{\{U > 0\}}$ est de loi de Bernoulli de paramètre $1/2$. On déduit des questions III.3 puis I.6 que $(\mathbf{1}_{\{S_{\tau_n(1)} > 0\}}, \dots, \mathbf{1}_{\{S_{\tau_n(n)} > 0\}})$ a même loi que $(\mathbf{1}_{\{S_1 > 0\}}, \dots, \mathbf{1}_{\{S_n > 0\}})$ et donc même loi que (Z_1, \dots, Z_n) .

5. Par linéarité, il vient $\mathbb{E}[T'_n] = \sum_{k=1}^n k/2 = n(n+1)/4$. En utilisant l'indépendance, il vient :

$$\text{Var}(T'_n) = \sum_{k=1}^n \text{Var}(kZ_k) = \sum_{k=1}^n k^2/4 = n(n+1)(2n+1)/24 = \sigma_n^2.$$

6. On a :

$$\begin{aligned}
 \psi_n(u) &= \mathbb{E} \left[e^{iu\xi_n} \right] = \mathbb{E} \left[e^{iu(T'_n - \frac{n(n+1)}{4})/\sigma_n} \right] \\
 &= \mathbb{E} \left[\prod_{k=1}^n e^{iuk(Z_k - \frac{1}{2})/\sigma_n} \right] \\
 &= \prod_{k=1}^n \mathbb{E} \left[e^{iuk(Z_k - \frac{1}{2})/\sigma_n} \right] \\
 &= \prod_{k=1}^n \frac{1}{2} \left(e^{-iuk/(2\sigma_n)} + e^{iuk/(2\sigma_n)} \right) \\
 &= \exp \left(\sum_{k=1}^n \log \left(\cos \left(\frac{uk}{2\sigma_n} \right) \right) \right),
 \end{aligned}$$

où l'on a utilisé que T_n a même loi que T'_n pour la deuxième égalité, et l'indépendance pour la quatrième égalité.

7. Pour n suffisamment grand, on a $|u|k/2\sigma_n \leq 1/2$ pour tout $k \in \{1, \dots, n\}$. En utilisant un développement limité des fonctions cosinus (au voisinage de 0) et logarithme (au voisinage de 1), on obtient :

$$\begin{aligned}
 \sum_{k=1}^n \log \left(\cos \left(\frac{uk}{2\sigma_n} \right) \right) &= \sum_{k=1}^n \log \left(1 - \frac{u^2 k^2}{8\sigma_n^2} + \frac{1}{n^2} g(n, k) \right) \\
 &= \sum_{k=1}^n -\frac{u^2 k^2}{8\sigma_n^2} + \frac{1}{n^2} h(n, k) \\
 &= -\frac{u^2}{2} + O(n^{-1}),
 \end{aligned}$$

où les fonctions g et h sont bornées sur $\{1 \leq k \leq n\}$. On en déduit que $\lim_{n \rightarrow \infty} \psi_n(u) = e^{-u^2/2}$. Ceci assure la convergence en loi de $(\xi_n, n \in \mathbb{N}^*)$ vers une variable aléatoire gaussienne $\mathcal{N}(0, 1)$.

8. La région critique est $W'_n = \{\xi_n \geq a\}$. Pour $a = \phi_{1-\alpha}$, le quantile d'ordre $1 - \alpha$ de la loi gaussienne $\mathcal{N}(0, 1)$, la question précédente assure que le test de région critique W'_n est de niveau asymptotique α . Le test est convergent car sous H_1 ,

$$\lim_{n \rightarrow \infty} \mathbb{P}(\xi_n \geq a) = 1.$$

9. La p -valeur asymptotique est donnée par $\mathbb{P}(G \geq \xi_n^{\text{obs}})$ où G est de loi gaussienne $\mathcal{N}(0, 1)$ et ξ_n^{obs} est la statistique de test évaluée sur les données. On a $\xi_n^{\text{obs}} \simeq 2.40$, la p -valeur asymptotique correspondante est environ 7.2%. Il est raisonnable d'accepter H_0 car la p -valeur n'est pas très faible. La p -valeur

exacte, utilisant la loi de ξ_n et non son approximation asymptotique, est de 0.7% ; on rejette alors H_0 , car la p -valeur est faible. On retrouve que l'erreur de l'approximation est, pour de petits échantillons, importante. Les p -valeurs sont semblables à celles de la question II.5. Pour comparer ces deux tests, il faudrait comparer les puissances.

▲

Exercice XII.12.

I Préliminaires

1. Il vient avec le changement de variable $y = \sqrt{1-v} x$:

$$\begin{aligned} \mathbb{E} \left[\exp \left(\frac{v}{2} G^2 \right) \right] &= \int_{\mathbb{R}} e^{-\frac{1}{2}(1-v)x^2} \frac{dx}{\sqrt{2\pi}} = \frac{1}{\sqrt{1-v}} \int_{\mathbb{R}} e^{-\frac{1}{2}(1-v)y^2} \frac{dy}{\sqrt{2\pi/(1-v)}} \\ &= \frac{1}{\sqrt{1-v}}, \end{aligned}$$

où l'on a reconnu que $e^{-\frac{1}{2}(1-v)y^2} / \sqrt{2\pi/(1-v)}$ est la densité de $(1-v)G$.

