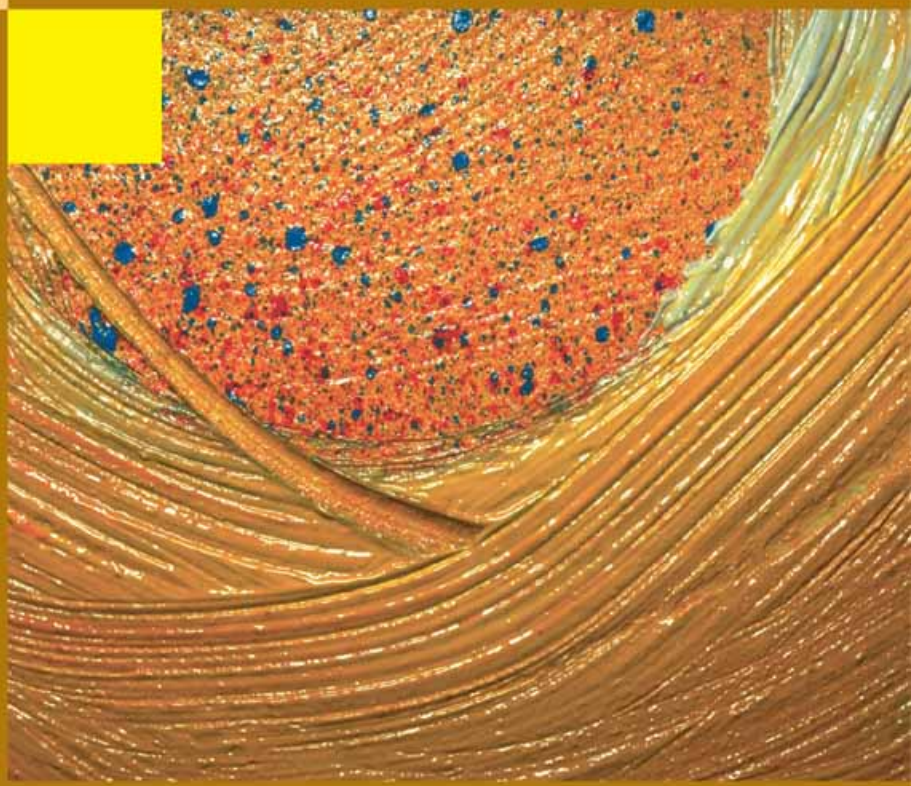
Collection
Statistique
et probabilités
appliquées

Eva Cantoni, Philippe Huber,
Elvezio Ronchetti

Maîtriser l'aléatoire

Exercices résolus de probabilités et statistique



$$\hat{\sigma}_{\bar{X}_1-\bar{X}_2}$$

$$S_1^2 = \frac{\sum_{i=1}^{n_1}(X_i-\bar{X}_1)^2}{n_1-1}$$

$$S_1 = \sqrt{S_1^2}$$

$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1}$$

$$S_1 = \sqrt{S_1^2} \quad \text{et} \quad S_2 = \sqrt{S$$

$$\bar{X}_1 = \frac{\sum_{i=1}^{n_1} X_i}{n_1} \quad \text{et} \quad \bar{X}_2 = \frac{\sum_{j=1}^{n_2}$$

Springer

Maîtriser l'aléatoire

Springer
Paris
Berlin
Heidelberg
New York
Hong Kong
Londres
Milan
Tokyo

Eva Cantoni
Philippe Huber
Elvezio Ronchetti

Maîtriser l'aléatoire

Exercices résolus de probabilités et statistique

Springer

Eva Cantoni
Département d'économétrie
Université de Genève
40, boulevard du Pont d'Arve
1211 Genève 4
Suisse

Philippe Huber
Département d'économétrie
Université de Genève
40, boulevard du Pont d'Arve
1211 Genève 4
Suisse

Elvezio Ronchetti
Département d'économétrie
Université de Genève
40, boulevard du Pont d'Arve
1211 Genève 4
Suisse

ISBN-10 : 2-287-34069-6 Springer Paris Berlin Heidelberg New York

ISBN-13 : 978-2287-34069-7 Springer Paris Berlin Heidelberg New York

SPIN : 11752523

Maquette de couverture : Jean-François Montmarché

Collection
Statistiques et probabilités appliquées
dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
2002 Neuchâtel - Suisse

Dans la même collection :

— *Statistique. La théorie et ses applications,*
Michel Lejeune, avril 2004
— *Le choix Bayésien. Principes et pratique,*
Christian P. Robert, novembre 2005

Préface

Cet ouvrage est le résultat de l'expérience pédagogique des dix dernières années à l'Université de Genève dans le cadre de deux cours de base semestriels de probabilités et statistique au niveau *bachelor*. Dans ces cours sont présentés les concepts de base des probabilités ainsi que de l'inférence statistique. Ces domaines ne sont pas des « sports pour des spectateurs » mais exigent un apprentissage actif des concepts présentés en classe. D'où la nécessité d'une grande quantité d'exercices qui permettent d'assimiler et maîtriser ces concepts.

Dans cet ouvrage, on a recueilli 212 problèmes résolus, structuré en huit chapitres pour des raisons de clarté. À l'intérieur de chaque chapitre, les exercices sont séparés par thème, avec, pour chaque thème, un niveau de difficulté qui va en augmentant. À la fin de chaque chapitre, sont indiquées quelques références qui couvrent les aspects théoriques (non traités ici).

Cet ouvrage peut être utilisé comme complément de tout livre de probabilités et statistique dans le cadre de cours de base dans des domaines aussi variés que les sciences économiques, la psychologie, les sciences sociales, les mathématiques, les sciences naturelles et la médecine. Il peut aussi être utilisé dans le cadre de cours de préparation à l'entrée dans un programme de *master* dans un domaine où ces connaissances sont exigées, ainsi que pour l'autoformation et la préparation d'examens.

Plusieurs personnes nous ont aidés et inspirés au cours de ce projet. Nous tenons à remercier les assistants qui ont collaboré au cours, F.X. de Rossi, V. Czellàr, D. Conne, S. Lô. Le choix de certains exercices a été largement influencé par les travaux pratiques de la Chaire de Statistique Appliquée de l'École Polytechnique Fédérale de Lausanne (EPFL) et aussi par l'excellent livre de S.M. Ross (1987) ; que toutes ces personnes trouvent ici l'expression de notre gratitude.

Pour terminer, nous tenons également à remercier Mme N. Huilleret, éditrice aux éditions Springer-Verlag France, et notre collègue le Professeur Y. Dodge, responsable de la Collection statistique et probabilités appliquées, pour leurs précieux conseils et leurs encouragements.

Genève, juin 2006

Eva Cantoni
Philippe Huber
Elvezio Ronchetti

Web: `http://www.unige.ch/ses/metri/cantoni/Maitriser/index.html`

Sommaire

Chapitre 1

Probabilités élémentaires

Introduction

Les exercices de ce chapitre concernent les règles de base du calcul des probabilités. Dans beaucoup de problèmes élémentaires on calcule la probabilité d'un événement comme {nombre de cas favorables à l'événement}/{nombre de cas possibles}. Cela implique la connaissance de quelques formules de base de l'analyse combinatoire. Un autre outil utile est la construction d'une structure à arbre qui représente graphiquement toutes les séquences possibles d'une expérience répétée. Dans ce contexte, la notion de probabilité conditionnelle permet de calculer des probabilités complexes à partir de situations plus simples. Enfin, le théorème de Bayes est un résultat fondamental qui permet d'« inverser » une probabilité conditionnelle (voir en particulier l'exercice 1.19 et la suite l'exercice 2.1).

Notes historiques
La notion d'aléatoire et le concept intuitif de probabilité remontent à l'antiquité mais c'est au XVI^{e} et au XVII^{e} siècle que des règles élémentaires de calcul sont développées. La fameuse correspondance entre les mathématiciens français Pascal et Fermat en 1654, concernant un problème du jeu au hasard proposé par un noble de l'époque, le Chevalier de Mérée (voir l'exercice 1.6), est considérée comme le point de départ du « calcul » des probabilités. Parmi les grands savants qui ont travaillé par la suite sur des problèmes de probabilités on peut mentionner Jacob Bernoulli avec son oeuvre *Ars Conjectandi* (1713) ainsi que d'autres membres de cette famille unique de mathématiciens suisses, de Moivre avec son oeuvre *Doctrine des chances* (1718) et ensuite Laplace, Euler, Gauss, Lagrange et Legendre. Jusqu'au début du XX^{e} siècle le domaine des probabilités resta un champ des mathématiques constitué d'un ensemble de résultats (intéressants et utiles) mais sans aucune base axiomatique. En 1900 Hilbert énonça son fameux programme qui contenait comme sixième problème

le développement d'une structure axiomatique pour les probabilités. En 1933 le mathématicien russe A.N. Kolmogorov releva le défi en publiant un article qui présentait les fameux axiomes à la base du calcul des probabilités. Les probabilités devenaient alors un domaine des mathématiques à part entière comme la géométrie, l'algèbre ou encore l'analyse.

Références (théorie)

Il existe beaucoup de livres sur les probabilités. Deux bonnes références sont Ross, chapitres 1 à 3 [1] et Pitman, chapitre 1 [2].

Exercices

Calculs simples de probabilités

1.1

Quatre hommes déposent leur chapeau au vestiaire en entrant dans un restaurant et choisissent au hasard en sortant 1 des 4 chapeaux. Calculer les probabilités suivantes.

1. Aucun des 4 hommes ne prend son propre chapeau.
2. Exactement 2 des 4 hommes prennent leur propre chapeau.

1.2

Aurélie et Nicolas jouent aux dés. Ils lancent tour à tour 2 dés et observent les chiffres sortis. Quand la somme est 7 ou le produit 6, Aurélie marque un point ; quand la somme est 6 ou le produit 4, Nicolas en marque 1. Pour qui parieriez-vous ?

1.3

Parmi les familles de 2 enfants, la moitié se trouve être bien répartie, c'est-à-dire composée d'autant de garçons que de filles. En est-il de même parmi les familles de 4 enfants ? (On suppose ici que chaque naissance donne avec équiprobabilité un garçon ou une fille.)

1.4

On tire au hasard 2 cartes d'un jeu de cartes de poker (52 cartes). Quelle est la probabilité qu'elles forment un *black jack*, ou autrement dit, que l'une soit un as et l'autre un dix, un valet, une dame ou un roi ?

1.5

On classe 5 hommes et 5 femmes selon leur résultats lors d'un examen. On fait l'hypothèse que tous les scores sont différents et que les 10! classements possibles ont tous la même probabilité de se réaliser. On désigne le rang de la meilleure femme par X (par exemple X vaudra 2 si le meilleur résultat a été obtenu par un homme et le suivant par une femme). Donner la fonction de fréquences de X, c'est-à-dire $P(X = i)$ pour $i = 1, \ldots, 10$.

1.6

Problème posé par le Chevalier de Mérée à Pascal en 1654.
Quel est l'événement le plus probable : obtenir au moins 1 fois 1 as en lançant 4 fois un dé ou obtenir au moins 1 fois 1 double as en lançant 24 fois 2 dés ?

1.7

On considère 3 événements A, B, et C.

1. À l'aide d'un dessin des ensembles A, B et C, trouver une formule permettant de calculer $P(A \cup B \cup C)$ si l'on connaît les probabilités de chacun de ces événements et les probabilités des intersections de ces événements.
2. Démontrer cette formule à partir des axiomes de la théorie des probabilités.

1.8

On considère une famille avec 2 enfants. On suppose que la venue d'une fille est aussi certaine que celle d'un garçon.

1. Quelle est la probabilité que les 2 enfants soient des garçons sachant que l'aîné est un garçon ?
2. Quelle est la probabilité que les 2 enfants soient des garçons sachant qu'au moins un des enfants est un garçon ?

1.9

On jette 2 dés équilibrés.

1. Quelle est la probabilité qu'au moins l'un d'entre eux montre 6, sachant que les 2 résultats sont différents ?
2. Quelle est la probabilité qu'au moins l'un d'entre eux montre 6, sachant que leur somme vaut i ? Calculer le résultat pour toutes les valeurs possibles de i.

Probabilités totales et théorème de Bayes

1.10

Un certain système a 5 composantes. Une panne du système est causée 35 %, 30 %, 20 %, 10 % et 5 % des fois par une panne dans les composantes

A, B, C, D et E, respectivement. On suppose que les pannes simultanées dans plus d'une composante à la fois sont si rares qu'on peut les négliger.

1. Si une panne du système n'est pas causée par A, quelle est la probabilité qu'elle soit causée par B?
2. Si une panne du système n'est causée ni par A, ni par B, quelle est la probabilité qu'elle soit causée par C ou D?

1.11

On compte respectivement 50, 75, et 100 employés dans 3 entrepôts A, B et C, les proportions des femmes étant respectivement égales à 50 %, 60 % et 70 %. Une démission a autant de chance de se produire chez tous les employés, indépendamment de leur sexe. Une employée donne sa démission. Quelle est la probabilité qu'elle vienne de l'entrepôt C?

1.12

Tous les meilleurs joueurs du monde sont inscrits au tournoi de tennis de Diamond City pour lequel le 1er prix est une rivière en diamants. On estime *a priori* que Roger Federer a 4 chances sur 10 de gagner, Andy Roddick 3 chances sur 10 et Leyton Hewitt 2 sur 10. Si par hasard Roger Federer se blesse et annule sa participation au dernier moment, que deviennent les chances respectives de Andy Roddick et Leyton Hewitt de remporter la rivière de diamants?

1.13

Dans un pays où il naît autant de filles que de garçons, le docteur Gluck prévoit le sexe des enfants à naître. Il se trompe 1 fois sur 10 si c'est un garçon et 1 fois sur 20 si c'est une fille. Aujourd'hui il vient de dire à Mme Parisod qu'elle aurait une fille. Quelle est la probabilité pour que cela soit vrai?

1.14

Une compagnie d'assurance répartit les assurés en 3 classes : personnes à bas risque, risque moyen et haut risque. Ses statistiques indiquent que la probabilité qu'une personne soit impliquée dans un accident sur une période d'un an est respectivement de 0,05, 0,15 et 0,30. On estime que 20 % de la population est à bas risque, 50 % à risque moyen et 30 % à haut risque.

1. Quelle est la proportion d'assurés qui ont eu un accident ou plus au cours d'une année donnée?

2. Si un certain assuré n'a pas eu d'accidents l'année passée, quelle est la probabilité qu'il fasse partie de la classe à bas risque ?

1.15

Un avion est porté disparu. On pense que l'accident a pu arriver aussi bien dans n'importe laquelle de 3 régions données. Notons par $1 - \alpha_i$ la probabilité qu'on découvre l'avion dans la région i s'il y est effectivement. Les valeurs α_i représentent donc la probabilité de manquer l'avion lors des recherches. On peut l'attribuer à diverses causes d'ordre géographique ou à la végétation propre à la région.

Quelle est la probabilité que l'avion se trouve dans la i^{e} région ($i = 1, 2, 3$) si les recherches dans la région 1 n'ont rien donné ?

1.16

À Londres il pleut en moyenne 1 jour sur 2 et donc la météo prévoit de la pluie la moitié des jours. Les prévisions sont correctes 2 fois sur 3, c'est-à-dire les probabilités qu'il pleuve quand on a prévu de la pluie et qu'il ne pleuve pas quand on a prévu du temps sec sont égales à 2/3. Quand la météo prévoit de la pluie, Mr. Pickwick prend toujours son parapluie. Quand la météo prévoit du temps sec il le prend avec probabilité 1/3. Calculer :

1. la probabilité que Mr. Pickwick prenne son parapluie un jour quelconque ;
2. la probabilité qu'il n'ait pas pris son parapluie un jour pluvieux ;
3. la probabilité qu'il ne pleuve pas sachant qu'il porte son parapluie.

1.17

Le sultan dit à Ali Baba : « Voici 2 urnes, 4 boules blanches (b) et 4 boules noires (n). Répartis les boules dans les urnes, mais je rendrai ensuite les urnes indiscernables. Tu auras la vie sauve en tirant une boule blanche. »

1. Quelle est la probabilité qu'Ali Baba ait la vie sauve s'il place les 4 boules blanches dans la 1^{re} urne et les 4 noires dans la 2^{e} ?
2. Idem avec 2b+2n dans la 1^{re} urne et 2b+2n dans la 2^{e}.
3. Idem avec 3b dans la 1^{re} urne et 1b+4n dans la 2^{e}.
4. Comment Ali Baba maximise-t-il ses chances ?

1.18

Les assistants sociaux travaillant pour une clinique psychiatrique sont si occupés qu'en moyenne seuls 60 % des patients prospectifs téléphonant pour

la 1re fois obtiendront une communication avec l'un de ces assistants. On demande aux autres de laisser leur numéro de téléphone. Trois fois sur 4 un assistant trouve le temps de rappeler le jour même, autrement le rappel a lieu le lendemain. L'expérience a montré que, dans cette clinique, la probabilité que le patient prospectif demande une consultation est de 0,8 s'il a pu parler immédiatement à un assistant, tandis qu'elle tombe à 0,6 et 0,4 respectivement s'il y a eu rappel du patient le jour même ou le lendemain.

1. Quel pourcentage des patients qui appellent demande une consultation ?
2. Quel pourcentage des patients en consultation n'a pas eu à attendre qu'on les rappelle ?

1.19

On a à disposition 2 tests sanguins pour le dépistage du HIV : d'une part l'ELISA, relativement bon marché (environ 20 €) et raisonnablement fiable, et d'autre part le Western Blot (WB), nettement meilleur mais beaucoup plus cher (environ 100 €).

Un patient vient vers vous, un médecin, avec des symptômes vous suggérant qu'il peut être HIV-positif. Pour ce patient, le prévalence du HIV est estimée par la littérature médicale à $P(A) = P(\text{il est HIV-positif }) = 0{,}01$. Les données concernant des personnes dont on connaît le statut HIV apportent :
$P(\text{ELISA positif} \mid \text{HIV-positif}) = 0{,}95$;
$P(\text{ELISA négatif} \mid \text{HIV-négatif}) = 0{,}98$.
En utilisant le théorème de Bayes, calculer :
$P(\text{HIV-positif} \mid \text{ELISA négatif})$ et $P(\text{HIV-négatif} \mid \text{ELISA positif})$.
Quelle(s) conséquence(s) peut-on en tirer sur l'utilisation de l'ELISA ?

1.20

L'hôpital de Jujuy, petite ville du Nord-Ouest de l'Argentine, compte parmi ses malades 4 % qui sont d'origine basque, 58 % d'origine espagnole, 32 % d'origine indienne et 6 % d'origine italienne. Sachant que 3 % des Indiens ont un sang de rhésus négatif, ainsi que 87 % des Basques et 22 % des populations d'origine latine, quelle est la probabilité pour qu'une éprouvette de sang de rhésus négatif provienne d'un malade d'origine basque ?

1.21

Depuis Genève (GVA) où il habite, Serge veut se rendre à Dublin (DUB) pour assister à un concert de U2. S'y étant pris un peu tard, tous les avions pour aller en Irlande sont presque pleins. Trois itinéraires différents et équiprobables s'offrent à lui : passer par Bruxelles (BRU), Munich (MUC) ou Francfort (FRA).

Nadine, qui est hôtesse d'accueil à l'aéroport, a une bonne expérience et fait l'estimation suivante :

- la correspondance partant de BRU a une probabilité de 1/5 d'être pleine ;
- celle partant de MUC, une probabilité de 1/4 ;
- celle partant de FRA, une probabilité de 1/2.

Il existe encore une possibilité supplémentaire. Si Serge décide de passer par FRA (et la liaison FRA-DUB est complète), il aura le temps de prendre un train rapide qui l'amènera à MUC à temps pour prendre le vol MUC-DUB (à condition qu'une place soit disponible dans l'avion, bien entendu).

Cinq jours plus tard, Serge rencontre David et lui témoigne le plaisir qu'il a eu de pouvoir assister au concert de U2. Quelle est la probabilité qu'il soit passé par MUC ?

1.22

Le petit David est très friand de bonbons ; il en a toujours quelques-uns dans les poches. Manquant d'esprit de décision quant à l'arôme qu'il préfère, il procède au jeu suivant. Dans sa poche gauche, il met 5 bonbons à l'orange et 3 à la fraise et, dans la droite, il en met 4 à l'orange et 2 à la fraise. Il tire ensuite une pièce et si elle donne pile, il pioche à gauche et si elle donne face, il se sert à droite. La pièce est bien sûr parfaitement équilibrée.

1. Quelle est la probabilité qu'après 2 jets, il ait mangé 2 bonbons ayant le même parfum ?
2. Il rentre ensuite chez lui et vide ses poches sur une table. Sa mère, au courant du jeu de son fils, trouve sur la table 7 bonbons à l'orange et 5 à la fraise. Aidez-la à trouver la séquence des 2 jets de pièce la plus probable qu'a eue David.
3. Le lendemain, David n'a plus que des bonbons à l'orange. Il en met 5 à gauche et 2 à droite. Il passe chez l'épicier pour en acheter à la fraise. Sachant qu'il les mettra tous dans la poche droite, combien doit-il en acheter pour qu'au prochain jet, il soit le plus près possible d'avoir autant de chances d'en tirer un à l'orange ou à la fraise ?

1.23

Un tribunal de 3 juges déclare un individu coupable lorsque 2 au moins des 3 juges estiment que cette décision est fondée. On admettra que si l'accusé est effectivement coupable, chaque juge se prononcera dans ce sens avec probabilité 0,7, ceci indépendamment des 2 autres. Cette probabilité tombe à 0,2 dans le cas où l'accusé est innocent. 70 % des accusés sont coupables. Calculer la probabilité que le juge 3 vote coupable dans chacune des situations suivantes :

1. les juges 1 et 2 l'ont fait ;
2. le juge 1 a voté coupable ou le juge 2 a voté coupable ;

3. les juges 1 et 2 ont voté tous deux non coupables.

1.24

Freddy fait une sauce au vin que le monde entier vient goûter. Comme elle est très délicate, il la rate 1 fois sur 10 s'il utilise du Bordeaux ou du Bourgogne et 1 fois sur 5 avec du Côtes-du-Rhône. Dans sa cuisine, Freddy a une bouteille ouverte dont il a perdu l'étiquette. Connaissant la proportion de ces 3 vins dans sa cave, il estime que les chances que cette bouteille soit un Bordeaux, un Bourgogne ou un Côtes-du-Rhône sont respectivement 40 %, 30 % et 30 %. Freddy utilise cette bouteille pour faire sa sauce et la rate. Quelles doivent être ses nouvelles estimations sur la provenance de la bouteille ?

Corrigés

1.1

On numérote les chapeaux $1, 2, 3$ et 4. Il y a $4! = 24$ issues possibles $w_1 = (1, 2, 3, 4)$, $w_2 = (1, 2, 4, 3), \ldots$, où $w_1, \ldots, w_{24}$ sont les issues possibles.

1. On compte les issues favorables, à savoir celles qui n'ont ni le 1 en 1re position, ni le 2 en 2e, ni le 3 en 3e, ni le 4 en 4e. On dénombre alors 9 issues favorables. La probabilité est donc de $\frac{9}{24} = \frac{3}{8}$.
2. On procède de la même manière en choisissant comme issues favorables celles qui ont exactement 2 chapeaux placés au bon endroit. On en dénombre 6 et la probabilité est de $\frac{6}{24} = \frac{1}{4}$.

1.2

Avec 2 dés de 6 faces, il y a 36 issues possibles. Aurélie marque 1 point si les dés montrent une des 8 combinaisons suivantes

(2,3) , (6,1) , (2,5) , (4,3) , (3,2) , (1,6) , (5,2) ou (3,4).

Nicolas marque 1 point si les dés donnent

(2,2) , (4,1) , (5,1) , (4,2) , (3,3) , (2,4) , (1,5) ou (1,4),

soit 8 issues favorables également. Donc

$$P(\text{Aurélie marque 1 point}) = P(\text{Nicolas marque 1 point}) = \frac{8}{36} = \frac{2}{9}.$$

Les probabilités de marquer 1 point sont égales.

1.3

Soient F l'événement « avoir une fille » et G l'événement « avoir un garçon ». Il y a $2^4 = 16$ issues possibles pour une famille de 4 enfants. On dénombre 6 issues « avoir 2 filles et 2 garçons » :

$(GGFF)$, $(GFGF)$, $(GFFG)$, $(FGGF)$, $(FGFG)$, $(FFGG)$,

et la probabilité cherchée est $\frac{6}{16} = \frac{3}{8}$. Il y a donc moins de familles bien réparties avec 4 enfants.

1.4

Le nombre d'issues possibles lorsqu'on tire 2 cartes parmi 52 est C_2^{52}. Le nombre d'issues favorables « (un 10 OU un valet OU une dame OU un roi) ET un as » est $C_1^4 \cdot C_1^{16}$. Ainsi

$$P(\textit{blackjack}) = \frac{\begin{pmatrix} 4 \\ 1 \end{pmatrix}\begin{pmatrix} 16 \\ 1 \end{pmatrix}}{\begin{pmatrix} 52 \\ 2 \end{pmatrix}} = \frac{32}{663}.$$

1.5

Soit X le rang de la meilleure femme. La probabilité d'avoir $X = 1$ est

$$P(X = 1) = P(\text{femme en premier}) = P(F\ldots) = \frac{1}{2}.$$

La probabilité d'avoir $X = 2$ correspond à la probabilité qu'un homme soit 1$^{\text{er}}$ et une femme 2$^{\text{e}}$:

$$P(X = 2) = P(HF\ldots) = \frac{1}{2} \cdot \frac{5}{9} = \frac{5}{18}.$$

On continue de la même manière

$$\begin{aligned}
P(X = 3) &= P(HHF\ldots) = \frac{1}{2} \cdot \frac{4}{9} \cdot \frac{5}{8} = \frac{5}{36} \\
P(X = 4) &= P(HHHF\ldots) = \frac{1}{2} \cdot \frac{4}{9} \cdot \frac{3}{8} \cdot \frac{5}{7} = \frac{5}{84} \\
P(X = 5) &= P(HHHHF\ldots) = \frac{1}{2} \cdot \frac{4}{9} \cdot \frac{3}{8} \cdot \frac{2}{7} \cdot \frac{5}{6} = \frac{5}{252} \\
P(X = 6) &= P(HHHHHF\ldots) = \frac{1}{2} \cdot \frac{4}{9} \cdot \frac{3}{8} \cdot \frac{2}{7} \cdot \frac{1}{6} \cdot \frac{5}{5} = \frac{1}{252}
\end{aligned}$$

Finalement, la meilleure femme ne pouvant être classée au-delà de la 6$^{\text{e}}$ place, $P(X = 7) = P(X = 8) = P(X = 9) = P(X = 10) = 0$.

1.6

Soient les événements A « obtenir au moins 1 as en lançant 4 fois un dé » et B « obtenir au moins une paire d'as en lançant 24 fois 2 dés ». Ainsi

$$P(A) = 1 - P(\overline{A}) = 1 - \left(\frac{5}{6}\right)^4 \simeq 0{,}518$$

et

$$P(B) = 1 - P(\overline{B}) = 1 - \left(\frac{35}{36}\right)^{24} \simeq 0{,}491.$$

L'événement le plus probable est A.

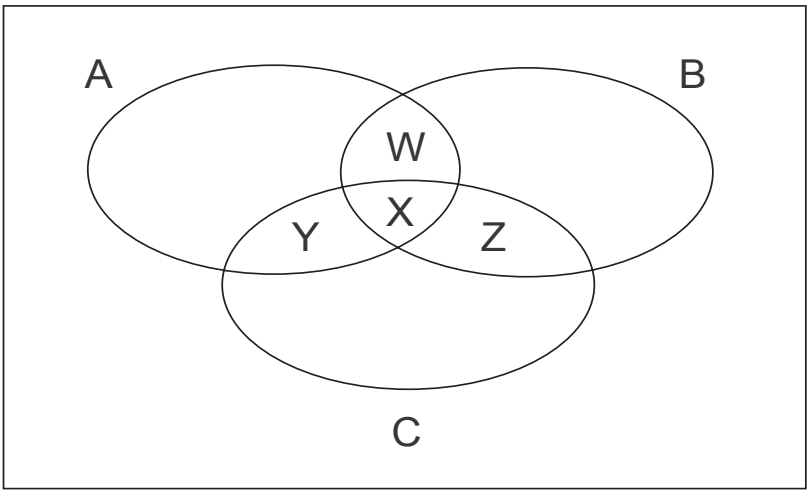


Fig. 1.1 – Diagramme de l'exercice 1.7.

1.7

Soient 3 événements A, B et C.

1. Dans le dessin des ensembles de la figure 1.1, W, X, Y et Z sont des surfaces qui représentent $P(A \cap B)$ par $W + X$, $P(A \cap C)$ par $X + Y$ et $P(B \cap C)$ par $X + Z$.
   Pour le calcul de $P(A \cup B \cup C)$, on commence par additionner les probabilités $P(A)$, $P(B)$ et $P(C)$. En procédant de la sorte, on a compté une fois de trop les aires $W + X$, $Y + X$ et $Z + X$ qu'il faut donc soustraire. Finalement, la surface X a été additionnée 3 fois, puis soustraite 3 fois également. Pour arriver au résultat final, il faut, par conséquent, l'ajouter encore une fois. Ainsi

$$\begin{aligned} &P(A \cup B \cup C) = \\ &\quad = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\qquad + P(A \cap B \cap C). \end{aligned}$$

2. Par les axiomes de la théorie des probabilités, on obtient le même résultat

$$\begin{aligned} &P(A \cup B \cup C) = \\ &\quad = P(A \cup B) + P(C) - P((A \cup B) \cap C) \\ &\quad = P(A) + P(B) - P(A \cap B) + P(C) - P((A \cap C) \cup (B \cap C)) \\ &\quad = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\qquad + P(A \cap B \cap C). \end{aligned}$$

1.8

Soient F l'événement « avoir une fille » et G l'événement « avoir un garçon ». Dans une famille avec 2 enfants, il y a 4 issues possibles

$$(GG\ ,\ GF\ ,\ FG\ ,\ FF).$$

1. On cherche
$$P(GG \mid GF \cup GG) = \frac{P(GG)}{P(GF) + P(GG)} = \frac{1/4}{1/4 + 1/4} = \frac{1}{2}.$$
2. Cette fois,
$$P(GG \mid GF \cup FG \cup GG) = \frac{1}{3}.$$

1.9

Soient « $\neq$ » l'événement « les 2 dés donnent des résultats différents » et S la variable aléatoire « somme des 2 dés ».

1.
$$P(6 \mid \neq) = \frac{P(6 \cap \neq)}{P(\neq)} = \frac{10/36}{30/36} = \frac{1}{3}.$$

2.
$$\begin{aligned}
&P(6 \mid S = i) = \frac{P(6 \cap S = i)}{P(S = i)} = 0 \quad i = 1, 2, 3, 4, 5, 6\\
&P(6 \mid S = 7) = \frac{1}{3}\\
&P(6 \mid S = 8) = \frac{2}{5}\\
&P(6 \mid S = 9) = \frac{1}{2}\\
&P(6 \mid S = 10) = \frac{2}{3}\\
&P(6 \mid S = 11) = 1\\
&P(6 \mid S = 12) = 1\\
&P(6 \mid S = i) = 0 \quad i > 12.
\end{aligned}$$

1.10

Soit A (respectivement B, C, D et E) l'événement « la panne provient de la composante A (respectivement B, C, D et E) ».

1.
$$P(B \mid \overline{A}) = \frac{P(B \cap \overline{A})}{P(\overline{A})}$$

Puisque $B \subset \overline{A}$

$$\frac{P(B \cap \overline{A})}{P(\overline{A})} = \frac{P(B)}{1 - P(A)} = \frac{0{,}3}{1 - 0{,}35} = \frac{30}{65} \simeq 0{,}46.$$

2. Par le même raisonnement, $(C \cup D) \subset (\overline{A} \cup \overline{B})$. De plus, les pannes ne peuvent pas être simultanées ce qui implique que tous les événements sont indépendants. Ainsi

$$P(C \cup D \mid \overline{A} \cup \overline{B}) = \frac{P(C \cup D)}{P(C \cup D \cup E)} = \frac{P(C) + P(D)}{P(C) + P(D) + P(E)} = \frac{30}{35} \simeq 0{,}86.$$

1.11

Soit A (respectivement B et C) l'événement « l'employé travaille dans l'entrepôt A » (respectivement B et C) et F l'événement « l'employé est une femme ». On trouve alors

$$\begin{aligned} P(C \mid F) &= \frac{P(F \mid C)P(C)}{P(F \mid A)P(A) + P(F \mid B)P(B) + P(F \mid C)P(C)} \\ &= \frac{0{,}7\frac{100}{225}}{0{,}5\frac{50}{225} + 0{,}6\frac{75}{225} + 0{,}7\frac{100}{225}} = \frac{1}{2}. \end{aligned}$$

1.12

Soit F (respectivement R et H) l'événement « R. Federer (respectivement A. Roddick et L. Hewitt) gagne le tournoi ». On cherche $P(R \mid \overline{F})$. De

$$P(R) = P(R \mid F)P(F) + P(R \mid \overline{F})P(\overline{F}),$$

on déduit

$$P(R \mid \overline{F}) = \frac{P(R) - P(R \mid F)P(F)}{P(\overline{F})} = \frac{3/10 - 0 \cdot 4/10}{6/10} = \frac{1}{2}.$$

De la même manière, on trouve

$$P(H \mid \overline{F}) = \frac{1}{3}.$$

1.13

Soient PA l'événement « Mme Parisod a une fille » et GL l'événement « le docteur Gluck a prévu une fille ». On trouve

$$\begin{aligned} P(PA \mid GL) &= \frac{P(GL \mid PA)P(PA)}{P(GL \mid PA)P(PA) + P(GL \mid \overline{PA})P(\overline{PA})} \\ &= \frac{19/20 \cdot 1/2}{19/20 \cdot 1/2 + 1/10 \cdot 1/2} = \frac{19}{21} \simeq 0{,}91. \end{aligned}$$

1.14

Soient A l'événement « avoir 1 accident ou plus » et B_1 (respectivement B_2 et B_3) l'événement « appartenir à la catégorie bas (respectivement moyen et haut) risque ».

1.

$$\begin{aligned} P(A) &= P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + P(A \mid B_3)P(B_3) \\ &= 0{,}05 \cdot 0{,}2 + 0{,}15 \cdot 0{,}5 + 0{,}3 \cdot 0{,}3 = 0{,}175. \end{aligned}$$

2.

$$P(B_1 \mid \overline{A}) = \frac{P(\overline{A} \mid B_1)P(B_1)}{P(\overline{A})} = \frac{0{,}95 \cdot 0{,}2}{0{,}825} \simeq 0{,}23.$$

1.15

Soient E_i l'événement « l'avion est dans la région i » et M_i l'événement « on ne trouve pas l'avion dans la région i », avec $i = 1\,,2\,,3$. On sait que

$$P(E_i) = \frac{1}{3} \quad \text{et} \quad P(M_i \mid E_i) = \alpha_i, \quad i = 1\,,2\,,3.$$

On cherche les probabilités que l'avion se trouve dans une des 3 régions en sachant que les recherches dans la zone 1 n'ont rien donné

$$P(E_1 \mid M_1) = \frac{P(E_1) \cdot P(M_1 \mid E_1)}{\sum_{i=1}^{3} P(M_1 \mid E_i) \cdot P(E_i)} = \frac{\alpha_1 \cdot 1/3}{\alpha_1 \cdot 1/3 + 1/3 + 1/3} = \frac{\alpha_1}{\alpha_1 + 2}.$$

Dans ce calcul, on a tenu compte du fait qu'il est certain de ne pas trouver l'avion dans la zone 1 si ce dernier se trouve en réalité dans les zones 2 ou 3 ($P(M_1 \mid E_2) = P(M_1 \mid E_3) = 1$). On procède de la même manière pour

$$P(E_2 \mid M_1) = \frac{P(E_2) \cdot P(M_1 \mid E_2)}{\sum_{i=1}^{3} P(M_1 \mid E_i) \cdot P(E_i)} = \frac{1}{\alpha_1 + 2} = P(E_3 \mid M_1).$$

1.16

Soient les événements $W =$ « la météo prévoit la pluie », $R=$« il pleut » et $U =$ « Mr. Pickwick prend son parapluie ». On sait que $P(R) = 1/2$. De plus, on peut dessiner l'arbre de la figure 1.2, avec lequel on calcule

1.

$$P(U) = \frac{1}{2}\frac{2}{3} + \frac{1}{2}\frac{1}{3} + \frac{1}{2}\frac{1}{3}\frac{1}{3} + \frac{1}{2}\frac{2}{3}\frac{1}{3} = \frac{2}{3},$$

2.

$$P(\overline{U} \mid R) = \frac{P(\overline{U} \cap R)}{P(R)} = \frac{1/2 \cdot 1/3 \cdot 2/3}{1/2} = \frac{2}{9},$$

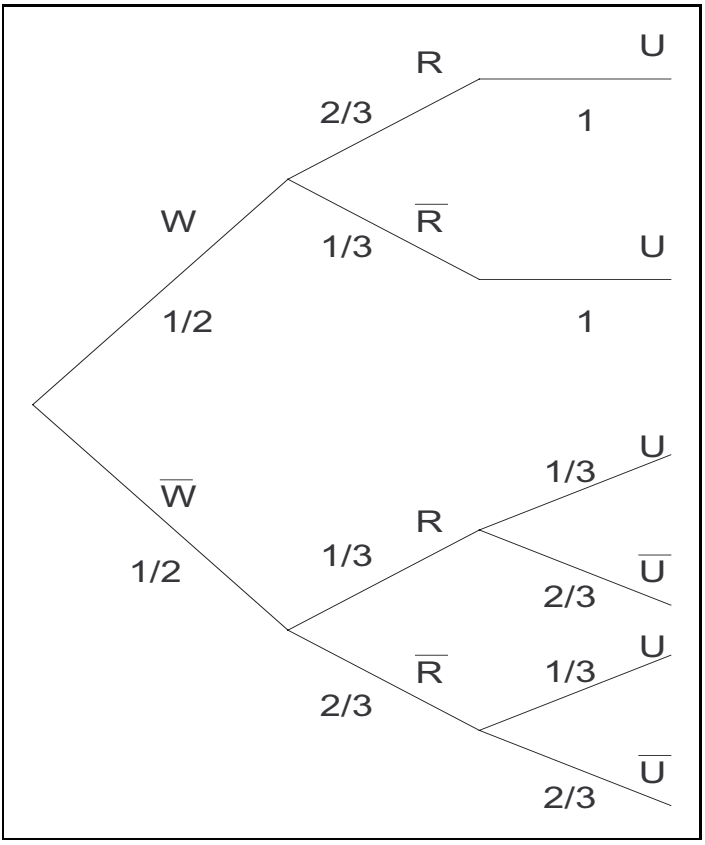


Fig. 1.2 – Arbre de l'exercice 1.16.

3.
$$P(\overline{R} \mid U) = \frac{P(\overline{R} \cap U)}{P(U)} = \frac{1/2 \cdot 2/3 \cdot 1/3 + 1/2 \cdot 1/3 \cdot 1}{2/3} = \frac{5}{12}.$$

1.17

Soient les événements U_1 (U_2) « Ali Baba tire dans la 1re (2e) urne » et S « Ali Baba a la vie sauve ». Dans tous les cas suivants, on aura

$$P(S) = P(S \mid U_1)P(U_1) + P(S \mid U_2)P(U_2).$$

1. Quatre boules blanches sont placées dans U_1 et 4 noires dans U_2

$$P(S) = 1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{2} = \frac{1}{2} \ ;$$

2. Deux boules blanches et 2 boules noires sont placées dans chacune des 2 urnes
$$P(S) = \frac{1}{2}\frac{1}{2} + \frac{1}{2}\frac{1}{2} = \frac{1}{2} \ ;$$
3. Trois boules blanches sont placées dans U_1 et le reste dans U_2
$$P(S) = 1 \cdot \frac{1}{2} + \frac{1}{5}\frac{1}{2} = \frac{3}{5} \ ;$$
4. Ali Baba maximise ses chances de rester en vie en plaçant 1 boule blanche dans la 1re urne et toutes les autres dans la 2e
$$P(S) = 1 \cdot \frac{1}{2} + \frac{3}{7}\frac{1}{2} = \frac{5}{7}.$$

1.18

Soient les événements A « le patient demande une consultation » et B_1 (respectivement B_2 et B_3) « le patient obtient immédiatement (respectivement le soir même et le lendemain) un entretien téléphonique ».

1. Le pourcentage des patients demandant une consultation parmi ceux qui appellent est
$$\begin{aligned} P(A) &= \sum_{i=1}^{3} P(A \mid B_i)P(B_i) = \\ &= 0{,}8 \cdot 0{,}6 + 0{,}6 \cdot 0{,}4 \cdot 0{,}75 + 0{,}4 \cdot 0{,}4 \cdot 0{,}25 = 0{,}7 = 70 \ \%. \end{aligned}$$
2. Le pourcentage des patients en consultation qui n'ont pas eu à attendre d'être rappelés est
$$P(B_1 \mid A) = \frac{P(A \mid B_1)P(B_1)}{P(A)} = \frac{0{,}8 \cdot 0{,}6}{0{,}7} \simeq 68{,}6 \ \%.$$

1.19

Soient les événements HIV^+ (respectivement HIV^-) « le patient est HIV-positif (respectivement négatif) » et E^+ (respectivement E^-) « l'ELISA est positif (respectivement négatif) ». On trouve
$$\begin{aligned} P(HIV^+ \mid E^-) &= \frac{P(E^- \mid HIV^+)P(HIV^+)}{P(E^- \mid HIV^+)P(HIV^+) + P(E^- \mid HIV^-)P(HIV^-)} \\ &= \frac{0{,}05 \cdot 0{,}01}{0{,}05 \cdot 0{,}01 + 0{,}98 \cdot 0{,}99} = 0{,}05 \ \% \end{aligned}$$
et
$$\begin{aligned} P(HIV^- \mid E^+) &= \frac{P(E^+ \mid HIV^-)P(HIV^-)}{P(E^+ \mid HIV^+)P(HIV^+) + P(E^+ \mid HIV^-)P(HIV^-)} \\ &= \frac{0{,}02 \cdot 0{,}99}{0{,}02 \cdot 0{,}99 + 0{,}95 \cdot 0{,}01} = 0{,}68. \end{aligned}$$

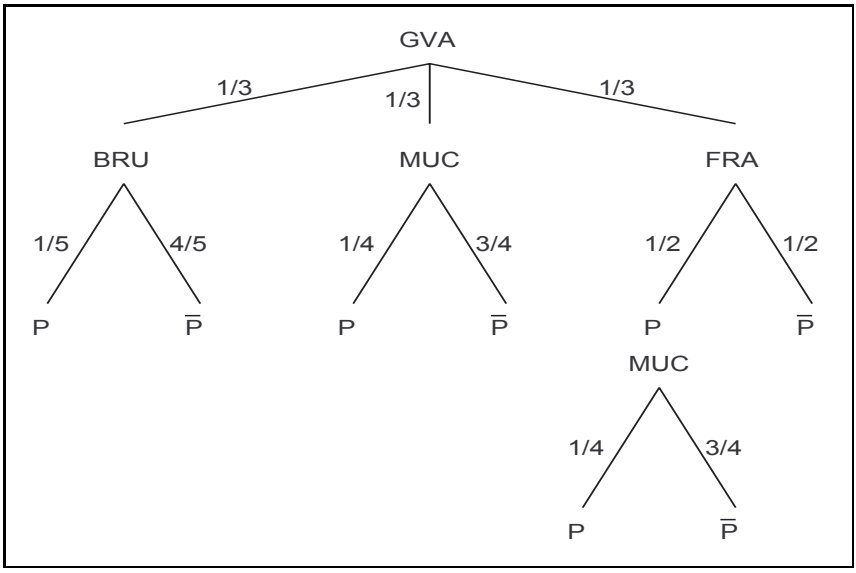


Fig. 1.3 – Arbre de l'exercice 1.21.

Au vu de ces résultats, il est évident que l'ELISA est un très bon test pour déterminer si un patient n'est pas HIV-positif. Par contre, il est inefficace pour décider de la contamination d'un patient. Dans l'exercice 2.1, un autre test est utilisé pour déterminer si un patient est vraiment malade.

1.20

Soient les événements B (respectivement I et L) « le malade est d'origine basque (respectivement indienne et latine) » et RH^+ (respectivement RH^-) « le sang contenu dans une éprouvette est de rhésus positif (respectivement négatif) ». On cherche la probabilité que le sang de rhésus négatif contenu dans une éprouvette provienne d'un malade d'origine basque :

$$\begin{aligned} P(B \mid RH^-) &= \frac{P(RH^- \mid B)P(B)}{P(RH^-)} \\ &= \frac{0{,}97 \cdot 0{,}04}{0{,}87 \cdot 0{,}04 + 0{,}03 \cdot 0{,}32 + 0{,}22 \cdot 0{,}64} \simeq 0{,}19. \end{aligned}$$

1.21

Soient les événements BRU (respectivement MUC et FRA) « Serge passe par Bruxelles (respectivement Munich et Francfort) » et P l'événement « Serge ne peut pas partir car l'avion est plein ». On construit l'arbre de la figure 1.3

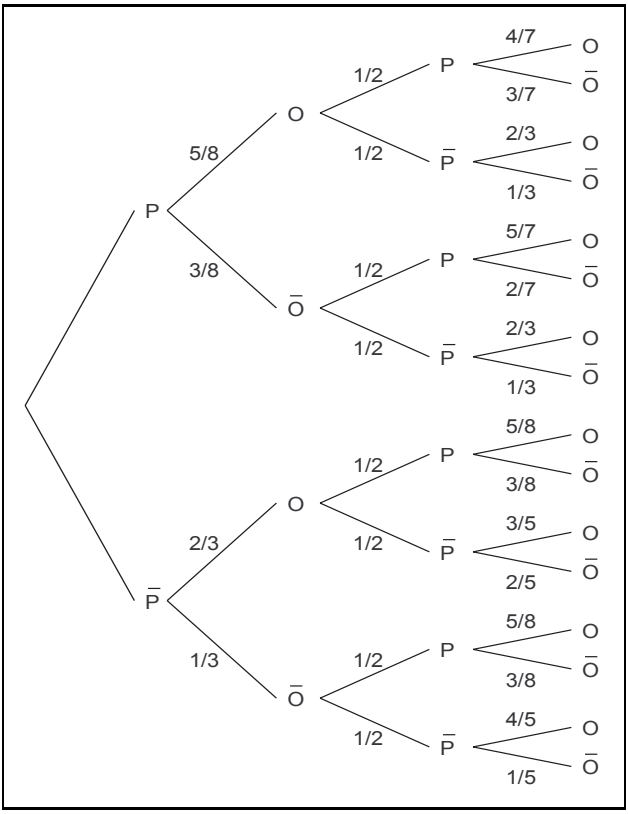


Fig. 1.4 – Arbre de l'exercice 1.22.

pour trouver

$$\begin{aligned} P(MUC \mid \overline{P}) &= \frac{P(MUC \cap \overline{P})}{P(\overline{P})} \\ &= \frac{1/3 \cdot 3/4 + 1/3 \cdot 1/2 \cdot 3/4}{1/3 \cdot 4/5 + 1/3 \cdot 3/4 + 1/3 \cdot 1/2 \cdot 3/4 + 1/3 \cdot 1/2} \\ &= \frac{45}{97} \simeq 0{,}46. \end{aligned}$$

1.22

Soient les événements P « la pièce donne pile » et O « le bonbon tiré est à l'orange ». Au départ, David a 5 bonbons à l'orange et 3 à la fraise dans la poche gauche ($5O+3\overline{O}$) et 4 à l'orange et 2 à la fraise dans la droite ($4O+2\overline{O}$).

On construit l'arbre de la figure 1.4 en faisant attention au fait que les tirages sont sans remise ; ce qui implique que les probabilités changent au fur

et à mesure des événements.

1. La probabilité de tirer 2 bonbons ayant le même parfum est alors

$$P(\text{2 parfums identiques}) = \\ = \frac{1}{4}\left(\frac{5}{8}\frac{4}{7} + \frac{5}{8}\frac{2}{3} + \frac{3}{8}\frac{2}{7} + \frac{3}{8}\frac{1}{3} + \frac{2}{3}\frac{5}{8} + \frac{2}{3}\frac{3}{5} + \frac{1}{3}\frac{3}{8} + \frac{1}{3}\frac{1}{5}\right) \simeq 0{,}50.$$

2. Il reste 7 bonbons à l'orange et 5 à la fraise. Cela signifie que David a tiré 2 bonbons à l'orange. Pour trouver la séquence la plus probable, on calcule la probabilité pour chacune des quatre séquences possibles

$$\begin{aligned} P(PP \mid OO) &= \frac{P(OO \mid PP)P(PP)}{P(OO)} \simeq \frac{5/8 \cdot 4/7 \cdot 1/4}{0{,}40} \simeq 0{,}23 \\ P(\overline{P}P \mid OO) &\simeq \frac{2/3 \cdot 5/8 \cdot 1/4}{0{,}40} \simeq 0{,}26 \\ P(P\overline{P} \mid OO) &\simeq \frac{2/3 \cdot 5/8 \cdot 1/4}{0{,}40} \simeq 0{,}26 \\ P(\overline{PP} \mid OO) &\simeq \frac{2/3 \cdot 3/5 \cdot 1/4}{0{,}40} \simeq 0{,}25 \end{aligned}$$

avec

$$P(OO) = \frac{1}{4}\left(\frac{5}{8}\frac{4}{7} + \frac{5}{8}\frac{2}{3} + \frac{2}{3}\frac{3}{5} + \frac{2}{3}\frac{5}{8}\right) \simeq 0{,}40.$$

Les 2 séquences les plus probables sont $P\overline{P}$ et $\overline{P}P$ (elles sont équivalentes).

3. Soit X le nombre de bonbons à la fraise que David doit acheter. On veut que $P(O) = 1/2$

$$P(O) = P(O \mid P)P(P) + P(O \mid \overline{P})P(\overline{P}) = 1 \cdot \frac{1}{2} + \frac{2}{2+X} \cdot \frac{1}{2} = \frac{1}{2}$$

$$\Rightarrow \frac{2}{2+X} = 0 \Rightarrow X \to \infty.$$

Il faudrait que David achète une infinité de bonbons pour équilibrer les chances de tirer un bonbon à l'orange ou à la fraise.

1.23

Soient les événements JC_i « le juge i vote coupable » ($i = 1\,,2\,,3$) et C « l'accusé est effectivement coupable ».

1. On cherche d'abord la probabilité que le 3e juge vote coupable en sachant que les 2 premiers votent coupable

$$P(JC_3 \mid JC_1 \cap JC_2) = \frac{P(JC_3 \cap JC_1 \cap JC_2)}{P(JC_1 \cap JC_2)}.$$

Il faut faire attention de ne pas utiliser l'indépendance dans cette équation directement. Il y a seulement indépendance entre les décisions des juges conditionnées à la culpabilité (ou non culpabilité) de l'accusé. Ainsi, il est nécessaire de faire apparaître des probabilités conditionnelles pour utiliser l'hypothèse d'indépendance :

$$\begin{aligned}
&P(JC_3 \mid JC_1 \cap JC_2) = \\
&= \frac{P(JC_3 \cap JC_1 \cap JC_2 \mid C)P(C) + P(JC_3 \cap JC_1 \cap JC_2 \mid \overline{C})P(\overline{C})}{P(JC_1 \cap JC_2 \mid C)P(C) + P(JC_1 \cap JC_2 \mid \overline{C})P(\overline{C})} \\
&= \frac{P(JC_3 \mid C) \cdot P(JC_1 \mid C) \cdot P(JC_2 \mid C) \cdot P(C)}{P(JC_1 \mid C) \cdot P(JC_2 \mid C) \cdot P(C) + \cdot P(JC_1 \mid \overline{C}) \cdot P(JC_2 \mid \overline{C}) \cdot P(\overline{C})} \\
&+ \frac{P(JC_3 \mid \overline{C}) \cdot P(JC_1 \mid \overline{C}) \cdot P(JC_2 \mid \overline{C}) \cdot P(\overline{C})}{P(JC_1 \mid C) \cdot P(JC_2 \mid C) \cdot P(C) + \cdot P(JC_1 \mid \overline{C}) \cdot P(JC_2 \mid \overline{C}) \cdot P(\overline{C})} \\
&= \frac{0{,}7^3 \cdot 0{,}7 + 0{,}2^3 \cdot 0{,}3}{0{,}7^2 \cdot 0{,}7 + 0{,}2^2 \cdot 0{,}3} = \frac{97}{142} \simeq 0{,}68.
\end{aligned}$$

2. Le raisonnement pour les 2 dernières questions est le même qu'en 1.

$$\begin{aligned}
&P(JC_3 \mid JC_1 \cap \overline{JC_2}) = \\
&= \frac{P(JC_3 \cap JC_1 \cap \overline{JC_2})}{P(JC_1 \cap \overline{JC_2})} = \frac{0{,}7^2 \cdot 0{,}3 \cdot 0{,}7 + 0{,}2^2 \cdot 0{,}8 \cdot 0{,}3}{0{,}7 \cdot 0{,}3 \cdot 0{,}7 + 0{,}2 \cdot 0{,}8 \cdot 0{,}3} \simeq 0{,}58
\end{aligned}$$

3.

$$\begin{aligned}
&P(JC_3 \mid \overline{JC_1} \cap \overline{JC_2}) = \\
&= \frac{P(JC_3 \cap \overline{JC_1} \cap \overline{JC_2})}{P(\overline{JC_1} \cap \overline{JC_2})} = \frac{0{,}7 \cdot 0{,}3^2 \cdot 0{,}7 + 0{,}2 \cdot 0{,}8^2 \cdot 0{,}3}{0{,}3^2 \cdot 0{,}7 + 0{,}8^2 \cdot 0{,}3} \simeq 0{,}32.
\end{aligned}$$

1.24

Soient les événements R « la sauce est ratée » et CR (respectivement BD et BG) « la bouteille est du Côtes-du-Rhône (respectivement du Bordeaux et du Bourgogne) ». On commence par calculer la probabilité que le vin soit du Bordeaux en sachant que la sauce est ratée :

$$\begin{aligned}
P(BD \mid R) &= \frac{P(R \mid BD)P(BD)}{P(R \mid BD)P(BD) + P(R \mid BG)P(BG) + P(R \mid CR)P(CR)} \\
&= \frac{0{,}1 \cdot 0{,}4}{0{,}1 \cdot 0{,}4 + 0{,}1 \cdot 0{,}3 + 0{,}2 \cdot 0{,}3} = \frac{4}{13} \simeq 0{,}31.
\end{aligned}$$

On procède de la même manière pour les 2 autres vins et on trouve

$$P(BG \mid R) = \frac{3}{13} \simeq 0{,}23 \quad \text{et} \quad P(CR \mid R) = \frac{6}{13} \simeq 0{,}46.$$

Ainsi, en sachant que la sauce a raté, la bouteille a environ 31 % de chance d'être du Bordeaux, 23 % d'être du Bourgogne et 46 % d'être du Côtes-du-Rhône.

Chapitre 2

Variables aléatoires discrètes

Introduction

La notion de variable aléatoire est fondamentale dans le calcul des probabilités. La première partie de ce chapitre présente des exercices sur les outils généraux de travail avec des variables aléatoires discrètes. Ceux-ci incluent la loi de probabilité, la fonction de répartition, l'espérance et la variance (exercices 2.1 à 2.4). Ensuite on a recueilli les exercices concernant quelques lois discrètes de probabilité. On traite en particulier les deux lois discrètes les plus importantes, à savoir la loi binomiale et la loi de Poisson (exercices 2.5 à 2.18).

La loi binomiale

La structure qui génère la loi binomiale peut être résumée de la façon suivante. Considérons une expérience aléatoire constituée de n épreuves indépendantes. À chaque épreuve deux issues sont possibles, « succès » avec probabilité p et « échec » avec probabilité $1-p$. La variable aléatoire X qui compte le nombre de « succès » dans la série de n épreuves suit une loi binomiale avec paramètres n et p, $X \sim Bin(n,p)$, c'est-à-dire

$$P(X=k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}, \qquad k = 1, \ldots, n.$$

On a $E(X) = np$ et $\text{var}(X) = np(1-p)$.

Cette structure est importante car elle se retrouve dans de nombreuses situations différentes, comme par exemple le contrôle de qualité, la modélisation en temps discret du prix d'un actif financier et les prévisions dans le cadre d'un sondage avant une votation.

La loi de Poisson

La loi de Poisson peut être vue comme un cas limite de la loi binomiale quand $n \to \infty$, $p \to 0$, et $np = \lambda = const.$ Ceci modélise la situation avec un grand nombre d'épreuves (individus, objets ...) et une petite probabilité p d'obtenir un « succès » à chaque épreuve. La loi de Poisson est utilisée pour modéliser par exemple le nombre de clients qui arrivent à un guichet, le nombre de pannes d'une machine, ou le nombres d'accidents de la circulation pendant une certaine période. La loi de probabilité de Poisson d'une variable aléatoire X, $X \sim P(\lambda)$, $\lambda > 0$, s'écrit

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

On a $E(X) = \text{var}(X) = \lambda$.

Notes historiques

Le cas spécial $n = 1$ de la loi binomiale est appelé la loi de Bernoulli (du nom du mathématicien suisse Jacques Bernoulli). La loi de Poisson fut introduite en 1837 par Siméon Denis Poisson et utilisée ensuite en 1898 par Bortkewitz pour modéliser « le nombre de soldats tués par ruades de cheval dans l'armée prussienne pendant 20 ans ».

Références (théorie)

Ross, chapitres 4 et 7 [1] et Pitman, chapitres 2 et 3 [2].

Exercices

Moments et probabilités

2.1 (Suite de l'exercice 1.19)

On note :
c_1 = coût du test ELISA = 20 €;
c_2 = coût du test WB = 100 €;
L_I = coût supplémentaire engendré par le faux diagnostic : conclure que le patient est HIV-négatif alors qu'il est HIV-positif = 1 000 000 €;
L_{II} = coût supplémentaire engendré par le faux diagnostic : conclure que le patient est HIV-positif alors qu'il est HIV-négatif = 1 000 €.

La 1re stratégie est la suivante : on fait le test de l'ELISA. S'il est négatif, on considère que le patient est HIV-négatif. Sinon qu'il est HIV-positif.

Une 2e stratégie est comme la 1re, sauf que si l'ELISA est positif, on fait en plus le WB. Si ce dernier est positif, on considère que le patient est HIV-positif, sinon qu'il est HIV-négatif.

On formalise la 2e stratégie ainsi
- les issues possibles sont composées de triplets (statut HIV, statut ELISA, statut WB). Par exemple, on considère $(+,+,+)$;
- une probabilité associée à chaque issue. Pour $(+,+,+)$ c'est 0,00945 ;
- à chaque issue est associée une valeur, définissant ainsi une variable aléatoire U, appelée utilité. Pour $(+,+,+)$, c'est $-c_1 - c_2 = -120$ €. Pour $(+,-,-)$, c'est $-c_1 - L_I = -1\,000\,020$ €.

1. Compléter le tableau suivant résumant la 2e stratégie.

Prob.	Vrai statut HIV	Statut ELISA	Statut WB	Utilité
0,00945	+	+	+	$-c_1 - c_2$
0,00005	+	+	-	$-c_1 - c_2 - L_I$
0,00004	+	-	(+)	$-c_1 - L_I$
	+	-	(-)	$-c_1 - L_I$
0,0001	-	+	+	$-c_1 - c_2 - L_{II}$
0,0197	-	+	-	
0,00095	-	-	(+)	$-c_1$
0,96925	-	-	(-)	

2. Calculer l'espérance de l'utilité.
3. Pour la 1re stratégie, faire de même (tableau et espérance de l'utilité).
4. Quelle stratégie faut-il adopter ? (Discuter en fonction de L_I et L_{II}.)

2.2

Vous vous promenez en forêt et vous tombez nez-à-nez avec un ours affamé. Heureusement, vous ne sortez jamais sans votre fusil, mais vous n'avez cependant que 3 cartouches en poche. Sachant qu'à chaque tir, vous avez une probabilité p d'atteindre votre objectif, c'est-à-dire de tuer l'ours, et en notant X la variable aléatoire représentant le nombre de tirs strictement nécessaires pour tuer l'ours, répondez aux questions suivantes.

1. Calculez $P(X = 1)$, $P(X = 2)$, et $P(X = 3)$ en fonction de p.
2. Donnez également $P(X \leq 3)$. Que représente cette probabilité ?
3. Sachant que $E(X) = 1/p$, quelle est la valeur minimale de p pour espérer sortir vivant de ce mauvais pas ?
4. En réalité, l'angoisse et le stress diminuent vos aptitudes au tir si bien que la probabilité d'atteindre l'ours au i^e essai est donnée par p^i. Donnez dans ce cas $P(X = 1)$, $P(X = 2)$ et $P(X = 3)$ et comparez ces probabilités avec celles obtenues au point 1.

2.3

Une personne possède 4 clefs parmi lesquelles une seule ouvre la porte. Elle les essaie au hasard en éliminant celles qui ne marchent pas. On pose X « le nombre d'essais pour ouvrir la porte ».

1. Calculer la loi de probabilité de X, c'est-à-dire $P(X = k)$ avec $k = 1, 2, 3, 4$.
2. Calculer $E(X)$ et $\mathrm{Var}(X)$.

2.4

Le fisc répartit les ménages en 5 classes de revenu. Les données de l'année fiscale 2005 lui apportent :
Classe 1 : 19 000 ménages.
Classe 2 : 45 000 ménages.
Classe 3 : 28 000 ménages.
Classe 4 : 9 000 ménages.
Classe 5 : 2 000 ménages.

Notons X la variable aléatoire « classe d'appartenance ».

1. Trouver la fonction de répartition de X.
2. Calculer $P(2 < X \leq 4)$ et $P(X > 4)$.
3. Calculer $E(X)$ et $\mathrm{Var}(X)$.

Loi binomiale et loi de Poisson

2.5

Un canal de transmission d'information ne peut traiter que des 0 et des 1. À cause de perturbations dues à l'électricité statique chaque chiffre transmis l'est avec une probabilité d'erreur de 0,2. Admettons que l'on veuille transmettre un message important limité à un signal binaire. Pour éviter une erreur on transmettra 00000 au lieu de 0 et 11111 au lieu de 1. Si le récepteur décode suivant la règle de la majorité, quelle est la probabilité que le message soit mal interprété ?

2.6

Un épicier reçoit un lot de pommes dont 25 % sont avariés. Il charge un employé de préparer des emballages de 5 pommes chacun. Celui-ci, négligent, ne se donne pas la peine de jeter les fruits avariés. Chaque client qui trouve, dans l'emballage qu'il achète, 2 fruits ou plus qui sont avariés, revient au magasin se plaindre.

1. Soit X le « nombre de pommes avariées dans un emballage ». Déterminer la loi de probabilité de X.
2. Quelle est la probabilité pour qu'un client donné se plaigne auprès de son épicier ?
3. Si l'épicier a 100 clients qui achètent des pommes ce jour-là, combien y aura-t-il de plaintes ?

2.7

Giovanni, dit Gianni, a décidé de parcourir cet été 10 000 km en Fiat Ritmo. Or la probabilité d'avoir un accident sur 1 km est de 1/10 000. En prenant connaissance de cette probabilité, Gianni décide alors d'annuler son long voyage en prétextant d'être sûr d'avoir un accident. Êtes-vous d'accord avec Gianni ? Si ce n'est pas le cas, quelle est l'erreur commise par notre ami Gianni ? Quelle est alors approximativement la probabilité que Gianni ait un accident ?

2.8

Un journaliste se voit remettre une liste de personnes à interviewer. Il doit interroger 5 personnes au moins. Les interviewés potentiels n'acceptent de parler qu'avec une probabilité de 2/3, indépendamment les uns des autres. Quelle est la probabilité qu'il puisse réaliser ses 5 entretiens si la liste compte 5 noms ? Et si elle en compte 8 ?

2.9

Des études effectuées par les compagnies aériennes montrent qu'il y a une probabilité de 0,05 que chaque passager ayant fait une réservation n'effectue pas le vol. À la suite de ça, SA Airlines vend toujours 94 billets pour ses avions à 90 sièges, tandis que BA Airlines vend toujours 188 billets pour ses avions à 180 sièges. Avec quelle compagnie un passager ayant réservé un siège risque-t-il le plus de ne pas pouvoir prendre place dans l'avion ?

2.10

Madame Gourmande prépare des biscuits aux raisins secs et aux pépites de chocolat. Elle mélange 600 raisins et 400 pépites de chocolat dans la pâte et prépare 500 biscuits. Dès que les biscuits sont prêts, Madame Gourmande en choisit un au hasard pour le goûter.

1. Calculer la probabilité qu'il n'y ait pas de raisins dans le biscuit.
2. Calculer la probabilité qu'il y ait exactement 2 pépites de chocolat.
3. Calculer la probabilité qu'il y ait au moins 2 morceaux (raisins ou pépites de chocolat).

2.11

1. Nadine part à la cueillette des champignons. Elle ne sait pas faire la différence entre un champignon comestible et un champignon toxique. On estime que la proportion de champignons toxiques se trouvant dans les bois s'élève à 0,7.
   (a) Nadine ramasse 6 champignons au hasard. Calculer la probabilité qu'elle ramasse exactement 4 champignons toxiques.
   (b) Nadine invite Serge à une cueillette. Serge connaît bien les champignons ; sur 10 champignons qu'il ramasse, 9 sont comestibles. Ce jour-là, il ramasse 4 champignons et Nadine en ramasse 3. Calculer la probabilité que tous les champignons soient comestibles.
2. Serge cueille en moyenne 12 champignons par heure.
   (a) Calculer la probabilité qu'il ramasse exactement 8 champignons en une heure.
   (b) Calculer la probabilité qu'il ramasse au moins 1 champignon en 20 minutes.

2.12

Les ingénieurs du son préposés à la sonorisation d'un concert en plein air hésitent entre deux solutions : 4 haut-parleurs de 4 000 watts chacun ou 2 haut-

parleurs de 8 000 watts chacun. On suppose que la probabilité qu'un haut-parleur tombe en panne est égale à $p = 0{,}2$ indépendamment du type de haut-parleur et que les pannes se produisent indépendamment les unes des autres. En admettant que le concert peut se dérouler avec au moins 8 000 watts, quelle solution conseillerez-vous à ces ingénieurs ?

Exercices combinés

2.13

Dix chasseurs guettent le passage d'un vol de canards. Lorsque les canards passent en groupe, les chasseurs font tous feu en même temps mais chacun choisit sa cible au hasard indépendamment des autres. On admet que chaque chasseur touche son canard avec la même probabilité p.

1. Combien de canards, en moyenne, survivront au tir lorsque le vol se compose de 20 canards ? Calculer cette moyenne pour diverses valeurs de p.
2. Quel sera le nombre de canards touchés si le vol se compose d'un nombre de canards suivant une loi de Poisson de paramètre 15 ?

2.14

Chacun des soldats d'une troupe de 500 hommes est porteur d'une certaine maladie avec probabilité 1/1 000. Cette maladie est détectable à l'aide d'un test sanguin et, pour faciliter les choses, on ne teste qu'un mélange du sang des 500 soldats.

1. Quelle est la probabilité que le test soit positif, indiquant par là qu'au moins une des personnes est malade ?
2. On suppose que le test a été positif. Quelle est la probabilité que dans ce cas, plus d'une personne soit malade ?