2. On remarque que les variables aléatoires $(\varepsilon_{2k-2}\varepsilon_{2k-1}, k \geq 2)$ sont indépendantes, de même loi intégrable et d'espérance nulle. On déduit de la loi forte des grands nombres que p.s. $\lim_{n \rightarrow +\infty} \frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} \varepsilon_{2k-2}\varepsilon_{2k-1} = 0$. De manière similaire, on obtient que p.s. $\lim_{n \rightarrow +\infty} \frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} \varepsilon_{2k-1}\varepsilon_{2k} = 0$. On en déduit donc que la suite $(T_n/n, n \in \mathbb{N}^*)$ converge p.s. vers 0.

3. Il découle de la question I.1 et de l'indépendance des variables aléatoires $(\varepsilon_k, k \in \mathbb{N}^*)$ que :

$$\mathbb{E}[N_n(u)] = \left(1 - \frac{u^2\sigma^2}{n} \right)^{-n/2}.$$

En remarquant que $N_n(u)^2 = N_n(\sqrt{2}u)$, on en déduit que :

$$\text{Var}(N_n(u)) = \left(1 - \frac{2u^2\sigma^2}{n} \right)^{-n/2} - \left(1 - \frac{u^2\sigma^2}{n} \right)^{-n} = e^{u^2\sigma^2+o(1)} - e^{u^2\sigma^2+o(1)}.$$

On a donc $\lim_{n \rightarrow +\infty} \text{Var}(N_n(u)) = 0$.

4. Comme $N_n(u)$ est une fonction de $\varepsilon_1, \dots, \varepsilon_{n-1}$, il vient :

$$\begin{aligned} \mathbb{E}[M_n(u) | \varepsilon_1, \dots, \varepsilon_{n-1}] &= N_n(u) e^{iu \frac{T_{n-1}}{\sqrt{n}}} \mathbb{E} \left[e^{iu\varepsilon_{n-1}\varepsilon_n/\sqrt{n}} | \varepsilon_{n-1} \right] \\ &= N_n(u) e^{iu \frac{T_{n-1}}{\sqrt{n}}} e^{-u^2\varepsilon_{n-1}^2\sigma^2/2n} \\ &= \exp \left(\frac{u^2\sigma^2}{2} \frac{V_{n-2}}{n} \right) \exp \left(iu \frac{T_{n-1}}{\sqrt{n}} \right). \end{aligned}$$

5. On a :

$$\mathbb{E}[M_n(u)] = \mathbb{E}[\mathbb{E}[M_n(u)|\varepsilon_1, \dots, \varepsilon_{n-1}]] = \mathbb{E} \left[\exp \left(\frac{u^2 \sigma^2}{2} \frac{V_{n-2}}{n} \right) \exp \left(iu \frac{T_{n-1}}{\sqrt{n}} \right) \right].$$

En itérant $n - 1$ fois, on obtient $\mathbb{E}[M_n(u)] = 1$.

6. En utilisant l'inégalité de Jensen deux fois, il vient :

$$\begin{aligned} \left| \psi_n(u) \mathbb{E}[N_n(u)] - \mathbb{E}[M_n(u)] \right| &\leq \mathbb{E} \left[\left| e^{iuT_n/n} (\mathbb{E}[N_n(u)] - N_n(u)) \right| \right] \\ &= \mathbb{E} \left[\left| \mathbb{E}[N_n(u)] - N_n(u) \right| \right] \\ &\leq \sqrt{\mathbb{E} \left[\left| \mathbb{E}[N_n(u)] - N_n(u) \right|^2 \right]} = \sqrt{\text{Var}(N_n(u))}. \end{aligned}$$

On déduit des questions I.5 et I.3 que :

$$\lim_{n \rightarrow +\infty} \left| \psi_n(u) \mathbb{E}[N_n(u)] - 1 \right| = 0.$$

Comme $\lim_{n \rightarrow +\infty} \mathbb{E}[N_n(u)] = e^{u^2 \sigma^4 / 2}$, on en déduit que $\lim_{n \rightarrow +\infty} \psi_n(u) = e^{-u^2 \sigma^4 / 2}$. Cette limite est valide pour tout $u \in \mathbb{R}$. Comme $e^{-u^2 \sigma^4 / 2}$ est la fonction caractéristique de $\sigma^2 G$, on en déduit que la suite $(T_n / \sqrt{n}, n \in \mathbb{N}^*)$ converge en loi vers $\sigma^2 G$.

II Estimation des paramètres

1. Il suffit de remarquer que, grâce à (XII.13), la densité de ε_k est donnée par :

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-z^2/2\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_k - \alpha x_{k-1} - \beta)^2 / 2\sigma^2}.$$

2. La log-vraisemblance est donnée par :

$$L_n(\varepsilon_1, \dots, \varepsilon_n; \alpha, \beta, \sigma) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^n (x_k - \alpha x_{k-1} - \beta)^2.$$

Il s'agit, à la constante $\frac{n}{2} \log(2\pi\sigma^2)$ près, d'une forme quadratique négative en α et β . Pour trouver les valeurs de (α, β) en lesquelles la log-vraisemblance est maximale, il suffit d'annuler ses dérivées en α et en β . On obtient alors le résultat.

3. La formule se démontre trivialement en sommant (XII.13) sur n de 1 à $n - 1$.

Pour démontrer que p.s. $\lim_{n \rightarrow +\infty} X_n/\sqrt{n} = 0$, on calcule par récurrence les moments successifs de $\mathbb{E}[X_n]$, $\mathbb{E}[X_n^2]$, $\mathbb{E}[X_n^3]$ et $\mathbb{E}[X_n^4]$ et l'on vérifie que pour $\alpha \in]-1, 1[$ ces quantités sont uniformément bornées en n . On en déduit que la variable aléatoire $\sum_{n \geq 1} X_n^4/n^2$ est intégrable et donc p.s. finie. Ceci assure que la suite $(X_n^4/n^2, n \geq 1)$ converge p.s. vers 0 et donc p.s. $\lim_{n \rightarrow +\infty} X_n/\sqrt{n} = 0$.

En utilisant la majoration précédente sur $\mathbb{E}[X_n^4]$, on vérifie que $\mathbb{E}[I_n^4]$ est de l'ordre de $1/n^2$. Ceci assure que la suite $\sum_{n \geq 1} I_n^4$ est intégrable et donc p.s. finie. On en déduit que la suite $(I_n, n \geq 1)$ converge p.s. vers 0.

1. La loi forte des grands nombres assure que p.s. $\lim_{n \rightarrow +\infty} \bar{\varepsilon}_n = 0$. On en déduit que p.s. $\lim_{n \rightarrow +\infty} \bar{X}_n(1 - \alpha) = \beta$.
2. On déduit de (XII.13) que :

$$X_{k+1}^2 = \beta^2 + \alpha^2 X_k^2 + \varepsilon_{k+1}^2 + 2\beta\alpha X_k + 2\beta\varepsilon_{k+1} + 2\alpha X_k \varepsilon_{k+1}.$$

En sommant cette relation pour $k \in \{0, \dots, n-1\}$, on obtient le résultat.

3. La loi forte des grands nombres assure que p.s. $\lim_{n \rightarrow +\infty} V_n/n = \mathbb{E}[\varepsilon_1^2] = \sigma^2$. On déduit de la question 4 que p.s. :

$$\lim_{n \rightarrow +\infty} \overline{X^2}_n(1 - \alpha^2) = \beta^2 + \sigma^2 + 2\beta\alpha \frac{\beta}{1 - \alpha}$$

et donc $b = \lim_{n \rightarrow +\infty} \overline{X^2}_n = \frac{\beta^2}{(1 - \alpha)^2} + \frac{\sigma^2}{1 - \alpha^2}$.

4. On déduit de (XII.13) que :

$$X_{k+1}X_k = \alpha X_k^2 + \beta X_k + X_k \varepsilon_{k+1}.$$

En sommant cette relation pour $k \in \{0, \dots, n-1\}$, on obtient :

$$\Delta_n = \alpha \overline{X^2}_{n-1} + \beta \bar{X}_{n-1} + I_n.$$

On déduit de ce qui précède que p.s. $\lim_{n \rightarrow +\infty} \Delta_n = \alpha b + \beta a$.

5. Comme $\beta a - a^2 = \alpha a^2$, on déduit que $(\hat{\alpha}_n, n \in \mathbb{N}^*)$ converge p.s. vers :

$$\frac{\alpha b + \beta a - a^2}{b - a^2} = \frac{\alpha b - \alpha a^2}{b - a^2} = \alpha.$$

6. On vérifie que l'estimateur du maximum de vraisemblance de σ^2 est :

$$\hat{\sigma}_n^2 = \frac{1}{n} \sum_{k=1}^n \hat{\varepsilon}_k^2 \quad \text{où} \quad \hat{\varepsilon}_k = X_k - \hat{\alpha}_n X_{k-1} - \hat{\beta}_n.$$

7. On a que $\hat{\sigma}_n^2$ est égal à :

$$\overline{X^2}_{n-1} + \frac{1}{n}(X_n^2 - X_0^2) + \hat{\alpha}_n^2 \overline{X^2}_{n-1} + \hat{\beta}_n^2 - 2\hat{\alpha}_n \Delta_n - 2\hat{\beta}_n(\bar{X}_{n-1} + \delta_n) + 2\hat{\alpha}_n \hat{\beta}_n \bar{X}_{n-1}.$$

On en déduit que p.s. :

$$\lim_{n \rightarrow +\infty} \hat{\sigma}_n^2 = b + \alpha^2 b + \beta^2 - 2\alpha(\alpha b + \beta a) - 2\beta a + 2\alpha\beta a = \sigma^2.$$

L'estimateur du maximum de vraisemblance de σ^2 est donc convergent.