2.15

À l'université, le taux de blocage de mâchoire provoqué par un bâillement profond est de 1 personne pour 1 000 et par mois. On suppose qu'un étudiant a au plus un blocage de mâchoire par mois et que les blocages de mâchoire sont indépendants.

1. Quelle est la probabilité qu'en un mois de l'année académique 2005-2006 il y ait 3 blocages de mâchoire ou plus parmi les étudiants de l'université ? *Indication :* supposer qu'il y a 4 000 étudiants.
2. Quelle est la probabilité qu'au cours de l'année académique 2005-2006 le nombre de mois comptant 3 blocages de mâchoire ou plus soit supérieur

ou égal à 4?
Indication : supposer qu'une année académique est constituée de 8 mois.

3. Le premier mois de l'année académique étant appelé mois 1, quelle est la probabilité que le premier mois où l'on enregistre 3 blocages de mâchoire ou plus soit le mois i, $i = 1, 2, ..., 8$?

2.16

Un jeux de dés très populaires dans les bars anglais est le *chuck-a-luck*. Il consiste pour la banque à jeter 3 dés. Un joueur peut parier sur n'importe quel résultat compris entre 1 et 6. Si exactement un de ces 3 dés montre le chiffre prévu, le joueur récupère sa mise plus un montant équivalent. Si 2 dés montrent ce résultat, le gain net est de 2 pour 1 ; si les 3 dés indiquent le chiffre prévu, le gain net est de 3 pour 1. Si aucun dé ne montre le chiffre choisi par le joueur, ce dernier perd sa mise.

1. Calculer l'espérance de gain lorsque l'enjeu est d'une livre.
2. Quel montant devrait recevoir le joueur si les 3 dés montrent le chiffre prévu pour que le jeu soit *fair* (c'est-à-dire pour que l'espérance de gain soit nulle)?

2.17

On observe l'arrivée de personnes à un guichet, 2 personnes ne pouvant arriver en même temps. Le nombre d'arrivées dans un intervalle de temps de longueur t est une variable aléatoire $N(t)$ distribuée selon une loi de Poisson de paramètre λt. Les arrivées se produisent indépendamment les unes des autres. On choisit un temps $t_0 = 0$. Soit T_k la variable aléatoire qui représente l'arrivée du k^{e} client à partir de t_0.

1. Quelle est la loi de T_1 ?
   Indication : la probabilité que T_1 soit supérieur à t est égale à la probabilité que personne n'arrive dans l'intervalle $[0, t]$.
2. Calculer la fonction de répartition de T_k.
3. Calculer la densité de T_k. De quelle loi s'agit-il?

2.18

Andi va skier et emprunte une des N perches d'un remonte-pente. Entre cet instant et la prochaine remontée, le nombre de skieurs qui se présentent suit une loi géométrique de paramètre p.
Quelle est la probabilité pour Andi de reprendre la même perche?

2.19 (Loi géométrique ou loi de Pascal)

On considère une série d'épreuves indépendantes. À chaque épreuve, on observe un « succès » avec probabilité p et un « échec » avec probabilité $1-p$. Soit X la variable aléatoire discrète suivante :

$X =$ nombre d'épreuves nécessaires pour obtenir le 1^{er} « succès ».

1. Calculer la loi de probabilité de X, c'est-à-dire $P(X = k)$, $k \in \mathbb{N}$. Cette loi est dite *loi géométrique*.
2. Vérifier que $E(X) = 1/p$.
3. Vérifier la propriété « sans mémoire » de la loi géométrique

$$P(X > k \mid X > j) = P(X > k - j), \quad k > j.$$

4. J'ai décidé de vendre ma maison et d'accepter la 1^{re} offre d'achat supérieure à K €. On suppose que les offres d'achat sont des variables aléatoires indépendantes avec fonction de répartition F. Soit N la variable aléatoire discrète suivante :

   $N =$ nombre d'offres d'achat reçues avant de vendre la maison.

   Donner la loi de probabilité de N, c'est-à-dire $P(N = n)$, $n \in \mathbb{N}$.

2.20 (Loi hypergéométrique)

On considère une urne avec N boules dont r sont rouges et $N-r$ sont noires. On tire n boules au hasard sans remise. Soit X la variable aléatoire discrète suivante

$X =$ « nombre de boules rouges parmi les n boules tirées ».

1. Calculer la loi de probabilité de X, c'est-à-dire $P(X = k)$ pour $k = 0, \ldots, \min(n, r)$.
2. Cette loi est dite *loi hypergéométrique*. Son espérance est $E(X) = np$ et sa variance est $\text{var}(X) = np(1-p)\frac{N-n}{N-1}$, où $p = \frac{r}{N}$. Comparer intuitivement avec la loi binomiale.

Corrigés

2.1

Soit U_1 la variable aléatoire « utilité dans la stratégie sans le test WB » et U_2 la variable aléatoire « utilité dans la stratégie avec le test WB ».

1. On complète le tableau donné dans l'énoncé. La somme de toutes les probabilités est égale à 1 ce qui implique que la probabilité manquante est $P(+,-,-) = 0{,}00046$. De plus, on trouve les utilités $U_2(-,+,-) = -c_1 - c_2$ et $U_2(-,-,-) = -c_1$.
2. On calcule l'espérance de U_2

$$\begin{aligned} E(U_2) &= \sum_{i,j,k=\pm} P(i,j,k) \cdot U_2(i,j,k) = \\ &= -c_1 - 0{,}0293c_2 - 0{,}00055L_I - 0{,}0001L_{II} = -573{,}03. \end{aligned}$$

3. On construit le tableau de probabilités et on calcule l'espérance pour U_1.

Prob.	Vrai statut HIV	Statut ELISA	Utilité
0,0095	+	+	$-c_1$
0,0005	+	−	$-c_1 - L_I$
0,0198	−	+	$-c_1 - L_{II}$
0,9702	−	−	$-c_1$

Par exemple, on a calculé

$$P(+,+) = P(+,+,+) + P(+,+,-) = 0{,}00945 + 0{,}00005 = 0{,}0095,$$

et l'espérance est

$$E(U_1) = -c_1 - 0{,}0005L_I - 0{,}0198L_{II} = -539{,}8.$$

4. On choisit la stratégie qui a la plus grande espérance, c'est-à-dire la 1re. Cette conclusion ne tient évidemment compte que de l'utilité. D'un point de vue médical, il est évident que la 2e stratégie est la meilleure puisqu'elle donne des diagnostics beaucoup plus sûrs. On pourrait imaginer d'augmenter L_I et L_{II} pour pénaliser les mauvais diagnostics.

2.2

Soient X le « nombre de tirs nécessaires pour tuer l'ours » et la probabilité $P(\text{tuer l'ours}) = p$.

1.

$$\begin{aligned} P(X=1) &= P(\text{tuer l'ours}) = p \\ P(X=2) &= P(\text{rater l'ours, et le tuer}) = (1-p)p \\ P(X=3) &= P(\text{rater l'ours 2 fois, et le tuer}) = (1-p)^2 p. \end{aligned}$$

2. La probabilité $P(X \leq 3)$ correspond à la probabilité que l'ours reste vivant. Elle vaut

$$P(X \leq 3) = \sum_{i=1}^{3} P(X = i) = p(3 - 3p + p^2).$$

3. On pose $E(X) = 1/p$. Pour rester vivant, il faut que le nombre de tirs disponibles soit plus grand ou égal à la moyenne du nombre de tirs nécessaires pour le tuer

$$E(X) \leq 3 \quad \Leftrightarrow \quad p \geq \frac{1}{3} \quad \Leftrightarrow \quad p_{min} = \frac{1}{3}.$$

4. Le stress diminue la probabilité de tuer l'ours : $P(\text{tuer l'ours au } i^{\text{e}} \text{ coup}) = p^i$. Ainsi

$$\begin{aligned}
\tilde{P}(X = 1) &= p = P(X = 1) \\
\tilde{P}(X = 2) &= (1 - p)p^2 < P(X = 2) \\
\tilde{P}(X = 3) &= (1 - p)(1 - p^2)p^3 < P(X = 3).
\end{aligned}$$

2.3

Soit X la variable aléatoire « nombre d'essais pour ouvrir la porte ».

1.

$$\begin{aligned}
P(X = 1) &= \frac{1}{4} \\
P(X = 2) &= \frac{3}{4} \cdot \frac{1}{3} = \frac{1}{4} \\
P(X = 3) &= \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{4} \\
P(X = 4) &= \frac{3}{4} \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{4}.
\end{aligned}$$

2. On calcule l'espérance et la variance de X

$$E(X) = 1 \cdot \frac{1}{4} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4} = \frac{5}{2}$$

$$V(X) = E(X^2) - E^2(X) = 1 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4} + 9 \cdot \frac{1}{4} + 16 \cdot \frac{1}{4} - \frac{25}{4} = \frac{5}{4}.$$

2.4

Soit X la variable aléatoire « classe d'appartenance ».

1. Pour calculer les probabilités $P(X = k)$, $k = 1, \ldots, 5$, on prend les fré-

quences mesurées.

$X = k$	1	2	3	4	5
$P(X = k)$	0,18	0,44	0,27	0,09	0,02
$F(x) = P(X < k)$	0	0,18	0,62	0,89	0,98

2.

$$\begin{aligned} P(2 < X \leq 4) &= P(X < 5) - P(X < 3) = 0{,}98 - 0{,}62 = 0{,}38. \\ P(X > 4) &= P(X = 5) = 0{,}02. \end{aligned}$$

3.

$$E(X) = \sum_{k=1}^{5} k \cdot P(X = k) = 2{,}33.$$

$$V(X) = E(X^2) - E^2(X) = \sum_{k=1}^{5} k^2 \cdot P(X = k) - 2{,}33^2 \simeq 0{,}88.$$

2.5

Soit X la variable aléatoire « nombre de chiffres transmis de façon incorrecte », qui suit une loi binomiale de paramètres (5, 0,2). La probabilité d'avoir une erreur est égale à la probabilité de recevoir au moins 3 chiffres incorrects. Ainsi

$$P(\text{erreur}) = P(X \geq 3) = C_3^5 0{,}2^3 \cdot 0{,}8^2 + C_4^5 0{,}2^4 \cdot 0{,}8 + C_5^5 0{,}2^5 \simeq 0{,}058.$$

2.6

Soit X la variable aléatoire « nombre de pommes avariées dans un emballage ».

1. La variable aléatoire X suit une loi binomiale de paramètres (5, 0,25).
2. Un client se plaint s'il trouve au moins 2 pommes avariées, donc

$$P(X \geq 2) = 1 - P(X < 2) = 1 - C_0^5 0{,}75^5 - C_1^5 0{,}25 \cdot 0{,}75^4 = \frac{47}{128} \simeq 0{,}37.$$

3. Soit Y la variable aléatoire « nombre de clients non satisfaits ». Elle suit une loi binomiale de paramètres (100, 47/128). On en cherche l'espérance :

$$E(Y) = 100 \cdot \frac{47}{128} \simeq 36{,}7.$$

Sur les 100 clients, environ 37 viendront se plaindre en moyenne.

2.7

Gianni se dit : « J'ai une chance sur 10 000 d'avoir un accident sur 1 km. Comme j'en parcours 10 000, j'ai une probabilité de $10\,000 \cdot \frac{1}{10\,000} = 1$ d'avoir un accident. »
Posons le problème : appelons X la variable aléatoire « nombre d'accidents sur 10 000 km ». Elle suit une loi binomiale de paramètres (10 000, 0,0001) qu'on peut approximer par une loi de Poisson de paramètre $(10\,000 \cdot \frac{1}{10\,000} = 1)$. On cherche la probabilité d'avoir au moins un accident, soit

$$P(X \geq 1) = 1 - P(X = 0) \simeq 1 - e^{-1} \simeq 0{,}63.$$

Gianni n'a « que » 63 % de chance d'avoir un accident sur 10 000 km.

2.8

Soit X la variable aléatoire « nombre de personnes qui acceptent l'interview ». Lorsqu'il y a 5 noms sur la liste, la variable aléatoire X suit une loi binomiale de paramètres $(5, 2/3)$. Donc

$$P(X = 5) = C_5^5 \left(\frac{2}{3}\right)^5 \simeq 0{,}13.$$

S'il y a 8 noms sur la liste, la variable aléatoire X suit une loi binomiale de paramètres $(8, 2/3)$ et la probabilité de pouvoir réaliser au moins 5 interviews est

$$\begin{aligned} P(X \geq 5) &= \\ &= C_5^8 \left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^3 + C_6^8 \left(\frac{2}{3}\right)^6 \left(\frac{1}{3}\right)^2 + C_7^8 \left(\frac{2}{3}\right)^7 \left(\frac{1}{3}\right) + \left(\frac{2}{3}\right)^8 \simeq 0{,}74. \end{aligned}$$

2.9

Soit X (respectivement Y) la variable aléatoire « nombre de passagers qui ont réservé une place avec la compagnie SA Airlines (respectivement BA Airlines) et n'effectuent pas le vol ». Les variables aléatoires X et Y suivent des lois binomiales de paramètres (94, 0,05) et (188, 0,05). On les approxime par des distributions de Poisson de paramètre $94 \cdot 0{,}05 = 4{,}7$ et $188 \cdot 0{,}05 = 9{,}4$ respectivement. Un passager qui a réservé un siège ne pourra pas prendre place dans un avion SA Airlines si moins de 4 personnes ayant réservé n'effectuent pas le vol (moins de 8 personnes pour la compagnie BA Airlines). Par conséquent, on

calcule

$$\begin{aligned} P(X \leq 4) &= \sum_{i=0}^{4} P(X = i) = \sum_{i=0}^{4} e^{-4,7} \frac{4,7^i}{i!} \simeq 0,49 \\ P(Y \leq 8) &= \sum_{i=0}^{8} P(Y = i) = \sum_{i=0}^{8} e^{-9,4} \frac{9,4^i}{i!} \simeq 0,40, \end{aligned}$$

On en conclut que c'est avec la compagnie BA Airlines qu'un passager ayant réservé risque le plus de ne pas pouvoir prendre l'avion.

2.10

Soit X (respectivement Y et Z) la variable aléatoire « nombre de raisins (respectivement pépites et morceaux [raisins + pépites]) dans un biscuit ».

1. La variable aléatoire X suit une loi binomiale de paramètres (600, 1/500) que l'on approxime par une loi de Poisson de paramètre $1/500 \cdot 600 = 1,2$. La probabilité qu'il n'y ait pas de raisin dans le biscuit est

$$P(X = 0) = e^{-1,2} \frac{1,2^0}{0!} \simeq 0,30.$$

2. Selon le même raisonnement, la variable aléatoire Y suit une distribution de Poisson de paramètre $1/500 \cdot 400 = 0,8$. Ainsi, la probabilité qu'il y ait exactement 2 pépites de chocolat est

$$P(Y = 2) = e^{-0,8} \frac{0,8^2}{2!} \simeq 0,14.$$

3. Finalement, la variable aléatoire Z suit elle aussi une loi de Poisson, de paramètre $1/500 \cdot 1\ 000 = 2$ et la probabilité de trouver au moins 2 morceaux dans le biscuit est

$$P(Z \geq 2) = 1 - P(Z < 2) = 1 - e^{-2} \frac{2^0}{0!} - e^{-2} \frac{2^1}{1!} \simeq 0,59.$$

2.11

1. Soit X la variable aléatoire « nombre de champignons toxiques », qui suit une loi binomiale de paramètres (6, 0,7).
   (a) La probabilité que Nadine ramasse exactement 4 champignons toxiques est
   $$P(X = 4) = C_4^6\ 0,7^4 \cdot 0,3^2 \simeq 0,32.$$

(b) On cherche la probabilité que tous les champignons ramassés soient comestibles. Les probabilités que Nadine et Serge ramassent des champignons comestibles sont indépendantes. Ainsi

$$\begin{aligned} &P(\text{tous comestibles}) = \\ &\quad = P(\text{Nadine ramasse 3 comestibles} \cap \text{Serge 4 comestibles}) \\ &\quad = 0{,}3^3 \cdot 0{,}9^4 \simeq 0{,}018. \end{aligned}$$

2. (a) Soit Y la variable aléatoire « nombre de champignons ramassés en une heure ». Elle suit une loi de Poisson de paramètre 12.

$$P(Y = 8) = \frac{12^8}{8!} \cdot e^{-12} \simeq 0{,}066.$$

(b) Soit Z la variable aléatoire « nombre de champignons ramassés en 20 minutes », qui suit une loi de Poisson de paramètre $\frac{12}{3} = 4$.

$$P(Z \geq 1) = 1 - P(Z = 0) = 1 - \frac{4^0}{0!} \cdot e^{-4} \simeq 0{,}98.$$

2.12

Appelons stratégie I (respectivement II) la stratégie qui consiste à utiliser 4 haut-parleurs de 4 000 W (respectivement 2 de 8 000 W). Soit V_I (respectivement V_{II}) l'événement « le concert peut se terminer avec la stratégie I (respectivement II) ». Soit X_I (respectivement X_{II}) la variable aléatoire « nombre de haut-parleurs en panne si on se trouve dans la stratégie I (respectivement II) ». Les variables aléatoires X_I et X_{II} suivent chacune une loi binomiale, de paramètre (4, 0,2) et (2, 0,2) respectivement. On veut choisir la stratégie avec la plus grande probabilité de voir le concert se terminer.

$$\begin{aligned} P(V_I) &= P(X_I \leq 2) = 1 - P(X_I = 3) - P(X_I = 4) = \\ &= 1 - \binom{4}{3} 0{,}2^3 \cdot 0{,}8 - \binom{4}{4} 0{,}2^4 \simeq 0{,}97 \end{aligned}$$

et

$$P(V_{II}) = P(X_{II} \leq 1) = 1 - P(X_{II} = 2) = 1 - \binom{2}{2} 0{,}2^2 = 0{,}96.$$

On choisit donc la stratégie I.

2.13

Pour $i = 1, \ldots, 20$, soit X_i une variable aléatoire telle que

$$X_i = \begin{cases} 1 & \text{si le } i^{\text{e}} \text{ canard survit ;} \\ 0 & \text{sinon.} \end{cases}$$

Un canard survit si les 10 chasseurs le manquent, donc

$$P(X_i = 1) = \left(1 - \frac{p}{20}\right)^{10}.$$

1. Soit Y la variable aléatoire « nombre de canards épargnés ». On en cherche l'espérance

$$\begin{aligned} E(Y) &= E\left(\sum_{i=1}^{20} X_i\right) = \sum_{i=1}^{20} (1 \cdot P(X_i = 1) + 0 \cdot P(X_i = 0)) = \\ &= 20\left(1 - \frac{p}{20}\right)^{10}. \end{aligned}$$

2. Soient N la variable aléatoire « nombre de canards » et Z la variable aléatoire « nombre de canards touchés ». On cherche l'espérance de Z. La difficulté du calcul provient du fait que Z dépend de la variable aléatoire N. On cherche

$$E(Z) = \sum_{k=0}^{\infty} kP(Z = k).$$

On ne connaît la probabilité $P(Z = k)$ que si N est fixé. On utilise donc la formule des probabilités totales

$$P(Z = k) = \sum_{i=1}^{\infty} P(Z = k \mid N = i)P(N = i)$$

pour trouver

$$\begin{aligned} E(Z) &= \sum_{k=0}^{\infty} k \sum_{i=1}^{\infty} P(Z = k \mid N = i)P(N = i) \\ &= \sum_{i=1}^{\infty} \left[\sum_{k=0}^{\infty} kP(Z = k \mid N = i)\right] P(N = i) \\ &= \sum_{i=1}^{\infty} E(Z \mid N = i)P(N = i). \end{aligned}$$

En utilisant le résultat de 1., l'espérance devient

$$E(Z) = \sum_{i=1}^{\infty} i\left(1 - \left(1 - \frac{p}{i}\right)^{10}\right) P(N = i).$$

Le tableau 2.1 résume les résultats de l'exercice pour différentes valeurs de la probabilité p.

Probabilité p	0,1	0,3	0,5	0,7	0,9
Vol de 20 canards	0,98	2,81	4,47	6	7,38
$N \sim P(15)$	0,97	2,73	4,27	5,62	6,81

Tableau 2.1 – Tableau d'espérances de l'exercice 2.13.

2.14

1. Soit X la variable aléatoire « nombre de soldats malades ». La variable aléatoire X suit une loi binomiale de paramètres (500, 0,001) et la probabilité que les test soit positif revient à calculer

$$P(X > 0) = 1 - P(X = 0) = 1 - 0{,}999^{500} \simeq 0{,}394.$$

   Il faut noter qu'il aurait été possible d'approximer la distribution de X par une loi de Poisson d'espérance 0,5 pour obtenir le même résultat.
2. On cherche à présent la probabilité qu'il y ait plus d'un malade alors que le test est positif et on obtient

$$\begin{aligned} P(X > 1 \mid X > 0) &= \frac{P(X > 1 \text{ et } X > 0)}{P(X > 0)} = \frac{P(X > 1)}{P(X > 0)} = \\ &= \frac{1 - P(X = 0) - P(X = 1)}{1 - P(X = 0)} \\ &\simeq \frac{0{,}394 - \binom{500}{1} \cdot 0{,}001 \cdot 0{,}999^{499}}{0{,}394} \simeq 0{,}230. \end{aligned}$$

2.15

Soit X la variable aléatoire « nombre de blocages de mâchoire par mois ». La variable aléatoire X suit une loi binomiale de paramètres (4 000, 0,001) qu'on approxime par une loi de Poisson de paramètre $4\,000 \cdot 0{,}001 = 4$.

1. La probabilité qu'il y ait au moins 3 blocages de mâchoires en 1 mois est

$$P(X \geq 3) = 1 - \sum_{k=0}^{2} P(X = k) = 1 - \sum_{k=0}^{2} e^{-4} \frac{4^k}{k!} \simeq 0{,}76.$$

2. Soit Y la variable aléatoire « nombre de mois comptant 3 blocages de mâchoire ou plus ». Cette variable suit une loi binomiale de paramètres (8, p), p étant le résultat du calcul précédent ($p = 0{,}76$). Ainsi

$$\begin{aligned} P(Y \geq 4) &= 1 - P(Y < 4) = 1 - \sum_{k=0}^{3} P(Y = k) = \\ &= 1 - \sum_{k=0}^{3} \binom{8}{k} p^k (1-p)^{8-k} \simeq 0{,}98. \end{aligned}$$

3. Soit N la variable aléatoire « numéro du premier mois qui voit au moins 3 blocages de mâchoire ». N suit une loi géométrique avec probabilité p (calculée en 1.). Par conséquent,

$$P(N = i) = (1 - p)^{i-1} \cdot p, \quad \text{avec } p = 0{,}76, \quad i = 1, 2, \ldots, 8,$$

et on trouve les probabilités du tableau 2.2.

i	1	2	3	4	5 . . . 8
$P(N = i)$	0,76	0,18	0,04	0,01	0,00

Tableau 2.2 – Tableau des probabilités de l'exercice 2.15.

2.16

Soient X la variable aléatoire « nombre de dés qui montrent le chiffre misé » et G la variable aléatoire « gain réel du joueur ». La variable aléatoire X suit une loi binomiale de paramètres (3, 1/6). La variable aléatoire G, qui dépend de X, peut prendre les valeurs -1, 1, 2 et 3, suivant qu'aucun, 1, 2 ou 3 dés montrent le résultat pronostiqué.

1. Ainsi, l'espérance du gain d'un joueur est

$$\begin{aligned} E(G(X)) &= \sum_{k=0}^{3} G(X = k)P(X = k) \\ &= -1 \cdot C_0^3 \left(\frac{5}{6}\right)^3 + 1 \cdot C_1^3 \frac{1}{6} \left(\frac{5}{6}\right)^2 + 2 \cdot C_2^3 \left(\frac{1}{6}\right)^2 \frac{5}{6} + 3 \cdot C_3^3 \left(\frac{1}{6}\right)^3 \\ &= -\frac{17}{216} \simeq -0{,}079£. \end{aligned}$$

2. Soit y le montant que l'on reçoit si les 3 dés montrent le chiffre prévu. Si on veut un jeu *fair*, il faut que

$$E(G(X)) = \frac{1}{216}(-1 \cdot 5 \cdot 5 \cdot 5 + 1 \cdot 3 \cdot 5 \cdot 5 + 2 \cdot 3 \cdot 5 + y \cdot 1) = 0,$$

donc

$$y = 20£.$$

2.17

Soit $N(t)$ la variable aléatoire « nombre d'arrivées dans l'intervalle t ». Elle suit une loi de Poisson de paramètre λt. Soit T_k la variable aléatoire « temps

avant l'arrivée de la k^{e} personne ».

1. On cherche la loi de T_1. Si aucune personne ne se présente au guichet avant l'instant t, cela signifie que $T_1 > t$. Ainsi

$$P(T_1 > t) = P(N(t) = 0)$$

et

$$F_{T_1}(t) = P(T_1 < t) = 1 - P(T_1 > t) = 1 - P(N(t) = 0) = 1 - e^{-\lambda t}.$$

On en déduit que T_1 suit une loi exponentielle de paramètre λ.

2. On cherche à présent la fonction de répartition de T_k. Si $T_k < t$, cela signifie que la k^{e} personne est arrivée avant l'instant t, ou que $N(t) \geq k$. Donc

$$F_{T_k} = P(T_k < t) = P(N(t) \geq k) = \sum_{j=k}^{\infty} P(N(t) = j) = \sum_{j=k}^{\infty} e^{-\lambda t} \frac{(\lambda t)^j}{j!}.$$

3. La fonction de densité de T_k est, par conséquent

$$\begin{aligned} f_{T_k}(t) &= \frac{\partial}{\partial t} F_{T_k} = \sum_{j=k}^{\infty} \left(-\lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} + j\lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{j!} \right) \\ &= \sum_{j=k}^{\infty} \left(\lambda e^{-\lambda t} \frac{(\lambda t)^{j-1}}{(j-1)!} - \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} \right) \\ &= \lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!} = \frac{\lambda^k}{\Gamma(k)} t^{k-1} e^{-\lambda t}. \end{aligned}$$

T_k suit une loi Gamma de paramètre $(n,\, k)$.

2.18

Soit X la variable aléatoire « nombre de skieurs qui se présentent entre 2 passages d'Andi ». La variable aléatoire X suit une loi géométrique de paramètre p. Andi reprend la même perche au passage suivant si $kN - 1$ skieurs se présentent entre ses 2 passages, avec $k = 1, 2, \ldots$ La probabilité qu'il reprenne la même perche est donc

$$\begin{aligned} P(\text{même perche}) &= \sum_{k=1}^{\infty} P(X = kN - 1) = \sum_{k=1}^{\infty} q^{(kN-1)-1} \cdot p = \frac{p}{q^2} \sum_{k=1}^{\infty} q^{kN} \\ &= \frac{p}{q^2} \sum_{k=1}^{\infty} \left(q^N\right)^k = \frac{p}{q^2} \left(\sum_{k=0}^{\infty} \left(q^N\right)^k - \left(q^N\right)^0 \right), \end{aligned}$$

où $q = 1 - p$. La dernière somme est une série géométrique de raison q^N. Par conséquent

$$\begin{aligned} P(\text{même perche}) &= \frac{p}{q^2}\left(\sum_{k=0}^{\infty}\left(q^N\right)^k - \left(q^N\right)^0\right) \\ &= \frac{p}{q^2}\left(\frac{1}{1-q^N} - 1\right) = \frac{p}{q^2}\frac{q^N}{1-q^N}. \end{aligned}$$

2.19

Soit X la variable aléatoire « nombre d'épreuves nécessaires pour obtenir le premier succès ».

1. $P(X = k) = P(k-1 \text{ échecs et un succès}) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \ldots$
2. En utilisant $0 < q = 1 - p < 1$, on a

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} k(p-1)^{k-1}p = p\sum_{k=1}^{\infty} kq^{k-1} = p\sum_{k=1}^{\infty}\left(\frac{d}{dq}q^k\right) \\ &= p\frac{d}{dq}\left(\sum_{k=1}^{\infty} q^k\right) = p\frac{d}{dq}\frac{q}{1-q} = p\frac{1}{(1-q)^2} = \frac{1}{p}. \end{aligned}$$

3.

$$\begin{aligned} P(X > k \mid X > j) &= \frac{P(X > k \cap X > j)}{P(X > j)} = \frac{P(X > k)}{P(X > j)} \\ &= \frac{(1-p)^k}{(1-p)^j} = (1-p)^{k-j} = P(X > k - j), \end{aligned}$$

où l'on a utilisé $P(X > k) = (1-p)^k$.

4. Soient Y la variable aléatoire « valeur de l'offre d'achat » avec fonction de répartition F et N la variable aléatoire « nombre d'offres d'achat reçues avant de vendre la maison ». Ainsi

$$P(N = n) = [P(Y < k)]^{n-1}\, P(Y > k) = F(k)^{n-1}(1 - F(k)).$$

2.20

On a N boules ; r sont rouges et $N - r$ sont noires. Soit X la variable aléatoire « nombre de boules rouges tirées ». On procède au tirage de n boules sans remise.

1.

$$P(X = k) = \frac{C_k^r \cdot C_{n-k}^{N-r}}{C_n^N} = \frac{\begin{pmatrix} r \\ k \end{pmatrix}\begin{pmatrix} N-r \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}, \quad k = 0, 1, \ldots, \min(n,r).$$

2. On veut comparer l'espérance et la variance de la loi hypergéométrique avec celles de la loi binomiale. Pour rappel, les tirages d'une loi binomiale sont avec remise. Soit Y une variable aléatoire de distribution binomiale avec paramètres (n, p). On sait que $E(Y) = np$ et $V(Y) = np(1-p)$. Ainsi

$$\begin{aligned} E(X) &= E(Y) \\ V(X) &= \frac{N-n}{N-1} V(Y) \leq V(Y). \end{aligned}$$

Si N est beaucoup plus grand que n, le rapport $(N-n)/(N-1)$ tend vers 1 et $V(X)$ vers $V(Y)$. En d'autres termes, en augmentant considérablement la quantité de boules dans l'urne, un tirage ne modifie presque plus la probabilité de tirer une boule rouge et on peut faire l'approximation que le tirage est sans remise.

Chapitre 3

Variables aléatoires continues

Introduction

Le passage des modèles discrets aux modèles continus intervient dans beaucoup de domaines de la science. Dans notre cas, cela consiste à considérer un passage à la limite d'une loi de probabilité discrète quand l'ensemble des valeurs possibles de la variable aléatoire augmente et devient toujours plus dense. Ceci conduit à caractériser une variable aléatoire continue par sa fonction de densité, une fonction non-négative et telle que l'intégrale (la surface totale) au-dessous de cette courbe soit égale à 1. La probabilité est représentée par la surface au-dessous de la courbe correspondant à un intervalle donné

$$P(X \in [a, b]) = \int_a^b f(x)dx \ ,$$

où $f(.)$ est la fonction de densité. Une variable aléatoire continue peut aussi être caractérisée par sa fonction de répartition $F(x) = P(X < x)$, par sa fonction de survie $S(x) = P(X > x)$ ou par sa fonction de risque $\lambda(x) = \frac{f(x)}{S(x)} = -\frac{d}{dx} \log S(x)$.

La première partie de ce chapitre présente des exercices sur ces outils de base (exercices 3.1 à 3.10). Comme dans le cas discret, on a ensuite recueilli les exercices concernant quelques lois continues de probabilité. On traite en particulier la loi de base la plus simple, la loi uniforme, ainsi que la loi la plus importante, la loi normale (exercices 3.11 à 3.17).

Des exercices concernant la loi d'une transformation d'une variable aléatoire continue (exercices 3.18 à 3.27) ainsi que d'autres aspects concernant les lois continues (exercices 3.28 à 3.39) concluent ce chapitre.

La loi uniforme

La loi uniforme est la loi continue la plus simple. Une variable aléatoire X suit une loi uniforme sur l'intervalle $[a, b]$, $X \sim \mathcal{U}(a, b)$, si sa fonction de densité est égale à $1/(b-a)$ sur l'intervalle $[a, b]$ et 0 ailleurs. La loi $\mathcal{U}(0,1)$ joue un rôle particulier car des algorithmes permettent la génération efficace par ordinateur de réalisations (« nombres aléatoires ») d'une variable aléatoire qui suit cette loi. Cela permet ensuite de générer des réalisations de variables aléatoires qui suivent d'autres lois discrètes ou continues (exercices 3.27, 3.31, 3.37 et 3.38). Ceci est à la base des techniques de simulation de processus aléatoires par ordinateur, un outil indispensable pour la modélisation statistique moderne.

La loi normale

La loi normale a ses racines et sa justification dans le théorème central limite (voir l'introduction du chapitre 5). La fonction de densité d'une variable aléatoire normale standard, $X \sim \mathcal{N}(0, 1)$, est donnée par

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}x^2\}.$$

On a $E(X) = 0$ et $\text{var}(X) = 1$. Notons comme curiosité que l'importance de cette loi se reflète aussi dans le fait que sa densité $\varphi(x)$ contient trois concepts fondamentaux en mathématiques, les nombres $\sqrt{2}, \pi, e$.

Notes historiques

En ce qui concerne la loi normale, voir l'introduction du chapitre 5.

Références (théorie)

Ross, chapitres 5 et 7 [1] et Pitman, chapitre 4 [2].

Exercices

Moments et probabilités

3.1

Soit X une variable aléatoire continue de fonction de densité $f(x)$.

1. Démontrer que $\text{var}(X) = E(X^2) - (E(X))^2$.
2. Démontrer que $\text{var}(aX) = a^2\text{var}(X)$.

3.2

Soit X une variable aléatoire dont la fonction de densité est

$$f(x) = \begin{cases} c(1-x^2) & -1 < x < 1 \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer la valeur de c.
2. Quelle est la fonction de répartition de X ?
3. Calculer $E(X)$.

3.3

Soit X une variable aléatoire dont la fonction de densité est

$$f(x) = \begin{cases} \sin(x) & 0 < x < \frac{\pi}{2} \\ 0 & \text{sinon.} \end{cases}$$

1. Quelle est la fonction de répartition de X ?
2. Calculer $E(X)$.
3. Calculer $\text{Var}(X)$.

3.4

On dit que la variable aléatoire X suit la loi exponentielle de paramètre λ, notée $\varepsilon(\lambda)$, si sa densité est

$$f(x) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer la fonction de répartition $F(x)$.
2. Que vaut le α-quantile $q = F^{-1}(\alpha)$?

3.5

Soit X une variable aléatoire de densité

$$f(x) = \begin{cases} c(a)(a^2 - x^2) & \text{si } |x| \leq a \\ 0 & \text{sinon.} \end{cases}$$

1. Déterminer la constante de normalisation $c(a)$.
2. Donner la fonction de répartition de X.
3. Calculer $E(X)$ et $\text{var}(X)$.

3.6

Soit $f(x)$ une fonction de densité définie sur $\mathcal{R}$. On définit

$$f_{\epsilon,b}(x) = (1-\epsilon)f(x) + \frac{\epsilon}{b} f\Big(\frac{x}{b}\Big),$$

où $0 < \epsilon < 0{,}05$ et $b > 1$, une perturbation de $f(x)$.

1. Montrer que $f_{\epsilon,b}(x)$ est une fonction de densité.
2. Soit $f(x)$ la densité d'une distribution normale avec espérance 0 et variance 1. Calculer $E(X)$ et $\text{var}(X)$, où $X \sim f_{\epsilon,b}(x)$.
   Indication : pour la dérivation de la variance, calculer $E(X^2)$.
3. Soit $b = 3$ et $\epsilon = 0{,}05$. Interpréter $f_{\epsilon,b}(x)$ du point 2. et comparer avec $f(x)$.

3.7

Tous les jours, Sébastien parcourt le même trajet de 40 km pour se rendre à son travail. Sa vitesse est une variable aléatoire V qui dépend des conditions météorologiques et de la circulation. Sa densité est de la forme

$$f_V(v) = \begin{cases} Cv\exp(-\lambda v) & v \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

Sébastien roule à une vitesse moyenne de 80 km/h.

1. Déterminer les valeurs de C et de λ.
2. La durée du trajet est décrite par la variable $T = \frac{40}{V}$. Déterminer la densité et l'espérance de T.

3.8

La fonction de densité de X, variable aléatoire représentant la durée de vie en heures d'un certain composant électronique, est donnée par

$$f(x) = \begin{cases} 10/x^2 & x > 10 \\ 0 & x \leq 10. \end{cases}$$

1. Trouver $P(X > 20)$.
2. Quelle est la fonction de répartition de X ?
3. Quelle est la probabilité que parmi 6 composants, au moins 3 d'entre eux fonctionnent au moins 15 heures? Quelles hypothèses faites-vous?
4. Calculer $E(X)$.

3.9

La quantité de pain (en centaines de kilos) qu'une boulangerie vend en 1 journée est une variable aléatoire X de fonction de densité

$$f(x) = \begin{cases} cx & 0 \leq x \leq 3 \\ c(6-x) & 3 \leq x \leq 6 \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer la valeur de c.
2. Quelle est la fonction de répartition de X ?
3. Soit A l'événement : « le nombre de kilos de pain vendus dans une journée est supérieur à 300 kg ». Soit B l'événement : « le nombre de kilos de pain vendus dans une journée est compris entre 150 et 450 kg ». Les événements sont-ils indépendants?

3.10

La durée de vie X en années d'une télévision suit une loi exponentielle de densité

$$f(x) = \frac{1}{8}e^{-\frac{x}{8}} \quad x \geq 0.$$

1. Calculer la probabilité que la télévision que vous venez d'acheter ait une durée de vie supérieure à 8 ans.
2. Vous possédez une telle télévision depuis 2 ans. Quelle est la probabilité que sa durée de vie soit encore de 8 ans à partir de maintenant ? Conclusion.
3. Quelle est la durée de vie moyenne $E(X)$ d'une télévision? Et la variance de cette durée de vie?

Loi normale

3.11

La longueur des pièces produites par une machine de type A varie selon une loi normale avec espérance 8 mm et variance 4 mm, et la longueur de celles

produites par une machine de type B varie selon une loi normale avec espérance 7,5 mm et variance 1 mm.

1. Si vous voulez produire des pièces de longueurs 8 ± 1 mm, quel type de machine choisiriez-vous ?
2. Si la moyenne des longueurs produites par la machine A reste 8 mm, quelle doit être sa variance pour qu'elle ait la même performance que la machine B ?

3.12

On suppose que la taille, en centimètres, d'un pygmée âgé de 25 ans est une variable aléatoire normale de paramètres $\mu = 140$ et $\sigma = 6$.

1. Quel est le pourcentage de pygmées de 25 ans ayant une taille supérieure à 150 cm ?
2. Parmi les pygmées mesurant plus de 145 cm, quel pourcentage dépasse 150 cm ?

3.13

On suppose que la taille, en centimètres, d'un homme âgé de 30 ans est une variable aléatoire normale de paramètres $\mu = 175$ et $\sigma^2 = 36$.

1. Quel est le pourcentage d'hommes de 30 ans ayant une taille supérieure à 185 cm ?
2. Parmi les hommes mesurant plus de 180 cm, quel pourcentage dépasse 192 cm ?

3.14

Un entrepreneur doit estimer le temps nécessaire à l'exécution d'un travail. Les incertitudes dues au marché du travail, à l'approvisionnement en matériaux, aux mauvaises conditions atmosphériques ... constituent une inconnue. Néanmoins, il affirme qu'il a une probabilité de 10 % de réaliser le travail en plus de 190 jours et une probabilité de 5 % que le travail soit terminé en moins de 50 jours. Soit X la variable aléatoire, supposée normale, désignant le nombre de jours nécessaires à l'exécution du travail.

1. Donner l'espérance et la variance de X.
2. Que vaut la probabilité que la durée du travail dépasse 200 jours ?

3.15

Le samedi soir, la police fait un alcootest à tous les conducteurs qui passent par une route principale. Quelle est la proportion d'automobilistes recevant une

amende (taux d'alcool > 0,08 %) si l'on suppose que le taux d'alcool chez les automobilistes est distribué selon une loi normale d'espérance $\mu = 0{,}07$ % et d'écart-type $\sigma = 0{,}01$ %? En plus de l'amende, les conducteurs ayant plus de 0,09 % ont un retrait de permis. Parmi les automobilistes réprimandés, quelle est la proportion de retraits de permis?

3.16

Lors d'un procès en attribution de paternité, un expert témoigne que la durée de la grossesse, en jours, est de distribution approximativement normale avec paramètres $\mu = 270$ et $\sigma^2 = 100$. L'un des pères putatifs est en mesure de prouver son absence du pays pendant une période s'étendant entre le 290^{e} et le 240^{e} jour précédent l'accouchement.
Quelle est la probabilité que la conception de l'enfant ait eu lieu plus de 290 jours avant sa naissance ou moins de 240 jours avant?

3.17

Sur une route principale où la vitesse est limitée à 80 km/h, un radar a mesuré la vitesse de toutes les automobiles pendant une journée. En supposant que les vitesses recueillies soient distribuées selon une loi normale avec une moyenne de 72 km/h et un écart-type de 8 km/h, répondez aux questions suivantes.

1. Quelle est la proportion de conducteurs qui devront payer une amende pour excès de vitesse?
2. Sachant qu'en plus de l'amende, un excès de plus de 30 km/h implique un retrait de permis, quelle est la proportion des conducteurs qui vont se faire retirer le permis parmi ceux qui vont avoir une amende?

Transformations de variables

3.18

Soit $X \sim \mathcal{N}(0{,}1)$ et définissons $Y = X^2$.
Trouver la fonction de répartition de Y et sa densité. Représenter graphiquement cette densité.
Cette distribution est appelée une loi de χ^2_1 (chi carré à un degré de liberté).

3.19

Soit $V \sim \mathcal{U}(0{,}1)$. Calculer la fonction de répartition de $W = \frac{-1}{\lambda}\ln(V)$ et sa densité. De quelle loi s'agit-il?

3.20

Soit X une variable aléatoire distribuée suivant la loi de Weibull de paramètres $\alpha > 0$ et $\beta > 0$. La densité de X est donc définie par la fonction

$$f(x) = \begin{cases} \left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta-1}\exp[-(\frac{x}{\alpha})^{\beta}] & \text{si } x > 0 \\ 0 & \text{sinon.} \end{cases}$$

Soit $Y = (X/\alpha)^{\beta}$. Calculer la fonction de densité de Y. De quelle loi s'agit-il?

3.21

Soit $X \sim \chi^2_1$ et définissons $Y = \ln(X)$. Trouver la fonction de répartition de Y et sa densité.

3.22

Supposons qu'une administration fiscale, dans la déclaration d'impôt d'un indépendant, ne prenne en compte des déductions professionnelles autorisées et justifiées qu'à partir de 5 000 €. Soit X (exprimé en milliers d'euros) le total des déductions d'une déclaration choisie au hasard. La variable aléatoire X est supposée suivre une loi dont la densité est $f_X(x) = \frac{k}{x^{\alpha}}$ si $x \geq 5$.

1. Déterminer la valeur de k. Quelle restriction sur α doit-on faire afin que $f(x)$ soit une densité?
2. Déterminer la fonction de répartition $F_X(x)$ de X.
3. Montrer que $Y = \ln(X/5)$ suit une loi exponentielle de paramètre $(\alpha - 1)$.

3.23

Une variable aléatoire X suit une distribution de Pareto de densité

$$f_{\alpha}(x) = \begin{cases} \alpha x_0^{\alpha} x^{-(1+\alpha)} & \text{si } x > x_0 \\ 0 & \text{sinon.} \end{cases}$$

avec $x_0 > 0$ et $\alpha > 1$.
Trouver la densité de la variable aléatoire $Y = \ln(X)$.

3.24

Soit V une variable aléatoire distribuée suivant la loi de Cauchy de densité

$$f(v) \quad = \quad \frac{1}{\pi}\frac{1}{(1+v^2)}, \quad v \in \mathcal{R}.$$

Soit $Z = 1/V$. Calculer la densité de Z. Que remarquez-vous?

3.25

Dans un atelier de céramique on produit des petites plaques carrées de largeur L. Cette valeur L est une variable aléatoire dont la fonction de densité est définie de la manière suivante

$$f_L(x) = \begin{cases} 6x(1-x) & 0 < x < 1 \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer la fonction de densité de la surface S d'une plaque.
2. Calculer l'espérance de la variable aléatoire S.

3.26

Dans une usine, on fabrique des boules pour la décoration de sapins de Noël. Le rayon de ces boules est une variable aléatoire R qui a pour fonction de densité

$$f_R(r) = \begin{cases} \frac{2}{9}r(3-r) & 0 \leq r \leq 3 \\ 0 & \text{sinon.} \end{cases}$$

1. L'usine s'intéresse à la quantité de matière nécessaire pour fabriquer ces boules et demande de calculer la fonction de densité de la surface $S = 4\pi R^2$ d'une boule.
2. Calculer l'espérance de la variable aléatoire S.

3.27

Soit X une variable aléatoire continue ayant une fonction de répartition F. On définit la variable aléatoire Y par $Y = F(X)$.
Montrer que Y est uniformément distribuée sur l'intervalle $(0, 1)$.

Exercices combinés

3.28

Les pièces d'une voiture sont souvent copiées et vendues comme pièces originales. On veut remplacer certaines pièces d'une voiture. Avec probabilité $1/4$, on achète une pièce piratée et avec probabilité $3/4$ on achète une pièce originale. La durée de vie est une variable aléatoire exponentielle avec espérance 2 pour une pièce piratée et avec espérance 5 pour une pièce originale. Appelons T la durée de vie de la pièce que l'on achète. Supposons que la pièce ait survécu jusqu'au temps t après son installation. Quelle est la probabilité $\pi(t)$ que cette pièce soit piratée? Trouver la limite de $\pi(t)$ lorsque $t \to \infty$.

3.29

Un point est choisi au hasard sur un segment de longueur L.
Trouver la probabilité que le rapport entre le plus petit et le plus grand segment soit inférieur à $1/4$.

3.30

Afin de définir une politique sociale à long terme, on désire étudier l'évolution du sida. On s'intéresse alors à modéliser la fonction de survie des malades. Notons $S(t) = 1 - F(t)$ la fonction de survie et $h(t) = -\frac{\partial \log S(t)}{\partial t}$ la fonction de risque *(hazard function)*.

1. Dans un 1$^{\text{er}}$ temps on va supposer que le risque est constant, c'est-à-dire que $h(t) = \lambda$. Trouver la fonction de densité des données. De quelle loi s'agit-il?
2. En 2$^{\text{e}}$ lieu on supposera que $h(t) = \lambda\kappa(\lambda t)^{\kappa-1}$. Trouver la fonction de densité des données. De quelle loi s'agit-il?

3.31

En analysant les temps de survie des *start-up*, on note que plus la société a été en activité et plus petite est la probabilité qu'elle fasse faillite le mois prochain. Ceci implique une fonction de risque $\lambda(t)$ décroissante avec le temps t. Un modèle utilisé est $\lambda(t) = (a + t)^{-1}$, $t > 0$, où a est un paramètre positif.

1. Calculer la fonction de survie $P(T > t)$, où T représente le temps de survie.
2. Quelle est la médiane de T?
3. En utilisant 2., donner une interprétation du paramètre a.
4. Proposer un algorithme pour simuler à l'ordinateur des temps de survie $t_1, \ldots, t_n$ à partir d'une série $u_1, \ldots, u_n$ de réalisations d'une variable aléatoire uniforme dans l'intervalle $(0, 1)$.

3.32

Un vendeur de journaux achète ses journaux 10 centimes et les revend 15 centimes. Cependant, il ne peut pas se faire rembourser les exemplaires invendus.
La demande journalière X est une variable aléatoire normale de paramètres $\mu = 100$ et $\sigma^2 = 200/3$.

1. Si s est le nombre de journaux qu'il achète, donner l'espérance de son bénéfice en fonction de s et de f_X, la fonction de densité de X.

2. Quel est approximativement le nombre de journaux qu'il doit acheter afin de maximiser l'espérance de son bénéfice?

3.33

Soient n parts budgétaires indépendantes $X_1, \ldots, X_n$ suivant une loi uniforme dans l'intervalle $(0, 1)$. Calculer la fonction de répartition et la densité de $Y = \min(X_1, \ldots, X_n)$.

3.34

Un analyste financier dispose de données historiques $X_1, \ldots, X_n$ de rendement d'une certaine action. On suppose par simplicité que ces observations sont indépendantes avec fonction de répartition $F_X(x)$ et densité $f_X(x)$. Soit $T = \max(X_1, \ldots, X_n)$ la variable aléatoire qui représente le rendement maximal.

1. Montrer que $F_T(t) = (F_X(t))^n$, où $F_T(t)$ est la fonction de répartition de T.
2. Soit
$$f_X(x) = \begin{cases} \frac{2}{\theta^2}x & \text{si } 0 < x < \theta \\ 0 & \text{sinon.} \end{cases}$$
Dans ce cas, calculer $E(T)$.
3. Sous les conditions du point 2., calculer la probabilité que le rendement maximal soit supérieur à un certain seuil a.

3.35

La fluctuation journalière du prix de l'action d'une société donnée, cotée en bourse, est une variable aléatoire d'espérance 0 et de variance σ^2. Cela veut dire que, si Y_n représente le prix de l'action du n^{e} jour

$$Y_n = Y_{n-1} + U_n, \;\; n > 1,$$

où $U_1, U_2, \ldots$ sont des variables aléatoires indépendantes identiquement distribuées d'espérance 0 et de variance σ^2. Supposons que le prix de l'action soit aujourd'hui de 100, c'est-à-dire $Y_1 = y_1 = 100$, et que $\sigma^2 = 1$.

Donner une borne inférieure pour la probabilité que le prix de l'action sera compris entre 95 et 105 dans 10 jours, en utilisant l'inégalité de Chebychev.

3.36

Lorsqu'on s'intéresse aux distributions de revenu, on peut faire un histogramme à partir des données. Pour étudier certaines propriétés, il est utile de disposer d'un modèle se présentant sous la forme d'une densité ajustant l'histogramme. Dans ce contexte, la distribution de Dagum se définit par

$$F_{\beta,\lambda,\delta}(x) = (1 + \lambda x^{-\delta})^{-\beta}$$

où x est positif, $\beta > 0$, $\lambda > 0$ et $\delta > 1$.

1. Quelle est sa densité?
2. On prend les paramètres $\beta = 0{,}71$, $\lambda = 7\ 883\ 000$, $\delta = 3{,}11$.
   Quelle est la médiane, c'est-à-dire x_M tel que $F(x_M) = \frac{1}{2}$?
   L'espérance sera-t-elle égale à la médiane?
   Quels sont les quartiles inférieurs et supérieurs?
   Indication : trouver la réciproque de F. (Pour cela, poser $F(x) = z$ et obtenir $x = G(z)$; G est la réciproque de F.)
3. Pourquoi l'emploi de la loi normale est-il inadapté à l'étude des distributions de revenu?
4. Si pour le même problème on change d'unité monétaire, disons que la nouvelle unité monétaire y vaut $0{,}80x$, quels sont les paramètres β', λ', et δ' de la distribution de Dagum associée à y? En déduire une interprétation pour le paramètre λ.

3.37

Si l'on dispose de 30 réalisations d'une loi uniforme $u_1, \ldots, u_{30}$, comment obtenir 30 réalisations d'une loi de Dagum? (*Indication* : s'inspirer des exercices 3.27 et 3.36.)

3.38

Un tremblement de terre de magnitude M libère une énergie X telle que $M = \log(X)$. Pour des tremblements de terre de magnitude supérieure à 3, on suppose que $(M-3)$ suit une distribution exponentielle avec espérance 2. Pour les tremblements de terre de magnitude supérieure à 3, calculer :

1. $E(M)$ et $Var(M)$;
2. la densité de M ;
3. la densité et la fonction de répartition de X ;
4. la probabilité que la magnitude du plus faible soit supérieure à 4, étant donné 2 tremblements de terre avec magnitudes M_1 et M_2 indépendantes.
5. On souhaite simuler à l'ordinateur les énergies libérées par un tremblement de terre. Construire un algorithme permettant de simuler n énergies

$x_1, \ldots, x_n$ à partir de n réalisations $u_1, \ldots, u_n$ d'une loi uniforme dans l'intervalle $(0, 1)$.

3.39

Soit S_t la valeur d'un actif à la fin de l'année t et $R_{0,n}$ le taux de rendement sur un horizon de n années, c'est-à-dire que $R_{0,n}$ est la solution de l'équation

$$S_n = S_0(1 + R_{0,n})^n.$$

Sous l'hypothèse que $\frac{S_t}{S_{t-1}}$ suit une loi log-normale avec paramètres μ et σ^2, calculer l'espérance et la variance de $R_{0,n}$.

$$\textit{Indication :}\ \frac{S_n}{S_0} = \frac{S_1}{S_0} \cdot \frac{S_2}{S_1} \ldots \frac{S_t}{S_{t-1}} \ldots \frac{S_n}{S_{n-1}} = \prod_{t=1}^{n} \frac{S_t}{S_{t-1}}$$

Que se passe-t-il quand $n \to \infty$?

Ce problème est basé sur l'article « *The Long-Term Expected Rate of Return : Setting it Right* », publié par O. de la Grandville dans le *Financial Analysts Journal* (1998, pages 75-80).

Corrigés

3.1

Soit $\mu = E(X)$.

1.

$$\begin{aligned} \text{var}(X) &= E\left[(X-\mu)^2\right] = E(X^2) - E(2\mu X) + E(\mu^2) \\ &= E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - \mu^2. \end{aligned}$$

2.

$$\begin{aligned} \text{var}(aX) &= E(a^2X^2) - (E(aX))^2 = a^2E(X^2) - a^2(E(X))^2 \\ &= a^2(E(X^2) - (E(X))^2) = a^2\text{var}(X). \end{aligned}$$

3.2

1. Pour calculer la valeur de c, on se sert du fait que l'intégrale de la fonction de densité sur tout son domaine est égale à 1.

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-1}^{1} c(1-x^2)dx = c\left(x - \frac{x^3}{3}\right)_{x=-1}^{x=1} = \frac{4}{3}c = 1$$

et donc $c = 3/4$.

2. Considérons le cas pour lequel $-1 < x < 1$

$$F(x) = P(X < x) = \int_{-1}^{x} f(t)dt = \frac{3}{4}\left(t - \frac{t^3}{3}\right)_{t=-1}^{t=x} = \frac{3}{4}x\left(1 - \frac{x^2}{3}\right) + \frac{1}{2}.$$

Par conséquent, la fonction de répartition est

$$F(x) = \begin{cases} 0 & \text{si } x \leq -1 \\ \frac{3}{4}x\left(1 - \frac{x^2}{3}\right) + \frac{1}{2} & \text{si } |x| < 1 \\ 1 & \text{si } x \geq 1. \end{cases}$$

3.

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{3}{4}\int_{-1}^{1}(x - x^3)dx = 0.$$

Ce résultat est évident car on intègre une fonction impaire ($xf(x)$) sur un domaine centré sur 0.

3.3

1. Si $0 < x < \pi/2$

$$F_X(x) = \int_0^x \sin(t)dt = -\cos(t)\Big|_{t=0}^{t=x} = 1 - \cos(x).$$

Par conséquent

$$F_X(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - \cos(x) & \text{si } 0 < x < \pi/2 \\ 1 & \text{sinon.} \end{cases}$$

2. L'espérance de X est

$$\begin{aligned} E(X) &= \int_0^{\pi/2} x\sin(x)dx = \\ &= \underbrace{-(x\cos(x))_0^{\pi/2}}_{=0} + \int_0^{\pi/2} \cos(x)dx = \sin(x)\Big|_{x=0}^{x=\pi/2} = 1. \end{aligned}$$

3. Calculons le 2e moment de X

$$\begin{aligned} E(X^2) &= \int_0^{\pi/2} x^2\sin(x)dx = \\ &= \underbrace{-\left(x^2\cos(x)\right)_0^{\pi/2}}_{=0} + 2\int_0^{\pi/2} x\cos(x)dx \\ &= \underbrace{2\,(x\sin(x))_0^{\pi/2}}_{=\pi} - 2\int_0^{\pi/2} \sin(x)dx = \pi - 2, \end{aligned}$$

et la variance vaut

$$\text{var}(X) = E(X^2) - E(X)^2 = \pi - 3.$$

3.4

1. La fonction de répartition de X pour $x \geq 0$ est

$$F_X(x) = \int_0^t \lambda e^{-\lambda t}dt = -e^{-\lambda t}\Big|_{t=0}^{t=x} = 1 - e^{-\lambda x},$$

et donc

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 - e^{-\lambda x} & \text{si } x \geq 0. \end{cases}$$

2. Le α-quantile q_α est défini comme

$$q_\alpha = F_X^{-1}(\alpha).$$

On en déduit que pour la loi exponentielle,

$$q_\alpha = -\frac{1}{\lambda}\log(1-\alpha).$$

3.5

1. Pour déterminer la constante de normalisation, on utilise le fait que $\int_{-\infty}^{\infty} f(x)dx = 1$. Ainsi

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-a}^{a} c(a)(a^2-x^2)dx = c(a)\left[a^2x - \frac{x^3}{3}\right]_{-a}^{a} = \frac{4}{3}c(a)a^3 = 1$$

et

$$c(a) = \frac{3}{4a^3}.$$

2. La fonction de répartition pour $-a \leq x \leq a$ est

$$F_X(x) = \int_{-a}^{x} \frac{3}{4a^3}(a^2-t^2)dt = \frac{3}{4a^3}\left(\frac{2a^3}{3} + a^2x - \frac{x^3}{3}\right),$$

et globalement

$$F_X(x) = \begin{cases} 0 & \text{si } x < -a \\ \frac{3}{4a^3}\left(\frac{2a^3}{3} + a^2x - \frac{x^3}{3}\right) & \text{si } |x| \leq a \\ 1 & \text{si } x > a. \end{cases}$$

3. L'espérance de X est

$$E(X) = \int_{-a}^{a} \frac{3}{4a^3}x(a^2-x^2)dx = \frac{3}{4a^3}\left(\frac{a^2x^2}{2} - \frac{x^4}{4}\right)\Big|_{x=-a}^{x=a} = 0$$

car la fonction à évaluer est paire en x et prise sur un intervalle symétrique autour de 0. Le 2$^{\text{e}}$ moment de X est alors égal à la variance de X, ce qui implique

$$\begin{aligned} \text{var}(X) &= E(X^2) = \\ &= \int_{-a}^{a} \frac{3}{4a^3}x^2(a^2-x^2)dx = \frac{3}{4a^3}\left(\frac{a^2x^3}{3} - \frac{x^5}{5}\right)\Big|_{x=-a}^{x=a} = \frac{a^2}{5}. \end{aligned}$$

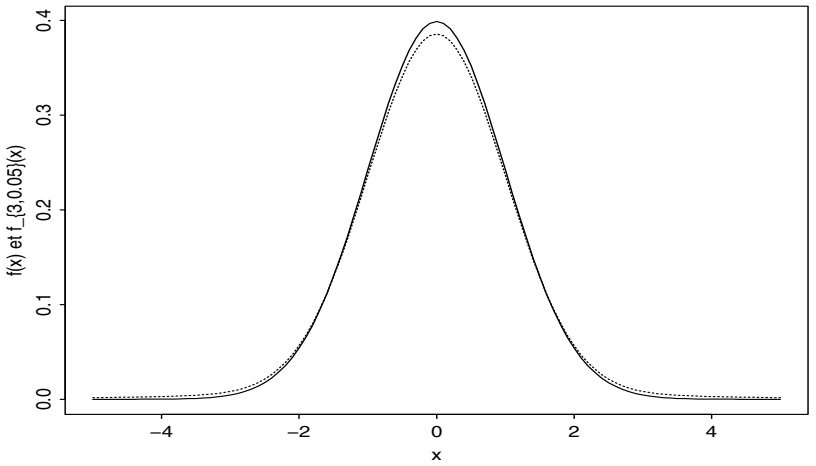


Fig. 3.1 – $f(x)$ en trait continu et $f_{\varepsilon,b}(x)$, avec $\varepsilon = 0{,}05$ et $b = 3$ en pointillés.

3.6

1. $f_{\epsilon,b}(x)$ est positive et

$$\begin{aligned}\int_{-\infty}^{+\infty} f_{\epsilon,b}(x)dx &= (1-\varepsilon)\int_{-\infty}^{+\infty} f(x)dx + \varepsilon \int_{-\infty}^{+\infty} \tfrac{1}{b} f(\tfrac{x}{b})dx \\ &= (1-\varepsilon) + \varepsilon \int_{-\infty}^{+\infty} \tfrac{1}{b} b f(t)dt = (1-\varepsilon) + \varepsilon \int_{-\infty}^{+\infty} f(t)dt = 1\end{aligned}$$

   par le changement de variable $t = x/b$.
2. $f(x) = \varphi(x)$, où φ est la densité d'une variable aléatoire $\mathcal{N}(0,1)$

$$\begin{aligned}E(X) &= \int_{-\infty}^{+\infty} x f_{\varepsilon,b}(x)dx = \int_{-\infty}^{+\infty} (1-\varepsilon)x\varphi(x)dx + \int_{-\infty}^{+\infty} \tfrac{\varepsilon}{b}\, x\varphi(\tfrac{x}{b})dx \\ &= (1-\varepsilon)\int_{-\infty}^{+\infty} x\varphi(x)dx + \varepsilon \int_{-\infty}^{+\infty} \tfrac{x}{b}\varphi(\tfrac{x}{b})dx = 0.\end{aligned}$$

$$\begin{aligned}\mathrm{var}(X) &= E(X^2) = (1-\varepsilon)\int_{-\infty}^{+\infty} x^2\varphi(x)dx + \varepsilon \int_{-\infty}^{+\infty} \tfrac{x^2}{b}\, \varphi(\tfrac{x}{b})dx \\ &= (1-\varepsilon) + \varepsilon b^2 = 1 + \varepsilon(b^2-1)\end{aligned}$$

3. $f_{\varepsilon,b}(x)$, avec $\varepsilon = 0{,}05$ et $b = 3$, est une densité centrée en 0 et avec des queues plus longues que $f(x)$, cf. figure 3.1.

3.7

1. Pour déterminer C et λ, on va utiliser

$$\int_0^\infty Cv\exp(-\lambda v)dv = 1$$

et

$$E(V) = \int_0^\infty Cv^2\exp(-\lambda v)dv = 80.$$

(a)

$$\begin{aligned}\int_0^\infty Cv\exp(-\lambda v)dv &= \\ &\overset{\text{parties}}{=} \underbrace{-\frac{1}{\lambda}Cv\exp(-\lambda v)\Big|_{v=0}^{v=\infty}}_{=0} + C\int_0^\infty \frac{1}{\lambda}\exp(-\lambda v)dv \\ &= -\frac{C}{\lambda^2}\exp(-\lambda v)\Big|_{v=0}^{v=\infty} = \frac{C}{\lambda^2} = 1 \quad \Leftrightarrow \quad C = \lambda^2,\end{aligned}$$

(b)

$$\begin{aligned}C\int_0^\infty v^2\exp(-\lambda v)dv &= \\ &\overset{\text{parties}}{=} \underbrace{-\frac{1}{\lambda}Cv^2\exp(-\lambda v)\Big|_{v=0}^{v=\infty}}_{=0} + C\int_0^\infty \frac{1}{\lambda}2v\exp(-\lambda v)dv \\ &= \frac{2}{\lambda}\underbrace{\int_0^\infty Cv\exp(-\lambda v)dv}_{=1} = \frac{2}{\lambda} = 80 \\ &\Leftrightarrow \quad \lambda = \frac{1}{40} \quad \Rightarrow C = \frac{1}{1\ 600}.\end{aligned}$$

2. Soit $T = \frac{40}{V}$. Sa fonction de répartition est

$$F_T(t) = P(T < t) = P\left(\frac{40}{V} < t\right) = P\left(V > \frac{40}{t}\right) = 1 - F_V\left(\frac{40}{t}\right).$$

On en déduit sa fonction de distribution :

$$f_T(t) = \frac{40}{t^2}f_V\left(\frac{40}{t}\right) = \frac{1}{t^3}\exp\left(-\frac{1}{t}\right),$$

et finalement, on calcule l'espérance de T :

$$\begin{aligned}E(T) &= \int_0^\infty t\frac{1}{t^3}\exp\left(-\frac{1}{t}\right)dt = \int_0^\infty \frac{1}{t^2}\exp\left(-\frac{1}{t}\right)dt \\ &= \exp\left(-\frac{1}{t}\right)\Big|_{t=0}^{t=\infty} = 1.\end{aligned}$$

3.8

1. $$P(X > 20) = 1 - \int_{10}^{20} \frac{10}{x^2} dx = 1 - \Big(-\frac{10}{x} \Big) \Big|_{x=10}^{x=20} = 1 - 0{,}5 = 0{,}5.$$

2. La primitive de $f(x)$ est
$$F(x) = -\frac{10}{x} + C.$$
Pour déterminer C, on utilise le fait que $F(x = 10) = 0$ et ainsi,
$$C = F(10) + \frac{10}{10} = 1.$$
Finalement
$$F_X(x) = \begin{cases} 0 & \text{si } x < 10 \\ 1 - 10/x & \text{si } x \geq 10. \end{cases}$$

3. On calcule en premier la probabilité qu'un composant fonctionne au moins 15 heures
$$P(X \geq 15) = 1 - F_X(15) = \frac{2}{3}.$$
On fait ensuite l'hypothèse que les durées de fonctionnement des composants sont indépendantes, ce qui implique que si Y est la variable aléatoire qui décrit le nombre de composants fonctionnant au moins 15 heures, elle suit une loi binomiale de paramètres $(6, 2/3)$. Par conséquent, la probabilité qu'au moins 3 composants fonctionnent plus de 15 heures est
$$P(Y \geq 3) = \sum_{k=3}^{6} P(Y = k) = \frac{656}{729} \simeq 0{,}90.$$

4. Calculons l'espérance de X
$$E(X) = \int_{10}^{\infty} \frac{10}{x} dx = 10 \log(x) \Big|_{x=10}^{x=\infty} = \infty.$$
L'espérance de X est infinie.

3.9

1. Pour trouver la valeur de c, on utilise le fait que l'intégrale de la fonction de distribution $f(x)$ sur le domaine de définition de x est égale à 1. Ainsi
$$\begin{aligned} \int_{-\infty}^{\infty} f(x)dx &= \int_0^3 cx dx + \int_3^6 c(6-x)dx \\ &= c\frac{x^2}{2}\Big|_{x=0}^{x=3} + c\left(6x - \frac{x^2}{2}\right)\Big|_{x=3}^{x=6} = \frac{9}{2}c + \frac{9}{2}c = 9c \\ &\Leftrightarrow \quad c = \frac{1}{9}. \end{aligned}$$

2. La fonction de répartition se calcule en 2 étapes ; d'abord pour $0 \leq x \leq 3$

$$F_X(x) = \int_0^x \frac{t}{9} dt = \frac{x^2}{18},$$

et ensuite pour $3 \leq x \leq 6$

$$\begin{aligned} F_X(x) &= \int_0^3 \frac{t}{9} dt + \int_3^x \frac{6-t}{9} dt = \frac{t^2}{18}\Big|_{t=0}^{t=3} + \frac{6t - t^2/2}{9}\Big|_{t=3}^{t=x} \\ &= \frac{6x - x^2/2}{9} - 1. \end{aligned}$$

On obtient alors la fonction de répartition

$$F_X(x) = \begin{cases} 0 & \text{si } x < 0 \\ x^2/18 & \text{si } 0 \leq x < 3 \\ \frac{1}{9}(6x - x^2/2) - 1 & \text{si } 3 \leq x < 6 \\ 1 & \text{si } x \geq 6. \end{cases}$$

3. Les événements A et B sont indépendants si

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

On calcule les 3 termes pour vérifier (ou infirmer) la dernière égalité :

$$P(A) = P(X \geq 3) = 1 - F_X(3) = \frac{1}{2},$$

puis

$$P(B) = P(1{,}5 \leq X \leq 4{,}5) = F_X(4{,}5) - F_X(1{,}5) = \frac{7}{8} - \frac{1}{8} = \frac{3}{4},$$

et finalement

$$P(A \cap B) = P(3 \leq X \leq 4{,}5) = F_X(4{,}5) - F_X(3) = \frac{7}{8} - \frac{1}{2} = \frac{3}{8}.$$

On vérifie donc bien que $P(A)P(B) = 3/8$, ce qui signifie que les 2 événements sont indépendants.

3.10

1.

$$P(X > 8) = \int_8^\infty f(x)dx = \int_8^\infty \frac{1}{8} e^{-\frac{x}{8}} dx = -e^{-\frac{x}{8}}\Big|_{x=8}^{x=\infty} = e^{-1} \simeq 0{,}37.$$

La probabilité que la télévision ait une durée de vie supérieure à 8 ans est d'approximativement 37 %.

2. On cherche la probabilité que la durée de vie d'une télévision soit supérieure à 10 ans en sachant qu'elle a déjà 2 ans

$$P(X > 10 \mid X > 2) = \frac{P(X > 10 \cap X > 2)}{P(X > 2)} = \frac{P(X > 10)}{P(X > 2)} = \frac{e^{-5/4}}{e^{-1/4}} = e^{-1}.$$

Le résultat est le même qu'en 1., car la loi exponentielle est *sans mémoire*.

3. Calculons l'espérance de X

$$E(X) = \int_0^\infty \frac{x}{8} e^{-\frac{x}{8}} dx = \underbrace{-xe^{-\frac{x}{8}}\Big|_{x=0}^{x=\infty}}_{=0} + \int_0^\infty e^{-\frac{x}{8}} dx = -8e^{-\frac{x}{8}}\Big|_{x=0}^{x=\infty} = 8.$$

Pour déterminer la variance de X, on va utiliser la relation $\mathrm{var}(X) = E(X^2) - E(X)^2$. Le 2e moment s'obtient de la manière suivante

$$E(X^2) = \int_0^\infty \frac{x^2}{8} e^{-\frac{x}{8}} dx = \underbrace{-x^2 e^{-\frac{x}{8}}\Big|_{x=0}^{x=\infty}}_{=0} + 16 \underbrace{\int_0^\infty \frac{x}{8} e^{-\frac{x}{8}} dx}_{=E(X)=8} = 2 \cdot 8^2,$$

et donc

$$\mathrm{var}(X) = E(X^2) - E(X)^2 = 2 \cdot 8^2 - 8^2 = 64.$$

On remarque que $\mathrm{var}(X) = E(X)^2$.

3.11

Soient X (respectivement Y) la longueur des pièces produites par la machine A (respectivement B). Les variables aléatoires X et Y suivent une loi normale de paramètres (8,4) et (7, 5,1).

1. On cherche la machine qui a la plus grande probabilité de fabriquer des pièces dont la longueur appartient à l'intervalle $[7, 9]$. Premièrement

$$P(7 \leq X \leq 9) = P(-0{,}5 \leq Z \leq 0{,}5) = 2\Phi(0{,}5) - 1 \simeq 0{,}38,$$

où Z est une variable aléatoire de distribution normale centrée et réduite. Pour la 2e machine

$$P(7 \leq Y \leq 9) = P(-0{,}5 \leq Z \leq 1{,}5) \simeq 0{,}62.$$

On en conclut que la machine B est meilleure que la machine A.

2. Soit X^* une nouvelle variable aléatoire qui décrit la taille des pièces produites par la machine A. Cette variable a la même espérance que X et on cherche sa variance σ^2 de telle sorte que les machines soient de qualité égale. On veut donc

$$P(7 \leq X^* \leq 9) = P\left(-\frac{1}{\sigma} \leq Z \leq \frac{1}{\sigma}\right) = 2P\left(Z \leq \frac{1}{\sigma}\right) - 1 = 0{,}62.$$

Ainsi, on trouve

$$P\left(Z \leq \frac{1}{\sigma}\right) = 0{,}81,$$

et finalement

$$\frac{1}{\sigma} \simeq 0{,}88 \quad \Leftrightarrow \quad \sigma^2 \simeq 1{,}29.$$

La variance de X^* devrait être d'environ 1,29 pour que la machine A ait la même probabilité que la machine B de produire une pièce de la bonne taille (cf. figure 3.2).

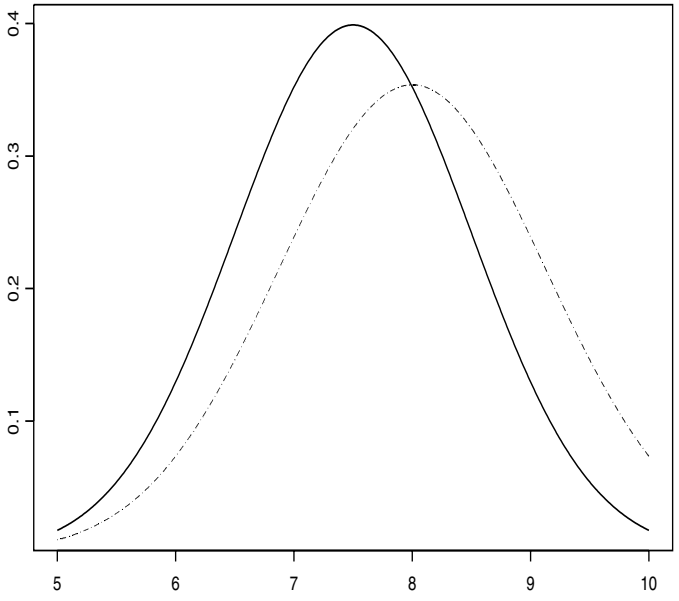


Fig. 3.2 – Fonctions de distribution des variables aléatoires Y (courbe pleine) et X^* (courbe pointillée) de l'exercice 3.11.

3.12

Soit X la taille en centimètres d'un pygmée âgé de 25 ans. Elle suit une loi normale d'espérance 140 et de variance 36.

1. Le pourcentage de pygmées qui mesurent plus que 150 cm est

$$P(X > 150) = 1 - \Phi\left(\frac{150 - 140}{6}\right) \simeq 1 - \Phi(1{,}67) \simeq 4{,}8\ \%.$$

2. On cherche la probabilité qu'un pygmée mesure plus de 150 cm sachant qu'il en mesure au moins 145

$$\begin{aligned} P(X > 150 \mid X > 145) &= \frac{P(X > 150 \cap X > 145)}{P(X > 145)} = \\ &= \frac{P(X > 150)}{P(X > 145)} \simeq \frac{1 - \Phi(1{,}67)}{1 - \Phi(0{,}83)} \simeq 23{,}5\ \%. \end{aligned}$$

3.13

Soit X la taille en centimètres d'un homme âgé de 30 ans. X suit une distribution normale de paramètres $(175, 36)$.

1. Le pourcentage d'hommes de 30 ans mesurant plus que 185 cm est

$$P(X > 185) = 1 - \Phi\left(\frac{185 - 175}{6}\right) \simeq 1 - \Phi(1{,}67) \simeq 4{,}8\ \%.$$

2. La probabilité qu'un homme mesurant plus de 180 cm dépasse 192 cm est

$$\begin{aligned} P(X > 192 \mid X > 180) &= \frac{P(X > 192 \cap X > 180)}{P(X > 180)} = \\ &= \frac{P(X > 192)}{P(X > 180)} \simeq \frac{1 - \Phi(2{,}83)}{1 - \Phi(0{,}83)} \simeq 1{,}1\ \%. \end{aligned}$$

3.14

Pour déterminer la variance σ^2 et l'espérance μ de X, on utilise l'information sur les probabilités pour écrire 2 équations, où Z est une variable provenant d'une loi normale centrée et réduite

$$\left\{ \begin{array}{l} P\left(Z > \frac{190-\mu}{\sigma}\right) = 0{,}1 \\ P\left(Z > \frac{50-\mu}{\sigma}\right) = 0{,}05 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{l} \frac{190-\mu}{\sigma} = \Phi(0{,}9) \simeq 1{,}28 \\ \frac{50-\mu}{\sigma} = \Phi(0{,}05) \simeq -1{,}65, \end{array} \right.$$

ce qui nous permet d'obtenir finalement $\mu \simeq 129$ et $\sigma \simeq 47{,}8$.

3.15

Soit X le taux d'alcool en pourcentage mesuré lors d'un alcootest. La variable aléatoire X a une distribution normale d'espérance $\mu = 0{,}07$ et d'écart-type $\sigma = 0{,}01$.

La proportion d'automobilistes recevant une amende est

$$P(X > 0{,}08) = 1 - \Phi\left(\frac{0{,}08 - 0{,}07}{0{,}01}\right) = 1 - \Phi(1) \simeq 0{,}16,$$

et celle d'automobilistes se faisant retirer le permis parmi ceux ayant reçu une amende est

$$\begin{aligned} P(X > 0{,}09 \mid X > 0{,}08) &= \frac{P(X > 0{,}09 \cap X > 0{,}08)}{P(X > 0{,}08)} = \\ &= \frac{P(X > 0{,}09)}{P(X > 0{,}08)} = \frac{1 - \Phi(2)}{1 - \Phi(1)} \simeq 0{,}14. \end{aligned}$$

3.16

Soit X la durée de grossesse. La variable aléatoire X suit une loi normale avec paramètres (270, 100). La probabilité que la conception de l'enfant ait eu lieu plus de 290 jours ou moins de 240 jours avant sa naissance est

$$\begin{aligned} &P(X > 290 \text{ ou } X < 240) = \\ &= P(X > 290) + P(X < 240) = (1 - \Phi(2)) + (1 - \Phi(3)) \simeq 0{,}021. \end{aligned}$$

3.17

Soit X la variable aléatoire représentant la vitesse mesurée. Cette variable aléatoire suit une loi normale (72, 64).

1. La proportion de conducteurs qui recevront une amende est égale à la probabilité que X soit plus grand que 80, soit

   $$P(X > 80) = P\left(Z > \frac{80 - 72}{8}\right) = 1 - \Phi(1) \simeq 16\ \%,$$

   Z suivant une loi normale centrée et réduite.
2. On cherche la probabilité conditionnelle de mesurer une vitesse supérieure à 110 en sachant que le conducteur est déjà passible d'une amende. On calcule donc

   $$\begin{aligned} P(X > 110 \mid X > 80) &= \frac{P(X > 110 \cap X > 80)}{P(X > 80)} = \\ &= \frac{P(X > 110)}{P(X > 80)} = \frac{1 - \Phi(4{,}75)}{1 - \Phi(1)} \simeq 0. \end{aligned}$$

   La probabilité est presque nulle.

3.18

La fonction de répartition de Y est

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(X^2 < y) = \\ &= P(-\sqrt{y} < X < \sqrt{y}) = F_X(-\sqrt{y}) - F_X(\sqrt{y}). \end{aligned}$$

Dérivons par rapport à y pour obtenir la densité

$$\frac{\partial F_Y(y)}{\partial y} = f_Y(y) = \frac{1}{2\sqrt{y}} f_X(-\sqrt{y}) + \frac{1}{2\sqrt{y}} f_X(\sqrt{y}) = \frac{1}{\sqrt{y}} f_X(\sqrt{y}),$$

où, dans la dernière égalité, on a utilisé la parité de la fonction de densité de la loi normale. Finalement, on trouve la fonction de distribution (illustrée par la figure 3.3)

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}} \exp\left(-\frac{y}{2}\right).$$

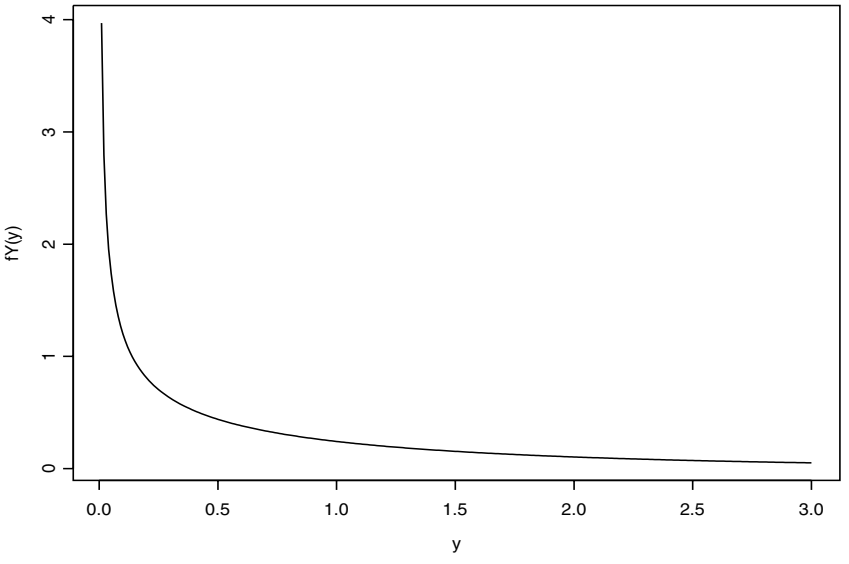


Fig. 3.3 – Fonction de distribution d'une variable aléatoire χ^2 à un degré de liberté (exercice 3.18).

3.19

La variable aléatoire V suit une loi uniforme sur l'intervalle $(0,1)$ et on cherche la fonction de répartition de $W = -\frac{1}{\lambda}\ln(V)$

$$\begin{aligned} F_W(w) = P(W < w) &= P\left(-\frac{1}{\lambda}\ln(V) < w\right) = P(V > \exp(-\lambda w)) \\ &= 1 - F_V(\exp(-\lambda w)) = 1 - \exp(\lambda w), \end{aligned}$$

si $w > 0$ et 0 sinon. La variable aléatoire W suit une loi exponentielle de paramètre λ et sa densité est

$$f_W(w) = \begin{cases} \lambda \exp(-\lambda w) & \text{si } w > 0 \\ 0 & \text{sinon.} \end{cases}$$

3.20

Soit $Y = (X/\alpha)^\beta$, où X suit une loi de Weibull. La fonction de répartition de Y est

$$F_Y(y) = P(Y < y) = P\left(\left(\frac{X}{\alpha}\right)^\beta < y\right) = P(X < \alpha y^{1/\beta}) = F_X(\alpha y^{1/\beta}),$$

et sa fonction de densité

$$\begin{aligned} f_Y(y) &= f_X(\alpha y^{1/\beta})\frac{\alpha}{\beta}y^{1/\beta-1} \\ &= \frac{\alpha}{\beta}y^{\frac{1-\beta}{\beta}}\frac{\beta}{\alpha}\left(\frac{\alpha y^{1/\beta}}{\alpha}\right)^{\beta-1}\exp\left(-\left(\frac{\alpha y^{1/\beta}}{\alpha}\right)^\beta\right) \\ &= \exp(-y). \end{aligned}$$

La variable aléatoire Y suit une loi exponentielle de paramètre $\lambda = 1$.

3.21

La variable aléatoire X suit une loi χ^2_1 et sa fonction de répartition est (voir l'exercice 3.18)

$$F_X(x) = 2\Phi(\sqrt{x}) - 1, \quad x \geq 0.$$

La fonction de répartition de $Y = \log(X)$ est ainsi

$$\begin{aligned} F_Y(y) = P(Y < y) = P(\log(X) < y) = \\ = P(X < \exp(y)) = F_X(\exp(y)) = 2\Phi(\exp(y/2)) - 1, \end{aligned}$$

et sa fonction de densité

$$f_Y(y) = \varphi(\exp(y/2))\exp(y/2) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}e^y + \frac{y}{2}\right),$$

où Φ est la fonction de répartition et φ la fonction de densité d'une variable aléatoire normale centrée et réduite.

3.22

1. L'égalité suivante doit être vérifiée
$$\int_5^\infty \frac{k}{x^\alpha} dx = 1,$$
ce qui implique
$$k = \frac{1}{\int_5^\infty x^{-\alpha} dx}.$$
L'intégrale s'écrit
$$\int_5^\infty x^{-\alpha} dx = \frac{1}{1-\alpha} x^{1-\alpha} \Big|_{t=5}^{t=\infty} = \frac{1}{\alpha - 1} 5^{1-\alpha}.$$
Il faut que $\alpha > 1$ pour que la fonction de répartition reste positive. On trouve par conséquent
$$k = (\alpha - 1) 5^{\alpha - 1}.$$
2. Calculons la fonction de répartition de X
$$F_X(x) = \int_5^x (\alpha - 1) 5^{\alpha-1} t^{-\alpha} dt = -5^{\alpha-1} t^{1-\alpha} \Big|_{t=5}^{t=x} = 1 - \left(\frac{5}{x}\right)^{\alpha - 1},$$
pour $x > 5$.
3. Soit $Y = \ln(X/5)$. Sa fonction de répartition est
$$\begin{aligned} F_Y(y) &= P(Y < y) = P\left(\ln\left(\frac{X}{5}\right) < y\right) \\ &= P(X < 5e^y) = F_X(5e^y) = 1 - \exp(-y(\alpha - 1)), \quad y > 0 \end{aligned}$$
et sa fonction de distribution
$$f_Y(y) = (\alpha - 1) \exp(-y(\alpha - 1)), \quad y > 0.$$
La variable aléatoire Y suit une loi exponentielle de paramètre $\alpha - 1$.

3.23

La variable aléatoire X suit une distribution de Pareto et on cherche la loi de la variable aléatoire $Y = \ln(X) = g(X)$. On utilise la formule de changement de variable pour transformations monotones
$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{\partial g^{-1}(y)}{\partial x} \right| \\ &= \alpha x_0^\alpha \exp(-y(1+\alpha)) \exp(y) = \alpha \exp(\alpha(y_0 - y)), \end{aligned}$$
avec $y > y_0 = \log x_0$.

3.24

Soit $Z = 1/V$ une nouvelle variable aléatoire dont on aimerait connaître la distribution. Sa fonction de répartition est

$$F_Z(z) = P(Z < z) = P\left(\frac{1}{V} < z\right) = P\left(V > \frac{1}{z}\right) = 1 - F_V\left(\frac{1}{z}\right).$$

On dérive cette fonction pour obtenir la densité de Z

$$f_Z(z) \quad = \quad -f_V\left(\frac{1}{z}\right)\left(-\frac{1}{z^2}\right) = \frac{1}{z^2}\frac{1}{\pi(1+\frac{1}{z^2})} = \frac{1}{\pi(z^2+1)}.$$

La variable aléatoire Z est aussi une variable aléatoire suivant une loi de Cauchy.

3.25

On va s'intéresser à la variable aléatoire $S = L^2$.

1. La fonction de densité de S peut être calculée à partir de la fonction de répartition de S

$$F_S(s) = P(S < s) = P(L^2 < s) = P(L < \sqrt{S}) = F_L(\sqrt{s})$$

$$\Rightarrow f_S(s) = \frac{1}{2\sqrt{s}} f_L(\sqrt{s}) = \frac{1}{2\sqrt{s}} 6\sqrt{s}(1-\sqrt{s}) = 3(1-\sqrt{s}), \quad 0 < s < 1.$$

2. L'espérance de S est alors

$$\begin{aligned} E(S) &= \int_0^1 s f_S(s) ds = \int_0^1 3s(1-\sqrt{s}) ds = 3\int_0^1 (s - s^{3/2}) ds \\ &= 3\left(\frac{1}{2}s^2 - \frac{2}{5}s^{5/2}\right)\Big|_{s=0}^{s=1} = \frac{3}{10}. \end{aligned}$$

3.26

1. Trouvons la fonction de répartition de $S = 4\pi R^2$:

$$\begin{aligned} F_S(s) &= P(S < s) = P(4\pi R^2 < s) \\ &= P\left(R < \sqrt{\frac{s}{4\pi}}\right) = F_R\left(\sqrt{\frac{s}{4\pi}}\right). \end{aligned}$$

Sa fonction de densité est donc

$$\begin{aligned} f_S(s) &= f_R\left(\sqrt{\frac{s}{4\pi}}\right)\frac{\partial}{\partial s}\sqrt{\frac{s}{4\pi}} = \frac{2}{9}\sqrt{\frac{s}{4\pi}}\left(3 - \sqrt{\frac{s}{4\pi}}\right)\frac{1}{4\sqrt{s\pi}} \\ &= \frac{1}{36\pi}\left(3 - \sqrt{\frac{s}{4\pi}}\right), \quad 0 \le s \le 36\pi. \end{aligned}$$

2. Calculons l'espérance de S

$$\begin{aligned} E(S) &= \int_0^{36\pi} s f_S(s) ds = \frac{1}{36\pi} \int_0^{36\pi} \left(3s - \frac{s^{3/2}}{\sqrt{4\pi}} \right) ds \\ &= \frac{1}{36\pi} \left(\frac{3}{2} s^2 - \frac{2}{5} \frac{s^{5/2}}{\sqrt{4\pi}} \right)_0^{36\pi} = \frac{54}{5}\pi. \end{aligned}$$

3.27

La variable aléatoire X a une fonction de répartition $F_X(x)$. Soit $Y = F_X(X)$ une nouvelle variable aléatoire. Sa fonction de répartition est

$$F_Y(y) = P(Y < y) = P(F_X(X) < y) = P(X < F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y,$$

pour $0 < y < 1$. Par conséquent, la variable aléatoire Y a une distribution uniforme sur l'intervalle $(0, 1)$.

3.28

Soient T la durée de vie d'une pièce et P l'événement « la pièce est piratée ». On sait que $P(P) = 1/4$, que $T \mid P$ suit une loi exponentielle avec espérance 5 et que $T \mid \overline{P}$ suit une loi exponentielle d'espérance 2. En sachant que la pièce a survécu jusqu'au temps t, la probabilité qu'elle soit piratée est

$$\begin{aligned} \pi(t) &= P(P \mid T > t) = \frac{P(T > t \mid P)P(P)}{P(T > t \mid P)P(P) + P(T > t \mid \overline{P})P(\overline{P})} \\ &= \frac{e^{-\frac{1}{2}t} \cdot \frac{1}{4}}{e^{-\frac{1}{2}t} \cdot \frac{1}{4} + e^{-\frac{1}{5}t} \cdot \frac{3}{4}} \\ &= \frac{1}{1 + 3e^{\frac{3}{10}t}}, \end{aligned}$$

car

$$P(T > t) = \int_t^{\infty} \lambda e^{-\lambda t} dt,$$

avec $\lambda = \frac{1}{2}$ ou $\lambda = \frac{1}{5}$ (cf. exercice 3.4). Lorsque t tend vers l'infini

$$\lim_{t \to \infty} \pi(t) = 0.$$

3.29

Soit X la position du point sur le segment. On suppose que X suit une loi uniforme $(0, L)$. Pour que le rapport entre le plus petit et le plus grand segment

soit inférieur à 1/4, il faut que $X/(L-X) < 1/4$ ou $(L-X)/X < 1/4$. On trouve alors la probabilité

$$\begin{aligned} P\left(\frac{X}{L-X} < \frac{1}{4} \text{ ou } \frac{L-X}{X} < \frac{1}{4}\right) &= P\left(X < \frac{L}{5}\right) + P\left(X > \frac{4L}{5}\right) \\ &= \frac{L/5}{L} + \left(1 - \frac{4L/5}{L}\right) = \frac{2}{5}. \end{aligned}$$

3.30

1. On suppose que le risque est constant :

   $$h(t) = \lambda = -\frac{\partial}{\partial t} \log S(t),$$

   donc

   $$-\lambda t = \log S(t) + C \quad \Leftrightarrow \quad S(t) = e^{-\lambda t - C}.$$

   On trouve alors la fonction $F(t)$ suivante

   $$F(t) = 1 - e^{-\lambda t - C}.$$

   Pour que $F(t)$ soit une fonction de répartition, il faut imposer $C = 0$ pour que $F(0) = 0$. La fonction $F(t)$ se trouve être la fonction de répartition d'une loi exponentielle de paramètre λ.
2. À présent, on cherche la densité de la fonction de survie des malades si la fonction de risque est

   $$h(t) = \lambda\kappa(\lambda t)^{\kappa-1}.$$

   On calcule une primitive $H(t)$ de la fonction $h(t)$:

   $$H(t) = (\lambda t)^{\kappa} + C$$

   et on en déduit

   $$S(t) = \exp(-(\lambda t)^{\kappa} - C).$$

   Comme précédemment, on impose $C = 0$ pour obtenir la fonction de répartition

   $$F(t) = 1 - \exp(-(\lambda t)^{\kappa})$$

   et la fonction de densité correspondante

   $$f(t) = \lambda\kappa(\lambda t)^{\kappa-1} \exp(-(\lambda t)^{\kappa}).$$

   Cette fonction est celle d'une variable aléatoire suivant une loi de Weibull.

3.31

1. Soit $S(t) = P(T > t)$ et $\lambda(t) = -\frac{d}{dt} \log S(t)$.
   Donc $S(t) = \exp\Big(- \int\limits_0^t \lambda(s) ds \Big)$. Dans notre cas

$$\int_0^t \lambda(s)\ ds = \int_0^t \frac{ds}{a+s} = \log\left(\frac{a+t}{a}\right)$$

   et $S(t) = \frac{a}{a+t}$.
2. Puisque $S(t) = 1 - F(t)$, la médiane est la valeur de t qui résout

$$\frac{a}{a+t} = \frac{1}{2},$$

   c'est-à-dire $t = a$.
3. Le paramètre a est le temps de passage à des probabilités de survie supérieures à 50 %.
4. Par le théorème des fonctions inverses

$$F_T(t) = 1 - S(t) = \frac{t}{a+t} = u$$

   pour $t > 0$. Donc

$$t_i = \frac{a u_i}{1 - u_i},$$

   pour $i = 1, \ldots, n$.

3.32

La demande journalière de journaux X suit une distribution normale de paramètres $(100, 200/3)$.

1. Le bénéfice B en centimes s'écrit

$$B = \begin{cases} 5s & \text{si } X \geq s \\ 5X - 10(s - X) & \text{sinon.} \end{cases}$$

   L'espérance de B est alors

$$\begin{aligned} E(B) &= \int_{-\infty}^{s} [5x - 10(s-x)] f_X(x) dx + \int_s^{\infty} 5s f_X(x) dx \\ &= 15 \int_{-\infty}^{s} x f_X(x) dx - 10s \int_{-\infty}^{s} f_X(x) dx + 5s \left[1 - \int_{-\infty}^{s} f_X(x) dx \right] \\ &= 5s - 15s F_X(s) + 15 \int_{-\infty}^{s} x f_X(x) dx. \end{aligned}$$

2. On cherche à présent la valeur $s = s_{max}$ qui maximise $E(B)$. On utilise

$$\frac{\partial}{\partial s}\int_{-\infty}^{s} x f_X(x)dx = s f_X(s)$$

pour calculer la dérivée par rapport à s de $E(B)$

$$\frac{\partial E(B)}{\partial s} = 5 - 15(F_X(s) + s f_X(s)) + 15 s f_X(s) = 5 - 15F_X(s).$$

La valeur s_{max} est la solution de $\partial E(B)/\partial s = 0$, donc

$$5 - 15F_X(s_{max}) = 0 \quad \Leftrightarrow \quad F_X(s_{max}) = \frac{1}{3},$$

et finalement

$$\Phi\left(\frac{s_{max} - 100}{\sqrt{\frac{200}{3}}}\right) = \frac{1}{3} \quad \Leftrightarrow \frac{s_{max} - 100}{\sqrt{\frac{200}{3}}} = -0{,}43 \quad \Leftrightarrow \quad s_{max} \simeq 97.$$

3.33

Commençons par la fonction de répartition de Y

$$F_Y(y) = P(Y < y) = P(\min(X_i) < y) = 1 - P(\min(X_i) > y).$$

Si $\min(X_i)$ est supérieur à y, cela signifie que le plus petit des X_i est supérieur à y, ou encore que tous les X_i sont plus grands que y. Par conséquent,

$$F_Y(y) = 1 - (P(X_1 > y))^n = 1 - (1 - F_X(y))^n = 1 - (1-y)^n.$$

En dérivant $F_Y(y)$ par rapport à y, on a la fonction de densité de y, à savoir

$$f_Y(y) = n(1-y)^{n-1}, \quad 0 < y < 1.$$

3.34

1. Étant donné l'indépendance de $X_1, \ldots, X_n$, on a

$$\begin{aligned} F_T(t) &= P(T < t) = P(\max(X_1, \ldots, X_n) < t) \\ &= P(X_1 < t, X_2 < t, \ldots, X_n < t) \\ &= (P(X_1 < t))^n = (F_X(t))^n. \end{aligned}$$

2. La densité de T est donnée par $f_T(t) = \frac{d}{dt}F_T(t) = n\left(F_X(t)\right)^{n-1} f_X(t)$. En utilisant la définition de $f_X(x)$ et

$$F_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{x^2}{\theta^2} & 0 < x \leq \theta \\ 1 & x > \theta, \end{cases}$$

on obtient

$$f_T(t) = \begin{cases} n\left(\frac{t}{\theta}\right)^{2(n-1)} \frac{2}{\theta^2} t & 0 < t < \theta \\ 0 & \text{sinon.} \end{cases}$$

Enfin

$$\begin{aligned} E(T) &= \int_0^\theta t f_T(t) dt = \int_0^\theta tn \frac{t^{2n-2}}{\theta^{2n-2}} \frac{2}{\theta^2} t dt \\ &= \frac{2n}{\theta^{2n}} \int_0^\theta t^{2n} dt = \frac{2n}{\theta^{2n}} \left(\frac{t^{2n+1}}{2n+1} \right) \Bigg|_{t=0}^{t=\theta} \\ &= \frac{2n}{2n+1} \theta. \end{aligned}$$

3. On calcule la probabilité

$$P(T > a) = 1 - F_T(a) = \begin{cases} 1 & a \leq 0 \\ 1 - \left(\frac{a}{\theta}\right)^{2n} & 0 < a < \theta \\ 0 & a \geq \theta. \end{cases}$$

3.35

Le prix de l'action dans 10 jours s'écrit

$$Y_{11} = Y_{10} + U_{11} = Y_1 + \sum_{i=2}^{11} U_i.$$

Son espérance est alors

$$E(Y_{11}) = E\left(Y_1 + \sum_{i=2}^{11} U_i \right) = 100 + \sum_{i=2}^{11} E(U_i) = 100$$

et sa variance

$$\text{var}(Y_{11}) = \text{var}(Y_1) + \sum_{i=2}^{11} \text{var}(U_i) = 10$$

car tous les U_i sont indépendants. La probabilité que l'action soit comprise entre 95 et 105 dans 10 jours devient

$$P(95 < Y_{11} < 105) = P(\mid Y_{11} - 100 \mid < 5) = 1 - P(\mid Y_{11} - 100 \mid > 5),$$

or, par l'inégalité de Chebychev

$$P(\mid Y_{11} - 100 \mid > 5) \leq \frac{10}{25} = 0{,}4.$$

Donc

$$P(95 < Y_{11} < 105) > 1 - 0{,}4 = 0{,}6,$$

c'est-à-dire qu'il y a 60 % de chance que le prix de l'action se trouve entre 95 et 105 dans 10 jours.

3.36

Soit X une variable aléatoire suivant une loi de Dagum.

1. La fonction de densité de X est la dérivée par rapport à x de la fonction de répartition, soit

$$f_{\beta,\lambda,\delta}(x) = \frac{\partial F_{\beta,\lambda,\delta}(x)}{\partial x} = \beta\lambda\delta x^{-\delta-1}(1+\lambda x^{-\delta})^{-\beta-1},$$

   avec $x > 0$.
2. On cherche premièrement l'équation d'un α-quantile q_α pour la loi de Dagum

$$\alpha = (1+\lambda q_\alpha^{-\delta})^{-\beta} \quad \Leftrightarrow \quad q_\alpha = \lambda^{1/\delta}(\alpha^{-1/\beta}-1)^{-1/\delta}.$$

   À l'aide des données de l'énoncé, on obtient

$$\begin{aligned} q_{0,5} &\simeq 140{,}4 \\ q_{0,25} &\simeq 92{,}5 \\ q_{0,75} &\simeq 206{,}3. \end{aligned}$$

   L'espérance n'est pas égale à la médiane parce que la distribution n'est pas symétrique.
3. L'emploi de la loi normale est inadapté dans ce cas car la distribution des revenus est asymétrique, à l'inverse de la loi normale.
4. Soit $y = 0{,}8x$ et sa fonction de répartition

$$F_{\beta',\lambda',\delta'}(y) = (1+\lambda(0{,}8x)^{-\delta})^{-\beta} = (1+0{,}8^{-\delta}\lambda x^{-\delta})^{-\beta} = (1+\lambda' x-\delta')^{-\beta'},$$

   où $\lambda' = 0{,}8^{-\delta}\lambda$, $\beta' = \beta$ et $\delta' = \delta$. Donc, λ est un paramètre d'échelle.

3.37

On sait que si une variable aléatoire X a une fonction de répartition $F_X(x)$, alors la variable aléatoire Y définie comme $Y = F_X(X)$ suit une loi uniforme sur l'intervalle $(0,1)$ (cf. exercice 3.27). Dans le cas présent, on dispose déjà d'un échantillon d'une loi uniforme $(0,1)$, donc de réalisations de Y, et on cherche à obtenir des réalisation d'une loi de Dagum. On définit alors $X = F^{-1}(Y)$ avec F la fonction de répartition de la loi de Dagum calculée dans l'exercice 3.36. En d'autres termes, on calcule les quantiles de $u_1, \ldots, u_{30}$

$$q_{u_i} = \lambda^{1/\delta}(u_i^{-1/\beta}-1)^{-1/\delta}.$$

3.38

1. On a que $(M-3) \sim \varepsilon(\lambda)$, avec $\lambda = \frac{1}{2}$. Cela implique $E(M-3) = \frac{1}{\lambda}$ et $\text{var}(M-3) = \frac{1}{\lambda^2}$. Dans notre cas $E(M) = 2+3 = 5$ et $\text{var}(M) = 4$.
2. La fonction de densité est $f_M(m) = \lambda e^{-\lambda(m-3)}$ pour $m \geq 3$.
3. Par la formule du changement de variable (transformation monotone), la densité de $X = \exp(M)$ est donnée par

$$\begin{aligned} f_X(x) &= f_M(m) \cdot \tfrac{1}{x} = \lambda e^{-\lambda(\log(x)-3)} \tfrac{1}{x} \\ &= \lambda e^{3\lambda} x^{-(\lambda+1)}, \quad \text{pour } x \geq e^3. \end{aligned}$$

   La fonction de répartition de X est donnée par

$$F_X(x) = \begin{cases} \int_{e^3}^{x} \lambda e^{3\lambda} y^{-(\lambda+1)} dy = 1 - e^{3\lambda} x^{-\lambda} & x \geq e^3 \\ 0 & \text{sinon.} \end{cases}$$

4. La plus faible magnitude est définie par $\min(M_1, M_2)$ et

$$\begin{aligned} P(\min(M_1, M_2) > 4) &= P(M_1 > 4 \text{ et } M_2 > 4) \\ &= P(M_1 > 4) P(M_2 > 4) \\ &= e^{-\lambda} e^{-\lambda} = e^{-1} \simeq 0{,}36. \end{aligned}$$

5. Par le théorème des fonctions inverses (cf. exercice 3.27) on a $u = F_X(x) = 1 - e^{3\lambda} x^{-\lambda}$ et donc $x_i = \frac{e^3}{(1-u_i)^{1/\lambda}}$, $i = 1, \ldots, n$.

3.39

On commence par isoler $R_{0,n}$ de l'équation de l'énoncé

$$R_{0,n} = \left(\frac{S_n}{S_0} \right)^{1/n} - 1.$$

On utilise ensuite l'indication pour obtenir

$$R_{0,n} = \prod_{t=1}^{n} \left(\frac{S_t}{S_{t-1}} \right)^{1/n} - 1 = \exp\left(\frac{1}{n} \sum_{t=1}^{n} \log\left(\frac{S_t}{S_{t-1}} \right) \right) - 1 = e^Y - 1.$$

Par hypothèse

$$\log\left(\frac{S_t}{S_{t-1}} \right) \sim \mathcal{N}(\mu, \sigma^2)$$

et par conséquent $Y \sim \mathcal{N}(\mu, \sigma^2/n)$, ou encore e^Y suit une loi log-normale d'espérance μ et de variance σ^2/n. On trouve dans les tables

$$E(R_{0,n}) = \exp\left(\mu + \frac{\sigma^2}{n} \right) - 1$$

et

$$\mathrm{var}(R_{0,n}) = \exp\left(2\left(\mu + \frac{\sigma^2}{n}\right)\right) - \exp\left(2\mu + \frac{\sigma^2}{n}\right).$$

Lorsque $n \to \infty$, $E(R_{0,n}) = \mu$ et $\mathrm{var}(R_{0,n}) = 0$.

Chapitre 4

Variables aléatoires multivariées

Introduction

Ce chapitre est consacré à l'étude des variables aléatoires multivariées. On discute les outils qui permettent de formaliser le comportement du point de vue probabiliste de plusieurs variables aléatoires simultanément. Les notions de base sont la distribution conjointe, la distribution marginale, la distribution conditionnelle et l'espérance conditionnelle. Ces concepts sont présentés dans les exercices 4.1 à 4.3 pour le cas discret et dans les exercices 4.4 à 4.11 pour le cas continu. On trouvera des exercices qui présentent le calcul d'intégrales doubles, un outil mathématique nécessaire pour calculer des probabilités conjointes.

La loi normale multivariée est discutée dans les exercices 4.12 à 4.14.

Les notions de covariance et corrélation sont traitées dans le cadre du calcul de la distribution d'une combinaison linéaire de variables aléatoires (exercices 4.15 à 4.20). Un cas intéressant dans le domaine de la finance concernant le choix d'un portefeuille d'actifs financiers est discuté en détail dans l'exercice 4.17.

Enfin la dernière partie (exercices 4.21 à 4.25) est consacrée à une série d'exercices plus complexes qui font appel à l'ensemble des concepts étudiés dans ce chapitre.

Références (théorie)

Ross, chapitres 6 et 7 [1] et Pitman, chapitres 5 et 6 [2].

Exercices

Lois conjointes, marginales et conditionnelles

4.1

On lance 3 fois une pièce de monnaie. Définissons X la variable aléatoire qui compte le nombre de piles et Y le numéro du lancement qui donne face pour la 1re fois. (On pose $Y = 0$ si l'on n'obtient pas de face en 3 jets.) Trouver les probabilités $P(X = i, Y = j)$ pour $(i, j) \in \{0, 1, 2, 3\} \times \{0, 1, 2, 3\}$. Constater que $\sum_{i,j} P(X = i, Y = j) = 1$.

4.2

Soit X un chiffre choisi aléatoirement parmi les entiers 1, 2, 3, 4. Soit Y un autre chiffre choisi aléatoirement parmi les chiffres au moins aussi grands que X, mais inférieurs ou égaux à 10.

1. Trouver la loi de probabilité de X et la loi de probabilité conditionnelle de $Y \mid X = x$ pour tout $x \in \{1, 2, 3, 4\}$.
2. Déterminer la loi de probabilité conjointe de (X, Y).

4.3

Je viens d'ouvrir une loterie écologique. Les clients en ont en effet assez de ces systèmes informatiques compliqués auxquels ils ne comprennent rien. Ici, il s'agit seulement de lancer une paire de dés ordinaires, après avoir payé 10 €. On note le plus petit des 2 résultats (ou le résultat unique en cas d'égalité). Si c'est 1 ou 2 on perd les 10 €. Si c'est 3 ou 4, on rejoue. Si c'est 5 ou 6, je donne une excellente bouteille de champagne, qui me revient à 50 €. Comme vous l'imaginez, la simplicité de mon système de loterie fait fureur (ainsi que son appellation écologique) et je gagne en moyenne 1 000 € par jour. À combien de lancers de dés par jour cela correspond?

4.4

Les variables aléatoires X et Y ont la densité conjointe

$$f(x,y) = \begin{cases} e^{-x^2 y} & \text{si } x \geq 1 \text{ et } y \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer $P(X^2 Y > 1)$.

2. Calculer les densités marginales $f_X(x)$ et $f_Y(y)$.
   X et Y sont-elles indépendantes?

4.5

Les variables aléatoires X et Y ont la densité conjointe

$$f(x,y) = \begin{cases} 2xy + \frac{3}{2}y^2 & 0 < x < 1,\ 0 < y < 1 \\ 0 & \text{sinon.} \end{cases}$$

1. Vérifier que $f(x,y)$ est une densité.
2. Trouver les densités marginales $f_X(x)$ et $f_Y(y)$.
3. Trouver les densités conditionnelles $f_{X|Y=y}(x)$ et $f_{Y|X=x}(y)$.
4. Calculer $P((X,Y) \in [0,\frac{1}{2}] \times [0,\frac{1}{2}])$.
5. Trouver $P(X < Y)$.
6. Trouver $E(Y \mid X = x)$.
7. Soit la variable aléatoire $Z = E(Y \mid X)$.
   (a) Quelle est la distribution de Z?
   (b) Trouver $E(Z)$.

4.6 (Problème de Buffon)

Sur un plan on trace des droites parallèles distantes d'une unité de longueur. On y jette au hasard une aiguille de longueur 1. On repère l'aiguille grâce à la distance X entre le milieu de celle-ci et la parallèle la plus proche et grâce à l'angle θ entre l'aiguille et une perpendiculaire aux lignes. On suppose que X suit la loi uniforme sur $(0, 1/2)$, que θ suit la loi uniforme sur $(0, \pi/2)$ et que X et θ sont indépendants. Donner la loi du couple (X, θ) et trouver la probabilité que l'aiguille rencontre une des droites (l'alternative étant que l'aiguille soit complètement située dans une des bandes délimitées par les lignes).

4.7

Marine et Laure se sont données rendez-vous à l'Arena à 20 h 30 environ. Si Marine arrive entre 20 h 15 et 20 h 45 et si Laure arrive indépendamment entre 20 h 00 et 21 h 00, trouver la probabilité que celle qui arrive en premier n'attende pas plus que 5 minutes. Quelle est la probabilité que Marine arrive en premier? (On considère que les arrivées sont uniformément distribuées.)

4.8

Selon l'horaire, un train doit arriver à la gare à 12 h 00 et la quitter à 12 h 07. Le retard X, en minutes, de l'arrivée à la gare est uniformément distribué sur

l'intervalle (-2, 3). Le temps de stationnement à la gare Y est indépendant de X et est uniformément distribué sur l'intervalle (3, 5). Le train ne quitte jamais la gare avant 12 h 07. Trouver la fonction de distribution de Z, le retard avec lequel le train quitte la gare.

Indication : exprimer Z en fonction de X et Y et utiliser une représentation graphique dans le plan (x, y).

4.9

La densité conjointe des variables aléatoires X et Y est donnée par

$$f(x,y) = \begin{cases} e^{-\frac{x}{y}} e^{-y}/y & \text{si } 0 < x < \infty,\ 0 < y < \infty \\ 0 & \text{sinon.} \end{cases}$$

Calculer $P(X > 1 \mid Y = y)$.

4.10

La densité conjointe de X et Y est donnée par

$$f(x,y) = \frac{3\sqrt{3}}{4\pi} \exp\left[-\frac{3}{2}(x^2 + y^2 - xy)\right], \quad (x,y) \in \mathcal{R}^2.$$

1. Trouver les densités marginales $f_X(x)$ et $f_Y(y)$.
   Indication : utiliser les propriétés de l'intégrale d'une densité normale.
2. Les variables aléatoires X et Y sont-elles indépendantes ?
3. Trouver la densité conditionnelle $f_{X|Y}(x \mid y)$.
4. Calculer $E(X \mid Y = y)$.
   Quelle est la distribution de la variable aléatoire $E(X \mid Y)$?

4.11

Soient X et Y les scores à 2 examens. Les données historiques montrent une dépendance entre X et Y qui est modélisée par la densité bivariée

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{756}(x^2 + xy), & \text{si } 0 < x < 6 \text{ et } 0 < y < 6 \\ 0 & \text{sinon.} \end{cases}$$

1. Trouver la densité marginale $f_X(x)$.
2. Trouver la densité conditionnelle $f_{Y|X}(y)$.
3. Calculer $P(X < Y)$.
4. Calculer le meilleur prédicteur $E(Y \mid X)$ et donner son espérance et sa distribution.

Loi normale multivariée

4.12

Soit $X_1, \ldots, X_n$ n variables aléatoires indépendantes de loi $\mathcal{N}(0,1)$.

1. Trouver la densité conjointe $f(x_1, \ldots, x_n)$ de $(X_1, \ldots, X_n)$.
2. Représenter les courbes de niveau (pour le cas $n = 2$) $f(x_1,x_2) = C$.

4.13

Les ventes annuelles du rayon jouets d'une grande surface sont représentées par une variable aléatoire normale X d'espérance μ_X et de variance σ_X^2. Le bénéfice l'est par une variable aléatoire normale Y d'espérance μ_Y et de variance σ_Y^2. Le paramètre ρ_{XY} désigne la corrélation entre les ventes et le bénéfice.

1. Si les ventes s'élèvent à un montant x, quelle est la densité de probabilité du bénéfice?
2. Quelle est la densité de $E(Y \mid X)$?

4.14

Supposons que l'indice *Dow Jones*, noté D, soit une variable aléatoire distribuée selon la loi $\mathcal{N}(\nu, \tau^2)$. Soit S le *Swiss Market Index*. La densité conditionnelle de S sachant que $D = d$ est une loi normale d'espérance d et de variance σ^2.

1. Quelle est la distribution conjointe du couple (D, S)?
2. Quelle est la distribution conditionnelle de D sachant que $S = s$?
3. À partir de données historiques on connaît les valeurs de ν, τ et σ. Ayant observé une valeur s pour le *Swiss Market Index*, calculer le meilleur prédicteur pour l'indice *Dow Jones*.

Combinaisons linéaires de variables aléatoires

4.15

Soit S une variable aléatoire binomiale de paramètres n et p. Représenter S comme une somme de variables aléatoires indépendantes et calculer l'espérance et la variance de S.

4.16

La variable aléatoire $S = \sum_{i=1}^{n} X_i$ est une somme de variables aléatoires indépendantes de même distribution. Calculer son espérance, sa variance et sa distribution lorsque :

1. chaque X_i suit une loi de Poisson de paramètre λ ;
2. chaque X_i suit une loi Gamma de paramètres λ et k.

4.17

On considère un portefeuille avec 2 actifs. Son rendement R_p s'exprime comme combinaison linéaire des rendements R_1 et R_2 des 2 actifs : $R_p = a_1 R_1 + (1 - a_1)R_2$. On connaît les rendements moyens μ_1 et μ_2, les risques σ_1 et σ_2 et la corrélation entre les 2 actifs ρ_{12}.

1. Exprimer le risque σ_p du portefeuille en fonction de la proportion a_1 du premier actif.
2. Quelle doit être la proportion a_1 du 1er actif pour que le risque du portefeuille soit minimal ?
3. On note μ_p le rendement moyen du portefeuille. Exprimer σ_p en fonction de μ_p.
4. Cas particulier : $\mu_1 = 14$ %, $\mu_2 = 8$ %, $\sigma_1 = 6$ %, $\sigma_2 = 3$ %.
   Pour $\rho_{12} = 0$, on a le tableau suivant :

a_1	0,00	0,20	0,40	0,60	0,80	1,00
μ_P	8,00	9,20	10,40	11,60	12,80	14,00
σ_P	3,00	2,68	3,00	3,79	4,84	6,00

   Faire des tableaux semblables pour $\rho_{12} = 1, 0, 5$ et -1.
5. Représenter les courbes des portefeuilles possibles dans l'espace des risques et des rendements moyens. Par convention dans ce domaine, σ_p est en abscisse et μ_p en ordonnée.

4.18

Le nombre de clients se rendant à un grand magasin donné dans l'espace d'une journée est une variable aléatoire suivant une loi de Poisson de paramètre 50. La somme dépensée par chacun des clients quotidiens du magasin est une variable aléatoire d'espérance 8 €. On admet que les dépenses d'un client ne dépendent ni de celles des autres clients, ni du nombre total de clients pour la journée.
Quelle est l'espérance du chiffre d'affaires quotidien du magasin ?

4.19

Soient X et Y 2 variables aléatoires indépendantes telles que $X \sim Bin(n,p)$ et $Y \sim Bin(m,p)$. On définit $Z = X + Y$.

1. Quelle est la distribution de Z ?
2. Quelle est la distribution de $X \mid Z$?
3. Trouver $E(X \mid Z)$.

4.20

Soient X et Y 2 variables aléatoires indépendantes de loi $\mathcal{U}(0,1)$. Quelle est la densité de $X + Y$?

Exercices combinés

4.21

Soit X une variable aléatoire binaire qui représente « l'état d'incertitude sur la scène internationale ». Supposons pour simplifier que X ne puisse admettre que 2 états notés N (« situation normale ») et E (« situation exceptionnelle », comme par exemple une guerre ou un attentat). Soit $P(X = N) = 0{,}95$. De plus, on considère le rendement R d'un actif financier qui suit le modèle stochastique

$$\log(R) = -\frac{\sigma^2}{2} + \sigma Y,$$

où $\sigma > 0$ et Y est une variable aléatoire dont la distribution conditionnelle sachant $\{X = N\}$ est $\mathcal{N}(0,1)$, et sachant $\{X = E\}$ est $\mathcal{N}(0,5^2)$.

1. Interpréter ce modèle : quel effet veut-on capturer avec la distribution conditionnelle postulée ?
2. Calculer l'espérance et la variance de Y.
3. En utilisant le point 2., déduire l'espérance et la variance de $Z = \log(R)$.
4. Écrire la densité $f_Y(y)$ de Y et, en utilisant le fait que la somme de 2 variables aléatoires normales reste une variable aléatoire normale, donner la distribution de Y.
5. Montrer que R suit la distribution

$$f_R(r) = \frac{1}{\sqrt{4{,}4\pi\sigma^2}} \frac{e^{-(\log r + \frac{\sigma^2}{2})^2/(4{,}4\sigma^2)}}{r},$$

c'est-à-dire la loi log-normale de paramètres $(-\sigma^2/2,\ 2{,}2\sigma^2)$.

4.22

Pour chaque automobiliste, les compagnies d'assurance déterminent un indicateur B de qualité de conduite qui est lié à la classe de bonus de l'automobiliste. Des petites valeurs de B caractérisent un bon automobiliste qui se trouve dans une classe à bas risque.
Des études actuarielles indiquent que la distribution de la variable aléatoire B dans la population des automobilistes d'une certaine région est une distribution exponentielle avec espérance 0,1.
Soit N_i le nombre d'accidents durant l'année i (à partir de la date de l'obtention du permis de conduire) pour un automobiliste donné. Le nombre annuel d'accidents causés par un automobiliste, étant donné $B = b$, suit une distribution de Poisson de paramètre b. On suppose aussi que, étant donné $B = b$, il y a indépendance entre le nombre d'accidents causés par un automobiliste durant une certaine année et ceux des années suivantes ou précédentes.

1. Quelle est la probabilité qu'un automobiliste tiré au hasard n'ait aucun accident pendant les 3 premières années d'observation ?
2. On observe un automobiliste tiré au hasard. Soit K la variable aléatoire qui désigne la 1re année pendant laquelle cet automobiliste cause un accident. Déterminer la loi de K, c'est-à-dire $P(K = k)$ pour $k = 0, 1, 2, \ldots$
3. Déterminer l'espérance conditionnelle $E(K \mid B)$.

4.23

On s'intéresse à évaluer la qualité des investisseurs institutionnels. On modélise celle-ci à l'aide d'une variable aléatoire X à 4 modalités (1, 2, 3, 4) : « 4 = très bonne connaissance du marché », « 3 = bonne connaissance du marché », « 2 = faible connaissance du marché » et « 1 = aucune connaissance du marché ». Pour comprendre la répartition des investisseurs entre ces 4 classes, on connaît la différence entre les prévisions données par un groupe d'investisseurs et la valeur réelle du marché mesurée par le *Dow Jones* et le *Nasdaq*. On notera D la variable aléatoire représentant cette différence entre le *Dow Jones* et la prévision d'un investisseur donné et N entre le *Nasdaq* et la prévision de ce même investisseur. On suppose que D suit une loi normale avec espérance 10 (unités en 10^3\$) et variance 2, et N une loi normale avec espérance 3 et variance 1. Enfin, soit $Y(\alpha)$ la variable aléatoire qui représente la somme pondérée des différences entre les prévisions et les valeurs réelles, c'est-à-dire $Y(\alpha) = \alpha D + (1 - \alpha)N$, où α est un coefficient positif connu.

1. Calculer la distribution de $E(Y(\alpha))$ et $\mathrm{var}(Y(\alpha))$ en sachant que la corrélation entre D et N est 0,8.
2. Trouver la distribution de $Y(\alpha)$ en supposant pour simplifier que D et N sont indépendantes.
3. En utilisant le point 2., calculer la densité de la variable aléatoire $Z = \exp(-Y(\alpha))$.

4. De plus, on connaît la relation entre X et $Y(\alpha)$ à travers la distribution conditionnelle de $X \mid Y(\alpha)$.

$j =$	1	2	3	4
$P(X = j \mid Y(\alpha) = y)$	$1 - 5z$	$2z$	$2z$	z

où $z = \exp(-y)$.
En utilisant 3., calculer $E(X)$.

5. Quelle est la probabilité que l'investisseur institutionnel ait une très bonne connaissance du marché?

4.24

Une compagnie pétrolière a décidé de forer dans 10 localités. La probabilité de trouver du pétrole dans une localité est seulement de 1/5, mais si du pétrole est trouvé, alors le profit P_i de la compagnie par la vente du pétrole d'une localité i (en excluant le coût du forage) suit une loi exponentielle d'espérance 5 Mio€.
Soit N la variable aléatoire qui représente le nombre de localités où l'on trouve du pétrole et Z la somme des profits tirés des ventes du pétrole de chaque localité.

1. Exprimer la variable aléatoire Z en fonction de N et des P_i.
2. Calculer $P(Z > 10\text{Mio} \mid N = 1)$ et $P(Z > 10\text{Mio} \mid N = 2)$.
3. Trouver $E(Z \mid N = n)$ et $E(Z)$.
4. Est-ce que la probabilité d'avoir un profit supérieur à 20 Mio€ en vendant le pétrole est plus grande que 1/2?

4.25

Une compagnie d'assurance modélise les montants de dédommagements lors d'accidents comme des variables aléatoires indépendantes $D_1, D_2, \ldots$ suivant une loi exponentielle avec espérance $1/\mu$, avec $\mu > 0$. En outre, le nombre d'accidents $N(t)$ dans un intervalle de temps $[0, t]$ est modélisé selon une variable aléatoire de Poisson $\mathcal{P}(\lambda t)$, avec $\lambda > 0$.

Soit $S = \sum_{i=1}^{N(t)} D_i$ le montant total des dédommagements dans l'intervalle $[0, t]$.

1. À l'aide du théorème central limite, donner une approximation de la distribution conditionnelle de $S \mid N(t)$.
2. Calculer l'espérance et la variance de S.
3. Exprimer la densité de S en fonction de la densité conditionnelle de $S \mid N(t)$ et de la loi de probabilité de $N(t)$.

4. Pour prévoir ses réserves, la compagnie doit connaître la probabilité que S excède une certaine limite s_0. Exprimer $P(S > s_0)$ à l'aide de la fonction de répartition de la distribution normale centrée réduite.
   Indication : utiliser le point 3.

Corrigés

4.1

Après avoir dénombré toutes les possibilités, on représente les probabilités de chaque occurrence des couples (X, Y) dans le tableau suivant.

	$X=0$	$X=1$	$X=2$	$X=3$	somme
$Y=0$	0	0	0	1/8	1/8
$Y=1$	1/8	1/4	1/8	0	1/2
$Y=2$	0	1/8	1/8	0	1/4
$Y=3$	0	0	1/8	0	1/8
somme	1/8	3/8	3/8	1/8	1

Par exemple

$$P(Y=1, X=1) = P(FPF \text{ ou } FFP),$$

où F et P représentent respectivement un tirage « face » et « pile ». On retrouve les lois marginales de X et Y dans la ligne et la colonne « somme ».

4.2

1. La loi de probabilité de X est

$$P(X=x) = \frac{1}{4}, \qquad x \in \{1,2,3,4\},$$

et celle de $Y \mid X=x$

$$P(Y=y \mid X=x) = \frac{1}{10-(x-1)}, \qquad y \in \{x, x+1, \ldots, 10\}.$$

2. La loi de probabilité conjointe de X et Y est obtenue par la relation

$$P(X=x, Y=y) = P(Y=y \mid X=x)P(X=x), \quad y \geq x, \quad x \in 1,2,3,4.$$

Sous forme de tableau récapitulatif, cela donne :

$Y=$	1	2	3	4	5	6	7	8	9	10
$X=1$	1/40	1/40	1/40	1/40	1/40	1/40	1/40	1/40	1/40	1/40
$X=2$	0	1/36	1/36	1/36	1/36	1/36	1/36	1/36	1/36	1/36
$X=3$	0	0	1/32	1/32	1/32	1/32	1/32	1/32	1/32	1/32
$X=4$	0	0	0	1/28	1/28	1/28	1/28	1/28	1/28	1/28

4.3

Soit X la variable aléatoire qui correspond au résultat du 1er dé et Y celle qui correspond au 2e. Sous forme de tableau, mes gains sont :

	$X=1$	$X=2$	$X=3$	$X=4$	$X=5$	$X=6$
$Y=1$	10	10	10	10	10	10
$Y=2$	10	10	10	10	10	10
$Y=3$	10	10	0	0	0	0
$Y=4$	10	10	0	0	0	0
$Y=5$	10	10	0	0	-40	-40
$Y=6$	10	10	0	0	-40	-40

Chaque occurrence arrive avec une probabilité de 1/36. Par conséquent, l'espérance de mes gains sur 1 lancer est la somme du tableau divisée par 36

$$E(G) = \frac{1}{36}(10 \cdot 20 + 0 \cdot 12 - 40 \cdot 4) = \frac{10}{9}.$$

Pour n lancers, l'espérance devient

$$E(\text{gain total}) = nE(G) = \frac{10}{9}n,$$

et comme on cherche n tel que le gain total soit égal à 1 000, on trouve $n = 900$.

4.4

1. Le domaine d'intégration est défini par les relations $X^2Y > 1$, $X > 1$ et $Y > 0$. Autrement dit, on intègre y de $1/x^2$ à l'infini et x sur tout son domaine.

$$\begin{aligned} P(X^2Y > 1) &= \int_1^\infty \left(\int_{1/x^2}^\infty \exp(-x^2 y) dy \right) dx \\ &= \int_1^\infty \left(-\frac{1}{x^2} \exp(-x^2 y) \right) \Big|_{y=1/x^2}^{y=\infty} dx \\ &= \int_1^\infty \frac{1}{x^2} e^{-1} dx = -e^{-1} \frac{1}{x} \Big|_{x=1}^{x=\infty} = e^{-1}. \end{aligned}$$

2. La densité marginale de X est

$$f_X(x) = \int_0^\infty \exp(-x^2 y) dy = \left[-\frac{1}{x^2} \exp(-x^2 y) \right]_{y=0}^{y=\infty} = \frac{1}{x^2}, \quad x \geq 1.$$

Pour la densité marginale de Y, il faut intégrer une fonction de type gaussien (e^{-x^2}). Pour y parvenir, nous allons faire apparaître la densité

d'une loi normale

$$f_Y(y) = \int_1^\infty \exp(-x^2 y)dx = \sqrt{\frac{2\pi}{2y}} \int_1^\infty \frac{1}{\sqrt{2\pi}(1/\sqrt{2y})} \exp(-x^2 y)dx.$$

Dans l'intégrale figure à présent la fonction de distribution d'une variable aléatoire X^* normale d'espérance nulle et de variance $1/\sqrt{2y}$. Par conséquent

$$f_Y(y) = \sqrt{\frac{\pi}{y}} P(X^* > 1) = \sqrt{\frac{\pi}{y}} P(Z > \sqrt{2y}),$$

où Z suit une loi normale centrée et réduite. Finalement

$$f_Y(y) = \sqrt{\frac{\pi}{y}}(1 - \Phi(\sqrt{2y})), \quad y > 0.$$

Comme $f_X(x)f_Y(y) \neq f(x,y)$, X et Y ne sont pas indépendantes.

4.5

1. Afin de vérifier que $f(x, y)$ est une densité, on l'intègre par rapport à x et y sur tout leur domaine

$$\int_0^1 \int_0^1 \left(2xy + \frac{3}{2}y^2\right) dxdy = \int_0^1 \left(x^2 y + \frac{3}{2}xy^2\right)\Big|_{x=0}^{x=1} dy$$
$$= \int_0^1 \left(y + \frac{3}{2}y^2\right) dy = \left(\frac{y^2}{2} + \frac{y^3}{2}\right)\Big|_{y=0}^{y=1} = 1.$$

   La fonction $f(x,y)$ est bien une densité.
2. La densité marginale $f_X(x)$ de X est

$$f_X(x) = \int_0^1 \left(2xy + \frac{3}{2}y^2\right) dy = x + \frac{1}{2}, \quad 0 < x < 1,$$

   et celle de Y

$$f_Y(y) = \int_0^1 \left(2xy + \frac{3}{2}y^2\right) dx = y + \frac{3}{2}y^2, \quad 0 < y < 1.$$

3. On trouve la densité conditionnelle $f_{X|Y=y}(x)$ à l'aide de la relation

$$f_{X|Y=y}(x) = \frac{f(x,y)}{f_Y(y)}.$$

   Ainsi

$$f_{X|Y=y}(x) = \frac{4x + 3y}{2 + 3y}, \quad 0 < x < 1.$$

   Et, de la même manière,

$$f_{Y|X=x}(y) = \frac{4xy + 3y^2}{2x + 1}, \quad 0 < y < 1.$$

4. La probabilité que X et Y appartiennent à l'intervalle $[0,1/2]\times[0,1/2]$ est

$$\begin{aligned} P((X,Y)\in[0,1/2]\times[0,1/2]) &= \int_0^{1/2}\int_0^{1/2}\left(2xy+\frac{3}{2}y^2\right)dxdy \\ &= \int_0^{1/2}\left(x^2y+\frac{3}{2}xy^2\right)\Big|_{x=0}^{x=1/2}dy \\ &= \int_0^{1/2}\left(\frac{1}{4}y+\frac{3}{4}y^2\right)dy \\ &= \left(\frac{1}{8}y^2+\frac{1}{4}y^3\right)\Big|_{y=0}^{y=1/2}=\frac{1}{16}. \end{aligned}$$

5. Pour obtenir la probabilité que Y soit supérieur à X, on intègre sur le domaine suivant

$$\begin{aligned} P(X<Y) &= \int_0^1\int_0^y\left(2xy+\frac{3}{2}y^2\right)dxdy=\int_0^1\left(x^2y+\frac{3}{2}xy^2\right)\Big|_{x=0}^{x=y}dy \\ &= \int_0^1\left(y^3+\frac{3}{2}y^3\right)dy=\frac{5}{8}y^4\Big|_{y=0}^{y=1}=\frac{5}{8}. \end{aligned}$$

6. Calculons l'espérance conditionnelle $E(Y\mid X=x)$:

$$\begin{aligned} E(Y\mid X=x) &= \int_0^1 y f_{Y|X=x}(y)dy=\int_0^1 y\frac{4xy+3y^2}{2x+1}dy \\ &= \frac{1}{2x+1}\left(\frac{4}{3}xy^3+\frac{3}{4}y^4\right)\Big|_{y=0}^{y=1}=\frac{1}{12}\frac{16x+9}{2x+1}. \end{aligned}$$

7. Soit $Z=E(Y\mid X=x)$. Puisque X est défini entre 0 et 1, le domaine de Z est $[25/36,3/4]$.
   (a) La fonction de répartition de Z est

$$\begin{aligned} F_Z(z) &= P(Z<z)=P\left(\frac{1}{12}\frac{16x+9}{2x+1}<z\right) \\ &= P\left(X<\frac{12z-9}{16-24z}\right)=F_X\left(\frac{12z-9}{16-24z}\right), \end{aligned}$$

   et sa fonction de densité

$$\begin{aligned} f_Z(z) &= f_X\left(\frac{12z-9}{16-24z}\right)\frac{3}{8}\frac{-1}{(2-3z)^2} \\ &= -\frac{3}{8}\frac{1}{(2-3z)^2}\left(\frac{3}{8}\left(\frac{4z-3}{2-3z}\right)+\frac{1}{2}\right) \\ &= \frac{3}{64(2-3z)^3}, \qquad \frac{25}{36}<z<\frac{3}{4}. \end{aligned}$$

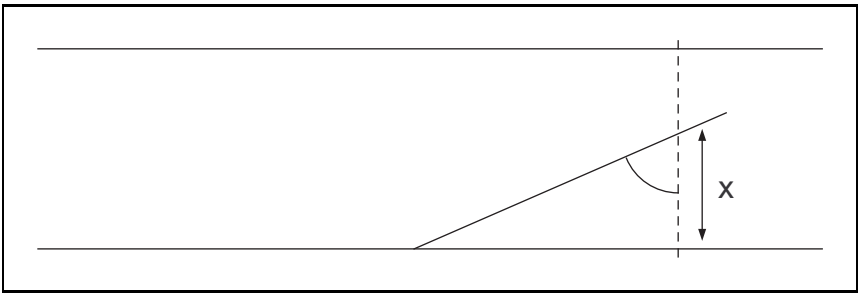


Fig. 4.1 – Illustration de l'exercice 4.6.

(b) L'espérance de Z peut se calculer de manière rapide pour éviter l'intégrale de $f_Z(z)$

$$E(Z) = E(E(Y \mid X)) = E(Y) = \int_0^1 \left(y^2 + \frac{3}{2}y^3\right) dy = \frac{17}{24}.$$

4.6

Les lois marginales de X et θ sont uniformes et leur fonction de densité sont

$$f_\theta(\theta) = \frac{2}{\pi}, \qquad 0 < \theta < \frac{\pi}{2}$$

et

$$f_X(x) = 2, \qquad 0 < x < \frac{1}{2}.$$

Les variables aléatoires X et θ sont indépendantes et, par conséquent, la densité conjointe de X et θ est

$$f(x,\theta) = \frac{4}{\pi}.$$

Au vu de la figure 4.1, il faut que l'hypoténuse du triangle en θ soit plus petite que $1/2$. De cette manière, l'aiguille chevauchera forcément la droite. On a donc la condition suivante sur X

$$\frac{x}{\cos\theta} < \frac{1}{2} \quad \Leftrightarrow \quad x < \frac{\cos\theta}{2}.$$

La probabilité que l'aiguille touche une droite est donc

$$P\left((\theta,x) \in \left[0, \frac{\pi}{2}\right] \times \left[0, \frac{\cos\theta}{2}\right]\right) = \int_0^{\pi/2} \int_0^{\cos\theta/2} \frac{4}{\pi} dx d\theta =$$
$$= \int_0^{\pi/2} \frac{4}{\pi} x \Big|_{x=0}^{x=\cos\theta/2} d\theta = \int_0^{\pi/2} \frac{2\cos\theta}{\pi} d\theta = \frac{2}{\pi} \sin\theta \Big|_{\theta=0}^{\theta=\pi/2} = \frac{2}{\pi}.$$

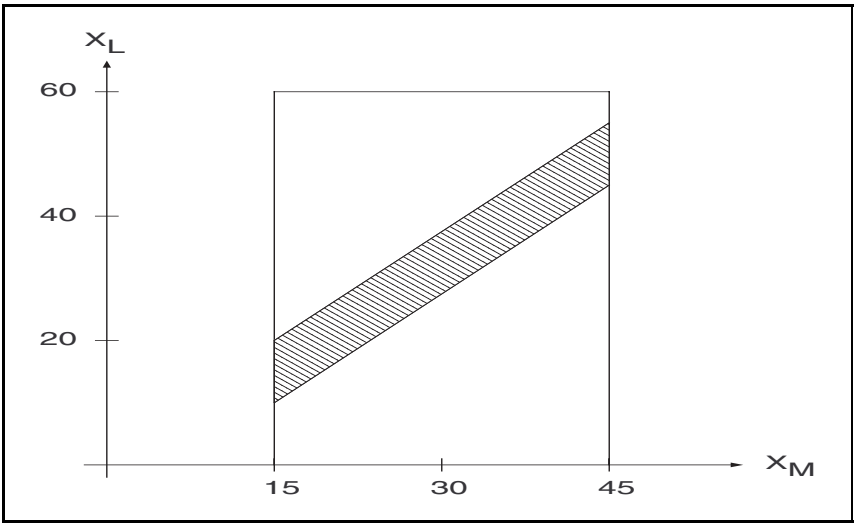


Fig. 4.2 – Domaine d'intégration de l'exercice 4.7.