III Test d'utilité du coefficient d'auto-régression

1. Un calcul élémentaire donne pour $\alpha = 0$:

$$\sqrt{n}(\Delta_n - (\bar{X}_{n-1} + \delta_n)\bar{X}_{n-1}) = \frac{T_n}{\sqrt{n}} + \frac{\beta\bar{\varepsilon}_n}{\sqrt{n}} - \sqrt{n}\frac{n-1}{n}\bar{\varepsilon}_n\bar{\varepsilon}_{n-1} + \frac{X_0}{\sqrt{n}}(\varepsilon_1 - \bar{\varepsilon}_n).$$

Comme $\sqrt{n}\bar{\varepsilon}_n$ converge en loi vers σG et que $\bar{\varepsilon}_{n-1}$ converge p.s. vers 0, on déduit du théorème de Slutsky que $\sqrt{n}\bar{\varepsilon}_n\bar{\varepsilon}_{n-1}$ converge en loi vers 0. On en déduit que $\sqrt{n}(\Delta_n - (\bar{X}_{n-1} + \delta_n)\bar{X}_{n-1}) = \frac{T_n}{\sqrt{n}} + R'_n$ où R'_n converge en loi vers 0. Par ailleurs, on a que $\overline{X^2}_{n-1} - \bar{X}_{n-1}^2$ est égal à :

$$\frac{V_{n-1}}{n} - \frac{(n-1)^2}{n^2}\bar{\varepsilon}_{n-1}^2 + \frac{X_0 - \beta}{n} \left[(X_0 - \beta)\left(1 - \frac{1}{n}\right) - 2\frac{n-1}{n}\bar{\varepsilon}_{n-1} \right].$$

On en déduit que $\overline{X^2}_{n-1} - \bar{X}_{n-1}^2 = \frac{V_{n-1}}{n} + R_n$ où R_n converge en loi vers 0.

2. La loi forte des grands nombres assure que la suite $(V_{n-1}/n, n \in \mathbb{N}^*)$ converge p.s. vers σ^2 . On déduit de la question I.8 et du théorème de Slutsky que sous H_0 la suite $(\zeta_n, n \in \mathbb{N}^*)$ converge en loi vers $\sigma^2 G/\sigma^2 = G$.
3. Sous H_1 , $\hat{\alpha}_n$ converge p.s. vers $\alpha \neq 0$. On en déduit que p.s. $\lim_{n \rightarrow +\infty} \zeta_n \in \{-\infty, +\infty\}$. La région critique du test est donc de la forme $W_n = \{|\zeta_n| \geq a\}$. Ce test est de niveau asymptotique η pour $a = \phi_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi gaussienne $\mathcal{N}(0, 1)$. Le test est convergent.
4. La p -valeur asymptotique de ce test est $p\text{-val} = \mathbb{P}(|G| \geq \zeta_n^{\text{obs}})$.
5. On a :

$$\hat{\alpha}_n \simeq 0.9317722, \quad \hat{\beta}_n \simeq 0.7845925, \quad \hat{\sigma}_n^2 \simeq 3.7443066, \quad \zeta_n^{\text{obs}} \simeq 17.801483.$$

La p -valeur est très faible ($< 10^{-23}$). On rejette donc très clairement H_0 .

▲

Exercice XII.13. I Préliminaires

1. On a :

$$\mathbb{E}[T_n^*] = \sum_{k=2}^n \mathbb{E}[T_k] = \sum_{k=2}^n \frac{2}{k(k-1)} = 2 \left(1 - \frac{1}{n}\right) \xrightarrow{n \rightarrow \infty} 2.$$

La date moyenne de naissance du dernier ancêtre est d'environ $2 \times 20 \times 10\,000 = 400\,000$ ans.

2. On a $\mathbb{E}[L_n] = \sum_{k=2}^n k \mathbb{E}[T_k] = 2h_{n-1}$ et, par indépendance :

$$\text{Var}(L_n) = \sum_{k=2}^n k^2 \text{Var}(T_k) = 4 \sum_{k=1}^{n-1} \frac{1}{k^2}.$$

Soit $\varepsilon > 0$. On déduit de l'inégalité de Tchebychev que :

$$\mathbb{P} \left(\left| \frac{L_n}{h_{n-1}} - 2 \right| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2 h_{n-1}^2} \mathbb{E}[(L_n - 2h_{n-1})^2] = \frac{1}{\varepsilon^2 h_{n-1}^2} \text{Var}(L_n) \xrightarrow{n \rightarrow \infty} 0.$$

On en déduit que la suite $(L_n/h_{n-1}, n \geq 2)$ converge en probabilité vers 2.

3. On a, en conditionnant par rapport à T_k :

$$\mathbb{E}_\theta[Y_k] = \mathbb{E}_\theta \left[\sum_{j=1}^k \mathbb{E}_\theta[Y_{j,k} | T_k] \right] = \frac{\theta}{2} \mathbb{E}_\theta[k T_k] = \frac{\theta}{k-1}.$$