4.7

Soient X_M et X_L les variables aléatoires donnant l'heure d'arrivée (en minutes après 20 h 00) de Marine et Laure. Elles suivent des lois uniformes respectivement de paramètres $(15, 45)$ et $(0, 60)$.

1. On cherche la probabilité que les 2 heures d'arrivée aient un écart inférieur à 5 minutes, soit

$$\begin{aligned} & P(\mid X_L - X_M \mid < 5) = \\ & \quad = \quad P(-5 < X_L - X_M < 5) = P(X_M - 5 < X_L < X_M + 5) \\ & \quad = \quad \iint\limits_{\mathcal{D}} \frac{1}{60}\frac{1}{30} dx_M dx_L, \end{aligned}$$

où le domaine de définition $\mathcal{D}$ correspond à l'aire hachurée représentée sur la figure 4.2. Ainsi

$$\begin{aligned} P(\mid X_L - X_M \mid < 5) &= \int_{15}^{45} \int_{x_M-5}^{x_M+5} \frac{1}{30}\frac{1}{60} dx_L dx_M \\ &= \int_{15}^{45} \frac{1}{1800}(x_M + 5 - x_M + 5) dx_M \\ &= \frac{1}{180}\int_{15}^{45} dx_M = \frac{1}{6}. \end{aligned}$$

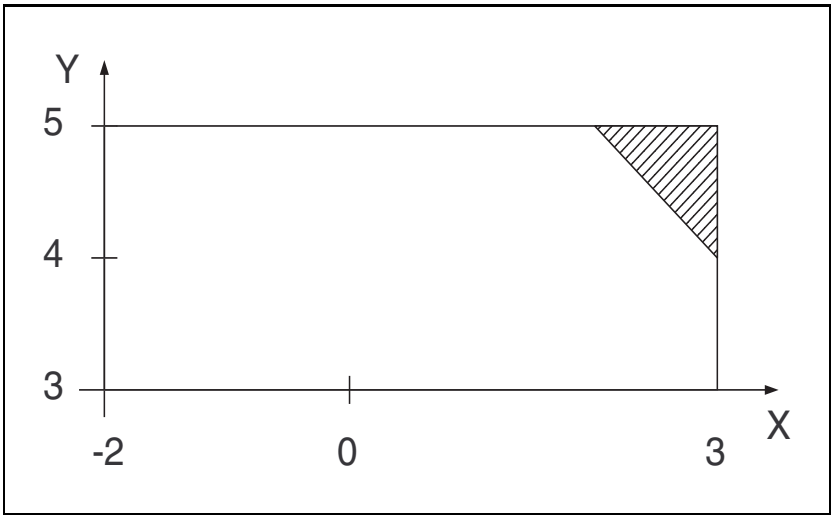


Fig. 4.3 – Représentation graphique du retard du train de l'exercice 4.8.

2. La probabilité que Marine arrive en premier est

$$P(X_M < X_L) = \int_{15}^{45} \int_{x_M}^{60} \frac{1}{30}\frac{1}{60} dx_L dx_M = \frac{1}{2}.$$

4.8

L'heure en minutes après 12 h 00 à laquelle le train quitte la gare est égale à la somme du retard X et du temps de stationnement Y. Or, cette somme ne peut pas être inférieure à 7 parce que le train ne part jamais en avance sur son horaire. Par conséquent, l'heure de départ sera égale à $\max(X+Y, 7)$ et le retard Z

$$Z = \max(X+Y, 7) - 7 = \max(X+Y-7, 0).$$

La figure 4.3 donne une représentation du retard dans le plan (x, y).

Nous savons que le retard Z est supérieur ou égal à 0 et qu'il ne peut pas dépasser 1 minute car $X+Y-7 \leq 1$. La fonction de répartition de Z est alors

$$F_Z(z) = P(Z < z) = \begin{cases} 0 & \text{si } z \leq 0 \\ P(X+Y-7 \leq z) & \text{si } 0 \leq z \leq 1 \\ 1 & \text{si } z \geq 1. \end{cases}$$

Il faut donc calculer $P(X+Y-7 \leq z)$ et nous allons utiliser un agrandissement (figure 4.4) de l'aire hachurée de la figure 4.3. Comme les distributions de X et Y sont uniformes, le calcul de cette probabilité revient à faire le rapport de

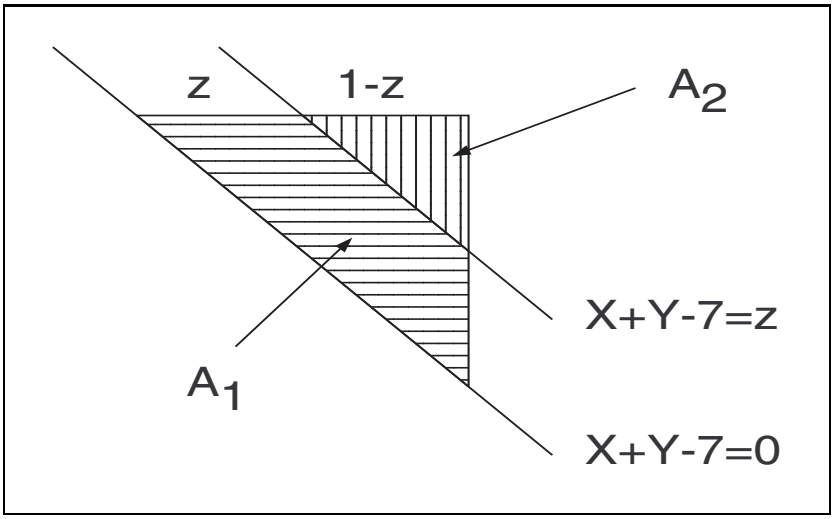


Fig. 4.4 – Agrandissement de l'aire hachurée de la figure 4.3 de l'exercice 4.8.

l'aire hachurée horizontalement (A_1) sur l'aire du triangle ($A_T = A_1 + A_2$) et on obtient

$$\begin{aligned} P(X+Y-7 \leq z) &= \frac{\text{aire } A_1}{\text{aire } A_T} = \frac{\text{aire } A_T - \text{aire } A_2}{\text{aire } A_T} \\ &= \frac{\frac{1}{2} - \frac{1}{2}(1-z)^2}{\frac{1}{2}} = 1 - (1-z)^2 = 2z - z^2. \end{aligned}$$

4.9

Pour obtenir la probabilité conditionnelle de $X \mid Y$, il faut calculer la densité marginale de Y

$$f_Y(y) = \int_0^\infty \frac{1}{y} \exp\left(-\frac{x}{y} - y\right) dx = \frac{e^{-y}}{y} \left(-y \exp\left(-\frac{x}{y}\right)\right) \Big|_{x=0}^{x=\infty} = e^{-y}, \quad y > 0$$

et ainsi obtenir la densité conditionnelle

$$f_{X|Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{1}{y} \exp\left(-\frac{x}{y}\right), \quad x > 0, \quad y > 0.$$

La probabilité cherchée est donc

$$P(X > 1 \mid Y = y) = \int_1^\infty \frac{1}{y} \exp\left(-\frac{x}{y}\right) dx = \exp\left(-\frac{1}{y}\right).$$

4.10

1. Pour déterminer la densité marginale de X, on intègre la densité conjointe sur le domaine de Y

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} \frac{3\sqrt{3}}{4\pi} \exp\left(-\frac{3}{2}(x^2+y^2-xy)\right) dy \\ &= \frac{3\sqrt{3}}{4\pi} \exp\left(-\frac{3}{2}x^2\right) \int_{-\infty}^{\infty} \exp\left(-\frac{3}{2}(y^2-xy)\right) dy. \end{aligned}$$

   Dans la dernière intégrale, on complète le carré afin de faire apparaître la densité d'une variable aléatoire normale et on obtient

$$\begin{aligned} &f_X(x) = \\ &= \frac{3\sqrt{3}}{4\pi} \exp\left(-\frac{3}{2}x^2+\frac{3}{8}x^2\right) \sqrt{\frac{2\pi}{3}} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\sqrt{3} \exp\left(-\frac{3}{2}\left(y-\frac{x}{2}\right)^2\right) dy}_{=1} \\ &= \frac{3}{2\sqrt{2\pi}} \exp\left(-\frac{9}{8}x^2\right). \end{aligned}$$

   On voit donc que $X \sim \mathcal{N}(0, 4/9)$. La fonction de densité conjointe est symétrique en x et y et il est donc inutile de calculer la densité marginale de Y ; c'est la même que celle de X

$$f_Y(y) = \frac{3}{2\sqrt{2\pi}} \exp\left(-\frac{9}{8}y^2\right).$$

   Ainsi, $Y \sim \mathcal{N}(0, 4/9)$.
2. On vérifie facilement que le produit des densités marginales est différent de la densité conjointe, ce qui signifie que les variables aléatoires X et Y ne sont pas indépendantes.
3. La densité conditionnelle de $X \mid Y = y$ s'écrit

$$\begin{aligned} f_{X|Y=y}(x) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\frac{3\sqrt{3}}{4\pi} \exp\left(-\frac{3}{2}(x^2+y^2-xy)\right)}{\frac{3}{2\sqrt{2\pi}} \exp\left(-\frac{9}{8}y^2\right)} \\ &= \sqrt{\frac{3}{2\pi}} \exp\left(-\frac{3}{2}\left(x^2+\frac{1}{4}y^2-xy\right)\right) \\ &= \sqrt{\frac{3}{2\pi}} \exp\left(-\frac{3}{2}\left(x-\frac{y}{2}\right)^2\right), \end{aligned}$$

   et par conséquent, $X \mid Y = y \sim \mathcal{N}(y/2, 1/3)$.
4. L'espérance de $X \mid Y = y$ est directement donnée par le calcul précédent : elle vaut $E(X \mid Y = y) = y/2$. La variable aléatoire $U = E(X \mid Y) = Y/2$ suit une loi normale d'espérance nulle et de variance $1/9$.

4.11

1. La densité marginale de X est égale à la densité conjointe intégrée sur le domaine de Y

$$f_X(x) = \int_0^6 \frac{1}{756}(x^2 + xy)dy = \frac{1}{126}(x^2 + 3x).$$

2. La densité conditionnelle de $Y \mid X = x$ est le rapport entre la densité conjointe et la densité marginale de X

$$f_{Y|X=x}(y) = \frac{1}{6}\frac{x+y}{x+3}.$$

3. On calcule à présent la probabilité que X soit inférieur à Y. On résout donc l'intégrale suivante

$$P(X < Y) = \int_0^6 \left[\int_0^y \frac{1}{756}(x^2 + xy)dx\right] dy = \int_0^6 \frac{5}{4536}y^3 dy = \frac{5}{14}.$$

4. Commençons par trouver l'espérance de $Y \mid X = x$

$$E(Y \mid X = x) = \int_0^6 y f_{Y|X=x}(y)dy = \int_0^6 \frac{1}{6}\frac{xy + y^2}{x+3}dy = 3\frac{x+4}{x+3}.$$

On crée maintenant une nouvelle variable aléatoire $Z = E(Y \mid X)$ définie sur le domaine $]10/3, 4[$. La fonction de répartition de Z est

$$\begin{aligned} F_Z(z) = P(Z < z) &= P\left(3\frac{x+4}{x+3} < z\right) = \\ &= P\left(X > 3\frac{z-4}{3-z}\right) = 1 - F_X\left(3\frac{z-4}{3-z}\right). \end{aligned}$$

On en déduit sa fonction de densité $f_Z(z)$

$$f_Z(z) = -f_X\left(3\frac{z-4}{3-z}\right)\left|\frac{d}{dz}\left(3\frac{z-4}{3-z}\right)\right| = -\frac{27}{142}3\frac{z-4}{(z-3)^4}.$$

Finalement

$$E(Z) = \int_{10/3}^4 z f_Z(z)dz = \frac{63}{71}.$$

4.12

1. Comme toutes les variables aléatoires sont indépendantes, la densité conjointe est le produit des densités marginales

$$f(x_1, \ldots, x_n) = \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right).$$

2. L'équation $f(x_1, x_2) = C$ donne

$$x_1^2 + x_2^2 = -2\log(2\pi C)$$

qui se trouve être l'équation d'un cercle centré sur l'origine et de rayon $\sqrt{-2\log(2\pi C)}$ pour $C < 1$.

4.13

La densité conjointe de X et Y est

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu_X, y-\mu_Y)\Sigma^{-1}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}\right),$$

avec

$$\Sigma = \begin{pmatrix} \sigma_X^2 & \sigma_X\sigma_Y\rho_{XY} \\ \sigma_X\sigma_Y\rho_{XY} & \sigma_Y^2 \end{pmatrix}$$

et

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma_X^2(1-\rho_{XY}^2)} & -\frac{\rho_{XY}}{\sigma_X\sigma_Y(1-\rho_{XY}^2)} \\ -\frac{\rho_{XY}}{\sigma_X\sigma_Y(1-\rho_{XY}^2)} & \frac{1}{\sigma_Y^2(1-\rho_{XY}^2)} \end{pmatrix}.$$

1. On calcule $f_{Y|X=x}(y)$

$$\begin{aligned} f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \\ = \quad & \frac{\frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \exp\left(-\frac{1}{2}(x-\mu_X, y-\mu_Y)\Sigma^{-1}\begin{pmatrix} x-\mu_X \\ y-\mu_Y \end{pmatrix}\right)}{\frac{1}{\sqrt{2\pi}\sigma_X}\exp\left(-\frac{1}{2\sigma_X^2}(x-\mu_X)^2\right)} \\ = \quad & \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho_{XY}^2)}} \exp\Bigg[-\frac{1}{2(1-\rho_{XY}^2)}\Bigg(\frac{(y-\mu_Y)^2}{\sigma_Y^2} \\ & -2\frac{\rho_{XY}}{\sigma_X\sigma_Y}(x-\mu_X)(y-\mu_Y) + \big(1-(1-\rho_{XY}^2)\big)\frac{(x-\mu_X)^2}{\sigma_X^2}\Bigg)\Bigg] \\ = \quad & \frac{1}{\sigma_Y\sqrt{2\pi(1-\rho_{XY}^2)}} \exp\left[-\frac{1}{2(1-\rho_{XY}^2)}\left(\frac{y-\mu_Y}{\sigma_Y} - \rho_{XY}\frac{x-\mu_X}{\sigma_X}\right)^2\right] \end{aligned}$$

et on voit que

$$Y \mid X \sim \mathcal{N}\left(\mu_Y + \frac{\sigma_Y}{\sigma_X}\rho_{XY}(x-\mu_X),\, \sigma_Y^2(1-\rho_{XY}^2)\right).$$

2. La variable aléatoire $Z = E(Y \mid X)$ est définie par

$$Z = E(Y \mid X) = \mu_Y + \frac{\sigma_Y}{\sigma_X}\rho_{XY}(X-\mu_X).$$

Notons que $X - \mu_X \sim \mathcal{N}(0, \sigma_X^2)$ et que par conséquent

$$\frac{\sigma_Y}{\sigma_X}\rho_{XY}(X - \mu_X) \sim \mathcal{N}(0, \sigma_Y^2\rho_{XY}^2).$$

Finalement

$$Z \sim \mathcal{N}(\mu_Y, \sigma_Y^2\rho_{XY}^2).$$

4.14

Nous savons que $D \sim \mathcal{N}(\nu, \tau^2)$ et $S \mid D = d \sim \mathcal{N}(d, \sigma^2)$.

1. Calculons la distribution conjointe de S et D

$$\begin{aligned}
f_{S,D}(s,d) = f_{S|D=d}(s)f_D(d) = \\
&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(s-d)^2\right)\frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{1}{2\tau^2}(d-\nu)^2\right) \\
&= \frac{1}{2\pi\sigma\tau}\exp\left(-\frac{1}{2}\left[\frac{s^2}{\sigma^2} + \left(\frac{1}{\sigma^2}+\frac{1}{\tau^2}\right)d^2 - 2\frac{sd}{\sigma^2} - 2\frac{d\nu}{\tau^2} + \frac{\nu^2}{\tau^2}\right]\right) \\
&= \frac{1}{2\pi\sigma\tau}\exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}(s-\nu)^2 + \frac{2}{\sigma^2}s\nu - \frac{1}{\sigma^2}\nu^2 + \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}(d-\nu)^2\right.\right. \\
&\quad \left.\left. +2\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}d\nu - \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\nu^2 - 2\frac{sd}{\sigma^2} - 2\frac{d\nu}{\tau^2} + \frac{\nu^2}{\tau^2}\right]\right) \\
&= \frac{1}{2\pi\sigma\tau}\exp\left(-\frac{1}{2}\left[\frac{1}{\sigma^2}(s-\nu)^2 + \frac{\sigma^2+\tau^2}{\sigma^2\tau^2}(d-\nu)^2\right.\right. \\
&\quad \left.\left. +2\frac{s\nu}{\sigma^2} + 2\frac{s\nu}{\sigma^2} - 2\frac{sd}{\sigma^2} - 2\frac{\nu^2}{\sigma^2}\right]\right) \\
&= \frac{1}{2\pi\sigma\tau}\exp\left(-\frac{1}{2}(s-\nu,d-\nu)\Sigma^{-1}\begin{pmatrix} s-\nu \\ d-\nu \end{pmatrix}\right),
\end{aligned}$$

où

$$\Sigma^{-1} = \begin{pmatrix} \frac{1}{\sigma^2} & -\frac{1}{\sigma^2} \\ -\frac{1}{\sigma^2} & \frac{\sigma^2+\tau^2}{\sigma^2\tau^2} \end{pmatrix}, \quad \text{car} \quad \Sigma = \begin{pmatrix} \sigma^2+\tau^2 & \tau^2 \\ \tau^2 & \tau^2 \end{pmatrix}.$$

On en conclut que

$$\begin{pmatrix} S \\ D \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \nu \\ \nu \end{pmatrix}, \Sigma\right),$$

et $S \sim \mathcal{N}(\nu, \sigma^2 + \tau^2)$.

2. De même, on calcule la distribution conditionnelle de $D \mid S = s$

$$\begin{aligned} f_{D|S=s}(s) &= \frac{f_{D,S}(s,d)}{f_S(s)} \\ &= \frac{\frac{1}{2\pi\sigma\tau}\exp\left(-\frac{1}{2}\left[\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}(d-\nu)^2 - 2\frac{(d-\nu)(s-\nu)}{\sigma^2} + \frac{1}{\sigma^2}(s-\nu)^2\right]\right)}{\frac{1}{\sqrt{2\pi(\sigma^2+\tau^2)}}\exp\left(-\frac{1}{2}\frac{1}{\sigma^2+\tau^2}(s-\nu)^2\right)} \\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}}\exp\left(-\frac{1}{2}\left[\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}(d-\nu)^2 \right.\right. \\ &\quad \left.\left. -2\frac{(d-\nu)(s-\nu)}{\sigma^2} + \frac{\tau^2}{\sigma^2(\sigma^2+\tau^2)}\right]\right) \\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}}\exp\left(-\frac{1}{2}\left[\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}}(d-\nu) - \frac{\tau}{\sigma\sqrt{\sigma^2+\tau^2}}\right]^2\right) \\ &= \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}}\exp\left(-\frac{1}{2}\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\left[d-\nu-\frac{\tau^2}{\sigma^2+\tau^2}(s-\nu)\right]^2\right). \end{aligned}$$

Par conséquent

$$D \mid S = s \sim \mathcal{N}\left(\nu + \frac{\tau^2}{\sigma^2+\tau^2}(s-\nu),\ \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right).$$

3. Le meilleur prédicteur est donné par l'espérance conditionnelle

$$E(D \mid S = s) = \nu + \frac{\tau^2}{\sigma^2+\tau^2}(s-\nu) = \frac{\sigma^2}{\sigma^2+\tau^2}\nu + \frac{\tau^2}{\sigma^2+\tau^2}s.$$

4.15

Soient $X_1, \ldots, X_n$ des variables aléatoires de Bernoulli indépendantes valant 1 avec une probabilité p. La variable aléatoire S est alors la somme de ces variables, $S = \sum_{i=1}^n X_i$, et par conséquent

$$E(S) = E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i) = np,$$

et

$$\mathrm{var}(S) = \sum_{i=1}^n \mathrm{var}(X_i) = np(1-p).$$

4.16

Notons $\phi_X(t) = E\left(e^{tx}\right)$ la fonction génératrice des moments d'une variable aléatoire X. On utilise pour cet exercice la relation

$$\phi_S(t) = \left(\phi_{X_1}(t)\right)^n,$$

où $S = \sum_{i=1}^{n} X_i$ et les X_i sont indépendants.

1. Si X_i suit une loi de Poisson de paramètre λ, sa fonction génératrice des moments est

$$\phi_{X_i}(t) = \sum_{k=0} e^{tk} \frac{\lambda^k e^{-\lambda}}{k!} = \exp(\lambda(e^{-t} - 1))$$

et celle de S

$$\phi_S(t) = \exp(n\lambda(e^{-t} - 1)).$$

On voit que $\phi_S(t)$ est la fonction génératrice des moments d'une variable aléatoire de distribution de Poisson avec paramètre $n\lambda$. Ainsi, $E(S) = \text{var}(S) = n\lambda$.

2. La fonction génératrice des moments de X_i est

$$\phi_{X_i}(t) = \int_0^\infty e^{tx} \frac{\lambda e^{-\lambda x}(\lambda x)^{\kappa-1}}{\Gamma(\kappa)} dx = \frac{\lambda^\kappa}{\Gamma(\kappa)} \int_0^\infty e^{(t-\lambda)x} x^{\kappa-1} dx.$$

On applique le changement de variable $y = -(t - \lambda)x$ pour obtenir

$$\phi_{X_i}(t) = \frac{\lambda^\kappa}{\Gamma(\kappa)} \frac{1}{(\lambda - t)^\kappa} \underbrace{\int_0^\infty e^{-y} y^{\kappa-1} dy}_{=\Gamma(\kappa)} = \left(\frac{\lambda}{\lambda - t}\right)^\kappa .$$

La fonction génératrice des moments de S devient

$$\phi_S(t) = \left(\frac{\lambda}{\lambda - t}\right)^{n\kappa},$$

et on voit que S suit donc une loi Gamma de paramètres $(\lambda, n\kappa)$.

4.17

1. Le risque σ_p du portefeuille s'écrit

$$\begin{aligned} \sigma_p^2 &= a_1^2\sigma_1^2 + (1 - a_1)^2\sigma_2^2 + 2a_1(1 - a_1)\sigma_1\sigma_2\rho_{12} \\ &= (\sigma_1^2 - 2\sigma_1\sigma_2\rho_{12} + \sigma_2^2)a_1^2 - 2(\sigma_2^2 - \sigma_1\sigma_2\rho_{12})a_1 + \sigma_2^2. \end{aligned}$$

2. Ce risque est minimal lorsque $\partial\sigma_p^2/\partial a_1 = 0$ car le coefficient multiplicatif $(\sigma_1^2 - 2\sigma_1\sigma_2\rho_{12} + \sigma_2^2)$ est positif. La dérivée partielle est

$$\frac{\partial\sigma_p^2}{\partial a_1} = 2a_1(\sigma_1^2 - 2\sigma_1\sigma_2\rho_{12} + \sigma_2^2) - 2(\sigma_2^2 - \sigma_1\sigma_2\rho_{12}) = 0,$$

et enfin

$$a_1 = \frac{\sigma_2^2 - \sigma_1\sigma_2\rho_{12}}{\sigma_1^2 - 2\sigma_1\sigma_2\rho_{12} + \sigma_2^2}.$$

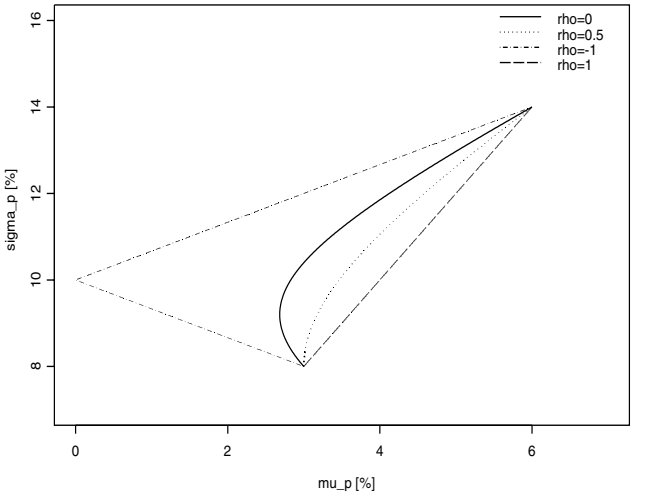


Fig. 4.5 – Courbes des portefeuilles possibles dans l'espace des risques et des rendements (exercice 4.17).

3. Le rendement moyen du portefeuille μ_p est une fonction de a_1 :

$$\mu_p = a_1\mu_1 + (1-a_1)\mu_2 \quad \Leftrightarrow \quad a_1 = \frac{\mu_p - \mu_2}{\mu_1 - \mu_2}.$$

On substitue cette dernière équation dans l'expression de σ_p^2 pour avoir

$$\begin{aligned}
\sigma_p^2 = & \\
&= (\sigma_1^2 - 2\sigma_1\sigma_2\rho_{12} + \sigma_2^2)\left(\frac{\mu_p - \mu_2}{\mu_1 - \mu_2}\right)^2 - 2(\sigma_2^2 - \sigma_1\sigma_2\rho_{12})\frac{\mu_p - \mu_2}{\mu_1 - \mu_2} + \sigma_2^2 \\
&= \frac{\sigma_1^2 - 2\sigma_1\sigma_2\rho_{12} + \sigma_2^2}{(\mu_1 - \mu_2)^2}\mu_p^2 - 2\frac{\mu_2\sigma_1^2 + \mu_1\sigma_2^2 - (\mu_1 - \mu_2)\sigma_1\sigma_2\rho_{12}}{(\mu_1 - \mu_2)^2}\mu_p \\
&\quad + \frac{\mu_2^2\sigma_1^2 + \mu_1^2\sigma_2^2 - 2\mu_1\mu_2\sigma_1\sigma_2\rho_{12}}{(\mu_1 - \mu_2)^2}
\end{aligned}$$

4. $\rho_{12} = 1$

a_1	0,00	0,20	0,40	0,60	0,80	1,00
μ_p	8,00	9,20	10,40	11,60	12,80	14,00
σ_p	3,00	3,60	4,20	4,80	5,40	6,00

$\rho_{12} = 0{,}5$

a_1	0,00	0,20	0,40	0,60	0,80	1,00
μ_p	8,00	9,20	10,40	11,60	12,80	14,00
σ_p	3,00	3,17	3,65	4,33	5,13	6,00

$\rho_{12} = -1$

a_1	0,00	0,20	0,40	0,60	0,80	1,00
μ_p	8,00	9,20	10,40	11,60	12,80	14,00
σ_p	3,00	1,20	0,60	2,40	4,20	6,00

5. Les courbes sont représentées dans la figure 4.5.

4.18

Soient N une variable aléatoire qui compte le nombre de clients par jour et X_i une autre qui mesure les dépenses en francs du i^e client. On sait d'ores et déjà que $E(N) = 50$ et $E(X_i) = 8$. Le chiffre d'affaire quotidien est $S = \sum_{i=1}^{N} X_i$. On ne peut en calculer directement l'espérance car la borne supérieure de sommation est une variable aléatoire. Pour contourner ce problème, on calcule d'abord l'espérance conditionnelle de $S \mid N = n$ et on en prend ensuite l'espérance par rapport à N

$$\begin{aligned} E_S(S) &= E_N(E_{S|N}(S \mid N)) = E_N\left(E_{S|N}\left(\sum_{i=1}^{n} X_i\right)\right) \\ &= E_N(NE(X_1)) = E(N)E(X_1) = 400, \end{aligned}$$

où l'on a utilisé l'indépendance entre X_1 et N. L'espérance du gain quotidien est de 400 €.

4.19

1. On veut trouver la distribution de la somme Z de 2 variables aléatoires binomiales indépendantes X et Y. Utilisons la fonction génératrice des moments

   $$\phi_Z(t) = \phi_X(t)\phi_Y(t) = (pe^t + 1 - p)^n (pe^t + 1 - p)^m = (pe^t + 1 - p)^{m+n},$$

   ce qui implique que Z suit une loi binomiale de paramètres $(m + n, p)$.
2. La probabilité d'observer $X = x \mid Z = z$ s'écrit

   $$P(X = x \mid Z = z) = \frac{P(X = x, Z = z)}{P(Z = z)}.$$

Comme Z et X ne sont pas indépendants, on utilise le fait que $Z = X+Y$ et on obtient

$$P(X = x \mid Z = z) = \frac{P(X = x)P(Y = z - x)}{P(Z = z)} = \frac{\left(\begin{array}{c} n \\ x \end{array} \right) \left(\begin{array}{c} m \\ z - x \end{array} \right)}{\left(\begin{array}{c} m+n \\ z \end{array} \right)},$$

d'où $X \mid Z$ suit une loi hypergéométrique.

3. L'espérance de $X \mid Z = z$ est

$$E(X \mid Z = z) = \frac{nz}{m+n},$$

et, par conséquent

$$E(X \mid Z) = \frac{n}{m+n} Z.$$

4.20

La fonction de répartition de Z est

$$\begin{aligned} F_Z(z) &= P(X + Y < z) = P(X < z - Y) \\ &= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{z-y} f_X(x)dx \right) f_Y(y)dy = \int_{-\infty}^{\infty} F_X(z-y) f_Y(y)dy, \end{aligned}$$

et sa dérivée

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)dy.$$

Ici, la difficulté réside dans les bornes de l'intégrale car, X et Y suivant des lois uniformes $(0, 1)$, l'intégrant est simplement égal à dy. On sait que Z peut varier entre 0 et 2 car $Z = X + Y$. Comme $0 < z - y < 1$ et $0 < y < 1$,

$$f_Z(z) = \begin{cases} \int_{z-1}^{0} 0dy + \int_{0}^{z} dy = z & \text{si } 0 \leq z < 1 \\ \int_{z-1}^{1} dy + \int_{1}^{z} 0dy = 2 - z & \text{si } 1 \leq z < 2 \\ 0 & \text{sinon.} \end{cases}$$

4.21

1. La distribution conditionnelle postulée signifie qu'il y a plus de volatilité en période d'incertitude.
2. Calculons l'espérance de Y en utilisant les espérances conditionnelles

$$\begin{aligned} E(Y) &= E(E(Y \mid X)) = \\ &= P(X = N) \cdot E(Y \mid X = N) + P(X = E) \cdot E(Y \mid X = E) \\ &= 0{,}95 \cdot 0 + 0{,}05 \cdot 0 = 0. \end{aligned}$$

Quant à la variance de Y

$$\begin{aligned}\text{var}(Y) &= E(Y^2) - E^2(Y) = E(Y^2) \\ &= P(X=N)\cdot E^2(Y \mid X=N) + P(X=E)\cdot E^2(Y \mid X=E) \\ &= 0{,}95\cdot 1 + 0{,}05\cdot 5^2 = 2{,}2.\end{aligned}$$

3. L'espérance de Z est

$$\begin{aligned}E(Z) &= E(\log(R)) = E\left(-\frac{\sigma^2}{2} + \sigma Y\right) \\ &= -\frac{\sigma^2}{2} + \sigma E(Y) = -\frac{\sigma^2}{2}\end{aligned}$$

et sa variance

$$\begin{aligned}\text{var}(Z) &= \text{var}(\log(R)) = \text{var}\left(-\frac{\sigma^2}{2} + \sigma Y\right) \\ &= \sigma^2 \text{var}(Y) = 2{,}2\sigma^2.\end{aligned}$$

4. En utilisant la formule des probabilités totales, on peut écrire

$$f_Y(y) = P(X=N)\cdot f_{Y|X=N}(y) + P(X=E)\cdot f_{Y|X=E}(y)$$

et comme on se retrouve avec une somme de 2 lois normales, on en déduit que

$$Y \sim \mathcal{N}(0,\, 2{,}2).$$

5. Nous savons que

$$Z = \log(R) \sim \mathcal{N}\left(-\frac{\sigma^2}{2},\, 2{,}2\sigma^2\right)$$

et donc

$$R = \exp(Z) \sim \log\mathcal{N}\left(-\frac{\sigma^2}{2},\, 2{,}2\sigma^2\right).$$

4.22

Soit N_i le nombre d'accidents durant l'année i pour un automobiliste et B un indicateur de la qualité de conduite. On sait que B suit une loi exponentielle avec espérance 0,1 et que $N_i \mid B = b$ suit une loi de Poisson de paramètre b. On suppose de plus que les $N_i \mid B = b$ sont indépendants.

1. On cherche la probabilité qu'un automobiliste n'ait pas d'accident pendant les 3 premières années d'observation

$$\begin{aligned}&P(N_1=0, N_2=0, N_3=0) = \\ &= \int_0^\infty P(N_1=0, N_2=0, N_3=0 \mid B=b) f_B(b) db \\ &= \int_0^\infty P(N_1=0 \mid B=b)^3 f_B(b) db.\end{aligned}$$

Lors de la dernière égalité, on a utilisé l'indépendance entre les $N_i \mid B = b$. On remplace maintenant les fonctions de distribution correspondantes pour obtenir

$$\begin{aligned} P(N_1 = 0, N_2 = 0, N_3 = 0) &= \int_0^\infty e^{-3b} \cdot 10e^{-10b} db \\ &= \int_0^\infty 10e^{-13b} db = -\frac{10}{13} e^{-b} \Big|_{b=0}^{b=\infty} = \frac{10}{13}. \end{aligned}$$

2. Si $K = k$, cela signifie que l'automobiliste n'a pas eu d'accident pendant les $k-1$ premières années et qu'il en a eu au moins un pendant la k^{e}. Donc

$$\begin{aligned} P(K = k) &= P(N_k \geq 1, N_{k-1} = 0, \ldots, N_1 = 0) \\ &= \int_0^\infty P(N_k \geq 1, N_{k-1} = 0, \ldots, N_1 = 0 \mid B = b) f_B(b) db \\ &= \int_0^\infty (1 - e^{-b}) e^{-b(k-1)} \cdot 10e^{-10b} db \\ &= \int_0^\infty 10 \left(e^{-(9+k)b} - e^{-(10+k)b} \right) db \\ &= 10 \left(-\frac{1}{9+k} e^{-(9+k)b} - \frac{1}{10+k} e^{-(10+k)b} \right) \Big|_{b=0}^{b=\infty} \\ &= \frac{10}{(9+k)(10+k)}. \end{aligned}$$

3. On peut éviter le calcul de l'espérance conditionnelle de $K \mid B = b$ en remarquant que $K \mid B = b$ suit une loi géométrique avec probabilité $p = 1 - e^{-b}$. On utilise alors simplement l'espérance d'une telle loi :

$$E(K \mid B = b) = \frac{1}{p} = \frac{1}{1 - e^{-b}}.$$

4.23

1. $E(Y(\alpha)) = E(\alpha D + (1-\alpha)N) = 10\alpha + (1-\alpha)3 = 7\alpha + 3$ et

$$\begin{aligned} \text{var}(Y(\alpha)) &= \text{var}(\alpha D) + \text{var}((1-\alpha)N) + 2\alpha(1-\alpha)\ \text{cov}\ (D,N) \\ &= 2\alpha^2 + (1-\alpha)^2 + 2\alpha(1-\alpha)\sqrt{2} \cdot 0{,}8 \\ &= 0{,}74\alpha^2 + 0{,}26\alpha + 1. \end{aligned}$$

2. La somme de variables aléatoires normales suit une loi normale. Donc, grâce au point 1., $Y(\alpha) \sim \mathcal{N}(7\alpha + 3,\ 3\alpha^2 - 2\alpha + 1)$.
3. On voit en 1^{er} lieu que $-Y(\alpha) \sim \mathcal{N}(-7\alpha - 3,\ 3\alpha^2 - 2\alpha + 1)$, et ensuite $Z = e^{-Y(\alpha)} \sim \log\mathcal{N}(-7\alpha - 3,\ 3\alpha^2 - 2\alpha + 1)$, par définition de la loi log-normale.

4. Par les propriétés des espérances conditionnelles

$$\begin{aligned} E(X) &= E(E(X \mid Y)) = E_Y(1 - 5Z + 4Z + 6Z + 4Z) \\ &= 1 + 9E(Z) \\ &= 1 + 9\exp(-7\alpha - 3 + \tfrac{3\alpha^2 - 2\alpha + 1}{2}) \\ &= 1 + 9\exp(\tfrac{3}{2}\alpha^2 - 8\alpha - \tfrac{5}{2}). \end{aligned}$$

5. Par la formule des probabilités totales

$$\begin{aligned} P(X = 4) &= \int_{-\infty}^{+\infty} P(X = 4 \mid y) f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} e^{-y} f_Y(y) dy = E(e^{-Y}) = M_Y(-1) \\ &= \exp\left(-7\alpha - 3 + \tfrac{3\alpha 2 - 2\alpha + 1}{2}\right) \\ &= \exp\left(\tfrac{3}{2}\alpha^2 - 8\alpha - \tfrac{5}{2}\right), \end{aligned}$$

où $M_Y(\cdot)$ est la fonction génératrice des moments d'une variable aléatoire $\mathcal{N}(7\alpha + 3, 3\alpha^2 - 2\alpha + 1)$.

4.24

Soit X_i la variable aléatoire de Bernoulli qui vaut 1 si on trouve du pétrole dans la localité i et 0 sinon, avec $i = 1, \ldots, 10$. La variable aléatoire $N = \sum_{i=1}^{10} X_i$ suit alors une distribution binomiale de paramètres $(10, 0{,}2)$. On sait de plus que le profit dans la localité P_i, si on y a trouvé du pétrole, suit une loi exponentielle d'espérance 5.

1. On écrit le profit total Z comme

$$Z = \sum_{i=1}^{N} P_i.$$

Notons que la borne supérieure de la somme est une variable aléatoire et que $Z \mid N = n$ suit une loi Gamma de paramètres $(1/5, n)$ car c'est une somme de variables suivant une loi exponentielle.

2. La variable aléatoire $Z \mid N = 1$ a une distribution exponentielle d'espérance 5, donc

$$P(Z > 10 \mid N = 1) = P(P_1 > 10) = 1 - F_{P_1}(10) = e^{-2} \simeq 0{,}1.3$$

Comme mentionné dans le point 1., la variable aléatoire $Z \mid N = 2$ suit

une distribution Gamma de paramètres $(5, 2)$. Par conséquent

$$\begin{aligned} P(Z > 10 \mid N = 2) &= P\left(\sum_{i=1}^{2} P_i > 10\right) = \\ &= 1 - \int_0^{10} \frac{1}{25} \exp\left(-\frac{1}{5}x\right) x dx \\ &= 1 - \frac{1}{25}\left(\left(-5x \exp\left(-\frac{1}{5}x\right)\right)\Big|_{x=0}^{x=10} + 5\int_0^{10} \exp\left(-\frac{1}{5}x\right) dx\right) \\ &= 3e^{-2} \simeq 0{,}41, \end{aligned}$$

où l'on a utilisé le théorème d'intégration par parties.

3. $$E(Z \mid N = n) = 5n$$

et

$$E(Z) = E_N(E_Z(Z \mid N)) = E(5N) = 5E(N) = 10.$$

4. Par l'inégalité de Markov

$$P(Z > 20) \leq \frac{E(Z)}{20} = 0{,}5.$$

La probabilité que le profit soit supérieur à 20 Mio€ n'est pas supérieure à 0,5.

4.25

1. Par le théorème central limite

$$S \mid N(t) \overset{approx.}{\sim} \mathcal{N}\left(\frac{N(t)}{\mu}, \frac{N(t)}{\mu^2}\right).$$

2. L'espérance de S est

$$E(S) = E(E(S \mid N(t))) = E\left(\frac{N(t)}{\mu}\right) = \frac{\lambda}{\mu}t$$

et sa variance

$$\begin{aligned} \text{var}(S) &= E(\text{var}(S \mid N(t))) + \text{var}(E(S \mid N(t))) \\ &= E\left(\frac{N(t)}{\mu^2}\right) + \text{var}\left(\frac{N(t)}{\mu}\right) \\ &= \frac{1}{\mu^2}\lambda t + \frac{1}{\mu^2}\lambda t = \frac{2\lambda t}{\mu^2}. \end{aligned}$$

3. Par définition des probabilités totales

$$f_S(s) = \sum_{n \in \mathbb{N}} f_{S|N(t)}(s \mid n) \cdot P(N(t) = n).$$

4.

$$\begin{aligned} P(S > s_0) &= \int_{s_0}^{\infty} f_S(s)ds = \sum_n P(N(t) = n) \cdot \int_{s_0}^{\infty} f_{S|N(t)}(s|n)ds \\ &= \sum_n \left[1 - \Phi\left(\frac{s_0 - \frac{n}{\mu}}{\frac{\sqrt{n}}{\mu}}\right)\right] \cdot P(N(t) = n) \\ &= \sum_n \left[1 - \Phi\left(\frac{s_0 \cdot \mu}{\sqrt{n}} - \sqrt{n}\right)\right] \frac{(\lambda t)^n}{n!} \cdot e^{-\lambda t}. \end{aligned}$$

Chapitre 5

Théorèmes limites

Introduction

Ce chapitre a trait à trois résultats importants de la théorie asymptotique des probabilités : la loi faible des grands nombres, la loi forte des grands nombres et le théorème central limite, dans sa version pour variables aléatoires indépendantes et identiquement distribuées ($X_i \sim F_\theta$ pour $i = 1, \ldots, n$). Ce sont des résultats qui traitent les propriétés de la distribution de la moyenne d'une suite de variables aléatoires ($\bar{X}_n$).

Les deux lois des grands nombres énoncent les conditions sous lesquelles la moyenne d'une suite de variables aléatoires converge vers leur espérance commune et expriment l'idée que lorsque le nombre d'observations augmente, la différence entre la valeur attendue ($\mu = E(X_i)$) et la valeur observée ($\bar{X}_n$) tend vers zéro. De son côté, le théorème central limite établit que la distribution standardisée d'une moyenne tends asymptotiquement vers une loi normale, et cela même si la distribution des variables sous-jacente est non normale. Ce résultat est central en probabilités et statistique et peut être facilement illustré (cf. figure 5.1). Indépendamment de la distribution sous-jacente des observations (ici une loi uniforme), lorsque n croît, la distribution de $\bar{X}_n$ tend vers une loi normale : on observe dans l'illustration la forme de plus en plus symétrique de la distribution ainsi que la concentration autour de l'espérance (ici $\mu = 0{,}5$) et la réduction de la variance.

Les retombées pratiques de ces résultats sont importantes. En effet, la moyenne de variables aléatoires est une quantité qui intervient dans plusieurs procédures statistiques. Aussi, le résultat du théorème central limite permet l'approximation des probabilités liées à des sommes de variables aléatoires. De plus, lorsque l'on considère des modèles statistiques, le terme d'erreur représente la somme de beaucoup d'erreurs (erreurs de mesure, variables non considérées, etc.). En prenant comme justification le théorème central limite, ce terme d'erreur est souvent supposé se comporter comme un loi normale.

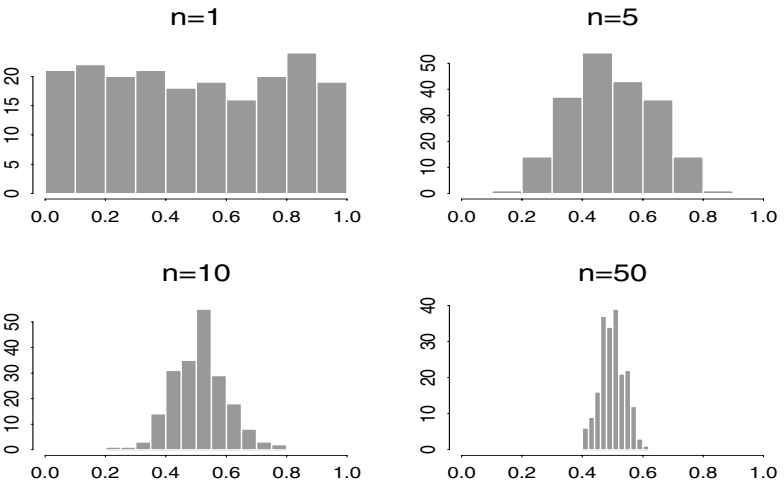


Fig. 5.1 – Illustration du théorème central limite : histogramme de la moyenne de 200 échantillons issus d'une loi uniforme sur l'intervalle $(0, 1)$ en fonction de la taille n de l'échantillon.

Les trois théorèmes

Loi faible des grands nombres
Soient $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées. On suppose que $E(|X_i|) < \infty$ et que touts les X_i admettent la même espérance $E(X_i) = \mu$. Pour tout $\epsilon > 0$

$$\lim_{n\to\infty} P\left(\mid \bar{X}_n - \mu \mid > \epsilon\right) = 0,$$

c'est-à-dire $\bar{X}_n$ converge en probabilité vers μ, ce qui en économétrie est souvent noté plim $\bar{X}_n = \mu$.

Loi forte des grands nombres
Soient $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées. On suppose que $E(|X_i|) < \infty$ et que touts les X_i admettent la même espérance $E(X_i) = \mu$. Alors, pour tout $\epsilon > 0$

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1.$$

On dit que $\bar{X}_n$ converge presque sûrement vers μ.

Théorème central limite
Soient $X_1, \ldots, X_n$ une suite de variables aléatoires indépendantes et identiquement distribuées, d'espérance μ et variance σ^2 finie. Alors

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \quad \stackrel{n\to\infty}{\longrightarrow} \quad \mathcal{N}(0,1), \tag{5.1}$$

en distribution.

On voit bien que, afin que la convergence se fasse, une standardisation est nécessaire : en effet, on peut voir le rapport dans (5.1) comme

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{\mathrm{var}(X_n)}}.$$

Notes historiques
La loi faible des grands nombres a été établie la première fois par J. Bernoulli pour le cas particulier d'une variable aléatoire binaire ne prenant que les valeurs 0 ou 1. Le résultat a été publié en 1713.

La loi forte des grands nombres est due au mathématicien E. Borel (1871-1956), d'où parfois son autre appellation : théorème de Borel.

Le théorème central limite a été formulé pour la première fois par A. de Moivre en 1733 pour approximer le nombre de « piles » dans le jet d'une pièce de monnaie équilibrée. Ce travail a été un peu oublié jusqu'à ce que P.S. Laplace ne l'étende à l'approximation d'une loi binomiale par la loi normale dans son ouvrage *Théorie analytique des probabilités* en 1812. C'est dans les premières années du XX^{e} siècle que A. Lyapounov l'a redéfini en termes généraux et prouvé avec rigueur.

Références (théorie)

Ross, chapitre 8 [1]; Lejeune, chapitre 5.8 [3], et Morgenthaler, chapitre 5 [4].

Exercices

5.1

Le nombre d'inscriptions à un cours d'économie politique est une variable aléatoire de Poisson de paramètre 100. Le professeur donnant ce cours a décidé que si le nombre d'inscriptions est au-delà de 120, il créera 2 sections et donnera donc 2 cours, tandis qu'en deçà une seule classe sera formée.
Quelle est la probabilité que ce professeur ait à donner 2 fois ce cours?

5.2

Supposons qu'on ait lancé 10 000 fois une pièce de monnaie bien équilibrée.

1. Trouver un intervalle symétrique autour de 5 000 où l'on puisse dire que le nombre de pile y appartient avec une probabilité supérieure à 0,99.
2. Comparer le résultat avec celui obtenu en appliquant l'inégalité de Chebychev.

5.3

Soient $X_1, \ldots, X_{1\ 000}$ des variables aléatoires suivant la loi uniforme sur l'intervalle $(0, 1)$. Soit N le nombre d'entre elles inférieures à 0,003. Calculer $P(N < 5)$ en utilisant l'approximation par la loi de Poisson. Soit M le nombre d'entre elles comprises entre $1/4$ et $3/4$. Déterminer par approximation normale $P(\mid M - 500 \mid > 20)$.

5.4

Christian vient d'ouvrir un salon de jeux. Pour l'instant, il n'a installé qu'une roulette et a simplifié le jeu de la manière suivante : on ne peut miser que 1 euro sur 1 seul des 37 numéros (de 0 à 36) à la fois. En cas de gain, Christian donne 35 €.

Si, après un mois, 5 000 personnes ont misé 1 €, quelle est la probabilité que Christian perde de l'argent ?

5.5

Quentin et ses 9 amis voudraient aller au cinéma. Ils décident de rassembler leur argent de poche et espèrent obtenir la somme totale nécessaire.

On peut supposer que l'argent de poche de chacun est une variable aléatoire X_i qui suit la loi exponentielle de paramètre $\lambda = 0{,}06$. Sa densité est

donc

$$f(x) = \begin{cases} 0{,}06\exp(-0{,}06x) & x \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

De plus, on admet que les X_i sont indépendants.

1. Écrire la loi exponentielle $\mathcal{E}(\lambda)$ comme une loi Gamma en donnant les paramètres de celle-ci.
2. Soit $S_{10} = \sum_{i=1}^{10} X_i$. Quelle est la densité de S_{10} ?
3. Sachant qu'un ticket de cinéma coûte 15 €, quelle est la probabilité que Quentin est ses amis puissent aller au cinéma?
4. Comment faut-il choisir $z > 0$ pour que la probabilité que la somme totale d'argent du groupe soit supérieur à z soit égale à 5 % ?

5.6

La moyenne des revenus annuels des employés d'une grande banque est égale à 50 000 €.

1. Donner une borne supérieure pour le pourcentage p des revenus supérieurs ou égaux à 80 000 €.
2. De plus, on sait que l'écart-type est égal à 10 000 €. Donner une borne supérieure plus petite pour p en utilisant cette information supplémentaire.

5.7

Nous avons 100 composants que nous allons employer les uns après les autres. Cela veut dire que le composant 1 sera d'abord utilisé, puis lorsqu'il tombera en panne, il sera remplacé par le composant 2, qui sera lui-même remplacé par le composant 3, et ainsi de suite. Si la durée de vie du composant i est distribuée selon une loi exponentielle avec espérance (en heures) $10 + i/10$ pour $i = 1, \ldots, 100$, et si les durées de vie sont indépendantes, estimer la probabilité que la durée de vie totale de l'ensemble des composants dépasse 1 200 heures.

Indication : $\displaystyle\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ et $\displaystyle\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$.

5.8

Samuel et Sonia jouent au jeu suivant. Chacun est appelé à lancer 100 fois un dé équilibré et à mesurer sa propre performance par le produit des scores obtenus à chaque lancer.
Sonia commence à jouer et totalise le score $\prod_{i=1}^{100} x_i = 4^{100}$. Soit Y_i la variable aléatoire qui représente la valeur obtenue par Samuel au i^{e} lancer. Donner une

approximation de la probabilité que Samuel obtienne un score plus élevé que Sonia, c'est-à-dire la probabilité

$$P\left(\prod_{i=1}^{100} Y_i > 4^{100}\right).$$

5.9

Certaines particules subissent des collisions qui causent leur division en 2 morceaux, chacun étant une fraction de la particule de départ. Supposons que la fraction X est distribuée uniformément sur l'intervalle $(0, 1)$. En suivant une seule particule après n divisions, on obtient une fraction de la particule de départ qu'on appelle $Z_n = X_1 \cdot X_2 \cdot \ldots \cdot X_n$ où chaque X_j est distribué uniformément sur l'intervalle $(0, 1)$. On veut trouver la distribution de Z_n.

1. Montrer que la variable aléatoire $Y_k = -\log(X_k)$ a une distribution exponentielle.
2. Utiliser le résultat précédent pour trouver la distribution de $S_n = Y_1 + Y_2 + \ldots + Y_n$. Écrire la fonction de densité de S_n.
3. Exprimer Z_n en fonction de S_n et montrer que la fonction de densité de Z_n est
$$f_n(z) = \frac{1}{(n-1)!}(-\log z)^{n-1}.$$

5.10

On génère à l'ordinateur 100 000 nombres aléatoires $u_1, \ldots, u_{100\ 000}$ selon une loi uniforme $(0, 1)$ et on calcule leur moyenne géométrique

$$(u_1 \cdot u_2 \cdot \ldots \cdot u_{100\ 000})^{1/100\ 000}.$$

Cette valeur sera très proche d'un certain nombre a. Calculer a et justifier votre réponse.

Corrigés

5.1

Soit X le nombre d'inscriptions à un cours d'économie politique. La variable aléatoire X suit une loi de Poisson d'espérance égale à 100. On cherche la probabilité qu'il y ait plus de 120 étudiants inscrits, soit

$$P(X > 120) = e^{-100} \sum_{i=121}^{\infty} \frac{100^i}{i!}.$$

Plutôt que de s'exposer à de grosses erreurs numériques dues à la taille énorme du numérateur et du dénominateur, il est possible d'utiliser l'approximation normale (théorème central limite) pour calculer cette probabilité

$$P(X > 120) \simeq P\left(\frac{X - 100}{\sqrt{100}} > \frac{120 - 100}{\sqrt{100}}\right) = 1 - \Phi(2) \simeq 0{,}023.$$

Ainsi, la probabilité que le professeur doive donner son cours 2 fois est environ égale à 2,3 %.

5.2

Soit X le nombre de fois où la pièce donne pile. La variable aléatoire X a une distribution binomiale de paramètres $(10\,000, 0{,}5)$. Son espérance est égale à $E(X) = np = 5\,000$ et sa variance à $\text{var}(X) = np(1-p) = 2\,500$.

1. Pour calculer un intervalle autour de 5 000 avec les bonnes propriétés, on utilise l'approximation normale :

$$\begin{aligned} &P\left(-z_{1-\alpha/2} < \frac{X - 5\,000}{50} < z_{1-\alpha/2}\right) = \\ &\quad = \quad P(-50z_{1-\alpha/2} + 5\,000 < X < 50z_{1-\alpha/2} + 5\,000) = 0{,}99. \end{aligned}$$

   On en déduit que l'intervalle $[4\,871\,,5\,129]$ contiendra le nombre de piles avec 99 % de probabilité.

2. Par l'inégalité de Chebychev, on a

$$P\left(\mid X - E(X) \mid \geq \varepsilon\right) \leq \frac{\text{var}(X)}{\varepsilon^2},$$

   ou plutôt dans le cas présent

$$P(-\varepsilon < X - E(X) < \varepsilon) > 1 - \frac{\text{var}(X)}{\varepsilon^2}.$$

On remplace alors ε par $\sigma(x) \cdot z_{1-\alpha/2}$, avec $\sigma(x)^2 = \text{var}(X)$, pour obtenir

$$\begin{aligned} P(E(X) - \sigma(X) \cdot z_{1-\alpha/2} < X > E(X) + \sigma(X) \cdot z_{1-\alpha/2}) & \\ > 1 - \frac{\text{var}(X)}{(\sigma z_{1-\alpha/2})^2} & \\ = 1 - \frac{1}{z_{1-\alpha/2}^2}. & \end{aligned}$$

Ainsi, puisque la probabilité précédente est supérieure à 0,99, $z_{1-\alpha/2}$ vaut au maximum 10 et l'intervalle de confiance est inclus dans $[4\ 500\,,5\ 500]$.

5.3

1. Soit N le nombre de réalisations de X inférieures à 0,003. La variable aléatoire N suit une loi binomiale de paramètres $(1\ 000\,, 0{,}003)$ que l'on l'approxime par une loi de Poisson d'espérance $\lambda = np = 3$. On calcule donc la probabilité

$$P(N < 5) = \sum_{k=0}^{4} \frac{\lambda^k e^{-\lambda}}{k!} \simeq 0{,}82.$$

   Il y a environ 82 % de chance que moins de 5 réalisations de la variable aléatoire X soient inférieures à 0,003.

2. Soit M le nombre des mêmes réalisations, mais cette fois-ci comprises entre 0,25 et 0,75. La variable aléatoire M a une distribution binomiale $(1\ 000\,, 0{,}5)$, son espérance vaut 500 et sa variance 250. On trouve alors

$$\begin{aligned} P(\mid M - 500 \mid > 20) &= 1 - P(-20 < M - 500 < 20) \\ &= 1 - P\left(-\frac{20}{\sqrt{250}} < \frac{M - 250}{\sqrt{250}} < \frac{20}{\sqrt{250}}\right) \\ &= 1 - P\left(-\frac{4}{\sqrt{10}} < Z < \frac{4}{\sqrt{10}}\right). \end{aligned}$$

   Par le théorème central limite, on fait l'approximation que Z suit une loi normale centrée et réduite. Par conséquent

$$P(\mid M - 500 \mid > 20) = 2 - 2P\left(Z < \frac{4}{\sqrt{10}}\right) \simeq 0{,}21,$$

   ce qui signifie que la probabilité qu'il y ait plus de 20 réalisations de X comprises entre 0,25 et 0,75 est d'environ 21 %.

5.4

Soit G_i le gain de Christian sur le joueur i. On cherche la probabilité que la somme des gains $S = \sum_{i=1}^{5\ 000} X_i$ sur 5 000 personnes soit inférieure à 0. Comme la taille de l'échantillon est importante, on utilise le théorème central limite pour approcher la distribution de S ; il nous faut donc trouver l'espérance et la variance de G_i

$$\begin{aligned} E(G_i) = P(\text{Christian gagne}) \cdot 1 + P(\text{Christian perd}) \cdot (-35) = \\ = \frac{36}{37} - \frac{1}{37} \cdot 35 = \frac{1}{37} \simeq 0{,}027, \end{aligned}$$

$$\text{var}(G_i) = E(G_i^2) - E^2(G_i) = \frac{36}{37} + \frac{1}{37} \cdot 35^2 - \frac{1}{37^2} \simeq 34.$$

On trouve donc

$$\begin{aligned} P\left(S < 0\right) &= P\left(\frac{S - nE(G_i)}{\sqrt{n\text{var}(G_i)}} < \frac{-nE(G_i)}{\sqrt{n\text{var}(G_i)}}\right) \\ &\simeq \Phi\left(-\frac{\sqrt{n}E(G_i)}{\sqrt{\text{var}(G_i)}}\right) \simeq 1 - \Phi(0{,}33) \simeq 0{,}37. \end{aligned}$$

Il y a 37 % de chance que Christian perde de l'argent.

5.5

1. La loi exponentielle $\mathcal{E}(\lambda)$ est une loi Gamma de paramètres $(\lambda, 1)$.
2. Comme S_{10} est la somme de 10 variables aléatoires qui suivent toutes la même loi exponentielle $\mathcal{E}(\lambda)$ et sont indépendantes, S_{10} suit une Gamma de paramètres $(\lambda, 10)$.
3. Afin de pouvoir calculer la probabilité que Quentin et ses amis puissent aller au cinéma, on utilise l'approximation normale

   $$P(S_{10} > 150) = 1 - P\left(\frac{S_{10} - \frac{10}{0{,}06}}{\frac{\sqrt{10}}{0{,}06}} < \frac{150 - \frac{10}{0{,}06}}{\frac{\sqrt{10}}{0{,}06}}\right) \simeq 1 - \Phi(-0{,}32) \simeq 0{,}63.$$

   La probabilité que Quentin et ses amis aillent au cinéma est d'environ 63 %.
4. On cherche z qui satisfait $P(S_{10} > z) = 0{,}05$. Ainsi

   $$P(S_{10} > z) = P\left(Z < \frac{z - \frac{10}{0{,}06}}{\frac{\sqrt{10}}{0{,}06}}\right) = 0{,}05,$$

   où $Z \sim \mathcal{N}(0{,}1)$. On en déduit finalement que

   $$\frac{0{,}06z - 10}{\sqrt{10}} = z_{0{,}95} \simeq 1{,}64 \quad \Leftrightarrow \quad z \simeq \frac{1{,}64\sqrt{10} + 10}{0{,}06} \simeq 253.$$

Il y a environ 5 % de chance que la somme du groupe soit supérieure à 253.

5.6

Soit X le revenu annuel d'un employé d'une grande banque et $E(X) = 50\ 000$ son espérance.

1. La seule information à disposition étant l'espérance de X, on utilise l'inégalité de Markov pour trouver une borne supérieure pour le pourcentage p des revenus supérieurs ou égaux à 80 000 €:

$$p = P(X > 80\ 000) \leq \frac{50\ 000}{80\ 000} = \frac{5}{8}.$$

   Moins de 63 % des employés de la banque gagnent plus que 80 000 €.

2. On connaît à présent l'écart-type de X et cela permet de calculer une borne supérieure pour p plus précise avec l'inégalité de Chebychev :

$$\begin{aligned} p &= P(X > 80\ 000) = P(X - 50\ 000 > 30\ 000) = \\ &= P(\mid X - 50\ 000 \mid > 30\ 000) \leq \frac{10\ 000^2}{30\ 000^2} = \frac{1}{9}. \end{aligned}$$

   Il y a au maximum 11 % des employés qui gagnent plus que 80 000 €. Dans le cas présent, la borne supérieure a été surestimée par l'inégalité de Markov.

5.7

Soit D_i la durée de vie du composant i. La variable aléatoire D_i suit une loi exponentielle d'espérance $10 + i/10$ heures. On s'intéresse à la probabilité que $S = \sum_{i=1}^{100} D_i$ soit supérieur à 1 200. L'échantillon étant grand, on applique le théorème central limite et il faut par conséquent calculer l'espérance et la variance de S

$$\begin{aligned} E(S) &= E\left(\sum_{i=1}^{100} D_i\right) = \sum_{i=1}^{100} E(D_i) = \sum_{i=1}^{100} \left(10 + \frac{i}{10}\right) \\ &= 1\ 000 + \frac{1}{10}\frac{100 \cdot 101}{2} = 1\ 505, \end{aligned}$$

et

$$\begin{aligned} \mathrm{var}(S) &= \sum_{i=1}^{100} \mathrm{var}(D_i) = \sum_{i=1}^{100} \left(10 + \frac{i}{10}\right)^2 = \sum_{i=1}^{100} \left(100 + 2i + \frac{i^2}{100}\right) \\ &= 10\ 000 + 2\frac{100 \cdot 101}{2} + \frac{1}{100}\frac{100 \cdot 101 \cdot 201}{6} \simeq 23\ 484. \end{aligned}$$

On utilise donc le théorème central limite pour trouver

$$P(S > 1\ 200) = P\left(Z > \frac{1\ 200 - 1\ 505}{\sqrt{23\ 484}}\right) \simeq \Phi(1{,}99) \simeq 0{,}98\ ;$$

La durée de vie de l'ensemble des composants a approximativement 98 % de chance d'être supérieure à 1 200 heures.

5.8

Comme le dé est lancé un grand nombre de fois, on aimerait pouvoir utiliser le théorème central limite pour calculer une approximation de la performance des joueurs. On transforme donc le produit des lancers par un logarithme et on définit ainsi

$$S_Y = \log\left(\prod_{i=1}^{n} Y_i\right) = \sum_{i=1}^{100} \log(Y_i).$$

L'espérance et la variance de S_Y s'écrivent

$$E(S_Y) = \sum_{i=1}^{100} E(\log(Y_i)) = 100 \sum_{j=1}^{6} \frac{1}{6} \log(j) = \frac{50}{3} \log(6!) \simeq 110,$$

et

$$\begin{aligned} \operatorname{var}(S_Y) &= \sum_{i=1}^{100} \operatorname{var}(\log(Y_i)) = 100\Big(E((\log(Y_1))^2) - E^2(\log(Y_1))\Big) \\ &= 100\left(\sum_{j=1}^{6} \frac{1}{6} \log^2(j) - \left(\frac{50}{3} \log(6!)\right)^2\right) \simeq 36{,}6. \end{aligned}$$

On utilise maintenant le théorème central limite pour trouver la probabilité que Samuel obtienne le meilleur score

$$\begin{aligned} P\left(\prod_{i=1}^{100} Y_i > 4^{100}\right) &= P\left(S_Y > 100 \log(4)\right) = P\left(Z > \frac{100 \log(4) - E(S_Y)}{\sqrt{\operatorname{var}(S_Y)}}\right) \\ &\simeq 1 - \Phi\left(\frac{100 \log(4) - 110}{\sqrt{36{,}6}}\right) \simeq 1 - \Phi(4{,}79) \simeq 0, \end{aligned}$$

où $Z \sim \mathcal{N}(0,1)$. Il est donc approximativement certain que Samuel ne batte pas Sonia.

5.9

1. Soit $Y_k = -\log(X_k)$. Sa fonction de répartition est

$$\begin{aligned} F_{Y_k}(y) &= P(Y_k < y) = P(-\log X_k < y) = \\ &= P\left(X_k > e^{-y}\right) = 1 - F_{X_k}\left(e^{-y}\right) = 1 - e^{-y}. \end{aligned}$$

Elle correspond à celle d'une variable aléatoire suivant une loi exponentielle d'espérance égale à 1.

2. La variable aléatoire S_n étant la somme de n variables aléatoires exponentielles de paramètre 1, elle suit une loi Gamma de paramètres $(1, n)$ dont la fonction de densité est

$$f_{S_n}(s) = \frac{1}{(n-1)!} e^{-s} s^{n-1}.$$

3. On trouve d'abord

$$Z_n = \prod_{i=1}^{n} X_i = \exp\left(-\sum_{i=1}^{n} \log(X_i)\right) = \exp\left(\sum_{i=1}^{n} Y_i\right) = \exp(-S_n).$$

On applique ensuite ce changement de variable sur la fonction de densité de S_n

$$f_{Z_n}(z) = f_{S_n}(-\log(z)) \left| \frac{\partial}{\partial z}(-\log(z)) \right| = \frac{1}{(n-1)!}(-\log(z))^{n-1}.$$

5.10

Soit $n = 100\ 000$. On cherche à calculer la moyenne géométrique σ_n de 100 000 nombres aléatoires u_i distribués selon une loi uniforme $(0, 1)$.

$$\sigma_n = (u_1 \cdot \ldots \cdot u_n)^{1/n}$$

Afin de trouver la valeur de σ_n, on en prend le logarithme

$$\log \sigma_n = \frac{1}{n} \sum_{i=1}^{n} \log(u_i).$$

Comme n est grand, on utilise la loi des grands nombres :

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log(u_i) = E(\log(u_1)).$$

Ainsi

$$E(\log(u_1)) = \int_0^1 \log(u_1) du_1 = \Big[\log(u_1) \cdot u_1\Big]_{u_1=0}^{u_1=1} - \int_0^1 du = -1$$

et

$$\sigma_n \simeq e^{-1}.$$

Chapitre 6

Principes d'induction statistique et échantillonnage

Introduction

À partir de ce chapitre, on quitte le monde des probabilités pour rentrer dans le monde de la statistique, où les résultats de probabilités sont un outil indispensable.

La démarche statistique consiste à utiliser l'information obtenue sur un échantillon pour pouvoir déduire de l'information sur la population (ou l'univers) d'intérêt (cf. illustration à la figure 6.1) : on *extrait* un échantillon de la population, on l'analyse et on *infère* sur la population.

La démarche peut se décrire en trois étapes :

1. choix d'un échantillon ;
2. estimation de quantités ponctuelles (chapitre 7) ;
3. estimation par intervalle et tests (chapitre 8).

La méthode du choix de l'échantillon est cruciale. Les meilleures méthodes sont basées sur l'introduction contrôlée du hasard. Cela permet d'éliminer les jugements subjectifs et le biais de sélection. Le cas le plus simple est appelé échantillonnage aléatoire simple et consiste à tirer de façon équiprobable n individus à partir de la population. Des schémas plus sophistiqués et efficaces qui contrôlent le biais peuvent être considérés, mais ne seront pas utilisés ici (échantillonnage stratifié, par grappe ...).

La qualité de l'information que l'on peut tirer d'un échantillon dépend de la variabilité sous-jacente des données et de la taille de l'échantillon, mais ne dépend quasiment pas de la taille de la population. La théorie de l'échantillon-

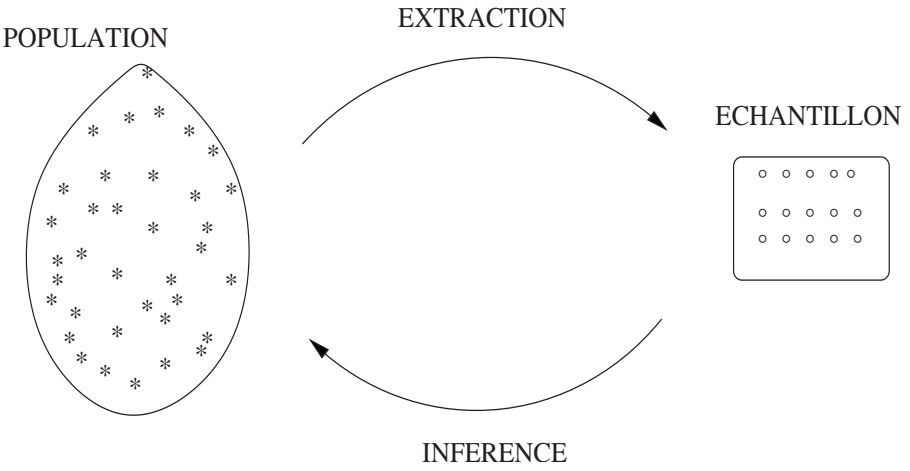


Fig. 6.1 – Illustration de l'échantillonnage.

nage permet de déterminer la taille de l'échantillon nécessaire pour atteindre un certain niveau de précision dans l'estimation qui en suivra.

Une fois l'échantillon extrait de la population, on répond aux questions d'intérêt (quelle est la médiane des revenus en Suisse, est-ce que cette valeur est égal à 60 000 francs, est-ce que le médicament A est plus efficace que le médicament B, etc.). La réponse statistique découle d'une approche probabiliste.

Notes historiques

L'utilisation de données partielles (l'échantillon) au lieu de l'information exhaustive (toute la population) a fait l'objet d'un grand débat au début du XX^{e} siècle. C'est le congrès de l'Institut International de Statistique en 1925 qui marque la reconnaissance officielle de la théorie des sondages. L'article de Neyman en 1934 est considéré comme un des textes fondateurs de la théorie des sondages. Ensuite la théorie des sondages se développe très rapidement grace aux contributions de Deming, Stephan, Cochran en particulier dans la première moitié du XX^{e} siècle. Les recherches se sont ensuite poursuivies en faisant appel à la statistique mathématique.

Références (théorie)

Dodge, chapitre 10 [5]; Ronchetti, Antille et Polla, chapitre 7 [6]; et Tillé, plus particulièrement chapitre 4 [7].

Exercices

6.1

S'agit-il de variables aléatoires?

1. Moyenne de la population.
2. Taille de la population.
3. Taille de l'échantillon.
4. Moyenne de l'échantillon.
5. Variance de la moyenne de l'échantillon.
6. Plus grande valeur de l'échantillon.
7. Variance de la population.
8. Variance estimée de la moyenne de l'échantillon.

6.2

Afin d'estimer leur espérance respective, on échantillonne 2 populations. On utilise un échantillon de taille n_1 pour la population I, qui présente un écart-type égal à σ_1. Pour la population II, dont l'écart-type vaut $\sigma_2 = 2\sigma_1$, on prend un échantillon de taille $n_2 = 2n_1$. Pour lequel des 2 échantillons est-ce que l'estimation de la moyenne de la population est la plus précise?

6.3

Le petit David a des problèmes de dyslexie. Des études montrent qu'un enfant atteint de dyslexie a une probabilité de 0,2 de ne pas obtenir 10 en comptant ses doigts. Pour vérifier empiriquement ce résultat, on décide de faire compter ses doigts à David. Quelle doit être la taille n de l'échantillon pour que l'écart-type de l'estimateur soit égal à 0,05?

6.4

Dans un certain canton suisse on veut estimer la proportion de familles vivant en dessous du seuil de pauvreté. Si cette proportion est environ 0,15, quelle est la taille de l'échantillon nécessaire pour que l'écart-type de l'estimateur soit égal à 0,02?

6.5

Dans un certain État américain on suppose que l'électorat est composé à 80 % de population urbaine et à 20 % de population rurale. Une proportion de

70 % des gens de la ville et seulement 25 % des gens de la campagne préfèrent le candidat D au candidat C. Dans un sondage effectué par l'éditeur d'un journal d'une petite ville, un électeur de la campagne a 6 fois plus de chance d'être choisi qu'un électeur de la ville. De ce fait, la proportion de l'échantillon en faveur de C sera un estimateur biaisé de la proportion de la population.

1. Combien vaut ce biais ?
2. Est-ce que le biais est suffisamment grand pour que la moyenne de l'échantillon soit fausse (dans le sens où la moyenne de l'échantillon « élit » un candidat différent que celui que la population choisit) ?

6.6

Un chercheur récolte un échantillon de 500 observations, mais perd les 180 dernières mesures. Il n'a donc plus que 320 observations pour calculer la moyenne de l'échantillon. Quelle est l'efficacité de son estimation par rapport à celle qu'il aurait pu obtenir avec les 500 observations ?

Corrigés

6.1

Dans une population, les paramètres tels que sa taille (N), sa moyenne (μ) ou sa variance (σ^2) sont connus et ne sont donc pas des variables aléatoires. De cette population, un nombre fixé (n) de variables aléatoires ($X_1, \ldots, X_n$) est tiré et constitue un échantillon. Toute quantité dépendant de ces variables aléatoires est également une variable aléatoire. Les réponses de l'exercice sont alors :

1. non (μ) ;
2. non (N) ;
3. non (n) ;
4. oui ($\overline{X} = \sum_{i=1}^{n} X_i$) ;
5. non (σ^2/n) ;
6. oui ($\max(X_i)$) ;
7. non (σ^2) ;
8. oui ($s^2(X_1, \ldots, X_n)/n$).

6.2

Afin de déterminer quel échantillon donnera une estimation de la moyenne la plus précise, on calcule la variance de chacune des moyenne. En général, on a

$$\mathrm{var}(\overline{X}) = \frac{\sigma^2}{n},$$

ce qui donne dans le cas présent

$$\mathrm{var}(\overline{X}_1) = \frac{\sigma_1^2}{n_1}$$

et

$$\mathrm{var}(\overline{X}_2) = \frac{\sigma_2^2}{n_2} = \frac{4\sigma_1^2}{2n_1} = 2\frac{\sigma_1^2}{n_1} = 2\mathrm{var}(\overline{X}_1).$$

Par conséquent, l'estimation de la moyenne est plus précise dans le premier échantillon. Si l'écart-type est 2 fois plus grand, il faut quadrupler la taille de l'échantillon pour obtenir la même précision dans l'estimation de l'espérance.

6.3

On sait que dans la population des enfants atteints de dyslexie, 20 % n'arrivent pas à compter leurs doigts. On s'intéresse à un échantillon construit

sur la répétition de l'expérience avec David. Soient $X_1, \ldots, X_n$ des variables aléatoires qui valent 1 si David n'arrive pas à compter ses doigts et 0 dans le cas contraire. Les variables aléatoires X_i suivent alors une distribution de Bernoulli avec probabilité $p = 0{,}2$. La probabilité estimée qu'un enfant dyslexique n'arrive pas à compter ses doigts sera

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}.$$

On aimerait que l'écart-type de l'estimateur $\hat{p}$ soit égal à 0,05, ou, autrement dit, que sa variance soit égale à $0{,}05^2$. Ainsi

$$\mathrm{var}(\hat{p}) = \mathrm{var}(\overline{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathrm{var}(X_i) = \frac{p(1-p)}{n} = \frac{0{,}16}{n} = 0{,}05^2,$$

et

$$n = 64.$$

La taille de l'échantillon est donc de 64.

6.4

Soit X_i la variable aléatoire qui vaut 1 si la famille i vit en dessous du seuil de pauvreté et 0 sinon. La proportion p de familles vivant en dessous du seuil de pauvreté dans l'échantillon est

$$p = \overline{X} = \sum_{i=1}^{n} X_i.$$

X_i suivant une loi de Bernoulli de probabilité p, on a

$$E(\overline{X}) \simeq p = 0{,}15$$

et

$$\mathrm{var}(\overline{X}) = \frac{p(1-p)}{n} \simeq \frac{0{,}128}{n}.$$

On veut que l'écart-type de $\overline{X}$ soit égal à 0,02, donc

$$\sqrt{\mathrm{var}(\overline{X})} = \sqrt{\frac{0{,}128}{n}} = 0{,}02 \quad \Leftrightarrow \quad n \simeq 319.$$

Il faut 319 familles pour avoir un écart-type d'environ 0,02.

6.5

Soit p la proportion de l'échantillon qui vote en faveur du candidat C et soit X_i une variable aléatoire qui vaut 1 si l'électeur i vote pour C et 0 sinon. Soit V (respectivement CA) l'événement « l'électeur vient de la ville (respectivement de la campagne) ».

1. On connaît la vraie valeur de p. En effet

$$\begin{aligned} p &= P(X_i = 1) = P(X_i = 1 \mid V)P(V) + P(X_i = 1 \mid CA)P(CA) \\ &= 0{,}3 \cdot 0{,}8 + 0{,}75 \cdot 0{,}2 = 0{,}39. \end{aligned}$$

L'estimateur $\hat{p}$ de p est $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Son espérance est

$$\begin{aligned} E(\hat{p}) &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} P(X_i = 1) \\ &= \frac{1}{n}\sum_{i=1}^{n} \Big(P(X_i = 1 \mid \tilde{V})P(\tilde{V}) + P(X_i = 1 \mid \widetilde{CA})P(\widetilde{CA}) \Big) \\ &= 0{,}3\frac{1}{7} + 0{,}75\frac{6}{7} \simeq 0{,}69, \end{aligned}$$

où $\tilde{V}$ (respectivement $\widetilde{CA}$) est l'événement « l'individu dans l'échantillon vient de la ville (respectivement de la campagne) ». On en déduit le biais de $\hat{p}$

$$\text{biais}(\hat{p}, p) = E(\hat{p}) - p \simeq 0{,}69 - 0{,}39 = 0{,}3.$$

2. Oui, le biais est suffisamment important pour que la moyenne de l'échantillon soit fausse.

6.6

Si σ^2 est la variance d'une observation X_i, la variance de la moyenne d'un échantillon de taille n est alors

$$\text{var}(\overline{X}) = \frac{\sigma^2}{n}.$$

L'efficacité entre les 2 estimateurs est le rapport entre les variances, soit

$$\text{eff} = \frac{\sigma^2/500}{\sigma^2/320} = \frac{320}{500} = 0{,}64.$$

Chapitre 7

Estimation ponctuelle

Introduction

Dans ce chapitre on considère les différentes méthodes d'estimation ainsi que leurs propriétés. On est souvent amenés à estimer une ou plusieurs caractéristiques de la population à partir des données (échantillon). Il peut s'agir de paramètre(s) de la distribution sous-jacente ou de caractéristique(s) de la population (espérance, variance, quantile, etc.), qui est (sont) bien entendu fonction des paramètres de la distribution. Par exemple, on peut vouloir estimer la médiane des revenus dans une région donnée, la proportion de la population qui vit en dessous du seuil de pauvreté, la proportion de votants qui donnera sa voix au candidat bleu, etc.

Les estimateurs peuvent être comparés sur la base de différents critères : le biais, la variance, l'erreur carrée moyenne et la convergence. La biais mesure l'écart entre la vraie valeur (inconnue, dans la population) de la quantité à estimer et la valeur que l'estimateur prend en espérance (c'est-à-dire en moyenne si l'on répétait l'expérience). La variance quantifie la variabilité autour de l'espérance. Idéalement, un bon estimateur possède un biais petit voire nul, et une petite variance. Il n'est pas possible de réduire simultanément le biais et la variance d'un estimateur, ce qui amène à la définition de l'erreur carrée moyenne qui exprime la combinaison de ces deux quantités. La minimisation de l'erreur carrée moyenne gère le compromis entre biais et variance. Ce compromis est une notion importante et apparaît dans beaucoup de situations en statistique. Finalement, un estimateur est dit convergent s'il converge en probabilité vers la vraie valeur lorsque la taille de l'échantillon augmente. Il est à noter que le biais, la variance et donc l'erreur carrée moyenne sont des mesures pour échantillons finis, alors que la convergence est une caractéristique asymptotique.

La borne de Cramér-Rao constitue un résultat important dans la comparaison d'estimateurs. Elle définit une borne inférieure à la variance de tout estimateur, en fonction de la distribution sous-jacente des observations. Si un

estimateur atteint la borne de Cramér-Rao (c'est-à-dire que sa variance est égale à la borne) on dit qu'il est efficace.

Les méthodes considérées ici sont la méthode des moments, la méthode du maximum de vraisemblance ainsi que la méthode des moindres carrés, que l'on résume de façon non formelle ci-dessous.

Méthode du maximum de vraisemblance. Comme son nom l'indique, cette méthode est basée sur la fonction de vraisemblance, qui est maximisée afin d'obtenir les estimateurs souhaités. Les estimations ainsi obtenues seront les valeurs les plus vraisemblable pour les paramètres étant donné les données que l'on a observées. Sous des conditions de régularité, les estimateurs du maximum de vraisemblance sont convergents et leur distribution est asymptotiquement normale. Ils sont asymptotiquement efficaces.

Méthode des moments. Cette méthode égalise les moments théoriques d'une population ($E(X^k)$, $k = 1, 2, \ldots$), qui sont une fonction des paramètres, avec les moments empiriques de l'échantillon ($1/n \sum_{i=1}^{n} X_i^k$, $k = 1, 2, \ldots$) afin d'obtenir des estimations des paramètres d'intérêt. Les estimateurs des moments sont convergents. Contrairement aux estimateurs du maximum de vraisemblance, les estimateurs des moments ne sont en général pas efficaces. Sous certaines conditions de régularité, ces estimateurs sont asymptotiquement normales.

Méthode des moindres carrés. Cette approche est fortement liée à l'analyse de régression. Dans ce cadre, on minimise les erreurs au carré afin d'obtenir les estimateurs des paramètres. Le théorème de Gauss-Markov assure que si l'on suppose que les erreurs du modèle ont espérance 0 et sont indépendantes avec même variance finie, l'estimateur des moindres carrés est le meilleur estimateur linéaire non biaisé. Plus généralement, le meilleur estimateur linéaire non biaisé de toute combinaison linéaire est son estimateur des moindres carrés.

Si l'on fait de plus l'hypothèse de distribution normale des erreurs dans le modèle de régression, alors les estimateurs des moindres carrés sont les estimateurs du maximum de vraisemblance et héritent donc de leur propriétés.

Notes historiques

La méthode du maximum de vraisemblance est due aux travaux de R. A. Fisher entre 1912 et 1922.

La méthode des moments a en premier lieu était discutée par K. Pearson en 1894. Elle a été généralisée par L. Hansen (1982), méthode généralisée des moments ou GMM.

La méthode des moindres carrés est née des intérêts de C. F. Gauss (1777-1855) en astronomie en 1795. Puisqu'il n'a jamais publié ses résultats, c'est souvent à A.-M. Legendre qu'est attribuée la paternité de la méthode des moindres carrés. Il l'a en effet publiée en 1806 dans son ouvrage *Nouvelles méthodes pour la détermination des orbites des comètes.*

La borne de Cramér-Rao port le nom du mathématicien suédois H. Cramér (1893-1985), et du statisticien indien C. R. Rao (né en 1920). En fait, l'inégalité a d'abord été énoncée par R. A. Fisher (qui était entre autre le directeur de thèse de Rao) en 1922. Ensuite, elle a été redérivée indépendamment par M. Fréchet (1943), G. Darmois (1945), H. Cramér (1946) et C. R. Rao (1945), raison pour laquelle on l'appelle aussi l'inégalité de Fréchet-Darmois-Cramér-Rao.

Références (théorie)

Casella et Berger, chapitre 7 [8]; Dodge, chapitre 10 [5]; Lejeune, chapitre 6 [3]; Morgenthaler, chapitres 6 et 7 [4]; et Rice chapitre 8 [9].

Exercices

Comparaisons d'estimateurs

7.1

1. Si $E(X_i) = \mu$ et $Var(X_i) = \sigma^2$, montrer que

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

est un estimateur non biaisé de σ^2.
Indication : développer $\sum_{i=1}^{n}(X_i - \bar{X})^2$.

2. Est-ce que $s = \sqrt{s^2}$ est un estimateur non biaisé de σ ?

7.2

On considère le modèle de régression simple

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n$$

où $\epsilon_1, \ldots, \epsilon_n$ sont n variables aléatoires (erreurs) indépendantes avec la même distribution et $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$.

1. Calculer $E(Y_i)$ et $Var(Y_i)$.
2. On désire estimer α et β par la méthode des moindres carrés. Donner les estimateurs $\hat{\alpha}_{MC}$ et $\hat{\beta}_{MC}$.
3. Montrer que $\hat{\alpha}_{MC}$ et $\hat{\beta}_{MC}$ sont des estimateurs sans biais.
4. Calculer $Var(\hat{\alpha}_{MC})$ et $Var(\hat{\beta}_{MC})$.

7.3

Pour estimer la moyenne μ d'une population on utilise souvent la moyenne de l'échantillon $\bar{X}$. Pour réduire la variance de $\bar{X}$, on envisage d'utiliser un estimateur du type $a\bar{X}$, $0 \leq a \leq 1$.

1. Calculer $\text{biais}^2(a\bar{X}, \mu)$, $\text{var}(a\bar{X})$ et l'erreur carré moyenne de $a\bar{X}$.
2. Faire un graphique de $\text{biais}^2(a\bar{X}, \mu)$ et $\text{var}(a\bar{X})$ en fonction de a, pour $0 \leq a \leq 1$.
3. À quel prix (en biais) peut-on réduire la variance de l'estimateur de μ ?
4. Pour quelle valeur de a a-t-on le meilleur compromis entre le biais au carré et la variance ?

7.4

Soient $X_1, \ldots, X_n$ distribués selon une loi $\mathcal{N}(\mu, \sigma^2)$. On considère les 2 estimateurs suivants de σ^2 (μ inconnu)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{et} \quad \tilde{s} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Comparez-les du point de vue de l'erreur carrée moyenne.

Indication : $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2$ suit une loi χ^2_{n-1}.

7.5

Soient $U_1, \ldots, U_n$ des variables aléatoires indépendantes qui sont distribuées uniformément sur l'intervalle $(0, \theta)$. Soit

$$T = 2 \cdot \frac{(U_1 + \ldots + U_n)}{n} \text{ et } S = \frac{n+1}{n} \cdot \max\{U_1, \ldots, U_n\}.$$

1. Trouver l'espérance et la variance de T.
2. Trouver l'espérance et la variance de S.
   Indication : trouver d'abord la distribution de $Y = \max\{U_1, \ldots, U_n\}$.
3. Comparer les 2 estimateurs du point de vue de l'erreur carrée moyenne.

7.6

Soit $X_1, \ldots, X_n$ un échantillon de variables indépendantes avec densité

$$f_X(x) = \begin{cases} \frac{2}{3x^2} & \text{si } -1 < x < 2 \\ 0 & \text{sinon.} \end{cases}$$

On dispose de 2 estimateurs de $\mu = E(X_i)$

$$T_1(X_1, \ldots, X_n) = \frac{X_1 + X_2}{2} \quad \text{et}$$
$$T_2(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i - \frac{X_1 - X_2}{2}.$$

Déterminer lequel des 2 estimateurs est le plus efficace.

7.7

1. Deux instituts de sondages effectuent un échantillonnage pour estimer la proportion π des Suisses qui sont favorables à la légalisation du cannabis. Le premier échantillon montre une proportion $P_1 = 60/200 = 30$ %

de personnes favorables. Dans le 2e échantillon la proportion est $P_2 = 240/1000 = 24$ %. Pour avoir une estimation globale, on prend simplement la moyenne $P^* = 27$ %. Quelle est la variance (estimée) de cet estimateur?

2. Le 1er sondage est clairement moins fiable que le 2e. On propose alors de ne pas le considérer et d'utiliser l'estimateur $P_2 = 24$ %. Quelle est la variance (estimée) de cet estimateur?
3. Le meilleur estimateur est celui qui donne le même poids à chaque observation (et non pas à chaque échantillon). Ceci veut dire qu'on prend l'estimation $P = (60 + 240)/(200 + 1\,000) = 25$ %. Quelle est la variance (estimée) de cet estimateur?
4. Donner les efficacités des estimateurs considérés ci-dessus (par rapport au meilleur).
5. Vrai ou faux?
   Il est important de connaître la fiabilité des sources des observations. Par exemple, si on donne trop de poids à l'information provenant de sources non fiables, on peut perturber toute l'estimation.

7.8

La moyenne $\bar{X}$ d'un échantillon de taille n est utilisée pour estimer la moyenne μ de la population. On désire déterminer n de sorte que l'erreur absolue $|\ \bar{X} - \mu\ |$ soit au plus égale à un nombre fixé d avec une grande probabilité $1 - \alpha$ (α donné). Soit un échantillon $X_1, \ldots, X_n$ issu d'une loi $\mathcal{N}(\mu, 4)$.

1. Existe-t-il pour tout α un tel n?
2. Trouver n si $\alpha = 0{,}05$ et $d = 1$.
3. Établir un graphe de n en fonction de $1-\alpha$ ($\alpha \leq 0{,}1$) en conservant $d = 1$.

7.9

Le rayon R d'un cercle est mesuré avec une erreur de mesure distribuée selon une loi $\mathcal{N}(0, \sigma^2)$, σ inconnu.
Trouver un estimateur non biaisé pour la surface S du cercle étant donné que l'on dispose de n mesures indépendantes du rayon.
Rappel : $S = \pi R^2$.

7.10

Pour estimer la proportion p d'étudiant(e)s genevois(es) inscrits à la Faculté de Sciences Économiques et Sociales (SES) de l'université de Genève, on possède un échantillon $X_1, \ldots, X_n$ tel que $X_i = 1$, si l'étudiant i est genevois

et $X_i = 0$ sinon.

1. Un estimateur naïf de p est $\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$. L'estimateur $\hat{p}_1$ est-il sans biais? Donner l'erreur carrée moyenne $\mathrm{ECM}(\hat{p}_1, p_1)$.
2. Un estimateur alternatif de p est donné par $\hat{p}_2 = \dfrac{n\hat{p}_1 + 1}{n+2}$. L'estimateur $\hat{p}_2$ est-il sans biais? Calculer l'erreur carrée moyenne $\mathrm{ECM}(\hat{p}_2, p_2)$.
3. Comparer $\mathrm{ECM}(\hat{p}_1, p_1)$ et $\mathrm{ECM}(\hat{p}_2, p_2)$ pour les valeurs suivantes de p: 0, 1/8, 1/4 et 1/2.

7.11

Soient $X_1, \ldots, X_n$ n variables aléatoires indépendantes, telles que $E(X_i) = \mu$ et $\mathrm{var}(X_i) = \sigma^2$ pour $i = 1, \ldots, n$.

1. Montrer que $\bar{X}$ est un estimateur sans biais pour μ.
2. Montrer que $\bar{X}^2$ n'est pas un estimateur sans biais de μ^2. Déterminer son biais.
3. Déterminer k tel que $\bar{X}^2 - ks^2$ soit un estimateur sans biais de μ^2, où $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

7.12

Soient $Y_1, \ldots, Y_n$ indépendants et identiquement distribués selon la loi Bernoulli(p), c'est-à-dire

$$Y_i = \begin{cases} 1 & \text{avec probabilité } p \\ 0 & \text{avec probabilité } 1-p. \end{cases}$$

On désire estimer $\mathrm{var}(Y_i)$ c'est-à-dire $\theta = p(1-p)$. Pour ce faire, on propose de s'inspirer de l'estimateur $\hat{p} = \bar{Y}$ de p et de considérer $\hat{\theta} = \bar{Y}(1-\bar{Y})$.

1. Calculer le biais de cet estimateur.
2. Comment proposez-vous de corriger ce biais?

Information de Fisher et borne de Cramér-Rao

7.13

On définit l'information de Fisher pour un paramètre θ par

$$J(\theta) = E\left(\left[\frac{\partial}{\partial\theta}\log(f_\theta(x))\right]^2\right).$$

Montrer que si θ appartient à un support compact

$$J(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2}\log(f_\theta(x))\right).$$

Indication : $E(\frac{\partial}{\partial\theta}\log(f_\theta(x))) = \int[\frac{\partial}{\partial\theta}\log(f_\theta(x))]f_\theta(x)dx = 0$.

7.14

Soient $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes distribuées selon la loi F_θ dont la densité est f_θ et θ appartient à un support compact. On notera $J(\theta)$ l'information de Fisher

$$J(\theta) = E\left(\left[\frac{\partial}{\partial\theta}\log(f_\theta(Y))\right]^2\right) = -E\left(\frac{\partial^2}{\partial\theta^2}\log(f_\theta(Y))\right).$$

Définissons $g_\theta(y_1, \ldots, y_n) = f_\theta(y_1)\cdot f_\theta(y_2)\cdot\ldots\cdot f_\theta(y_n)$, la densité conjointe de $(Y_1, \ldots, Y_n)$.

1. Calculer l'information de Fisher $\widetilde{J}(\theta)$ par rapport à $g_\theta(y_1, \ldots, y_n)$.
2. Montrer que $\widetilde{J}(\theta) = nJ(\theta)$.

7.15

1. Calculer l'information de Fisher $J(p)$ et déduire la borne de Cramér-Rao pour des estimateurs non biaisés lorsque $X_1, \ldots, X_n$ sont tels que

$$X_i = \begin{cases} 1 & \text{avec probabilité } p \\ 0 & \text{avec probabilité } q = 1-p. \end{cases}$$

   Donner un estimateur de p qui atteint cette borne.
2. Calculer $J(\sigma^2)$ et la borne de Cramér-Rao lorsque $X_1, \ldots, X_n$ suivent une loi $\mathcal{N}(0, \sigma^2)$.
   Indication : utiliser la formule démontrée à l'exercice 7.13.

7.16

Calculer la borne de Cramér-Rao pour $X_1, \ldots, X_n$ provenant d'une loi de Pareto avec densité

$$f_\theta(x) = \begin{cases} c^\theta\theta x^{-(1+\theta)} & x > c \\ 0 & \text{sinon,} \end{cases}$$

avec $\theta > 1$ et $c > 0$.

7.17

1. Calculer la borne de Cramér-Rao pour la variance d'un estimateur sans biais du paramètre λ d'une loi de Poisson $\mathcal{P}(\lambda)$.
2. Pour estimer λ, on utilise $\bar{X}$. S'agit-il d'un « bon » estimateur?

7.18

Dans le cas d'une loi normale de paramètres (μ, σ^2), σ^2 étant connu, $\overline{X}$ est utilisé pour estimer μ. Atteint-il la borne de Cramér-Rao?

Méthodes d'estimation

7.19

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes avec densité

$$f_\theta(x) = \begin{cases} \frac{1}{2}(1+\theta x) & -1 \leq x \leq 1 \\ 0 & \text{sinon.} \end{cases}$$

1. Calculer l'estimateur des moments $\hat{\theta}_M$ de θ.
2. Calculer le biais et la variance de $\hat{\theta}_M$.
3. Pour étudier les propriétés statistiques de $\hat{\theta}_M$, on effectue une simulation d'échantillonnage. On génère 1 000 échantillons de taille $n = 10$ suivant la loi $f_\theta(x)$, où $\theta = 1$, et on calcule pour chaque échantillon la valeur observée de l'estimateur $\hat{\theta}_M$. Quel est en moyenne le nombre d'échantillons où la valeur observée de l'estimateur est supérieure à la valeur du paramètre $\theta = 1$?

7.20

Soit $X_1, \ldots, X_n$ un échantillon aléatoire provenant d'une distribution géométrique qui s'écrit

$$P(X = x) = (1-p)^{x-1}p, \;\; x = 1, 2, 3, \ldots$$

1. Trouver l'estimateur du maximum de vraisemblance de p.
2. Trouver l'estimateur des moments de p.

7.21

Soit une variable aléatoire X distribuée selon la loi binomiale $\mathcal{B}(n, p)$.

1. Calculer l'estimateur du maximum de vraisemblance pour p.
2. Calculer la borne de Cramér-Rao.
3. Montrer que l'estimateur du maximum de vraisemblance calculé en 1. atteint la borne de Cramér-Rao.

7.22

On suppose que $X_1, \ldots, X_n$ sont indépendants et identiquement distribués selon une loi $\mathcal{N}(\mu, \sigma^2)$.

1. Pour μ connu, donner l'estimateur du maximum de vraisemblance de σ^2.
2. Dessiner le logarithme de la fonction de vraisemblance de μ dans le cas où σ^2 est connu. Donner l'estimateur du maximum de vraisemblance de μ.
3. Dans le dernier cas ci-dessus (σ^2 connu), existe-t-il un autre estimateur de μ qui possède une variance plus petite?

7.23

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes suivant une loi avec fonction de densité

$$f_\theta(x) = (\theta + 1)x^\theta, \text{ pour } 0 \leq x \leq 1.$$

1. Trouver l'estimateur des moments de θ.
2. Calculer l'estimateur du maximum de vraisemblance de θ.

7.24

Si $X_1, \ldots, X_n$ sont issus d'une loi uniforme $\mathcal{U}(a, b)$, calculer $\hat{a}_M$ et $\hat{b}_M$, les estimateurs des moments de a et de b.

7.25

Considérons $X_1, \ldots, X_n$ indépendants provenant d'une loi uniforme $\mathcal{U}(0, \theta)$.

1. Calculer Y l'estimateur du maximum de vraisemblance de θ.
   Indication : dessiner la fonction de vraisemblance de θ et en déduire le maximum.
2. Déterminer la distribution de Y, c'est-à-dire calculer $F_Y(y) = P(Y < y)$. (Considérer d'abord le cas $n = 2$.)
3. Calculer le biais de l'estimateur Y et proposer un estimateur sans biais de la forme aY, avec $a \in \mathcal{R}$.

7.26

Soient $X_1, \ldots, X_n$ provenant d'une loi de Pareto avec densité

$$f_\alpha(x) = \begin{cases} \alpha x_0^\alpha x^{-(1+\alpha)} & x > x_0 \\ 0 & \text{sinon,} \end{cases}$$

avec $x_0 > 0$ et $\alpha > 1$.

1. Calculer l'estimateur des moments $\hat{\alpha}_M$ de α.
2. Donner l'estimateur du maximum de vraisemblance $\hat{\alpha}_{MV}$ de α.
3. On décide de changer de paramétrisation en prenant $\alpha = 1/\eta$. La densité s'écrit donc
$$f_\eta(x) = \begin{cases} \frac{1}{\eta} x_0^{1/\eta} x^{-(1+1/\eta)} & x > x_0 \\ 0 & \text{sinon,} \end{cases}$$
avec $x_0 > 0$ et $\eta < 1$.
Calculer l'estimateur du maximum de vraisemblance $\hat{\eta}_{MV}$ de η.
4. Quelle est la relation entre $\hat{\alpha}_{MV}$ et $\hat{\eta}_{MV}$? Commenter.

7.27

On considère le modèle de régression

$$Y_i = \alpha + \beta x_i + \epsilon_i \quad i = 1, \ldots, n,$$

où $x_1, \ldots, x_n$ sont fixés et $\epsilon_1, \ldots, \epsilon_n$ sont des variables aléatoires indépendantes suivant une loi $\mathcal{N}(0, \sigma^2)$.

1. Démontrer que les estimateurs des moindres carrés pour α et β sont les estimateurs du maximum de vraisemblance.
2. Donner la densité de Y_i.
3. Supposons par simplicité que α est connu. Calculer l'estimateur des moindres carrés de β et sa variance.
4. Donner $g_\beta(y_1, \ldots, y_n)$, la densité conjointe de $(Y_1, \ldots, Y_n)$.
5. Calculer l'information de Fisher $\widetilde{J}(\beta)$ par rapport à $g_\beta(y_1, \ldots, y_n)$.
6. Conclure en utilisant :
- la borne de Cramér-Rao ;
- l'exercice 7.14.

7.28

Soient $X_1, \ldots, X_n$ les revenus annuels des jeunes diplômé(e)s de l'Université de Genève durant leur 1re année de travail. On suppose que chaque X_i suit la loi de Weibull de densité

$$f_\theta(x) = \begin{cases} \frac{c}{\theta} x^{c-1} \exp(-\frac{x^c}{\theta}) & x > 0 \\ 0 & \text{sinon,} \end{cases}$$

avec $c > 0$ fixé.
L'espérance d'une variable aléatoire X_i qui suit la loi de Weibull vaut $\mu = \theta^{1/c}\Gamma(1 + 1/c)$, où $\Gamma(\alpha)$ est la fonction Gamma.

1. Trouver l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ de θ.
2. Calculer son espérance et sa variance.
   Indication: pour ce calcul, trouver d'abord la densité de $Y_i = X_i^c$ et ensuite $E(X_i^c)$ et $\text{var}(X_i^c)$.
3. Calculer la borne de Cramér-Rao pour des estimateurs non biaisés de θ. Est-ce que $\hat{\theta}_{MV}$ est un bon estimateur?

7.29

Soit X une variable aléatoire log-normale de paramètres μ et et σ^2, c'est-à-dire vérifiant la propriété $\log(X) \sim \mathcal{N}(\mu, \sigma^2)$. La densité d'une telle distribution s'écrit

$$f(x) = \frac{c}{x} \exp\Big\{ - \frac{\big[\log(x) - \mu\big]^2}{2\sigma^2} \Big\}, \quad x > 0$$

1. Calculez l'estimateur du maximum de vraisemblance $\hat{\mu}_{MV}$ de μ basé sur $X_1, \ldots, X_n$, sachant que c ne dépend pas de μ.
2. Calculer $E(\hat{\mu}_{MV})$. S'agit-il d'un estimateur non biaisé?
3. Calculer $Var(\hat{\mu}_{MV})$ et la borne de Cramér-Rao. Est-ce que $\hat{\mu}_{MV}$ est un bon estimateur? Justifiez votre réponse.

7.30

Lorsqu'on désire estimer une distribution de revenus, on utilise souvent un modèle basé sur la loi log-normale, c'est-à-dire $Y \sim \text{log-normale}(\mu, \sigma^2)$ ou $\log(Y) \sim N(\mu, \sigma^2)$. De plus, il arrive souvent qu'on exprime μ par une combinaison linéaire de variables explicatives. Prenons le cas simple d'une seule variable explicative, c'est-à-dire $\mu = \beta x$.

1. Étant donné un échantillon d'observations (x_i, y_i), calculer les estimateurs du maximum de vraisemblance $(\hat{\beta}_{MV}, \hat{\sigma}^2_{MV})$ de (β, σ^2).
2. Calculer la borne de Cramér-Rao pour σ^2 (β étant constant) et la comparer à la variance de $\hat{\sigma}^2_{MV}$. Commenter.

7.31

On veut modéliser des parts budgétaires par une distribution $\text{Beta}(\alpha, 2)$ avec densité

$$f_\alpha(x) = \alpha(\alpha + 1)x^{\alpha-1}(1 - x), \quad 0 < x < 1, \, \alpha > 0.$$

Soient $X_1, \ldots, X_n$ n observations indépendantes issues de cette distribution.

1. Calculer l'estimateur T_n de α obtenu par la méthode des moments.
2. En sachant que $V(T_n) \simeq \frac{\alpha(\alpha+2)^2}{2n(\alpha+3)}$, montrer que T_n n'est pas efficace en calculant la borne de Cramér-Rao.

7.32

Une usine qui produit des pièces électroniques a décidé de marquer chaque pièce avec un numéro de série. Les numéros de série commencent à 1 et se terminent à N, où N est le nombre de pièces produites. Deux pièces sont tirées au hasard et leur numéros de série sont 888 et 751 respectivement. Quelle est l'estimation de N par la méthode des moments ? Et quelle est l'estimation par la méthode du maximum de vraisemblance ?

7.33

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées dont la densité est définie par

$$f_\theta(x) = \begin{cases} \frac{x}{\theta}\exp\left(-\frac{x^2}{2\theta}\right) & \text{si } \ x > 0 \\ 0 & \text{sinon.} \end{cases}$$

1. Déterminer l'estimateur des moments de θ.
2. Déterminer θ_{MV}, l'estimateur du maximum de vraisemblance de θ.
3. Utiliser ce résultat pour montrer que θ_{MV} est un estimateur sans biais pour θ.
4. Pour $n = 10$, existe-t-il un autre estimateur non biaisé de θ dont la variance est strictement plus petite que celle de θ_{MV} ?

7.34

Lorsqu'une variable expliquée (endogène) est dichotomique, on utilise souvent un modèle logistique pour la modéliser. Soit

$$P(Y = 1 \mid X = x) = p = \frac{1}{1 + e^{-\beta x}}$$
$$P(Y = 0 \mid X = x) = 1 - p = \frac{1}{1 + e^{\beta x}},$$

où x est une variable explicative.
On dispose de n observations (y_i, x_i), $i = 1, \ldots, n$.

1. Montrer que la log-vraisemblance s'écrit

$$l(\beta \mid y_1, \ldots, y_n; x_1, \ldots, x_n) = \\ = \sum_{i=1}^{n} \left\{ y_i \log \left(\frac{1}{1 + e^{-\beta x_i}} \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + e^{-\beta x_i}} \right) \right\}.$$

2. Écrire l'équation de vraisemblance.
3. Comparer l'équation du point 2. avec l'équation normale pour l'estimation des moindres carrés de β de la régression linéaire classique $y_i = \beta x_i + \varepsilon_i$. Que remarquez-vous ?

Corrigés

7.1

1. Pour montrer que s^2 est un estimateur non biaisé de σ^2, il faut vérifier que son espérance est égale à σ^2

$$\begin{aligned} E(s^2) = & \\ = & \frac{1}{n-1} E\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n-1} E\left(\sum_{i=1}^{n} X_i^2 - 2\overline{X}\sum_{i=1}^{n} X_i + n\overline{X}^2\right) \\ = & \frac{1}{n-1}\left(\sum_{i=1}^{n} E(X_i^2) - nE(\overline{X}^2)\right) \\ = & \frac{1}{n-1}\left(\sum_{i=1}^{n}(\mathrm{var}(X_i) + E^2(X_i)) - n(\mathrm{var}(\overline{X}) + E^2(\overline{X}))\right) \\ = & \frac{1}{n-1}\left(n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2)\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2. \end{aligned}$$

   L'estimateur s^2 est sans biais pour σ^2.
2. Par contre, on montre à l'aide de l'inégalité de Jensen que S n'est pas un estimateur sans biais de σ

$$E(s) = E(\sqrt{s^2}) \neq \sqrt{E(s^2)} = \sigma.$$

7.2

1. Le calcul de l'espérance et de la variance de Y_i donne

$$E(Y_i) = E(\alpha + \beta x_i + \epsilon_i) = \alpha + \beta x_i + E(\epsilon_i) = \alpha + \beta x_i$$

   et

$$\mathrm{var}(Y_i) = \mathrm{var}(\alpha + \beta x_i + \epsilon_i) = \mathrm{var}(\epsilon_i) = \sigma^2.$$

2. On aimerait estimer α et β par la méthode des moindres carrés. Pour cela, il faut minimiser la quantité $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$ par rapport à α et β. On obtient l'ensemble d'équations

$$\begin{cases} 2\sum_{i=1}^{n}(y_i - \hat{\alpha}_{MC} - \hat{\beta}_{MC} x_i)\cdot(-1) = 0 \\ 2\sum_{i=1}^{n}(y_i - \hat{\alpha}_{MC} - \hat{\beta}_{MC} x_i)\cdot(-x_i) = 0. \end{cases}$$

   De la 1re, on déduit que

$$\begin{aligned} \Leftrightarrow & \quad \sum_{i=1}^{n} y_i - n\hat{\alpha}_{MC} - \hat{\beta}_{MC}\sum_{i=1}^{n} x_i = 0 \\ \Leftrightarrow & \quad \hat{\alpha}_{MC} = \overline{y} - \hat{\beta}_{MC}\overline{x}, \end{aligned}$$

et en utilisant ce résultat dans la 2^{e}

$$\begin{aligned}
&\Leftrightarrow \quad \sum_{i=1}^{n} y_i x_i - \hat{\alpha}_{MC} \sum_{i=1}^{n} x_i - \hat{\beta}_{MC} \sum_{i=1}^{n} x_i^2 = 0 \\
&\Leftrightarrow \quad \sum_{i=1}^{n} y_i x_i - (\overline{y} - \hat{\beta}_{MC}\overline{x}) \sum_{i=1}^{n} x_i - \hat{\beta}_{MC} \sum_{i=1}^{n} x_i^2 = 0 \\
&\Leftrightarrow \quad \hat{\beta}_{MC} = \frac{\sum_{i=1}^{n} y_i (x_i - \overline{x})}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = \frac{\sum_{i=1}^{n} y_i (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \\
&\qquad = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}.
\end{aligned}$$

3. Vérifions que les 2 estimateurs sont sans biais

$$\begin{aligned}
E(\hat{\beta}_{MC}) &= E\left(\frac{\sum_{i=1}^{n} y_i (x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \\
&= \frac{1}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sum_{i=1}^{n} (x_i - \overline{x}) E(y_i) \\
&= \frac{1}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \sum_{i=1}^{n} (x_i - \overline{x})(\alpha + \beta x_i) \\
&= \frac{1}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \left(\alpha \underbrace{\sum_{i=1}^{n} (x_i - \overline{x})}_{=0} + \beta \sum_{i=1}^{n} (x_i - \overline{x}) x_i \right) \\
&= \beta \frac{\sum_{i=1}^{n} (x_i - \overline{x}) x_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \beta \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} = \beta,
\end{aligned}$$

et

$$\begin{aligned}
E(\hat{\alpha}_{MC}) &= E(\overline{y} - \hat{\beta}_{MC}\overline{x}) = E(\overline{y}) - \overline{x} E(\hat{\beta}_{MC}) = \frac{1}{n} \sum_{i=1}^{n} E(y_i) - \overline{x}\beta \\
&= \frac{1}{n} \sum_{i=1}^{n} (\alpha + \beta x_i) - \beta\overline{x} = \alpha + \beta\overline{x} - \beta\overline{x} = \alpha.
\end{aligned}$$

4. Pour terminer, on calcule la variance de chacun des 2 estimateurs

$$\begin{aligned}
\mathrm{var}(\hat{\beta}_{MC}) &= \mathrm{var}\left(\frac{\sum_{i=1}^{n} (x_i - \overline{x}) y_i}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right) \\
&= \frac{1}{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right)^2} \sum_{i=1}^{n} \mathrm{var}\left((x_i - \overline{x}) y_i \right) \\
&= \mathrm{var}(y_i) \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{\left(\sum_{i=1}^{n} (x_i - \overline{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2},
\end{aligned}$$

où l'on a utilisé l'indépendance des y_i. Pour $\hat{\alpha}_{MC}$

$$\text{var}(\hat{\alpha}_{MC}) = \text{var}(\overline{y} - \hat{\beta}_{MC}\overline{x}) = \text{var}(\overline{y}) + \overline{x}^2\text{var}(\hat{\beta}_{MC}) + 2\overline{x}\text{cov}(\overline{y}, \hat{\beta}_{MC}).$$

La covariance entre $\overline{y}$ et $\hat{\beta}_{MC}$ se calcule de la manière suivante

$$\begin{aligned}
\text{cov}(\overline{y}, \hat{\beta}_{MC}) &= \text{cov}\left(\frac{1}{n}\sum_{j=1}^{n} y_j\,,\, \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right) \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\sum_{j=1}^{n}\text{cov}(y_j, y_i) \\
&= \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\text{var}(y_i, y_i) \\
&= \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}\underbrace{\sum_{i=1}^{n}(x_i - \overline{x})}_{=0} = 0,
\end{aligned}$$

où, par l'indépendance des y_i, $\text{cov}(y_j, y_i) = 0$ si $i \neq j$ et $\text{cov}(y_i, y_i) = \sigma^2$. On trouve alors la variance de $\hat{\alpha}_{MC}$

$$\begin{aligned}
\text{var}(\hat{\alpha}_{MC}) &= \frac{\sigma^2}{n} + \overline{x}^2\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right) \\
&= \frac{\sigma^2}{n}\frac{\sum_{i=1}^{n}x_i^2 - n\overline{x}^2 + n\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sigma^2}{n}\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.
\end{aligned}$$

7.3

1. Pour calculer $\text{biais}^2(a\overline{X}, \mu)$, on a besoin de $E(a\overline{X})$

$$E(a\overline{X}) = a \cdot E\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) = a \cdot \mu,$$

ce qui implique

$$\text{biais}^2(a\overline{X}, \mu) = (a\mu - \mu)^2 = (a-1)^2\mu^2.$$

La variance et l'erreur carrée moyenne sont alors

$$\text{var}(a\overline{X}) = a^2\text{var}(\overline{X}) = a^2\frac{\sigma^2}{n},$$

et

$$\text{ECM}(a\overline{X}, \mu) = a^2\frac{\sigma^2}{n} + (a-1)^2\mu^2.$$

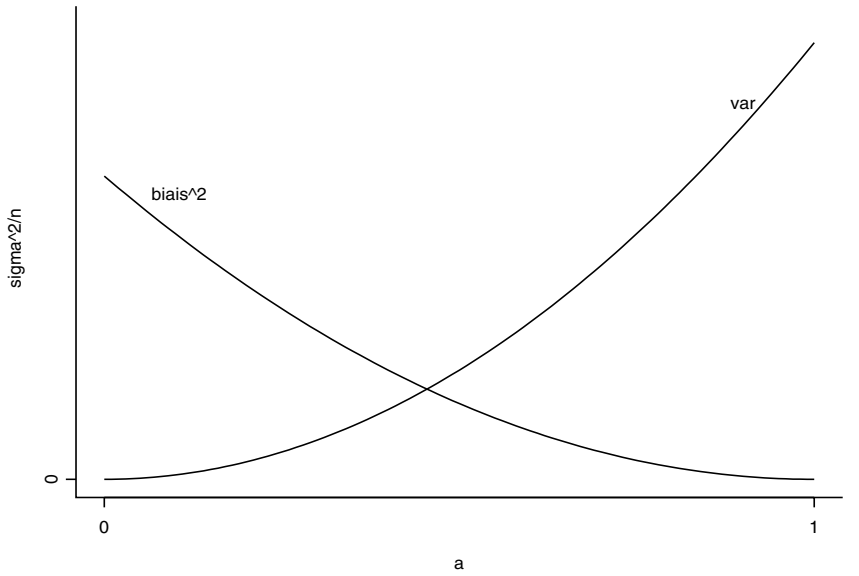


Fig. 7.1 – Graphe de $\text{biais}^2(a\overline{X}, \mu)$ et $\text{var}(a\overline{X})$ en fonction de a (exercice 7.3).

2. La figure 7.1 montre le graphique de $\text{biais}^2(a\overline{X}, \mu)$ et $\text{var}(a\overline{X})$ en fonction de a.

3. Pour diminuer la variance, il faut réduire la valeur de a. Cela implique une augmentation du biais en valeur absolue.

4. Il existe 2 manières différentes pour choisir a :

   (a) a tel que $\text{biais}^2(a\overline{X}, \mu) = \text{var}(a\overline{X})$

$$\Leftrightarrow \quad (a-1)^2\mu^2 = a^2\frac{\sigma^2}{n}$$
$$\Leftrightarrow \quad a = \frac{\mu^2 \pm \frac{\sigma}{\sqrt{n}}\mu}{\mu^2 - \frac{\sigma^2}{n}}$$

(b) a tel que $\mathrm{ECM}(a\overline{X}, \mu)$ est minimal

$$\begin{aligned} &\Leftrightarrow \quad \frac{\partial}{\partial a}\mathrm{ECM}(a\overline{X}, \mu) = 0 \\ &\Leftrightarrow \quad 2a\frac{\sigma^2}{n} + 2(a-1)\mu^2 = 0 \\ &\Leftrightarrow \quad a = \frac{\mu^2}{\sigma^2/n + \mu^2} \end{aligned}$$

Les 2 méthodes donnent des résultats différents.

7.4

Soit Z une variable aléatoire distribuée selon une loi χ^2 à $(n-1)$ degrés de liberté. Son espérance est $(n-1)$ et sa variance $2(n-1)$. Calculons l'espérance de l'estimateur s^2

$$\begin{aligned} E(s^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\ &= \frac{\sigma^2}{n-1}E\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{\sigma^2}{n-1}E(Z) = \sigma^2. \end{aligned}$$

La variance de l'estimateur est

$$\begin{aligned} \mathrm{var}(s^2) &= \mathrm{var}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) \\ &= \frac{\sigma^4}{(n-1)^2}\mathrm{var}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{\sigma^4}{(n-1)^2}\mathrm{var}(Z) = \frac{2\sigma^4}{n-1}, \end{aligned}$$

et, par conséquent, son erreur carrée moyenne est

$$\mathrm{ECM}(s^2, \sigma^2) = \frac{2\sigma^4}{n-1}.$$

L'estimateur $\tilde{s}^2$ s'écrit comme une fonction de s^2

$$\tilde{s}^2 = \frac{n-1}{n}s^2.$$

On en déduit son espérance, son biais, sa variance et son erreur carrée moyenne

$$\begin{aligned} &E(\tilde{s}^2) = \frac{n-1}{n}\sigma^2, \\ &\mathrm{biais}(\tilde{s}^2, \sigma^2) = -\frac{\sigma^2}{n}, \\ &\mathrm{var}(\tilde{s}^2) = \frac{2(n-1)}{n^2}\sigma^4, \\ &\mathrm{ECM}(\tilde{s}^2, \sigma^2) = \frac{2(n-1)}{n^2}\sigma^4 + \frac{1}{n^2}\sigma^4 = \frac{2n-1}{n^2}\sigma^4. \end{aligned}$$

Comparons les erreurs carrées moyennes

$$\mathrm{ECM}(s^2,\sigma^2) - \mathrm{ECM}(\tilde{s}^2,\sigma^2) \quad = \quad \frac{2}{n-1}\sigma^4 - \frac{2n-1}{n^2}\sigma^4 = \frac{3n-1}{n^2(n-1)}\sigma^4.$$

Le dernier résultat est toujours positif car la taille de l'échantillon n est obligatoirement plus grande que 1. Cela implique donc que $\tilde{s}^2$ est plus efficace que s^2 du point de vue de l'erreur carrée moyenne, et ceci malgré son biais.

7.5

Les variables aléatoires $U_1, \ldots, U_n$ suivent une loi uniforme de paramètres $(0,\theta)$. Leur espérance est $\theta/2$ et leur variance $\theta^2/12$.

1. $$E(T) = 2E(\overline{U}) = 2\frac{\theta}{2} = \theta.$$
$$\mathrm{var}(T) = 4\mathrm{var}(\overline{U}) = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$
2. Déterminons la fonction de répartition de $Y = \max(U_1, \ldots, U_n)$
$$F_Y(y) = P(\max U_i < y) = P(U_1 < y)^n = (F_U(y))^n = \left(\frac{y}{\theta}\right)^n.$$
Ainsi, la fonction de densité de Y est
$$f_Y(y) = \frac{n}{\theta^n}y^{n-1}.$$
Le calcul de l'espérance de Y donne
$$E(Y) = \int_0^\theta y\frac{n}{\theta^n}y^{n-1}dy = \frac{n}{\theta^n}\frac{y^{n+1}}{n+1}\Big|_{y=0}^{y=\theta} = \frac{n}{n+1}\theta,$$
et implique
$$E(S) = \frac{n+1}{n}E(Y) = \theta.$$
L'estimateur S de θ est sans biais. Nous cherchons à présent sa variance.
$$E(Y^2) = \int_0^\theta y^2\frac{n}{\theta^n}y^{n-1}dy = \frac{n}{\theta^n}\frac{y^{n+2}}{n+2}\Big|_{y=0}^{y=\theta} = \frac{n}{n+2}\theta^2.$$
La variance de Y est alors
$$\mathrm{var}(Y) = E(Y^2) - E^2(Y) =$$
$$= \quad \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 = \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2}\theta^2$$
et celle de S
$$\mathrm{var}(S) = \frac{(n+1)^2}{n^2}\mathrm{var}(Y) = \frac{1}{n(n+2)}\theta^2.$$

3. Les 2 estimateurs sont sans biais. Le plus efficace est donc celui qui a la variance la plus petite. Comparons-les

$$\mathrm{var}(T) - \mathrm{var}(S) = \left(\frac{1}{3n} - \frac{1}{n(n+2)}\right)\theta^2 = \frac{n-1}{3n(n+2)}\theta^2.$$

Ce dernier résultat est toujours positif, n étant forcément plus grand que 1. L'estimateur S est donc le plus efficace du point de vue de l'erreur carrée moyenne.

7.6

L'estimateur le plus efficace est celui qui a l'erreur carrée moyenne la plus petite. Commençons par calculer l'espérance des 2 estimateurs pour en avoir le biais

$$E(T_1) = \frac{1}{2}(E(X_1) + E(X_2)) = \mu,$$

et

$$\begin{aligned} E(T_2) &= \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) - \frac{1}{2}(E(X_1) - E(X_2)) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) - \frac{1}{2}(\mu - \mu) = \frac{1}{n}n\mu = \mu. \end{aligned}$$

Les 2 estimateurs sont des estimateurs sans biais de μ. Le plus efficace des 2 sera, par conséquent, celui dont la variance est la plus petite. Le calcul des variances donne

$$\mathrm{var}(T_1) = \frac{1}{4}(\mathrm{var}(X_1) + \mathrm{var}(X_2)) = \frac{\sigma^2}{2},$$

où $\sigma^2 = \mathrm{var}(X_i)$. On a utilisé l'indépendance de X_1 et X_2 pour le calcul de cette variance. Dans le cas de $\mathrm{var}(T_2)$, on a

$$\begin{aligned} \mathrm{var}(T_2) &= \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{2}X_1 + \frac{1}{2}X_2\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(X_i) + \frac{1}{4}(\mathrm{var}(X_1) + \mathrm{var}(X_2)) \\ &\quad + \frac{1}{n}\mathrm{cov}\left(\sum_{i=1}^{n} X_i\,, X_1\right) - \frac{1}{n}\mathrm{cov}\left(\sum_{i=1}^{n} X_i\,, X_2\right) \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{2}. \end{aligned}$$

Dans le dernier calcul, il faut noter que les 2 covariances proviennent de la présence des variables aléatoires X_1 et X_2 dans $\sum_{i=1}^{n} X_i$. On en déduit que

$$\text{var}(T_2) - \text{var}(T_1) = \frac{\sigma^2}{n} + \frac{\sigma^2}{2} - \frac{\sigma^2}{2} = \frac{\sigma^2}{n} > 0.$$

La variance de T_2 est donc supérieure à celle de T_1, ce qui implique que T_1 est l'estimateur le plus efficace, malgré le fait qu'il n'utilise que l'information de X_1 et X_2.

7.7

1. Soit l'estimateur P^* tel que

$$P^* = \frac{1}{2}(P_1 + P_2).$$

Sa variance estimée est

$$\begin{aligned} \widehat{\text{var}}(P^*) &= \frac{1}{4}(\widehat{\text{var}}(P_1) + \widehat{\text{var}}(P_2)) = \\ &= \frac{1}{4}\left(\frac{0{,}3 \cdot 0{,}7}{200} + \frac{0{,}24 \cdot 0{,}76}{1\,000}\right) \simeq 3{,}08 \cdot 10^{-4}. \end{aligned}$$

2. Le calcul de la variance estimée de P_2 donne

$$\widehat{\text{var}}(P_2) = \frac{0{,}24 \cdot 0{,}76}{1\,000} \simeq 1{,}82 \cdot 10^{-4}.$$

3. Celui de P donne

$$\widehat{\text{var}}(P) = \frac{0{,}25 \cdot 0{,}75}{1\,200} \simeq 1{,}56 \cdot 10^{-4}.$$

4. Calculons l'efficacité de P^* et P_2 par rapport à P

$$\begin{aligned} \text{eff}(P^*, P) &\simeq \frac{1{,}56 \cdot 10^{-4}}{3{,}08 \cdot 10^{-4}} \simeq 0{,}51, \\ \text{eff}(P_2, P) &\simeq \frac{1{,}56 \cdot 10^{-4}}{1{,}82 \cdot 10^{-4}} \simeq 0{,}86. \end{aligned}$$

5. C'est vrai.

7.8

Soit un échantillon $X_1, \ldots, X_n$ tiré d'une loi normale de paramètres $(\mu, 4)$.

1. On cherche n tel que l'erreur absolue $\mid \overline{X} - \mu \mid$ soit au plus égale à un

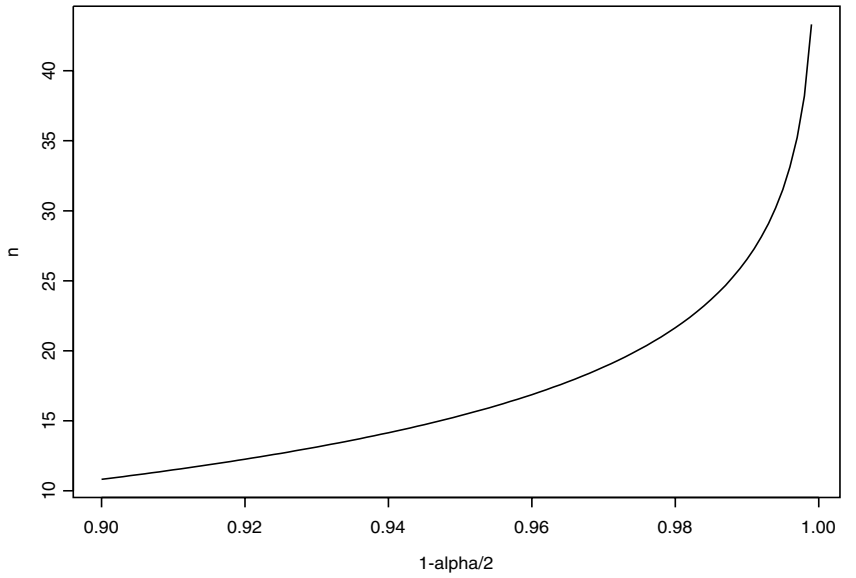


Fig. 7.2 – Graphe de $1-\alpha/2$ en fonction de n de l'exercice 7.8.

nombre fixé d avec une probabilité $1-\alpha$. Autrement dit

$$\begin{aligned} & P(\mid \overline{X} - \mu \mid \leq d) = 1-\alpha \\ \Leftrightarrow \quad & P\left(\frac{\mid \overline{X} - \mu \mid}{2/\sqrt{n}} \leq \frac{d}{2}\sqrt{n}\right) = 1-\alpha \\ \Leftrightarrow \quad & P\left(-\frac{d}{2}\sqrt{n} \leq \frac{\overline{X}-\mu}{2/\sqrt{n}} \leq \frac{d}{2}\sqrt{n}\right) = 1-\alpha \\ \Leftrightarrow \quad & \frac{d}{2}\sqrt{n} = z_{1-\alpha/2} \quad \Leftrightarrow \quad n = \left(\frac{2z_{1-\alpha/2}}{d}\right)^2, \end{aligned}$$

où $z_{1-\alpha/2}$ est le quantile de la loi normale. Il existe donc un n pour tout α.

2. Si $\alpha = 0{,}05$ et $d = 1$, alors $z_{1-\alpha/2} = 1{,}96$ et

$$n = \left(\frac{2 \cdot 1{,}96}{1}\right)^2 \simeq 15.$$

3. La figure 7.2 contient le graphe de n en fonction de $1-\alpha/2$.

7.9

Soient $X_1, \ldots, X_n$ des mesures du rayon dont la distribution est normale d'espérance R et de variance σ^2. On cherche un estimateur non biaisé de la surface S d'un cercle de rayon R. Commençons par calculer l'espérance de la nouvelle variable aléatoire $T = \pi X^2$

$$E(T) = E(\pi X^2) = \pi(\text{var}(X) + E^2(X)) = \pi R^2 + \pi\sigma^2.$$

On voit que T est un estimateur biaisé de la surface du cercle. On corrige ce biais en utilisant l'estimateur non biaisé s^2 de la variance défini par

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

pour obtenir le nouvel estimateur $\tilde{S}$ non biaisé de la surface du cercle tel que

$$\tilde{S} = \pi(X^2 - s^2).$$

7.10

1. Soit l'estimateur $\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$. Calculons son espérance

$$E(\hat{p}_1) = \frac{1}{n} n E(X_i) = p.$$

L'estimateur $\hat{p}_1$ est donc sans biais pour p. Sa variance est

$$\text{var}(\hat{p}_1) = \frac{p(1-p)}{n}$$

et son erreur carrée moyenne

$$\text{ECM}(\hat{p}_1, p_1) = \frac{p(1-p)}{n}.$$

2. Soit le nouvel estimateur $\hat{p}_2 = \frac{n\hat{p}_1+1}{n+2}$. Le calcul de son espérance donne

$$E(\hat{p}_2) = \frac{nE(\hat{p}_1) + 1}{n+2} = \frac{np+1}{n+2}.$$

Ainsi, son biais devient

$$\text{biais}(\hat{p}_2, p_2) = \frac{np+1}{n+2} - p = \frac{1-2p}{n+2}.$$

De plus, sa variance est

$$\text{var}(\hat{p}_2) = \frac{n^2}{(n+2)^2}\text{var}(\hat{p}_1) = \frac{np(1-p)}{(n+2)^2}$$

ce qui implique la valeur de son erreur carrée moyenne :

$$\mathrm{ECM}(\hat{p}_2, p_2) = \frac{np(1-p) + (1-2p)^2}{(n+2)^2}.$$

3. On compare maintenant $\mathrm{ECM}(\hat{p}_1, p_1)$ et $\mathrm{ECM}(\hat{p}_2, p_2)$ pour différentes valeurs de p. Pour $p = 0$, $\mathrm{ECM}(\hat{p}_1, p_1) = 0$ et $\mathrm{ECM}(\hat{p}_2, p_2) = \frac{1}{(n+2)^2}$, donc $\mathrm{ECM}(\hat{p}_1, p_1) < \mathrm{ECM}(\hat{p}_2, p_2)$. Pour les autres valeurs de p, il faut regarder le signe de $\mathrm{ECM}(\hat{p}_1, p_1) - \mathrm{ECM}(\hat{p}_2, p_2)$

$$\mathrm{sign}(\mathrm{ECM}(\hat{p}_1, p_1) - \mathrm{ECM}(\hat{p}_2, p_2)) = \mathrm{sign}(p(1-p)(8n+4) - n).$$

Ainsi
   - $p = 1/8$: $\mathrm{ECM}(\hat{p}_1, p_1) < \mathrm{ECM}(\hat{p}_2, p_2), \quad n > 3,$
   - $p = 1/4$: $\mathrm{ECM}(\hat{p}_1, p_1) > \mathrm{ECM}(\hat{p}_2, p_2), \quad \forall n,$
   - $p = 1/2$: $\mathrm{ECM}(\hat{p}_1, p_1) > \mathrm{ECM}(\hat{p}_2, p_2), \quad \forall n.$

7.11

1. L'espérance de $\overline{X}$ vaut

$$E(\overline{X}) = E(X_i) = \mu.$$

L'estimateur $\overline{X}$ est donc sans biais pour μ.

2. On calcule l'espérance de $\overline{X}^2$

$$E(\overline{X}^2) = \mathrm{var}(\overline{X}) + E^2(\overline{X}) = \frac{\sigma^2}{n} + \mu^2,$$

et

$$\mathrm{biais}(\overline{X}^2, \mu^2) = \frac{\sigma^2}{n} + \mu^2 - \mu^2 = \frac{\sigma^2}{n}.$$

3. Soit $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. On veut obtenir un nouvel estimateur de μ^2 de la forme $\overline{X}^2 - ks^2$. Déterminons k

$$E(\overline{X}^2 - ks^2) = E(\overline{X}^2) - kE(s^2) = \frac{\sigma^2}{n} + \mu^2 - k\sigma^2 = \mu^2 + \sigma^2\left(\frac{1}{n} - k\right).$$

Pour que cet estimateur soit sans biais, il faut que $k = \frac{1}{n}$.

7.12

Soit l'estimateur $\hat{\theta} = \overline{Y}(1 - \overline{Y})$.