4. On a, en conditionnant par rapport à T_k :

$$\begin{aligned} \psi_{Y_{j,k}, T_k}(u, v) &= \mathbb{E}_\theta \left[e^{iuY_{j,k} + ivT_k} \right] \\ &= \mathbb{E} \left[\mathbb{E}_\theta \left[e^{iuY_{j,k} + ivT_k} \mid T_k \right] \right] \\ &= \mathbb{E} \left[e^{ivT_k} \mathbb{E}_\theta \left[e^{iuY_{j,k}} \mid T_k \right] \right] \\ &= \mathbb{E} \left[e^{ivT_k - \frac{\theta}{2}(1-e^{iu})T_k} \right] \\ &= \frac{k(k-1)/2}{\frac{k(k-1)}{2} - iv + \frac{\theta}{2}(1-e^{iu})} = \frac{k(k-1)}{k(k-1) - 2iv + \theta(1-e^{iu})}. \end{aligned}$$

5. En utilisant le fait que, conditionnellement à T_k , les variables aléatoires $Y_{1,k}, \dots, Y_{k,k}$ sont indépendantes et de même loi, il vient :

$$\begin{aligned} \psi_{Y_k, T_k}(u, v) &= \mathbb{E}_\theta \left[e^{iu \sum_{j=1}^k Y_{j,k} + ivT_k} \right] = \mathbb{E} \left[e^{ivT_k} \mathbb{E}_\theta \left[e^{iu \sum_{j=1}^k Y_{j,k}} \mid T_k \right] \right] \\ &= \mathbb{E} \left[e^{ivT_k} \mathbb{E}_\theta \left[e^{iuY_{1,k}} \mid T_k \right]^k \right] \\ &= \mathbb{E} \left[e^{ivT_k - k \frac{\theta}{2}(1-e^{iu})T_k} \right] \\ &= \frac{k-1}{(k-1) - 2i \frac{v}{k} + \theta(1-e^{iu})}. \end{aligned}$$

6. On a :

$$\psi_{Y_k+1}(u) = e^{iu} \psi_{Y_k}(u) = e^{iu} \psi_{Y_k, T_k}(u, 0) = \frac{p_k e^{iu}}{1 - (1 - p_k) e^{iu}},$$

avec

$$p_k = \frac{k - 1}{\theta + (k - 1)}.$$

On reconnaît la fonction caractéristique de la loi géométrique de paramètre p_k , ce qui implique que $Y_k + 1$ est de loi géométrique de paramètre p_k .

7. Par linéarité, on a $\mathbb{E}_\theta[S_n] = \sum_{k=2}^n \mathbb{E}_\theta[Y_k] = \theta h_{n-1}$. Par indépendance, et en utilisant le fait que $\text{Var}(Y_k) = \text{Var}(Y_k + 1) = (1 - p_k)/p_k^2$, il vient :

$$\text{Var}_\theta(S_n) = \sum_{k=2}^n \text{Var}_\theta(Y_k) = \sum_{k=2}^n \frac{\theta}{\theta + (k - 1)} \frac{(\theta + (k - 1))^2}{(k - 1)^2} = \theta h_{n-1} + \theta^2 \sum_{k=1}^{n-1} \frac{1}{k^2}.$$

II Propriétés de l'estimateur de Watterson

1. L'estimateur de Watterson de θ est sans biais car $\mathbb{E}_\theta[\hat{\theta}_n] = \theta$.

2. Soit $\varepsilon > 0$. On déduit de l'inégalité de Tchebychev que :

$$\mathbb{P}_\theta \left(\left| \hat{\theta}_n - \theta \right| \geq \varepsilon \right) \leq \frac{1}{\varepsilon^2} \text{Var}_\theta(\hat{\theta}_n) = \frac{1}{\varepsilon^2 h_{n-1}^2} \text{Var}_\theta(S_n) = O(1/\log(n)) \xrightarrow{n \rightarrow \infty} 0.$$

On en déduit que la suite $(\hat{\theta}_n, n \geq 2)$ converge en probabilité vers θ .

3. On rappelle que $\mathbb{E}_\theta [e^{iuY_k}] = \left(1 + \frac{\theta}{k-1}(1 - e^{iu})\right)^{-1}$. On en déduit que :

$$\begin{aligned} & \mathbb{E}_\theta \left[\exp \left(i \frac{u}{\sqrt{\theta h_{n-1}}} \left(Y_k - \frac{\theta}{k-1} \right) \right) \right] \\ &= \left(e^{iu\sqrt{\theta}/(k-1)\sqrt{h_{n-1}}} \left(1 + \frac{\theta}{k-1}(1 - e^{iu/\sqrt{\theta h_{n-1}}}) \right) \right)^{-1} \\ &= \left(1 + \frac{iu\sqrt{\theta}}{(k-1)\sqrt{h_{n-1}}} - \frac{u^2\theta}{2(k-1)^2 h_{n-1}} + O\left((k-1)^{-3} h_{n-1}^{-3/2} \right) \right)^{-1} \\ & \quad \left(1 - \frac{iu\sqrt{\theta}}{(k-1)\sqrt{h_{n-1}}} + \frac{u^2}{2(k-1)h_{n-1}} + \frac{1}{k-1} O\left(h_{n-1}^{-3/2} \right) \right)^{-1} \\ &= \left(1 + \frac{u^2}{2(k-1)h_{n-1}} + \frac{u^2\theta}{2(k-1)^2 h_{n-1}} + O\left((k-1)^{-1} h_{n-1}^{-3/2} \right) \right)^{-1} \\ &= \left(1 + \frac{u^2}{2(k-1)h_{n-1}} + O\left((k-1)^{-1} h_{n-1}^{-3/2} + (k-1)^{-2} h_{n-1}^{-1} \right) \right)^{-1}. \end{aligned}$$

Pour conclure, il suffit de remarquer que :

$$\begin{aligned} e^{-u^2/(2(k-1)h_{n-1})} &= \left(1 + \frac{u^2}{2(k-1)h_{n-1}} + O\left((k-1)^{-2}h_{n-1}^{-2}\right) \right)^{-1} \\ &= \left(1 + \frac{u^2}{2(k-1)h_{n-1}} + O\left((k-1)^{-1}h_{n-1}^{-3/2} + (k-1)^{-2}h_{n-1}^{-1}\right) \right)^{-1}. \end{aligned}$$

4. Par indépendance, on déduit de ce qui précède que :

$$\begin{aligned} \left| \mathbb{E}_\theta \left[e^{iuZ_n(\theta)} \right] - e^{-u^2/2} \right| &= \left| \prod_{k=2}^n \mathbb{E}_\theta \left[e^{i \frac{u}{\sqrt{\theta h_{n-1}}} \left(Y_k - \frac{\theta}{k-1} \right)} \right] - \prod_{k=2}^n e^{-\frac{u^2}{2(k-1)h_{n-1}}} \right| \\ &\leq \sum_{k=2}^n \left| \mathbb{E}_\theta \left[e^{i \frac{u}{\sqrt{\theta h_{n-1}}} \left(Y_k - \frac{\theta}{k-1} \right)} \right] - e^{-\frac{u^2}{2(k-1)h_{n-1}}} \right| \\ &= \sum_{k=2}^n O\left((k-1)^{-1}h_{n-1}^{-3/2} + (k-1)^{-2}h_{n-1}^{-1} \right) \\ &= O\left(h_{n-1}^{-1/2} \right) = O(\log(n)^{-1/2}). \end{aligned}$$

5. On a :

$$\lim_{n \rightarrow +\infty} \mathbb{E}_\theta \left[e^{iuZ_n(\theta)} \right] = e^{-u^2/2},$$

autrement dit la suite $(Z_n(\theta), n \geq 2)$ converge en loi vers G . La suite $(S_n/(\theta h_{n-1}), n \geq 2)$ converge en probabilité vers 1. On déduit donc du théorème de Slutsky que $((S_n - \theta h_{n-1})/\sqrt{S_n}, n \geq 2)$ converge en loi vers G .

6. Soit $\phi_{1-\alpha/2}$ le quantile d'ordre $1 - \alpha/2$ de la loi $\mathcal{N}(0, 1)$. On a :

$$\frac{S_n - \theta h_{n-1}}{\sqrt{S_n}} \in [\pm \phi_{1-\alpha/2}] \iff \theta \in \left[\frac{S_n}{h_{n-1}} \pm \frac{\sqrt{S_n}}{h_{n-1}} \phi_{1-\alpha/2} \right].$$

Il vient :

$$\begin{aligned} \mathbb{P}_\theta \left(\theta \in \left[\frac{S_n}{h_{n-1}} \pm \frac{\sqrt{S_n}}{h_{n-1}} \phi_{1-\alpha/2} \right] \right) &= \mathbb{P}_\theta \left(\frac{S_n - \theta h_{n-1}}{\sqrt{S_n}} \in [\pm \phi_{1-\alpha/2}] \right) \\ &\xrightarrow[n \rightarrow \infty]{} \mathbb{P}(G \in [\pm \phi_{1-\alpha/2}]) = 1 - \alpha. \end{aligned}$$

On en déduit que :

$$I_n = \left[\frac{S_n}{h_{n-1}} \pm \frac{\sqrt{S_n}}{h_{n-1}} \phi_{1-\alpha/2} \right] = \left[\hat{\theta}_n \pm \sqrt{\frac{\hat{\theta}_n}{h_{n-1}}} \phi_{1-\alpha/2} \right]$$

est un intervalle de confiance pour θ de niveau asymptotique $1 - \alpha$.

7. L'application numérique donne : $\phi_{1-\alpha/2} \simeq 1.96$, $\hat{\theta}_n = 5.51$ et $I_n = [3.4, 7.6]$. Le taux de mutation par nucléotide est $\hat{\theta}_n/K = 0.0153$ et l'intervalle de confiance de niveau asymptotique $1 - \alpha$ correspondant est $I_n/K = [0.0094, 0.0212]$.

III Comparaison d'estimateurs

1. On a par indépendance :

$$\begin{aligned}\mathbb{P}_\theta(Y_{1,k} = y_{1,k}, \dots, Y_{k,k} = y_{k,k} | T_k = t_k) &= \prod_{j=1}^k \mathbb{P}_\theta(Y_{j,k} = y_{j,k} | T_k = t_k) \\ &= \prod_{j=1}^k e^{\theta t_k/2} \frac{(\theta t_k/2)^{y_{j,k}}}{y_{j,k}!}.\end{aligned}$$

On en déduit que la vraisemblance de $(T_k, Y_{1,k}, \dots, Y_{k,k})$ est :

$$p(t_k, y_{1,k}, \dots, y_{k,k}) = \frac{k(k-1)}{2} e^{-k(k-1)t_k/2} \prod_{j=1}^k e^{\theta t_k/2} \frac{(\theta t_k/2)^{y_{j,k}}}{y_{j,k}!}.$$

On utilise l'indépendance des variables aléatoires $(T_k, Y_{1,k}, \dots, Y_{k,k})$, $k \in \{2, \dots, n\}$ pour conclure.