1. Pour calculer son biais, on calcule d'abord son espérance

$$\begin{aligned} E(\hat{\theta}) &= E(\overline{Y}(1-\overline{Y})) = E(\overline{Y}) - E(\overline{Y}^2) = p - (\mathrm{var}(\overline{Y}) + E^2(\overline{Y})) \\ &= p - \left(\frac{p(1-p)}{n} + p^2\right) = \frac{n-1}{n}p(1-p) = \frac{n-1}{n}\theta. \end{aligned}$$

Son biais est donc

$$\text{biais}(\hat{\theta}, \theta) = \frac{n-1}{n}\theta - \theta = -\frac{1}{n}\theta.$$

2. Pour corriger ce biais, on prendra l'estimateur

$$\hat{\theta}_c = \frac{n}{n-1}\overline{Y}(1-\overline{Y}).$$

7.13

Partons de l'indication

$$E\left(\frac{\partial}{\partial\theta}\log(f_\theta(x))\right) = \int\left[\frac{\partial}{\partial\theta}\log(f_\theta(x))\right] f_\theta(x)dx = 0$$

et dérivons-la par rapport à θ (permuter l'intégration avec la dérivation est permis car x et θ sont indépendants)

$$\int\left[\frac{\partial^2}{\partial\theta^2}\log(f_\theta(x))\right] f_\theta(x)dx + \int\frac{\partial}{\partial\theta}\log(f_\theta(x))\frac{\partial}{\partial\theta}f_\theta(x)dx = 0. \qquad (7.1)$$

On utilise alors

$$\frac{\partial}{\partial\theta}\log(f_\theta(x)) = \frac{1}{f_\theta(x)}\frac{\partial}{\partial\theta}f_\theta(x)$$

pour obtenir

$$\begin{aligned}\int\frac{\partial}{\partial\theta}\log(f_\theta(x))\frac{\partial}{\partial\theta}f_\theta(x)dx &= \int\left(\frac{\partial}{\partial\theta}\log(f_\theta(x))\right)^2 f_\theta(x)dx \\ &= E\left(\left(\frac{\partial}{\partial\theta}\log(f_\theta(x))\right)^2\right) = J(\theta).\end{aligned}$$

Et finalement, grâce à (7.1)

$$J(\theta) = -\int\left[\frac{\partial^2}{\partial\theta^2}\log(f_\theta(x))\right] f_\theta(x)dx = -E\left(\frac{\partial^2}{\partial\theta^2}\log(f_\theta(x))\right).$$

7.14

1. Calculons l'information de Fisher pour la distribution $g_\theta(y_1, \ldots, y_n)$. Le logarithme de cette distribution s'écrit

$$\log(g_\theta(y_1, \ldots, y_n)) = \log\left(\prod_{i=1}^n f_\theta(y_i)\right) = \sum_{i=1}^n \log(f_\theta(y_i)),$$

par conséquent

$$\begin{aligned}\tilde{J}(\theta) &= -E\left(\frac{\partial^2}{\partial\theta^2}\log(g_\theta(y_1,\ldots,y_n))\right)\\ &= -E\left(\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^{n}\log(f_\theta(y_i))\right)\\ &= -\sum_{i=1}^{n}E\left(\frac{\partial^2}{\partial\theta^2}\log(f_\theta(y_i))\right).\end{aligned}$$

2. Les Y_i ont la même distribution, donc

$$\tilde{J}(\theta) = -nE\left(\frac{\partial^2}{\partial\theta^2}\log(f_\theta(y_i))\right) = nJ(\theta).$$

7.15

1. La loi de probabilité d'une variable aléatoire X suivant une loi de Bernoulli de probabilité p est

$$P(X=x) = p^x(1-p)^{1-x}.$$

On cherche la borne de Cramér-Rao pour p

$$\begin{aligned}&\log(P(X=x)) = x\log(p) + (1-x)\log(1-p)\\ \Rightarrow\quad &\frac{\partial\log(P(X=x))}{\partial p} = \frac{x}{p} - \frac{1-x}{1-p}\\ \Rightarrow\quad &\frac{\partial^2\log(P(X=x))}{\partial p^2} = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2}.\end{aligned}$$

L'information de Fisher est donc

$$J(p) = -E\left(-\frac{x}{p^2} - \frac{1-x}{(1-p)^2}\right) = \frac{1}{p^2}E(X) - \frac{1-E(X)}{(1-p)^2} = \frac{1}{p(1-p)},$$

et la borne de Cramér-Rao pour p

$$\text{BCR} = \frac{p(1-p)}{n}.$$

L'estimateur $\hat{p} = \overline{X}$ atteint la borne de Cramér-Rao.

2. On procède de la même manière pour un estimateur non biaisé de la variance d'une loi normale de paramètre $(0,\sigma^2)$ dont la fonction de densité est

$$f_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}x^2\right).$$

On calcule l'information de Fisher afin de trouver la borne de Cramér-Rao

$$\begin{aligned} & \log(f_{\sigma^2}(x)) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}x^2 \\ \Rightarrow \quad & \frac{\partial \log(f_{\sigma^2}(x))}{\partial(\sigma^2)} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}x^2 \\ \Rightarrow \quad & \frac{\partial^2 \log(f_{\sigma^2}(x))}{\partial(\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}x^2. \end{aligned}$$

Donc

$$J(\sigma^2) = -E\left(\frac{1}{2\sigma^4} - \frac{1}{\sigma^6}x^2\right) = -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}E(X^2) = -\frac{1}{2\sigma^4} + \frac{\sigma^2}{\sigma^6} = \frac{1}{2\sigma^4}$$

et la borne de Cramér-Rao est

$$\text{BCR} = \frac{2\sigma^4}{n}.$$

7.16

Pour obtenir la borne de Cramér-Rao d'un estimateur non biaisé du paramètre θ d'une loi de Pareto, on calcule en premier lieu l'information de Fisher

$$\begin{aligned} & \log(f_\theta(x)) = \theta\log(c) + \log(\theta) - (1+\theta)\log(x) \\ \Rightarrow \quad & \frac{\partial f_\theta(x))}{\partial\theta} = \log(c) + \frac{1}{\theta} - \log(x) \\ \Rightarrow \quad & \frac{\partial^2 f_\theta(x))}{\partial\theta^2} = -\frac{1}{\theta^2}, \end{aligned}$$

et

$$J(\theta) = -E\left(-\frac{1}{\theta^2}\right) = \frac{1}{\theta^2}.$$

La borne de Cramér-Rao est par conséquent

$$\text{BCR} = \frac{\theta^2}{n}.$$

7.17

La loi de probabilité d'une variable aléatoire X suivant une loi de Poisson de paramètre λ est

$$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

1. On en calcule la borne de Cramér-Rao pour un estimateur sans biais.

$$\begin{aligned} & \log(P(X=x)) = -\lambda + x\log(\lambda) - \log(x!) \\ \Rightarrow \quad & \frac{\partial \log(P(X=x))}{\partial \lambda} = -1 + \frac{x}{\lambda} \\ \Rightarrow \quad & \frac{\partial^2 \log(P(X=x))}{\partial \lambda^2} = -\frac{x}{\lambda^2}. \end{aligned}$$

Ainsi, l'information de Fisher est

$$J(\lambda) = -E\left(\frac{\partial^2 \log(f_\lambda(x))}{\partial \lambda^2}\right) = -E\left(\frac{x}{\lambda^2}\right) = \frac{1}{\lambda^2}E(X) = \frac{1}{\lambda},$$

et la borne de Cramér-Rao

$$\text{BCR} = \frac{\lambda}{n}.$$

2. Calculons la variance de $\overline{X}$

$$\text{var}(\overline{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{1}{n}\text{var}(X_i) = \frac{\lambda}{n}.$$

La variance de $\overline{X}$ atteint la borne de Cramér-Rao. Cela signifie que c'est un estimateur à variance minimale, donc un bon estimateur du point de vue de la variance.

7.18

Rappelons que la fonction de densité d'une loi normale de paramètres (μ, σ^2) est

$$f_\mu(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2).$$

On cherche la borne de Cramér-Rao pour μ

$$\begin{aligned} & \log(f_\mu(x)) = -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}(x-\mu)^2 \\ \Rightarrow \quad & \frac{\partial \log(f_\mu(x))}{\partial \mu} = \frac{1}{\sigma^2}(x-\mu) \\ \Rightarrow \quad & \frac{\partial^2 \log(f_\mu(x))}{\partial \mu^2} = -\frac{1}{\sigma^2}, \end{aligned}$$

ce qui donne l'information de Fisher

$$J(\mu) = -E\left(\frac{\partial^2 \log(f_\mu(x))}{\partial \mu^2}\right) = E\left(\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}.$$

On en déduit la borne de Cramér-Rao pour μ

$$\text{BCR} = \frac{\sigma^2}{n}.$$

La variance de $\overline{X}$ est

$$\text{var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{1}{n}\text{var}(X_i) = \frac{\sigma^2}{n}.$$

L'estimateur $\overline{X}$ atteint la borne de Cramér-Rao pour μ. De plus, rappelons que $\overline{X}$ est un estimateur non biaisé de μ. Il est donc le meilleur estimateur du point de vue de l'erreur carrée moyenne également.

7.19

1. L'espérance de la variable aléatoire X_i est

$$E(X_i) = \int_{-1}^{1}\frac{1}{2}x(1+\theta x)dx = \frac{1}{2}\left(\frac{x^2}{2} + \theta\frac{x^3}{3}\right)\Big|_{x=-1}^{x=1} = \frac{\theta}{3}.$$

   L'estimateur des moments $\hat{\theta}_M$ de θ est donc

$$\hat{\theta}_M = 3\overline{X}.$$

2. C'est un estimateur sans biais

$$\text{biais}(\hat{\theta}_M, \theta) = E(\hat{\theta}_M) - \theta = 3E(\overline{X}) - \theta = 3\cdot\frac{\theta}{3} - \theta = 0.$$

   Pour trouver sa variance, il est nécessaire de calculer d'abord $E(X_i^2)$

$$E(X_i^2) = \int_{-1}^{1}\frac{1}{2}x^2(1+\theta x)dx = \frac{1}{2}\left(\frac{x^3}{3} + \theta\frac{x^4}{4}\right)\Big|_{x=-1}^{x=1} = \frac{1}{3},$$

$$\Rightarrow \text{var}(X_i) = E(X_i^2) - E^2(X_i) = \frac{1}{3} - \frac{\theta^2}{9} = \frac{3-\theta^2}{9}.$$

   La variance de l'estimateur $\hat{\theta}_M$ est alors

$$\text{var}(\hat{\theta}_M) = \frac{9}{n}\text{var}(X_i) = \frac{3-\theta^2}{n}.$$

3. Par le théorème central limite, on sait que

$$\frac{\overline{X} - E(\overline{X})}{\sqrt{\text{var}(\overline{X})}} = \frac{\overline{X} - \theta/3}{\sqrt{\frac{3-\theta^2}{9n}}} \overset{n\to\infty}{\sim} \mathcal{N}(0,1).$$

Pour trouver le nombre moyen d'échantillons pour lesquels $\hat{\theta}_M$ est supérieur à la valeur de θ, on cherche la probabilité que $\hat{\theta}_M$ soit plus grand que 1, en sachant que $\theta = 1$

$$P(\hat{\theta}_M > 1) = P(\overline{X} > \frac{1}{3}) =$$
$$\overset{\theta=1}{=} \quad P\left(\frac{\overline{X} - \theta/3}{\sqrt{\frac{3-\theta^2}{9n}}} > \frac{1/3 - 1/3}{\sqrt{\frac{2}{9n}}}\right) = P(Z > 0) = \frac{1}{2},$$

où Z est une variable aléatoire issue d'une loi normale de paramètres $(0, 1)$. On en déduit qu'il y aura en moyenne $1\ 000 \cdot \frac{1}{2} = 500$ échantillons qui donneront un estimateur de θ supérieur à la vraie valeur du paramètre. Notons que, dans le cas présent, ce résultat ne dépend pas de la taille des échantillons.

7.20

1. Cherchons l'estimateur du maximum de vraisemblance $\hat{p}_{MV}$ de p. La vraisemblance est

$$L(p \mid x_1, \ldots, x_n) = (1-p)^{\sum_{i=1}^n x_i - n} p^n$$

et la log-vraisemblance

$$l(p \mid x_1, \ldots, x_n) = \log(1-p)\left(\sum_{i=1}^n x_i - n\right) + n \log p.$$

En maximisant $l(p \mid x_1, \ldots, x_n)$ par rapport à p, on obtient

$$\frac{\partial}{\partial p} l(p \mid x_1, \ldots, x_n) = \frac{n - \sum_{i=1}^n x_i}{1-p} + \frac{n}{p} = 0$$

et l'estimateur

$$\hat{p}_{MV} = \frac{1}{\overline{X}}.$$

2. L'espérance d'une variable aléatoire X suivant une loi géométrique est

$$E(X) = \frac{1}{p}.$$

On en tire l'estimateur des moments $\hat{p}_M$ de p :

$$\hat{p}_M = \frac{1}{\overline{X}}.$$

Les 2 estimateurs sont identiques.

7.21

1. Comme il n'y a qu'une seule mesure, la vraisemblance est égale à la fonction de distribution de X

$$L(p \mid x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

La log-vraisemblance s'écrit

$$l(p \mid x) = \log \binom{n}{x} + x \log p + (n-x) \log(1-p),$$

et sa maximisation par rapport à p donne

$$\frac{\partial}{\partial p} l(p \mid x) = \frac{x}{p} - \frac{n-x}{1-p} = 0.$$

On en déduit

$$\hat{p}_{MV} = \frac{X}{n}.$$

2. Dérivons une 2^{e} fois $L(p \mid x)$ par rapport à p

$$\frac{\partial^2}{\partial p^2} L(p \mid x) = -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}.$$

L'information de Fisher est alors

$$J(p) = -E\left(-\frac{x}{p^2} - \frac{n-x}{(1-p)^2}\right) = \frac{np}{p^2} - \frac{n-np}{(1-p)^2} = \frac{n}{p(1-p)},$$

et la borne de Cramér-Rao pour un estimateur non biaisé de p est

$$\text{BCR} = \frac{p(1-p)}{n}.$$

3. La variance de $\hat{p}_{MV}$ est

$$\text{var}(\hat{p}_{MV}) = \text{var}\left(\frac{X}{n}\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

L'estimateur atteint bien la borne de Cramér-Rao.

7.22

1. La fonction de distribution d'une variable aléatoire normale d'espérance μ et de variance σ^2 est

$$f_{\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

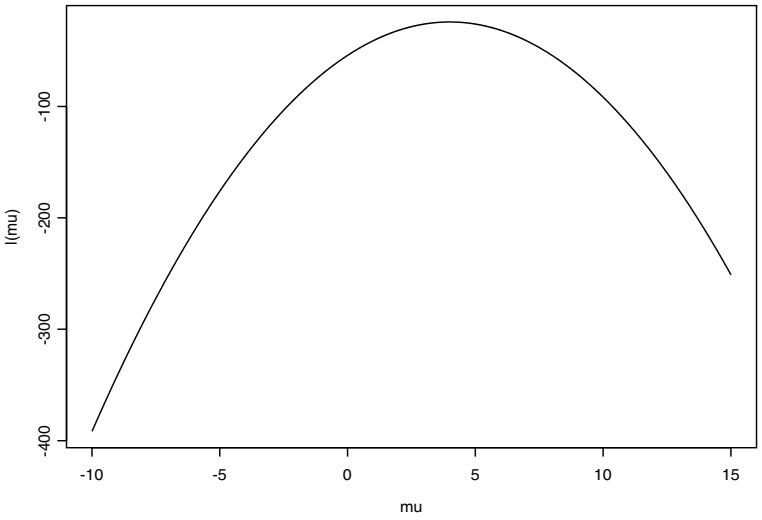


Fig. 7.3 – Graphe de la log-vraisemblance de μ de l'exercice 7.22.

La vraisemblance d'un échantillon de ces variables aléatoires est

$$L(\sigma^2 \mid x_1, \ldots, x_n) = (2\pi)^{-n/2}(\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right),$$

et la log-vraisemblance

$$l(\sigma^2 \mid x_1, \ldots, x_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Dérivons-la

$$\frac{\partial}{\partial(\sigma^2)} l(\sigma^2 \mid x_1, \ldots, x_n) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2,$$

d'où l'on obtient l'estimateur du maximum de vraisemblance $\hat{\sigma}^2_{MV}$ de σ^2

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.$$

2. On peut récrire la log-vraisemblance comme

$$\begin{aligned} l(\mu \mid x_1, \ldots, x_n) = \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \overline{x})^2 - \frac{n}{2\sigma^2}(\overline{x} - \mu)^2 \\ &= \text{const} - \frac{n}{2\sigma^2}(\overline{x} - \mu)^2. \end{aligned}$$

Le graphique de cette fonction est présenté à la figure 7.3. La log-vraisemblance a un maximum en $\mu = \overline{x}$ ce que l'on vérifie par l'optimisation suivante

$$\frac{\partial}{\partial\mu} l(\mu \mid x_1, \ldots, x_n) = \frac{n}{\sigma^2}(\overline{x} - \mu)^2 = 0,$$

qui implique

$$\hat{\mu}_{MV} = \overline{x}.$$

Pour savoir si l'estimateur du maximum de vraisemblance $\hat{\mu}_{MV}$ de μ est celui qui a la variance minimale, il faut calculer la borne de Cramér-Rao. Cherchons l'information de Fisher

$$\frac{\partial}{\partial\mu}\log(f_\mu(x)) = \frac{x - \mu}{\sigma^2}$$

$$\Rightarrow \frac{\partial^2}{\partial\mu^2}\log(f_\mu(x)) = -\frac{1}{\sigma^2}$$

$$\Rightarrow J(\mu) = -E\left(-\frac{1}{\sigma^2}\right) = \frac{1}{\sigma^2}.$$

La borne de Cramér-Rao pour un estimateur sans biais de μ est par conséquent

$$\text{BCR} = \frac{\sigma^2}{n}.$$

Or

$$\text{var}(\hat{\mu}_{MV}) = \text{var}(\overline{X}) = \frac{\sigma^2}{n}.$$

L'estimateur $\hat{\mu}_{MV}$ atteint la borne ; il n'existe donc pas d'estimateurs de μ ayant une variance plus petite.

7.23

1. Pour trouver l'estimateur des moments $\hat{\theta}_M$ de θ, calculons l'espérance de la variable aléatoire X

$$E(X) = \int_0^1 x(\theta + 1)x^\theta dx = \frac{\theta + 1}{\theta + 2}x^{\theta+2}\Big|_{x=0}^{x=1} = \frac{\theta + 1}{\theta + 2}.$$

On en déduit l'estimateur des moments

$$\frac{\hat{\theta}_M + 1}{\hat{\theta}_M + 2} = \overline{X} \quad \Leftrightarrow \quad \hat{\theta}_M = \frac{2\overline{X} - 1}{1 - \overline{X}}.$$

2. La log-vraisemblance de l'échantillon est

$$l(\theta \mid x_1, \ldots, x_n) = n \log(\theta + 1) + \theta \sum_{i=1}^{n} \log x_i.$$

On la dérive par rapport à θ et on cherche la racine pour obtenir l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ de θ

$$\frac{\partial}{\partial \theta} l(\theta \mid x_1, \ldots, x_n) = \frac{n}{\theta + 1} + \sum_{i=1}^{n} \log x_i = 0.$$

D'où

$$\hat{\theta}_{MV} = -\frac{n}{\sum_{i=1}^{n} \log x_i} - 1.$$

7.24

L'espérance et la variance d'une variable aléatoire X suivant une loi uniforme de paramètres (a, b) sont

$$E(X) = \frac{a+b}{2} \qquad \operatorname{var}(X) = \frac{(b-a)^2}{12}.$$

Le 2e moment de X est alors

$$E(X^2) = \operatorname{var}(X) + E^2(X) = \frac{a^2 + ab + b^2}{3}.$$

Les estimateurs des moments $\hat{a}_M$ et $\hat{b}_M$ de a et b sont les solutions du système

$$\begin{cases} \frac{\hat{a}_M + \hat{b}_M}{2} = \overline{X} \\ \frac{\hat{a}_M^2 + \hat{a}_M \hat{b}_M + \hat{b}_M^2}{3} = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \end{cases}$$

La 1re équation donne

$$\hat{a}_M = 2\overline{X} - \hat{b}_M,$$

ce qui conduit par substitution dans la 2e équation à

$$\hat{b}_M = \overline{X} \pm \sqrt{\frac{3}{n} \sum_{i=1}^{n} X_i^2 - 3\overline{X}^2}.$$

7.25

1. La fonction de distribution d'une variable aléatoire suivant une loi uniforme de paramètres $(0, \theta)$ est

$$f_\theta(x) = \begin{cases} 1/\theta & \text{si } 0 \leq x \leq \theta \\ 0 & \text{sinon.} \end{cases}$$

La vraisemblance d'un échantillon de taille n issu de cette loi s'écrit

$$L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) =$$
$$= \begin{cases} 1/\theta^n & \text{si } \max(x_1, \ldots, x_n) \leq \theta \text{ et } \min(x_1, \ldots, x_n) > 0 \\ 0 & \text{sinon.} \end{cases}$$

Cette fonction est représentée sur la figure 7.4. On voit qu'elle n'est pas dérivable en son maximum. Il est donc inutile de chercher à l'optimiser par une dérivation. La lecture du graphique indique clairement qu'elle atteint son maximum en $Y = \max(X_1, \ldots, X_n)$.

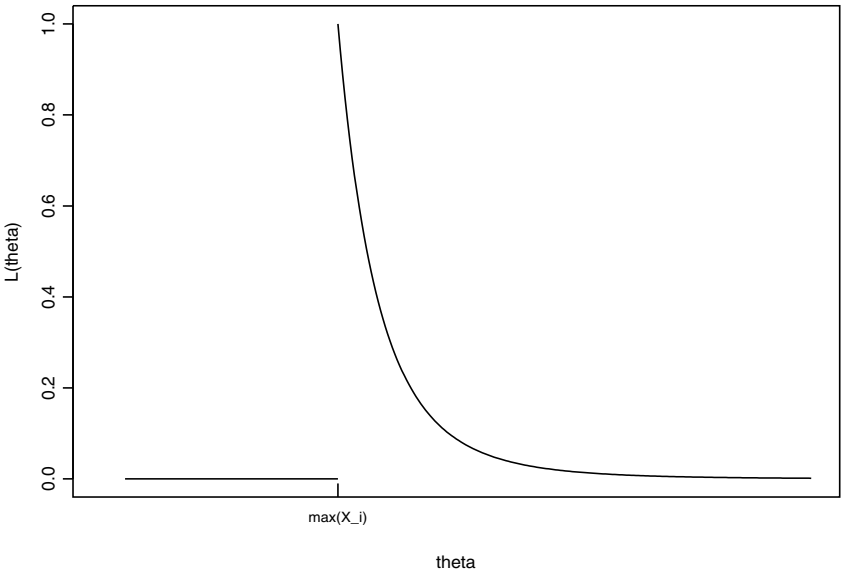


Fig. 7.4 – Graphe de la vraisemblance de θ de l'exercice 7.25.

2. Calculons la distribution de Y pour $n = 2$.

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(\max(X_1, X_2) < y) \\ &= P(X_1 < y, X_2 < y) \\ &\overset{\text{ind}}{=} P(X_1 < y) \cdot P(X_2 < y) = [P(X_1 < y)]^2 . \end{aligned}$$

De manière générale

$$F_Y(y) = [P(X_1 < y)]^n .$$

Par conséquent

$$F_Y(y) = \begin{cases} 0 & \text{si } y < 0 \\ (y/\theta)^n & \text{si } 0 \leq y \leq \theta \\ 1 & \text{si } y > \theta. \end{cases}$$

3. Afin de trouver le biais de l'estimateur Y, on calcule son espérance. Pour cela, notons d'abord que

$$f_Y(y) = \begin{cases} \theta^{-n} n y^{n-1} & \text{si } 0 \leq y \leq \theta \\ 0 & \text{sinon.} \end{cases}$$

Ainsi

$$E(Y) = \int_0^\theta \frac{n}{\theta^n} y^n dy = \frac{n}{\theta^n} \frac{y^{n+1}}{n+1}\Big|_{y=0}^{y=\theta} = \frac{n}{n+1}\theta.$$

L'estimateur Y est donc biaisé

$$\text{biais}(Y, \theta) = -\frac{\theta}{n+1}.$$

On propose alors le nouvel estimateur $\tilde{Y}$ sans biais suivant

$$\tilde{Y} = \frac{n+1}{n} Y = \frac{n+1}{n} \max(X_1, \ldots, X_n).$$

7.26

1. L'espérance d'une variable aléatoire X qui suit une distribution de Pareto est

$$E(X) = \frac{\alpha}{\alpha - 1} x_0.$$

L'estimateur des moments $\hat{\alpha}_M$ du paramètre α est par conséquent

$$\hat{\alpha}_M = \frac{\overline{X}}{\overline{X} - x_0}.$$

2. La fonction de densité conjointe de tout l'échantillon est

$$f_\alpha(x_1, \ldots, x_n) = \alpha^n x_0^{\alpha n} \left(\prod_{i=1}^{n} x_i \right)^{-(1+\alpha)},$$

et la log-vraisemblance est

$$\begin{aligned} L(\alpha \mid x_1, \ldots, x_n) &= \log(f_\alpha(x_1, \ldots, x_n)) = \\ &= n \log \alpha + \alpha n \log x_0 - (1+\alpha) \sum_{i=1}^{n} \log x_i. \end{aligned}$$

L'estimateur du maximum de vraisemblance $\hat{\alpha}_{MV}$ de α est la solution du maximum de L

$$\frac{\partial L(\alpha \mid x_1, \ldots, x_n)}{\partial \alpha} = \frac{n}{\alpha} + n \log x_0 - \sum_{i=1}^{n} \log x_i = 0.$$

Par conséquent

$$\hat{\alpha}_{MV} = \frac{n}{\sum_{i=1}^{n} \log x_i - n \log x_0}.$$

3. Avec la nouvelle paramétrisation, la fonction de densité conjointe est

$$f_\eta(x_1, \ldots, x_n) = \eta^{-n} x_0^{n/\eta} \left(\prod_{i=1}^{n} x_i \right)^{-1+1/\eta}.$$

La log-vraisemblance devient

$$L(\eta \mid x_1, \ldots, x_n) = -n \log \eta + \frac{n}{\eta} \log x_0 - (1 + \frac{1}{\eta}) \sum_{i=1}^{n} \log x_i,$$

et l'estimateur du maximum de vraisemblance $\hat{\eta}_{MV}$ de η devient la solution de

$$\frac{\partial L(\eta \mid x_1, \ldots, x_n)}{\partial \eta} = -\frac{n}{\eta} - \frac{n}{\eta^2} \log x_0 + \frac{1}{\eta^2} \sum_{i=1}^{n} \log x_i = 0.$$

Ainsi

$$\hat{\eta}_{MV} = \frac{\sum_{i=1}^{n} \log x_i - n \log x_0}{n}.$$

4. On remarque que

$$\hat{\alpha}_{MV} = \frac{1}{\hat{\eta}_{MV}}.$$

D'une manière générale, $\widehat{h(\theta)}_{MV} = h(\hat{\theta}_{MV})$ si h est monotone.

7.27

Comme les ε_i suivent une distribution normale de paramètres $(0, \sigma^2)$, les Y_i suivent eux aussi une distribution normale, mais de paramètres $(\alpha+\beta x_i, \sigma^2)$. La fonction de vraisemblance d'un échantillon de taille n d'observations y_i s'écrit alors

$$L(\alpha,\beta \mid x_1, \ldots, x_n, y_1, \ldots, y_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right)$$

et la log-vraisemblance

$$l(\alpha,\beta \mid x_1, \ldots, x_n, y_1, \ldots, y_n) = -n\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2.$$

Maximiser la log-vraisemblance par rapport à α et β revient à minimiser $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$, ce qui est le principe même de la détermination des estimateurs des moindres carrés.

7.28

1. On cherche l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ de θ. La fonction de vraisemblance d'un échantillon de n variables aléatoires distribuées selon une loi de Weibull est

   $$L(\theta \mid x_1, \ldots, x_n) = c^n\theta^{-n}\prod_{i=1}^{n} x_i^{c-1}\exp\left(-\frac{\sum_{i=1}^{n} x_i^c}{\theta}\right)$$

   et sa log-vraisemblance est

   $$l(\theta \mid x_1, \ldots, x_n) = n\log c - n\log\theta + (c-1)\sum_{i=1}^{n} x_i - \frac{1}{\theta}\sum_{i=1}^{n} x_i^c.$$

   Dérivons cette dernière pour trouver l'estimateur $\hat{\theta}_{MV}$:

   $$\frac{\partial}{\partial\theta} l(\theta \mid x_1, \ldots, x_n) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i^c$$

   $$\Rightarrow \hat{\theta}_{MV} = \frac{1}{n}\sum_{i=1}^{n} x_i^c.$$

2. Cherchons la fonction de densité de $Y = X^c$.

   $$F_Y(y) = P(Y < y) = P(X^c < y) = P(X < y^{1/c}) = F_X(y^{1/c})$$

$$\Rightarrow f_Y(y) = f_X(y^{1/c})\frac{1}{c}y^{1/c-1} = \frac{1}{\theta}\exp\left(-\frac{y}{\theta}\right).$$

Y est distribuée selon une loi exponentielle de paramètre $1/\theta$. On en déduit

$$E(\hat{\theta}_{MV}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^c\right) = \frac{1}{n}n\theta = \theta,$$

et

$$\mathrm{var}(\hat{\theta}_{MV}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^c\right) = \frac{1}{n^2}n\theta^2 = \frac{\theta^2}{n}.$$

3. Calculons l'information de Fisher

$$\frac{\partial}{\partial\theta}\log f_\theta(x) = -\frac{1}{\theta} + \frac{x^c}{\theta^2}$$

$$\Rightarrow \frac{\partial^2}{\partial\theta^2}\log f_\theta(x) = \frac{1}{\theta^2} - \frac{2x^c}{\theta^3}$$

$$\Rightarrow J(\theta) = -E\left(\frac{1}{\theta^2} - \frac{2X^c}{\theta^3}\right) = -\frac{1}{\theta^2} + \frac{2}{\theta^3}E(X^c) = \frac{1}{\theta^2}.$$

La borne de Cramér-Rao pour un estimateur non biaisé de θ est par conséquent

$$\mathrm{BCR} = \frac{\theta^2}{n}.$$

L'estimateur $\hat{\theta}_{MV}$ est donc à variance minimale.

7.29

1. La vraisemblance d'un échantillon de n variables aléatoires distribuées selon une loi log-normale d'espérance μ et de variance σ^2 s'écrit

$$L(\mu \mid x_1, \ldots, x_n) = c^n \prod_{i=1}^{n}\left[\frac{1}{x_i}\exp\left(-\frac{1}{2\sigma^2}(\log x_i - \mu)^2\right)\right]$$

et la log-vraisemblance

$$l(\mu \mid x_1, \ldots, x_n) = n\log c - \sum_{i=1}^{n}\log x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\log x_i - \mu)^2.$$

L'estimateur du maximum de vraisemblance $\hat{\mu}_{MV}$ de μ est alors la solution de

$$\frac{\partial}{\partial\mu}l(\mu \mid x_1, \ldots, x_n) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(\log x_i - \mu) = 0.$$

Donc

$$\hat{\mu}_{MV} = \frac{1}{n}\sum_{i=1}^{n}\log x_i.$$

2. Calculons l'espérance de l'estimateur $\hat{\mu}_{MV}$:

$$E(\hat{\mu}_{MV}) = \frac{1}{n}\sum_{i=1}^{n} E(\log X_i) = \mu.$$

L'estimateur n'est pas biaisé.

3. Sa variance est

$$\mathrm{var}(\hat{\mu}_{MV}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}(\log X_i) = \frac{\sigma^2}{n}.$$

On aimerait la comparer à la borne de Cramér-Rao pour un estimateur non biaisé de μ.

$$\log f_\mu(x) = \log c - \log x - \frac{1}{2\sigma^2}(\log x_i - \mu)^2$$

$$\Rightarrow \frac{\partial}{\partial \mu}\log f_\mu(x) = -\frac{1}{\sigma^2}(\log x_i - \mu)$$

$$\Rightarrow \frac{\partial^2}{\partial \mu^2}\log f_\mu(x) = -\frac{1}{\sigma^2}$$

$$\Rightarrow J(\mu) = \frac{1}{\sigma^2}$$

$$\Leftrightarrow \mathrm{BCR} = \frac{\sigma^2}{n}.$$

L'estimateur $\hat{\mu}_{MV}$ est un bon estimateur car sa variance atteint la borne de Cramér-Rao.

7.30

1. On veut trouver les estimateurs du maximum de vraisemblance $\hat{\beta}_{MV}$ de β et $\hat{\sigma}^2_{MV}$ de σ^2. À partir de la fonction de densité

$$f(y \mid \beta, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma y}\exp\left(-\frac{1}{2\sigma^2}(\log(y) - \mu)^2\right)$$

on calcule la vraisemblance

$$L(\beta, \sigma^2 \mid x, y) = \left(\frac{1}{(2\pi)^{n/2}\sigma^n \prod_i y_i}\right)\exp\left(-\frac{1}{2\sigma^2}\sum_i(\log(y_i) - \beta x_i)^2\right)$$

et la log-vraisemblance

$$\begin{aligned}&\log(L(\beta, \sigma^2 \mid x, y)) = \\ &\quad = -\sum_i \log(y_i) - \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i(\log(y_i) - \beta x_i)^2.\end{aligned}$$

Les estimateurs du maximum de vraisemblance sont les solutions des conditions de 1er ordre. Par conséquent, l'estimateur de β est

$$\begin{aligned}
&\frac{\partial \log L}{\partial \beta} = -\frac{1}{\sigma^2}\sum_i (\log(y_i) - \hat{\beta}_{MV} x_i)\cdot(-x_i) = 0 \\
\Leftrightarrow\quad &\sum_i \log(y_i)x_i - \hat{\beta}_{MV}\sum_i x_i^2 = 0 \\
\Leftrightarrow\quad &\hat{\beta}_{MV} = \frac{\sum_i \log(y_i)x_i}{\sum_i x_i^2},
\end{aligned}$$

et celui de σ^2 est

$$\begin{aligned}
&\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^2_{MV}} + \frac{1}{2\hat{\sigma}^4_{MV}}\sum_i (\log(y_i) - \hat{\beta}_{MV} x_i)^2 = 0 \\
\Leftrightarrow\quad &\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_i (\log(y_i) - \hat{\beta}_{MV} x_i)^2
\end{aligned}$$

2. Calculons l'information de Fisher pour σ^2

$$\begin{aligned}
&\frac{\partial \log(f(y))}{\partial(\sigma^2)} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(\log y - \mu)^2 \\
\Leftrightarrow\quad &\frac{\partial^2 \log(f(y))}{\partial(\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(\log y - \mu)^2 = \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}(\log y - E(\log y))^2 \\
\Leftrightarrow\quad &J(\sigma^2) = -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}E(\log y - E(\log y))^2 = -\frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\cdot\sigma^2 = \frac{1}{2\sigma^4}.
\end{aligned}$$

Nous avons déjà vu que l'espérance de l'estimateur $\hat{\sigma}^2_{MV}$ est

$$E(\hat{\sigma}^2_{MV}) = \frac{n-1}{n}\sigma^2.$$

Ainsi, si on pose $g(\hat{\sigma}^2_{MV}) = \frac{n-1}{n}\hat{\sigma}^2_{MV}$, on trouve la borne de Cramér-Rao correspondante, à savoir

$$\mathrm{BCR} = \frac{(g'(\hat{\sigma}^2_{MV}))^2}{nJ(\sigma^2)} = \frac{2(n-1)^2\sigma^4}{n^3}.$$

La variance de $\hat{\sigma}^2_{MV}$ est

$$\begin{aligned}
\mathrm{var}(\hat{\sigma}^2_{MV}) &= \mathrm{var}\left(\frac{1}{n}\sum_i (\log(y_i) - \hat{\beta}_{MV} x_i)^2\right) \\
&= \mathrm{var}\left(\frac{1}{n}\sum_i (\log(y_i) - \overline{\log y})^2\right) = \frac{2(n-1)}{n^2}\sigma^4.
\end{aligned}$$

Lors de la dernière égalité, on utilise la variance de $\tilde{s}^2$ trouvée dans l'exercice 7.4. On en déduit que $\hat{\sigma}^2_{MV}$ n'atteint pas la borne de Cramér-Rao.

7.31

1. L'espérance de la variable aléatoire X est

$$\begin{aligned} E(X) &= \int_0^1 x\alpha(\alpha+1)x^{\alpha-1}(1-x)dx \\ &= \alpha(\alpha+1)\left[\frac{x^{\alpha+1}}{\alpha+1} - \frac{x^{\alpha+2}}{\alpha+2}\right]_{x=0}^{x=1} = \frac{\alpha}{\alpha+2}. \end{aligned}$$

On en tire l'estimateur des moments $\hat{\alpha}_M$ de α

$$\frac{\hat{\alpha}_M}{\hat{\alpha}_M+2} = \overline{X} \quad \Leftrightarrow \quad \hat{\alpha}_M = \frac{2\overline{X}}{1-\overline{X}}.$$

2. Calculons en 1er lieu l'information de Fisher.

$$\begin{aligned} \log f_\alpha(x) &= \log\alpha + \log(\alpha+1) - (\alpha-1)\log x + \log(1-x) \\ \Rightarrow \frac{\partial}{\partial\alpha}\log f_\alpha(x) &= \frac{1}{\alpha} + \frac{1}{\alpha+1} + -\log x \\ \Rightarrow \frac{\partial^2}{\partial\alpha^2}\log f_\alpha(x) &= -\frac{1}{\alpha^2} - \frac{1}{(\alpha+1)^2} \\ \Rightarrow J(\alpha) &= \frac{(\alpha+1)^2+\alpha^2}{\alpha^2(\alpha+1)^2}. \end{aligned}$$

Ainsi, la borne de Cramér-Rao pour un estimateur non biaisé de α est

$$\text{BCR} = \frac{\alpha^2(\alpha+1)^2}{n(\alpha^2+(\alpha+1)^2)}.$$

Comparons-la à la variance de l'estimateur $\hat{\alpha}_M$:

$$\begin{aligned} \text{var}(\hat{\alpha}) - \text{BCR} &= \frac{\alpha(\alpha+2)^2}{2(\alpha+3)} - \frac{\alpha^2(\alpha+1)^2}{\alpha^2+(\alpha+1)^2} \\ &= \frac{\alpha(3\alpha^2+6\alpha+4)}{2(\alpha+3)(2\alpha^2+2\alpha+1)} > 0, \end{aligned}$$

car tous les facteurs sont positifs. La borne de Cramér-Rao est bien inférieure à la variance de l'estimateur $\hat{\alpha}_M$; l'estimateur n'est pas efficace.

7.32

Soit X_i une variable aléatoire correspondant au numéro de série d'une pièce électronique. La variable aléatoire X_i est distribuée selon une loi uniforme discrète de paramètres $(0, N)$. Cherchons l'estimateur des moments $\hat{N}_M$ de N

$$E(X) = \sum_{i=1}^{n} \frac{1}{N} i = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$$

$$\Leftrightarrow \frac{\hat{N}_M + 1}{2} = \overline{X} \quad \Leftrightarrow \hat{N}_M = 2\overline{X} - 1.$$

Les valeurs mesurées donnent $\overline{X} = 819{,}5$, donc $\hat{N}_M = 1\ 638$. Passons à l'estimateur du maximum de vraisemblance. Il est important de noter que, dans cet exercice, la fonction de vraisemblance n'est pas une fonction continue car N est un nombre entier. En d'autres termes, la maximisation par dérivation n'est pas faisable car le support n'est pas compact :

$$L(N \mid x_1, x_2) = \begin{cases} N^{-2} & \text{si } \max(x_i) \leq N \text{ et } \min(x_i) \geq 0 \\ 0 & \text{sinon.} \end{cases}$$

On se rend malgré tout bien compte que cette fonction atteint sa valeur maximale lorsque $\max(x_i) = N$. Par conséquent, l'estimateur du maximum de vraisemblance $\hat{N}_{MV}$ de N est

$$\hat{N}_{MV} = \max(X_i) = 888.$$

7.33

1. Le calcul de l'espérance de X donne

$$\begin{aligned} E(X) &= \int_0^\infty x \frac{x}{\theta} \exp\left(-\frac{x^2}{2\theta}\right) dx \\ &\overset{\text{par parties}}{=} -\frac{x}{\theta}\theta \exp\left(-\frac{x^2}{2\theta}\right)\Big|_{x=0}^{x=\infty} + \int_0^\infty \frac{1}{\theta}\theta \exp\left(-\frac{x^2}{2\theta}\right) dx \\ &= \sqrt{2\pi\theta} \int_0^\infty \frac{1}{\sqrt{2\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right) dx \\ &= \sqrt{2\pi\theta} \cdot \frac{1}{2} = \sqrt{\frac{\pi\theta}{2}}. \end{aligned}$$

   L'estimateur des moments $\hat{\theta}_M$ de θ est donc

$$\sqrt{\frac{\pi\hat{\theta}_M}{2}} = \overline{X}$$

$$\Leftrightarrow \hat{\theta}_M = \frac{2\overline{X}^2}{\pi}.$$

2. La vraisemblance d'un échantillon de n variables aléatoires X_i est

$$L(\theta \mid x_1, \ldots, x_n) = \frac{1}{\theta^n} \left(\prod_{i=1}^n x_i\right) \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\theta}\right)$$

   et la log-vraisemblance correspondante

$$l(\theta \mid x_1, \ldots, x_n) = -n \log\theta + \sum_{i=1}^n \log x_i - \frac{1}{2\theta} \sum_{i=1}^n x_i^2.$$

Dérivons-la par rapport à θ

$$\frac{\partial}{\partial\theta} l(\theta \mid x_1, \ldots, x_n) = -\frac{n}{\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n} x_i^2.$$

On obtient alors l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ de θ suivant

$$\hat{\theta}_{MV} = \frac{1}{2n}\sum_{i=1}^{n} x_i^2.$$

3. Nous savons que $X_i^2/2$ suit une loi exponentielle de paramètre $1/\theta$. Son espérance vaut donc θ et celle de l'estimateur

$$E(\hat{\theta}_{MV}) = \frac{1}{n}\sum_{i=1}^{n} E\left(\frac{X_i^2}{2}\right) = \theta.$$

L'estimateur est sans biais.

4. La variance de l'estimateur vaut

$$\mathrm{var}(\hat{\theta}_{MV}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}\left(\frac{X_i^2}{2}\right) = \frac{1}{n^2} n\theta^2 = \frac{\theta^2}{n}.$$

Calculons l'information de Fisher et la borne de Cramér-Rao pour un estimateur non biaisé de θ

$$\log f_\theta(x) = -\log\theta + \log x - \frac{x^2}{2\theta}$$

$$\Rightarrow \frac{\partial}{\partial\theta}\log f_\theta(x) = -\frac{1}{\theta} + \frac{x^2}{2\theta^2}$$

$$\Rightarrow \frac{\partial^2}{\partial\theta^2}\log f_\theta(x) = \frac{1}{\theta^2} - \frac{x^2}{\theta^3}$$

$$\Rightarrow J(\theta) = -\frac{1}{\theta^2} + \frac{2}{\theta^3} E\left(\frac{X^2}{2}\right) = -\frac{1}{\theta^2} + \frac{2}{\theta^2} = \frac{1}{\theta^2}$$

$$\Leftrightarrow \mathrm{BCR} = \frac{\theta^2}{n}.$$

L'estimateur atteint donc la borne de Cramér-Rao. Il n'en existe pas d'autres avec une variance inférieure, et ceci quelle que soit la valeur de n.

7.34

1. La fonction de vraisemblance s'écrit

$$L(\beta \mid y_1, \ldots, y_n, x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}.$$

Par conséquent, la log-vraisemblance est

$$\begin{aligned} & l(\beta \mid y_1, \ldots, y_n, x_1, \ldots, x_n) \\ &= \log L(\beta \mid y_1, \ldots, y_n, x_1, \ldots, x_n)) \\ &= \sum_{i=1}^{n} (y_i \log p + (1-y_i)\log(1-p)) \\ &= \sum_{i=1}^{n} \left(y_i \log \left(\frac{1}{1+e^{-\beta x_i}} \right) + (1-y_i) \log \left(1 - \frac{1}{1+e^{-\beta x_i}} \right) \right). \end{aligned}$$

2. Établissons l'équation de vraisemblance

$$\frac{\partial \log L(\beta \mid y_1, \ldots, y_n, x_1, \ldots, x_n)}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - \frac{1}{1+e^{-\hat{\beta}_{MV} x_i}} \right) x_i = 0.$$

Notons que cette équation peut s'écrire sous la forme $\sum_{i=1}^{n} r_i x_i = 0$, où

$$r_i = y_i - \frac{1}{1+e^{-\hat{\beta}_{MV} x_i}}$$

est un résidu.

3. L'équation normale pour l'estimation des moindres carrés de β de la régression linéaire classique est

$$\sum_{i=1}^{n} r_i x_i = 0,$$

où $r_i = y_i - \hat{\beta}_{MV} x_i$.

Chapitre 8

Inférence

Introduction

Ce chapitre est d'abord une extension du chapitre précédent dans le sens où il étend la notion d'estimation ponctuelle à la notion d'estimation par intervalles. Ensuite, il aborde la théorie des tests. À partir de l'échantillon on veut répondre à la question sur un paramètre ou une caractéristique de la population : est-ce que les dépense mensuelles en cosmétiques dépendent du statut professionnel, est-ce que les garçons ont plus de chance d'être prématurés que les filles, etc.

Formellement, on teste une hypothèse H_0 que l'on appelle l'hypothèse nulle contre une hypothèse H_A (ou H_1), l'hypothèse alternative. La statistique de test permet de rejeter ou pas l'hypothèse H_0.

Les statistiques considérées ici sont les tests basés sur la moyenne, le test t de Student, le test du χ^2 (test d'adéquation et test d'indépendance), les tests sur les paramètres de régression et les tests dérivés de la théorie de Neyman-Pearson. Ces tests permettent de traiter plusieurs situations courantes.

Les propriétés d'une statistique de test se quantifient avec la notion d'erreur de 1re espèce (α, la probabilité de rejeter H_0 alors qu'elle est vraie), l'erreur de 2^{e} espèce ($1-\beta$, la probabilité de ne pas rejeter H_0 alors qu'elle est fausse) et la notion de puissance (β, qui est le complément à 1 de l'erreur de 2^{e} espèce).

Deux approches s'affrontent historiquement : l'approche par la p-valeur (Fisher) et l'approche par les valeurs critiques, qui définissent une région de rejet (Neyman-Pearson). Dans les deux cas, il s'agit de comparer la valeur que prend la statistique sur l'échantillon à la distribution de la statistique si l'hypothèse nulle était vraie (cf. figure 8.1). Avec la p-valeur on calcule la probabilité sous H_0 que l'on observe une valeur autant voire plus extrême que celle que l'on a observée sur l'échantillon. Si cette probabilité est faible, cela indique que nous sommes face à un événement peu probable sous H_0 et produit de l'évidence contre H_0. Dans l'approche par la valeur critique, après avoir fixé d'avance le

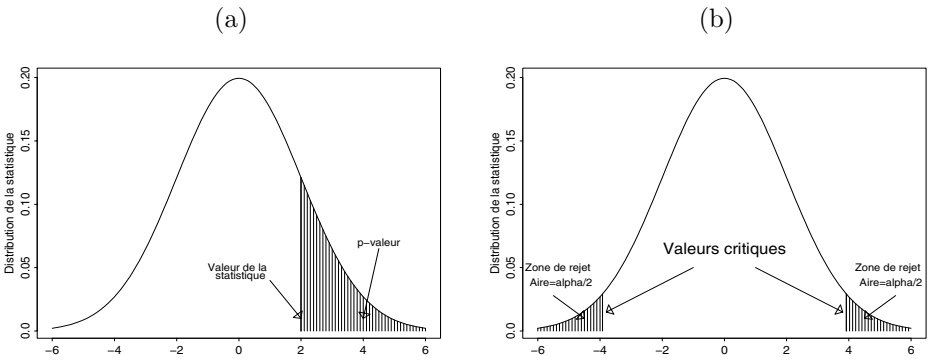


Fig. 8.1 – Tests statistiques : approche par la p-valeur (a) et par les valeurs critiques (b).

seuil α (généralement 1, 5, ou 10 %), on détermine les valeurs critiques (pour la statistique) à partir desquelles on rejettera H_0.

Il existe un lien entre l'estimation par intervalle et les tests statistiques, raison pour laquelle les deux sujets ont été regroupés ici. En effet, si la statistique de test est un pivot (c'est-à-dire que sa distribution ne dépend pas des paramètres) alors tester l'hypothèse nulle selon laquelle un certain paramètre est égal à une valeur donnée (avec alternative bilatérale) revient à regarder si cette valeur donnée appartient à l'intervalle de confiance.

Notes historiques

La théorie des tests a été formalisée par J. Neyman et E. Pearson à la fin des années 1920 et au début des années 1930. Il est à noter que le test du chi-carré d'adéquation avait été proposé préalablement par K. Pearson (vers 1900).

Le test t a été développé par Student, pseudonyme de W. S. Gossett (1876-1937), alors qu'il travaillait dans une brasserie. Le test du χ^2 d'adéquation et d'indépendance est dû à K. Pearson (1857-1936). Il a été désigné comme une des vingt découvertes majeures qui ont marqué le XX^e siècle (cf. *20 Discoveries That Changed Our Lives*, Special Issue of *Science 84* en 1984).

Références (théorie)

Dodge, chapitres 11 à 13 [5]; Lejeune, chapitres 7 et 9 [3]; Moore et McCabe, chapitres 6 à 8 [10]; Morgenthaler, chapitres 8 et 9 [4]; et Rice, chapitre 9 [9].

Exercices

Intervalles de confiance

8.1

Un fabricant de voitures fait un test d'efficacité d'un certain modèle. On mesure les litres de consommation d'essence pour 100 kilomètres

14,60 11,21 15,56 11,37 13,68 11,06 26,58
13,37 15,98 12,07 13,22 12,01 15,07.

On connaît la variance de la population : elle vaut $\sigma^2 = 16$.

1. Calculer un intervalle de confiance à 95 % pour la moyenne.
2. Quel niveau correspond à un intervalle de longueur 3 litres/100 km ?
3. Combien d'observations additionnelles sont nécessaires pour avoir un intervalle à 99 % de longueur 2 litres/100 km ?

8.2

On considère un portefeuille avec n actifs $R_1, \ldots, R_n$ indépendants, où l'on suppose que $R_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$.
Soit R_p le rendement du portefeuille : $R_p = \sum_{i=1}^n a_i R_i$ avec $\sum_{i=1}^n a_i = 1$.

1. Donner la distribution de R_p.
2. Supposons maintenant que les σ_i soient connus. Calculer un intervalle de confiance à 95 % pour $\mu_p = E(R_p)$.
3. Application

   $n = 16$, $a_i = \begin{cases} 2{,}5\ \% & i = 1, \ldots, 8 \\ 10\ \% & i = 9, \ldots, 16 \end{cases}$ et $\sigma_i = \begin{cases} 4\ \% & i = 1, \ldots, 8 \\ 1\ \% & i = 9, \ldots, 16 \end{cases}$

   On a observé les réalisations $r_i = \begin{cases} 6\ \% & i = 1, \ldots, 8 \\ 9\ \% & i = 9, \ldots, 16 \end{cases}$

   Calculer l'intervalle de confiance à 95 % pour μ_p.

8.3

Dans un test de fabrication de composants d'une chaîne hi-fi, la baisse de puissance de sortie des circuits électriques après 2 000 heures d'utilisation a été mesurée. Un essai sur 80 composants identiques a donné les résultats suivants : la baisse de puissance moyenne est de 12 watts. Par ailleurs il est connu que l'écart-type de la baisse de puissance pour ce type de circuit électrique est $\sigma = 2$ watts.

1. Calculer l'intervalle de confiance (approximatif) à 95 % de la baisse de puissance moyenne de la fabrication.

2. Recalculer l'intervalle pour un niveau de confiance plus élevé, notamment 99 %.
3. Vérifier que l'intervalle obtenu en 2. est plus large que celui obtenu en 1. Expliquer ce fait.

8.4

On voudrait estimer la vitesse moyenne μ à laquelle est parcouru le contournement de Genève. Les variables aléatoires $X_1, \ldots, X_n$ représentent les mesures effectuées sur un échantillon pris uniquement parmi les hommes. On suppose qu'elles proviennent d'une loi normale d'espérance inconnue μ et de variance connue σ_H^2. En deuxième lieu on récolte un échantillon de mesures de même taille uniquement parmi les femmes $Y_1, \ldots, Y_n$. Ces mesures proviennent d'une loi normale d'espérance inconnue μ et de variance connue σ_F^2. Pour estimer μ on considère l'estimateur $\hat{\mu} = \frac{1}{2}(\bar{X} + \bar{Y})$. Donner un intervalle de confiance à 95 % pour μ basé sur $\hat{\mu}$.

8.5

Soient $X_1, \ldots, X_n$ n observations indépendantes provenant d'une loi avec densité

$$f_\lambda(x) = \begin{cases} \lambda^2 x \exp(-\lambda x) & x > 0 \\ 0 & \text{sinon.} \end{cases}$$

1. Sachant que $\mu = E(X_i) = 2/\lambda$, exprimer $\mathrm{var}(X_i)$ en fonction de μ.
2. Donner un intervalle de confiance approximé, au degré 95 %, pour μ basé sur $\sum_{i=1}^n X_i$.

8.6

Un institut de recherche planifie un sondage pour estimer la proportion de Suisses favorables à la révision du système de sécurité sociale (AVS). Le financement de l'étude permet d'analyser un échantillon de 200 personnes au maximum. On aimerait obtenir un intervalle de confiance à 95 % avec une longueur maximale de 10 % pour la vraie proportion de personnes favorables à la proposition.

1. Déterminer la faisabilité de l'étude sur la base du financement prévu.
2. Quel degré de confiance peut être obtenu pour l'intervalle sur la base du financement ?

8.7

On simule à l'ordinateur 1 000 échantillons de taille 100 à partir de la loi $\mathcal{N}(5,1)$. Pour chaque échantillon on calcule l'intervalle

$$I = [\bar{Y} - 0{,}1\,;\, \bar{Y} + 0{,}1],$$

où $\bar{Y}$ est la moyenne de l'échantillon.

Approximativement combien d'intervalles contiendront la valeur 5 ? Justifiez votre réponse.

8.8

Soient $Y_1, \ldots, Y_n$ des variables aléatoires indépendantes et identiquement distribuées suivant une loi normale d'espérance μ et de variance 1. Soit

$$[\bar{Y} - b_1/\sqrt{n}; \bar{Y} + b_2/\sqrt{n}]$$

un intervalle de confiance pour μ, où b_1 et b_2 sont des constantes positives.

1. Exprimer le degré de confiance de cet intervalle en fonction de b_1 et b_2.
2. Sous la contrainte d'un degré de confiance fixé à 90 %, quelles sont les valeurs b_1 et b_2 qui minimisent la longueur de l'intervalle ?

8.9

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes provenant d'une distribution uniforme sur l'intervalle $(k\theta\,, (k+1)\theta)$, où k est une constante fixée d'avance. La fonction de densité de chaque X_i est

$$f_\theta(x) = \begin{cases} \frac{1}{\theta} & \text{si } k\theta \leq x \leq (k+1)\theta \\ 0 & \text{sinon.} \end{cases}$$

1. Calculez $\hat{\theta}_M$ l'estimateur des moments pour θ. Est-il consistant ?
2. Donner l'erreur carrée moyenne $\mathrm{ECM}(\hat{\theta}_M, \theta)$.
3. Quelle est la distribution approximative (n grand) de $\hat{\theta}_M$?
4. Construire un intervalle de confiance (approximatif) pour θ au degré de confiance de 95 %.
5. Que peut-on dire de la longueur de cet intervalle lorsque k augmente ?

8.10

Soient $X_1, \ldots, X_n$ des observations indépendantes et identiquement distribuées selon une distribution dont la densité est

$$f_\theta(x) = \frac{\theta^3}{2} x^2 e^{-\theta x}, \quad \theta > 0,\, x \geq 0.$$

1. Trouver l'estimateur des moments de θ.
2. Construire un intervalle de confiance approximatif au seuil de confiance de 95 % pour θ basé sur la moyenne des observations.

Indication : la distribution de X_i est une loi Gamma $(3, \theta)$.

8.11

Un biochimiste étudie un type de moisissure qui attaque les cultures de blé. La toxine contenue dans cette moisissure est obtenue sous forme d'une solution organique. On mesure la quantité de substance par gramme de solution. Sur 9 extraits on a obtenu les mesures suivantes exprimées en milligrammes

$$1{,}2 \quad 0{,}8 \quad 0{,}6 \quad 1{,}1 \quad 1{,}2 \quad 0{,}9 \quad 1{,}5 \quad 0{,}9 \quad 1{,}0.$$

On suppose que l'écart-type est égal à $\sigma = 0{,}3$.

1. Calculer la moyenne de cet échantillon.
2. Déterminer un intervalle de confiance à 95 % pour l'espérance de la quantité de substance toxique par gramme de solution.
3. Le biochimiste trouve que l'intervalle obtenu n'est pas satisfaisant car il est trop long. Que doit-il faire pour obtenir une estimation plus précise?

Concepts de base sur les tests

8.12

Les situations suivantes requièrent un test d'hypothèse pour l'espérance μ d'une population. On suppose dans les 2 cas que les données proviennent d'une loi $\mathcal{N}(\mu, \sigma^2)$, avec σ^2 connu.

– Un certain type d'expérience sur les rats mesure le temps qu'un rat met pour sortir d'un labyrinthe. Pour un certain type de labyrinthe le temps moyen est de 18 secondes. Un chercheur soupçonne qu'un bruit très fort aura comme conséquence de faire diminuer cette durée. Il mesure le temps mis par 10 rats pour sortir du labyrinthe dans ces nouvelles conditions (on suppose que $\sigma^2 = 2{,}25$).

$$16{,}0 \quad 19{,}2 \quad 16{,}9 \quad 18{,}5 \quad 17{,}2 \quad 15{,}5 \quad 18{,}9 \quad 14{,}3 \quad 17{,}3 \quad 17{,}5$$

– La surface moyenne de plusieurs appartements d'un nouveau complexe résidentiel est annoncée par le propriétaire comme étant de 100 mètres carrés. La gérance chargée de la vente de ces locaux pense que les appartements sont plus petits. Elle envoie son ingénieur pour mesurer la surface d'un échantillon de ces appartements ($\sigma^2 = 25$).

$$96 \quad 92 \quad 100 \quad 99 \quad 103 \quad 101 \quad 96 \quad 109 \quad 106 \quad 94 \quad 95 \quad 100 \quad 109 \quad 98$$

Pour chacun des deux cas ci-dessus :

1. donner l'hypothèse nulle et proposer la procédure pour la vérifier ;
2. calculer la p-valeur et conclure.

8.13

Pour déterminer si le contenu effectif de nicotine d'une certaine marque de cigarettes est plus élevé que ce qui est annoncé sur le paquet (1,4 mg), on procède à un test sur un échantillon de taille $n = 25$ cigarettes. Pour cet échantillon, on obtient $\bar{x} = 1{,}79$. On suppose que la variance d'une mesure vaut $\sigma^2 = 1$.

1. Formuler l'hypothèse nulle et l'hypothèse alternative dans le cas de cette étude.
2. Définir et calculer la p-valeur.
3. Est-ce que le résultat est significatif au niveau de 5 % ?
4. Est-ce que le résultat est significatif au niveau de 1 % ?

8.14

Pour vérifier si la naissance de prématurés est liée au sexe du bébé, le Service de Maternité de l'Hôpital d'une grande ville nous fournit les données concernant l'année 1995. Il y a eu 512 cas de prématurés, dont 284 garçons.
Décrire une procédure pour répondre à la question : « Est-ce que les garçons ont plus de chance d'être prématurés que les filles ? »

8.15

Sur un tronçon autoroutier on a observé en moyenne 8 accidents par année pendant les dernières années. Cette année il y a eu 5 accidents. En postulant une distribution de Poisson pour le nombre d'accidents par année, est-il possible de conclure au seuil de 5% que le taux d'accidents a diminué cette année ?

8.16

Dans les années 1970, les athlètes féminines d'Allemagne de l'Est étaient réputées pour leur forte corpulence. Le comité d'éthique olympique de l'époque, mettant en doute cette étonnante « virilité », avait fait appel aux services du docteur Volker Fischbach. Celui-ci sélectionna 9 athlètes féminines présentant des caractéristiques quasiment identiques, puis effectua des analyses mesurant leur quantité de substances hormonales virilisantes (dites androgènes) par litre de sang. Les résultats sont les suivants

3,22 3,07 3,17 2,91 3,40 3,58 3,23 3,11 3,62.

(On peut considérer que ces mesures sont indépendantes et proviennent de la loi normale.) On veut tester l'hypothèse : « les athlètes allemandes ne sont pas dopées. »En sachant que chez une femme la quantité moyenne d'androgènes est de $\mu = 3{,}1$, le docteur Fischbach a-t-il rejeté l'hypothèse ?

8.17

Soient 20 variables aléatoires $X_1, \ldots, X_{20}$ indépendantes et identiquement distribuées selon une loi $\mathcal{N}(\mu, \sigma^2)$. On veut tester l'hypothèse $H_0 : \sigma^2 = \sigma_0^2$ et pour cela on utilise

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

1. Quelle est la distribution de $\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2$ sous H_0 ?
2. Calculer la valeur critique k_α pour un test unilatéral ($H_1 : \sigma^2 > \sigma_0^2$) au seuil de 5 %.

8.18

Douze paires d'agneaux jumeaux sont sélectionnés pour comparer 2 régimes I et II ; dans chaque paire les agneaux ont été nourris avec des régimes différents. Les poids des agneaux après 8 mois sont les suivants

paire	1	2	3	4	5	6	7	8	9	10	11	12
I	86	96	82	103	91	88	94	90	97	87	97	105
II	106	118	106	91	96	102	100	100	108	105	118	93

En admettant qu'à la base les 12 paires sont semblables et en considérant un modèle normal, tester H_0: « les régimes I et II sont équivalents » contre H_1: « le régime I est différent du régime II. »

8.19

Les teneurs en azote de 30 échantillons d'un même terrain ont été mesurées par 2 méthodes d'analyse différentes : 15 échantillons par la méthode A et 15 autres par la méthode B. Sur la base des résultats suivants et en considérant un modèle normal, peut-on admettre, au seuil de 5 %, que les 2 méthodes donnent en moyenne des résultats analogues ?

- méthode A
  3,51 3,01 3,33 3,31 3,54 3,17 3,50 2,72 3,24 3,48 1,97 2,85 2,51 2,93 1,83
- méthode B
  2,34 3,12 3,30 2,15 3,13 2,84 2,97 3,65 3,89 3,23 2,46 2,70 2,44 2,41 2,85

8.20

Une entreprise pharmaceutique désire tester l'effet de 2 somnifères (Sopo et Dodo). Le médecin chargé de l'étude vous apporte les mesures (nombre d'heures de sommeil additionnel par rapport à la moyenne habituelle) effectuées sur 20 individus.

individu	1	2	3	4	5	6	7	8	9	10
Sopo	-0,5	-0,1	0,9	1,0	1,5	1,8	1,9	3,1	4,5	4,8
individu	11	12	13	14	15	16	17	18	19	20
Dodo	-1,1	0,0	0,1	1,2	1,2	0,4	1,1	2,1	2,3	3,2

On veut tester l'égalité des effets des 2 somnifères. Nous supposons en plus que les données suivent une distribution normale.

1. Définissez le test que vous pouvez faire dans ce cas, en posant les conditions nécessaires, ainsi que l'hypothèse nulle et l'hypothèse alternative.
2. Effectuez le test pour un seuil de $\alpha = 5$ %. Concluez.
3. En présentant les résultats de l'étude au médecin qui vous a chargé de la mener, vous réalisez que vous avez mal compris et qu'il s'agit en fait d'une expérience où chacun des 10 individus a testé chacun des 2 somnifères, ce qui donne le tableau suivant

individu	1	2	3	4	5	6	7	8	9	10
Sopo	-0,5	-0,1	0,9	1,0	1,5	1,8	1,9	3,1	4,5	4,8
Dodo	-1,1	0,0	0,1	1,2	1,2	0,4	1,1	2,1	2,3	3,2

   Ainsi on veut tester de nouveau l'égalité de l'effet des 2 somnifères.
   (a) Définissez de nouveau les hypothèses nulle et alternative, et effectuez le test au seuil de $\alpha = 5$ %.
   (b) Lequel des 2 tests effectués est correct ? Justifiez.

8.21

Le nombre d'arrivées par heure au guichet d'une grande banque suit une loi de Poisson $\mathcal{P}(\lambda)$. Pour améliorer le service à la clientèle, la banque nous confie la tâche de vérifier si on peut supposer que le nombre d'arrivés par heure est égale à 25, en sachant que pour $n = 30$ échantillons observés la semaine passée on a obtenu $\bar{x} = 27$. Que peut-on lui répondre ?

Propriétés des tests

8.22

Les bouteilles d'une boisson très populaire devraient contenir 300 mL. Comme la machine qui remplit ces bouteilles n'est pas très précise, il existe des différences d'une bouteille à l'autre. On peut supposer que la distribution du

contenu d'une bouteille est normale avec écart-type égal à 3 mL. Un étudiant qui soupçonne que les bouteilles sont moins remplies qu'elles le doivent, mesure le contenu de 6 bouteilles et obtient les valeurs

$$299{,}4 \quad 297{,}7 \quad 301{,}0 \quad 298{,}9 \quad 300{,}2 \quad 297{,}0.$$

Est-ce qu'on peut affirmer que le contenu des bouteilles est inférieur à 300 mL ?

Un calcul de puissance nous permet de déterminer à partir de quelle quantité de liquide manquant le test arrive à détecter le mauvais remplissage. Pour cela on suppose un niveau de test $\alpha = 5$ %.

1. Déterminer la puissance du test par rapport à l'alternative $\mu = 299$.
2. Déterminer la puissance du test par rapport à l'alternative $\mu = 295$.
3. Est-ce que la puissance par rapport à l'alternative $\mu = 290$ va être supérieure ou inférieure à celle trouvée en 2. ? Expliquer les raisons de ce comportement.

8.23

Deux types de pièces de monnaie sont produites dans une fabrique : des pièces homogènes et des pièces mal équilibrées, lesquelles montrent la face pile dans 55 % des cas. Supposons que nous possédons une pièce dont nous ignorons la provenance. Pour pouvoir déterminer de quelle pièce il s'agit, nous effectuons le test suivant : la pièce est lancée 1 000 fois ; si l'on obtient pile 525 fois ou plus, alors on conclut que c'est une pièce mal équilibrée, tandis que si l'on obtient pile moins de 525 fois, alors on conclut que c'est une pièce homogène.

1. Si la pièce est réellement homogène, quelle est la probabilité que l'on aboutisse à une conclusion fausse ?
2. Si la pièce est réellement mal équilibrée, quelle est la probabilité que l'on aboutisse à une conclusion fausse ?

8.24

On fait passer un test d'aptitude à 500 étudiants genevois. Les résultats indépendants de ces 500 étudiants donnent une moyenne de $\overline{x} = 461$.

1. En supposant que l'écart-type de la population est connu et égal à 100, peut-on dire que l'espérance des résultats de tous les étudiants de Genève n'est pas supérieure à 450 ?
2. Est-ce que le test utilisé au seuil de 5 % est suffisamment puissant pour détecter une augmentation de 10 points dans le résultat des étudiants ?

8.25

On désire tester si la durée de vie moyenne d'un tube électronique est égale à 1 600 heures ou si elle est plutôt inférieure à cette valeur. Les observations sur un échantillon de taille 16 suivent une loi normale avec $\overline{X} = 1\ 590$ heures et $\hat{\sigma} = s = 30$ heures.

1. Donner les hypothèses H_0 et H_A.
2. Quelle statistique de test utilisez-vous ?
3. Peut-on rejeter H_0 à un seuil de 1 % ?
4. Calculer l'erreur de 2^{e} espèce et la puissance du test au seuil de 1 % pour $\mu = 1\ 570$.

8.26

Soient $X_1, \ldots, X_n$ provenant d'une loi Gamma (θ, α). Donner le test le plus puissant pour tester :

1. $H_0 : \theta = \theta_0$ contre $H_A : \theta = \theta_1$;
2. $H_0 : \theta = \theta_0$ contre $H_A : \theta > \theta_0$.

8.27

La durée de vie des ampoules servant à l'illumination des auditoriums est modélisée par une loi exponentielle $\mathcal{E}(\lambda)$ avec densité

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & x > 0 \\ 0 & \text{sinon.} \end{cases}$$

Pour optimiser le service d'entretien (coupures budgétaires obligent), on veut savoir si la durée de vie de ces ampoules est de 30 jours. La service d'entretien procède à des mesures sur un échantillon de taille $n = 400$ pour lequel il observe $\bar{x} = 27{,}6$. Avec ces données peut-on infirmer que la durée est de 30 jours ?