2. La vraisemblance est de la forme $h(x)\Psi(\sum_{k=2}^n kt_k, \sum_{k=2}^n \sum_{i=1}^k y_{i,k}; \theta)$. On a $L_n = \sum_{k=2}^n kT_k$ et $S_n = \sum_{k=2}^n \sum_{i=1}^k Y_{i,k}$. Le théorème de factorisation assure donc que (S_n, L_n) est une statistique exhaustive. (Cette statistique est en fait totale.)

3. On a :

$$\partial_\theta \mathcal{L}_n(x; \theta) = 0 \iff -\frac{1}{2} \sum_{k=2}^n kt_k + \frac{1}{\theta} \sum_{k=2}^n \sum_{i=1}^k y_{i,k} = 0 \iff \theta = \frac{2 \sum_{k=2}^n \sum_{i=1}^k y_{i,k}}{\sum_{k=2}^n kt_k}.$$

On a également $\partial_\theta^2 \mathcal{L}_n(x; \theta) = -\frac{1}{\theta^2} \sum_{k=2}^n \sum_{i=1}^k y_{i,k} \leq 0$, ce qui assure que la fonction \mathcal{L}_n est concave en θ . Elle possède donc un seul maximum dans $]0, +\infty[$ si $\sum_{k=2}^n \sum_{i=1}^k y_{i,k} > 0$. On en déduit que l'estimateur du maximum de vraisemblance est :

$$\tilde{\theta}_n = 2 \frac{\sum_{k=2}^n \sum_{i=1}^k Y_{i,k}}{\sum_{k=2}^n kT_k} = 2 \frac{S_n}{L_n}.$$

La convergence de l'estimateur se déduit des questions I.2 et II.2 .

4. On a :

$$I_n(\theta) = -\mathbb{E}_\theta[\partial_\theta^2 \mathcal{L}_n(X; \theta)] = \frac{1}{\theta^2} \mathbb{E}_\theta[S_n] = \frac{h_{n-1}}{\theta}.$$

L'estimateur de Watterson est sans biais de variance :

$$\text{Var}_\theta(\hat{\theta}_n) = \frac{\theta}{h_{n-1}} + \frac{\theta^2}{h_{n-1}^2} \sum_{k=1}^{n-1} \frac{1}{k^2} > \frac{1}{I_n(\theta)}.$$

On en déduit que l'estimateur de Watterson n'est pas efficace. En revanche il est fonction de la statistique exhaustive (S_n, L_n) (qui est en fait minimale car totale). Il ne peut donc pas être amélioré; il est donc optimal.

5. La question II.4 assure que $\left(\sqrt{h_{n-1}/\theta}(\hat{\theta}_n - \theta), n \geq 2\right)$ converge en loi vers G ; autrement dit, $\left(\sqrt{h_{n-1}}(\hat{\theta}_n - \theta), n \geq 2\right)$ converge en loi vers $\sqrt{\theta} G$. On en déduit que l'estimateur de Watterson est asymptotiquement normal de vitesse de convergence $\sqrt{h_{n-1}}$ et de variance asymptotique $\Sigma(\theta) = \theta$. Comme $s_n^2/I_n(\theta) = \Sigma(\theta)$, on obtient que l'estimateur de Watterson est asymptotiquement efficace.
6. L'estimateur du maximum de vraisemblance est asymptotiquement efficace ainsi que l'estimateur de Watterson. Ils ont donc asymptotiquement la même précision.

IV Comparaison du taux de mutations

1. L'hypothèse alternative correspond, d'après le texte, à un taux plus élevé de mutation soit $H_1 = \{\theta > \theta_0\}$.
2. D'après la question II.4, sous H_0 , la statistique de test ζ_n converge en loi vers G . On déduit de la question II.2 que, sous H_1 , la suite $(S_n/h_{n-1} - \theta_0, n \geq 2)$ converge en probabilité vers $\theta - \theta_0 > 0$. On en déduit que sous H_1 la statistique de test ζ_n converge en probabilité vers $+\infty$.
3. La région critique est donc de la forme $W_n = \{\zeta_n \geq c\}$. Comme :

$$\mathbb{P}_{\theta_0}(W_n) = \mathbb{P}_{\theta_0}(\zeta_n \geq c) \xrightarrow[n \rightarrow \infty]{} \mathbb{P}(G \geq c),$$

on en déduit que la région critique est de niveau asymptotique α pour $c = \phi_{1-\alpha}$.