Pour un seuil $\alpha = 5$ % fixé, calculer la puissance du test pour les valeurs suivantes de l'alternative : durée de vie de 25, 28 et 35 jours.

8.28

Le point de fusion de 16 échantillons d'une marque d'huile végétale a été déterminé et on a obtenu $\overline{x} = 94{,}32$. On suppose que la distribution de ce point de fusion suit une loi normale avec espérance μ et écart-type $\sigma = 1{,}2$.

1. Tester $H_0 : \mu = 95$ contre $H_1 : \mu < 95$ avec un seuil $\alpha = 0{,}01$.
2. Si la vraie valeur de μ est 94, quelle est la valeur de la puissance pour le test défini en 1. ?

3. Quelle est la valeur de n nécessaire pour assurer que l'erreur de 2^e espèce soit égale à 0,01 pour un niveau $\alpha = 0{,}01$?

8.29

Une banque désire vérifier l'hypothèse que l'omission des frais sur les cartes de crédits des clients qui ont un chiffre d'affaire annuel supérieur à 5 200 € amène à une augmentation du chiffre d'affaire annuel moyen. Cette offre d'omission de frais est accordée à un échantillon aléatoire de 200 clients et le chiffre d'affaire obtenu dans l'année courante est comparé avec celui de l'année précédente. L'augmentation moyenne du chiffre d'affaire parmi les clients de l'échantillon est égale à 332 € avec une variance estimée *de l'augmentation moyenne* égale à 108^2 €^{2}.

1. Peut-on dire, à un seuil de 1 %, que l'offre de la banque a généré une augmentation du chiffre d'affaire ?
2. Donner un intervalle de confiance au degré 99 % pour l'espérance de l'augmentation annuelle μ du chiffre d'affaire.
3. Utiliser l'approximation normale pour déterminer la puissance du test effectué au point 1. par rapport à l'alternative que l'espérance de l'augmentation μ est égale à 150 €.
4. Quelle est la taille n de l'échantillon que la banque doit choisir pour que la puissance du test par rapport à l'alternative $\mu = 150$ soit égale à 80 % ?

8.30

Vingt informaticiens ont installé chacun soit Linux, soit WinNT. Le temps nécessaire (en minutes) à chacun pour l'installation est répertorié dans le tableau suivant.

Linux	WinNT
154	145
164	162
198	156
168	152
180	168
172	157
142	155
165	140
172	145
158	160

On suppose que les données proviennent de la loi normale.

1. Calculer l'intervalle de confiance de la durée moyenne d'installation de chacun des 2 logiciels.

2. Par un test statistique, déterminer au seuil de 5 % si la durée d'installation de Linux est supérieure à celle de WinNT.
3. Supposons maintenant que les variances des temps d'installation sont connues : $\sigma_L^2 = 225$ pour Linux et $\sigma_W^2 = 100$ pour WinNT.
   (a) Que devient la statistique de test utilisée au point 2. ?
   (b) Quelle est sa distribution ?
   (c) Quelle taille d'échantillon faudrait-il pour garantir une puissance de 80 % afin de détecter des différences de durée de 10 minutes ?

Test du chi-carré

8.31

Une étude a examiné le lien qui existe entre le fait de fumer et le fait d'être droitier ou gaucher. Un échantillon de 400 gauchers a révélé que 190 fumaient, alors que dans un échantillon de 800 droitiers on a trouvé 300 fumeurs. Tester si la proportion de fumeurs est identique dans les 2 catégories.

8.32

Le score Z d'un test d'intelligence est supposé être distribué selon une loi normale de moyenne $\mu = 100$ et de variance $\sigma^2 = 225$. Un échantillon de 1 000 personnes est testé. On obtient les scores suivants

Score	[0,70)	[70,85)	[85,100)	[100,115)	[115,130)	[130,∞)
Nombre	34	114	360	344	120	28

Sur la base de cet échantillon, tester au seuil $\alpha = 5$ % que la variable Z suit une loi $\mathcal{N}(100, 225)$.

8.33

À la suite du croisement de 2 lignées d'une certaine variété de plantes, on devrait obtenir 75 % de plantes à fleurs rouges et 25 % de plantes à fleurs blanches, si la couleur obéit à la loi de Mendel. Pour vérifier si cette loi s'applique à la couleur des fleurs dans le cas présent, nous avons mis en culture 500 plantes, dont 350 ont donné des fleurs rouges. Peut-on affirmer que la loi de Mendel a été respectée ?

8.34

On considère ici les évaluations fournies par les étudiants au cours de Statistique de l'Université de Genève. Pour chacune des 2 questions ci-dessous justifiez votre réponse par un test d'hypothèse. Veuillez préciser quelle est l'hypothèse nulle, quelle statistique de test vous utilisez et quelle est la conclusion. Fixez le seuil α à 5 %.

1. À la question sur l'organisation générale du cours, 12 étudiants se sont dits très satisfaits, 17 satisfaits, 19 mitigés, 11 insatisfaits et 8 très insatisfaits. Peut-on affirmer, au vu de ces résultats, que les réponses sont uniformément distribuées parmi les 5 possibilités de réponse ?
2. Les réponses concernant la relation entre les étudiants et le professeur sont résumées dans le tableau ci-dessous

	Bonnes	Moyennes	Mauvaises
Hommes	12	13	8
Femmes	17	10	7

   Peut-on affirmer, au vu de ces résultats, que l'appréciation du professeur par les hommes est différente (indépendante) de celle des femmes ?

8.35

Pour comparer 2 bières on fait une expérience avec 100 amateurs de chaque marque. Chaque groupe affirme connaître la différence entre les 2 et préférer nettement la sienne. On demande à chaque sujet d'identifier sa préférence, après avoir goûté les 2. Voici les résultats.

		habituellement A	boivent B	total
ont préféré	A	65	45	110
	B	35	55	90
total		100	100	200

Est-ce que l'habitude et la préférence sont des caractères indépendants ?

8.36

Un pronostiqueur a observé les résultats de 144 courses de chevaux. Le tableau suivant donne le nombre de vainqueurs pour les 8 positions de départ numérotées de l'intérieur vers l'extérieur.

position de départ	1	2	3	4	5	6	7	8
nombre de vainqueurs	29	19	18	25	17	10	15	11

En supposant que tous les chevaux ont les mêmes capacités, on aimerait savoir si chaque concurrent a autant de chance de gagner la course (indépendamment de la position de départ).

1. Définir l'hypothèse H_0 à tester.
2. Choisir le bon test.
3. Calculer la p-valeur et conclure.

8.37

Le tableau suivant (extrait de *Marketing Research*, de A. Parasuraman, Addison-Wesley, 1986) reprend des données concernant le niveau des dépenses mensuelles pour l'achat de produits cosmétiques observé sur un échantillon aléatoire simple de 500 femmes adultes différenciées selon leur statut professionnel.

	plein temps	temps partiel	sans profession
Moins de 10 $	30	20	60
De 10 $ à 25 $	55	60	65
Plus de 25 $	55	80	75

Sur la base de ce tableau, peut-on dire qu'il y a indépendance entre le niveau de dépenses et le statut professionnel (utilisez un risque de 1re espèce $\alpha = 5\,\%$)?

Quelle recommandation peut-on faire sur cette base aux responsables du marketing de produits cosmétiques?

8.38

Mille personnes ont été classifiées selon leur sexe et selon le fait d'être daltonien ou pas. Les résultats sont les suivants.

	homme	femme	total
normal	442	514	956
daltonien	38	6	44
total	480	520	1 000

1. Peut-on affirmer que le fait d'être daltonien est indépendant du sexe d'une personne?
2. Selon un modèle génétique les chiffres dans le tableau précédent devraient avoir des fréquences relatives données par le tableau suivant

$\frac{p}{2}$	$\frac{p^2}{2} + pq$
$\frac{q}{2}$	$\frac{q^2}{2}$

où q est la proportion des daltoniens dans la population et $p + q = 1$. Est-ce que les observations sont cohérentes avec ce modèle?

8.39

Un grand fabricant de cigarettes désire savoir si le fait de fumer est lié à la consommation de chocolat. Mille personnes sont questionnées et les résultats sont les suivants : parmi 600 mangeurs de chocolat, 200 personnes ne fument pas et 100 personnes ne consomment ni chocolat, ni cigarettes.

Quelles conclusions statistiques peut-on tirer de ce sondage ?

8.40

Ayant soupçonné que certains postes sont inaccessibles aux femmes, le collectif des femmes d'une société engage un statisticien pour mener une étude sur la répartition des postes suivant le sexe. Ce dernier recueille un échantillon de 166 personnes présenté dans une table de contingence entre le secteur d'activité des personnes et leur sexe.

	hommes	femmes	total
management	28	20	48
ventes	18	32	50
service	8	12	20
autres	40	8	48
total	94	72	166

Que va conclure le statisticien au vu de ces résultats (au niveau $\alpha = 5\ \%$) ?

8.41

Quatre marques de peinture sont comparées. Les marques A et B sont meilleur marché que les marques C et D. Plusieurs plaquettes sont peintes puis exposées pendant 6 mois aux conditions météorologiques. Chaque plaquette est ensuite jugée selon différents critères et un score lui est attribué.

peinture	scores					
A	84	86	91	93	84	88
B	90	88	92	84	94	
C	86	87	85	91	93	
D	81	83	92	84	87	81

Construire une statistique de test basée sur la variance pour tester l'hypothèse : « les 4 marques de peinture ont le même comportement face aux intempéries. » Quelle est votre conclusion ?

Régression linéaire

8.42

Soit le modèle de régression

$$Y_i = \theta x_i^2 + \epsilon_i,$$

pour $i = 1, \ldots, n$. On fait l'hypothèse que les erreurs ϵ_i sont indépendantes et distribuées selon la loi normale avec espérance 0 et variance σ^2.

1. Donner la distribution de Y_i.
2. Calculer l'estimateur des moindres carrés de θ, son biais et sa variance.
3. Calculer l'information de Fisher $\tilde{J}(\theta)$ par rapport à la distribution conjointe de $(Y_1, \ldots, Y_n)$ et en déduire la borne de Cramér-Rao pour des estimateur sans biais.
4. Donner l'efficacité de l'estimateur des moindres carrés.
5. Donner un intervalle de confiance au degré $(1 - \alpha)$ pour θ basé sur la statistique $(\sum_{i=1}^n x_i^2 Y_i) / \sum_{i=1}^n x_i^4$.

8.43

On considère le modèle de régression simple

$$Y_i = \alpha + \beta x_i + \epsilon_i, \quad i = 1, \ldots, n$$

où $\epsilon_1, \ldots, \epsilon_n$ sont n variables aléatoires telles que $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

1. Construire des intervalles de confiance pour α et β.
   Indication : utiliser l'exercice 7.2.
2. Tester l'hypothèse $H_0 : \beta = 0$.

8.44

Est-ce qu'il y a un lien entre l'intelligence et la taille du cerveau ? Une étude de Willerman *et al.* (1991) a étudié cette question en mesurant le QI (*Performance IQ de Wechsler (1981)*) et la taille du cerveau (TC, mesurée par résonance magnétique) de 40 individus. On suppose le modèle de régression suivant

$$\mathrm{QI}_i = \alpha + \beta \mathrm{TC}_i + \epsilon_i,$$

pour $i = 1, \ldots, 40$, où ϵ_i est le terme d'erreur tel que $E(\epsilon_i) = 0$ et $Var(\epsilon_i) = \sigma^2$. Sur la figure 8.2 sont représentées les données.

En sachant que l'estimation par la méthode des moindres carrés donne $\hat{\alpha}_{MC} = 1{,}74$, $\hat{\beta}_{MC} = 0{,}00012$ et $\hat{\sigma} = 20{,}99$, veuillez répondre aux questions suivantes.

1. Quelle est la valeur estimée du QI pour un individu pour lequel TC = 910 000 ?

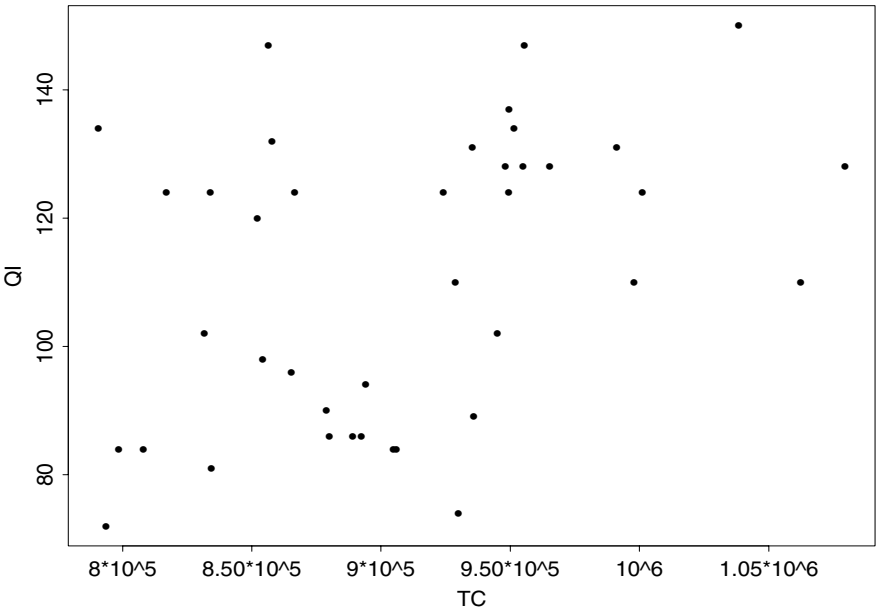


Fig. 8.2 – *Représentation des données de l'exercice 8.44.*

2. Donner un intervalle de confiance pour β. (Utiliser le fait que $\sqrt{S_{xx}} = 457\ 152$, où $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$.)
3. Est-ce que le QI dépend de la taille du cerveau ? Autrement dit, est-ce que β est significativement différent de 0 ?

8.45

Vous étudiez le marché de l'immobilier à travers la relation entre le logarithme du prix de vente d'une maison $Y_i = \log(p_i)$ et l'âge de la maison x_i, $i = 1, \ldots, n$. Sur les données dont vous disposez, vous ajustez un modèle linéaire

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n.$$

Le graphique des résidus se présente selon la figure 8.3.

1. Est-ce qu'il s'agit d'un bon ajustement ? Expliquez.
2. Si le modèle précédent n'est pas satisfaisant, proposez-en un meilleur.
3. Écrivez les équations d'estimation par la méthode des moindres carrés de ce nouveau modèle.

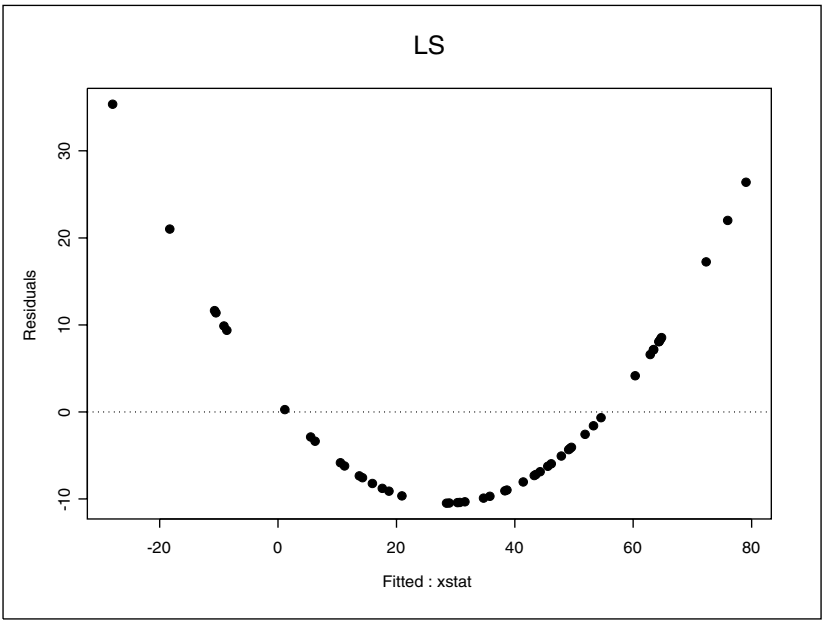


Fig. 8.3 – Représentation des résidus de l'exercice 8.45.

4. Donnez une interprétation de ce modèle dans le cadre du marché de l'immobilier.

Exercices combinés

8.46

Dans le cadre d'une étude de marketing, on analyse la consommation de lessive des ménages. Sur la base d'un échantillon aléatoire simple de 100 ménages, on a observé une consommation moyenne de 14 kilos avec un écart-type estimé s de 2 kilos. D'autre part, 30 % des personnes interrogées accordaient leur préférence à un produit sans phosphates.

1. Construire un intervalle de confiance à 95 % pour l'espérance de la quantité consommée de lessive.
2. Un fabricant affirme que 25 % des ménages préfèrent la lessive sans phosphates alors que le consultant responsable de l'étude estime qu'il y en a plus que 25 %. Réaliser un test pour départager ces deux opinions. (Utiliser un risque de 1re espèce $\alpha = 5$ %.)

8.47

Pour modéliser les valeurs extrêmes d'une action, on peut utiliser la loi de Gumbel. Sa fonction de répartition est

$$F(x) = \exp\left(-\exp\left(-\frac{x-\xi}{\sqrt{\theta}}\right)\right).$$

Soit $X_1, \ldots, X_n$ un échantillon d'observations indépendantes telles que chaque $X_i \sim F$.

1. Calculer l'estimateur des moments pour ξ et θ.
   Indication : $E(X_i) = \xi + \gamma\sqrt{\theta}$ et $V(X_i) = \frac{\pi^2}{6}\theta$, avec $\gamma = 0{,}57722$.
2. Les estimateurs sont-ils convergent ?
3. L'estimateur des moments pour θ est-il biaisé ?

Pour les questions suivantes, on suppose que $\theta = 6$.

4. En utilisant l'estimateur des moments pour ξ, construire un intervalle de confiance au degré de confiance $(1-\alpha)$ pour le paramètre ξ.
5. Exprimer le 0,9-quantile de la distribution de Gumbel en fonction des paramètres ξ et $\theta = 6$, et trouver un estimateur de ce quantile en remplaçant ξ par son estimateur des moments.
6. En utilisant 5., dériver un intervalle de confiance au degré de confiance $(1-\alpha)$ pour le 0,9-quantile de cette distribution.
7. On dispose de 20 observations historiques sur les valeurs extrêmes de l'action entre 1978 et 2000. La moyenne de ces valeurs est 372,05 €. Calculer la valeur estimée du 0,9-quantile.
8. L'année passée, on a observé une valeur maximale de l'action de 374,80 €. Au vu des données disponibles, peut-on considérer le niveau maximal atteint par l'action comme exceptionnel ?

8.48

Soit $X_1, \ldots, X_n$ un échantillon de n variables aléatoires qui proviennent d'une distribution exponentielle avec densité

$$f(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right)$$

pour $x \geq 0$ et avec $\theta > 0$. On considère les 3 estimateurs suivants de θ :
$\hat{\theta}_1 = \sum_{i=1}^{n} X_i/n$, $\hat{\theta}_2 = \sum_{i=1}^{n} X_i/(n+1)$, et $\hat{\theta}_3 = n \cdot \min(X_1, \ldots, X_n)$.

1. Calculer l'espérance et la variance de $\hat{\theta}_1$ et de $\hat{\theta}_2$ ainsi que leur erreur quadratique moyenne.
2. De même pour $\hat{\theta}_3$. Comparer l'efficacité de ces 3 estimateurs du point de vue de l'erreur quadratique moyenne.
   Indication : trouver la distribution de $Y = \min(X_1, \ldots, X_n)$, c'est-à-dire calculer $P(Y < y)$. De quelle loi s'agit-il ?

3. Les estimateurs $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ sont-ils convergents?
4. Définissons l'intervalle de confiance suivant pour $\lambda = 1/\theta$

$$I = [a/\min(X_i)\,, b/\min(X_i)].$$

Calculer le degré de confiance de cet intervalle en fonction de a et de b. *Indication :* montrer que $Z = \lambda \cdot \min(X_i)$ est distribué selon la loi exponentielle de paramètre n et utiliser ce résultat.

8.49

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes et identiquement distribuées selon une loi uniforme $U(0\,, \theta)$. On définit les statistiques suivantes

$$\hat{\theta}_1 = 2\bar{X} \quad \text{et} \quad \hat{\theta}_2 = \max(X_1, \ldots, X_n)$$

1. Montrer que $\hat{\theta}_1$ est un estimateur non biaisé de θ.
2. Calculer l'espérance de $\hat{\theta}_2$, puis en déduire un estimateur non biaisé de θ que l'on notera $\hat{\theta}_3$.
3. Comparer l'efficacité des ces 3 estimateurs du point de vue de l'erreur quadratique moyenne.
4. Les estimateurs $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ sont-ils consistants?
5. On considère 2 candidats pour construire un intervalle de confiance pour θ

$$I_1 = [a\,\hat{\theta}_2\ ,\ b\,\hat{\theta}_2] \quad \text{et} \quad I_2 = [\hat{\theta}_2 + c\ ,\ \hat{\theta}_2 + d],$$

avec $1 \leq a < b$ et $0 \leq c < d \leq \theta$.

   (a) Quel est le taux de couverture de I_1, c'est-à-dire la probabilité que $\theta \in I_1$ (en fonction de a et b)?
   (b) De même, quel est le taux de couverture de I_2 (en fonction de c et d)? Qu'observez-vous par rapport à I_1?
   (c) Si on prend $b = 2a$ dans I_1, comment faut-il choisir a pour que le taux de couverture soit égal à 95 %?

8.50

On dispose des résultats d'un test auquel on a soumis n personnes pendant les dernières années. Chaque personne pouvait se présenter au test plusieurs fois jusqu'au succès.

Soit θ, $0 < \theta < 1$, la probabilité qu'une personne de cette population réussisse le test lors d'un essai et X la variable aléatoire qui représente le nombre d'essais indépendants nécessaires jusqu'au succès.

1. Interpréter le modèle de probabilité

$$P(X = k) = \theta \cdot (1 - \theta)^{k-1} \quad k = 1, 2, 3, \ldots$$

2. On dispose des données $k_1, \ldots, k_n$ concernant le nombre d'essais effectués par chaque personne jusqu'au succès. Calculer l'estimateur du maximum de vraisemblance $\hat{\theta}_{MV}$ pour θ.
3. Donner une approximation de la variance de $\hat{\theta}_{MV}$.
4. Calculer l'estimateur des moments pour θ sur la base de $k_1, \ldots, k_n$.
5. En utilisant le modèle 1., exprimer la probabilité que plus de 2 essais soient nécessaires pour réussir le test.
6. Sur la base des données disponibles la moyenne observée du nombre d'essais est $\overline{k} = 3{,}3$. Utiliser les résultats obtenus aux points 2. et 5. pour estimer la probabilité que dans le futur plus de 2 essais soient nécessaires à une personne pour réussir le test.
7. Sur la base des données $k_1, \ldots, k_n$ on souhaite tester l'hypothèse nulle $H_0 : \theta = \frac{1}{2}$ contre l'alternative $H_A : \theta = \frac{2}{3}$. Un analyste propose d'utiliser $\hat{\theta}_{MV}$ comme statistique pour effectuer le test. S'agit-il d'un choix optimal? Si cela n'est pas le cas, proposer une meilleure statistique.
8. Proposer une façon de calculer la valeur critique approximée du test optimal obtenu au point 7.

8.51

Un syndicat de boulangers fait une enquête sur la consommation mensuelle de pain chez les ménages de plus de 2 personnes. Cette consommation, exprimée en kilos, est désignée par X. Cent ménages qui ont bien voulu se prêter à l'enquête sont suivis pendant 1 mois, et on note pour chacun d'entre eux la quantité totale de pain consommée. On obtient ainsi un échantillon aléatoire relatif à la variable X.

1. Pour cet échantillon, on trouve une moyenne estimée $\bar{x} = 21{,}5$ kilos et un écart-type estimé $s = 1{,}8$ kilo. Donner un intervalle de confiance au niveau 95 % pour $\mu = E(X)$.
2. Ces données permettent-elles de rejeter l'hypothèse nulle $H_0 : \mu = 21$ contre $H_1 : \mu \neq 21$ au seuil $\alpha = 5$ %?
3. Sur les 100 ménages, 23 ont consommé moins de 20 kilos de pain dans le mois. Donner un intervalle de confiance à 95 % pour la proportion de ménages qui consomment mensuellement moins de 20 kilos de pain.
4. Un boulanger du quartier fait une enquête analogue, mais il n'a pu s'assurer que de la coopération de 11 ménages. Il voudrait obtenir un intervalle de confiance pour $\mu = E(X)$ au niveau de 95 %. On suppose que sur son échantillon, il a obtenu $\bar{x} = 21{,}5$ kilos et $s = 1{,}8$ kilo. Construire un intervalle de confiance au niveau de 95 % pour μ. Comparer ce dernier avec l'intervalle trouvé en 1. et discuter.

8.52

Dans le cadre de la modélisation de distributions de revenu, on utilise souvent la distribution de Dagum qui est définie par la fonction de répartition suivante

$$F_\beta(x) = P(X < x) = \left(1 + \frac{1}{x^2}\right)^{-\beta},$$

où $x > 0$ et $\beta > 0$.

1. Sur la base d'un échantillon $X_1, \ldots, X_n$ de revenus, calculer l'estimateur du maximum de vraisemblance $\hat{\beta}$ pour le paramètre β.
2. Calculer l'information de Fisher de ce modèle et donner la distribution de $\sqrt{n}(\hat{\beta} - \beta)$ lorsque $n \to \infty$.
3. On génère à l'ordinateur 100 000 échantillons de 16 observations suivant la loi de Dagum avec paramètre $\beta = 1$. Pour chaque échantillon, on calcule la valeur de l'estimateur $\hat{\beta}$. Donner le nombre approximé d'échantillons où la valeur de l'estimateur se trouve dans l'intervalle [0,5, 1,5].
4. Construire un intervalle de confiance approximé au degré 95 % pour β. *Indication :* utiliser le résultat du point 2. et la distribution de $\sqrt{n}(\frac{\hat{\beta}}{\beta} - 1)$ lorsque $n \to \infty$.

8.53

Soient $X_1, \ldots, X_n$ un échantillon de n variables aléatoires indépendantes qui proviennent d'une distribution dont la densité est

$$f_\alpha(x) = \begin{cases} \alpha x^{\alpha-1} & \text{si } 0 < x < 1 \\ 0 & \text{sinon,} \end{cases}$$

où $\alpha > 0$.

1. Calculer l'estimateur des moments $\hat{\alpha}_M$ de α et l'écrire comme une fonction $\hat{\alpha}_M = g(\overline{X})$.
2. En utilisant le théorème central limite, donner la distribution approximée de $\sqrt{n}(\overline{X} - \mu)$, où $\mu = E(X)$.
3. Donner une approximation de la variance de $\hat{\alpha}_M$ en utilisant le résultat

$$\sqrt{n}(g(\overline{X}) - g(\mu)) \overset{n\to\infty}{\sim} \mathcal{N}\left(0\ ,\ \left(\frac{\partial g(\mu)}{\partial \mu}\right)^2 \sigma^2\right),$$

   où σ^2 est la variance de $\sqrt{n}(\overline{X} - \mu)$ calculée au point 2.
4. L'estimateur $\hat{\alpha}_M$ est-il convergent ? Justifier votre réponse.
5. L'estimateur $\hat{\alpha}_M$ est-il efficace ? Justifier votre réponse.
6. À l'aide du résultat du point 3., construire un intervalle de confiance approximé au degré 95 % pour le paramètre α.

8.54

Soient $X_1, \ldots, X_n$ des variables aléatoires indépendantes telle que $X_i \sim \mathcal{N}(\mu, \sigma^2)$ avec σ^2 connu. On veut tester l'hypothèse nulle $H_0 : \mu = \mu_0$ contre l'alternative $H_A : \mu > \mu_0$ à l'aide de la statistique $\bar{X}$. La p-valeur observée est définie par $pv = P_{H_0}(\bar{X} > \bar{x})$, où $\bar{x}$ est la valeur observée de la statistique.

1. Expliquer pourquoi la statistique $\bar{X}$ est une bonne statistique pour tester H_0.
2. Écrivez pv comme une fonction $g(\bar{x}, \mu_0, \sigma, n)$ à l'aide de $\Phi(.)$, la fonction de répartition d'une distribution $\mathcal{N}(0, 1)$.
3. On considère maintenant la variable aléatoire $PV = g(\bar{X}, \mu_0, \sigma, n)$. Calculer sa fonction de répartition et sa densité sous une alternative $\mu = \mu_1 (> \mu_0)$.
   Indication : développer $P_{\mu_1}(PV > a) = P_{\mu_1}(g(\bar{X}, \mu_0, \sigma, n) > a)$ pour une valeur de a donnée.
4. Quelle est la distribution de PV quand $\mu_1 = \mu_0$?

Corrigés

8.1

On a un échantillon de taille $n = 13$ avec $\overline{x} \simeq 14{,}3$ et on sait que $\sigma = 4$.

1. Par le théorème central limite, on fait l'hypothèse que la variable aléatoire $\overline{X}$ suit approximativement une loi normale de paramètres $(0, 1)$. Par conséquent, on trouve l'intervalle de confiance à 95 % suivant

$$\text{IC} = \left[\overline{x} \pm z_{0{,}975}\frac{\sigma}{\sqrt{n}}\right] \simeq [12{,}1\ ;\ 16{,}5].$$

2. Si l'intervalle est de longueur 3, alors

$$z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 1{,}5$$

et

$$z_{1-\alpha/2} \simeq 1{,}35 \quad \Leftrightarrow \quad \alpha \simeq 0{,}18.$$

On obtient un niveau de confiance de 82 %.

3. On aimerait que

$$z_{0{,}995}\frac{\sigma}{\sqrt{n}} = 1.$$

Cela implique que

$$n = (z_{0{,}995}\sigma)^2 \simeq 106.$$

Il faut par conséquent 93 observations additionnelles.

8.2

1. La variable aléatoire R_p suit une loi normale car elle est une somme de variables aléatoires normales. Son espérance est

$$E(R_p) = E\left(\sum_{i=1}^{n} a_i R_i\right) = \sum_{i=1}^{n} a_i E(R_i) = \sum_{i=1}^{n} a_i \mu_i = \mu_p,$$

et sa variance

$$\text{var}(R_p) = \text{var}\left(\sum_{i=1}^{n} a_i R_i\right) = \sum_{i=1}^{n} a_i^2 \text{var}(R_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2 = \sigma_p^2.$$

La variable aléatoire R_p est distribuée selon une loi normale de paramètres (μ_p, σ_p^2).

2. On cherche un intervalle de confiance à 95 % pour μ_p. Comme R_p suit une loi normale
$$-z_{1-\alpha/2} \leq \frac{R_p - \mu_p}{\sigma_p} \leq z_{1-\alpha/2},$$
alors
$$\text{IC} = \left[\sum_{i=1}^{n} a_i R_i \pm 1{,}96 \sqrt{\sum_{i=1}^{n} a_i^2 \sigma_i^2} \right].$$
3. L'application numérique donne l'intervalle de confiance [0,076 ; 0,092].

8.3

1. On cherche un intervalle de confiance pour la moyenne de la baisse de puissance de la fabrication. Par le théorème central limite, on sait qu'approximativement
$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1),$$
où μ et σ^2 sont l'espérance et la variance de chacun des X_i. Pour un intervalle de confiance à 95 %, on va donc chercher un nombre a tel que
$$P(-a < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < a) = 0{,}95.$$
Par le théorème central limite, on trouve
$$P(-a < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < a) \simeq 2\Phi(a) - 1 = 0{,}95 \quad \Leftrightarrow \quad a = 1{,}96.$$
De ce résultat, on déduit l'intervalle de confiance suivant pour μ (avec les valeurs des paramètres de l'énoncé)
$$\text{IC} = \left[\overline{x} - a\frac{\sigma}{\sqrt{n}} \ ; \ \overline{x} + a\frac{\sigma}{\sqrt{n}} \right] = [11{,}56 \ ; \ 12{,}44].$$
L'intervalle [11,56 ; 12,44] contient μ avec une probabilité de 95 %.
2. On procède de la même manière qu'au point 1., mais avec un intervalle de confiance à 99 %. Dans ce cas, $a = 2{,}58$ et l'intervalle qui contient μ avec une probabilité de 99 % est [11,42 ; 12,58].
3. Le 2e intervalle, qui mesure 1,16, est plus long que le premier (0,88). Cela s'explique par le fait que la précision diminue lorsque la certitude augmente.

8.4

Soit $\hat{\mu}$ l'estimateur de μ tel que

$$\hat{\mu} = \frac{1}{2}(\overline{X} + \overline{Y}).$$

Les variables aléatoires X_i et Y_i suivent des lois normales de variances connues. Pour calculer l'intervalle de confiance basé sur $\hat{\mu}$, il faut calculer sa variance

$$\begin{aligned} \mathrm{var}(\hat{\mu}) = \mathrm{var}\Big(\frac{1}{2}(\overline{X} + \overline{Y})\Big) &= \frac{1}{4}\Big(\mathrm{var}(\overline{X}) + \mathrm{var}(\overline{Y})\Big) \\ = \quad \frac{1}{4}\left(\frac{\sigma_H^2}{n} + \frac{\sigma_F^2}{n}\right) &= \frac{1}{4n}(\sigma_H^2 + \sigma_F^2), \end{aligned}$$

où l'on a utilisé l'indépendance des X_i avec les Y_i. L'intervalle de confiance est par conséquent

$$\mathrm{IC} = \left[\hat{\mu} \pm z_{1-\alpha/2}\sqrt{\mathrm{var}(\hat{\mu})}\right] = \left[\frac{1}{2}(\overline{X} + \overline{Y}) \pm \frac{1}{2} z_{1-\alpha/2}\sqrt{\frac{\sigma_H^2 + \sigma_F^2}{n}}\right].$$

8.5

1. On sait que $E(X_i) = 2/\lambda$. Pour calculer la variance de X_i, il faut connaître $E(X_i^2)$

$$E(X_i^2) \quad = \quad \int_0^\infty x^2 \lambda^2 x \exp(-\lambda x) dx = \lambda^2 \int_0^\infty x^3 \exp(-\lambda x) dx.$$

L'intégrale se fait par parties

$$\lambda^2 \int_0^\infty x^3 \exp(-\lambda x) dx = 3\lambda^2 \int_0^\infty x^2 \left(\frac{1}{\lambda}\exp(-\lambda x)\right) dx.$$

Soit Z une variable aléatoire suivant une distribution exponentielle de paramètre λ. L'intégrale précédente est alors égale à l'espérance de Z^2. Comme $\mathrm{var}(Z) = 1/\lambda^2$ et $E(Z) = 1/\lambda$, on en déduit que $E(Z^2) = \mathrm{var}(Z) + E^2(Z) = 2/\lambda^2$. Finalement

$$E(X_i^2) = 3\lambda^2 \frac{2}{\lambda^2} = 6.$$

Ce résultat permet d'obtenir $\mathrm{var}(\sum_{i=1}^n X_i)$ en fonction de μ

$$\mathrm{var}(\sum_{i=1}^n X_i) = n\mathrm{var}(X_i) = n(E(X_i^2) - E^2(X_i)) = n(6 - \mu^2).$$

2. Par le théorème central limite

$$-z_{1-\alpha/2} < \frac{\sum_{i=1}^n X_i - E(\sum_{i=1}^n X_i)}{\sqrt{\mathrm{var}(\sum_{i=1}^n X_i)}} < z_{1-\alpha/2},$$

avec $\alpha = 5$ %. D'après les résultats du point 1.,

$$-z_{1-\alpha/2} < \frac{\sum_{i=1}^n X_i - n\mu}{\sqrt{n/2}\,\mu} < z_{1-\alpha/2},$$

donc

$$\frac{\sum_{i=1}^n X_i}{n + \sqrt{n/2}z_{1-\alpha/2}} < \mu < \frac{\sum_{i=1}^n X_i}{n - \sqrt{n/2}z_{1-\alpha/2}}.$$

8.6

La demi-longueur a de l'intervalle de confiance au degré $(1-\alpha)$ est

$$a = z_{1-\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}},$$

où p est la vraie proportion de personnes favorables à la proposition, n la taille de l'échantillon et $z_{1-\frac{\alpha}{2}}$ le $(1-\frac{\alpha}{2})$ quantile de la distribution $\mathcal{N}(0,1)$. La demi-longueur maximale est obtenue pour $p = \frac{1}{2}$,

$$a = z_{1-\frac{\alpha}{2}}\frac{1}{2\sqrt{n}}.$$

1. Dans ce cas $(1-\alpha) = 95$ %, $z_{1-\frac{\alpha}{2}} = z_{0,975} = 1{,}96$, $a = 0{,}10/2$. On obtient

$$n = \frac{z_{1-\frac{\alpha}{2}}^2}{4a^2} = 384.$$

L'étude n'est donc pas faisable à ces conditions.

2. Ici $z_{1-\frac{\alpha}{2}} = 2a\sqrt{n} = 2 \cdot 0{,}05 \cdot \sqrt{200} = \sqrt{2} \simeq 1{,}412$.
Donc $1 - \frac{\alpha}{2} = 0{,}92$, $\alpha = 0{,}16$ et $1 - \alpha = 0{,}84$.

8.7

Par le théorème central limite, on a

$$\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

lorsque $n \to \infty$. Avec $\mu = 5$, $\sigma = 1$ et $n = 100$ on obtient donc

$$\begin{aligned} P(\bar{Y} - 0{,}1 < 5 < \bar{Y} + 0{,}1) &= P(-0{,}1 < \bar{Y} - 5 < 0{,}1) \\ &= P\left(\frac{-0{,}1}{1/\sqrt{100}} < Z < \frac{0{,}1}{1/\sqrt{100}}\right) \\ &= P(-1 < Z < 1) \\ &= 2 \cdot P(Z < 1) \\ &= 0{,}682. \end{aligned}$$

Sur les 1 000 échantillons simulés, 682 intervalles de la forme $[\bar{Y} - 0{,}1\,;\, \bar{Y} + 0{,}1]$ contiendront la valeur 5.

8.8

1. Le degré de confiance est la probabilité de couverture de l'intervalle de confiance, c'est-à-dire la probabilité que μ appartienne à l'intervalle

$$\begin{aligned} P\left(\overline{y} - \frac{b_1}{\sqrt{n}} < \mu < \overline{y} + \frac{b_2}{\sqrt{n}}\right) &= P\left(-b_1 < \frac{\mu - \overline{y}}{1/\sqrt{n}} < b_2\right) \\ &= P\left(-b_2 < \frac{\overline{y} - \mu}{1/\sqrt{n}} < b_1\right) \\ &= \Phi(b_1) - \Phi(-b_2), \end{aligned}$$

   où $\Phi(\cdot)$ est la fonction de répartition de la loi normale centrée et réduite.
2. La longueur de l'intervalle de confiance est

$$\overline{y} + \frac{b_2}{\sqrt{n}} - \left(\overline{y} - \frac{b_1}{\sqrt{n}}\right) = \frac{1}{\sqrt{n}}(b_2 + b_1).$$

   On veut donc minimiser $(b_1 + b_2)$ sous la contrainte $\Phi(b_1) - \Phi(-b_2) = 0{,}9$. Construisons le lagrangien $\mathcal{L}$

$$\mathcal{L} = b_1 + b_2 + \lambda\left[\Phi(b_1) - \Phi(-b_2) - 0{,}9\right],$$

   où λ est un multiplicateur de Lagrange. La minimisation de $\mathcal{L}$ donne le système suivant

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial b_1} &= 1 + \lambda\Phi'(b_1) = 0 \\ \frac{\partial \mathcal{L}}{\partial b_2} &= 1 + \lambda\Phi'(-b_2) = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= \Phi(b_1) - \Phi(b_2) - 0{,}9 = 0, \end{aligned}$$

   qui, par différence de ces 2 premières équations, conduit à

$$\Phi'(b_1) = -\Phi'(-b_2).$$

On dérive les 2 membres de cette égalité et on obtient

$$\varphi(b_1) = \varphi(-b_2),$$

où $\varphi(\cdot)$ est la densité de la loi normale centrée et réduite. Par symétrie de la fonction $\varphi(\cdot)$, on obtient

$$b_1 = b_2 = b.$$

La contrainte implique alors

$$\Phi(b) - \Phi(-b) = \Phi(b) - (1 - \Phi(b)) = 2\Phi(b) - 1 = 0{,}9$$

$$\Leftrightarrow \quad \Phi(b) = 0{,}95 \quad \Leftrightarrow b \simeq 1{,}64.$$

Les valeurs qui minimisent la longueur de l'intervalle sont $b_1 = b_2 \simeq 1{,}64$.

8.9

1. Pour déterminer l'estimateur des moments, il faut d'abord calculer son espérance

$$E(X) = \int_{\theta k}^{\theta(k+1)} x f(x) dx = \int_{\theta k}^{\theta(k+1)} \frac{x}{\theta} dx = \frac{x^2}{2\theta}\Big|_{x=\theta k}^{x=\theta(k+1)} = \frac{2k+1}{2}\theta.$$

L'estimateur doit satisfaire

$$\frac{2k+1}{2}\,\hat{\theta}_M = \bar{X},$$

donc $\hat{\theta}_M = \frac{2\bar{X}}{2k+1}$.
Puisque

$$\begin{aligned} \text{plim}_{n\to\infty}\hat{\theta}_M &= \frac{2}{2k+1}\text{plim}_{n\to\infty}\bar{X} = \frac{2}{2k+1}E(X_i) = \\ &= \frac{2}{2k+1}\cdot\frac{2k+1}{2}\theta = \theta \end{aligned}$$

l'estimateur $\hat{\theta}_M$ est convergent.

2. L'estimateur $\hat{\theta}_M$ est sans biais et sa variance vaut

$$\begin{aligned} \text{var}(\hat{\theta}_M) &= \frac{4}{(2k+1)^2}\text{var}(\bar{X}) = \frac{4}{(2k+1)^2}\frac{\text{var}(X_i)}{n} \\ &= \frac{4}{(2k+1)^2}\cdot\frac{1}{n}\cdot\frac{\theta^2}{12} \\ &= \frac{\theta^2}{3n(2k+1)^2}. \end{aligned}$$

On a

$$\mathrm{ECM}(\hat{\theta}_M, \theta) = \mathrm{var}(\hat{\theta}_M) + \mathrm{biais}^2(\hat{\theta}_M, \theta) = \frac{\theta^2}{3n(2k+1)^2}.$$

3. Puisque $\hat{\theta}_M$ est fonction de $\bar{X}$, la distribution asymptotique de $\hat{\theta}_M$ s'obtient à l'aide du théorème central limite

$$\frac{\hat{\theta}_M - E(\hat{\theta}_M)}{\sqrt{\mathrm{var}(\hat{\theta}_M)}} \sim \mathcal{N}(0,1)$$

$$\frac{\hat{\theta}_M - \theta}{\frac{\theta}{(2k+1)\sqrt{3n}}} \sim \mathcal{N}(0,1)$$

4. Pour construire l'intervalle de confiance, on utilise le résultat du point 3., d'où l'on tire que

$$-z_{0,975} < \frac{\hat{\theta}_M - \theta}{\frac{\theta}{(2k+1)\sqrt{3n}}} < z_{0,975}$$

et ensuite

$$1 - \frac{1{,}96}{(2k+1)\sqrt{3n}} < \frac{\hat{\theta}_M}{\theta} < 1 + \frac{1{,}96}{(2k+1)\sqrt{3n}}$$

$$\frac{(2k+1)\sqrt{3n} - 1{,}96}{(2k+1)\sqrt{3n}} < \frac{\hat{\theta}_M}{\theta} < \frac{(2k+1)\sqrt{3n} + 1{,}96}{(2k+1)\sqrt{3n}}$$

$$\frac{(2k+1)\sqrt{3n}\ \hat{\theta}_M}{(2k+1)\sqrt{3n} + 1{,}96} < \theta < \frac{(2k+1)\sqrt{3n}\ \hat{\theta}_M}{(2k+1)\sqrt{3n} - 1{,}96}$$

$$\frac{2\sqrt{3n}\ \bar{X}}{(2k+1)\sqrt{3n} + 1{,}96} < \theta < \frac{2\sqrt{3n}\ \bar{X}}{(2k+1)\sqrt{3n} - 1{,}96}$$

5. La longueur de l'intervalle calculé au point 4. est

$$\begin{aligned} l(k) &= \frac{2\sqrt{3n}\ \bar{X}}{(2k+1)^2\sqrt{3n} - 1{,}96} - \frac{2\sqrt{3n}\ \bar{X}}{(2k+1)\sqrt{3n} + 1{,}96} \\ &= \frac{4 \cdot 1{,}96\sqrt{3n}\bar{X}}{3n(2k+1)^2 - (1{,}96)^2} \end{aligned}$$

Il s'agit d'une fonction décroissante en k : si k augmente, $l(k)$ diminue.

8.10

1. Puisque X_i suit une loi Gamma de paramètres $(3, \theta)$, son espérance est

$$E(X_i) = \frac{3}{\theta}.$$

On trouve alors l'estimateur des moments $\hat{\theta}_M$ de θ :

$$\hat{\theta}_M = \frac{3}{\overline{X}}.$$

2. Pour construire un intervalle de confiance pour μ, on utilise le théorème central limite

$$\frac{\overline{X} - E(\overline{X})}{\sqrt{\text{var}(\overline{X})}} \sim N(0,1),$$

avec $E(\overline{X}) = 3/\theta$ et $\text{var}(\overline{X}) = \frac{3}{n\theta^2}$. Ainsi, pour un intervalle de confiance à 95 %, on obtient

$$\begin{aligned} & -1{,}96 < \frac{\overline{X} - \frac{3}{\theta}}{\sqrt{\frac{3}{n\theta^2}}} < 1{,}96 \\ \Leftrightarrow \quad & -1{,}96 < \sqrt{\frac{n}{3}}\overline{X}\theta - \sqrt{3n} < 1{,}96 \\ \Leftrightarrow \quad & \text{IC} = \left[\frac{3}{\overline{X}} \pm 1{,}96\sqrt{\frac{3}{n}}\frac{1}{\overline{X}}\right]. \end{aligned}$$

Avec $n = 100$ et $\overline{x} = 6$, l'intervalle devient [0,444 ; 0,556].

8.11

Soit x_i la i^e mesure avec $i = 1, \ldots, 9$.

1. La moyenne de l'échantillon est $\overline{x} \simeq 1{,}02$.
2. Pour déterminer un intervalle de confiance pour $\mu = E(X)$, on fait l'hypothèse que $\overline{X}$ suit approximativement une loi normale de paramètres $(1{,}02, 0{,}09)$. Ainsi, l'intervalle de confiance à 95 % pour μ est

$$\text{IC} = \left[\overline{x} \pm z_{0{,}975}\frac{\sigma}{\sqrt{n}}\right] \simeq [0{,}82\,;\, 1{,}22],$$

où $z_{0{,}975} \simeq 1{,}96$ est le 0,975-quantile de la loi normale centrée et réduite.

3. Pour réduire la taille de l'intervalle de confiance, le biochimiste doit augmenter la taille de l'échantillon ou réduire le niveau de confiance.

8.12

Premier cas.

1. Les hypothèses sont
$$\begin{array}{rcl} H_0 & : & \mu = 18 \\ H_A & : & \mu < 18. \end{array}$$

2. Soit la statistique de test $T = \overline{X}$, avec $X_1, \ldots, X_{10}$ les temps mesurés. Calculons la p-valeur
$$\begin{aligned} p\text{-valeur} &= P_{H_0}(T < 17{,}13) = \\ &= P\left(Z < \frac{17{,}13 - 18}{1{,}5/\sqrt{10}}\right) \simeq P(Z < -1{,}83) \simeq 0{,}03, \end{aligned}$$
où $Z = \frac{T-\mu}{\sigma/\sqrt{n}}$ suit une loi normale de paramètres $(0, 1)$ sous l'hypothèse de normalité des X_i. Ce résultat implique que si le seuil a été fixé à 5 %, on rejette l'hypothèse nulle, et que s'il a été fixé à 1 %, on ne la rejette pas.

Deuxième cas.

1. Les hypothèses sont cette fois
$$\begin{array}{rcl} H_0 & : & \mu = 100 \\ H_A & : & \mu < 100. \end{array}$$

2. Soit la statistique de test $T = \overline{X}$, où $X_1, \ldots, X_{14}$ sont les surfaces mesurées. La p-valeur est
$$\begin{aligned} p\text{-valeur} &= P_{H_0}(T < 99{,}9) = \\ &= P\left(Z < \frac{99{,}9 - 100}{5/\sqrt{14}}\right) \simeq P(Z < -0{,}07) \simeq 0{,}47, \end{aligned}$$
avec $Z = \frac{T-\mu}{\sigma/\sqrt{n}}$ qui suit une loi normale de paramètres $(0, 1)$ sous l'hypothèse de normalité des X_i. Pour un seuil $\alpha = 5$ %, on ne rejette pas l'hypothèse nulle.

8.13

1. Les hypothèses formulées sont
$$\begin{array}{rcl} H_0 & : & \mu = 1{,}4 \\ H_A & : & \mu > 1{,}4. \end{array}$$

2. On définit la statistique de test
$$T = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

La statistique T suit une loi normale centrée et réduite. La valeur observée t_{obs} de la statistique sur l'échantillon est

$$t_{obs} = \frac{1{,}79 - 1{,}4}{1{,}02/\sqrt{25}} \simeq 1{,}91.$$

Le quantile à 95 % de la loi normale vaut environ 1,96 et est supérieur à la valeur observée. On ne rejette donc pas l'hypothèse nulle.

3. Le quantile à 99 % vaut environ 2,49. Encore une fois, l'hypothèse nulle n'est pas rejetée.

8.14

Soit la variable aléatoire X_i telle que

$$X_i = \begin{cases} 0 & \text{si l'enfant } i \text{ est une fille} \\ 1 & \text{si l'enfant } i \text{ est un garçon,} \end{cases}$$

et définissons $Y = \sum_{i=1}^{n} X_i$. La variable aléatoire Y suit une loi binomiale de paramètres $(512, p)$, où p est la probabilité qu'un nouveau-né soit un garçon. On désire tester l'hypothèse H_0 contre H_A telles que

$$\begin{array}{rcl} H_0 & : & p = 1/2 \\ H_A & : & p > 1/2. \end{array}$$

On va calculer la p-valeur. La distribution exacte est connue, mais nous allons voir que le calcul de la probabilité serait laborieux si on ne disposait pas d'un ordinateur. Par conséquent, la distribution de Y est approximée par une loi normale centrée et réduite en appliquant le théorème central limite.

$$\begin{array}{rcl} p\text{-valeur} & = & P_{H_0}\left(\displaystyle\sum_{i=1}^{512} X_i > 284\right) \simeq P_{H_0}\left(\frac{\sum_{i=1}^{512} X_i - np}{\sqrt{np(1-p)}} > \frac{284 - np}{\sqrt{np(1-p)}}\right) \\ & \simeq & P(Z > 2{,}48) \simeq 0{,}007, \end{array}$$

où Z est distribuée selon une loi normale de paramètres (0, 1). On rejette donc au seuil de 5 % l'hypothèse affirmant que les filles ont autant de chance que les garçons d'être prématurées.

8.15

Soit X le nombre d'accidents par année, $X \sim \mathcal{P}(\lambda)$. L'hypothèse nulle est $H_0 : \lambda = 8$ et on observe $X = 5$. La p-valeur est

$$\begin{array}{rcl} p - \text{valeur} & = & P_{H_0}(X < 5) \\ & = & e^{-8}\left(1 + 8 + \frac{8^2}{2!} + \frac{8^3}{3!} + \frac{8^4}{4!}\right) \simeq 0{,}10. \end{array}$$

Par l'approximation normale

$$p - \text{valeur } = P_{H_0}\left(\frac{X-8}{\sqrt{8}} < \frac{5-8}{\sqrt{8}}\right) \simeq \Phi(-1{,}0607) \approx 0{,}14.$$

Le test n'est pas significatif au seuil de 5 % et l'hypothèse nulle n'est pas rejetée.

8.16

Les hypothèses sont

$$\begin{array}{lcl} H_0 & : & \mu = 3{,}1 \\ H_A & : & \mu > 3{,}1. \end{array}$$

La moyenne et l'écart-type estimé de l'échantillon sont $\overline{X} = 3{,}26$ et $\hat{\sigma} = 0{,}24$. Soit la statistique de test

$$T = \frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}}.$$

La statistique T suit une loi de Student à $n-1$ degrés de liberté. La valeur observée t_{obs} sur l'échantillon est

$$t_{obs} \simeq \frac{3{,}26 - 3{,}1}{0{,}24/\sqrt{9}} = 2.$$

Les quantiles à 95 % et 99 % d'une loi de Student à 8 degrés de liberté valent respectivement 1,86 et 2,9. Par conséquent, le docteur Fischbach rejette l'hypothèse nulle avec un seuil à 5 % et ne la rejette pas s'il choisit un seuil à 1 %.

8.17

1. Sous H_0, $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2$ suit une loi χ^2 à $(n-1)$ degrés de liberté.
2. On cherche k_α tel que

$$P_{H_0}(\hat{\sigma}^2 < k_\alpha) = 0{,}95$$

et, en utilisant le point 1., on trouve

$$\begin{aligned} P_{H_0}(\hat{\sigma}^2 < k_\alpha) &= P_{H_0}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 < k_\alpha\right) \\ &= P_{H_0}\left(\frac{1}{\sigma_0^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 < \frac{n-1}{\sigma_0^2}k_\alpha\right) \\ &= P_{H_0}\left(U < \frac{n-1}{\sigma_0^2}k_\alpha\right), \end{aligned}$$

avec $U \sim \chi^2_{n-1}$. À l'aide du quantile à 95 % $\chi^2_{n-1;\ 0,95}$, on détermine k_α :

$$k_\alpha = \frac{\chi^2_{n-1;\ 0,95}}{n-1}\sigma_0^2 \simeq 1{,}59\sigma_0^2.$$

8.18

Soit D_i la différence de poids entre 2 agneaux de la paire i ; on fait l'hypothèse que D_i suit une loi normale d'espérance μ et de variance σ^2. Soient les hypothèses suivantes

$$\begin{array}{lcl} H_0 & : & \mu = 0 \\ H_A & : & \mu \neq 0 \end{array}$$

et T la statistique de test définie par

$$T = \frac{\sqrt{n}(\overline{D} - \mu)}{s}$$

avec $s = \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \overline{D})^2$ et n la taille de l'échantillon. La statistique T suit une loi de Student à 11 degrés de liberté. On observe sur l'échantillon

$$t_{obs} \simeq \frac{\sqrt{12}(-10{,}58)}{12{,}24} \simeq -3$$

et la p-valeur correspondante est

$$p\text{-valeur} \simeq P(\mid T \mid > 3) \simeq 0{,}006.$$

On rejette l'hypothèse nulle au seuil de 5 % ce qui signifie que les régimes alimentaires sont différents.

8.19

On considère les variances comme étant égales. Soient les hypothèses

$$\begin{array}{lcl} H_0 & : & \mu_A = \mu_B \\ H_A & : & \mu_A \neq \mu_B. \end{array}$$

On choisit la statistique de test T définie par la relation

$$T = \frac{\overline{X} - \overline{Y}}{\hat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}},$$

où m et n sont la taille des échantillons et

$$\hat{\sigma}^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{i=1}^{m}(Y_i - \overline{Y})^2\right).$$

Sous l'hypothèse nulle, la statistique T suit une loi de Student à $m+n-2$ degrés de liberté. Dans le cas de l'exercice, les valeurs observées sont $m = n = 15$, $\overline{x} \simeq 2{,}99$, $\overline{y} \simeq 2{,}90$, $\hat{\sigma}_x^2 \simeq 0{,}29$ et $\hat{\sigma}_y^2 \simeq 0{,}25$. Ainsi, $t_{obs} \simeq 0{,}47$ et la p-valeur est

$$p\text{-valeur} = P(\mid T \mid > t_{obs}) \simeq 0{,}64.$$

On ne rejette pas l'hypothèse nulle, on ne peut pas affirmer que les moyennes des échantillons sont différentes.

8.20

1. Puisque les 2 échantillons sont indépendants et les observations issues d'une loi normale, nous effectuons un test de Student à 2 échantillons, pour tester
$$\begin{array}{rcl} H_0 & : & \mu_S = \mu_D \\ H_A & : & \mu_S \neq \mu_D, \end{array}$$
où μ_S (respectivement μ_D) est le nombre moyen d'heures de sommeil additionnel pour les utilisateurs de Sopo (respectivement Dodo).
2. On calcule $\bar{x}_D = 1{,}05$, $\bar{x}_S = 1{,}89$ et les écart-types estimés $\hat{\sigma}_D^2 = 1{,}60$ et $\hat{\sigma}_S^2 = 3{,}15$. La statistique de test observée est
$$t_{obs} = \frac{\bar{x}_S - \bar{x}_D}{\sqrt{\frac{\hat{\sigma}_S^2 + \hat{\sigma}_D^2}{n}}} = \frac{1{,}03 - 1{,}05}{\sqrt{\frac{1{,}60-3{,}15}{10}}} \simeq 1{,}22$$
Puisque le quantile 0,975 de la loi t_9 est égal à 2,26, on n'a pas assez d'évidence pour rejeter H_0.
3. (a) Définissons la différence $\Delta = \mu_S - \mu_D$. On teste alors
$$\begin{array}{rcl} H_0 & : & \Delta = 0 \\ H_A & : & \Delta \neq 0 \end{array}$$
avec la statistique de test
$$T = \frac{\bar{D}}{\frac{\hat{\sigma}_\Delta}{\sqrt{n}}}$$
où $D_i = X_i^S - X_i^D$, $\bar{D} = \bar{X}_S - \bar{X}_D$ et $\hat{\sigma}_\Delta^2 = 1/(n-1)\sum_{i=1}^n (D_i - \bar{D})^2$. Sous H_0 on a que $T \sim t_{(n-1)}$.
La valeur observée de la statistique sur l'échantillon vaut
$$t_{obs} = \frac{0{,}84}{\sqrt{\frac{0{,}56}{10}}} = 3{,}55,$$
ce qui est plus grand que $t_{0{,}975;9} = 2{,}26$, le quantile 0,975 de la loi $t_{(n-1)}$. On rejette H_0 au profit de H_A. Il y a bien une différence entre les 2 somnifères.
   (b) Le dernier test est plus approprié que le premier parce que nous tenons compte de la physiologie de chaque individu.

8.21

L'échantillon provenant d'une loi de Poisson de paramètre λ, on sait que $E(X) = \text{var}(X) = \lambda$. On en tire les hypothèses de test

$$\begin{array}{lcl} H_0 & : & \lambda = \lambda_0 = 25 \\ H_A & : & \lambda \neq 25. \end{array}$$

La statistique de test est $\overline{X}$ et sa valeur observée $\overline{x} = 27$. On se sert de la nouvelle variable aléatoire normale centrée et réduite Z définie comme

$$Z = \frac{\overline{X} - \lambda}{\sqrt{n\lambda}}$$

pour trouver la p-valeur

$$p\text{-valeur} = P_{H_0}\left(\left|\frac{\overline{X} - \lambda_0}{\sqrt{n\lambda_0}}\right| > \left|\frac{27 - \lambda_0}{\sqrt{n\lambda_0}}\right|\right) \simeq P(|\ Z\ | > 2{,}19) \simeq 0{,}029.$$

On rejette donc l'hypothèse nulle à 5 %, mais pas à 1 %.

8.22

Soient les hypothèses

$$\begin{array}{lcl} H_0 & : & \mu = 300 \\ H_A & : & \mu < 300. \end{array}$$

La valeur moyenne mesurée sur l'échantillon est de 299,03 mL. Soit la statistique de test

$$T = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Sous l'hypothèse de normalité des X_i, la statistique T suit une loi normale de paramètres (0,1). La p-valeur est

$$\begin{aligned} p\text{-valeur} &= P_{H_0}(\overline{X} < 299{,}03) = \\ &= \quad P\left(T < \frac{299{,}03 - 300}{3/\sqrt{6}}\right) \simeq P(T < -0{,}792) \simeq 0{,}21. \end{aligned}$$

On ne rejette donc pas l'hypothèse nulle. Pour déterminer la puissance du test, il est nécessaire de calculer la valeur critique k_α avec un seuil $\alpha = 5$ %

$$P_{H_0}(\overline{X} < k_{0{,}05}) = P\left(T < \frac{k_{0{,}05} - 300}{3/\sqrt{6}}\right) = 0{,}05$$

$$\Leftrightarrow \quad \frac{k_{0{,}05} - 300}{3/\sqrt{6}} \simeq -1{,}64 \quad \Leftrightarrow \quad k_{0{,}05} \simeq 298.$$

On trouve la puissance du test

$$\beta(\mu) = P_{H_A}(\overline{X} < 298) = P\left(T < \frac{298-\mu}{3/\sqrt{6}}\right) = \Phi\left(\frac{298-\mu}{3}\sqrt{6}\right).$$

où $\Phi(\cdot)$ est la fonction de répartition de la loi normale. On obtient alors

1. $\beta(299) \simeq 0{,}21$;
2. $\beta(295) \simeq 0{,}99$;
3. $\beta(290) \simeq 1$.

8.23

Soit $X_1, \ldots, X_{1\ 000}$ des variables aléatoires qui valent 1 si on obtient pile et 0 sinon. Elles suivent une loi de Bernoulli avec probabilité p, p valant $1/2$ si la pièce est bien équilibrée et 0,55 dans le cas contraire. La somme $S = \sum_{i=1}^{1\ 000} X_i$ suit alors une loi binomiale de paramètres $(1\ 000, p)$.

1. Si la pièce est réellement homogène

$$P\left(S > 525\right) = P\left(\frac{S-500}{\sqrt{250}} > \frac{525-500}{\sqrt{250}}\right) \simeq 1 - P(Z < 1{,}58) \simeq 0{,}06,$$

   où Z suit une loi normale centrée et réduite par approximation.
2. Si la pièce est mal équilibrée

$$P(S < 525) = P\left(\frac{S-550}{\sqrt{247{,}5}} < \frac{525-550}{\sqrt{247{,}5}}\right) \simeq P(Z < -1{,}59) \simeq 0{,}06.$$

Les 2 résultats sont égaux parce que le nombre de piles mesuré se trouve exactement au milieu de l'espérance de S pour chacun des cas, et les variances ne diffèrent que peu.

8.24

1. Les hypothèses formulées sont

$$\begin{array}{lcl} H_0 & : & \mu = \mu_0 = 450 \\ H_A & : & \mu > 450. \end{array}$$

   Soit la statistique de test $\overline{X}$. La valeur observée est $\overline{x} = 461$. Calculons la p-valeur

$$\begin{aligned} p\text{-valeur} &= P_{H_0}(\overline{X} > \overline{x}) = P\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}}\right) \\ &= P(Z > 2{,}46) \simeq 0{,}007, \end{aligned}$$

   où Z suit une loi normale de paramètres $(0, 1)$ par le théorème central limite. On rejette donc H_0 et on accepte l'hypothèse alternative. En d'autres termes, on peut dire que la moyenne est supérieure à 450.

2. Pour déterminer la puissance du test $\overline{X}$, il faut d'abord calculer le seuil critique k_α

$$\frac{k_\alpha - \mu_0}{\sigma/\sqrt{n}} = z_{1-\alpha} \quad \Leftrightarrow \quad k_\alpha = \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}} \simeq 457{,}4.$$

Calculons à présent la probabilité d'erreur de 2^e espèce

$$\begin{aligned} P_{H_A}(\overline{X} < k_\alpha) &= P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{k_\alpha - \mu}{\sigma/\sqrt{n}}\right) \\ &= P\left(Z < \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{1-\alpha}\right) \\ &= P\left(Z < \frac{-10}{100/\sqrt{500}} + 1{,}645\right) \\ &\simeq P(Z < -0{,}59) \simeq 0{,}28. \end{aligned}$$

La puissance du test au seuil de 5 % pour détecter une augmentation de 10 points est donc

$$\beta \simeq 1 - 0{,}28 = 0{,}72.$$

Cela signifie qu'on détectera une différence de 10 points environ 72 fois sur 100.

8.25

1. On formule les hypothèses suivantes

$$\begin{aligned} H_0 &: \mu = 1\,600 \\ H_A &: \mu < 1\,600. \end{aligned}$$

2. Soit la statistique de test

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}},$$

qui suit une loi de Student à $n - 1$ degrés de liberté.

3. La valeur observée de la statistique est $t_{obs} = 1\,590$ et la p-valeur correspondante

$$p\text{-valeur} = P(\overline{X} < t_{obs}) = P\left(T < -\frac{4}{3}\right) \simeq 0{,}101.$$

À un seuil de 1 %, on ne rejette pas l'hypothèse nulle.

4. Commençons par déterminer la valeur critique k_α.

$$\begin{aligned} \frac{k_{0,01} - \mu_0}{s/\sqrt{n}} &= t_{n-1;0,01} \simeq -2{,}6 \\ \Leftrightarrow k_{0,01} &\simeq 1\,581. \end{aligned}$$

L'erreur de 2^{e} espèce est

$$P_{H_A}(\overline{X} > k_{0,01}) = P_{H_A}(\overline{X} > k_{0,01}) = P\left(T > \frac{1581 - 1570}{30/\sqrt{16}}\right) \simeq 0{,}09,$$

et la puissance du test

$$\beta \simeq 0{,}91.$$

8.26

1. La densité conjointe de l'échantillon est le produit

$$\prod_{i=1}^{n} f_\theta(x_i) = \frac{1}{\Gamma^n(\alpha)}\theta^{n\alpha}\left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \exp\left(-\theta\sum_{i=1}^{n} x_i\right).$$

Pour trouver le test le plus puissant, on se sert du lemme de Neyman-Pearson qui définit la région de rejet

$$\frac{\prod_{i=1}^{n} f_{\theta_1}(x_i)}{\prod_{i=1}^{n} f_{\theta_0}(x_i)} = \frac{\frac{1}{\Gamma^n(\alpha)}\theta_1^{n\alpha}\left(\prod_{i=1}^{n} x_i\right)^{\alpha-1}\exp\left(-\theta_1\sum_{i=1}^{n} x_i\right)}{\frac{1}{\Gamma^n(\alpha)}\theta_0^{n\alpha}\left(\prod_{i=1}^{n} x_i\right)^{\alpha-1}\exp\left(-\theta_0\sum_{i=1}^{n} x_i\right)} > k_\alpha,$$

et comme $\theta_1 > 0$ et $\theta_0 > 0$, on a la région de rejet équivalente

$$\exp\left((\theta_0 - \theta_1)\sum_{i=1}^{n} x_i\right) > \tilde{k}_\alpha$$

Finalement, la statistique de test est $T = \sum_{i=1}^{n} X_i$ et

$$T > c_\alpha, \quad \text{si } \theta_0 > \theta_1$$

$$T < c_\alpha, \quad \text{si } \theta_0 < \theta_1.$$

La valeur critique c_α est déterminée en approximant la distribution de T par une loi normale.

2. Si l'hypothèse alternative est $H_A : \theta > \theta_0$, on procède de la même manière et on arrive à

$$T = \sum_{i=1}^{n} X_i < c_\alpha.$$

8.27

Comme $E(X) = \lambda^{-1}$, les hypothèses sont

$$\begin{array}{lcl} H_0 & : & \lambda = \lambda_0 = 1/30 \\ H_A & : & \lambda \neq 1/30. \end{array}$$

La statistique de test est

$$Z = \frac{\overline{X} - \lambda^{-1}}{\frac{1}{\sqrt{\lambda^2 n}}} = \sqrt{n}(\lambda \overline{X} - 1),$$

La variable aléatoire Z suivant une loi normale centrée et réduite par hypothèse. La statistique de test observée est $\overline{x} \simeq 27{,}6$ et la p-valeur vaut

$$p\text{-valeur} \simeq 2\,(1 - P(|\ Z\ | < 1{,}6)) \simeq 0{,}11.$$

On ne peut pas rejeter l'hypothèse nulle H_0.

Les seuils critiques $k_{0,025}$ et $k_{0,975}$ pour $\alpha = 5\ \%$ sont déterminés par

$$\sqrt{n}(k_{0,025}\lambda_0 - 1) = z_{0,025}$$

et

$$\sqrt{n}(k_{0,975}\lambda_0 - 1) = z_{0,975}.$$

On trouve ainsi $k_{0,025} \simeq 27{,}06$ et $k_{0,975} \simeq 32{,}94$.

La puissance du test β pour l'alternative λ s'écrit

$$\begin{aligned} \beta(\lambda) &= 1 - P_{H_A}(27{,}06 < \overline{X} < 32{,}94) = \\ &= \quad 1 - P\left(\sqrt{n}(27{,}06\lambda - 1) < Z < \sqrt{n}(32{,}94\lambda - 1)\right), \end{aligned}$$

ce qui nous donne $\beta(1/25) \simeq 95\ \%$, $\beta(1/28) \simeq 25\ \%$ et $\beta(1/35) \simeq 88\ \%$.

8.28

Soient les hypothèses

$$\begin{array}{lcl} H_0 & : & \mu = 95 \\ H_A & : & \mu < 95, \end{array}$$

la statistique de test $T = \overline{X}$ et la variable aléatoire Z telle que

$$Z = \frac{T - \mu}{\sigma/\sqrt{n}}.$$

La statistique Z suit une loi normale de paramètres $(0, 1)$ par hypothèse.

1. Calculons la p-valeur

$$p\text{-valeur} = P_{H_0}(T < 94{,}32) = P\left(Z < \frac{94{,}32 - 95}{1{,}2/\sqrt{16}}\right) \simeq 0{,}012.$$

Au seuil de 1 %, on ne rejette pas l'hypothèse nulle.

2. Pour calculer la puissance du test, il faut trouver la valeur critique $k_{0,01}$:

$$P_{H_0}(T < k_{0,01}) = P\left(Z < \frac{k_{0,01} - 95}{1{,}2/\sqrt{16}}\right) = 0{,}01.$$

On obtient

$$k_{0,01} = \frac{1{,}2}{4} \cdot z_{0,01} + 95 \simeq 94{,}30.$$

Si la vraie valeur de μ est 94, l'erreur de 2^{e} espèce est

$$P_{H_A}(T > k_{0,01}) \simeq P\left(Z > \frac{94{,}30 - 94}{1{,}2/4}\right) \simeq 0{,}16$$

et, par conséquent, la puissance du test est

$$\beta(94) \simeq 0{,}84.$$

3. On aimerait trouver la valeur de n qui assure une puissance de test d'au moins 0,99. On utilise la relation suivante

$$n = \left(\frac{z_{1-\alpha} - z_{1-\beta}}{(\mu_A - \mu_0)/\sigma}\right)^2,$$

où β est la puissance du test et α son seuil. Ici, on obtient

$$n = \left(\frac{z_{0,99} - z_{0,01}}{(94 - 95)/1{,}2}\right)^2 \simeq 31{,}3.$$

La taille de l'échantillon n doit valoir au moins 32 pour assurer une puissance de 0,99.

8.29

Soit μ l'espérance de l'augmentation du chiffre d'affaire parmi les clients de l'échantillon et soient les hypothèses

$$\begin{array}{lcl} H_0 & : & \mu = 0 \\ H_A & : & \mu > 0. \end{array}$$

Soit la statistique de test $T = \overline{X}$ et la variable aléatoire Z définie par

$$Z = \frac{T - \mu}{\sqrt{\widehat{\text{var}}(\overline{X})}}.$$

Par le théorème central limite, Z suit une loi normale de paramètres (0, 1).

1. Calculons la p-valeur

$$p\text{-valeur} = P_{H_0}(T > 332) = P\left(Z > \frac{332}{108}\right) \simeq P(Z > 3{,}07) \simeq 0{,}11\ \%.$$

On rejette l'hypothèse nulle, l'offre de la banque génère une augmentation du chiffre d'affaire.

2. Puisque Z suit approximativement une loi normale centrée réduite, on cherche un intervalle de confiance pour μ tel que

$$P(-z_{0,995} < Z < z_{0,995}) = 0{,}99.$$

On trouve donc

$$\text{IC} = \left[\overline{X} \pm z_{0,995}\hat{\sigma}\right] = [53{,}36\,;\,610{,}64],$$

qui, comme attendu, ne contient pas la valeur 0.

3. Calculons la valeur critique $k_{0,99}$

$$P_{H_0}(T > k_{0,99}) = 0{,}01 \quad \Leftrightarrow \quad k_{0,99} = \hat{\sigma} \cdot z_{0,99} \simeq 251{,}64.$$

On en déduit la puissance du test par rapport à l'alternative $\mu = 150$

$$\begin{aligned}\beta(150) &= P_{\mu=150}(T > k_{0,99}) = \\ &= P\left(Z > \frac{251{,}64 - 150}{108}\right) \simeq P(Z > 0{,}94) \simeq 0{,}17.\end{aligned}$$

4. Pour déterminer la taille n de l'échantillon qui assure une puissance de test égale à 80 %, on utilise la relation suivante

$$n = \left(\frac{z_{1-\alpha} - z_{1-\beta}}{(\mu_A - \mu_0)/(\hat{\sigma} \cdot \sqrt{N})}\right)^2,$$

où $N = 200$ est la taille du premier échantillon. On a donc

$$n \simeq \left(\frac{2{,}33 + 0{,}84}{150/(108\sqrt{200})}\right)^2 \simeq 1\ 022.$$

8.30

1. Soient $\overline{X}_L$ et $\overline{X}_W$ les temps moyens d'installation de Linux et de WinNT. En utilisant le théorème central limite, on a pour l'installation de Linux que

$$Z_L = \frac{\overline{X}_L - \mu_L}{\hat{\sigma}_L/\sqrt{n}} \sim \mathcal{N}(0{,}1)$$

et la même chose avec Z_W pour WinNT.
Pour Linux, on a $\overline{x}_L \simeq 167$ et $\hat{\sigma}_L \simeq 14{,}4$ ce qui donne l'intervalle de confiance à 95 % suivant

$$\text{IC} \simeq 167 \pm 1{,}96\frac{14{,}4}{\sqrt{10}} \simeq [158\,;\,176]\,.$$

En ce qui concerne WinNT, $\overline{x}_W \simeq 154$, $\hat{\sigma}_W \simeq 8{,}19$ et IC $\simeq [149\,;\,159]$.

2. On procède à un test de Student pour 2 échantillons. Les hypothèses sont

$$\begin{array}{rcl} H_0 & : & \mu_L = \mu_W \\ H_A & : & \mu_L > \mu_W. \end{array}$$

Les variances étant différentes, on utilise la statistique de test

$$T = \frac{\overline{X}_L - \overline{X}_W}{\sqrt{\frac{\hat{\sigma}_L^2 + \hat{\sigma}_W^2}{n}}}$$

qui suit approximativement une loi de Student à $n-1$ degrés de liberté. La valeur observée de cette statistique est $t_{obs} \simeq 2{,}54$ et elle est supérieure au quantile $t_{0,95} \simeq 1{,}83$ ce qui signifie que l'on rejette H_0. On en déduit que les durées d'installation des 2 logiciels semblent différentes.

3. (a) Maintenant que les variances sont connues, la statistique de test devient

$$\tilde{T} = \frac{\overline{X}_L - \overline{X}_W}{\sqrt{\frac{\sigma_L^2 + \sigma_W^2}{n}}}.$$

   (b) $\tilde{T}$ suit une loi normale centrée et réduite.

   (c) si $z_{1-\alpha}$ est le α-quantile de la loi normale centrée réduite, la puissance du test s'écrit

$$\begin{aligned} \beta = P_{H_A}(T > z_{1-\alpha}) &= P_{H_A}\left(\frac{\overline{X}_L - \overline{X}_W}{\sqrt{\frac{\sigma_L^2 + \sigma_W^2}{n}}} > z_{1-\alpha} \right) \\ &= 1 - P_{H_A}\left(\frac{(\overline{X}_L - \overline{X}_W) - (\mu_L - \mu_W)}{\sqrt{\frac{\sigma_L^2 + \sigma_W^2}{n}}} < z_{1-\alpha} - \frac{(\mu_L - \mu_W)}{\sqrt{\frac{\sigma_L^2 + \sigma_W^2}{n}}} \right) \\ &= 1 - \Phi\left(z_{1-\alpha} - \frac{(\mu_L - \mu_W)}{\sqrt{\frac{\sigma_L^2 + \sigma_W^2}{n}}} \right). \end{aligned}$$

Cela signifie que

$$1 - \beta = \Phi\left(z_{1-\alpha} - \frac{(\mu_L - \mu_W)}{\sqrt{\sigma_L^2 + \sigma_W^2}} \sqrt{n} \right),$$

donc

$$z_{1-\beta} = z_{1-\alpha} - \frac{(\mu_L - \mu_W)}{\sqrt{\sigma_L^2 + \sigma_W^2}} \sqrt{n},$$

et finalement

$$n = \left(\frac{z_{1-\alpha} - z_{1-\beta}}{\frac{\mu_L - \mu_W}{\sqrt{\sigma_L^2 + \sigma_W^2}}} \right)^2 \simeq \left(\frac{1{,}64 + 0{,}84}{\frac{10}{\sqrt{225+100}}} \right)^2 \simeq 20.$$

8.31

On a la table de contingence

	Gauchers	Droitiers	Total
Fume	190	300	490
Ne fume pas	210	500	710
Total	400	800	1 200

On teste l'hypothèse nulle H_0: « la proportion de fumeurs est identique chez les droitiers et les gauchers. » Soit O_{ij} la quantité réellement observée et E_{ij} la quantité espérée sous l'hypothèse nulle pour la catégorie (i, j), où $i = 1$ si la personne est gauchère et $j = 1$ si elle fume. On obtient

O_{ij}	E_{ij}	$(O_{ij} - E_{ij})^2/E_{ij}$
190	163	4,37
210	237	3,01
300	327	2,18
500	473	1,51

Soit la statistique $X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ qui suit une loi χ^2 à 1 degré de liberté. La valeur observée de la statistique est $x^2_{obs} \simeq 11{,}1$ et on en déduit la p-valeur

$$p\text{-valeur} = P(X^2 > 11{,}1) \simeq 0{,}001.$$

On rejette donc H_0 au seuil de 5 %.

8.32

Sous l'hypothèse nulle, $Z \sim \mathcal{N}(100, 225)$. On compare les valeurs espérées sous l'hypothèse nulle par rapport aux valeurs réellement observées. Par exemple, pour la 1re classe, la valeur espérée est

$$E_1 = E_{[0,70)} = 1\ 000 \cdot P_{H_0}\left(Z \in [0, 70)\right) \simeq 1\ 000 \cdot 0{,}0228 = 22{,}8.$$

Score	observés (O_i)	espérés (E_i)	$(O_i - E_i)^2/E_i$
[0,70)	34	22,8	5,50
[70,85)	114	135,9	3,53
[85,100)	360	341,3	1,02
[100,115)	344	341,3	0,02
[115,130)	120	135,9	1,86
[130,∞)	28	22,8	1,19

La somme $X^2 = \sum_{i=1}^{6}(O_i - E_i)^2/E_i = 13{,}12$ est la valeur sur l'échantillon de la statistique de test qui suit une loi χ^2_5. La p-valeur est ainsi

$$p\text{-valeur} = P_{H_0}(T > 13{,}12) \simeq 0{,}022.$$

On rejette donc l'hypothèse nulle au seuil de 5 % et on ne la rejette pas au seuil de 1 %.

8.33

Soit X_i une variable aléatoire qui vaut 1 avec une probabilité p si la fleur i est rouge et 0 sinon. Les hypothèses que l'on veut tester sont

$$\begin{array}{lcl} H_0 & : & p = \frac{3}{4} \\ H_A & : & p \neq \frac{3}{4}. \end{array}$$

Soit $S = \sum_{i=1}^{500} X_i$ qui suit une loi binomiale de paramètres $(500, p)$ et T la statistique de test définie par

$$T = \frac{S - E(S)}{\sqrt{\text{var}(S)}}.$$

Par approximation du théorème central limite, T suit une loi normale centrée et réduite. La valeur observée de la statistique sur l'échantillon est

$$t_{obs} = \frac{350 - 500\frac{3}{4}}{\sqrt{500\frac{3}{4}\frac{1}{4}}} \simeq -2{,}58$$

et on en déduit la p-valeur

$$p\text{-valeur} = P_{H_0}(\mid T \mid > \mid t_{obs} \mid) = P(\mid T \mid > \mid -2{,}58 \mid) = 2 - 2\Phi(2{,}58) \simeq 0{,}01.$$

On rejette l'hypothèse nulle à 5 %, mais la décision n'est pas claire à 1 %. En d'autres termes, la loi de Mendel n'est pas vérifiée pour un seuil de 5 %.

8.34

1. Il faut procéder à un test d'adéquation. On teste

$$\begin{array}{lcl} H_0 & : & \text{les réponses sont uniformément distribuées} \\ H_A & : & H_0 \text{ est fausse.} \end{array}$$

On utilise la statistique du test d'adéquation du χ^2

$$X^2 = \sum_{i=1}^{5} \frac{(O_i - E_i)}{E_i} \sim X_4^2.$$

Par l'hypothèse d'uniformité sous H_0, les valeurs observées E_i sont toutes égales à 13,4, ce qui donne

$$x_{obs}^2 = \frac{1}{13{,}4}(1{,}96 + 12{,}96 + 31{,}36 + 5{,}76 + 29{,}16) = \frac{81{,}2}{13{,}4} = 6{,}06.$$

Or, le quantile à 95 % de la loi χ^2 avec 4 degrés de liberté est égal à 9,49 qui est plus grand que x_{obs}^2. On ne peut donc pas rejeter H_0.

2. Il s'agit ici de faire un test d'indépendance pour tester

H_0 : l'appréciation est indépendante du sexe de l'étudiant
H_A : H_0 est fausse.

Sur la base du tableau des fréquences espérées (E_{ij}) suivant

	Bonnes	Moyennes	Mauvaise	Total
Hommes	14,28	11,33	7,39	33
Femmes	14,72	11,67	7,61	34
Total	29	23	15	

on procède à un test d'indépendance du χ^2, avec la statistique

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

qui suit une loi χ^2_2 sous H_0. La valeur observée x^2_{obs} est 1,30, plus petite que le quantile à 95 % de la loi χ^2_2 qui vaut 5,99. On ne peut donc pas rejeter H_0.

8.35

Soit l'hypothèse nulle H_0 « l'habitude et la préférence sont indépendants. » Sous H_0, on aurait dû observer

		habituellement A	boivent B	total
ont préféré	A	55	55	110
	B	45	45	90
total		100	100	200

Soit X^2 la statistique de test définie par

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(E_{ij} - O_{ij})^2}{E_{ij}},$$

où E_{ij} et O_{ij} sont respectivement les quantités espérées et observées dans la catégorie (i, j). Cette statistique suit une loi χ^2 à 1 degré de liberté et la valeur que l'on observe sur l'échantillon est $x^2_{obs} \simeq 8{,}08$. La p-valeur correspondante est alors

$$p\text{-valeur} = P(X > 8{,}08) \simeq 0{,}0045$$

et elle implique que l'hypothèse nulle est rejetée ; l'habitude et la préférence sont dépendants.

8.36

1. L'hypothèse nulle H_0 est « la position de départ n'influence pas le résultat. »
2. On fait un test d'adéquation du χ^2.
3. Les valeurs espérées sous H_0 sont

position de départ	1	2	3	4	5	6	7	8
nombre de vainqueurs	18	18	18	18	18	18	18	18

Soit X la statistique de test que nous écrivons comme

$$X = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i},$$

où O_i et E_i sont les $i^{\text{èmes}}$ valeurs observées et espérées. La statistique X suit une loi χ^2 à 7 degrés de liberté et sa valeur réalisée sur l'échantillon est $x^2_{obs} \simeq 16{,}33$. On en déduit la p-valeur suivante

$$p\text{-valeur} = P(X > 16{,}33) \simeq 0{,}022$$

qui conduit à rejeter l'hypothèse nulle à un seuil de 5 % et à l'« accepter » à un seuil de 1 %.

8.37

Soit l'hypothèse nulle H_0 : « Il y a indépendance entre le niveau de dépenses et le statut professionnel. » Soit X la statistique de test telle que

$$X^2 = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

La statistique X^2 suit une loi χ^2 à 4 degrés de liberté sous l'hypothèse nulle. Le tableau des effectifs espérés E_{ij} est le suivant

	plein temps	temps partiel	sans profession
Moins de 10 $	30,8	35,2	44
De 10 $ à 25 $	50,4	57,6	72
Plus de 25 $	58,8	67,2	84

ce qui donne une valeur observée pour la statistique X^2 de $x^2_{obs} \simeq 17{,}25$. On en déduit la p-valeur

$$p\text{-valeur} = P(X > 17{,}25) \simeq 0{,}17\ \%.$$

Par conséquent, on rejette H_0 au seuil de 5 % ou, en d'autres termes, le niveau de dépense et le statut professionnel ne sont pas indépendants. On pourrait donc conseiller aux responsables marketing de produits cosmétiques d'orienter leur publicité suivant le statut professionnel des femmes.

8.38

1. À l'aide d'une table de contingence, on aimerait tester l'hypothèse nulle qui suppose l'indépendance entre le fait d'être daltonien et le sexe d'un patient. Récrivons la table sous la forme suivante

	homme	femme	total
normal	O_{11}	O_{12}	O_{1+}
daltonien	O_{21}	O_{22}	O_{2+}
total	O_{+1}	O_{+2}	O_{++}

On définit $E_{ij} = O_{i+}O_{+j}/O_{++}$ et la statistique de test X^2 comme

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Sous l'hypothèse nulle, X^2 suit une distribution χ^2 avec $(2-1)\times(2-1) = 1$ degré de liberté. La valeur observée de la statistique est $x^2_{obs} \simeq 27{,}14$. La valeur critique de la statistique à un seuil de 1 % est environ 6,63. Par conséquent, on rejette l'hypothèse nulle ; le daltonisme dépend du sexe du patient.

2. On veut maintenant tester l'hypothèse nulle affirmant que les mesures suivent le modèle de fréquences tel qu'il est donné dans l'énoncé. La proportion de daltoniens mesurée dans l'échantillon est $\hat{q} = 0{,}044$. Soit la statistique de test X^2 définie comme

$$X^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i},$$

où E_i est la quantité espérée selon le modèle et O_i celle qui est effectivement observée pour l'observation i. La statistique X^2 suit approximativement une loi χ^2 à 2 degrés de liberté sous l'hypothèse nulle. La valeur observée de la statistique est $x^2_{obs} \simeq 40$ et donne une p-valeur égale à 0. L'hypothèse nulle est donc rejetée ; les mesures ne suivent pas le modèle génétique.

8.39

Soit le tableau de contingence

	Fumeur	Non fumeur	Total
Chocolat	400	200	600
Pas de chocolat	300	100	400
Total	700	300	1 000

qu'on paramétrise de la manière suivante

	Fumeur	Non fumeur	Total
Chocolat	O_{11}	O_{12}	O_{1+}
Pas de chocolat	O_{21}	O_{22}	O_{2+}
Total	O_{+1}	O_{+2}	O_{++}

On fait l'hypothèse nulle d'indépendance entre le fait d'être fumeur et de manger du chocolat. Définissons $E_{ij} = O_{i+}O_{+j}/O_{++}$ et la statistique de test X^2 comme

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}.$$

Sous l'hypothèse nulle, X^2 suit une distribution χ^2 à $(2-1)\times(2-1) = 1$ degré de liberté. La valeur observée de la statistique est $x^2_{obs} \simeq 7{,}9$ qui donne la p-valeur

$$p\text{-valeur} = P_{H_0}(X^2 > 7{,}9) = 0{,}5\ \%.$$

On rejette donc l'hypothèse nulle aux seuils de 1 et 5 %. Autrement dit, la consommation de chocolat est dépendante du fait d'être fumeur ou non.

8.40

On teste

H_0 : le poste est indépendant du sexe dans cette entreprise
H_A : H_0 est fausse.

La statistique de test

$$X^2 = \sum_{i=1}^{4}\sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})}{E_{ij}}$$

calculée à l'aide du tableau des fréquences espérées suivantes

	homme	femme
management	27,18	20,82
vente	28,31	21,68
service	11,33	8,67
autres	27,18	20,82

donne $x^2_{obs} = 24{,}90$, qui doit être comparé au quantile 0,95 de la loi χ^2_3 qui vaut 7,8. On rejette donc H_0: la répartition des postes est bel et bien liée au sexe.

8.41

On fait les hypothèses suivantes sur les peintures :

$$\begin{array}{lcl} H_0 & : & \text{même comportement} \\ H_A & : & \text{comportement différent.} \end{array}$$

On a 22 observations réparties dans $g = 4$ groupes ($i = 1, \ldots, 4$) : $n_1 = n_4 = 6$ et $n_2 = n_3 = 5$. Les moyennes des observations de chacun des groupes sont $y_{1\cdot} \simeq 87{,}67$, $y_{2\cdot} = 89{,}6$, $y_{3\cdot} = 88{,}4$ et $y_{4\cdot} \simeq 84{,}67$. La moyenne de toutes les observations est $y_{\cdot\cdot} \simeq 87{,}45$. Soient les deux statistiques

$$\begin{aligned} SS_B &= \sum_{i=1}^{4} n_j (y_{i\cdot} - y_{\cdot\cdot})^2 \\ SS_W &= \sum_{i=1}^{4} \sum_{j=1}^{n_i} (y_{ij} - y_{i\cdot})^2. \end{aligned}$$

Les valeurs observées de ces statistiques sont $ss_{B,obs} \simeq 74{,}39$ et $ss_{W,obs} \simeq 265{,}07$. Soit F une nouvelle statistique définie par

$$F = \frac{SS_B/(g-1)}{SS_W/(N-g)},$$

où $N = \sum_{i=1}^{4} n_i$.

Sous l'hypothèse nulle, F suit une loi de Fisher à $g - 1 = 3$ degrés de liberté au numérateur et $N - g = 22 - 4 = 18$ degrés de liberté au dénominateur. La valeur observée de la statistique est $f_{obs} \simeq 1{,}68$. Le quantile à 95 % de la loi de Fisher vaut environ 3,16. Par conséquent, on ne rejette pas l'hypothèse nulle ; on ne peut pas prouver que les peintures se comportent différemment.

8.42

1. La variable aléatoire Y_i suit une loi normale d'espérance θx_i^2 et de variance σ^2.
2. L'estimateur des moindres carrés $\hat{\theta}_{MC}$ de θ est celui qui minimise la quantité

$$S(\theta) = \sum_{i=1}^{n} (Y_i - \theta x_i)^2.$$

On calcule la dérivée de $S(\theta)$ par rapport à θ

$$\frac{\partial}{\partial \theta} S(\theta) = -2 \sum_{i=1}^{n} x_i^2 (Y_i - \theta x_i^2),$$

et on trouve $\hat{\theta}_{MC}$ en égalant cette dernière à zéro

$$\sum_{i=1}^{n} Y_i x_i^2 - \hat{\theta}_{MC} \sum_{i=1}^{n} x_i^4 = 0$$

$$\Leftrightarrow \hat{\theta}_{MC} = \frac{\sum_{i=1}^{n} Y_i x_i^2}{\sum_{i=1}^{n} x_i^4}.$$

L'estimateur est sans biais car

$$E(\hat{\theta}_{MC}) = \frac{1}{\sum_{i=1}^{n} x_i^4} \sum_{i=1}^{n} x_i^2 E(Y_i) = \frac{1}{\sum_{i=1}^{n} x_i^4} \theta \sum_{i=1}^{n} x_i^4 = \theta$$

et sa variance vaut

$$\text{var}(\hat{\theta}_{MC}) = \frac{1}{\left(\sum_{i=1}^{n} x_i^4\right)^2} \sum_{i=1}^{n} x_i^4 \text{var}(Y_i) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^4}.$$

3. Pour obtenir l'information de Fisher, commençons par écrire la log-vraisemblance $l(\theta \mid y_1, \ldots, y_n)$ de l'échantillon

$$\begin{aligned} l(\theta \mid y_1, \ldots, y_n) &= \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y_i - \theta x_i^2)^2\right)\right) \\ &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \theta x_i^2)^2. \end{aligned}$$

Dérivons-la une 1re fois

$$\frac{\partial}{\partial\theta} l(\theta \mid y_1, \ldots, y_n) = \frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^2(y_i - \theta x_i^2)$$

puis une 2e fois (il faut noter que $\hat{\theta}_{MC}$ est également l'estimateur du maximum de vraisemblance de θ)

$$\frac{\partial^2}{\partial\theta^2} l(\theta \mid y_1, \ldots, y_n) = -\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^4.$$

L'information de Fisher $J(\theta)$ est par conséquent

$$J(\theta) = -E\left(-\frac{1}{\sigma^2}\sum_{i=1}^{n} x_i^4\right) = \frac{\sum_{i=1}^{n} x_i^4}{\sigma^2}$$

et la borne de Cramér-Rao

$$\text{BCR} = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^4}.$$

4. La variance de $\hat{\theta}_{MC}$ atteint la borne de Cramér-Rao. Son efficacité est donc égale à 1.
5. Soit T la statistique définie selon

$$T = \frac{\sum_{i=1}^{n} x_i^2 Y_i}{\sum_{i=1}^{n} x_i^4} = \hat{\theta}_{MC},$$

avec $T \sim \mathcal{N}\left(\theta, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^4}\right)$. On construit l'intervalle de confiance au degré $(1-\alpha)$ depuis

$$P\left(-z_{1-\alpha/2} < \frac{\frac{\sum_{i=1}^{n} x_i^2 Y_i}{\sum_{i=1}^{n} x_i^4} - \theta}{\frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^4}}} < z_{1-\alpha/2}\right) = 1-\alpha$$

$$\Leftrightarrow \quad \text{IC} = \left[\frac{\sum_{i=1}^{n} x_i^2 Y_i}{\sum_{i=1}^{n} x_i^4} \pm \frac{\sigma z_{1-\alpha/2}}{\sqrt{\sum_{i=1}^{n} x_i^4}}\right].$$

8.43

1. Construisons un intervalle de confiance pour α. Soit $\hat{\alpha}_{MV}$ l'estimateur du maximum de vraisemblance de α. Soit la statistique de test T telle que

$$T = \frac{\hat{\alpha}_{MV} - \alpha}{\sqrt{\widehat{\text{var}}(\hat{\alpha}_{MV})}}.$$

La statistique T suit une loi de Student à $n-2$ degrés de liberté. On trouve alors l'intervalle de confiance suivant

$$\text{IC} = \left[\hat{\alpha}_{MV} \pm t_{n-2;1-\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\alpha}_{MV})}\right],$$

où $t_{n-2;1-\alpha/2}$ est le $\alpha/2$-quantile d'une loi de Student à $n-2$ degrés de liberté. De la même manière, on trouve un intervalle de confiance pour le paramètre β :

$$\text{IC} = \left[\hat{\beta}_{MV} \pm t_{n-2;1-\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\beta}_{MV})}\right].$$

2. Faisons maintenant l'hypothèse nulle $H_0 : \beta = 0$. La statistique de test T pour β sous l'hypothèse nulle est

$$T = \frac{\hat{\beta}_{MV}}{\sqrt{\widehat{\text{var}}(\hat{\beta}_{MV})}}.$$

Cette statistique suit elle aussi une loi de Student à $n-2$ degrés de liberté. On rejette l'hypothèse nulle si la p-valeur est inférieure au seuil α choisi.

8.44

1. Nous avons la relation

$$\mathrm{QI}_i = \alpha + \beta \mathrm{TC}_i + \epsilon_i,$$

dans laquelle nous remplaçons la variable TC par 910 000 pour obtenir

$$\widehat{\mathrm{QI}} = \hat{\alpha}_{MC} + \hat{\beta}_{MC} \cdot \mathrm{TC} = 1{,}74 + 0{,}00012 \cdot 910\ 000 = 110{,}94.$$

2. Calculons à présent un intervalle de confiance à 95 % pour β. On sait que la statistique T définie par

$$T = \frac{\hat{\beta}_{MC} - \beta}{\sigma / \sqrt{S_{xx}}}$$

suit une loi de Student avec $n - 2 = 38$ degrés de liberté et on en déduit l'intervalle de confiance

$$\begin{aligned} \mathrm{IC} &= \left[\hat{\beta}_{MC} \pm t_{0,975} \frac{\hat{\sigma}}{\sqrt{S_{xx}}} \right] \\ &= \left[0{,}00012 \pm 2{,}02 \frac{20{,}99}{457\ 152} \right] \simeq [0{,}00003\,;\, 0.00021]\,. \end{aligned}$$

3. Dans le point 2., on voit clairement que l'intervalle à 95 % pour β ne comprend pas 0 ce qui signifie que le paramètre est significatif et que le QI dépend de la taille du cerveau.

8.45

1. L'ajustement n'est pas bon car les résidus ne sont pas distribués de manière aléatoire mais plutôt selon une forme quadratique.
2. Proposons un nouveau modèle qui tienne compte d'une relation quadratique, à savoir

$$Y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i.$$

3. Soit X la matrice construite comme

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

et l'équation du modèle devient sous forme matricielle :

$$(X^T X)\hat{\beta}_{MC} = X^T Y.$$

Soit $L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_i - \beta_2 x_i^2)^2$ et les conditions de premier ordre :

$$\left\{ \begin{array}{l} \frac{\partial L}{\partial \alpha} = 0 \\ \frac{\partial L}{\partial \beta_1} = 0 \\ \frac{\partial L}{\partial \beta_2} = 0 \end{array} \right. \quad \Rightarrow \quad \left\{ \begin{array}{l} \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_i - \beta_2 x_i^2) = 0 \\ \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_i - \beta_2 x_i^2) x_i = 0 \\ \sum_{i=1}^{n} (y_i - \alpha - \beta_1 x_i - \beta_2 x_i^2) x_i^2 = 0. \end{array} \right.$$

4. Vu la forme quadratique et convexe des résidus, on peut supposer que, dans le marché immobilier, les maisons neuves aussi bien que les maisons anciennes ont des prix élevés contrairement aux maisons d'âge moyen.

8.46

1. L'intervalle de confiance à 95 % pour μ vaut

$$\begin{aligned} IC(\mu, 95\ \%) &= \bar{X} \pm \frac{\hat{\sigma}}{\sqrt{n}} z_{0,975} \\ &= 14 \pm \frac{2}{\sqrt{100}} 1{,}96 \\ &= 14 \pm 0{,}392 \\ &= [13{,}61\,;\, 14{,}39] \end{aligned}$$

2. Soit p =« la proportion de ménages préférant un produit sans phosphates », et

$$X_i = \left\{ \begin{array}{ll} 1 & \text{si le ménage } i \text{ préfère un produit sans phosphates} \\ 0 & \text{sinon.} \end{array} \right.$$

On estime la proportion de ménages qui préfèrent un produit sans phosphates par $\sum_{i=1}^{n} X_i$ qui est une variable aléatoire distribuée selon une loi binomiale de paramètres $(100, p)$.
On veut tester

$$\begin{aligned} H_0 &: p = 0{,}25 \\ H_A &: p > 0{,}25, \end{aligned}$$

à l'aide de la statistique

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{np(1-p)}}$$

qui suit approximativement une loi normale centrée et réduite.
La p-valeur est

$$P_{H_0}\Big(\sum_{i=1}^{n} X_i > 30\Big) = P\Big(Z > \frac{30 - 25}{\sqrt{18{,}75}}\Big) = P(Z > 1{,}154) = 0{,}124,$$

ce qui ne fournit pas assez d'évidence pour rejeter H_0 au seuil de 5 %.

8.47

1. Pour trouver l'estimateur des moments pour ξ et θ, on a

$$E(X_i) = \xi + \gamma\sqrt{\theta}$$

et

$$\text{var}(X_i) = \frac{\pi^2}{6}\theta.$$

On trouve donc

$$\hat{\xi}_M = \overline{X} - \gamma\sqrt{\hat{\theta}_M}$$

et

$$\hat{\theta}_M = \frac{6}{n\pi^2}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

2. Les estimateurs sont convergents s'ils convergent en probabilité vers la vraie valeur des paramètres lorsque $n \to \infty$. Qu'en est-il pour $\hat{\theta}_M$?

$$\text{plim}_{n\to\infty}\hat{\theta}_M = \text{plim}_{n\to\infty}\frac{6}{n\pi^2}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

On utilise ici la transformation suivante

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}\left((X_i - \mu)^2 - (\overline{X} - \mu)^2\right) = \sum_{i=1}^{n}(X_i - \mu)^2 - n(\overline{X} - \mu)^2.$$

On obtient alors

$$\begin{aligned}
&\text{plim}_{n\to\infty}\frac{6}{n\pi^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \text{plim}_{n\to\infty}\frac{6}{\pi^2}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{1}{n}n(\overline{X} - \mu)^2\right) \\
&= \frac{6}{\pi^2}(\text{var}(X_i) - (E(X_i) - \mu)^2) \\
&= \frac{6}{\pi^2}\frac{\pi^2}{6}\theta - \frac{6}{\pi^2}(\mu - \mu)^2 = \theta.
\end{aligned}$$

L'estimateur $\hat{\theta}_M$ est convergent. Nous procédons maintenant au calcul équivalent pour l'estimateur $\hat{\xi}_M$

$$\begin{aligned}
\text{plim}_{n\to\infty}\hat{\xi}_M &= \text{plim}_{n\to\infty}(\overline{X} - \gamma\sqrt{\hat{\theta}_M}) = E(X_i) - \gamma\sqrt{\theta} \\
&= \xi + \gamma\sqrt{\theta} - \gamma\sqrt{\theta} = \xi.
\end{aligned}$$

L'estimateur $\hat{\xi}_M$ est convergent également.

3. On cherche le biais de l'estimateur $\hat{\theta}_M$. Calculons son espérance

$$\begin{aligned} E(\hat{\theta}_M) &= E\left(\frac{6}{n\pi^2}\sum_{i=1}^{n}(X_i-\overline{X})^2\right) \\ &= \frac{6}{n\pi^2}E(\sum_{i=1}^{n}X_i^2 - n\overline{X}^2) = \frac{6}{\pi^2}(E(X_i^2)-E(\overline{X}^2)) \\ &= \frac{6}{\pi^2}(\mathrm{var}(X_i)+E^2(X_i)-\mathrm{var}(\overline{X})-E^2(\overline{X})) \\ &= \frac{6}{\pi^2}\left(\frac{\pi^2}{6}\theta+(\xi+\gamma\sqrt{\theta})^2-\frac{\pi^2}{6n}\theta-(\xi+\gamma\sqrt{\theta})^2\right) \\ &= \theta-\frac{\theta}{n}=\frac{n-1}{n}\theta. \end{aligned}$$

L'estimateur $\hat{\theta}_M$ est donc biaisé

$$\mathrm{biais}(\hat{\theta}_M,\theta) = -\frac{\theta}{n}.$$

4. On suppose à présent que $\theta = 6$. L'estimateur de ξ devient

$$\hat{\xi}_M = \overline{X} - \sqrt{6}\gamma.$$

Par le théorème central limite, on a approximativement

$$\frac{\hat{\xi}_M - E(\hat{\xi}_M)}{\sqrt{\mathrm{var}(\hat{\xi}_M)}} \sim \mathcal{N}(0,1),$$

où

$$\begin{aligned} E(\hat{\xi}_M) &= E(\overline{X}) - \sqrt{6}\gamma = \xi + \sqrt{6}\gamma - \sqrt{6}\gamma = \xi \\ \mathrm{var}(\hat{\xi}_M) &= \mathrm{var}(\overline{X}) = \frac{\pi^2}{n}. \end{aligned}$$

Un intervalle de confiance au degré de confiance $(1-\alpha)$ pour le paramètre ξ est défini par les inégalités

$$-z_{1-\alpha/2} < \frac{\hat{\xi}_M - \xi}{\sqrt{\pi^2/n}} < z_{1-\alpha/2}.$$

Finalement

$$\mathrm{IC} = \left[\hat{\xi}_M \pm z_{1-\alpha/2}\frac{\pi}{\sqrt{n}}\right].$$

5. La fonction réciproque de la fonction de répartition de la loi de Gumbel est

$$q(\alpha) = \xi - \sqrt{6}\log(-\log(\alpha)),$$

et $q(\alpha)$ est le α-quantile de la distribution de Gumbel. Le 0,9-quantile peut par conséquent être estimé par

$$\hat{q}_{0,9} = \hat{\xi}_M - \sqrt{6}\log(-\log(0{,}9)) = \overline{X} - \sqrt{6}(\gamma + \log(-\log(0{,}9))).$$

6. Une fois encore, le théorème central limite implique qu'approximativement
$$\frac{\hat{q}_{0,9} - E(\hat{q}_{0,9})}{\sqrt{\mathrm{var}(\hat{q}_{0,9})}} \sim \mathcal{N}(0,1),$$
avec
$$E(\hat{q}_{0,9}) = E(\overline{X}) - \sqrt{6}(\gamma - \log(-\log(0{,}9))) = \xi \pm \sqrt{6}\log(-\log(0{,}9)) = q_{0,9},$$
et
$$\mathrm{var}(\hat{q}_{0,9}) = \mathrm{var}(\overline{X}) = \frac{\pi^2}{n}.$$
Par conséquent, on obtient un intervalle de confiance au degré $(1-\alpha)$ pour le quantile $q_{0,9}$
$$-z_{1-\alpha/2} < \frac{\hat{q}_{0,9} - q_{0,9}}{\sqrt{\pi^2/n}} < z_{1-\alpha/2}$$
$$\Leftrightarrow \mathrm{IC} = [\hat{q}_{0,9} \pm \frac{\pi}{\sqrt{n}} z_{1-\alpha/2}].$$
7. Au vu des données observées, la valeur estimée du 0,9-quantile est
$$\hat{q}_{0,9} = 372{,}05 - 1{,}41 + 5{,}51 = 376{,}15\ €.$$
8. La valeur de 374,80 € est en dessous du 0,9-quantile ; ce n'est pas un niveau exceptionnel.

8.48

Soient les 3 estimateurs suivants

$$\hat{\theta}_1 = \overline{X}, \quad \hat{\theta}_2 = \frac{1}{n+1}\sum_{i=1}^{n} X_i \quad \text{et} \quad \hat{\theta}_3 = n \cdot \min(X_1, \ldots, X_n).$$

1. Les X_i provenant d'une loi exponentielle de paramètre θ^{-1}, on a $E(X_i) = \theta$ et $\mathrm{var}(X_i) = \theta^2$. On calcule alors l'espérance des estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$:
$$E(\hat{\theta}_1) = E(\overline{X}) = E(X_i) = \theta,$$
et
$$E(\hat{\theta}_2) = E\left(\frac{1}{n+1}\sum_{i=1}^{n} E(X_i)\right) = \frac{n}{n+1}\theta.$$

Le 1er estimateur est sans biais et le 2e a le biais suivant

$$\text{biais}(\hat{\theta}_2, \theta) = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta.$$

On procède ensuite aux calculs des variances

$$\text{var}(\hat{\theta}_1) = \text{var}(\overline{X}) = \frac{\theta^2}{n},$$

et

$$\text{var}(\hat{\theta}_2) = \frac{1}{(n+1)^2}\sum_{i=1}^{n}\text{var}(X_i) = \frac{n}{(n+1)^2}\theta^2.$$

On en déduit l'erreur carrée moyenne pour chacun des 2 estimateurs

$$\text{ECM}(\hat{\theta}_1, \theta) = \frac{\theta^2}{n},$$

et

$$\text{ECM}(\hat{\theta}_2, \theta) = \frac{\theta^2}{n+1}.$$

2. Soit $Y = \min(X_1, \ldots, X_n)$ une nouvelle variable aléatoire . On cherche sa fonction de répartition $F_Y(y)$

$$\begin{aligned} F_Y(y) &= P(Y < y) = P(\min_{1\le i\le n}(X_i) < y) = 1 - P(\min_{1\le i\le n}(X_i) > y) \\ &= 1 - P(X_i > y\ \forall i) = 1 - (P(X_1 > y))^n \\ &= 1 - (1 - P(X_1 < y))^n = 1 - (1 - F_{X_1}(y))^n \\ &= 1 - (1 - (1 - \exp(-y/\theta)))^n = 1 - \exp(-y\frac{n}{\theta}). \end{aligned}$$

Ce résultat implique que Y suit une loi exponentielle de paramètre $\frac{n}{\theta}$. Ainsi

$$E(\hat{\theta}_3) = n\frac{\theta}{n} = \theta \quad \text{(estimateur sans biais)},$$

$$\text{var}(\hat{\theta}_3) = n^2\frac{\theta^2}{n^2} = \theta^2,$$

et

$$\text{ECM}(\hat{\theta}_3, \theta) = \theta^2.$$

La comparaison des ECM donne alors

$$\text{ECM}(\hat{\theta}_3, \theta) > \text{ECM}(\hat{\theta}_1, \theta) > \text{ECM}(\hat{\theta}_2, \theta)\ ;$$

le 2e estimateur est le plus efficace du point de vue de l'erreur carrée moyenne.

3. Pour tester la convergence des 3 estimateurs, on étudie le comportement de leur ECM lorsque $n \to \infty$

$$\lim_{n\to\infty} \text{ECM}(\hat{\theta}_1, \theta) = 0,$$
$$\lim_{n\to\infty} \text{ECM}(\hat{\theta}_2, \theta) = 0,$$
$$\lim_{n\to\infty} \text{ECM}(\hat{\theta}_3, \theta) = \theta^2,$$

Les estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ sont donc convergents, mais nous ne pouvons pas conclure pour $\hat{\theta}_3$. Il faut alors regarder la fonction de répartition de $\hat{\theta}_3$ qui est

$$F_{\hat{\theta}_3}(t) = P(\hat{\theta}_3 < t) = P(Y < t/n) = F_Y(t/n) = 1 - \exp(t/\theta),$$

où l'on a utilisé le résultat du point 2. Puisque $\lim_{n\to\infty} F_{\hat{\theta}_3}(t) \neq 1$, l'estimateur $\hat{\theta}_3$ n'est pas convergent.

4. Soit $\lambda = 1/\theta$ et $Z = \lambda \cdot \min(X_1, \ldots, X_n)$ une nouvelle variable aléatoire. La fonction de répartition $F_Z(z)$ de Z est

$$\begin{aligned} F_Z(z) &= P(Z < z) = P(\lambda \cdot \min(X_1, \ldots, X_n) < z) \\ &= P\left(Y < \frac{z}{\lambda}\right) = F_Y\left(\frac{z}{\lambda}\right) \\ &= 1 - \exp\left(-\frac{z}{\lambda} n\lambda\right) = 1 - \exp(-zn). \end{aligned}$$

La variable aléatoire Z suit bien une loi exponentielle de paramètre n. Par conséquent, le degré de confiance de l'intervalle de confiance pour λ devient

$$\begin{aligned} &P(a/\min(X_1, \ldots, X_n) < \lambda < b/\min(X_1, \ldots, X_n)) = \\ &= P(a < \lambda \min(X_1, \ldots, X_n) < b) \\ &= F_Z(b) - F_Z(a) = 1 - \exp(-bn) - (1 - \exp(-an)) \\ &= \exp(-an) - \exp(-bn). \end{aligned}$$

8.49

1. Puisque

$$E(\hat{\theta}_1) = E(2\bar{X}) = 2E(X_i) = 2 \cdot \frac{\theta}{2} = \theta,$$

l'estimateur $\hat{\theta}_1$ est non biaisé.

2. Pour $0 \leq t \leq \theta$, la fonction de répartition de $\hat{\theta}_2$ est :

$$\begin{aligned} F_{\hat{\theta}_2}(t) &= P(\max(X_1, \ldots, X_n) \leq t) = P(X_1 < t, \ldots, X_n \leq t) \\ &= (P(X_i \leq t))^n = \left(\frac{t}{\theta}\right)^n. \end{aligned}$$

Aussi, la fonction de densité pour $0 \leq t \leq \theta$ est

$$f_{\hat{\theta}_2}(t) = \frac{\partial F_{\hat{\theta}_2}(t)}{\partial t} = n\frac{t^{n-1}}{\theta^n},$$

d'où

$$E(\hat{\theta}_2) = \int_0^\theta t f_{\hat{\theta}_2}(t)dt = \int_0^\theta n\frac{t^n}{\theta^n}dt = \frac{n}{n+1}\frac{t^{n+1}}{\theta^n}\Big|_{t=0}^{t=\theta} = \frac{n\theta}{n+1}.$$

On déduit $\hat{\theta}_3 = \frac{n+1}{n}\hat{\theta}_2 = \frac{n+1}{n} \cdot \max(X_1, \ldots, X_n)$.

3. On calcule les erreurs carrées moyennes. D'abord celle de $\hat{\theta}_1$

$$\mathrm{ECM}(\hat{\theta}_1, \theta) = \mathrm{var}(\hat{\theta}_1) = \mathrm{var}(2\bar{X}) = 4\frac{\mathrm{var}(X_i)}{n} = \frac{4}{n}\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Pour obtenir $\mathrm{var}(\hat{\theta}_2)$, il faut calculer

$$E(\hat{\theta}_2^2) = \int_0^\theta t^2 f_{\hat{\theta}_2}(t)dt = \int_0^\theta n\frac{t^{n+1}}{\theta^n}dt = \frac{n}{n+2}\frac{t^{n+2}}{\theta^n}\Big|_{t=0}^{t=\theta} = \frac{n\theta^2}{n+2},$$

d'où

$$\begin{aligned}
\mathrm{ECM}(\hat{\theta}_2, \theta) &= \mathrm{var}(\hat{\theta}_2) + \mathrm{biais}^2(\hat{\theta}_2, \theta) \\
&= E(\hat{\theta}_2^2) - E^2(\hat{\theta}_2) + \mathrm{biais}^2(\hat{\theta}_2, \theta) \\
&= \frac{n}{n+2}\theta^2 - \Big(\frac{n}{n+1}\Big)^2\theta^2 + \Big(\frac{-\theta}{n+1}\Big)^2 \\
&= \frac{n\theta^2}{(n+1)^2(n+2)} + \frac{\theta^2}{(n+1)^2} \\
&= \frac{2\ \theta^2}{(n+1)(n+2)}.
\end{aligned}$$

Finalement

$$\begin{aligned}
\mathrm{ECM}(\hat{\theta}_3, \theta) &= \frac{(n+1)^2}{n^2}\mathrm{var}(\hat{\theta}_2) \\
&= \frac{1}{n(n+2)}\theta^2.
\end{aligned}$$

Pour $n \geq 1$, on conclut que ECM $(\hat{\theta}_3) \leq \mathrm{ECM}(\hat{\theta}_1) \leq \mathrm{ECM}(\hat{\theta}_2)$.

4. Les 3 estimateurs sont convergents. En effet

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{ECM}(\hat{\theta}_1, \theta) &= \lim_{n\to\infty} \frac{\theta^2}{3n} = 0, \\
\lim_{n\to\infty} \mathrm{ECM}(\hat{\theta}_2, \theta) &= \lim_{n\to\infty} \frac{2\theta^2}{(n+1)(n+1)} = 0, \\
\lim_{n\to\infty} \mathrm{ECM}(\hat{\theta}_3, \theta) &= \lim_{n\to\infty} \frac{\theta^2}{n(n+2)} = 0.
\end{aligned}$$

5. (a) Le taux de couverture de I_1 est

$$\begin{aligned} P(\theta \in I_1) &= P(a\hat{\theta}_2 < \theta < b\hat{\theta}_2) \\ &= P\left(\frac{\theta}{b} < \hat{\theta}_2 < \frac{\theta}{a}\right) \\ &= \left(\frac{1}{a}\right)^n - \left(\frac{1}{b}\right)^n, \end{aligned}$$

où l'on a utilisé la fonction de répartition de $\hat{\theta}_2$ calculée en 2. pour obtenir la dernière ligne ci-dessus (par la condition $1 \leq a < b$, on a bien que $\frac{\theta}{a} \leq \theta$ et $\frac{\theta}{b} \leq \theta$).

(b) Le taux de couverture de I_2 est

$$\begin{aligned} P(\theta \in I_2) &= P(\hat{\theta}_2 + c \leq \theta \leq \hat{\theta}_2 + d) \\ &= P(\theta - d \leq \hat{\theta}_2 \leq \theta - c) \\ &= \left(\frac{\theta - c}{\theta}\right)^n - \left(\frac{\theta - d}{\theta}\right)^n, \end{aligned}$$

où l'on a bien $0 \leq \theta - c \leq \theta$ et $0 \leq \theta - d \leq \theta$, car $0 \leq c < d \leq \theta$. On remarque que le taux de couverture de l'intervalle I_2 dépend de θ, alors que I_1 en est indépendant.

(c) Posons $b = 2a$. On résout

$$\begin{aligned} & \left(\frac{1}{a}\right)^n - \left(\frac{1}{2a}\right)^n = 0{,}95 \\ \Leftrightarrow \quad & \frac{1}{a^n}\left(1 - \frac{1}{2^n}\right) = 0{,}95 \\ \Leftrightarrow \quad & a^n = \frac{1}{0{,}95}\left(1 - \frac{1}{2^n}\right) \\ \Leftrightarrow \quad & a = \frac{1}{\sqrt[n]{0{,}95}} \sqrt[n]{1 - \frac{1}{2^n}} \qquad (a \geq 1). \end{aligned}$$

8.50

1. $(k-1)$ échecs avant un succès.

$$\begin{array}{lll} k-1 \text{ échecs} & : & \text{prob } = (1-\theta)\cdot(1-\theta)\ldots(1-\theta) = (1-\theta)^{k-1} \\ 1 \text{ succès} & : & \text{prob } = \theta \end{array}$$

2. La log-vraisemblance s'écrit

$$\log P(X = k) = \log\theta + (k-1)\log(1-\theta),$$

d'où l'on déduit les équations du maximum de vraisemblance pour $\hat{\theta}_{MV}$:

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log P(X = k_i) = \sum_{i=1}^{n} \left(\frac{1}{\theta} - \frac{k_i - 1}{1-\theta} \right) = \sum_{i=1}^{n} \frac{1 - k_i\theta}{\theta(1-\theta)} = 0.$$

Donc $\hat{\theta}_{MV} = \frac{1}{\overline{k}}$, où $\overline{k} = \frac{1}{n} \sum_{i=1}^{n} k_i$.

3. Par les propriétés asymptotiques de l'estimateur du maximum de vraisemblance, on a que $\mathrm{var}(\hat{\theta}_{MV}) \approx \frac{1}{n} \frac{1}{J(\theta)}$, où

$$J(\theta) = -E\left(\frac{\partial^2}{\partial\theta^2} \log P(X = k) \right)$$

est l'information de Fisher. Ici on a que $\frac{\partial^2}{\partial\theta^2} \log P(X = k) = -\frac{1}{\theta^2} - \frac{k-1}{(1-\theta)^2}$ et que

$$\begin{aligned} J(\theta) &= \frac{1}{\theta^2} + \frac{1}{(1-\theta)^2}(E[X] - 1) \\ &= \frac{1}{\theta^2} + \frac{1}{(1-\theta)^2}\left(\frac{1}{\theta} - 1 \right) = \frac{1}{\theta^2(1-\theta)}, \end{aligned}$$

et donc

$$\mathrm{var}(\hat{\theta}_{MV}) \approx \frac{\theta^2(1-\theta)}{n}.$$

4. Puisque $E(X) = \frac{1}{\theta}$ on obtient $\hat{\theta}$ en résolvant $\frac{1}{\theta} = \bar{k}$, et dans ce cas l'estimateur des moments est égal à l'estimateur du maximum de vraisemblance.
5. On calcule

$$\begin{aligned} P(X > 2) &= 1 - P(X = 1) - P(X = 2) \\ &= 1 - \theta - \theta(1-\theta) = (1-\theta)^2. \end{aligned}$$

6. $\widehat{P(X > 2)} = (1 - \hat{\theta}_{MV})^2 = \left(\frac{7}{10}\right)^2 = 0{,}49$.
7. La statistique optimale est donnée par le lemme de Neyman-Pearson :

$$\begin{aligned} T &= \prod_{i=1}^{n} \frac{P_{\theta_1}(X = k_i)}{P_{\theta_0}(X = k_i)} = \prod_{i=1}^{n} \frac{\theta_1}{\theta_0} \left(\frac{1-\theta_1}{1-\theta_0} \right)^{k_i - 1} \\ &= \left(\frac{\theta_1}{\theta_0} \right)^n \prod_{i=1}^{n} \left(\frac{1-\theta_1}{1-\theta_0} \right)^{k_i - 1}, \end{aligned}$$

ou d'une manière équivalente

$$\log(T) = n \log\left(\frac{\theta_1}{\theta_0} \right) + \left(\frac{1-\theta_1}{1-\theta_0} \right) \sum_{i=1, k_i \geq 2}^{n} \log(k_i - 1).$$

La statistique optimale est $\tilde{T}(k_1, \ldots, k_n) = \sum_{i=1}^{n} \log(k_i - 1)$, où $k_i \geq 2$, et $\hat{\theta}_{MV} = \frac{1}{k}$ n'est pas optimale.

8. On peut utiliser le théorème central limite. Pour cela il est nécessaire de calculer $E(\tilde{T})$ et $\text{var}(\tilde{T})$, et donc $E(\log(X-1))$ et $\text{var}(\log(X-1))$, où $X \geq 2$. Ceci demande le calcul des séries

$$\begin{aligned} E(\log(X-1)) &= \sum_{k=2}^{\infty} \log(k-1)\theta(1-\theta)^{k-1} \\ &= \theta \sum_{k=1}^{\infty} (\log k)(1-\theta)^k. \end{aligned}$$

et

$$\begin{aligned} E\big((\log(X-1))^2\big) &= \sum_{k=2}^{\infty} (\log(k-1))^2 \theta(1-\theta)^{k-1} \\ &= \theta \sum_{k=1}^{\infty} (\log k)^2 (1-\theta)^k. \end{aligned}$$

8.51

1. L'intervalle de confiance à 95 % pour μ vaut

$$\begin{aligned} IC(\mu, 95\ \%) &= \bar{X} \pm z_{0,975} \frac{s}{\sqrt{n}} \\ &= 21{,}5 \pm 1{,}96 \frac{1{,}8}{\sqrt{100}} \\ &= [21{,}15\,;\, 21{,}85] \end{aligned}$$

2. On veut tester

$$\begin{aligned} H_0 &: \mu = 21 \\ H_A &: \mu \neq 21. \end{aligned}$$

Puisque la valeur 21 n'appartient pas à l'intervalle de confiance dérivé en 1., on rejette H_0 au seuil de 5 %.

3. Soit π la vraie proportion de ménages qui consomment moins de 20 kg de pain par mois. On définit

$$X_i = \begin{cases} 1 & \text{si le ménage } i \text{ consomme moins de 20 kg/mois} \\ 0 & \text{sinon.} \end{cases}$$

On utilise $\hat{\pi} = \bar{x}$ pour construire l'intervalle de confiance

$$\begin{aligned} IC(\pi, 95\ \%) &= \hat{\pi} \pm z_{0,975} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \\ &= [0{,}23 \pm 1{,}96 \cdot 0{,}042] \\ &= [0{,}15\,;\, 031] \end{aligned}$$

4. On a l'intervalle de confiance

$$\begin{aligned} IC(\mu, 95\ \%) &= \bar{x} \pm z_{1-\alpha/2}\frac{s}{\sqrt{n}} \\ &= 21{,}5 \pm 1{,}96\frac{1{,}8}{\sqrt{11}} \\ &= [20{,}44\,;\,22{,}56] \end{aligned}$$

Cet intervalle est plus large que celui trouvé en 1. En effet n est plus petit, l'IC est donc moins précis.

8.52

1. La fonction de distribution d'une variable aléatoire X suivant une loi de Dagum s'obtient par la dérivation de $F_\beta(x)$ par rapport à x

$$f_\beta(x) = \frac{2\beta}{x^3}\left(1+\frac{1}{x^2}\right)^{-\beta-1}.$$

La vraisemblance d'un échantillon de taille n de variables aléatoires suivant une loi de Dagum est alors

$$L(\beta \mid x_1, \ldots, x_n) = 2^n\beta^n \prod_{i=1}^{n}\left[x_i^{-3}\left(1+\frac{1}{x_i^2}\right)^{-\beta-1}\right]$$

et sa log-vraisemblance

$$l(\beta \mid x_1, \ldots, x_n) = n\log 2 + n\log\beta - 3\sum_{i=1}^{n}\log x_i - (\beta+1)\sum_{i=1}^{n}\log\left(1+\frac{1}{x_i^2}\right).$$

On obtient l'estimateur du maximum de vraisemblance $\hat{\beta}$ de β en optimisant la log-vraisemblance

$$\frac{\partial}{\partial\beta} l(\beta \mid x_1, \ldots, x_n) = \frac{n}{\beta} - \sum_{i=1}^{n}\log\left(1+\frac{1}{x_i^2}\right)$$

$$\Rightarrow \hat{\beta} = \frac{n}{\sum_{i=1}^{n}\log\left(1+\frac{1}{x_i^2}\right)}.$$

2. Pour calculer l'information de Fisher $J(\beta)$, on dérive 2 fois la fonction $\log f_\beta(x)$ par rapport à β

$$\frac{\partial}{\partial\beta}\log f_\beta(x) = \frac{1}{\beta} - \log\left(1+\frac{1}{x^2}\right)$$

$$\Rightarrow \frac{\partial^2}{\partial\beta^2}\log f_\beta(x) = -\frac{1}{\beta^2},$$

qui donne

$$J(\beta) = \frac{1}{\beta^2}.$$

Par conséquent

$$\sqrt{n}(\hat{\beta} - \beta) \overset{n}{\rightsquigarrow}{}^{\infty} N(0,\beta^2).$$

3. On a généré 100 000 échantillons de taille $n = 16$ avec $\beta = 1$. Calculons la probabilité d'obtenir une valeur de l'estimateur $\hat{\beta}$ comprise entre 0,5 et 1,5

$$\begin{aligned} P(0{,}5 < \hat{\beta} < 1{,}5) &= P\left(\frac{\sqrt{n}(0{,}5-\beta)}{\sqrt{\mathrm{var}\hat{\beta}}} < \frac{\sqrt{n}(\hat{\beta}-\beta)}{\sqrt{\mathrm{var}\hat{\beta}}} < \frac{\sqrt{n}(1{,}5-\beta)}{\sqrt{\mathrm{var}\hat{\beta}}}\right) \\ &\overset{\beta=1}{=} P\left(-\frac{1}{8} < \frac{\sqrt{n}(\hat{\beta}-\beta)}{\sqrt{\mathrm{var}\hat{\beta}}} < \frac{1}{8}\right) \\ &\overset{n \text{ grand}}{=} 2\Phi\left(\frac{1}{8}\right) - 1 \simeq 0{,}1. \end{aligned}$$

On aura approximativement 10 000 échantillons où la valeur de l'estimateur sera comprise dans l'intervalle $[0{,}5\,;\ 1{,}5]$.

4. On peut écrire

$$\sqrt{n}\left(\frac{\hat{\beta}}{\beta} - 1\right) = \frac{1}{\beta}\sqrt{n}(\hat{\beta} - \beta).$$

Donc, $\sqrt{n}\left(\frac{\hat{\beta}}{\beta} - 1\right)$ suit approximativement une loi normale d'espérance nulle et de variance 1. Construisons un intervalle de confiance au degré 95 % pour β

$$P\left(-z_{1-\alpha/2} < \sqrt{n}\left(\frac{\hat{\beta}}{\beta} - 1\right) < z_{1-\alpha/2}\right) = \alpha,$$

où $z_{1-\alpha/2}$ est le $\alpha/2$-quantile de la loi normale centrée et réduite. On déduit donc de l'inégalité précédente

$$\mathrm{IC} = \left[\frac{\sqrt{n}\hat{\beta}}{\sqrt{n} \pm z_{1-\alpha/2}}\right].$$

8.53

1. Calculons l'espérance de X

$$E(X) = \int_0^1 \alpha x^{\alpha} dx = \frac{\alpha}{\alpha+1}\ x^{\alpha+1}\Big|_{x=0}^{x=1} = \frac{\alpha}{\alpha+1}.$$

L'estimateur des moments $\hat{\alpha}_M$ de α est donc

$$\hat{\alpha}_M = \frac{\overline{X}}{1-\overline{X}} = g(\overline{X}).$$

2. Pour connaître la variance de X, il faut chercher son 2^e moment

$$E(X^2) = \int_0^1 \alpha x^{\alpha+1} dx = \frac{\alpha}{\alpha+2}\, x^{\alpha+2}\Big|_{x=0}^{x=1} = \frac{\alpha}{\alpha+2}.$$

On en déduit

$$\text{var}(X) = E(X^2) - E^2(X) = \frac{\alpha}{(\alpha+2)(\alpha+1)^2}.$$

Finalement

$$\sqrt{n}(\overline{X} - \mu) \sim \mathcal{N}\left(0\,, \frac{\alpha}{(\alpha+2)(\alpha+1)^2}\right)$$

car $\overline{X}$ n'est pas biaisé.

3. La dérivée de la fonction $g(\mu)$ par rapport à μ est

$$\frac{\partial}{\partial\mu} g(\mu) = \frac{\partial}{\partial\mu}\frac{\mu}{1-\mu} = \frac{\partial}{\partial\mu}\left(\frac{1}{1-\mu} - 1\right) = \frac{1}{(1-\mu)^2}.$$

La variance approximée de $\hat{\alpha}_M$ est par conséquent

$$\text{var}(\hat{\alpha}_M) \simeq \left(\frac{\partial g(\mu)}{\partial\mu}\right)^2 \frac{\sigma^2}{n} = \left[\frac{1}{(1-\mu)^2}\right]^2 \frac{\sigma^2}{n} = \frac{\alpha}{n(\alpha+2)}.$$

4. Pour tester la convergence de l'estimateur $\hat{\alpha}_M$, calculons sa limite en probabilité

$$\text{plim}_{n\to\infty}\frac{\overline{X}}{1-\overline{X}} = \frac{\text{plim}_{n\to\infty}\overline{X}}{1-\text{plim}_{n\to\infty}\overline{X}} = \frac{\frac{\alpha}{\alpha+1}}{1-\frac{\alpha}{\alpha+1}} = \alpha.$$

L'estimateur est convergent.

5. L'estimateur est efficace si sa variance atteint la borne de Cramér-Rao. Il faut donc calculer cette dernière. Reprenons la fonction $\log f_\alpha(x)$ et dérivons-la à 2 reprises par rapport à α

$$\frac{\partial}{\partial\alpha}\log f_\alpha(x) = \frac{\partial}{\partial\alpha}\left(\log\alpha + (\alpha-1)\log x\right) = \frac{1}{\alpha} + \log x$$

$$\Rightarrow \frac{\partial^2}{\partial\alpha^2}\log f_\alpha(x) = -\frac{1}{\alpha^2}.$$

On en déduit l'information de Fisher $J(\alpha)$

$$J(\alpha) = \frac{1}{\alpha^2},$$

et la borne de Cramér-Rao pour un estimateur sans biais de α

$$\text{BCR} = \frac{\alpha^2}{n}.$$

Comparons-la à la variance de $\hat{\alpha}_M$

$$\text{var}(\hat{\alpha}_M) - \text{BCR} = \frac{\alpha^2}{n} - \frac{\alpha}{n(\alpha+2)} = \frac{\alpha^3+1}{n(\alpha+2)}.$$

Ce résultat est toujours positif car α est positif. Ainsi, l'estimateur n'atteint pas la borne ; il n'est pas efficient.

6. Par le théorème central limite

$$\frac{\sqrt{n}(\hat{\alpha}_M - \alpha)}{\sqrt{\text{var}(\hat{\alpha}_M)}} \sim N(0,1).$$

On peut donc construire un intervalle de confiance approximé au degré 95 % pour α :

$$P\left(-z_{0,975} < \frac{\sqrt{n}(\hat{\alpha}_M - \alpha)}{\sqrt{\text{var}(\hat{\alpha}_M)}} < z_{0,975}\right) = 0{,}95$$

$$\Leftrightarrow \text{IC} = \left[\hat{\alpha}_M \pm z_{0,975}\sqrt{\frac{\text{var}(\hat{\alpha}_M)}{n}}\right].$$

8.54

1. Le statistique $\bar{X}$ donne le test le plus puissant lorsque $X_i \sim \mathcal{N}(\mu, \sigma^2)$ (cf. lemme de Neyman-Pearson).
2.

$$\begin{aligned} pv &= P_{H_0}(\bar{X} > \bar{x}) \\ &= P_{H_0}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) \\ &= P_{H_0}\left(Z < \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) \\ &= 1 - P_{H_0}\left(Z < \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) \\ &= 1 - \Phi\left(\frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}\right) = g(\bar{x}, \mu_0, \sigma, n) \end{aligned}$$

3. D'après 1., on a que $PV = 1 - \Phi\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right)$. Donc

$$\begin{aligned}
P_\mu(PV > a) &= P_\mu(1 - \Phi\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right) > a) \\
&= P_\mu\left(\Phi\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}\right) < 1-a\right) \\
&= P_\mu\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < \Phi^{-1}(1-a)\right) \\
&= P_\mu\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < \Phi^{-1}(1-a) + \frac{\mu_0}{\sigma/\sqrt{n}}\right) \\
&= P_\mu\left(Z < \Phi^{-1}(1-a) + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(\Phi^{-1}(1-a) + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

4. Lorsque $\mu_1 = \mu_0$, la distribution de PV devient

$$\begin{aligned}
P_{\mu_1}(PV < a) &= 1 - P_\mu(PV > a) \\
&= 1 - \Phi\left(\Phi^{-1}(1-a)\right) \\
&= 1 - (1-a) \\
&= a,
\end{aligned}$$

ce qui veut dire que la distribution de PV est uniforme sur l'intervalle $(0, 1)$.

Bibliographie

[1] Sheldon M. Ross, *Initiation aux probabilités*. Presses polytechniques romandes, Lausanne, 1987.

[2] Jim Pitman, *Probability*. Springer, New York, 1993.

[3] Michel Lejeune, *Statistique La Théorie et ses applications*. Springer, Paris, 2004.

[4] Stephan Morgenthaler, *Introduction à la statistique*. Presses polytechniques et universitaires romandes, Lausanne, 2e édition, 2001.

[5] Yadolah Dodge, *Premiers pas en statistique*. Springer, Paris, 2003.

[6] Elvezio Ronchetti, Gabrielle Antille et Maurice Polla, *STEP I - STatistique Et Probabilités : une Introduction*. Presses polytechniques et universitaires romandes, Lausanne, 2e édition, 1991.

[7] Yves Tillé, *Théorie des sondages : échantillonnage et estimation en population finie : cours et exercices avec solutions*. Dunod, Paris, 2001.

[8] George Casella et Roger L. Berger, *Statistical Inference*. Duxbury Press, Pacific Grove (CA), 1990.

[9] John A. Rice, *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont (CA), 2e édition, 1995.

[10] David S. Moore et George P. McCabe, *Introduction to the Practice of Statistics*. W.H. Freeman, New York, 4e édition, 2003.

Index