4. La p -valeur asymptotique est définie par :

$$p\text{-val} = \mathbb{P}(G \geq \zeta_n^{\text{obs}}).$$

L'application numérique donne $\zeta_n^{\text{obs}} \simeq 62$ et $p\text{-val} \leq \mathbb{P}(G \geq 10) \leq 10^{-23}$. On rejette donc H_0 . Le taux de mutation de l'ADNmt est donc supérieur à celui de l'ADN chromosomique.



Index

- ADN, 148
- anniversaire, 3, 4, 71

- Banach (allumettes de), 12
- Bertrand (paradoxe de), 7
- Black-Sholes, 91
- bombes sur Londres, 37
- Bonferroni (inégalités de), 4
- Box-Muller (transformation de), 46
- Buffon (aiguille de), 24

- cavalerie, 65
- censure, 49, 106
- circuit (série/parallèle), 22
- clefs, 10
- code, 96
- collectionneur, 71, 72, 74
- compression, 16
- contamination au mercure, 40
- convergence du maximum, 35, 40
- Cox-Ross-Rubinstein, 91

- dé, 10, 11
- détection de contamination, 54
- dilution, 54
- données censurées, 107
- duplication pour la fonction Γ , 26

- entropie, 16
- estimateur
 - Bailey, 104, 105
 - Petersen, 104
 - Watterson, 145

- famille
 - 2 enfants, 6
 - 5 enfants, 65
- finance, 91
- fonction
 - génératrice, 9–11
 - pivotale, 68
 - répartition, 70
- formule du crible, 5

- généalogie, 144
- génétique, 7
- gâteau avec
 - cerise, 20
 - fève, 19
- Galton-Watson, 83
- geyser, 62
- GPS, 24
- grandes déviations, 33
- grandes déviations (majoration), 34

- jeu de cartes, 3, 7
- jeu de dé, 3
- jeu de la porte d'or, 8

- Kingman, 144
- Kolmogorov-Smirnov (test de), 339

- Laplace, 33
- Legendre, 34
- loi
 - χ^2 , 21
 - Bernoulli, 12, 14, 34
 - béta, 49, 134, 182, 184
 - biaisée par la taille, 77
 - binomiale, 33, 37, 39, 236

- binomiale négative, 12, 13
- Bose-Einstein, 86
- Cauchy, 21, 27, 32
- conditionnelle, 21
- défaut de forme, 29
- exponentielle, 20, 27, 31, 45, 48, 49, 56, 179
- exponentielle symétrique, 20, 27, 199
- Fréchet, 40
- géométrique, 13–15, 18, 31, 45, 51
- gamma, 20, 26, 27
- gaussienne, 21, 24, 28, 29, 43, 52, 56, 57, 60, 61, 196
- Gumbel, 277
- hypergéométrique, 33, 50
- invariante par rotation, 29
- log-normale, 208
- Maxwell, 26
- Pareto, 119
- Poisson, 6, 9, 13–15, 23, 35, 37, 45, 48, 59, 65, 206
- Rayleigh, 19, 24, 51
- sans mémoire, 15
- symétrique, 28
- uniforme, 20, 21, 23, 31, 33, 35, 70
- Weibull, 106, 124
- Yule, 88, 90, 131

- méthode MPN, 54
- Mann et Whitney, 81
- mariage, 100
- méthode du rejet, 46
- modèle auto-regressif, 143
- modèle non-paramétrique, 112
- de Montmort (problème de), 6
- moyenne, 47
- mutation, 144, 148

- nombres normaux, 39
- nombre inconnu, 25

- optimisation de coûts, 15

- paradoxe
 - Bertrand (de), 23
 - bus (du), 76, 78
 - Saint-Petersbourg (de), 36
- paramètre de décalage, 112, 137
- permutation, 100
- Poincaré, 70
- population, 83

- rapport de vraisemblance, 55
- renouvellement, 18

- sélection de juré, 58
- secrétaire, 100
- sensibilité, 7
- Shannon, 96
- somme aléatoire, 28
- sondage, 36, 37, 59, 88, 128
- spécificité, 7
- stabilisation de la variance, 68
- statistique
 - bayésienne, 134
 - d'ordre, 72, 123
- Stirling, 5
- stratégie optimale, 100
- stratification, 65, 129
- suites typiques, 16
- suite consécutive, 4

- taux de transmission, 94
- température, 143
- test
 - χ^2 , 61, 62, 64–66, 261
 - d'égalité des marginales, 64
 - générateur, 61
 - Kolmogorov-Smirnov, 339
 - Mann et Whitney, 113
 - Mc Nemar, 64
 - Neyman, 55
 - non-paramétrique, 137
 - signes, 137
 - UPPS, 58, 61
 - Wilcoxon, 137
- test médical, 7
- théorème
 - Cochran, 43
 - Weierstrass, 39
- théorie de l'information, 94
- tirage
 - avec remise, 4, 11, 88, 162
 - sans remise, 4, 33, 50, 88, 163, 299
- transformée
 - Laplace, 33, 95
 - Legendre, 34, 95
- transformation de Box-Muller, 46
- triangle, 20

- urne, 4, 11

- vaccin, 59, 64

variance inter-strate, 361

variance intra-strate, 361

vecteur gaussien, 70

vitesse d'un gaz, 26

Weierstrass (théorème de), 39

Weldon (dés de), 65